Inference in High-Dimensional Panel Models:
Two-Way Dependence and Unobserved Heterogeneity

Kaicheng Chen

Department of Economics
Michigan State University

Midwest Econometrics Group
Lexington, KY
Nov 2, 2024

Table of Contents

1 Introduction

2 Two-Way Cluster LASSO

3 Panel-DML

4 Unobserved Heterogeneity

5 Discussion

Motivation: Why High Dimensionality Matters in Economics and Panel Models?

- High dimensionality: a **large number of parameters** relative to the sample.
- Three common scenarios:
  - **Many potentially relevant variables**: e.g., provisions in trade agreements, price of relevant goods.
  - **Nonparameric or semiparametric modeling**: example
  - **Heterogeneity**: e.g. heterogeneous effects across units or covariates(non-homothetic preference; relaxing/testing parallel trend assumption), unobserved heterogeneous effects in nonlinear models.
- Common to consider unobserved heterogeneity in **panel data models**, making high dimensionality a practical issue rather than just a theoretical concern.
- Research is limited in high-dimensional methods for panel data models.

Preview of Results

- For a high-dimensional regression model, I propose a (post) **LASSO** approach, robust to **two-way cluster dependence** in panel data.
- Due to the high dimensionality and the cluster dependence driven by the underlying components, the **rate of convergence is slow** even in the oracle case.
- Nonetheless, using a **panel sample-splitting/cross-fitting** approach, it is possible to establish inferential theory on low-dimensional treatment parameters.
- Both the LASSO and cross-fitting for panel data are of independent interests.

Preview of Results

- Together, they extend the **double/debiased machine learning**(DML) approach to panel data models.
- Additionally, I address the challenges posed by **unobserved heterogeneity**: a subtle issue for cross-fitting.
- In a panel data application, high dimensionality can be hidden. The proposed toolkit enables **flexible modeling** and **robust inference**.

Example 1: The Hidden High Dimensionality

- Consider the **estimation of a multiplier**: the percentage increase in output that results from the 1 percent increase in government spending.
- Consider a **panel data framework** by Nakamura and Steinsson (2014): one of the most cited empirical-macro papers on AER.
- Identification is through IV and only a few control variables are considered.
- Why is high dimensionality relevant here?

Example 1: The Hidden High Dimensionality

- Researchers often start with a baseline model:

$$Y_{it} = \theta_0 D_{it} + \pi X_{it} + c_i + d_t + U_{it},\ \mathrm{E}[Z_{it} U_{it}] = 0$$

- More robustness in function forms and identification:

$$Y_{it} = \theta_0 D_{it} + g(X_{it}, c_i, d_t) + U_{it}.$$

- Cost: noisy or infeasible estimations with limited sample sizes (51 states with 39 periods).
- Approaches that allow for high dimensionality and robust to dependence in panel data?

Yes! But Challenge One...

- We need a regularized estimator, valid for panel data models.
- For LASSO, convergence rates can be derived under **Gaussian error** (Bickel et al. (2009); Chetverikov et al. (2021)).
- Belloni et al. (2012): normality is not essential and instead utilizing **self-normalization penalty weights**.
- Their approach along with its extensions (Belloni et al., 2016) breaks down under two-way dependence in the panel.
- Analysis of **regularization methods with dependent data** is limited.
- My solution: a (post) LASSO estimator using **two-way robust penalty weights**.

Okay! But Challenge Two...

- The rates of convergence are slow with the underlying components: slower than $O_P(1/\sqrt{N \wedge T})$.
- Such slow convergence makes inference challenging.
- In the example, a two-step estimation using LASSO first-step requires $o_P(1/\sqrt{N \wedge T})$ (sufficient) for inference.
- Off-the-shelf (bias-correction) inference procedures may not be sufficient or valid: e.g. the low-dimensional projection in Zhang and Zhang (2014), the de-sparsification in Van de Geer et al. (2014), the decorrelating matrix adjustment in Javanmard and Montanari (2014), the decorrelated score in Ning and Liu (2017), the double lasso in Belloni et al. (2014), the double/debiased machine learning (DML) in Chernozhukov et al. (2018), Chiang et al., 2022, etc.
- My solution: a **new cross-fitting** procedure robust to two-way cluster dependence in panel. The inferential theory is **possible with a slow rate of convergence**.

Fair! But One More Challenge...

- Complication due to **unobserved heterogeneous effects**: endogeneity, high dimensionality, etc.
- Common approaches, e.g. within-transformation, fixed effects, and Mundlak device, carry sample averages into regressors.
- Regressors carrying the **history of the data** introduce dependence to cross-fitting sub-samples.
- Without cross-fitting, asymptotic normality is difficult in general, not just under two-way dependence.

My Solution

- **Feasible conditions**: linear additive unobserved heterogeneous effects; sub-sample Mundlak device.
- **Without cross-fitting**, a sufficient rate of convergence for the nuisance estimator is derived in a partial linear model: $o_P(1/\sqrt{N \wedge T})$.
- Still an **open question** for inference when cross-fitting is not valid and two-way cluster dependence is present.
- An alternative idea is to project out the underlying components: illustrated in special cases.

The Rest of the Talk

- More details on the **two-way cluster-LASSO** and the underlying issue of the slow convergence.
- **Panel cross-fitting** algorithm.
- The subtle issue of the **unobserved heterogeneity**.
- Overall strategies under various scenarios and the implications.

Table of Contents

1 Introduction

2 Two-Way Cluster LASSO

3 Panel-DML

4 Unobserved Heterogeneity

5 Discussion

- Consider the following high-dimensional regression model:

$$Y_{it} = f(X_{it}) + V_{it}, \ E[V_{it}|X_{it}] = 0.$$

- I take a sparse approximation approach as in Belloni et al. (2012, 2014, 2016):

$$f(X_{it}) = f_{it}\zeta_0 + r_{it}.$$

- Rewrite the model as

$$Y_{it} = f_{it}\zeta_0 + r_{it} + V_{it}, \ E[V_{it}|X_{it}] = 0.$$

- Apply $l1$ regularization:

$$\tilde{\zeta} = \arg\min_{\zeta} \frac{1}{NT} \sum_{i=1}^{N} \sum_{t=1}^{T} (Y_{it} - f_{it}\zeta)^2 + \frac{\lambda}{NT} \|\omega^{1/2}\zeta\|_1,$$

where ω is some $p \times p$ diagonal matrix of penalty weights.

