Estimating Spillovers under Factor-induced Constraints

Hao Dong *
Southern Methodist University

Qiwei Yao
London School of Economics

Abstract

In the literature on the estimation for spillover effects, prior knowledge on the spillover structure is often incorporated as a constraint on the estimation. This paper proposes a new method to construct constraints on the spillover effects using the latent factor structure of the variables that generate the spillovers. The method improves the performance of the existing methods such as LASSO. We have derived the L_2 error bound for the LASSO estimator under factor-induced constraints. Comparing it with that of the unconstrained LASSO estimator, the new LASSO estimator has an approximately sharper L_2 error bound when factors are strong. Simulation results demonstrate our findings.

Keywords: Factor analysis, Constrained estimation, LASSO

JEL: C13, C23, C38, C52

*haod@mail.smu.edu

1 Introduction

The spillover effect, also known as externality, is the effect of a treatment assigned to one individual from the outcomes of others. It is prevalent in economic studies. Common examples include technological adoptions, peer effects of education, and return and volatility spillovers in stock markets. Ignoring the spillover effects could lead to severe bias in the estimation of the treatment effects, while the structure of interactions driving the spillovers could also be of its own interest.

In the estimation of spillover effects, one of the challenges is due to the relative large number of pair-specific parameters in comparison to the limited sample size. In such a situation, a panel data structure can relieve the problem because the variations across the additional time dimension can be used as a source of the knowledge on the spillover structure. However, given the prevalence of the short panel, which is the case when there are more cross-sectional units than the number of periods, the spillover effects could still be under-identified. Therefore, as pointed out in Blume et al. (2015), prior knowledge on the spillover structure is usually needed.

Most existing works focus on the case when the spillover structure is observed, see De Paula (2017). There is a recent surge focusing on the cases when the spillover structure is unobserved but known to be sparse, including, among others, Manresa (2013), and Lam and Souza (2013). Specifically, they assume that each individual is only connected with a limited number of others in the population. Compared to the observed spillover structure assumption, the sparsity assumption is more restrictive on the number of spillovers but less restrictive on the identity and intensity. Under the sparsity assumption, the estimation can be facilitated by some adequate penalized regression methods, such as LASSO of Tibshirani (1996), and adaptive LASSO of Zou (2006).

All the aforementioned work relies on the assumption of certain direct knowledge on the spillover structure. This paper takes a different perspective. By imposing a latent factor

structure on the variables which generate spillovers, we show that such a factor structure is in fact an indirect knowledge on the spillover structure. Similar to the direct knowledge, the factor structure also implies constraints on the parameters characterizing the spillovers. Therefore, we can improve the performance of existing spillover estimation methods, like LASSO, by adding these factor-induced constraints. The motivation for our setting can be understood from, for example, the fact that for technological spillovers from the R&D investments to the productivities of different firms in the same market, the R&D investments of different firms are typically driven by some macro factors of the market.

The remainder of this paper is organized as follows. Section 2 introduces the model used to characterize the spillovers and the factor structure, and provides the intuition on how the factor structure could be treated as a knowledge about the spillover structure. Section 3 discusses the general construction of the LASSO estimator based on the factor-induced constraints. Section 4 derives the properties of the proposed estimator. Section 5 investigates the performance of the new estimation by simulation, and Section 6 concludes.

2 Spillovers and Factor-induced Constraints

Consider linear regression model

$$y_t = \beta' x_t + \gamma' z_t + \epsilon_t, \quad t = 1, \cdots, T, \tag{1}$$

where $x_t = (x_{1t}, \cdots, x_{Nt})'$ is an N-vector of the variables generating the spillovers, $z_t = (z_{1t}, \cdots, z_{kt})'$ is a k-vector of the additional controls, and ϵ_t is a regression error. In particular, we focus on the case when $N > T$, which makes (1) a high-dimensional problem in the sense that the number of unknown parameters is larger than the number of the observations. For the estimation of spillover effects using a panel dataset, β in (1) can be interpreted as capturing the spillovers from a specific individual, that is $\{y_t, x_t, z_t, \epsilon_t, \beta, \gamma\}$ in (1) should all be indexed by i in general. To keep notations simple,

the discussion will focus on (1) rather than a general panel data model throughout the rest of this paper.

For the factor structure of x_t, we consider the factor model as follows.

$$x_t = Af_t + u_t, \tag{2}$$

where A is a $N \times r$ matrix of the factor loadings, f_t is a r-vector of the factors, and u_t is a N-vector of the idiosyncratic shocks. To avoid the potential bias caused by omitting relevant factors, we treat all factors f_t as latent in this paper. In the case when f_t is partly observed, x_t then could be treated as residuals from the regression on the observed factors.

To understand how (2) could be the indirect knowledge on β, notice that substituting (2) into (1) gives the reduced form as follows.

$$y_t = \psi' f_t + \gamma' z_t + e_t, \tag{3}$$

where $\psi = A'\beta$ and $e_t = \beta' u_t + \epsilon_t$. If A and ψ were observed, $\psi = A'\beta$ are actually r linear constraints on β. Even though A and ψ may not be directly observed in practice, both of them can be estimated at a minor cost, and the constraints can then be established by substituting the estimates of A and ψ.

3 Estimation

In this section, we construct the estimator of β making use of the latent factor structure (2). Let $\hat{A}$, $\hat{\psi}$, and $\hat{\gamma}$ denote, respectively, the estimator of A, ψ, and γ. For expository purposes, we first treat them as if they had been defined in Section 3.1. The details of their construction are discussed later in Section 3.2.

3.1 LASSO under Factor-induced Constraints

If the spillover structure is sparse, penalized regression methods can be used to estimate β in (1) when $N > T$. One popular method is LASSO, which places an L_1 penalty on β. Specifically, let $\tilde{\beta}$ denote the LASSO estimator of β, which is defined as the solution to the following minimization problem.

$$\min_b \|Y - Xb - Z\hat{\gamma}\|_2^2/2 + \lambda\|\hat{D}b\|_1, \tag{4}$$

where $Y = (y_1, \cdots, y_T)'$, $X = (x_1, \cdots, x_T)'$, $Z = (z_1, \cdots, z_T)'$, and $\hat{D}$ is a diagonal matrix introduced to normalize X. In comparison with the other penalized regression methods, LASSO is featured by its ability of shrinking small regression coefficients to exact zero.

Knowing the latent factor structure (2) in x_t, which implies linear constraints on β, we can improve the performance of the LASSO estimator by adding these factor-induced constraints. The LASSO estimator under these factor-induced constraints, denoted as $\hat{\beta}$, is then defined as the solution to the following minimization problem.

$$\min_b \|Y - Xb - Z\hat{\gamma}\|_2^2/2 + \lambda\|\hat{D}b\|_1 \quad \text{s.t.} \quad \hat{\psi} = \hat{A}'b \tag{5}$$

In the situation when the true parameters satisfy the constraints, James et al. (2012) shows that the constrained LASSO outperforms the unconstrained one in the sense that it has a sharper L_2 error bound. This implies that the infeasible constrained LASSO estimator, which is the LASSO estimator under the infeasible factor-induced constraints as if A and ψ were observed, denoted as $\check{\beta}$, should outperform the unconstrained LASSO estimator $\tilde{\beta}$. In our setting, even though the true value of β may not satisfy the constraints in (5) exactly due to the estimation errors of $\hat{A}$ and $\hat{\psi}$, the gap would decrease as the

sample size grows. In particular, if $\hat{A}$ and $\hat{\psi}$ converge sufficiently fast, we can show that the L_2 error bound of the feasible constrained LASSO estimator $\hat{\beta}$ is close to that of the infeasible constrained LASSO estimator $\check{\beta}$, which is strictly sharper than that of the unconstrained LASSO estimator $\tilde{\beta}$.

3.2 Construction of Constraints

In the rest of this section, we provide details on $\hat{A}$, $\hat{\psi}$, and $\hat{\gamma}$, which are needed in the construction of the feasible constrained LASSO estimator $\hat{\beta}$ defined in (5). Specifically, we will present the estimation of the factor model (2), which yields $\hat{A}$, and the corresponding factor-augmented regression, which yields $\hat{\psi}$ and $\hat{\gamma}$.

The estimation for the factor model (2), which is known as the large factor model due to the fact that $N > T$, is well documented in the literature. Early attempts include the principle components method (PC) in Bai and Ng (2002) and the generalized principle components method (GPC) in Choi (2012) using the estimated covariance matrix suggested by Bai and Liao (2013) as weights. When the data exhibits non-zero serial correlation, Lam et al. (2011) develops a new approach based on the information from the autocovariance matrix at non-zero lags, instead of the covariance matrix as in PC and GPC. Compared with PC and GPC, Lam et al. (2011) has a better performance when there exists strong cross-correlation over different component series. In this paper, we deal with x_t which exhibits strong autocorrelations. This makes the method of Lam et al. (2011) a pertinent choice for our inference.

