ORACLE EFFICIENT VARIABLE SELECTION IN RANDOM AND FIXED EFFECTS PANEL DATA MODELS

ANDERS BREDAHL KOCK
AARHUS UNIVERSITY AND CREATES

Abstract. This paper generalizes the results for the Bridge estimator of Huang et al. (2008) to linear random and fixed effects panel data models which are allowed to grow in both dimensions. In particular, we show that the Bridge estimator is oracle efficient. It can correctly distinguish between relevant and irrelevant variables and the asymptotic distribution of the estimators of the coefficients of the relevant variables is the same as if only these had been included in the model, i.e. as if an oracle had revealed the true model prior to estimation.

In the case of more explanatory variables than observations we prove that the Marginal Bridge estimator can asymptotically correctly distinguish between relevant and irrelevant explanatory variables if the error terms are Gaussian. Furthermore, a partial orthogonality condition of the same type as in Huang et al. (2008) is needed to restrict the dependence between relevant and irrelevant variables.

Key words: Panel data, high dimensional modeling, variable selection, Bridge estimators, oracle property.
JEL codes: C1, C23.

1. Introduction

When building a model one of the first steps is to decide which variables to include. Sometimes theory can guide the researcher towards a set of potential explanatory variables but which variables in this set are relevant and which are to be left out? Huang et al. (2008) showed that the Bridge estimator is able to discriminate between relevant and irrelevant explanatory variables in a cross section setting with fixed covariates whose number is allowed to increase with the sample size. In fact, oracle efficient estimation has received quite some attention in the statistics literature in the recent years, see (among others) Zou (2006), Candes and Tao (2007), Fan and Lv (2008), and Meinshausen and Yu (2009). However, we are not aware of any similar results for panel data models. For the case of fewer explanatory variables than observations we show that the oracle efficiency of the Bridge estimator carries over to linear panel data models with random regressors

Date: August 31, 2011.
The author wishes to thank Svend Erik Graversen, Niels Haldrup, Michael Jansson, Jørgen Hoffmann-Jørgensen, Adrian Pagan, Timo Teräsvirta, Allan Würtz, the co-editor and an anonymous referee for help, comments and suggestions. Also thanks to Joel Horowitz for responding quickly to my emails. All errors and shortcomings are my responsibility. Financial support from CREATES funded by the Danish National Research Foundation is gratefully acknowledged.
Address for correspondence: Anders Bredahl Kock, School of Economics and Management, Aarhus University and CREATES. email: akock@creates.au.dk.

in the random and fixed effects settings. More precisely, it suffices that either the number of cross sectional units (N) or the number of observations within each cross sectional unit (T_N) goes to infinity in order to establish consistency and correct elimination of irrelevant variables. To obtain the oracle efficient asymptotic distribution (the distribution obtained by only including the relevant covariates) of the estimators of the nonzero coefficients, further restrictions are needed. In the classical setting of fixed T_N and large N these restrictions are satisfied. Further sufficient conditions for oracle efficiency are given. By fixing T_N and the number of covariates we obtain as a corollary that the asymptotic distribution of the estimators of the non-zero coefficients is exactly the classical fixed effects or random effects limit law.

If the set of potential explanatory variables is larger than the number of observations we show that the Marginal Bridge estimator of Huang et al. (2008) can be used to distinguish between relevant and irrelevant variables in random and fixed effects panel data models. A partial orthogonality condition restricting the dependence between the relevant and the irrelevant variables of the same type as in Huang et al. (2008) is imposed. Furthermore, the error terms must be Gaussian – a price paid for letting the covariates be random. The random covariates also rendered the maximum inequalities based on exponential Orlicz norms used in Huang et al. (2008) inapplicable. However, more simple maximum inequalities in L^q spaces can still be applied but the result is that the number of irrelevant variables must be $o(N^{q/2})$ for some $q \geq 1$ (this is for fixed T_N for comparability to the known cross sectional results) as opposed to $\exp(o(N))$ (a subexponential rate). Since q is arbitrary this still allows the number of irrelevant variables to increase at any polynomial rate. The number of relevant variables may still be $o(N^{1/2})$ (again T_N is considered fixed for comparison).

Furthermore, the Marginal Bridge estimator is very fast to implement which also makes it useful as an initial screening device to weed out the most irrelevant variables before initiating the actual modeling stage.

Since cross section data can be viewed as panel data with only one observation per individual, all our results are also valid for cross section data and hence generalize the results for these.

The plan of the paper is as follows. Section 2 puts forward the general framework. Section 3 introduces the Bridge estimator and its properties while Section 4 discusses the Marginal Bridge estimator. Section 5 illustrates the results by simulation and Section 6 concludes. Section 7 contains proofs of the propositions.

2. Setup and assumptions

Consider the following linear panel data model.

$$\tilde{y}_{it} = \tilde{\mathbf{x}}_{it}'\beta_0 + c_i + \tilde{\epsilon}_{it},\ i = 1, ..., N,\ t = 1, ..., T_N \tag{2.1}$$

$\tilde{\mathbf{x}}_{it}$ is a $p_N \times 1$ vector of covariates indicating that the number of covariates is allowed to increase with the sample size. The interpretation of (2.1) is that N individuals are observed in T_N time periods, totaling NT_N observations. The c_i indicate the unobserved heterogeneity, i.e. unobserved time invariant variables such as intelligence of an individual or start up capital of a firm. The $\tilde{\epsilon}_{it}$ are the idiosyncratic error terms. Some of the elements of β_0 may be zero. It is our objective to locate these while still estimating the nonzero coefficients consistently.

N as well as T_N are allowed to tend to infinity. However, all results are valid as long as N tends to infinity. Hence, the traditional large N, fixed T_N setting is covered. Notice that T_N is indexed by N. Some of our results put no restrictions on how T_N depends on N.

Equation (2.1) can equivalently be written as

$$\tilde{\mathbf{Y}}_{iN} = \tilde{\mathbf{X}}_{iN}\beta_0 + \mathbf{c}_{iN} + \tilde{\epsilon}_{iN},\ i = 1, ..., N, \tag{2.2}$$

where $\tilde{\mathbf{Y}}_{iN} = (\tilde{y}_{i1}, ..., \tilde{y}_{iT_N})',\ \tilde{\mathbf{X}}_{iN} = (\tilde{\mathbf{x}}_{i1}, ..., \tilde{\mathbf{x}}_{iT_N})',\ \tilde{\epsilon}_{iN} = (\tilde{\epsilon}_{i1}, ..., \tilde{\epsilon}_{iT_N})',\ \mathbf{c}_{iN} = c_{iN}\iota_{T_N},\ \iota'_{T_N} = (1, ..., 1),\ i = 1, ..., N.$

2.1. **Fixed Effects.** In the fixed effects setting one assumes:

(FE1) Random sampling: $(\tilde{\mathbf{X}}_{iN}, \mathbf{c}_{iN}, \tilde{\epsilon}_{iN})_{i=1}^N$ is i.i.d.

(FE2) $E\left(\tilde{x}_{itl}^4\right),\ E\left(\tilde{\epsilon}_{it}^4\right) < \infty,\ i = 1, ..., N,\ t = 1, ..., T_N,\ l = 1, ..., p_N$

(FE3) a) $E(\tilde{\epsilon}_{iN}|\tilde{\mathbf{X}}_{iN}, \mathbf{c}_{iN}) = 0$ and b) $E(\tilde{\epsilon}_{iN}\tilde{\epsilon}'_{iN}|\tilde{\mathbf{X}}_{iN}, \mathbf{c}_{iN}) = \sigma^2\mathbf{I}_{T_N},\ i = 1, ..., N$

where $\tilde{x}_{itl}$ is the l'th covariate of individual i in period t. For our proofs we may replace (FE3) by $E(\tilde{\epsilon}_{iN}|\tilde{\mathbf{X}}_{iN}) = 0$ and $E(\tilde{\epsilon}_{iN}\tilde{\epsilon}'_{iN}|\tilde{\mathbf{X}}_{iN}) = \sigma^2\mathbf{I}_{T_N}$ which is less restrictive but since (FE3) is standard in the literature we stick to this. Next, we carry out the forward orthogonal deviations transform of Arellano (2003), (page 17). This transformation removes the unobserved heterogeneity while keeping the error terms uncorrelated. In particular, define the $(T_N - 1) \times T_N$ matrix

$$\mathbf{D} = \begin{pmatrix} -1 & 1 & 0 & \cdots & 0 & 0 \\ 0 & -1 & 1 & & 0 & 0 \\ \vdots & & & \ddots & & \vdots \\ 0 & 0 & 0 & \cdots & -1 & 1 \end{pmatrix}$$

and multiply (2.2) through by $\left(\mathbf{DD}'\right)^{-1/2}\mathbf{D}$ to get

$$\mathbf{Y}_{iN} = \mathbf{X}_{iN}\beta_0 + \epsilon_{iN},\ i = 1, ..., N, \tag{2.3}$$

where $\mathbf{Y}_{iN} = \left(\mathbf{DD}'\right)^{-1/2}\mathbf{D}\tilde{\mathbf{Y}}_{iN},\ \mathbf{X}_{iN} = \left(\mathbf{DD}'\right)^{-1/2}\mathbf{D}\tilde{\mathbf{X}}_{iN}$ and $\epsilon_{iN} = \left(\mathbf{DD}'\right)^{-1/2}\mathbf{D}\tilde{\epsilon}_{iN}$. Clearly, (FE3) implies

(FE3') a) $E(\epsilon_{iN}|\mathbf{X}_{iN}) = 0$ and b) $E(\epsilon_{iN}\epsilon'_{iN}|\mathbf{X}_{iN}) = \sigma^2\mathbf{I}_{T_N-1}$

which is what will be used in the proofs. Arellano (2003) gives the specific form of the entries of $\left(\mathbf{DD}'\right)^{-1/2}\mathbf{D}$. The number of time series observations for each individual is reduced from T_N to $T_N - 1$ by the forward orthogonal deviations transform. However, for notational convenience, we will keep using T_N for the number of time series observations in the transformed model. In a cross section setting this transform does not need to be carried out.

Assumption (FE3) b) $E(\tilde{\epsilon}_{iN}\tilde{\epsilon}'_{iN}|\tilde{\mathbf{X}}_{iN}, \mathbf{c}_{iN}) = \sigma^2\mathbf{I}_{T_N}$ restricts the $\tilde{\epsilon}_{it}$ to be uncorrelated. This may be relaxed to $E(\tilde{\epsilon}_{iN}\tilde{\epsilon}'_{iN}|\tilde{\mathbf{X}}_{iN}, \mathbf{c}_{iN}) = \mathbf{S}$ where $\mathbf{S}$ is a known covariance matrix. In this case the forward orthogonal deviations transform is replaced by $\left(\mathbf{DD}'\right)^{-1/2}\mathbf{DS}^{-1/2}$. So heteroskedasticity can be handled as long as the structure of it is known – the transformation applied simply changes accordingly. If the heteroskedasticity is ignored the situation is more subtle and we will discuss this in more detail in Section 3.