Two-way cluster dependence

- **Assumption 1**(AHK representation): Random elements $W_{it} = (Y_{it}, X_{it}, V_{it})$ are generated by the underlying process:

$$W_{it} = \mu + h(\alpha_i, \gamma_t, \varepsilon_{it}), \quad \forall i \geq 1, t \geq 1,$$

  where $\mu = E[W_{it}]$; h is unknown; vector components $(\alpha_i)_{i\geq 1}$, $(\gamma_t)_{t\geq 1}$, and $(\varepsilon_{it})_{i\geq 1, t\geq 1}$ are mutually independent; α_i is i.i.d across i, ε_{it} is i.i.d across i and t, and γ_t is strictly stationary.

- **Assumption 2** (*beta*-mixing of $\{\gamma_t\}_{t>1}$)

Regularization Event

- Choose λ and ω for the event:

$$\max_{j=1,\ldots,p} \left| \frac{1}{NT} \sum_{i=1}^{N} \sum_{t=1}^{T} \omega_j^{-1/2} f_{it,j} V_{it} \right| \leq \frac{\lambda}{2c_1 NT}.$$

- (Conditional) Gaussian or sub-Gaussian errors?.
- Moderate deviation theorem for self-normalized sums?
- I decompose $f_{it,j} V_{it} = a_{i,j} + g_{t,j} + e_{it.j}$ where

$$\begin{aligned} a_{i,j} :=& \mathrm{E}[f_{it,j} V_{it} | \alpha_i], \; g_{t,j} := \mathrm{E}[f_{it,j} V_{it} | \gamma_t], \\ e_{it.j} =& f_{it,j} V_{it} - a_{i,j} - g_{t,j}. \end{aligned}$$

- Rely on a moderate deviation theorems for weakly dependent random variables due to Gao et al. (2022).

My Construction of Weights

- I consider the following choice of penalty level λ and (infeasible) penalty weights ω: for each $j = 1, ..., p$,

$$\lambda = 6c_1 \frac{NT}{(N \wedge T)^{1/2}} \Phi^{-1}\left(1 - \frac{\gamma}{2p}\right),$$

$$\omega_j = \frac{N \wedge V}{N^2} \sum_{i=1}^{N} a_{i,j}^2 + \frac{N \wedge V}{T^2} \sum_{b=1}^{B} \left(\sum_{t \in H_b} g_{t,j}\right)^2.$$

- B is the number of blocks; H_b is an index set for block b.
- γ is a tuning parameter sufficiently small: $\log(1/\gamma) \simeq \log(p \vee NT)$.
- Asymptotically equivalent (point-wise) feasible weights.

Main Theorem 1 and Slow Rate of Convergence

- **Theorem 1**: Given desired tuning parameters and other regular conditions, the l^2 rate of convergence for the two-way cluster-LASSO is $O_P\left(\sqrt{\frac{s\log(p\vee NT)}{N\wedge T}}\right)$. Comparison.
- Consider the sample mean estimator $\widehat{\theta} = \frac{1}{NT}\sum_{i=1}^{N}\sum_{t=1}^{T} Y_{it}$. Rewrite the estimator through a Hajek projection:

$$\widehat{\theta} - \theta_0 = \frac{1}{N}\sum_{i=1}^{N} a_i + \frac{1}{T}\sum_{t=1}^{T} g_t + \frac{1}{NT}\sum_{i=1}^{N}\sum_{t=1}^{T} e_{it},$$

  where $a_i := \mathrm{E}[Y_{it} - \theta_0|\alpha_i], g_t := \mathrm{E}[Y_{it} - \theta_0|\gamma_t]$, and $e_{it} := Y_{it} - \theta_0 - a_i - g_t$.
- Under some regularity conditions, for each j, $\widehat{\theta}_j = O_P\left(\frac{1}{\sqrt{N\wedge T}}\right)$ and $\|\widehat{\theta} - \theta_0\|_2 = O_P\left(\sqrt{\frac{s}{N\wedge T}}\right)$.

Table of Contents

1 Introduction

2 Two-Way Cluster LASSO

3 Panel-DML

4 Unobserved Heterogeneity

5 Discussion

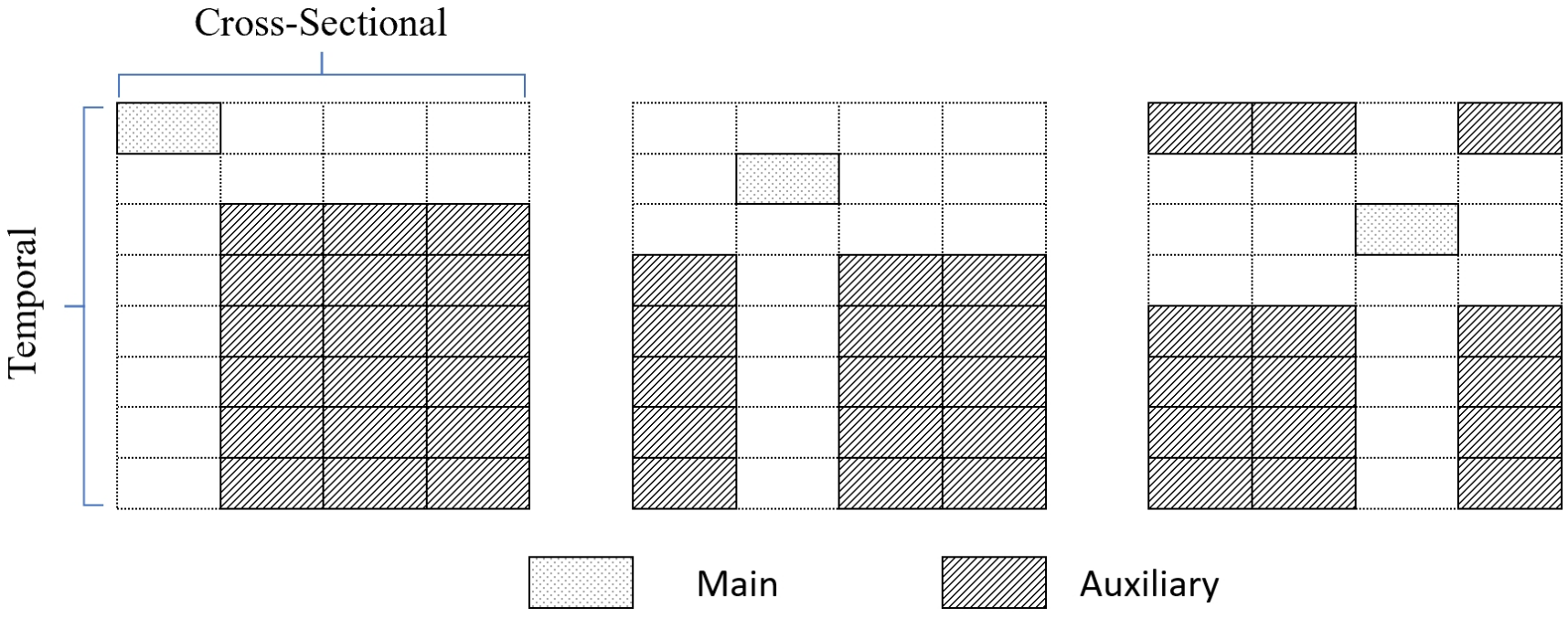
Inference with High-Dimensional Nuisance Estimation