When the number of factors r is known, Lam et al. (2011) suggests the following estimators of A and f_t.

$$\hat{A} = (\hat{s}_1, \cdots, \hat{s}_r) \quad \text{and} \quad \hat{f}_t = \hat{A}' x_t, \tag{6}$$

where $(\hat{s}_1, \cdots, \hat{s}_r)$ are the orthonormal eigenvectors of $\hat{H}_x$ corresponding to its r largest

eigenvalues, and

$$\hat{H}_x = \sum_{k=1}^{k_0} \widehat{\Sigma}_x(k)\widehat{\Sigma}_x(k)', \quad \widehat{\Sigma}_x(k) = \frac{1}{T}\sum_{j=1}^{T-k}(x_{t+j} - \bar{x})(x_t - \bar{x})', \quad \bar{x} = \frac{1}{T}\sum_{t=1}^{T} x_t \tag{7}$$

with $k_0 \geq 1$ being a prespecified integer.

Due to the rotational indeterminacy of the factor model [1], instead of the estimator of a specific loading matrix, $\hat{A}$ is an estimator for the factor loading space $\mathcal{M}(A)$, which is the r-dimensional linear space spanned by the columns of A. However, it is worth noting that different choices of A and f_t lead to equivalent factor-induced constraints $\psi = A'\beta$, which implies that the feasible constrained LASSO estimator $\hat{\beta}$ essentially does not depend on the choice of A and corresponding f_t.

In practice, we need to estimate the number of factors r to implement the method of Lam et al. (2011). There exist mainly two types of estimation methods. One is based on information criteria, for example see Bai and Ng (2002). The other is based on the distribution of eigenvalues, including Onatski (2009), Lam et al. (2012), and Ahn and Horenstein (2013). Following Lam et al. (2012), we estimate the number of factors r by the relative magnitude of the ratios of eigenvalues, which is defined as follows.

$$\hat{r} = \arg\min_{1 \leq j \leq R} \frac{\hat{\lambda}_{j+1}}{\hat{\lambda}_j}, \tag{8}$$

where $r < R < N$ and $\hat{\lambda}_1 \geq \cdots \geq \hat{\lambda}_N$ are the decreasingly ordered eigenvalues of $\hat{H}_x$.

To estimate ψ and γ, we consider the factor-augmented regression as follows.

$$y_t \quad = \quad \psi'\hat{f}_t + \gamma' z_t + e_t \tag{9}$$

Note that the number of unknown parameters in (9) is $r + k < T$. Thus, $\hat{\psi}$ and $\hat{\gamma}$ can be

[1] $Af_t = AHH'f_t$ for any orthnormal matrix $H \in \mathbb{R}^{r \times r}$.

obtained by least squares method. Specifically, $\hat{\psi}$ and $\hat{\gamma}$ are defined as the solution to the following minimization problem with respect to b_1 and b_2.

$$\min_{b_1, b_2} \sum_{t=1}^{T} (y_t - b_1' \hat{f}_t - b_2' z_t)^2 \tag{10}$$

4 Theoretical Results

In this section we develop the theoretical properties of the feasible constrained LASSO estimator $\hat{\beta}$. We introduce some notations first. For a d-vector $v = (v_1, \cdots, v_d)$, let $\|v\|_p$ be its L_p norm with $p \geq 1$, and $\text{supp}(v) = \{j : v_j \neq 0\}$. Given a set of indices $I \subset \{1, \cdots, d\}$, let v_I be the d-vector with the j-th component $(v_I)_j = v_j \mathbb{1}\{j \in I\}$ for $j = 1, \ldots, d$. For a $d_1 \times d_2$ matrix W, let $\|W\|_2 = \sqrt{\lambda_{\max}(W'W)}$ be its spectral norm and $\|W\|_{\min}$ be the square root of the smallest non-zero eigenvalue of WW', where $\lambda_{\max}(\cdot)$ and $\lambda_{\min}(\cdot)$ denote, respectively, the largest and the smallest eigenvalues of the matrix. Given a set of indices $I_2 \subset \{1, \cdots, d_2\}$, we denote by W_{I_2} the columns of W associated with I_2, and $W_{I_2^c}$ corresponds to the remaining columns of W. We also use the notation $a \asymp b$ to denote the situation when $a = O(b)$ and $b = O(a)$ hold simultaneously. To simplify the exposition, we assume $z_t \equiv 1$ in (1). The results with more general z_t can be obtained in a similar manner.

The construction of $\hat{\beta}$ and its performance critically depend on the accuracy of $\hat{A}$ and $\hat{\psi}$. Intuitively the closer $\hat{A}$ and $\hat{\psi}$ are to A and ψ, the closer the constraints in (5) would be to the infeasible factor-induced constraints $\psi = A'b$, under which the infeasible constrained LASSO estimator $\check{\beta}$ would have sharper L_2 error bound than the unconstrained LASSO estimator $\hat{\beta}$. To specify their accuracy, we study the convergence rate of $\hat{A}$ and $\hat{\psi}$.

The convergence rate of $\hat{A}$ has been shown in Theorem 1 and 2 of Lam et al. (2011). Thus, we focus on $\hat{\psi}$. To derive the convergence rate of $\hat{\psi}$, we need to impose assumptions as follow.

Assumption 1

(i) $A'A = I_r$, f_t *is weakly stationary,* u_t *is a white noise with zero mean and variance* Σ_u, *and* $\text{cov}(f_t, u_s) = 0$ *for* $t \leq s$.

(ii) $\|\Sigma_f(k)\|_2 \asymp N^{1-\nu} \asymp \|\Sigma_f(k)\|_{\min}$, $\|\Sigma_u\|_2 = O(N^\rho)$, *and* $\|\Sigma_{fu}(k)\|_2 = O\left(N^\rho\right)$ *with* $\rho < 1 - \nu$ *for some* $\nu \in [0, 1]$ *and* $k = 0, 1, \cdots, k_0$, *where* $\Sigma_f(k) = \text{cov}[f_{t+k}, f_t]$ *and* $\Sigma_{fu}(k) = \text{cov}[f_{t+k}, u_t]$; $\lambda_{\min}\left(A\Sigma_{fu}(0)\right) = o\left(N^{1-v}\right)$.

(iii) $\{x_t', \epsilon_t'\}$ *is a stationary* α*-mixing process with* $E\|(x_t', \epsilon_t')\|^{2+\gamma} < \infty$ *elementwisely for some* $\gamma > 0$, *and the mixing coefficients* $\alpha(t)$ *satisfying* $\sum_{t\geq 1}^{\infty} \alpha(t)^{\frac{\gamma}{2+\gamma}} < \infty$.

(iv) $\|\Sigma_{f\epsilon}\|_2 = O\left(N^{\frac{1}{2}}T^{-\frac{1}{2}}\right)$ *and* $\|\Sigma_{u\epsilon}\|_2 = O\left(N^{\frac{1}{2}}T^{-\frac{1}{2}}\right)$, *where* $\Sigma_{f\epsilon} = cov[f_t, \epsilon_t]$ *and* $\Sigma_{u\epsilon} = cov[u_t, \epsilon_t]$.

Assumption 1 (i) is always fulfilled via a normalization on factor loadings, see Lam et al. (2011). Due to the rotational indeterminacy of factor model (2), only the linear space spanned by the columns of A, denoted as $\mathcal{M}(A)$, is identified.[2]. Following Lam et al. (2011), we specify $\tilde{A}$ as QV, where Q is a $N \times r$ matrix that comes from the thin Q-R decomposition of A, and satisfies $Q'Q = I_r$, and V is a r-dimensional orthonormal matrix that comes from the spectral decomposition $\sum_{k=1}^{k_0}\{\Sigma_f(k)Q' + \Sigma_{f,u}(k)\}\{\Sigma_f(k)Q' + \Sigma_{f,u}(k)\}' = VDV'$ [3]. Once we specify the target factor loading matrix as above, the corresponding factor process follows by $V'Rf_t$, where R is a $r \times r$ matrix that comes from the same thin Q-R decomposition of A that generates Q. To distinguish the objects based on the original factors from those based on the chosen factors, we add superscript "o" to the objects based on the original factors, leaving the objects based on the chosen factors with no superscript.

Assumption 1 (ii) follows by $\|\Sigma_f^o(k)\|_2 \asymp \|\Sigma_f^o(k)\|_{\min} \asymp 1$, $\Sigma_{f,u}^o(k) = O(1)$ elementwise, and $\|a_i\|_2^2 = N^{1-\nu}$ for $i = 1, \cdots, r$ and $0 \leq \nu \leq 1$, where a_i is the ith column of A, see

[2]Even though $A'A = I_r$ has been restrictive, it still cannot pin down A because $A'H'HA = I_r$ for any $r \times r$ orthonormal matrix H.