2.2. **Random Effects.** In the random effects setting (FE1)-(FE3) are maintained while

(RE4) a) $E\left(\mathbf{c}_{iN}|\tilde{\mathbf{X}}_{iN}\right)=0$, b) $E\left(\mathbf{c}_{iN}\mathbf{c}_{iN}'|\tilde{\mathbf{X}}_{iN}\right)=\sigma_c^2\iota_{T_N}\iota_{T_N}'$ and c) σ and σ_c are known and finite[1]

are added. Part a) of this extra assumption restricts the dependence between $\mathbf{X}_{iN}$ and $\mathbf{c}_{iN}$ sufficiently in order to allow merging the latter with the error term while still being able to prove the desired results. Part b) specifies the conditional covariance structure of $\mathbf{c}_i$. RE4c) is needed to enable us to carry out the GLS transform below. The gain from these stronger assumptions is that they (as opposed to fixed effects) allow for the inclusion of a covariate which is constant over time and only varies over individuals. Defining $\mathbf{v}_{iN}=\mathbf{c}_{iN}+\tilde{\epsilon}_{iN}$, (FE3) and (RE4) imply $E\left(\mathbf{v}_{iN}|\tilde{\mathbf{X}}_{iN}\right)=0$ and

$$E(\mathbf{v}_{iN}\mathbf{v}_{iN}'|\tilde{\mathbf{X}}_{iN})=E([\mathbf{c}_{iN}+\tilde{\epsilon}_{iN}][\mathbf{c}_{iN}+\tilde{\epsilon}_{iN}]'|\tilde{\mathbf{X}}_{iN})=\begin{pmatrix}\sigma_c^2+\sigma^2 & \sigma_c^2 & \cdots & \sigma_c^2\\ \sigma_c^2 & \sigma_c^2+\sigma^2 & \cdots & \vdots\\ \vdots & & \ddots & \sigma_c^2\\ \sigma_c^2 & & & \sigma_c^2+\sigma^2\end{pmatrix}=\Omega$$

The presence of the unobserved heterogeneity renders the error terms correlated. Since the structure of the correlation is known, the correlation is easily removed by premultiplying (2.2) by[2] $\sigma\Omega^{-1/2}$ (GLS transform). This yields

$$\mathbf{Y}_{iN}=\mathbf{X}_{iN}\beta_0+\epsilon_{iN},\ i=1,...,N, \tag{2.4}$$

where $\mathbf{Y}_{iN}=\sigma\Omega^{-1/2}\tilde{\mathbf{Y}}_{iN},\ \mathbf{X}_{iN}=\sigma\Omega^{-1/2}\tilde{\mathbf{X}}_{iN}$ and $\epsilon_{iN}=\sigma\Omega^{-1/2}\mathbf{v}_{iN}$. Hence,

(RE3') a) $E(\epsilon_{iN}|\mathbf{X}_{iN})=0$ and b) $E(\epsilon_{iN}\epsilon_{iN}'|\mathbf{X}_{iN})=\sigma^2\mathbf{I}_{T_N}$

which is what will be used in the proofs. In a cross section setting the random effects transform does not need to be carried out. As was the case in the fixed effects setting, any known heteroskedasticity structure can be handled in the random effects setting as well, as long as Ω is known.

3. The Bridge estimator

The Bridge estimator estimates β_0 by minimizing

$$L_N(\beta)=\sum_{i=1}^{N}\sum_{t=1}^{T_N}\left(y_{it}-\mathbf{x}_{it}'\beta\right)^2+\lambda_N\sum_{k=1}^{p_N}|\beta_k|^\gamma \tag{3.1}$$

$$=\sum_{j=1}^{NT_N}\left(y_j-\mathbf{x}_j'\beta\right)^2+\lambda_N\sum_{k=1}^{p_N}|\beta_k|^\gamma,\qquad \gamma>0 \tag{3.2}$$

[1] In principle it suffices for most purposes that ratio σ_c/σ is known but it is hard to imagine a situation where the ratio is known while σ_c and σ are not. I wish to thank the co-editor for pointing this out.

[2] The sole reason for multiplying $\Omega^{-1/2}$ by σ is that (FE3') and (RE3') become identical except for the dimension of the covariance matrix. Since (FE3') and (RE3') are the assumptions used in the proofs this indicates that the proofs only have to be carried out in either the fixed or the random effects setting.

where summation from 1 to NT_N indicates summation over all time periods for each individual (So the first T_N terms in the sum correspond to all T_N observation on individual 1, the next T_N terms to all observations on individual 2 and so on. This convention is adopted in the sequel.). The Bridge estimator, denoted $\hat{\beta}_N$, may hence be seen as a sort of penalized/regularized least squares. The objective function consists of two parts; the first part being the least squares objective function and the second part penalizing parameters different from 0. The larger λ_N, the larger the penalty. For $\gamma = 1$ the minimizer of (3.2) could be called the LASSO panel estimator, (Tibshirani (1996)). For $\gamma = 2$ it could be called the ridge regression estimator (Tikhonov regularization) for panel data models. In a cross sectional setting the ridge regression is frequently used to deal with multicollinearity. The Tikhonov regularization is more generally used to solve ill-conditioned (singular) overdetermined systems of linear equations.

Let β_0 denote the true value of β where the dependence on N is suppressed as in Huang et al. (2008). Partition β_0 as $\beta_0 = (\beta_{10}', \beta_{20}')'$ where $\beta_{10} \neq 0$ is $k_N \times 1$ and $\beta_{20} = 0$ is $m_N \times 1$. Hence, the β_{10} are the coefficients corresponding to the relevant variables denoted $\mathbf{w}_{it}$. β_{20} are the coefficients of the irrelevant variables denoted $\mathbf{z}_{it}$. So $\mathbf{x}_{it}$ is partitioned as $\mathbf{x}_{it} = (\mathbf{w}_{it}', \mathbf{z}_{it}')'$. Accordingly, we define $\mathbf{X}_N = \left(\mathbf{x}_{11}, ..., \mathbf{x}_{NT_N}\right)'$, $\mathbf{W}_N = \left(\mathbf{w}_{11}, ..., \mathbf{w}_{NT_N}\right)'$ and $\mathbf{Z}_N = \left(\mathbf{z}_{11}, ..., \mathbf{z}_{NT_N}\right)'$. $\mathbf{\Sigma}_N = (NT_N)^{-1}\mathbf{X}_N'\mathbf{X}_N$ as well as $\mathbf{\Sigma}_{1N} = (NT_N)^{-1}\mathbf{W}_N'\mathbf{W}_N$ are the scaled Gram matrices of $\mathbf{X}_N$ and $\mathbf{W}_N$, respectively. Let ρ_{1N} and ρ_{2N} be the smallest and the largest eigenvalue of $\mathbf{\Sigma}_N$. Similarly, define τ_{1N} and τ_{2N} as the smallest and the largest eigenvalue of $\mathbf{\Sigma}_{1N}$. Set $\mathbf{W}_{iN} = \left(\mathbf{w}_{i1}, ..., \mathbf{w}_{iT_N}\right)'$ and for $\mathbf{x} \in \mathbf{R}^p$ $\|\mathbf{x}\| = \sqrt{\sum_{k=1}^p x_k^2}$ denotes the Euclidean norm on $\mathbf{R}^p$ stemming from the dot product. Finally, $\mathbf{x}_k = (x_{1,k}, ..., x_{NT_N,k})'$ is the vector containing all observations of the k'th explanatory variable.

Next, we state and discuss the assumptions needed to establish consistency and oracle efficiency of Bridge estimators in random and fixed effects panel data models. Notice how N and T_N enter symmetrically indicating that what matters is their product, i.e. the total number of observations, and not whether it is N or T_N which gets large (however, some theorems require further assumptions restricting the rate at which T_N increases relative to N).

(A1) $\frac{1}{NT_Np_N}\sum_{i=1}^N\sum_{t=1}^{T_N}\sum_{k=1}^{p_N} x_{itk}^2$ is bounded in L^1, i.e.,

$$\sup_{1\leq N<\infty} E\left(\frac{1}{NT_Np_N}\sum_{i=1}^N\sum_{t=1}^{T_N}\sum_{k=1}^{p_N} x_{itk}^2\right) = \sup_{1\leq N<\infty}\frac{1}{T_Np_N}\sum_{t=1}^{T_N}\sum_{k=1}^{p_N} E\left(x_{1tk}^2\right) = K < \infty$$

(A2) There exist constants $0 < \tau_1 < \tau_2 < \infty$ such that $\tau_1 \leq \tau_{1N} \leq \tau_{2N} \leq \tau_2$ *almost surely*

(A3) $\lambda_N(k_N/(NT_N))^{1/2} \to 0$

(A4) $\lambda_N\rho_{1N}^{2-\gamma}(NT_N)^{-\gamma/2}p_N^{\gamma/2-1} \to \infty$ *almost surely*

(A5) There exist constants $0 < b_0 < b_1 < \infty$ such that $b_0 \leq \min\left\{\left|\beta_{10j}\right| \;|1 \leq j \leq k_N\right\} \leq \max\left\{\left|\beta_{10j}\right| \;|1 \leq j \leq k_N\right\} \leq b_1$

(A6) $(p_N + \lambda_Nk_N)/(NT_N\rho_{1N}) \to 0$ *almost surely*

(A7) $\frac{\rho_{1N}\rho_{2N}^{1/2}}{p_N^{1/2}} \in O_p(1)$

Assumption (A1) may be dropped altogether if the covariates are normalized as $\frac{1}{NT_N}\sum_{j=1}^{NT_N} x_{jk}^2 = \frac{1}{NT_N}\sum_{i=1}^{N}\sum_{t=1}^{T_N} x_{itk}^2 = 1$ for all $1 \leq k \leq p_N$. Alternatively, (A1) is satisfied if $\{x_{itk}\}$ is bounded in L^2 – this is in turns satisfied if, e.g, the covariates are uniformly bounded or if T_N and p_N are fixed constants. If the covariates are identically distributed over time, then the assumption reduces to boundedness of the Cesàro sum $\frac{1}{p_N}\sum_{k=1}^{p_N} E\left(x_{11k}^2\right)$. Finiteness of p_N or convergence of $\left\{E\left(x_{11k}^2\right)\right\}_{k=1}^{\infty}$ are sufficient for this. Finally, it may be noted that convergence of $\frac{1}{T_N p_N}\sum_{t=1}^{T_N}\sum_{k=1}^{p_N} E\left(x_{1tk}^2\right)$ is also sufficient for the desired boundedness in L^1.

Huang et al. (2008) mention that assumption (A2) is likely to be satisfied in sparse systems, where k_N is relatively small.

Regarding condition (A3) one notices that if the number of relevant covariates k_N stays fixed $\lambda_N/(NT_N)^{1/2} \to 0$. Hence, $\lambda_N \in o((NT_N)^{1/2})$.

Assumption (A4): Assume $0 < a_1 < \rho_{1N} \leq \rho_{2N} < a_2 < \infty$ for some constants a_1 and a_2 and that the number of covariates stays constant. Then it must be the case that $\lambda_N(NT_N)^{-\gamma/2} \to \infty$. This excludes $\gamma \geq 1$ by (A3). Hence, $0 < \gamma < 1$ and $\lambda_N \in o((NT_N)^{1/2}) \cap \omega((NT_N)^{\gamma/2})$ where $\omega(g(N))$ is the set of functions that diverge to infinity when divided by $g(N)$ as $N \to \infty$.

Assumption (A5) requires that the non-zero coefficients are uniformly bounded away from 0 and infinity. This is trivially satisfied if the number of covariates is finite. Also note that all results remain valid (with slight modifications) if b_1 is replaced by a sequence b_{1N} which is allowed to tend to infinity.

By assumption (A3), assumption (A6) is satisfied if $0 < a_1 < \rho_{1N} < \rho_{2N} < a_2 < \infty$ for some constants a_1 and a_2 and the number of covariates is finite. Since the Gramian Σ_N is positive semidefinite (A6) also implies that $\rho_{1N} > 0$ in order for the condition to be well defined. This excludes $p_N > NT_N$ since the rank of Σ_N can be no larger than NT_N.

Assumption (A7) is satisfied if $0 < a_1 < \rho_{1N} < \rho_{2N} < a_2 < \infty$ for some constants a_1 and a_2.

Assumptions (A2)-(A6) are identical[3] to assumptions made in Huang et al. (2008). (A1) and (A7) are not made by Huang et al. (2008) but both these assumptions are redundant if the covariates are normalized as $\frac{1}{NT_N}\sum_{j=1}^{NT_N} x_{jk}^2 = 1$ for all $1 \leq k \leq p_N$ as done by these authors.

Our first theorem states that the Bridge estimator is consistent in the random as well as the fixed effects setting. Throughout we will assume that (FE1)-(FE3) (fixed effects setting) or (FE1)-(FE3) and (RE4) (random effects setting) are satisfied.