- For example,

$$Y = \theta_0 D + g_0(X) + U,\ \mathrm{E}[U|D, X] = 0.$$

- Orthogonalized score, $(D - \mathrm{E}[D|X])(Y - \theta_0 D - g_0(X))$, is not sufficient for asymptotic normality in general.
- To further relax the rate requirement, I consider a sample-splitting/cross-fitting approach.
- Splitting the sample in a suitable way to eliminate the dependence between the two steps. Intuitively, it overcomes over-fitting.
- However, existing sample-splitting/cross-fitting procedure breaks down with two-way clustered panel.
- Inspired by Chiang et al. (2022) and Semenova et al. (2023), I propose the following cross-fitting scheme:

Cross-Fitting: My Proposal



Lemma 1 (validity of the cross-fitting): Under Assumptions 1 (AHK) and 2 (beta-mixing), the cross-fitting sub-samples are "approximately" independent as $N, T \to \infty$ with $\log(N)/T \to 0$.

panel-DML algorithm

Main Theorem 2 and Intuition

- **Theorem 2**: Given L^2 rate of convergence for the first-step, $o\left((N \wedge T)^{-1/4}\right)$, and regular conditions, $\sqrt{N}\left(\widehat{\theta} - \theta_0\right) \Rightarrow \mathrm{N}(0,1)$.
- Consistent estimation of the asymptotic variance
- A slower rate requirement than prototypical DML.
- $\sqrt{N}$-consistent instead of $\sqrt{NT}$-consistent.
- Cluster dependence introduced by the unit and time components does not decay over time or space: the information carried by data is accumulated slower.
- A common feature in cluster robust inference. Also see the inferential theory under strong cross-sectional dependence as in Gonçalves (2011).

Table of Contents

1 Introduction

2 Two-Way Cluster LASSO

3 Panel-DML

4 Unobserved Heterogeneity

5 Discussion

A Partial Linear Model (with Endogeneity)

- Consider the following partial linear model:

$$Y_{it} = D_{it}\theta_0 + g(X_{it}, c_i, d_t) + U_{it},\ \mathrm{E}[U_{it}|X_{it}, c_i, d_t] = 0.$$

- Z_{it} has the same dimension of D_{it}; $\mathrm{E}[Z_{it}U_{it}] = 0$. As a special case, $Z_{it} = D_{it}$.
- All three sources of high dimensionality
- (c_i, d_t) are in general different objects from (α_i, γ_t).
- The (infeasible) orthogonal moment is given by

$$\mathrm{E}\left(Z_{it} - \mathrm{E}\left[Z_{it}|X_{it}, c_i, d_t\right]\right) \times \left(Y_{it} - \mathrm{E}\left[Y_{it}|X_{it}, c_i, d_t\right] - \theta_0\left(D_{it} - \mathrm{E}\left[D_{it}|X_{it}, c_i, d_t\right]\right)\right) = 0.$$

Existing Approaches

- Suppose the unknown function g is linear in c_i and d_t:

$$Y_{it} = D_{it}\theta_0 + g(X_{it}) + c_i + d_t + U_{it}.$$

- Then, a sparse approximation and within-transformation would do the trick. (Belloni et al., 2016).
- Alternatively, fixed-effect approach with sparsity assumption on the fixed effects. (Kock and Tang, 2019).

CRE approach

- Generalized Mundlak device:

$$c_i = h_c(\bar{X}_i) + \epsilon_i^c, \ \bar{X}_i \text{ is sample average over t},$$
$$d_t = h_d(\bar{X}_t) + \epsilon_t^d, \ \bar{X}_t \text{ is sample average over i},$$

- $(\epsilon_i^c, \epsilon_t^d)$ are independent of $\mathbf{X}$.
- Not as strong as it seems: the equivalence exists among within-transformation, fixed-effect, and Mundlak device with POLS in a linear model (Mundlak, 1978; Wooldridge, 1997).
- Generalized by a flexible function. Also see Wooldridge and Zhu (2020).
- The proxy is more sensible with high dimensionality in X.

A Subtle Issue

- Now it seems like one can simply apply the two-way cluster LASSO with cross-fitting.
- However, the Mundlak device uses the full history of the covariates, introducing dependence across the cross-fitting sub-samples.
- If g is linear in c_i and d_t, then a within-transformation in sub-samples would work.
- One could assume the generalized Mundlak device holds in each sub-sample: For each $i \in I_k$ and $t \in S_l$,

$$c_i = h_c(\bar{X}_{i,l}) + \epsilon^c_{i,l},$$
$$d_t = h_d(\bar{X}_{t,k}) + \epsilon^d_{t,k},$$

where $(\epsilon^c_{i,l}, \epsilon^d_{t,k})$ are independent of $\mathbf{X}$; $\bar{X}_{t,k}$ and $\bar{X}_{i,l}$ are sub-sample unit and time averages.

Full-Sample Mundlak Approximation?

- Alternatively, one could maintain full-sample Mundlak assumption but, for each $(i,t) \in W(-k,-l)$, approximate $(\bar{X}_t, \bar{X}_i)$ using $(\bar{X}_{t,k}, \bar{X}_{i,l})$, the sub-sample averages.
- The sub-sample averages use a big portion of the full-sample so the difference between the sub-sample and full-sample averages should vanish as the sample sizes (N,T) grow.
- However, $(\bar{X}_i, \bar{X}_t)$ are also high-dimensional vectors and the approximation error does not vanish fast enough.