[3]Since the spectral decomposition does not pinned down the sign of V, we need to allow $\hat{A}$ to adjust sign to adept to the sign of chosen V.

Lemma 1 of Lam et al. (2011). Assumption 1 (iii) is introduced in order to capture the upper bounds of $\|\hat{\Sigma}_f^o(k) - \Sigma_f^o(k)\|_2$ and $\|\hat{\Sigma}_{f,u}^o(k) - \Sigma_{f,u}^o(k)\|_2$, which follows by using the Frobenius norm as upper bound of the spectral norm, and using the central limit theorem of α-mixing process elementwisely, for example Theorem 0 of Bradley (1985). Since we deal with an otherwise linear regression model, putting restrictions on the correlation between f_t and ϵ_t is necessary, and this is specified in Assumption 1 (iv).

Under Assumption 1, the convergence rate of $\hat{A}$ in the spectral norm is given by Lam et al. (2011), and is stated in Lemma 1 as follows.

Lemma 1 *Under Assumption 1,* $\|\hat{A} - A\|_2 = O_p(N^\nu T^{-1/2})$.

The linear regression with the estimated factors such as (9) has been studied by Stock and Watson (2002) and Bai and Ng (2006). Both works rely on the factors estimated by the variance-covariance based method such as Bai and Ng (2002). In this paper, we use a different approach to estimate the factor model (2). Thus, the result on the convergence rate of $\hat{\psi}$ is new. Specifically, under Assumption 1, the convergence rate of $\hat{\psi}$ is obtained and this result is summarized in Theorem 1 as follows. The proof of Theorem 1 is relegated to Appendix A.

Theorem 1 *Let Assumption 1 hold and* $N^\nu T^{-1/2} = o(1)$. *Then it holds that*

$$\|\hat{\psi} - \psi\|_2 = O_p(N^\nu T^{-\frac{1}{2}}).$$

Theorem 1 provides the convergence rate of $\hat{\psi}$ when both N and T go to infinity. The rate depends on the strength of factors ν. When $\nu = 0$, which is the case when the factors are strong, $\hat{\psi}$ attains the root-T rate, which is the same as if f_t is directly observed. If $\nu > 0$, which is the case when the factors are weak, the convergence rate of $\hat{\psi}$ is slowed down by N^ν. In this sense, the behavior of $\hat{\psi}$ is very similar to that of $\hat{A}$.

To derive the error bound of $\hat{\beta}$, we need further assumptions. The key is to regularize the Gram matrix $M_x = X'X/T$. M_x is singular when $N > T$, and it is impossible to require its smallest eigenvalue is bounded off zero. In such a situation, instead of the standard eigenvalue, following the literature of the penalized regression, we impose the lower bound assumption on the restricted eigenvalue of M_x. To characterize the restricted eigenvalue, let $\mathcal{T}$ be a subset of $\{1, \cdots, N\}$ such that $A'_{\mathcal{T}}$ is invertible, $\mathcal{T}^{\mathrm{c}} = \{1, \cdots, N\} \backslash \mathcal{T}$, and $\mathcal{S} = \mathrm{supp}(\beta_{\mathcal{T}^{\mathrm{c}}})$. Then, the restricted eigenvalue of M_x is defined in (11) as follows.

$$\kappa_{\zeta}(M_x) = \min_{\delta \in \Delta_{\zeta}} \frac{\delta^{'} M_x \delta}{\|\delta\|_2^2} \tag{11}$$

where

$$\Delta_{\zeta} = \left\{ \delta \in \mathbb{R}^N : \|\delta_{\mathcal{S}^{\mathrm{c}} \cap \mathcal{T}^{\mathrm{c}}}\|_1 \leq \frac{\zeta + 1}{\zeta - 1} \left(\|\delta_{\mathcal{T}}\|_1 + \|\delta_{\mathcal{S} \cap \mathcal{T}^{\mathrm{c}}}\|_1 + \|\delta_f\|_1 \right) \right\}, \tag{12}$$

and

$$\delta_f = (\hat{A}'_{\mathcal{T}})^{-1} \hat{\psi} - (A'_{\mathcal{T}})^{-1} \psi - [(\hat{A}'_{\mathcal{T}})^{-1} \hat{A}'_{\mathcal{T}^{\mathrm{c}}} - (A'_{\mathcal{T}})^{-1} A'_{\mathcal{T}^{\mathrm{c}}}] \beta_{\mathcal{T}^{\mathrm{c}}}.$$

In Appendix A, we show that the estimation errors of $\hat{\beta}$ belong to Δ_{ζ}, which justifies the sufficiency of $\kappa_{\zeta}(M_x) > 0$ in the study of the error bound of $\hat{\beta}$. Compared to James et al. (2012), our restricted set Δ_{ζ} is larger due to the estimation errors of $\hat{A}$ and $\hat{\psi}$, which are characterized by δ_f. If we know A and ψ exactly, $\delta_f = 0$ and Δ_{ζ} coincides with the restricted set defined in James et al. (2012). Moreover, if there is no restriction, $\mathcal{T} = \emptyset$ and $\zeta = 2$, which implies that Δ_{ζ} degenerates to the restricted set of standard LASSO, see Bickel et al. (2009).

Assumption 2

(i) *$r \leq s = o(T)$, where $r = \mathrm{rank}(A)$ and $s = \|\phi^*\|_0$.*

(ii) *$A'_{\mathcal{T}}$ is non-singular for some $\mathcal{T} \subset \mathrm{supp}(\beta)$.*

(iii) *For any $\zeta > 0$, there exists a finite constrant $\kappa > 0$, which does not depend on T but may depend on ζ, such that $\kappa_\zeta(M_x) \geq \kappa$ with probability approaching 1 as $T \to \infty$.*

(iv) *$\{x_{jt}, \epsilon_t\}_{t=1}^T$ is a strong mixing sequence with $E[x_{jt}\epsilon_t] = 0$ for $j = 1, \cdots, N$ with mixing coefficient $\alpha(t)$ satisfying $\alpha(t) \leq \exp(-ct^{\eta_1})$ for some $\eta_1 > 0$ and $c > 0$; $\sup_{1\leq j\leq r} \sup_t P(|x_{jt}\epsilon_t| > x) \leq \exp(1-x^{\eta_2})$ for some $\eta_2 > 0$; $0 < l \leq \min_{1\leq j\leq N} V_j \leq \max_{1\leq j\leq N} V_j \leq u < \infty$, where $V_j = \sup_{1\leq t\leq T} \left(Ex_{jt}^2\epsilon_t^2 + 2\sum_{s>t} |E(x_{js}x_{jt}\epsilon_s\epsilon_t)|\right)$.*

(v) $\log N = o(T^{1/3})$.

Assumption 2 (i) requires that the number of the factors is no greater than the number of essentially relevant regressors. This is a reasonable condition, as in practice the number of latent factors is usually small or very small. Assumption 2 (ii) requires the loadings of the essentially relevant regressors are linearly independent with each other. Assumption 2 (iv) is a technical assumption to facilitate the Fuk-Nagaev inequality, which is used to control the probability when the maximum score is beyond a specific penalty level. Assumption 2 (v) is mild and is a standard condition for the penalized regression methods such as LASSO. In particular, it requires that the size of the cross-sectional dimension is no greater than the exponential of the sample size. To simplify the exposition, we set $\gamma = 0$ and $\hat{D} = I$ in the rest part of this section. The general results can be obtained in a similar manner.

Theorem 2 *Let Assumption 1 and 2 hold. If $\lambda = K\sqrt{T \log N}$ with $K > 4\zeta\sqrt{\frac{2u}{\log 2}}$ and some constant $\zeta > 1$, it holds that*

$$\|\hat{\beta} - \beta\|_2^2 \leq T_{n,1} + \sqrt{T_{n,1}^2 + T_{n,2}}, \tag{13}$$

where

$$T_{n,1} = \frac{\sqrt{2}(\zeta+1)\max\{\sqrt{s-r}, \sqrt{r}\}\lambda}{\kappa^2\zeta T},$$

$$T_{n,2} = -\frac{2(\zeta+1)\sqrt{r}\lambda\|\delta_f\|_2}{\kappa^2\zeta T}.$$

When factors are strong (i.e. $\nu = 0$), $T_{n,2}$ is of order $\frac{\sqrt{\log N}}{T}$, and is dominated by $T_{n,1}^2$ whose order is $\frac{\log N}{T}$. In such a situation, the error bound of our constrained LASSO estimator $\hat{\beta}$ is close to that of the infeasible constrained LASSO estimator $\check{\beta}$, see Theorem 1 of James et al. (2012), which is known to be sharper than that of the standard LASSO estimator $\tilde{\beta}$, see Theorem 1 of Negahban et al. (2009). When factors are weak (i.e. $\nu > 0$), $T_{n,2}$ is of order $N^{\nu}\sqrt{\log N}/T$, which implies that $T_{n,2}$ dominates $T_{n,1}^2$, and the error bound of our feasible constrained LASSO estimator $\hat{\beta}$ will not be close to that of the infeasible constrained LASSO estimator $\check{\beta}$ even when T is large.