Theorem 1. *Let $\hat{\beta}_N$ denote the minimizer of (3.2). Suppose that $\gamma > 0$ and that conditions (A1), (A3), (A5), and (A6) hold. Then $||\hat{\beta}_N - \beta_N|| \in O_p(\min(h_N, h'_N))$ where $h_N = \rho_{1N}^{-1}(p_N/(NT_N))^{1/2}$ and $h'_N = \left[(p_N + \lambda_N k_N)/(NT_N\rho_{1N})\right]^{1/2}$.*

Theorem 1 shows the consistency of the Bridge estimator by assumption (A6). By considering h_N it follows that if there exists a constant a_1 such that $0 < a_1 < \rho_{1N}$ and p_N is constant the Bridge estimator converges at the same rate as the least squares estimator. The faster the arrival rate of new explanatory variables (p_N increases) the slower the rate of convergence of the Bridge estimator since h_N as well

[3] Since we allow for random covariates some of our assumptions must hold in an almost sure sense while the equivalent assumptions in Huang et al. (2008) must hold surely.

as h'_N are increasing in p_N. If ρ_{1N} tends to 0 (approaching a singular design) the convergence rate is also slowed down. It is also seen that N and T_N enter symmetrically. This is not immediate on the outset since only independence of $\{\mathbf{X}_{iN}\}_{i=1}^{\infty}$ has been assumed while the T_N rows of each $\mathbf{X}_{iN}$ may have any dependence structure between them. What provides the result is that $E(\epsilon_{iN}\epsilon'_{iN}|\mathbf{X}_{iN}) = \sigma^2\mathbf{I}_{T_N}$, i.e. the conditional uncorrelatedness of the error terms of each individual. This underscores the importance of orthogonalizing (in L^2) the error terms prior to applying the Bridge estimator. In the presence of unknown arbitrary heteroskedasticity, $E(\epsilon_{iN}\epsilon'_{iN}|\mathbf{X}_{iN}) = \mathbf{S}$, (which can not be orthogonalized) we were only able to prove that $||\hat{\beta}_N - \beta_N|| \in O_p(h'_N + (T_N-1)/\rho_{1N})$. So even in the situation of fixed T_N, consistency requires $\rho_{1N} \to \infty$ which is impossible.

The next theorem reveals that the Bridge estimator performs variable selection and gives the limiting law of the estimator of the nonzero coefficients.

Let $U_{1N} = \alpha' \left(E\left[\frac{1}{T_N}\mathbf{W}'_{1N}\mathbf{W}_{1N}\right]\right)^{-1} T_N^{-1/2}\mathbf{W}'_{1N}\epsilon_{1N}$.

Theorem 2. *Assume $0 < \gamma < 1$. Then under (A1)-(A7),*

(i) $\hat{\beta}_{2N} = 0$ *with probability converging to 1.*

(ii) *Let k_N be a fixed number k, α be a $k \times 1$ vector, and $s_N = \sqrt{\sigma^2\alpha'\mathbf{\Sigma}_{1N}^{-1}\alpha}$. If $\left\{U_{1N}^2\right\}_{N=1}^{\infty}$ is uniformly integrable,*

$$\frac{\max_{1\leq t\leq T_N} Var\left(w_{1tl}w_{1tm}\right)}{N} \to 0 \text{ for all } 1 \leq l, m \leq k$$

and

$$\lim_{N\to\infty} \frac{1}{NT_N}\sum_{j=1}^{NT_N} E\left(\mathbf{w}_j\mathbf{w}'_j\right) = \lim_{N\to\infty} E\left(\frac{1}{T_N}\mathbf{W}'_{1N}\mathbf{W}_{1N}\right)$$

exists then,

$$(NT_N)^{1/2}s_N^{-1}\alpha'\left(\hat{\beta}_{1N} - \beta_{10}\right) \xrightarrow{d} N(0,1)$$

Part (i) states that not only does $\hat{\beta}_{2N} \to 0$ in probability (Theorem 1), the Bridge estimator actually sets $\hat{\beta}_{2N} = 0$ with probability converging to 1. The latter of course implies the former while the converse is not true. The fact that $\hat{\beta}_{2N}$ is set exactly equal to 0 with probability converging to 1 means that the Bridge estimator performs variable selection.

Part (ii) of the theorem states that the asymptotic distribution of the estimators of the non zero coefficients is the same as if the true model had been known in advance – i.e. as if an oracle had revealed which variables to include and which to exclude. This is a very useful result in practice. One simply includes the whole set of potential explanatory variables. The irrelevant ones will be kicked out ($\hat{\beta}_{2N} = 0$ with probability converging to 1) while the relevant ones are estimated with the same asymptotic efficiency as if the irrelevant ones had been left out from the outset. However, notice that the price paid for letting the covariates be random is that k_N must be fixed. Alternatively, one may continue to let k_N increase in N while conditioning on the covariates and establish the limiting law along the lines of Huang et al. (2008).

Next we discuss conditions under which the requirements of part (ii) of Theorem 2 hold. The following Theorem gives sufficient conditions under which $\left\{U_{1N}^2\right\}_{N=1}^{\infty}$ is uniformly integrable.

Theorem 3. *$\left\{U_{1N}^2\right\}_{N=1}^{\infty}$ is uniformly integrable if either of the following conditions is satisfied.*

(i) *$T_N = T$ for a fixed T*
(ii) *The rows in $\mathbf{W}_{1N}$ are identically distributed and $\mathbf{W}_{iN} \perp\!\!\!\perp \epsilon_{iN},\ i = 1, ..., N$.*
(iii) *$\mathbf{W}_{1N}$ and ϵ_{1N} are uniformly bounded in N.*

The assumption $\max_{1 \leq t \leq T_N} Var\left(w_{1tl}w_{1tm}\right)/N \to 0$ for all $1 \leq l, m \leq k$ in part (ii) of Theorem 2 is not restrictive. It is clearly satisfied if T_N is fixed. It is also satisfied if $\max_{1 \leq t \leq T_N} E\left(\left[w_{1tl}w_{1tm}\right]^2\right) \leq M < \infty$ for all T_N and $1 \leq l, m \leq k$ (second moments uniformly bounded in t) which in turn is satisfied if the variables themselves are uniformly bounded in t. The assumption is also satisfied if $\mathbf{w}_{1t}$ are identically distributed across t. If the variances are linearly increasing, i.e. $Var\left(w_{1tl}w_{1tm}\right) = a_{lm}t$ for some $a_{lm} > 0$, it suffices that $T_N/N \to 0$.[4]

If T_N is fixed, $\lim_{N\to\infty} \frac{1}{NT_N}\sum_{j=1}^{NT_N} E\left(\mathbf{w}_j\mathbf{w}_j'\right) = \lim_{N\to\infty} E\left(\frac{1}{T_N}\mathbf{W}_{1N}'\mathbf{W}_{1N}\right)$ exists. The same is true if $\mathbf{w}_{1t}$ is identically distributed across t.

Part (ii) of Theorem 2 is made more precise in the following corollary which considers the classical situation of fixed T_N. Let $\tilde{\mathbf{W}}_1$ denote the matrix containing the k untransformed relevant variables of individual 1 in all time periods and $\ddot{\tilde{\mathbf{W}}}_1$ its column demeaned version.

Corollary 1. *Under the assumptions of Theorem 2, T_N fixed*

(i) *and (FE1)-(FE3) and the forward orthogonal deviations transform*

$$N^{1/2}\left(\hat{\beta}_{1N} - \beta_{10}\right) \xrightarrow{d} N\left(0, \sigma^2\left[E\left(\ddot{\tilde{\mathbf{W}}}_1'\ddot{\tilde{\mathbf{W}}}_1\right)\right]^{-1}\right) \tag{3.3}$$

(ii) *and (FE1)-(FE3), (RE4) and the GLS transform*

$$N^{1/2}\left(\hat{\beta}_{1N} - \beta_{10}\right) \xrightarrow{d} N\left(0, \sigma^2\left[E\left(\tilde{\mathbf{W}}_1'\mathbf{\Omega}^{-1}\tilde{\mathbf{W}}_1\right)\right]^{-1}\right) \tag{3.4}$$

Notice that the asymptotic distribution in (3.3) is the same as for a fixed effects estimator with *known* sparsity pattern of β_0. This underscores the oracle property of the panel Bridge estimator. Similarly, (3.4) is the asymptotic distribution of the random effects estimator with *known* sparsity pattern of β_0.

4. The Marginal Bridge estimator

Since the Bridge estimator is not applicable when $p_N > NT_N$ (though it does allow $p_N \to \infty$) a different approach is needed for this situation. As in Huang et al. (2008) we will employ the Marginal Bridge estimator which estimates β_0 by minimizing

[4] More generally, if $Var\left(w_{1tl}w_{1tm}\right) \in O\left(g(t)\right)$ for all $1 \leq l, m \leq k$ for some positive increasing function g it suffices that $\frac{g(T_N)}{N} \to 0$.

$$U_N(\beta) = \sum_{k=1}^{p_N} \sum_{j=1}^{NT_N} \left(y_j - x_{jk}\beta_k\right)^2 + \lambda_N \sum_{k=1}^{p_N} |\beta_k|^\gamma \tag{4.1}$$

$$= \sum_{k=1}^{p_N} \left(\sum_{j=1}^{NT_N} \left(y_j - x_{jk}\beta_k\right)^2 + \lambda_N |\beta_k|^\gamma \right) \tag{4.2}$$

From (4.2) it is clear that the objective function is nothing else than the sum of the marginal objective functions for each variable – hence the name Marginal Bridge estimator. Let $\tilde{\beta}_N$ denote the minimizer of (4.2). We show that the Marginal Bridge estimator is able to correctly distinguish between relevant and irrelevant variables even when there are more explanatory variables than observations ($p_N > NT_N$). Maintain (FE1), assume $\tilde{\epsilon}_{iN}$ is normally distributed for all $1 \leq i \leq N$ and replace (FE2) and (FE3) by

(FE2MB) $\left(\tilde{\mathbf{X}}_{iN}, \mathbf{c}_{iN}\right) \perp\!\!\!\perp \tilde{\epsilon}_{iN}$, with $E\left(\tilde{\epsilon}_{it}\right) = 0$ and $E\left(\tilde{\epsilon}_{iN}\tilde{\epsilon}_{iN}'\right) = \sigma^2 \mathbf{I}_{T_N}$[5].

(FE2MB) clearly implies (FE3) while the reverse need not be the case (see e.g. Stoyanov (1997) for an example). However, this strengthening is not likely to be of any practical importance since it is hard to imagine *practical* examples where (FE3) is satisfied while (FE2MB) is not. After carrying out either the fixed effects or the random effects transform (FE2MB) implies that $\mathbf{X}_{iN} \perp\!\!\!\perp \epsilon_{iN}$, $\epsilon_{i1}, ..., \epsilon_{iT_N}$ is i.i.d. gaussian for all $1 \leq i \leq N$, $E\left(\epsilon_{it}\right) = 0$, and $E\left(\epsilon_{it}^2\right) = \sigma^2$. The gaussianity assumption on the error terms is a price we must pay for working with L^q-norms instead of the exponential Orlicz-norms in Huang et al. (2008) in the proofs. Working with exponential Orlicz-norms did not turn out to be fruitful due to the presence of random covariates which rendered some otherwise useful maximal inequalities inapplicable.