Without Cross-Fitting

- Without cross-fitting, not clear if valid inference in possible in general.
- Recall that the cross-fitting is used to relax the rate requirement and sparsity condition.
- What is the rate requirement without cross-fitting?
- **Theorem 3**: The asymptotic normality can be established in the partial linear model without cross-fitting, with a sufficient rate requirement $o_P\left((N \wedge T)^{-1/2}\right)$ on the nuisance parameters.
- Not achievable in general under two-way cluster dependence and high-dimensionality.
- Alternative idea is to project-out the underlying components to remove the two-way cluster dependence.

Summary of Various Scenarios

Table: Summary of Results

Conditions on (c_i, d_t)	Approaches for (c_i, d_t)	First-Step Estimation	First-Step Sufficient Rates	Sparsity TW CL LASSO
Not Present	NA	Panel CF	$o\left((N \wedge T)^{-1/4}\right)$ in $L^2(P)$ norm	$s = o\left(\frac{(N \wedge T)^{1/2}}{\log(p \vee NT)}\right)$
Linear	Within-Transformation	Panel CF	$o\left((N \wedge T)^{-1/4}\right)$ in $L^2(P)$ norm	$s = o\left(\frac{(N \wedge T)^{1/2}}{\log(p \vee NT)}\right)$
Non-Additive	Sub-sample Mundlak	Panel CF	$o\left((N \wedge T)^{-1/4}\right)$ in $L^2(P)$ norm	$s = o\left(\frac{(N \wedge T)^{1/2}}{\log(p \vee NT)}\right)$
Non-Additive	Full-sample Mundlak	Panel CF	Not Valid	NA
Non-Additive	Full-sample Mundlak	Full Sample	$o_P\left((N \wedge T)^{-1/2}\right)$ in l^2 norm	Not Achievable

Table of Contents

1 Introduction

2 Two-Way Cluster LASSO

3 Panel-DML

4 Unobserved Heterogeneity

5 Discussion

Summary

- The inferential theory for high-dimensional models is particularly relevant in panel data setting.
- This paper enriches the toolbox of researchers in dealing with high-dimensional panel models.
- I develop a high-dimensional estimator for panel data models and a general inference procedure for semiparametric models with high dimensional controls and two-way dependence.
- Unobserved heterogeneity complicates the inference. I provide different strategies under various conditions.
- I illustrate in a multiplier estimation that high dimensionality can be hidden and how proposed approaches can enrich the analysis as a robustness check.

High Dimensionality from Flexible Modeling

- Suppose X is $k \times 1$. Let L^τ be $\tau-$th order polynomial transformation and let r denote the approximation error.
- Then, the high dimensionality is realized as follows:

model	sparse approx.	dim. of unknown param.
$Y = f(X) + U$	no approx.	∞,
$Y = L^\tau(X)\beta + r + U$	$\tau = 2$	$k^2/2 + 3k/2$
$Y = L^\tau(X)\beta + r + U$	$\tau = 3$	$k^3/6 + k^2 + 11k/6$

Back

Panel-DML: Orthogonalized Moment Condition

- Let $\varphi(W_{it}; \theta, \eta)$ be an identifying moment condition:

$$E[\varphi(W_{it}; \theta_0, \eta_0)] = 0$$

  where W_{it} are random elements; θ are the low-dimensional parameters of interest and η are nuisance parameters.

- Let $\psi(W_{it}; \theta, \eta)$ be a corresponding orthogonal moment condition such that

$$\begin{aligned} E[\psi(W; \theta_0, \eta_0)] &= 0, \\ \partial_\eta E[\psi(W; \theta_0, \eta_0)][\eta - \eta_0] &= 0. \end{aligned}$$

Back

Implementation

Definition (Panel-DML algorithm)

(i) Given identifying moment functions $\varphi(W;\theta,\eta)$, find the orthogonalized moment functions $\psi(W,\theta,\eta)$.

(ii) Select (K,L). Randomly partition the cross-sectional indices $\{1,2,...,N\}$ into K equal-size index sets $\{I_1,I_2,...,I_K\}$ and partition the temporal indices $\{1,2,...,T\}$ into L adjacent equal-size index sets $\{S_1,S_2,...,S_L\}$ [a]. Construct the main sample $W(k,l)=\{W_{it}: i\in I_k, t\in S_l\}$, and the auxiliary sample

$$W(-k,-l)=\left\{W_{it}: i\in \bigcup_{k'\neq k} I_{k'}, t\in \bigcup_{l'\neq l,l\pm 1} S_{l'}\right\}.$$

[a] assume N, T are divisible by K, L for simplicity.

Implementation

Definition (Panel-DML algorithm (continues))

(iii) For each k and l, use the sample $W(-k,-l)$ for first step estimation and obtain $\widehat{\eta}_{k,l}$, then construct the averages of scores $\overline{\psi}_{k,l}(\theta) = \frac{1}{N_k \times T_l} \sum_{i \in I_k, t \in S_l} \psi(W_{it}; \theta, \widehat{\eta}_{k,l})$ for all (k,l).

(iv) Obtain the DML estimator $\widehat{\theta}$ as the solution to

$$\frac{1}{KL} \sum_{i=1}^{K} \sum_{l=1}^{L} \overline{\psi}_{k,l}(\theta) = 0.$$

Back

Absolute Regularity

Let $\|\nu\|_{TV}$ denote the total variation norm of a signed measure ν on a measurable space (S, Σ) where Σ is a σ-algebra on S:

$$\|\nu\|_{TV} = \sup_{A \in \Sigma} \nu(A) - \nu(A^c)$$

Define the dependence coefficient of X and Y as:

$$\beta(X, Y) = \frac{1}{2}\|P_{X,Y} - P_X \times P_Y\|_{TV}$$

Assumption 2 (Absolute Regularity of $\{\gamma_t\}_{t \geq 1}$)

The sequence $\{\gamma_t\}_{t \geq 1}$ is beta-mixing at a geometric rate:

$$\beta_\gamma(q) = \sup_{s \leq T} \beta\left(\{\gamma_t\}_{t \leq s}, \{\gamma_t\}_{t \geq s+q}\right) \leq c_\kappa exp(-\kappa q), \forall q \in \mathbb{Z}^+,$$

for some constants $\kappa > 0$ and $c_\kappa \geq 0$.