5 Numerical Results

In this section, we report some simulation results which illustrate the finite sample properties of the estimators proposed in Section 3.

5.1 Simulation Design

Throughout this section, we consider three data generation processes as follow.

Example 1: For the factor model (2), let $r = 1$, $A = (a_1, \cdots, a_N)'$ with $a_i = 2\cos(2\pi i/N)$, $f_t = 0.4f_{t-1} + \omega_t$ with $\omega_t \sim i.i.N(0,1)$, and $u_t \sim i.i.N(0, I_N)$. For (1), we set $\beta_j = 5\mathbb{1}_{\{j=1,2,3\}}$, $x_t \sim i.i.N(0,5)$, $\gamma = 5$, $z_t = 1$, and $\epsilon_t \sim i.i.N(0, 1/\sqrt{2})$.

Example 2.1: For the factor model (2), let $r = 3$. The elements of A is generated randomly from the $U(-5,5)$ distribution. $f_{1t} = v_t$, $f_{2t} = v_{t-1}$, $f_{3t} = v_{t-2}$ for $v_t = 0.5\omega_{t-1} + \omega_t$ with $\omega_t \sim i.i.N(0,1)$, and $u_t \sim i.i.N(0, I_N)$. For (1), we set $\beta_j = 5\mathbb{1}_{\{j=1,2,3,4,5\}}$, $x_t \sim i.i.N(0,5)$, $\gamma = 5$, $z_t = 1$, and $\epsilon_t \sim i.i.N(0, 1/\sqrt{2})$.

Example 2.2: $u_t \sim i.i.N(0, \Sigma_u)$, where the (i,j) element of Σ_u is defined as

$$\sigma_{i,j} = \begin{cases} \frac{1}{2}\left\{(|i-j|+1)^{2H} - 2|i-j|^{2H} + (|i-j|-1)^{2H}\right\}, & i \neq j; \\ 1, & i = j. \end{cases}$$

with $H = 0.9$ is the Hurst parameter. Everything else is the same as Example 2.1.

5.2 Simulation Results

First, we provide results for $\hat{A}$, $\hat{\gamma}$, and $\hat{\psi}$. Let $N \in \{400, 500\}$ and $T \in \{100, 200, 400\}$. For each data generation process, the rooted mean square errors (RMSE) are reported in two situations: $\nu = 0$ and $\nu = 0.5$. The experiments are replicated 100 times for each setting.

In Table 1 and 2, RMSE of $\hat{A}$, $\hat{\gamma}$, and $\hat{\psi}$ are reported separately when data is generated from Example 1 under $\mu = 0$ and $\mu = 0.5$. In the case when $\nu = 0$, i.e. factors are strong, Table 1 shows that RMSE of $\hat{A}$, $\hat{\gamma}$, and $\hat{\psi}$ decrease as T grows from 100 to 400, while the increase of N from 400 to 500 does not have much effect when T is large. In the case when $\nu = 0.5$, i.e. factors are weak, Table 2 shows that RMSE of $\hat{A}$, $\hat{\gamma}$, and $\hat{\psi}$ decrease much slower than the case when $\mu = 0$. Moreover, the increase of N from 400 to 500 does slightly raise RMSE even when T is large.

Table 1: RMSE of $\hat{A}$, $\hat{\gamma}$, and $\hat{\psi}$ for Example 1 with $\nu = 0$.

	$\|\hat{A} - A\|_2$		$\|\hat{\gamma} - \gamma\|_2$		$\|\hat{\psi} - \psi\|_2$	
	$N = 400$	$N = 500$	$N = 400$	$N = 500$	$N = 400$	$N = 500$
$T = 100$	176_{53}	186_{60}	126_{94}	155_{114}	27_{22}	24_{19}
$T = 200$	120_{24}	113_{20}	90_{78}	107_{87}	16_{12}	16_{11}
$T = 400$	84_{9}	83_{13}	67_{56}	77_{64}	12_{9}	11_{9}

Note: Means and standard deviations are reported for each case, and the reported values are actual values multiplied by 1000.

Table 2: RMSE of $\hat{A}$, $\hat{\gamma}$, and $\hat{\psi}$ for Example 1 with $\nu = 0.5$.

	$\|\hat{A} - A\|_2$		$\|\hat{\gamma} - \gamma\|_2$		$\|\hat{\psi} - \psi\|_2$	
	$N = 400$	$N = 500$	$N = 400$	$N = 500$	$N = 400$	$N = 500$
$T = 100$	530_{73}	555_{70}	127_{95}	156_{113}	126_{102}	105_{87}
$T = 200$	425_{56}	426_{46}	91_{77}	108_{87}	84_{67}	88_{60}
$T = 400$	330_{31}	340_{40}	67_{56}	76_{64}	59_{46}	61_{46}

Note: Means and standard deviations are reported for each case, and the reported values are actual values multiplied by 1000.

Table 3 and 4 reports RMSE of $\hat{A}$, $\hat{\gamma}$, and $\hat{\psi}$ when data is generated from Example 2.1 and 2.2 under $\mu = 0$ and $\mu = 0.5$ respectively. The difference between Example 2.1 and 2.2 is the cross-correlations of the idiosyncractic shocks u_t in factor model (2), where Example 2.2 has stronger cross-correlations than Example 2.1. When factors are strong (i.e. $\nu = 0$), Table 3 again shows that RMSE of $\hat{A}$, $\hat{\gamma}$, and $\hat{\psi}$ decrease as T grows, and the increase of N does not have much effect. Moreover, comparing the results from Example 2.1 and Example 2.2, Table 3 also shows that the cross-correlations among the idiosyncratic shocks do not have much effect on the convergence rates characterized by RMSE. Similar results are reported in Table 4 when factors are weak (i.e. $\nu = 0.5$), except that RMSE decrease much slower than the case when $\nu = 0$.

Table 3: RMSE of $\hat{A}$, $\hat{\gamma}$, and $\hat{\psi}$ for Example 2.1 and 2.2 with $\nu = 0$.

		$\|\hat{A} - A\|_2$		$\|\hat{\gamma} - \gamma\|_2$		$\|\hat{\psi} - \psi\|_2$	
		$N = 400$	$N = 500$	$N = 400$	$N = 500$	$N = 400$	$N = 500$
DGP2.1	$T = 100$	408_{176}	450_{256}	203_{148}	169_{129}	108_{83}	101_{91}
	$T = 200$	354_{196}	349_{223}	126_{91}	147_{114}	71_{58}	84_{149}
	$T = 400$	234_{106}	266_{116}	95_{67}	87_{70}	49_{39}	41_{33}
DGP2.2	$T = 100$	438_{305}	402_{249}	363_{280}	298_{241}	181_{282}	121_{150}
	$T = 200$	343_{211}	331_{188}	212_{155}	284_{201}	87_{60}	79_{61}
	$T = 400$	229_{121}	265_{125}	189_{137}	166_{117}	57_{35}	49_{35}

Note: Means and standard deviations are reported for each case, and the reported values are actual values multiplied by 1000.

Results for the feasible constrained LASSO estimator $\hat{\beta}$ using data generated as in Example 1 are reported in Figure 1 and 2 as follow. For $T \in \{400, 800\}$ and $N \in \{1.2T, 2.4T\}$, the simulated L_2 errors of the standard LASSO estimator $\tilde{\beta}$, infeasible constrained LASSO estimator $\check{\beta}$, and the feasible constrained LASSO estimator $\hat{\beta}$ are separately visualized along the path of the tunning parameter λ. In particular, we replicate the experiment 100 times in each setting and report the mean of the realized L_2 error. To compute the LASSO estimator under linear constraints, we employ the alternating direction method of multipliers (ADMM) by Gaines and Zhou (2016).

Table 4: RMSE of $\hat{A}$, $\hat{\gamma}$, and $\hat{\psi}$ for Example 2.1 and 2.2 with $\nu = 0.5$.