Let $K_N = (1, ..., k_N)$ denote the active set, i.e. the set of indices of the relevant variables, and $J_N = (k_N + 1, ..., p_N)$ the inactive set, i.e. the set of indices of the irrelevant variables. Standardize the covariates such that $\frac{1}{NT_N}\sum_{j=1}^{NT_N} x_{jk}^2 = 1$ for all $k = 1, ..., p_N$. This implies that the covariates have moments of any order since $\left|x_{jk}\right| \leq \sqrt{\sum_{j=1}^{NT_N} x_{jk}^2} = (NT)^{1/2}$. Since the gaussian error terms also have moments of any order, this is the real reason enabling us to refrain from any moment assumptions like (FE2). Finally, define $\xi_{Nk} = \frac{1}{NT_N}\sum_{j=1}^{NT_N} \mathbf{w}_j'\beta_{10}x_{jk}$. Assume

(B1) There exists a constant $\xi_0 > 0$ such that $\min_{k \in K_N} |\xi_{Nk}| > \xi_0$ with probability approaching 1.

(B2) $\lambda_N/(NT_N) \to 0$.

(B3) $\frac{k_N}{\left(\lambda_N (NT_N)^{-\gamma/2}\right)^{1/(2-\gamma)}} \to 0$.

(B4) $\frac{m_N}{\left(\lambda_N (NT_N)^{-\gamma/2}\right)^{q/(2-\gamma)}} \to 0$ for some $q \geq 1$.

[5] It is sufficient to assume $\tilde{\mathbf{X}}_{iN} \perp\!\!\!\perp \tilde{\epsilon}_{iN}$ for all $1 \leq i \leq N$ but for comparison with (FE3) we refrain from this (see also the comment after (FE3)). Furthermore, as was the case in the $p_N < NT_N$ setting in Sections 2 and 3 $(\tilde{\epsilon}_{i1}, ..., \tilde{\epsilon}_{iT_N})$ may have any covariance structure as long as it is known so that it can be handled by an appropriate transformation. The most common choice is a diagonal matrix.

(B5) For all $\delta > 0$ there exists a $c_0 > 0$ and a $N_0 \in \mathbf{N}$ such that
$$P\left(\frac{\sum_{j=1}^{NT_N} x_{jk}x_{jl}}{(NT_N)^{1/2}} \leq c_0,\ k \in K_N,\ l \in J_N\right) \geq 1 - \delta \text{ for } N \geq N_0.$$

(B6) There exists a constant $0 < b_1 < \infty$ such that $\max_{k \in K_N} |\beta_{10k}| \leq b_1$.

Assumption (B1) is a technical assumption needed to prove that no variables from the active set will be discarded by the Marginal Bridge. In a fixed regressor setting it is similar to assuming that the covariance between the left hand side variable and the relevant covariates is bounded away from 0.

Assumption (B2) requires that $\lambda_N \in o(NT_N)$.

Assumption (B3) combined with assumption (B2) implies that $k_N \in o((NT_N)^{1/2})$. In the classical case of fixed T_N this amounts to $k_N \in o(N^{1/2})$. This is in line with the results of Huang et al. (2008).

For $0 < \gamma < 2$, (B3) also implies that $\lambda_N (NT_N)^{-\gamma/2} \to \infty$. Together with (B2) this yields that $\lambda_N \in o(NT_N) \cap \omega\left((NT_N)^{\gamma/2}\right)$.

Using (B2) in (B4) implies $m_N \in o\left((NT_N)^{q/2}\right)$. Since q is arbitrary this says that the number of irrelevant variables must be asymptotically dominated by some polynomial in in the sample size. Notice that the number of irrelevant variables can not tend to infinity as fast as in Huang et al. (2008) where $m_N \in \exp(o(N))$. As indicated above the reason is that the exponential Orlicz-norms did not carry over straightforwardly to the random covariate setting and so we had to settle with maximal inequalities based on L^q norms which don't give us as sharp results. However, $m_N \in o(N^{q/2})$ (T_N fixed) is not very restrictive in practice since it still allows the number of irrelevant variables to increase at a much higher rate than the sample size as long as this rate is polynomial.

Assumption (B5) is a partial orthogonality assumption limiting the dependence between the variables in the active and the inactive set. It rules out correlations of -1 or 1[6]. However, it is not too restrictive and as will be seen from the Monte Carlo simulations in Section 5 the Marginal Bridge also works quite well even when the covariates in the active and inactive set are highly correlated.

Assumption (B6) is a uniform bound on the size of the coefficients belonging to the relevant variables. This assumption may be relaxed in the same way as assumption (A5) for the Bridge estimator at the price of a lower growth rate of the number of relevant variables.

Assumption (B1)-(B6) are similar to the assumptions made in Huang et al. (2008). However, we must assume gaussianity of the error terms instead of sub-Gaussianity in Huang et al. (2008)[7]. As indicated above, this is a price we pay for letting the covariates be random. The properties of the Marginal Bridge are given in the following Theorem.

Theorem 4. *Under assumption (B1)-(B6) and if* $0 < \gamma < 1$,
$$P\left(\tilde{\beta}_{2N} = 0\right) \to 1 \quad \textit{and} \quad P\left(\tilde{\beta}_{1Nk} = 0,\ k \in K_N\right) \to 0 \tag{4.3}$$

[6] If x_{j1} and x_{j2} are perfectly correlated and (assume for simplicity) have an empirical mean of zero $x_{j2} = bx_{j1}$ *a.s.* for some constant b. Then $\frac{\sum_{j=1}^{NT_N} x_{j1}x_{j2}}{(NT_N)^{1/2}} = b(NT_N)^{1/2}$ which violates (B5).

[7] A random variable X is said to be sub-Gaussian if there exist positive constants C and K such that $P\left(|X| \geq x\right) \leq C\exp(-Kx^2)$

Hence, the Marginal Bridge estimator is able to screen out the irrelevant variables while retaining the relevant ones.

The nonzero coefficients are not estimated consistently. In order to obtain consistent estimates the same two step procedure as in Huang et al. (2008) can be applied. In the first step the Marginal Bridge estimator is applied to distinguish between the relevant and irrelevant variables. In the second step, where only the relevant variables are left, these may be estimated by any consistent estimator (e.g. least squares or the Bridge estimator).

5. Simulations

In this section the finite sample properties of the proposed estimators will be investigated. The Bridge estimator will be implemented by means of the MM-algorithm of Hunter and Li (2005) which in the present case reduces to a series of ridge regressions.

Implementing the Marginal Bridge is very fast. Since $\sum_{j=1}^{NT_N} x_{jk}^2 = NT_N$ for $k = 1, ..., p_N$ it follows from Lemma A in Knight and Fu (2000) that $\beta_k = 0$ iff

$$\frac{\lambda_N}{NT_N} > c_\gamma \left| \frac{\sum_{j=1}^{NT_N} y_j x_{jk}}{NT_N} \right|^{2-\gamma} \tag{5.1}$$

where $c_\gamma = \left(\frac{2}{2-\gamma}\right)\left(\frac{2(1-\gamma)}{2-\gamma}\right)^{1-\gamma}$. Hence, variable selection is extremely fast[8] even in vast dimensional models, since the inclusion of a variable is solely based on the criterion (5.1) which roughly amounts to checking whether the correlation between the left hand side variable and the covariate is sufficiently high to deem the latter relevant. Notice how only marginal information is used to decide whether a variable is to be included or not. Having decided on the sparsity pattern the second step estimates of β_{10} are found by means of least squares[9].

The following issues will be investigated

(1) How often do the Bridge and the Marginal Bridge estimator select the correct sparsity pattern, i.e. how good are they at distinguishing the active from the inactive set? This is highly relevant in applied work investigating which variables help explaining the left hand side variable.
(2) The median number of variables included, i.e how well do the Bridge and the Marginal Bridge reduce the dimension of the problem? This median is ideally equal to the cardinality of the active set.
(3) The explanatory power of the Bridge and the Marginal Bridge. To investigate this, the estimated parameters are used to fit values on a validation data set drawn from the same distribution as the training set.
(4) In connection to the explanatory power it is investigated how often the procedures retain all relevant explanatory variables. As can be expected, retention of all relevant explanatory variables is important for achieving a good fit. It is also highly desirable if the procedures are to be used as initial screening devices in vast dimensional data sets.

[8] A model with 100 observations and 2500 potential explanatory variables takes between 0.2 and 0.3 seconds to estimate on a 2.66 GHz i7 processor.

[9] The Bridge estimator was also tried in the second step but did not outperform least squares while being considerably slower.

(5) The precision of the parameter estimates using the mean square error of $\hat{\beta}$.
(6) The asymptotic distribution of the estimator of the non-zero β_0's. This is done by comparing the standard deviation of $\hat{\beta}_1$ to the corresponding quantities for the least squares estimator with only the active set included. The latter (in practice infeasible) estimator will be called the OLS Oracle henceforth.
(7) In the $p_N < NT_N$ setting the coverage probabilities of 95% confidence intervals are reported for the Bridge estimator and the OLS Oracle to assess theorem 2 [10].

The Bridge and the Marginal Bridge estimators will be compared to the LASSO estimated by pathwise coordinate descent, the Schwarz information criterion (BIC), the OLS Oracle, and OLS on the system including all covariates. Only the Marginal Bridge, the LASSO and the OLS Oracle are applied when $p_N > NT_N$. To limit the computational burden, BIC is only applied for the designs with 15 or fewer covariates which implies a maximum of $2^{15} - 1 = 32.767$ regressions per Monte Carlo replication. All experiments are carried out with 1.000 replications.

The data is generated from equation (2.2). In all experiments $T_N = 10$. Initial experiments indicated that $\gamma = 0.5$ works quite well for the Bridge as well as the Marginal Bridge estimator and this value will be used throughout. Larger values of γ resulted in larger models. In light of the fact that the LASSO turns out to select larger models than the Bridge Estimator, it is no surprise that as γ approaches 1 the median number of variables included by the Bridge Estimator increases too. $\tilde{\epsilon}_{it}$ and c_i are $N(0,1)$ in all experiments.

The regularization parameter λ_N is usually chosen by 10-fold cross validation. Here we try this as well as the significantly faster BIC to determine λ_N for the Bridge, the Marginal Bridge and the LASSO.

5.1. The experiments.

(A) N=10, $\beta_0 = (1,1,1,1,1,0,0,0,0,0,0,0,0,0,0)'$ and the covariates are independent N(0,1) variables.
(B) β_0 is as in (A). The correlation between the k'th and l'th covariate is $\rho^{|k-l|}$ with $\rho = 0.50$.
(C) As (B) but with $\rho = 0.95$.
(D) As (A) but with N=100.
(E) As (B) but with N=100.
(F) As (C) but with N=100.
(G) N=10 and 5 relevant explanatory variables with a coefficient of 1. 245 irrelevant variables. All covariates are independent.
(H) N=10 and 5 relevant explanatory variables with a coefficient of 1. 495 irrelevant variables. All covariates are independent.
(I) N=10 and 5 relevant explanatory variables with a coefficient of 1. 2495 irrelevant variables. All covariates are independent.

Note that even though the covariates are independent in Experiments G-I the maximum spurious sample correlation, i.e. the maximum observed sample correlation between covariates, may still be very high (see Fan and Lv (2008) for examples). In particular, if a relevant and irrelevant covariate are highly correlated it will be difficult to distinguish between these.

[10] The author wishes to thank an anonymous referee for suggesting this.