Back

Cross Fitting Validity

Lemma (Independent Coupling)

Consider the main sample $W(k,l)$ and auxiliary sample $W(-k,-l)$ for $k = 1, ..., K$ and $l = 1, ..., L$. Suppose Assumptions 1-2 hold for $\{W_{it}\}$. Then, if $\log N/T \to 0$ as $N, T \to \infty$, we can construct $\tilde{W}(k,l)$ and $\tilde{W}(-k,-l)$ such that:

- *They are independent of each other;*
- *They have the same marginal distribution as $W(k,l)$ and $W(-k,-l)$, respectively;*

and
$$Pr\left\{(W(k,l), W(-k,-l)) \neq \left(\tilde{W}(k,l), \tilde{W}(-k,-l)\right), \text{ for some } (k,l)\right\} = o(1)$$

Back

Variances and Non-Degeneracy

Define
$$a_i := E\left[\psi\left(W_{it};\theta_0,\eta_0\right)|\alpha_i\right],$$
$$g_t := E\left[\psi\left(W_{it};\theta_0,\eta_0\right)|\gamma_t\right],$$
$$e_{it} := \psi\left(W_{it};\theta_0,\eta_0\right) - a_i - g_t.$$

and
$$\Lambda_a\Lambda_a' = E[a_i a_i']$$
$$\Lambda_g\Lambda_g' = \sum_{l=-\infty}^{\infty} E[g_t g_{t+l}']$$
$$\Lambda_e\Lambda_e' = \sum_{l=-\infty}^{\infty} E[e_{it} e_{i,t+l}']$$

Let $\lambda_{min}[.]$ denote the smallest eigenvalue of a square matrix.

Assumption 3 (Non-degeneracy)

Either $\lambda_{min}[\Lambda_a\Lambda_a'] > 0$ *or* $\lambda_{min}[\Lambda_g\Lambda_g'] > 0$.

Assumption 4 (Linear Orthogonal Scores)

For any $P \in \mathcal{P}_n$, the following conditions hold:

(i) $\psi(W;\theta,\eta) = \psi^a(W,\eta)\theta + \psi^b(W,\eta),\ \forall\ W \in \mathcal{W}, \theta \in \Theta, \eta \in \mathcal{T}$.

(ii) $\psi(W;\theta,\eta)$ *satisfy the Neyman orthogonality conditions, or more generally, by a* λ_n *near-orthogonality condition:* $\lambda_n := \sup_{\eta\in\mathcal{T}_n} \|\partial_r E[\psi(W;\theta_0,\eta_0 + r(\eta-\eta_0))]|_{r=0}\| \leq \delta_n/\sqrt{N}$, *where* $\mathcal{T}_n \in \mathcal{T}$ *is a nuisance realization set.*

(iii) *The map* $\eta \to E_P[\psi(W_{it};\theta,\eta)]$ *is twice continuously Gateaux-differentiable on* $\mathcal{T}$.

(iv) *The singular values of the matrix* $A_0 := E_P[\psi^a(W_{it};\eta_0)]$ *are bounded between* c_0 *and* c_1.

Back

Assumption 5 (Statistical Rates and Score Regularity)

For some $(\Delta_n)_{n\geq 1}$ *that* $\Delta_n > 0$ *and* $\Delta_n \to 0$, *we have*

(i) *For each* (k, l), *the nuisance estimator* $\widehat{\eta}_{k,l}$ *belongs to the realization set* $\mathcal{T}_n$ *with probability* $1 - \Delta_n$, *where* $\mathcal{T}_n$ *contains* η_0.

(ii) *For all* $i \geq 1$, $t \geq 1$, *and some* $p > 2$, *the following moment conditions hold:*

$$m_n := \sup_{\eta \in \mathcal{T}_n} \left(E_P \left\|\psi(W_{it}; \theta_0, \eta)\right\|^p\right)^{1/p} \leq c_1, \tag{1}$$

$$m'_n := \sup_{\eta \in \mathcal{T}_n} \left(E_P \left\|\psi^a(W_{it}; \eta)\right\|^p\right)^{1/p} \leq c_1. \tag{2}$$

Assumption 5 (Statistical Rates and Score Regularity)

(iii) *The following conditions on the statistical rates r_n, r_n', λ_n' hold for all $i \geq 1$, $t \geq 1$:*

$$r_n := \sup_{\eta \in \mathcal{T}_n} \|E_P[\psi^a(W_{it}; \eta) - \psi^a(W_{it}; \eta_0)]\| \leq \delta_n,$$

$$r_n' := \sup_{\eta \in \mathcal{T}_n} \left(E_P \|\psi(W_{it}; \theta_0, \eta) - \psi(W_{it}; \theta_0, \eta_0)\|^2\right)^{1/2} \leq \delta_n,$$

$$\lambda_n' := \sup_{r \in (0,1), \eta \in \mathcal{T}_n} \left\|\partial_r^2 E_P[\psi(W_{it}; \theta_0, \eta_0 + r(\eta - \eta_0))]\right\| \leq \delta_n / \sqrt{N}.$$

Back

Variance Estimator

I consider two variance estimators $\widehat{V}_{CHS}$ and $\widehat{V}_{DKA}$ which are derived under two-way dependence in low-dimensional panel model and here adjusted for the cross-fitting. formulas

Theorem 2

Suppose that Assumptions of Theorem 1 hold for $P = P_{NT}$ for each (N, T) with some $p > 4$ (higher moments) and $M/T^{1/2} = o(1)$. Then, as $N, T \to \infty$, we have,

$$\widehat{V}_{CHS} = V + o_P(1),$$
$$\widehat{V}_{DKA} = \widehat{V}_{CHS} + o_P(1).$$

Back

Variance Estimators

$$\widehat{V}_{CHS} = \widehat{A}^{-1}\widehat{\Omega}_{CHS}\widehat{A}^{-1'}, \qquad \widehat{V}_{DKA} = \widehat{A}^{-1}\widehat{\Omega}_{DKA}\widehat{A}^{-1'}$$
$$\widehat{\Omega}_{CHS} = \widehat{\Omega}_A + \widehat{\Omega}_{DK} - \widehat{\Omega}_{NW}, \quad \widehat{\Omega}_{DKA} = \widehat{\Omega}_A + \widehat{\Omega}_{DK}.$$

where $\widehat{A} = \frac{1}{KL}\sum_{k=1}^{K}\sum_{l=1}^{L}\frac{1}{N_k T_l}\sum_{i\in I_k, s\in S_l}\psi^a(W_{it};\widehat{\eta}_{kl})$, and

$$\widehat{\Omega}_A := \frac{1}{KL}\sum_{k=1}^{K}\sum_{l=1}^{L}\frac{1}{N_k T_l^2}\sum_{i\in I_k, t\in S_l, r\in S_l}\psi(W_{it};\widehat{\theta},\widehat{\eta}_{kl})\psi(W_{ir};\widehat{\theta},\widehat{\eta}_{kl})',$$