		$\|\hat{A} - A\|_2$		$\|\hat{\gamma} - \gamma\|_2$		$\|\hat{\psi} - \psi\|_2$	
		$N = 400$	$N = 500$	$N = 400$	$N = 500$	$N = 400$	$N = 500$
DGP2.1	$T = 100$	653_{220}	709_{255}	205_{150}	168_{130}	216_{123}	217_{166}
	$T = 200$	626_{164}	676_{270}	125_{91}	148_{114}	168_{119}	186_{191}
	$T = 400$	552_{161}	575_{92}	95_{67}	87_{70}	127_{120}	107_{64}
DGP2.2	$T = 100$	849_{370}	930_{400}	339_{269}	292_{228}	740_{428}	716_{408}
	$T = 200$	801_{338}	824_{357}	193_{148}	278_{201}	745_{460}	639_{377}
	$T = 400$	646_{285}	784_{367}	176_{136}	164_{112}	561_{370}	571_{335}

Note: Means and standard deviations are reported for each case, and the reported values are actual values multiplied by 1000.

Figure 1: Simulated L_2 Errors ($\nu = 0$)

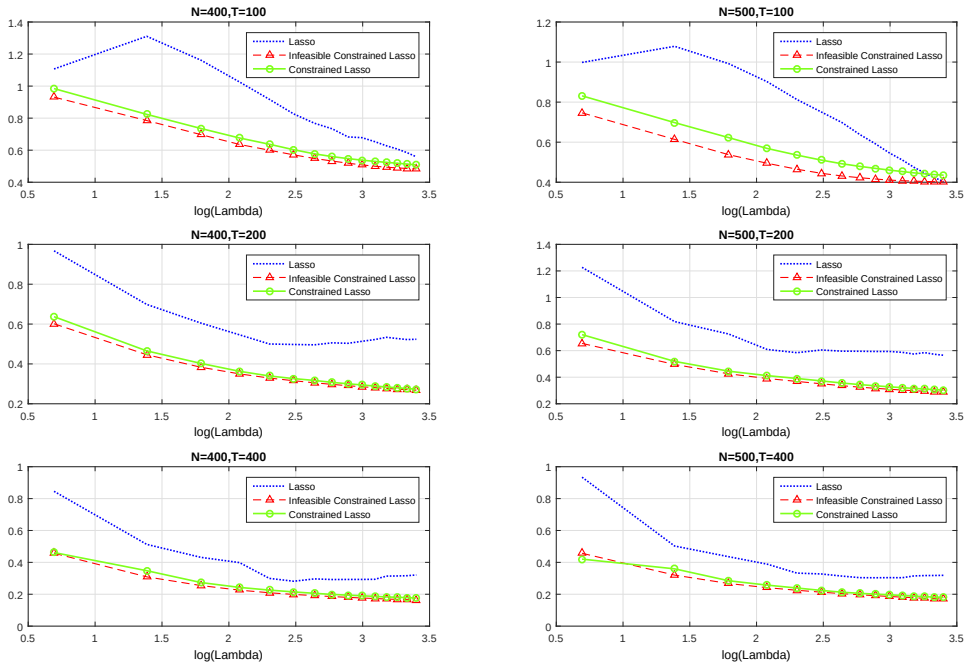
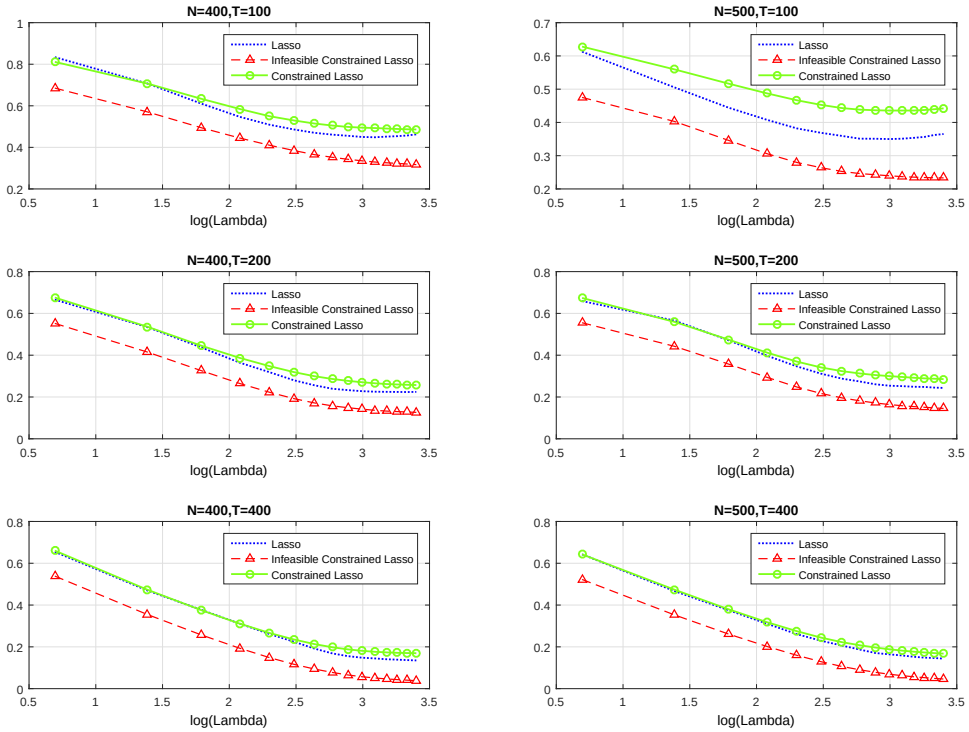


Figure 1 shows that when factors are strong (i.e. $\nu = 0$), the simulated L_2 error of the feasible constrained LASSO estimator $\hat{\beta}$ is close to that of the infeasible constrained LASSO estimator $\check{\beta}$ along the path of the tunning parameter λ, and is much smaller than the simulated L_2 error of the unconstrained LASSO estimator $\tilde{\beta}$. However, in Figure 2, we find that the performance of the feasible constrained LASSO estimator $\hat{\beta}$ is unsatisfactory,

Figure 2: Simulated L_2 Errors ($\nu = 0.5$)



and could be even worse than the unconstrained LASSO estimator $\tilde{\beta}$ at some values of the tunning parameter λ. This finding, however, is expected, which is due to the presence of weak factors as discussed in Section 4. When there are weak factors, the dominant term of the error bound is the term capturing the estimation error of the factor model, which has slower convergence rate than that of the term capturing the estimation error of the LASSO estimator.

6 Conclusions

This paper proposes a method to improve the performance of existing spillover estimators by using a latent factor structure in the variables that generate the spillovers. Specifically, a latent factor structure implies linear constraints on the parameters characterizing the spillovers, and these factor-induced constraints can be incorporated into the estimation

process to improve the estimation of spillovers. In particular, the L_2 error bound of the LASSO estimator under the feasible factor-induced constraints is derived. Compared with the unconstrained estimator, the LASSO estimator under the feasible factor-induced constraints is more accurate in the sense that it has approximately sharper error bound. Also, we note that the strength of the latent factors is critical. In particular, strong factors always provide better constraints than weak factors, which is due to the fact that the estimated factor loadings and corresponding factors have faster convergence rate when the factors are strong.

References

Ahn, S. C. and A. R. Horenstein (2013). Eigenvalue ratio test for the number of factors. *Econometrica 81*(3), 1203–1227.

Bai, J. and Y. Liao (2013). Statistical inferences using large estimated covariances for panel data and factor models.

Bai, J. and S. Ng (2002). Determining the number of factors in approximate factor models. *Econometrica 70*(1), 191–221.

Bai, J. and S. Ng (2006). Confidence intervals for diffusion index forecasts and inference for factor-augmented regressions. *Econometrica 74*(4), 1133–1150.

Bickel, P. J., Y. Ritov, and A. B. Tsybakov (2009). Simultaneous analysis of lasso and dantzig selector. *The Annals of Statistics*, 1705–1732.

Blume, L. E., W. A. Brock, S. N. Durlauf, and R. Jayaraman (2015). Linear social interactions models. *Journal of Political Economy 123*(2), 444–496.

Bradley, R. C. (1985). On the central limit question under absolute regularity. *The Annals of Probability*, 1314–1325.

Choi, I. (2012). Efficient estimation of factor models. *Econometric Theory 28*(02), 274–308.

De Paula, A. (2017). Econometrics of network models. In *Advances in Economics and Econometrics: Theory and Applications, Eleventh World Congress*, pp. 268–323. Cambridge University Press Cambridge.

Gaines, B. R. and H. Zhou (2016). Algorithms for fitting the constrained lasso. *arXiv preprint arXiv:1611.01511*.

James, G. M., C. Paulson, and P. Rusmevichientong (2012). The constrained lasso. Technical report, Citeseer.

Lam, C. and P. Souza (2013). Regularization for high-dimensional spatial models using the adaptive lasso. Technical report, LSE Working Paper.

Lam, C., Q. Yao, et al. (2012). Factor modeling for high-dimensional time series: inference for the number of factors. *The Annals of Statistics 40*(2), 694–726.

Lam, C., Q. Yao, and N. Bathia (2011). Estimation of latent factors for high-dimensional time series. *Biometrika 98*(4), 901–918.