		Cross Validation			BIC					
			Marg			Marg			OLS	OLS
		Bridge	Bridge	LASSO	Bridge	Bridge	LASSO	BIC	Oracle	All
Experiment A	Sparsity pattern	0.6040	0.5190	0.0130	0.6700	0.7830	0.0990	0.6590	1.0000	0.0000
	Median #Var	5.0000	5.0000	10.0000	5.0000	5.0000	8.0000	5.0000	5.0000	15.0000
	Loss	2.1128	2.1062	2.1544	2.0947	2.0794	2.1643	2.1007	2.0667	2.2125
	Relevant retained	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000
	Median Beta MSE	0.0764	0.0725	0.0950	0.0706	0.0647	0.0971	0.0712	0.0584	0.1131
	Stdv	0.1188	0.1145	0.1206	0.1172	0.1145	0.1236	0.1148	0.1135	0.1198
	Cov. Prob	0.9140			0.9180				0.9390	
Experiment B	Sparsity pattern	0.6560	0.6520	0.0540	0.7170	0.9220	0.2910	0.6870	1.0000	0.0000
	Median #Var	5.0000	5.0000	8.0000	5.0000	5.0000	6.0000	5.0000	5.0000	15.0000
	Loss	2.1048	2.1009	2.1288	2.0898	2.0740	2.1374	2.0959	2.0667	2.2125
	Relevant retained	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000
	Median Beta MSE	0.0863	0.0843	0.0952	0.0827	0.0740	0.0927	0.0841	0.0710	0.1441
	Stdv	0.1387	0.1327	0.1362	0.1351	0.1306	0.1393	0.1314	0.1301	0.1365
	Cov. Prob	0.9300			0.9420				0.9520	
Experiment C	Sparsity pattern	0.0760	0.5600	0.1370	0.0180	0.6400	0.2750	0.0070	1.0000	0.0000
	Median #Var	5.0000	5.0000	7.0000	4.0000	5.0000	6.0000	4.0000	5.0000	15.0000
	Loss	2.1566	2.1106	2.1048	2.1664	2.1011	2.0984	2.1828	2.0667	2.2125
	Relevant retained	0.3230	0.8920	0.8960	0.0290	0.6870	0.8840	0.0110	1.0000	1.0000
	Median Beta MSE	0.4038	0.2854	0.2625	0.4809	0.2840	0.2496	0.4862	0.2189	0.4799
	Stdv	0.5042	0.4080	0.3548	0.6046	0.5138	0.3551	0.6123	0.3497	0.3680
	Cov. Prob	0.7350			0.5710				0.9540	
		Cross Validation			BIC					
			Marg			Marg			OLS	OLS
		Bridge	Bridge	LASSO	Bridge	Bridge	LASSO	BIC	Oracle	All
Experiment D	Sparsity pattern	0.7570	0.6590	0.0190	0.9350	0.9860	0.4570	0.9110	1.0000	0.0000
	Median #Var	5.0000	5.0000	10.0000	5.0000	5.0000	6.0000	5.0000	5.0000	15.0000
	Loss	2.0089	2.0092	2.0135	2.0073	2.0066	2.0195	2.0074	2.0065	2.0176
	Relevant retained	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000
	Median Beta MSE	0.0210	0.0209	0.0279	0.0191	0.0179	0.0335	0.0186	0.0178	0.0326
	Stdv	0.0334	0.0331	0.0340	0.0333	0.0330	0.0354	0.0330	0.0330	0.0331
	Cov. Prob	0.9470			0.9470				0.9480	
Experiment E	Sparsity pattern	0.7660	0.7100	0.1180	0.9340	0.9930	0.6230	0.9110	1.0000	0.0000
	Median #Var	5.0000	5.0000	8.0000	5.0000	5.0000	5.0000	5.0000	5.0000	15.0000
	Loss	2.0085	2.0084	2.0114	2.0070	2.0066	2.0136	2.0073	2.0065	2.0176
	Relevant retained	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000
	Median Beta MSE	0.0239	0.0247	0.0278	0.0222	0.0215	0.0276	0.0225	0.0214	0.0410
	Stdv	0.0391	0.0387	0.0391	0.0390	0.0386	0.0395	0.0387	0.0386	0.0389
	Cov. Prob	0.9440			0.9420				0.9510	
Experiment F	Sparsity pattern	0.6960	0.7020	0.1570	0.9310	0.9930	0.4690	0.9240	1.0000	0.0000
	Median #Var	5.0000	5.0000	7.0000	5.0000	5.0000	6.0000	5.0000	5.0000	15.0000
	Loss	2.0092	2.0088	2.0094	2.0075	2.0066	2.0096	2.0072	2.0065	2.0176
	Relevant retained	1.0000	1.0000	1.0000	0.9990	1.0000	1.0000	0.9990	1.0000	1.0000
	Median Beta MSE	0.0802	0.0795	0.0775	0.0721	0.0662	0.0725	0.0684	0.0659	0.1362
	Stdv	0.1134	0.1067	0.1069	0.1120	0.1066	0.1079	0.1066	0.1066	0.1078
	Cov. Prob	0.9340			0.9380				0.9510	

TABLE 1. Top panel: Experiments A-C (N=10). Bottom panel: Experiments D-F (N=100). Cross Validation and BIC indicate which procedure was used to determine λ_N. Sparsity pattern: The fraction of times the correct sparsity pattern is detected. Median #Var: The median number of variables included. Loss: The MSE when using the estimated parameters on a validation data set drawn from the same distribution as the training set. Relevant retained: The fraction of relevant variables retained in the model. Median Beta MSE: Calculated as explained in the main text. Stdv: Standard deviation of the estimated coefficient of the first variable (which is always in the active set). Cov. Prob: Coverage probability of 95% confidence interval of the estimated coefficient of the first variable.

5.2. **Results.** Table 1 holds the results for experiments A-F, where $p_N < NT_N$.

Experiment A reveals that the Bridge, Marginal Bridge and Schwarz information criterion all perform quite well in the independent covariates setting. They all detect the correct sparsity pattern in more than half of the cases irrespective of whether cross validation or BIC is used to determine λ_N. In all respects their performance is comparable to the OLS Oracle.

As seen from Experiment B making the covariates moderately correlated does not deteriorate the performance of the procedures with respect to the fraction of times the right sparsity pattern is chosen or the fraction of times all relevant covariates are retained. However, all procedures get more imprecise. Since this is also the case for the OLS Oracle this is not a particular artefact of the Bridge class of estimators.

Experiment C reveals that as the correlation gets very high the performance of the Bridge and BIC deteriorate. On the other hand the Marginal Bridge continues to detect the right sparsity pattern in more than half of the cases. However, even the latter fails to retain all relevant variables in all cases. The coverage probabilities of the Bridge confidence intervals are significantly below 95% – this is the case in particular when BIC is used to select λ_N. Taking into account that it only retains all relevant variables in 3% of the simulations, this result is not too surprising.

Experiments D-F illuminate the asymptotic properties of the Bridge and the Marginal Bridge. In particular the Marginal Bridge with BIC used to determine λ_N detects the correct sparsity pattern in almost all cases irrespective of the correlation structure imposed on the covariates. The performance of the Bridge also gets significantly better as the sample size is increased while the LASSO only improves moderately. The Loss of all procedures is reduced and the parameters are estimated more precisely. Finally, the coverage probabilities of the Bridge are now close to the nominal rate of 95%.

Notice that the Marginal Bridge performs quite well even in the high correlation experiments C and F indicating that the partial orthogonality assumption (B5) is not overly restrictive.

It is seen that in general the BIC is a better way of determining λ_N than cross validation. BIC detects the correct sparsity pattern more often and only in Experiment C one finds that cross validation is superior with respect to the number of relevant variables retained as well as coverage probabilities.

Table 2 contains the results for the Experiments G-I which investigate the performance of the Marginal Bridge in the $p_N > NT_N$ case. As can be expected the correct sparsity pattern is detected less frequently. However, all relevant variables are retained very often while only few irrelevant variables are kept in the model. Hence, the Marginal Bridge is still a very effective tool for dimension reduction.

The LASSO and the Marginal Bridge perform equally well in Experiments G and H (slight advantage for the LASSO) while the LASSO is superior in Experiment I. However, the LASSO also takes a lot longer to compute and the models it chooses are bigger. The following idea which builds on the thoughts of Fan and Lv (2008) could potentially improve the performance of the Marginal Bridge: estimate the Marginal Bridge one or several times more using the residuals from the first (previous) step as left hand side variables. This will lower the priority of those irrelevant variables which seemed relevant only through their high correlation with some of the relevant variables already included.

		Cross Validation			BIC					
			Marg			Marg			OLS	OLS
		Bridge	Bridge	LASSO	Bridge	Bridge	LASSO	BIC	Oracle	All
Experiment G	Sparsity pattern		0.1380	0.0020		0.3380	0.0420		1.0000	
	Median #Var		9.0000	20.0000		6.0000	9.0000		5.0000	
	Loss		2.3302	2.4744		2.2263	2.6162		2.0784	
	Relevant retained		0.9820	1.0000		0.9390	0.9990		1.0000	
	Median Beta MSE		0.0299	0.0413		0.0204	0.0470		0.0141	
	Stdv		0.1465	0.1358		0.1612	0.1448		0.1081	
Experiment H	Sparsity pattern		0.0710	0.0010		0.2300	0.0290		1.0000	
	Median #Var		12.0000	23.5000		7.0000	10.0000		5.0000	
	Loss		2.5162	2.5763		2.3593	2.7642		2.0891	
	Relevant retained		0.9350	1.0000		0.8830	0.9980		1.0000	
	Median Beta MSE		0.0257	0.0320		0.0177	0.0368		0.0101	
	Stdv		0.1714	0.1414		0.2045	0.1553		0.1100	
Experiment I	Sparsity pattern		0.0100	0.0000		0.0540	0.0130		1.0000	
	Median #Var		19.0000	35.0000		9.0000	11.0000		5.0000	
	Loss		3.1359	2.8733		2.8479	3.2288		2.0331	
	Relevant retained		0.7450	0.9950		0.6730	0.9540		1.0000	
	Median Beta MSE		0.0178	0.0183		0.0147	0.0213		0.0045	
	Stdv		0.2428	0.1470		0.2657	0.1690		0.1056	

TABLE 2. Cross Validation and BIC indicate which procedure was used to determine λ_N. Sparsity pattern: The fraction of times the correct sparsity pattern is detected. Median #Var: The median number of variables included. Loss: The MSE when using the estimated parameters on a validation data set drawn from the same distribution as the training set. Relevant retained: The fraction of relevant variables retained in the model. Median Beta MSE: Calculated as explained in the main text. Stdv: Standard deviation of the estimated coefficient of the first variable (which is always in the active set).

6. Conclusions

This paper introduces the Bridge and Marginal Bridge estimator in a linear panel data setting allowing for random as well as fixed effects. When $p < NT_N$ it is shown that the Bridge estimator (and Marginal Bridge) has the oracle property. It sets all coefficients that are truly zero to zero and the asymptotic distribution of the estimator of the non zero coefficients is the same as if the sparsity pattern had been known. Monte Carlo experiments underscore this conclusion and are used to investigate the finite sample properties of the procedures. They also reveal that the Schwarz information criterion is more useful than 10-fold cross validation for selecting λ_N. This is encouraging since BIC is also faster than cross validation.

When $p > NT_N$ it is shown that the Marginal Bridge estimator still detects the correct sparsity pattern with probability converging to one. This is true in the random as well as the fixed effects setting under a partial orthogonality assumption on the covariates. However, the Marginal Bridge works well even when the relevant and irrelevant covariates are highly correlated. Furthermore, the Marginal Bridge estimates are extremely fast to compute since only marginal information is used to decide whether a variable is relevant or not. The Marginal Bridge is also shown to perform well in the $p_N < NT_N$ setting. In the $p_N > NT_N$ setting the Marginal Bridge does not always retain all relevant variables. An iterative procedure was proposed to solve this problem. Working out the properties of this procedure is left for future research.