$$\widehat{\Omega}_{DK} := \frac{1}{KL}\sum_{k=1}^{K}\sum_{l=1}^{L}\frac{K/L}{N_k T_l^2}\sum_{t\in S_l, r\in S_l} k\left(\frac{|t-r|}{M}\right)\sum_{i\in I_k, j\in I_k}\psi(W_{it};\widehat{\theta},\widehat{\eta}_{kl})\psi(W_{jr};\widehat{\theta},\widehat{\eta}_{kl}$$

$$\widehat{\Omega}_{NW} := \frac{1}{KL}\sum_{k=1}^{K}\sum_{l=1}^{L}\frac{K/L}{N_k T_l^2}\sum_{i\in I_k, t\in S_l, r\in S_l} k\left(\frac{|t-r|}{M}\right)\psi(W_{it};\widehat{\theta},\widehat{\eta}_{kl})\psi(W_{ir};\widehat{\theta},\widehat{\eta}_{kl})'.$$

where $k\left(\frac{m}{M}\right) = 1 - \frac{m}{M}$ is the Bartlett kernel and M is the bandwidth parameter. Back

Assumption 2 (Approximate Sparse Model)

The unknown function f can be well-approximated by a dictionary of transformations $f_{it} = F(X_{it})$ where f_{it} is a $p \times 1$ vector with $p >> NT$ allowed, and F is a measurable map, such that

$$f(X_{it}) = f_{it}\zeta_0 + r_{it}$$

where the coefficients ζ_0 and the approximation error r_{it} satisfy

$$\|\zeta_0\|_0 \leq s = o(N \wedge T), \ \|r_{it}\|_{2,NT} \equiv R = O_P\left(\sqrt{\frac{s}{N \wedge T}}\right).$$

Back

- If the dimension of f_{it} to be larger than the sample size, the empirical Gram matrix $\frac{1}{NT}\sum_{i=1}^{N}\sum_{t=1}^{T} f_{it}f_{it}'$ is singular.
- However, it turns out we only need its certain sub-matrices to be non-singular.

Assumption 3 (Sparse Eigenvalues)

For any $C > 0$, there exists constants $0 < \kappa_1 < \kappa_2 < \infty$ such that with probability approaching one as $(N, T) \to \infty$ jointly,

$$\kappa_1 \leq \min_{\delta \in \Delta(m)} \delta' M_f \delta < \max_{\delta \in \Delta(m)} \delta' M_f \delta \leq \kappa_2,$$

where $\Delta(m) = \{\delta : \|\delta\|_0 = m, \|\delta\|_2 = 1\}$ and $M_f = \frac{1}{NT}\sum_{i=1}^{N}\sum_{t=1}^{T} f_{it}f_{it}'$.

Back

- The sparse eigenvalue assumption follows from Belloni et al. (2012). It implies a restricted eigenvalue condition, which represents a modulous of continuity between the prediction norm and the norm of δ within a restricted set, and it is also needed in the proof.
- Here we assume the sparse eigenvalue condition because it is needed to show l^2 convergence rate for the LASSO estimator and it implies the restricted eigenvalue condition as shown in Bickel et al. (2009).
- More primitive conditions for both types of assumptions are given in Belloni et al. (2012).

Back

Assumption 4 (Regularity Conditions)

(i) $\left[\mathrm{E}(a_{i,j})^2\right]^{1/2}/\left[\mathrm{E}(a_{i,j})^3\right]^{1/3} = O(1)$ *where* $a_{i,j} := \mathrm{E}[f_{it,j}V_{it}|\alpha_i]$ *for* $j = 1, ..., p$. *(ii)* $\log p = o\left((T)^{1/6}/\log T\right)$, $p = o\left((T)^{13/12}/(\log T)^{1/2}\right)$. *(iii) For some* $s > 1, \delta > 0, \mu > 0$, $E[\|f_{it,j}\|^{8(s+\delta)}] < \infty$, $E[\|V_{it}\|^{8(s+\delta)}] < \infty$ *and* $\mathrm{E}\left[\sum_{t=r}^{r+m} f_{it,j}V_{it}\right]^2 \geq \mu^2 m$ *for all* j *and* $r \geq 0, m \geq 1$. *(iv) Either* $\lambda_{a,j} := \left[E(a_{i,j}^2)\right]^{1/2} > 0$ *or* $\lambda_{g,j} := \left[\sum_{\ell=-\infty}^{\infty} E[g_{t,j}g_{t+\ell,j}]\right]^{1/2} > 0$.

- Assumption 4(iii) restricts the dimension of f_{it}: it still allows the regressor to be larger than the sample size, but in a slower rate than that of Belloni et al. (2016) $\log^3(p) = o(NT)$.
- This is because the weak dependence in our setting makes the tail probability to decay at a slower rate, as shown in the results of Gao et al. (2022).
- The second part of Assumption 4(iii) serves a similar purpose and is due to the a remainder term in decomposing $f_{it,j}V_{it}$.
- The second part of Assumption 4(iii) is binding when the

$$\tilde{\omega}_j^{\mathrm{CHS}} = \tilde{\omega}_j^{\mathrm{A}} + \tilde{\omega}_j^{\mathrm{DK}} - \tilde{\omega}_j^{\mathrm{NW}},$$

where, with $\tilde{v}_{it,j} \equiv f_{it,j}\tilde{V}_{it}$,

$$\begin{aligned}
\tilde{\omega}_j^{\mathrm{A}} &= \frac{N \wedge V}{N^2 T^2} \sum_{i=1}^{N} \sum_{t=1}^{T} \sum_{s=1}^{T} \tilde{v}_{it,j}\tilde{v}_{is,j}, \\
\tilde{\omega}_j^{\mathrm{DK}} &= \frac{N \wedge V}{N^2 T^2} \sum_{t=1}^{T} \sum_{s=1}^{T} k\left(\frac{|t-s|}{M}\right) \left(\sum_{i=1}^{N} \tilde{v}_{it,j}\right) \left(\sum_{l=1}^{N} \tilde{v}_{ls,j}\right), \\
\tilde{\omega}_j^{\mathrm{NW}} &= \frac{N \wedge V}{N^2 T^2} \sum_{i=1}^{N} \sum_{t=1}^{T} \sum_{s=1}^{T} k\left(\frac{|t-s|}{M}\right) \tilde{v}_{it,j}\tilde{v}_{is,j}.
\end{aligned}$$