Manresa, E. (2013). Estimating the structure of social interactions using panel data. *Unpublished Manuscript. CEMFI, Madrid*.

Negahban, S., B. Yu, M. J. Wainwright, and P. K. Ravikumar (2009). A unified framework for high-dimensional analysis of m-estimators with decomposable regularizers. In *Advances in Neural Information Processing Systems*, pp. 1348–1356.

Onatski, A. (2009). Testing hypotheses about the number of factors in large factor models. *Econometrica 77*(5), 1447–1479.

Rio, E. (1999). *Théorie asymptotique des processus aléatoires faiblement dépendants*, Volume 31. Springer Science & Business Media.

Stock, J. H. and M. W. Watson (2002). Forecasting using principal components from a large number of predictors. *Journal of the American Statistical Association 97*(460), 1167–1179.

Tibshirani, R. (1996). Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society. Series B (Methodological)*, 267–288.

Zou, H. (2006). The adaptive lasso and its oracle properties. *Journal of the American Statistical Association 101*(476), 1418–1429.

Appendix A Proofs of Main Results

Let $Y = (y_1, \cdots, y_T)' \in \mathbb{R}^T$, $\epsilon = (\epsilon_1, \cdots, \epsilon_T)' \in \mathbb{R}^T$, $X = (x_1, \cdots, x_T)' \in \mathbb{R}^{T\times N}$, $F = (f_1, \cdots, f_T)' \in \mathbb{R}^{T\times r}$, $U = (u_1, \cdots, u_T)' \in \mathbb{R}^{T\times N}$, and $e = U\beta + \epsilon \in \mathbb{R}^T$. We can then rewrite (1) as $X = FA' + U$ and (3) as $Y = i_T\gamma + F\psi + e$. Moreover, let $\hat{F} = (\hat{f}_1, \cdots, \hat{f}_T)' \in \mathbb{R}^{T\times r}$, $M_T = I_T - T^{-1}i_T i_T'$, where I_T is the $T \times T$ identity matrix and i_T is the T-vector whose components are all one.

Proof of Theorem 1: By $Y = i_T\gamma + \hat{F}\psi + [(F - \hat{F})\psi + U\beta + \epsilon]$, we have

$$\hat{\psi} - \psi = \left(\hat{F}'M_T\hat{F}\right)^{-1}\hat{F}'M_T[(F - \hat{F})\psi + U\beta + \epsilon].$$

Given the results in Lemma 2 (2), it is suffice to study the rate of $\|\hat{F}'M_T(F - \hat{F})\|_2$. Let $R_0 = (\hat{A} - A)'X'M_TXA + A'X'M_TX(\hat{A} - A) + (\hat{A} - A)'X'M_TX(\hat{A} - A)$, which gives $\hat{F}'M_T(F - \hat{F}) = -A'X'M_TUA + (\hat{A} - A)'X'M_TF - R_0$. Following similar arguments as in the proof of Lemma 1, for some positive constant $\rho < 1 - v$, we have

$$\|A'X'M_TUA\|_2 = O_p(TN^\rho + \sqrt{T}N),$$
$$\|(\hat{A} - A)'X'M_TF - R_1\|_2 = O_p(\sqrt{T}N),$$

which implies $\|\hat{F}'M_T(F - \hat{F})\|_2 = O_p(TN^\rho + \sqrt{T}N)$. Thus, the conclusion follows by Lemma 2 (1). ■

Proof of Theorem 2: Let $\hat{\delta} \equiv \hat{\beta} - \beta$. Also, for a d-vector v and an index set $I \subset \{1, \cdots, d\}$, let $\Pi_I(v) = v_I$.

Step 1: Under Assumption 2, $\hat{\delta} \in \Delta_\zeta$ if $\lambda \geq \zeta\|X'\epsilon\|_\infty$ for some constant $\zeta \geq 1$ and $\|X_\mathcal{T}\|_2^2/\lambda = 0$.

Since $\|\hat{A} - A\|_2 \xrightarrow{p} 0$, we focus on the case when $\hat{A}'_{\mathcal{T}}$, which gives

$$b_{\mathcal{T}} = (\hat{A}'_{\mathcal{T}})^{-1}(\hat{\psi} - \hat{A}'_{\mathcal{T}^c} b_{\mathcal{T}^c}),$$

for any $b \in \mathbb{R}^N$ satisfying $\hat{\psi} = \hat{A}'b$. Thus, the constrained problem (5) with respect to $b \in \mathbb{R}^N$ is equivalent to the unconstrained problem with respect to $b_2 \in \mathbb{R}^{N-r}$ as follows.

$$\min_{b_2 \in \mathbb{R}^{N-r}} \frac{1}{2}\|\tilde{Y} - \tilde{X}_{\mathcal{T}^c} b_2\|_2^2 + \lambda \left(\|b_{\mathcal{T}}(b_2)\|_1 + \|b_2\|_1\right), \tag{A.1}$$

where $\tilde{Y} = Y - X_{\mathcal{T}}(\hat{A}'_{\mathcal{T}})^{-1}\hat{\psi}$, $\tilde{X}_{\mathcal{T}^c} = X_{\mathcal{T}^c} - X_{\mathcal{T}}(\hat{A}'_{\mathcal{T}})^{-1}(\hat{A}'_{\mathcal{T}^c})$, and $b_{\mathcal{T}}(b_2) = (\hat{A}'_{\mathcal{T}})^{-1}\hat{\psi} - (\hat{A}'_{\mathcal{T}})^{-1}\hat{A}'_{\mathcal{T}^c} b_2$. Let $\hat{b}_2$ denote the solution to (A.1). The equivalence between (5) and (A.1) thus implies $\hat{\beta} = (b_{\mathcal{T}}(\hat{b}_2), \hat{b}_2)$. By the optimality of $\hat{b}_2$, we have

$$\begin{aligned}
0 \geq & -(\tilde{Y} - \tilde{X}_{\mathcal{T}^c}\beta_{\mathcal{T}^c})'\tilde{X}_{\mathcal{T}^c}(\hat{b}_2 - \beta_{\mathcal{T}^c}) + \frac{1}{2}\|\tilde{X}_{\mathcal{T}^c}(\hat{b}_2 - \beta_{\mathcal{T}^c})\|_2^2 \\
& + \lambda\left(\|b_{\mathcal{T}}(\hat{b}_2)\|_1 - \|b_{\mathcal{T}}(\beta_{\mathcal{T}^c})\|_1 + \|\hat{b}_2\|_1 - \|\beta_{\mathcal{T}^c}\|_1\right) \\
= & -(\epsilon - X_{\mathcal{T}}\delta_f)'(X\hat{\delta} - X_{\mathcal{T}}\delta_f) + \frac{1}{2}\|X\hat{\delta} - X_{\mathcal{T}}\delta_f\|_2^2 \\
& + \lambda\left(\|\hat{\beta}\|_1 - \|\beta_{\mathcal{T}} + \delta_f\|_1 - \|\beta_{\mathcal{T}^c}\|_1\right) \\
\geq & -\epsilon' X\hat{\delta} + \epsilon' X_{\mathcal{T}}\delta_f - \frac{1}{2}\|X_{\mathcal{T}}\delta_f\|_2^2 + \frac{1}{2}\|X\hat{\delta}\|_2^2 \\
& + \lambda\left(\|\hat{\beta}\|_1 - \|\beta\|_1 - \|\delta_f\|_1\right),
\end{aligned} \tag{A.2}$$

where the equality follows by

$$\begin{aligned}
\tilde{Y} - \tilde{X}_{\mathcal{T}^c}\beta_{\mathcal{T}^c} &= \epsilon - X_{\mathcal{T}}\delta_f, \\
\tilde{X}_{\mathcal{T}^c}(\hat{b}_2 - \beta_{\mathcal{T}^c}) &= X\hat{\delta} - X_{\mathcal{T}}\delta_f, \\
b_{\mathcal{T}}(\beta_{\mathcal{T}^c}) &= \beta_{\mathcal{T}} + \delta_f,
\end{aligned}$$

with $\delta_f = (\hat{A}'_{\mathcal{T}})^{-1}\hat{\psi} - (A'_{\mathcal{T}})^{-1}\psi - \left[(\hat{A}'_{\mathcal{T}})^{-1}\hat{A}'_{\mathcal{T}^c} - (A'_{\mathcal{T}})^{-1}A'_{\mathcal{T}^c}\right]\beta_{\mathcal{T}^c}$.