7. Appendix

Lemma 1. *Let* $\mathbf{u}$ *be a* $p_N \times 1$ *vector. Then*

$$E\left(\sup_{\|\mathbf{u}\|\leq\delta}\left|\sum_{j=1}^{NT_N}\epsilon_j\mathbf{x}_j'\mathbf{u}\right|\middle|\mathbf{X}_N\right)\leq\delta\sigma\left(NT_Np_N\right)^{\frac{1}{2}}\left(\frac{1}{NT_Np_N}\sum_{i=1}^{N}\sum_{t=1}^{T_N}\sum_{k=1}^{p_N}x_{itk}^2\right)^{\frac{1}{2}}$$

Proof.

$$E\left(\sup_{\|\mathbf{u}\|\leq\delta}\left|\sum_{j=1}^{NT_N}\epsilon_j{\mathbf{x}'}_j\mathbf{u}\right|^2\middle|\mathbf{X}_N\right)\leq E\left[\sup_{\|\mathbf{u}\|\leq\delta}\|\mathbf{u}\|^2\left\|\left(\sum_{j=1}^{NT_N}\epsilon_j\mathbf{x}_j\right)\right\|^2\middle|\mathbf{X}_N\right]$$

$$\leq\delta^2E\left[\left(\sum_{j=1}^{NT_N}\epsilon_j\mathbf{x}_j'\right)\left(\sum_{j=1}^{NT_N}\epsilon_j\mathbf{x}_j\right)\middle|\mathbf{X}_N\right]$$

where the first inequality follows from the Cauchy-Schwarz inequality. Since

$$E\left(\epsilon_{it}\mathbf{x}_{it}'\epsilon_{is}\mathbf{x}_{is}|\mathbf{X}_N\right)=E\left(\mathbf{x}_{it}'\mathbf{x}_{is}\epsilon_{it}\epsilon_{is}|\mathbf{X}_{iN}\right)=\mathbf{x}_{it}'\mathbf{x}_{is}E\left(\epsilon_{it}\epsilon_{is}|\mathbf{X}_{iN}\right)=0,\quad t\neq s$$

$$E\left(\epsilon_{it}\mathbf{x}_{it}'\epsilon_{js}\mathbf{x}_{js}|\mathbf{X}_N\right)=E\Big(E\left[\epsilon_{it}\mathbf{x}_{it}'\epsilon_{js}\mathbf{x}_{js}|\mathbf{X}_N,\epsilon_{it}\right]|\mathbf{X}_N\Big)$$

$$=E\Big(\epsilon_{it}\mathbf{x}_{it}'\mathbf{x}_{js}E\left[\epsilon_{js}|\mathbf{X}_N,\epsilon_{it}\right]|\mathbf{X}_N\Big)=E\Big(\epsilon_{it}\mathbf{x}_{it}'\mathbf{x}_{js}E\left[\epsilon_{js}|\mathbf{X}_{jN}\right]|\mathbf{X}_N\Big)=0,\quad i\neq j$$

$$E\left(\epsilon_{it}\mathbf{x}_{it}'\epsilon_{it}\mathbf{x}_{it}|\mathbf{X}_N\right)=E\left[\mathbf{x}_{it}'\mathbf{x}_{it}\epsilon_{it}\epsilon_{it}|\mathbf{X}_{iN}\right]=\mathbf{x}_{it}'\mathbf{x}_{it}E\left[\epsilon_{it}\epsilon_{it}|\mathbf{X}_{iN}\right]=\sigma^2\mathbf{x}_{it}'\mathbf{x}_{it}$$

Hence,

$$E\left(\sup_{\|\mathbf{u}\|\leq\delta}\left|\sum_{j=1}^{NT_N}\epsilon_j\mathbf{x}_j\mathbf{u}\right|^2\middle|\mathbf{X}_N\right)\leq\delta^2\sigma^2\sum_{j=1}^{NT_N}\mathbf{x}_j'\mathbf{x}_j=\delta^2\sigma^2\sum_{i=1}^{N}\sum_{t=1}^{T_N}\mathbf{x}_{it}'\mathbf{x}_{it}$$

$$=\delta^2\sigma^2NT_Np_N\frac{1}{NT_Np_N}\sum_{i=1}^{N}\sum_{t=1}^{T_N}\sum_{k=1}^{p_N}x_{itk}^2$$

and the result follows from the conditional Jensen inequality. □

Lemma 2. *Let* $\{X_n\}_{n\in\mathbb{N}}$ *and* $\{Y_n\}_{n\in\mathbb{N}}$ *be sequences of nonnegative random variables. If there exists an integer* N_0 *and a constant* C *such that for* $n\geq N_0$

$$E\left(\frac{X_n}{Y_n}\right)\leq C$$

then

$$X_n\in O_p(Y_n)$$

Proof. It suffices to show that for any $\epsilon>0$ $P\left(\left\{\frac{X_n}{Y_n}>\frac{C}{\epsilon}\right\}\right)\leq\epsilon$ for $n\geq N_0$. Assume the opposite is true for some $\epsilon>0$ to reach a contradiction. Then,

$$E\left(\frac{X_n}{Y_n}\right)=\int\frac{X_n}{Y_n}dP\geq\int_{\left\{\frac{X_n}{Y_n}>\frac{C}{\epsilon}\right\}}\frac{X_n}{Y_n}dP\geq\int_{\left\{\frac{X_n}{Y_n}>\frac{C}{\epsilon}\right\}}\frac{C}{\epsilon}dP\geq\frac{C}{\epsilon}P\left(\left\{\frac{X_n}{Y_n}>\frac{C}{\epsilon}\right\}\right)>C$$

which is the desired contradiction. □

Proof of Theorem 1. We first show that $\left\|\hat{\beta}_N - \beta_0\right\| \in O_p\left(\left[\frac{p_N+\lambda_N k_N}{NT_N\rho_{1N}}\right]^{\frac{1}{2}}\right)$. Since $\hat{\beta}_N$ minimizes (3.2)

$$\sum_{j=1}^{NT_N}\left(y_j - \mathbf{x}_j'\hat{\beta}_N\right)^2 + \lambda_N\sum_{k=1}^{p_N}|\hat{\beta}_{Nk}|^\gamma \leq \sum_{j=1}^{NT_N}\left(y_j - \mathbf{x}_j'\beta_0\right)^2 + \lambda_N\sum_{k=1}^{p_N}|\beta_{0k}|^\gamma$$

Defining $\eta_N = \lambda_N\sum_{k=1}^{p_N}|\beta_{0k}|^\gamma$ this implies:

$$\begin{aligned}\eta_N &\geq \sum_{j=1}^{NT_N}\left(y_j - \mathbf{x}_j'\hat{\beta}_N\right)^2 - \sum_{j=1}^{NT_N}\left(y_j - \mathbf{x}_j'\beta_0\right)^2 \\ &= \sum_{j=1}^{NT_N}\left(\left[y_j - \mathbf{x}_j'\hat{\beta}_N\right] - \left[y_j - \mathbf{x}_j'\beta_0\right]\right)\left(\left[y_j - \mathbf{x}_j'\hat{\beta}_N\right] + \left[y_j - \mathbf{x}_j'\beta_0\right]\right) \\ &= \sum_{j=1}^{NT_N}\left[\mathbf{x}_j'(\beta_0 - \hat{\beta}_N)\right]^2 + 2\sum_{j=1}^{NT_N}\epsilon_j\mathbf{x}_j'\left(\beta_0 - \hat{\beta}_N\right)\end{aligned}$$

Now define $\delta_N = (NT_N)^{1/2}\Sigma_N^{1/2}(\hat{\beta}_N - \beta_0)$, $\mathbf{D}_N = (NT_N)^{-1/2}\Sigma_N^{-1/2}\mathbf{X}_N'$ and $\epsilon_N = (\epsilon_1, ..., \epsilon_{NT_N})'$. With these definitions

$$\begin{aligned}\sum_{j=1}^{NT_N}\left[\mathbf{x}_j'(\beta_0 - \hat{\beta}_N)\right]^2 + 2\sum_{j=1}^{NT_N}\epsilon_j\mathbf{x}_j'\left(\beta_0 - \hat{\beta}_N\right) &= \delta_N'\delta_N - 2(\mathbf{D}_N\epsilon_N)'\delta_N \\ &= ||\delta_N - \mathbf{D}_N\epsilon_N||^2 - ||\mathbf{D}_N\epsilon_N||^2\end{aligned}$$

which implies

$$||\delta_N - \mathbf{D}_N\epsilon_N||^2 - ||\mathbf{D}_N\epsilon_N||^2 - \eta_N \leq 0$$

and by the sub additivity of $x \mapsto x^{1/2}$ yields $||\delta_N - \mathbf{D}_N\epsilon_N|| \leq ||\mathbf{D}_N\epsilon_N|| + \eta_N^{1/2}$. By sub additivity of the norm $\|\cdot\|$ this implies $||\delta_N|| \leq ||\delta_N - \mathbf{D}_N\epsilon_N|| + ||\mathbf{D}_N\epsilon_N|| \leq 2||\mathbf{D}_N\epsilon_N|| + \eta_N^{1/2}$. Since $(x+y)^2 \leq 2x^2 + 2y^2$ for $x, y \in \mathbf{R}$ by the convexity of $x \mapsto x^2$ one has

$$||\delta_N||^2 \leq 4||\mathbf{D}_N\epsilon_N||^2 + 2\eta_N$$

Letting $\mathbf{d}_j$ denote the j'th column of $\mathbf{D}_N$ we may write $\mathbf{D}_N\epsilon_N = \sum_{j=1}^{NT_N}\mathbf{d}_j\epsilon_j$. Using that $\mathbf{D}_N$ is measurable with respect to $\mathbf{X}_N$, conclude

$$\begin{aligned}&E\left(\mathbf{d}_{it}'\epsilon_{it}\mathbf{d}_{is}\epsilon_{is}\right) = E\left(\mathbf{d}_{it}'\mathbf{d}_{is}E\left[\epsilon_{it}\epsilon_{is}|\mathbf{X}_N\right]\right) = E\left(\mathbf{d}_{it}'\mathbf{d}_{is}E\left[\epsilon_{it}\epsilon_{is}|\mathbf{X}_{iN}\right]\right) = 0,\ s \neq t \\ &E\left(\mathbf{d}_{it}'\epsilon_{it}\mathbf{d}_{js}\epsilon_{js}\right) = E\left(\mathbf{d}_{it}'\epsilon_{it}\mathbf{d}_{js}E\left[\epsilon_{js}|\mathbf{X}_N, \epsilon_{it}\right]\right) = E\left(\mathbf{d}_{it}'\epsilon_{it}\mathbf{d}_{js}\left[\epsilon_{js}|\mathbf{X}_{jN}\right]\right) = 0,\ i \neq j \\ &E\left(\mathbf{d}_{it}'\epsilon_{it}\mathbf{d}_{it}\epsilon_{it}\right) = E\left(\mathbf{d}_{it}'\mathbf{d}_{it}E\left[\epsilon_{it}\epsilon_{it}|\mathbf{X}_{iN}\right]\right) = E\left(\mathbf{d}_{it}'\mathbf{d}_{it}\right) = \sigma^2 E\left(||\mathbf{d}_{it}||^2\right)\end{aligned}$$

Hence,

$$\begin{aligned}E\left(\|\mathbf{D}_N\epsilon_N\|^2\right) &= E\left(\left\|\sum_{j=1}^{NT_N}\mathbf{d}_j\epsilon_j\right\|^2\right) = E\left[\left(\sum_{j=1}^{NT_N}\mathbf{d}_j\epsilon_j\right)'\left(\sum_{j=1}^{NT_N}\mathbf{d}_j\epsilon_j\right)\right]\\ &= \sigma^2 E\left(\sum_{j=1}^{NT_N}\|\mathbf{d}_j\|^2\right) = \sigma^2 E\left(\mathrm{tr}\left(\mathbf{D}_N'\mathbf{D}_N\right)\right) = \sigma^2 E\left(\mathrm{tr}\left(\mathbf{D}_N\mathbf{D}_N'\right)\right)\\ &= \sigma^2\mathrm{tr}\left(\mathbf{I}_{p_N}\right) = \sigma^2 p_N\end{aligned}$$

And so, $E\left(||\delta_N||^2\right) \leq 4\sigma^2 p_N + 2\eta_N$. Hence,

$$(NT_N)E\left(\left(\hat{\beta}_N-\beta_0\right)'\Sigma_N\left(\hat{\beta}_N-\beta_0\right)\right) = E(\delta_N'\delta_N) = E(||\delta_N||^2) \leq 4\sigma^2 p_N + 2\eta_N$$