Asymptotic valid feasible weights: there exists $0 < l \leq 1$ and $1 \leq u < \infty$ such that $l \to 1$ and $l\omega_j^{1/2} \leq \widehat{\omega}_j^{1/2} \leq u\omega_j^{1/2}$, uniformly over $j = 1, ..., p$. Back

Theorem

Suppose AHK and beta-mixing | generalized Mundlak device | approximate sparse model | sparse eigenvalues | regularity conditions *hold for* $P = P_{NT}$ *for each* (N, T)*, then* $\widehat{\theta}$ *from Definition 5.1 obeys, as* $N, T \to \infty$ *with* $N/T \to c$*,* $0 < c < \infty$*,*

$$\sqrt{N}(\widehat{\theta} - \theta_0) \xrightarrow{d} N(0, A_0^{-1} \Omega A_0^{-1})$$

where $A_0 = \mathrm{E}_P[(Z_{it} - f_{it}\zeta_0)(D_{it} - f_{it}\pi)]$*,* $\Omega = \Lambda_a^2 + c\Lambda_g^2$ *with* Λ_a^2 *being the variance of* $\mathrm{E}_P[\psi|\alpha]$ *and* Λ_g^2 *being the long-run variance of* $\mathrm{E}_P[\psi|\gamma]$*.*

Two variance estimators | Consistency

Assumption 1 (AHK Representation and Absolutely Regularity)

Assume the following underlying process:

$$(Y_{it}, D_{it}, X_{it}, U_{it}^{Y}, U_{it}^{D}) = \mu + h(\alpha_i, \gamma_t, \varepsilon_{it}), \quad \forall i \geq 1, t \geq 1,$$

where $\mu = E_P[W_{it}]$, h *is some unknown measurable function;* $(\alpha_i)_{i\geq 1}$, $(\gamma_t)_{t\geq 1}$, *and* $(\varepsilon_{it})_{i\geq 1, t\geq 1}$ *are mutually independent sequences,* α_i *is i.i.d across* i, ε_{it} *is i.i.d across* i *and* t, γ_t *is strictly stationary and beta-mixing at a geometric rate:*

$$\beta_\gamma(q) = \sup_{s\leq T} \beta\left(\{\gamma_t\}_{t\leq s}, \{\gamma_t\}_{t\geq s+q}\right) \leq c_\kappa exp(-\kappa q), \forall q \in \mathbb{Z}^+,$$

for some constants $\kappa > 0$ *and* $c_\kappa \geq 0$.

Back

Assumption 3 (Approximate Sparse Model)

The unknown functions are well-approximated by a linear combination of polynomial transformation as above, such that for some positive integer $s = o\left(\frac{T}{\log(p \vee NT)}\right)$,

$$\|\eta_{Y1}\|_0 \leq s, \ \|r_{it}^Y\|_{NT,2} = O_P\left(\sqrt{\frac{s}{NT}}\right)$$

$$\|\eta_{D1}\|_0 \leq s, \ \|r_{it}^D\|_{NT,2} = O_P\left(\sqrt{\frac{s}{NT}}\right),$$

$$\|\eta_{Z1}\|_0 \leq s, \ \|r_{it}^Z\|_{NT,2} = O_P\left(\sqrt{\frac{s}{NT}}\right).$$

Assumption 4 (Sparse Eigenvalues)

For any $C > 0$, there exists constants $0 < \kappa_1 < \kappa_2 < \infty$ such that with probability approaching one as $(N, T) \to \infty$ jointly,

$$\kappa_1 \leq \min_{\delta \in \Delta(m)} \delta' M_f \delta < \max_{\delta \in \Delta(m)} \delta' M_f \delta \leq \kappa_2, \text{ for } .$$

where $\Delta(m) = \{\delta : \|\delta\|_0 = m, \|\delta\|_2 = 1\}$ and $M_f = \mathbb{E}_{NT}[f_{it}' f_{it}]$.

Back

Assumption 5 (Regularity Conditions)

(i) *For some* $s > 1$, $\delta > 0$, $\mathrm{E}\|D_{it}\|^{8(s+\delta)} < \infty$, $\mathrm{E}\|Z_{it}\|^{8(s+\delta)} < \infty$, $\mathrm{E}\|U_{it}\|^{8(s+\delta)} < \infty$, $\mathrm{E}\|V_{it}^{\iota}\|^{8(s+\delta)} < \infty$ *and* $\max_j \mathrm{E}\|f_{it,j} V_{it}^{\iota}\|^{4(s+\delta)} < \infty$, *for* $\iota = Y, D, Z$. *Moreover, there exist neighborhoods* $\mathcal{N}_m(\xi_0)$ *with* $0 < m < \infty$, *such that* $\mathrm{E}\left[\sup_{\beta \in \mathcal{N}_m(\xi_0)} |f_{it}\xi|\right]^4 < \infty$ *for* $\xi = \beta, \pi, \text{or } \zeta$.

(ii) $\lambda_{min}[\Lambda_a \Lambda_a'] > 0$ *or* $\lambda_{min}[\Lambda_g \Lambda_g'] > 0$, *where* $\Lambda_a \Lambda_a' = \mathrm{E}[a_i a_i']$, $\Lambda_b \Lambda_b' = \sum_{l=-\infty}^{\infty} \mathrm{E}[g_t g_{t+l}']$, *and* $a_i = \mathrm{E}[\psi(W_{it}; \theta_0, \eta_0)|\alpha_i]$, $g_t = \mathrm{E}[\psi(W_{it}; \theta_0, \eta_0)|\gamma_t]$.

(iii) $\mathrm{E}[(Z_{it} - f_{it}\zeta_0)(D_{it} - f_{it}\pi_0)]$ *is non-singular.*

(iv) *For* $\iota = Y, D, Z$ *and* $j = 1, ..., p$, $\left[\mathrm{E}(a_{i,j}^{\iota})^2\right]^{1/2} / \left[\mathrm{E}(a_{i,j}^{\iota})^3\right]^{1/3} = O(1)$ *where* $a_{i,j}^{\iota} := \mathrm{E}[f_{it,j} V_{it}^{\iota}|\alpha_i]$; *For some* $\mu > 0$, $\mathrm{E}\left[\sum_{t=r}^{r+m} Z_{it,j} V_{it}^{\iota}\right]^2 \geq \mu^2 m$ *for all* $r \geq 0, m \geq 1$.