By Hölder's inequality and $\lambda \geq \zeta \|X'\epsilon\|_\infty$, we have

$$\begin{aligned} -\epsilon' X\hat{\delta} &\geq -|\epsilon' X\hat{\delta}| \geq -\|\epsilon' X\|_\infty \|\hat{\delta}\|_1 \geq -\frac{\lambda}{\zeta}\|\hat{\delta}\|_1, \\ \epsilon' X_{\mathcal{T}}\delta_f &\geq -|\epsilon' X_{\mathcal{T}}\delta_f| \geq -\|\epsilon' X_{\mathcal{T}}\|_\infty \|\delta_f\|_1 \geq -\frac{\lambda}{\zeta}\|\delta_f\|_1, \end{aligned}$$

Thus, noting $\|X\hat{\delta}\|_2^2 \geq 0$ and $\|X_{\mathcal{T}}\|_2^2/\lambda = 0$, (A.2) implies

$$0 \geq \lambda \left(\|\hat{\beta}\|_1 - \|\beta\|_1 - \frac{1}{\zeta}\|\hat{\delta}\|_1 - \frac{\zeta+1}{\zeta}\|\delta_f\|_1 \right). \tag{A.3}$$

Let $\mathcal{S} = \mathrm{supp}(\beta_{\mathcal{T}^c})$. By Lemma 5 of James et al. (2012), we have

$$\begin{aligned} &\|\hat{\beta}\|_1 - \|\beta\|_1 - \frac{1}{\zeta}\|\hat{\delta}\|_1 - \frac{\zeta+1}{\zeta}\|\delta_f\|_1 \\ &= \left(\|b_{\mathcal{T}}(\hat{b}_2)\|_1 - \|\beta_{\mathcal{T}}\|_1 - \frac{1}{\zeta}\|b_{\mathcal{T}}(\hat{b}_2) - \beta_{\mathcal{T}}\|_1 \right) + \left(\|\hat{b}_2\|_1 - \|\beta_{\mathcal{T}^c}\|_1 - \frac{1}{\zeta}\|\hat{b}_2 - \beta_{\mathcal{T}^c}\|_1 \right) - \frac{\zeta+1}{\zeta}\|\delta_f\|_1 \\ &\geq -\frac{\zeta+1}{\zeta}\|b_{\mathcal{T}}(\hat{b}_2) - \beta_{\mathcal{T}}\|_1 - \frac{\zeta+1}{\zeta}\|\Pi_{\mathcal{S}}(\hat{b}_2 - \beta_{\mathcal{T}^c})\|_1 + \frac{\zeta-1}{\zeta}\|\Pi_{\mathcal{S}^c}(\hat{b}_2 - \beta_{\mathcal{T}^c})\|_1 - \frac{\zeta+1}{\zeta}\|\delta_f\|_1. \end{aligned} \tag{A.4}$$

The conclusion follows by (A.3), (A.4), and $\lambda > 0$ and $\zeta > 1$.

Step 2: Under Assumption 2, $\lambda > \zeta\|\epsilon' X\|_\infty$ and $\|X_{\mathcal{T}}\|_2^2/\lambda = 0$ w.p.a.1.

First statement follows by Lemma 3 and

$$\begin{aligned} P\left(\lambda < \zeta\|\epsilon' X\|_\infty\right) &\leq N \max_{1\leq j\leq N} P\left(\left| \sum_{t=1}^{T} \epsilon_t X_{t,j} \right| > \frac{\lambda}{\zeta} \right) \\ &\leq 4N \max_{1\leq j\leq N} \left\{ \exp\left[-\frac{\lambda^2 \log 2}{32\zeta^2 T V_j} \right] + 4C\zeta T\lambda^{-1} \exp\left[-\frac{c^2(4\zeta T V_j)^\phi}{\lambda^\phi} \right] \right\} \\ &\leq 4N \exp\left[-\frac{\lambda^2 \log 2}{32\zeta^2 T u} \right] + 16C\zeta N T\lambda^{-1} \exp\left[-\frac{c^2(4\zeta T\, l)^\phi}{\lambda^\phi} \right] \\ &= 4\epsilon_p + C_1 N \sqrt{\frac{T}{\log(\frac{N}{\epsilon_p})}} \exp\left(-C_2 \left[\frac{T}{\log(\frac{N}{\epsilon_p})} \right]^{\frac{\phi}{2}} \right) \\ &\to 0, \end{aligned}$$

where the first inequality comes from the union bound, the second inequality follows by Lemma 3 and Assumption 2 (4), the third inequality follows by $l \leq \min_{1\leq j\leq N} V_j \leq \max_{1\leq j\leq N} V_j < u$ from Assumption 2 (4), and the equality follows by choosing $\lambda = K\sqrt{T\log(\frac{N}{\epsilon_p})}$ with $K = 4\zeta\sqrt{\frac{2u}{\log 2}}$, $\log(\frac{N}{\epsilon_p}) = o(T)$, and $\epsilon_p \to 0$, and $\log(\frac{N}{\epsilon_p}) = O(\log N)$, $\log N = o(T^{1/3})$ and $\phi > 1$ from Assumption 2. By a similar argument, we can show $\frac{\sum_{j=1}^{r}\sum_{t=1}^{T} X^2_{\mathcal{T}_{t,j}}}{\lambda} \xrightarrow{a.s.} 0$, and the second statement follows.

Step 3: Note that

$$\begin{aligned}
&\|\hat{\beta}\|_1 - \|\beta\|_1 - \frac{1}{\zeta}\|\hat{\delta}\|_1 \\
&\geq \frac{\zeta+1}{\zeta}\|b_{\mathcal{T}}(\hat{b}_2) - \beta_{\mathcal{T}}\|_1 - \frac{\zeta+1}{\zeta}\|\Pi_{\mathcal{S}}(\hat{b}_2 - \beta_{\mathcal{T}^c})\|_1 \\
&\geq -\frac{(\zeta+1)\max\{\sqrt{s-r}, \sqrt{r}\}}{\zeta}(\|b_{\mathcal{T}}(\hat{b}_2) - \beta_{\mathcal{T}}\|_2 + \|\Pi_{\mathcal{S}}(\hat{b}_2 - \beta_{\mathcal{T}^c})\|_2) \\
&\geq -\frac{\sqrt{2}(\zeta+1)\max\{\sqrt{s-r}, \sqrt{r}\}}{\zeta}\|\hat{\delta}\|_2,
\end{aligned} \tag{A.5}$$

where the first inequality follows by similar arguments as in (A.4) and $\zeta > 1$, the second inequality follows by $\|b_{\mathcal{T}}(\hat{b}_2) - \beta_{\mathcal{T}}\|_1 \leq \sqrt{r}\|b_{\mathcal{T}}(\hat{b}_2) - \beta_{\mathcal{T}}\|_2$ and $\|\Pi_{\mathcal{S}}(\hat{b}_2 - \beta_{\mathcal{T}^c})\|_1 \leq \sqrt{s-r}\|\Pi_{\mathcal{S}}(\hat{b}_2 - \beta_{\mathcal{T}^c})\|_2$, and the third inequality follows by $\|b_{\mathcal{T}}(\hat{b}_2) - \beta_{\mathcal{T}}\|_2 + \|\Pi_{\mathcal{S}}(\hat{b}_2 - \beta_{\mathcal{T}^c})\|_2 \leq \sqrt{2}\|\hat{\delta}\|_2$.

By similar argument as in (A.2), we have

$$\begin{aligned}
0 &\geq \|X\hat{\delta}\|_2^2 - \|X_{\mathcal{T}}\delta_f\|_2^2 + 2\lambda\left(\|\hat{\beta}\|_1 - \|\beta\|_1 - \frac{1}{\zeta}\|\hat{\delta}\|_1 - \frac{\zeta+1}{\zeta}\|\delta_f\|_1\right) \\
&\geq \|X\hat{\delta}\|_2^2 - \frac{2\sqrt{2}(\zeta+1)\max\{\sqrt{s-r}, \sqrt{r}\}\lambda}{\zeta}\|\hat{\delta}\|_2 - \|X_{\mathcal{T}}\delta_f\|_2^2 - \frac{2(\zeta+1)\lambda}{\zeta}\|\delta_f\|_1 \\
&\geq T\kappa^2\|\hat{\delta}\|_2^2 - \frac{2\sqrt{2}(\zeta+1)\max\{\sqrt{s-r}, \sqrt{r}\}\lambda}{\zeta}\|\hat{\delta}\|_2 - \frac{2(\zeta+1)\sqrt{r}\lambda}{\zeta}\|\delta_f\|_2,
\end{aligned} \tag{A.6}$$

where the second inequality follows by (A.5), and the last inequality follows by Assumption 2 (3), $\|X_{\mathcal{T}}\delta_f\|_2^2 \leq \|X_{\mathcal{T}}\|_2^2\|\delta_f\|_2^2$ with $\|X_{\mathcal{T}}\|_2^2/\lambda = 0$ w.p.a.1, and $\sqrt{r}\|\delta_f\|_2 \geq \|\delta_f\|_1$. The conclusion follows by solving (A.6) as an inequality with respect to $\|\hat{\delta}\|_2$. ■

Appendix B Proofs of Lemmas

Lemma 2 *Under Assumption 1,*

(i) $T^{-1}N^{\nu-1}\hat{F}'M_T\hat{F} \xrightarrow{p} M_F$, *where* $M_F \in \mathbb{R}^{r\times r}$ *is a positive definite matrix with* $\lambda_{\min}(M_F) \geq c$ *for some* $c > 0$.