Since the number of non zero coefficients is k_N

$$\eta_N = \lambda_N\sum_{k=1}^{p_N}|\beta_{0j}|^\gamma = \lambda_N\sum_{k=1}^{k_N}|\beta_{0j}|^\gamma \leq \lambda_N k_N b_1^\gamma$$

where the inequality is a consequence of assumption (A5). Since ρ_{1N} is the smallest eigenvalue of $\mathbf{\Sigma}_N$

$$\rho_{1N}||\hat{\beta}_N-\beta_0||^2 = \rho_{1N}\left(\hat{\beta}_N-\beta_0\right)'\left(\hat{\beta}_N-\beta_0\right) \leq \left(\hat{\beta}_N-\beta_0\right)'\Sigma_N\left(\hat{\beta}_N-\beta_0\right)$$

Hence,

$$\begin{aligned}E\left(\rho_{1N}||\hat{\beta}_N-\beta_0||^2\right) &\leq \frac{NT_N}{NT_N}E\left(\left(\hat{\beta}_N-\beta_0\right)'\Sigma_N\left(\hat{\beta}_N-\beta_0\right)\right)\\ &\leq \frac{4\sigma^2 p_N+2\eta_N}{NT_N} \leq \frac{4\sigma^2 p_N + 2\lambda_N k_N b_1^\gamma}{NT_N}\\ &\leq C\frac{p_N+\lambda_N k_N b_1^\gamma}{NT_N}\end{aligned}$$

for $C = \max\left(4\sigma^2, 2b_1^\gamma\right)$. This implies

$$E\left(\frac{||\hat{\beta}_N-\beta_0||^2}{\left(\frac{p_N+\lambda_N k_N}{NT_N\rho_{1N}}\right)}\right) \leq C$$

By Lemma 2 this establishes $||\hat{\beta}_N-\beta_0|| \in O_p\left(\left[\frac{p_N+\lambda_N k_N}{NT_N\rho_{1N}}\right]^{\frac{1}{2}}\right)$. Next we show that $||\hat{\beta}_N-\beta_0|| \in O_p\left(\rho_{1N}^{-1}\left(p_N/(NT_N)\right)^{1/2}\right)$. Like Huang et al. (2008) we use the idea from the proof of Theorem 3.2.5 in Van der Vaart and Wellner (1996). Let $r_N = \rho_{1N}^{-1}\left(p_N/(NT_N)\right)^{1/2}$. For every N partition the parameter space (excluding β_0) into the disjoint shells $S_{l,N} = \left\{\beta : 2^{l-1} < \|\beta-\beta_0\|/r_N \leq 2^l\right\}$ where $l \in \mathbb{Z}$. If $2^M < \left\|\hat{\beta}_N-\beta_0\right\|/r_N$ for a given integer M then $\hat{\beta}_N \in \bigcup_{l>M} S_{l,N}$. For the shell which $\hat{\beta}_N$ belongs to, the infimum of the map $\beta \mapsto L_N(\beta) - L_N(\beta_0)$ is non positive. Hence, for any $\delta > 0$ [11]

$$
\begin{aligned}
&P\left(\left\|\hat{\beta}_N - \beta_0\right\| / r_N > 2^M\right) \\
&\leq \sum_{\substack{l>M \\ 2^{l-1}<\delta/r_N}} P\left(\inf_{\beta\in S_{l,N}} \left(L_N(\beta) - L_N(\beta_0)\right) \leq 0\right) + P\left(\left\|\hat{\beta}_N - \beta_0\right\| > \delta\right) \qquad (7.1)
\end{aligned}
$$

The last term in (7.1) converges to 0 by the consistency of $\hat{\beta}_N$ shown in the first part of the theorem. The theorem is established by proving that the first term on the right hand side can be made arbitrarily small by choosing M sufficiently large. To this is end let $\beta \in S_{l,N}$ for an arbitrary l summed over, and notice that

$$
\begin{aligned}
L_N(\beta) - L_N(\beta_0) &= \sum_{j=1}^{NT_N} \left(y_j - \mathbf{x}_j'\beta\right)^2 + \lambda_N \sum_{k=1}^{k_N} |\beta_{1k}|^\gamma + \lambda_N \sum_{k=1}^{m_N} |\beta_{2k}|^\gamma \\
&- \sum_{j=1}^{NT_N} \left(y_j - \mathbf{x}_j'\beta_0\right)^2 - \lambda_N \sum_{k=1}^{k_N} |\beta_{01k}|^\gamma \\
&\geq \sum_{j=1}^{NT_N} \left(y_j - \mathbf{x}_j'\beta\right)^2 + \lambda_N \sum_{k=1}^{k_N} |\beta_{1k}|^\gamma - \sum_{j=1}^{NT_N} \left(y_j - \mathbf{x}_j'\beta_0\right)^2 - \lambda_N \sum_{k=1}^{k_N} |\beta_{01k}|^\gamma \\
&= \sum_{j=1}^{NT_N} \left(\mathbf{x}_j' \left[\beta - \beta_0\right]\right)^2 - 2 \sum_{j=1}^{NT_N} \epsilon_j \mathbf{x}_j' \left(\beta - \beta_0\right) + \lambda_N \sum_{k=1}^{k_N} \left(|\beta_{1k}|^\gamma - |\beta_{01k}|^\gamma\right) \qquad (7.2)
\end{aligned}
$$

Regarding the first term in (7.2),

$$
\begin{aligned}
&\sum_{j=1}^{NT_N} \left(\mathbf{x}_j' \left[\beta - \beta_0\right]\right)^2 = \left[\beta - \beta_0\right]' \sum_{j=1}^{NT_N} \mathbf{x}_j \mathbf{x}_j' \left[\beta - \beta_0\right] = NT_N \left[\beta - \beta_0\right]' \mathbf{\Sigma}_N \left[\beta - \beta_0\right] \\
&\geq NT_N \|\beta - \beta_0\|^2 \rho_{1N} > NT_N 2^{2(l-1)} r_N^2 \rho_{1N}
\end{aligned}
$$

Regarding the third term in (7.2) we notice that $\beta \in S_{l,N}$ and $2^{l-1} < \delta/r_N$ implies that $\|\beta - \beta_0\| / r_N \leq 2^l < 2\delta/r_N$. Hence, it suffices to consider βs satisfying

[11] Note that

$$
\begin{aligned}
&\left\{\left\|\hat{\beta}_N - \beta_0\right\| / r_N > 2^M\right\} \subseteq \bigcup_{l>M} \left\{\inf_{\beta\in S_{l,N}} \left(L_N(\beta) - L_N(\beta_0)\right) \leq 0\right\} \\
&\subseteq \left(\bigcup_{l>M} \left\{\inf_{\beta\in S_{l,N}} \left(L_N(\beta) - L_N(\beta_0)\right) \leq 0\right\} \cap \left\{\left\|\hat{\beta}_N - \beta_0\right\| \leq \delta\right\}\right) \cup \left\{\left\|\hat{\beta}_N - \beta_0\right\| > \delta\right\} \\
&= \left(\bigcup_{l>M} \left\{\inf_{\beta\in S_{l,N}} \left(L_N(\beta) - L_N(\beta_0)\right) \leq 0\right\} \cap \left\{\left\|\hat{\beta}_N - \beta_0\right\| / r_N \leq \delta / r_N\right\}\right) \cup \left\{\left\|\hat{\beta}_N - \beta_0\right\| > \delta\right\} \\
&\subseteq \bigcup_{\substack{l>M \\ 2^{l-1}<\delta/r_N}} \left\{\inf_{\beta\in S_{l,N}} \left(L_N(\beta) - L_N(\beta_0)\right) \leq 0\right\} \cup \left\{\left\|\hat{\beta}_N - \beta_0\right\| > \delta\right\}
\end{aligned}
$$

and conclude using the subadditivity of P.

$\|\beta - \beta_0\| < 2\delta$. Since $\delta > 0$ is arbitrary and the entries of β_{01} are bounded uniformly away from the 0 by b_0 the mean value theorem may be applied to conclude that for some ζ_k between β_{1k} and β_{01k}

$$\left|\lambda_N \sum_{k=1}^{k_N} \left(|\beta_{1k}|^\gamma - |\beta_{01k}|^\gamma\right)\right| = \left|\lambda_N \gamma \sum_{k=1}^{k_N} |\zeta_k|^{\gamma-1} \text{sign}(\zeta_k)\left(\beta_{1k} - \beta_{01k}\right)\right|$$
$$\leq c\lambda_N \sum_{k=1}^{k_N} |\beta_{1k} - \beta_{01k}| \leq c\lambda_N k_N^{1/2} \|\beta - \beta_0\| \leq c\lambda_N k_N^{1/2} 2^l r_N$$

where $c = \gamma(b_0 - 2\delta)^{\gamma-1}$ and the second to last estimate follows from Jensen's inequality. Hence, on $S_{l,N}$, $\lambda_N \sum_{k=1}^{k_N} \left(|\beta_{1k}|^\gamma - |\beta_{01k}|^\gamma\right) \geq -c\lambda_N k_N^{1/2} 2^l r_N$. Therefore, on $S_{l,N}$,

$$L_N(\beta) - L_N(\beta_0) \geq -\left|2\sum_{j=1}^{NT_N} \epsilon_j \mathbf{x}_j' \left(\beta - \beta_0\right)\right| + \rho_{1N} NT_N 2^{2(l-1)} r_N^2 - c\lambda_N k_N^{1/2} 2^l r_N$$

Hence, by the conditional Markov inequality and Lemma 1

$$P\left(\inf_{\beta \in S_{l,N}} \left(L_N(\beta) - L_N(\beta_0)\right) \leq 0 \middle| \mathbf{X}_N\right)$$
$$\leq P\left(\sup_{\beta \in S_{l,N}} \left|2\sum_{j=1}^{NT_N} \epsilon_j \mathbf{x}_j' \left(\beta - \beta_0\right)\right| \geq \rho_{1N} NT_N 2^{2(l-1)} r_N^2 - c\lambda_N k_N^{1/2} 2^l r_N \middle| \mathbf{X}_N\right)$$
$$\leq \frac{E\left(\sup_{\beta \in S_{l,N}} \left|2\sum_{j=1}^{NT_N} \epsilon_j \mathbf{x}_j' \left(\beta - \beta_0\right)\right| \middle| \mathbf{X}_N\right)}{\rho_{1N} NT_N 2^{2(l-1)} r_N^2 - c\lambda_N k_N^{1/2} 2^l r_N}$$
$$\leq \frac{\sigma (NT_N p_N)^{1/2} 2^l r_N \left(\frac{1}{NT_N p_N} \sum_{i=1}^N \sum_{t=1}^{T_N} \sum_{k=1}^{p_N} x_{itk}^2\right)^{1/2}}{\rho_{1N} NT_N 2^{2(l-1)} r_N^2 - c\lambda_N k_N^{1/2} 2^l r_N}$$
$$= \frac{2\sigma \left(\frac{1}{NT_N p_N} \sum_{i=1}^N \sum_{t=1}^{T_N} \sum_{k=1}^{p_N} x_{itk}^2\right)^{1/2}}{2^{l-2} - c\lambda_N \left(k_N/(p_N NT_N)\right)^{1/2}}$$

By assumption (A3) $\lambda_N \left(k_N/(p_N NT_N)\right)^{1/2} \to 0$ and so $2^{l-2} - c\lambda_N \left(k_N/(NT_N)\right)^{1/2} \geq 2^{l-3}$ for N sufficiently large. Hence, by iterated expectations and assumption (A1)

$$P\left(\inf_{\beta \in S_{l,N}} \left(L_N(\beta) - L_N(\beta_0)\right) \leq 0\right) \leq \frac{\sigma\sqrt{K}}{2^{l-4}}$$

Finally, this implies that

$$\sum_{\substack{l>M\\2^{l-1}<\delta/r_N}} P\left(\inf_{\beta\in S_{l,N}}\left(L_N(\beta)-L_N(\beta_0)\right)\leq 0\right)\leq\sum_{l>M}\frac{\sigma\sqrt{K}}{2^{l-4}}$$

which is convergent and so the tail can be made arbitrarily small by choosing M sufficiently large. □