(v) $\log p = o\left((T)^{1/6} / \log T\right)$, $p = o\left((T)^{13/12} / (\log T)^{1/2}\right)$.

- To estimate the asymptotic variance, we again use CHS and DKA variance estimators, except this time we don't need cross-fitting.

$$\widehat{V}_{\mathrm{CHS}} = \widehat{A}_{NT}^{-1}\widehat{\Omega}_{\mathrm{CHS}}\widehat{A}_{NT}^{-1'}, \quad \widehat{\Omega}_{\mathrm{CHS}} = \widehat{\Omega}_{\mathrm{A}} + \widehat{\Omega}_{\mathrm{DK}} - \widehat{\Omega}_{\mathrm{NW}},$$
$$\widehat{V}_{\mathrm{DKA}} = \widehat{A}_{NT}^{-1}\widehat{\Omega}_{\mathrm{DKA}}\widehat{A}_{NT}^{-1'}, \quad \widehat{\Omega}_{\mathrm{DKA}} = \widehat{\Omega}_{\mathrm{A}} + \widehat{\Omega}_{\mathrm{DK}},$$

where

$$\widehat{A}_{NT} := \frac{1}{NT}\sum_{i=1}^{N}\sum_{t=1}^{T}(Z_{it} - f_{it}\tilde{\zeta})(D_{it} - f_{it}\tilde{\pi}),$$

$$\widehat{\Omega}_{\mathrm{A}} := \frac{1}{NT^2}\sum_{i=1}^{N}\sum_{t=1}^{T}\sum_{r=1}^{T}\psi(W_{it};\widehat{\theta},\tilde{\eta})\psi(W_{ir};\widehat{\theta},\tilde{\eta})',$$

$$\widehat{\Omega}_{\mathrm{DK}} := \frac{1}{NT^2}\sum_{t=1}^{T}\sum_{r=1}^{T}k\left(\frac{|t-r|}{M}\right)\sum_{i=1}^{N}\sum_{j=1}^{N}\psi(W_{it};\widehat{\theta},\tilde{\eta})\psi(W_{jr};\widehat{\theta},\tilde{\eta})',$$

$$\widehat{\Omega}_{\mathrm{NW}} := \frac{1}{NT^2}\sum_{i=1}^{N}\sum_{t=1}^{T}\sum_{r=1}^{T}k\left(\frac{|t-r|}{M}\right)\psi(W_{it};\widehat{\theta},\tilde{\eta})\psi(W_{ir};\widehat{\theta},\tilde{\eta})'.$$

Back

Theorem

Suppose assumptions for the previous theorem holds for $P = P_{NT}$ for each (N, T) and $M/T^{1/2} = o(1)$. Then, $(N, T) \to \infty$ and $N/T \to c$ where $0 < c < \infty$,

$$\widehat{V}_{\mathrm{CHS}} = A_0^{-1} \Omega A_0^{-1} + o_P(1),$$
$$\widehat{V}_{\mathrm{DKA}} = \widehat{V}_{\mathrm{CHS}} + o_P(1).$$

Back

Belloni, A., Chen, D., Chernozhukov, V., and Hansen, C. (2012). Sparse models and methods for optimal instruments with an application to eminent domain. *Econometrica*, 80(6):2369–2429.

Belloni, A., Chernozhukov, V., and Hansen, C. (2014). Inference on treatment effects after selection among high-dimensional controls. *Review of Economic Studies*, 81(2):608–650.

Belloni, A., Chernozhukov, V., Hansen, C., and Kozbur, D. (2016). Inference in high-dimensional panel models with an application to gun control. *Journal of Business & Economic Statistics*, 34(4):590–605.

Bickel, P. J., Ritov, Y., and Tsybakov, A. B. (2009). Simultaneous analysis of lasso and dantzig selector.

Chernozhukov, V., Chetverikov, D., Demirer, M., Duflo, E., Hansen, C., Newey, W., and Robins, J. (2018). Double/debiased machine learning for treatment and structural parameters. *Econometrics Journal*, 21(1):C1–C68.

Chetverikov, D., Liao, Z., and Chernozhukov, V. (2021). On

cross-validated lasso in high dimensions. *The Annals of Statistics*, 49(3):1300–1317.

Chiang, H. D., Kato, K., Ma, Y., and Sasaki, Y. (2022). Multiway Cluster Robust Double/Debiased Machine Learning. *Journal of Business and Economic Statistics*, 40(3):1046–1056.

Gao, L., Shao, Q.-M., and Shi, J. (2022). Refined cramér-type moderate deviation theorems for general self-normalized sums with applications to dependent random variables and winsorized mean. *The Annals of Statistics*, 50(2):673–697.

Gonçalves, S. (2011). The moving blocks bootstrap for panel linear regression models with individual fixed effects. *Econometric Theory*, 27(5):1048–1082.

Javanmard, A. and Montanari, A. (2014). Confidence intervals and hypothesis testing for high-dimensional regression. *The Journal of Machine Learning Research*, 15(1):2869–2909.

Kock, A. B. and Tang, H. (2019). Uniform inference in high-dimensional dynamic panel data models with approximately sparse fixed effects. *Econometric Theory*, 35(2):295–359.

Mundlak, Y. (1978). On the pooling of cross-section and time-series data. *Econometrica*, 46(69):X6.

Nakamura, E. and Steinsson, J. (2014). Fiscal stimulus in a monetary union: Evidence from us regions. *American Economic Review*, 104(3):753–792.

Ning, Y. and Liu, H. (2017). A general theory of hypothesis tests and confidence regions for sparse high dimensional models.

Semenova, V., Goldman, M., Chernozhukov, V., and Taddy, M. (2023). Inference on heterogeneous treatment effects in high-dimensional dynamic panels under weak dependence. *Quantitative Economics*, 14(2):471–510.

Van de Geer, S., Bühlmann, P., Ritov, Y., and Dezeure, R. (2014). On asymptotically optimal confidence regions and tests for high-dimensional models.

Wooldridge, J. M. (1997). Multiplicative panel data models without the strict exogeneity assumption. *Econometric Theory*, 13(5):667–678.

Wooldridge, J. M. and Zhu, Y. (2020). Inference in approximately sparse correlated random effects probit models with panel data. *Journal of Business & Economic Statistics*, 38(1):1–18.

Zhang, C.-H. and Zhang, S. S. (2014). Confidence intervals for low dimensional parameters in high dimensional linear models. *Journal of the Royal Statistical Society Series B: Statistical Methodology*, 76(1):217–242.