(ii) $\|\hat{F}'M_TU\|_2 = O_p(TN^{\rho} + \sqrt{T}N)$ *for a positive constant* $\rho < 1 - v$, *and* $\|\hat{F}'M_T\epsilon\| = O_p(\sqrt{NT})$.

Proof of Lemma 2: For (i), let $R_1 = (\hat{A} - A)'X'M_TXA + A'X'M_TX(\hat{A} - A) + (\hat{A} - A)'X'M_TX(\hat{A} - A)$. By $\hat{F} = X\hat{A}$, we have

$$\hat{F}'M_T\hat{F} = A'X'M_TXA + R_1.$$

Since $\|A\|_2 = 1$ and $\|\hat{A} - A\|_2 = O_p(N^{\nu}T^{-1/2}) = o_p(1)$, R_1 is dominated by $A'X'M_TXA$ as $T \to \infty$. By $X = FA' + U$ and $A'A = I_r$, we have

$$\begin{aligned}
& T^{-1}N^{\nu-1}A'X'M_TXA \\
= \; & T^{-1}N^{\nu-1}F'M_TF + T^{-1}N^{\nu-1}F'M_TUA + T^{-1}N^{\nu-1}A'U'M_TF + T^{-1}N^{\nu-1}A'U'M_TUA \\
= \; & N^{\nu-1}\Sigma_f + N^{\nu-1}\Sigma_{fu}A + N^{\nu-1}A'\Sigma_{uf} + N^{\nu-1}A'\Sigma_uA + R_2
\end{aligned}$$

where $R_2 = N^{\nu-1}\left(T^{-1}F'M_TF - \Sigma_F\right) + N^{\nu-1}\left(T^{-1}F'M_TU - \Sigma_{fu}\right)A + N^{\nu-1}A'\left(T^{-1}U'M_TF - \Sigma_{uf}\right) + N^{\nu-1}A'\left(T^{-1}U'M_TU - \Sigma_u\right)A$. By Lemma 2 of Lam et al. (2011), if $N^{\nu}T^{-1/2} = o(1)$, $\|R_2\|_2 = o(1)$. Thus, (1) follows by

$$\begin{aligned}
& \lambda_{\min}\left(N^{\nu-1}\Sigma_f + N^{\nu-1}\Sigma_{fu}A + N^{\nu-1}A'\Sigma_{uf} + N^{\nu-1}A'\Sigma_uA\right) \\
\geq \; & N^{\nu-1}\lambda_{\min}\left(\Sigma_f\right) + N^{\nu-1}\lambda_{\min}\left(A'\Sigma_uA\right) + 2N^{\nu-1}\lambda_{\min}\left(\Sigma_{fu}A\right) \\
\geq \; & N^{\nu-1}\|\Sigma_f\|_{\min} + N^{\nu-1}\|\Sigma_u\|_{\min} + o(1) \\
\geq \; & c,
\end{aligned}$$

for some $c > 0$, where the first inequality follows by $\lambda_{\min}(G_1+G_2) \geq \lambda_{\min}(G_1)+\lambda_{\min}(G_2)$,

and the second inequality follows by $\lambda_{\min}(G) = \|G\|_{\min}$ for real symmetric matrix G, $\|G_1 G_2\|_{\min} \geq \|G_1\|_{\min}\|G_2\|_{\min}$, $\|A\|_{\min} = 1$, and $\lambda_{\min}(A\Sigma_{fu}) = o(N^{1-v})$ from Assumption 1 (2), and the equality follows by $\|\Sigma_f\|_{\min} \asymp N^{1-\nu}$.

For (ii), let $R_3 = (\hat{A} - A)'X'M_T U$, which gives $\hat{F}'M_T U = A'X'M_T U + R_3$. Since $A'X'M_T U$ dominates R_3, we focus on the rate of $A'X'M_T U$, which follows by

$$\begin{aligned} T^{-1}\|A'X'M_T U\|_2 &\leq \|T^{-1}F'M_T U\|_2 + \|T^{-1}U'M_T U\|_2 \\ &\leq \|\Sigma_{fu}\|_2 + \|\Sigma_u\|_2 + \|T^{-1}F'M_T U - \Sigma_{fu}\|_2 + \|T^{-1}U'M_T U - \Sigma_u\|_2 \\ &= O(N^{\rho}) + O(N^{\rho}) + O_p(N^{1-\frac{\nu}{2}}T^{-\frac{1}{2}}) + O_p(NT^{-\frac{1}{2}}) \\ &= O_p(N^{\rho} + NT^{-\frac{1}{2}}), \end{aligned}$$

where the first inequality follows by $X = FA' + U$, $A'A = I_r$, $\|G_1 G_2\|_2 \leq \|G_1\|_2\|G_2\|_2$, $\|A\|_2 = 1$, and the triangular inequality, the first equality follows by $\|\Sigma_{fu}\|_2 = O(N^{\rho})$ and $\|\Sigma_u\|_2 = O(N^{\rho})$ with $\rho < 1 - \nu$, and Lemma 2 of Lam et al. (2011).

For the remaining part of (2), let $R_4 = (\hat{A} - A)'X'M_T\epsilon$, which gives $\hat{F}'M_T\epsilon = A'X'M_T\epsilon + R_3$. Since $A'X'M_T\epsilon$ dominates R_3, we focus on the rate of $A'X'M_T\epsilon$, which follows by

$$\begin{aligned} T^{-1}\|A'X'M_T\epsilon\|_2 &\leq \|T^{-1}F'M_T\epsilon\|_2 + \|T^{-1}U'M_T\epsilon\|_2 \\ &\leq \|\Sigma_{f,\epsilon}\|_2 + \|\Sigma_{u,\epsilon}\|_2 + \|T^{-1}F'M_T\epsilon - \Sigma_{f,\epsilon}\|_2 + \|T^{-1}U'M_T\epsilon - \Sigma_{u,\epsilon}\|_2 \\ &= O_p(N^{\frac{1}{2}}T^{-\frac{1}{2}}) + O_p(N^{\frac{1}{2}}T^{-\frac{1}{2}}) + O_p(N^{\frac{1-\nu}{2}}T^{-\frac{1}{2}}) + O_p(N^{\frac{1}{2}}T^{-\frac{1}{2}}) \\ &= O_p(N^{\frac{1}{2}}T^{-\frac{1}{2}}), \end{aligned}$$

where the first inequality follows by $X = FA' + U$, $A'A = I_r$, $\|G_1 G_2\|_2 \leq \|G_1\|_2\|G_2\|_2$, $\|A\|_2 = 1$, and the triangular inequality, the second inequality follows by subtracting $\Sigma_{f,\epsilon}$ and $\Sigma_{u,\epsilon}$, and the triangular inequality, and the first equality follows by $\|\Sigma_{f\epsilon}\|_2 = O\left(N^{\frac{1}{2}}T^{-\frac{1}{2}}\right)$ and $\|\Sigma_{u\epsilon}\|_2 = O\left(N^{\frac{1}{2}}T^{-\frac{1}{2}}\right)$ from Assumption 1 (ii), and similar arguments as in Lemma 2 of Lam et al. (2011). ■

Lemma 3 *Let $\{U_t\}_{t=1}^T$ be a strongly mixing sequence of real-valued and centered random variables with mixing coefficient $\alpha(t)$. There are constants ϕ_1 and $c > 0$ such that $\alpha(t) \leq \exp(-ct^{\phi_1})$, and there is a constant $\phi_2 > 0$ such that $\sup_t P(|U_t| > u) \leq \exp(1 - u^{\phi_2})$. Then, for any $\lambda \geq (TV)^{1/2}$, we have*

$$P\left(\left|\sum_{t=1}^{T} U_t\right| \geq 4\lambda\right) \leq 4\exp\left(-\frac{\lambda^2 \log 2}{2TV}\right) + 4CT\lambda^{-1}\exp\left(-\frac{c^2(TV)^{\phi}}{\lambda^{\phi}}\right)$$

, where $\phi = \frac{\phi_1\phi_2}{\phi_1+\phi_2}$ and $V = \sup_{1\leq t\leq T}\left(EU_t^2 + 2\sum_{s>t}|E(U_sU_t)|\right)$.

Proof of Lemma 3: This is an immediate corollary of Theorem 6.2 in Rio (1999). ■