Lemma 3. *Suppose* $0<\gamma<1$. *Let* $\hat{\beta}_N=\left(\hat{\beta}_{1N},\ \hat{\beta}_{2N}\right)$. *Then* $\hat{\beta}_{2N}=0$ *with probability converging to* 1 *under assumptions (A1)-(A7).*

Proof. By Theorem 1 $||\hat{\beta}_N-\beta_0||\in O_p\left(h_N\right)$ with $h_N=\rho_{1N}^{-1}(p_N/(NT_N))^{1/2}$ so for all $\epsilon>0$ there exists a constant C such that for N sufficiently large

$$P\left(||\hat{\beta}_N-\beta_0||/h_N>C\right)<\epsilon\Leftrightarrow P\left(||\hat{\beta}_N-\beta_0||\leq Ch_N\right)\geq 1-\epsilon$$

Put differently, $\hat{\beta}_N\in\left\{\beta:||\beta-\beta_0||\leq Ch_N\right\}$ with probability converging to 1. Let $\hat{\beta}_{1N}=\beta_{10}+h_N\mathbf{u}_1$ and $\hat{\beta}_{2N}=\beta_{20}+h_N\mathbf{u}_2=h_N\mathbf{u}_2$. Choosing $\hat{\beta}_N$ is then equivalent to choosing $\mathbf{u}_1$ and $\mathbf{u}_2$. For $\mathbf{u}=(\mathbf{u}_1',\mathbf{u}_2')'$ one has $||\mathbf{u}||=||\hat{\beta}_N-\beta_0||/h_N$ which is bounded by C with probability approaching 1. Hence, we may assume $||\mathbf{u}||^2=||\mathbf{u}_1||^2+||\mathbf{u}_2||^2\leq C^2$ and define

$$V_N(\mathbf{u}_1,\mathbf{u}_2)=L_N(\hat{\beta}_{1N},\hat{\beta}_{2N})=L_N(\beta_{10}+h_N\mathbf{u}_1,h_N\mathbf{u}_2)$$

To establish the lemma it now suffices to show that for any $\mathbf{u}$ with $\|\mathbf{u}\|\leq C$, $V_N(\mathbf{u}_1,\mathbf{u}_2)-V_N(\mathbf{u}_1,\mathbf{0})>0$ with probability converging to 1 if $\mathbf{u}_2\neq 0$. Now,

$$\begin{aligned}
V_N(\mathbf{u}_1,\mathbf{u}_2)-V_N(\mathbf{u}_1,\mathbf{0}) &= \sum_{j=1}^{NT_N}\left(y_j-\beta_{01}'\mathbf{w}_j-h_N\mathbf{u}_1'\mathbf{w}_j-h_N\mathbf{u}_2'\mathbf{z}_j\right)^2+\lambda_N\sum_{k=1}^{k_N}|\beta_{01k}+h_Nu_{1k}|^\gamma\\
&+\lambda_N\sum_{k=1}^{m_N}|h_Nu_{2k}|^\gamma-\sum_{j=1}^{NT_N}\left(y_j-\beta_{01}'\mathbf{w}_j-h_N\mathbf{u}_1'\mathbf{w}_j\right)^2-\lambda_N\sum_{k=1}^{k_N}|\beta_{01k}+h_Nu_{1k}|^\gamma\\
&=\sum_{j=1}^{NT_N}-h_N(\mathbf{u}_2'\mathbf{z}_j)\left[2(y_j-\beta_{01}'\mathbf{w}_j-h_N\mathbf{u}_1'\mathbf{w}_j)-h_N\mathbf{u}_2'\mathbf{z}_j\right]+\lambda_N\sum_{k=1}^{m_N}|h_Nu_{2k}|^\gamma\\
&=h_N^2\sum_{j=1}^{NT_N}(\mathbf{z}_j'\mathbf{u}_2)^2+2h_N^2\sum_{j=1}^{NT_N}(\mathbf{w}_j'\mathbf{u}_1)(\mathbf{z}_j'\mathbf{u}_2)-2h_N\sum_{j=1}^{NT_N}(\mathbf{z}_j'\mathbf{u}_2)\epsilon_j+\lambda_Nh_N^\gamma\sum_{k=1}^{m_N}|u_{2j}|^\gamma
\end{aligned}$$

Regarding the sum of the first two terms since $2xy\geq-(x^2+y^2)$

$$\begin{aligned}
h_N^2\sum_{j=1}^{NT_N}(\mathbf{z}_j'\mathbf{u}_2)^2+2h_N^2\sum_{j=1}^{NT_N}(\mathbf{w}_j'\mathbf{u}_1)(\mathbf{z}_j'\mathbf{u}_2)&\geq h_N^2\sum_{j=1}^{NT_N}(\mathbf{z}_j'\mathbf{u}_2)^2-h_N^2\sum_{j=1}^{NT_N}\left[(\mathbf{w}_j'\mathbf{u}_1)^2+(\mathbf{z}_j'\mathbf{u}_2)^2\right]\\
&=-h_N^2NT_N\mathbf{u}_1'\Sigma_{1N}\mathbf{u}_1\geq-\rho_{1N}^{-2}p_N\tau_2C^2
\end{aligned}$$

where the last inequality follows from assumption (A2) and the fact that $||\mathbf{u}_1||\leq C$. Hence,

$$\frac{h_N^2\sum_{j=1}^{NT_N}(\mathbf{z}_i'\mathbf{u}_2)^2+2h_N^2\sum_{j=1}^{NT_N}(\mathbf{w}_i'\mathbf{u}_1)(\mathbf{z}_i'\mathbf{u}_2)}{p_N\rho_N^{-2}}\geq-\tau_2C^2$$

Regarding the third term it follows from Jensen's inequality (conditional version)

$$E\left(\left|\sum_{j=1}^{NT_N}(\mathbf{z}_j'\mathbf{u}_2)\epsilon_j\right|\middle|\mathbf{X}_N\right) \leq \left[E\left(\left(\sum_{j=1}^{NT_N}(\mathbf{z}_j'\mathbf{u}_2)\epsilon_j\right)^2\middle|\mathbf{X}_N\right]^{1/2} = \left[\sigma^2\sum_{j=1}^{NT_N}(\mathbf{z}_j'\mathbf{u}_2)^2\right]^{1/2}$$

$$= \sigma\left[\sum_{j=1}^{NT_N}\mathbf{u}_2'\mathbf{z}_j\mathbf{z}_j'\mathbf{u}_2\right]^{1/2} = \sigma(NT_N)^{1/2}\left(\mathbf{u}_2'\Sigma_{2N}\mathbf{u}_2\right)^{1/2} \leq \sigma(NT_N)^{1/2}\rho_{2N}^{1/2}C$$

where the last inequality used that

$$\rho_{2N} = \max_{\mathbf{u}\in\mathbf{R}^{p_N}}\frac{\mathbf{u}'\Sigma_N\mathbf{u}}{\mathbf{u}'\mathbf{u}} \geq \max_{\mathbf{u}_2\in\mathbf{R}^{m_N}}\frac{(\mathbf{0}',\mathbf{u}_2')\Sigma_N(\mathbf{0}',\mathbf{u}_2')'}{(\mathbf{0}',\mathbf{u}_2')(\mathbf{0}',\mathbf{u}_2')'} = \max_{\mathbf{u}_2\in\mathbf{R}^{m_N}}\frac{\mathbf{u}'\Sigma_{2N}\mathbf{u_2}}{\mathbf{u}_2'\mathbf{u}_2}$$

Hence, since h_N is measurable wrt. $\sigma(\mathbf{X}_N)$,

$$E\left(\left|-2h_N\sum_{j=1}^{NT_N}(\mathbf{z}_j'\mathbf{u}_2)\epsilon_j\right|\middle|\mathbf{X}_N\right) \leq 2\sigma h_N(NT_N)^{1/2}\rho_{2N}^{1/2}C$$

and so

$$E\left(\frac{\left|-2h_N\sum_{j=1}^{NT_N}(\mathbf{z}_j'\mathbf{u}_2)\epsilon_j\right|}{h_N(NT_N)^{1/2}\rho_{2N}^{1/2}}\right) \leq 2\sigma C$$

which by Lemma 2 shows that $-2h_N\sum_{j=1}^{NT_N}(\mathbf{z}_j'\mathbf{u}_2)\epsilon_j$ is $O_p(h_N(NT_N)^{1/2}\rho_{2N}^{1/2}) = O_p(\rho_{1N}^{-1}\rho_{2N}^{1/2}p_N^{1/2})$. Hence,

$$\frac{\left|-2h_N\sum_{j=1}^{NT_N}(\mathbf{z}_j'\mathbf{u}_2)\epsilon_j\right|}{p_N\rho_{1N}^{-2}} \in O_p\left(\frac{\rho_{1N}^{-1}\rho_{2N}^{1/2}p_N^{1/2}}{p_N\rho_{1N}^{-2}}\right) = O_p\left(\frac{\rho_{1N}\rho_{2N}^{1/2}}{p_N^{1/2}}\right) \subseteq O_p(1)$$

by assumption (A7). Regarding the fourth term since $\mathbf{u}_2 \neq 0$ we have

$$\lambda_N h_N^{\gamma}/(p_N\rho_{1N}^{-2}) = \lambda_N\left[\rho_{1N}^{-1}(p_N/NT_N)^{1/2}\right]^{\gamma}/(p_N\rho_{1N}^{-2}) = \lambda_N\rho_{1N}^{2-\gamma}(NT_N)^{-\gamma/2}p_N^{\gamma/2-1} \to \infty$$

by assumption (A4) and so the fourth term diverges to infinity. Since $p_N\rho_{1N}^{-2} \in \Omega_p(1)$ this completes the proof. □

Proof of Theorem 2. The first part has been established in Lemma 3. Since $\hat{\beta}_N$ is consistent it follows from assumption (A5) that for an arbitrary $\epsilon > 0$

$$P\left(\left\{\min\left\{|\hat{\beta}_{1Nj}|\,|1 \leq j \leq k\right\} + \epsilon < b_0\right\}\right) = P\left(\bigcup_{j=1}^{k}\left\{|\hat{\beta}_{1Nj}| + \epsilon < b_0\right\}\right)$$

$$= P\left(\bigcup_{j=1}^{k}\left\{b_0 - |\hat{\beta}_{1Nj}| > \epsilon\right\}\right) \leq P\left(\bigcup_{j=1}^{k}\left\{|\beta_{10j}| - |\hat{\beta}_{1Nj}| > \epsilon\right\}\right)$$

$$\leq P\left(\bigcup_{j=1}^{k}\left\{|\beta_{10j} - \hat{\beta}_{1Nj}| > \epsilon\right\}\right) \leq P\left(||\beta_{10} - \hat{\beta}_{1Nj}|| > \epsilon\right) \to 0$$

Choosing $\epsilon = b_0/2$ shows that with probability converging to one $\min\left\{|\hat{\beta}_{1Nj}|\ |1 \le j \le k\right\} \ge b_0/2$ and so $\hat{\beta}_{1N}$ is bounded away from 0. Hence L_N is differentiable at $\hat{\beta}_{1N}$ with probability converging to one. And so $\hat{\beta}_{1N}$ satisfies

$$\frac{\partial}{\partial \beta_1} L_N(\hat{\beta}_{1N}, \hat{\beta}_{2N}) = 0$$

That is,

$$-2\sum_{j=1}^{NT_N}\left(y_j - \mathbf{w}_j'\hat{\beta}_{1N} - \mathbf{z}_j'\hat{\beta}_{2N}\right)\mathbf{w}_j + \lambda_N\gamma\psi_N = 0$$

with probability converging to 1 where ψ_N is a $k \times 1$ vector with l'th entry given by $\psi_{Nl} = |\hat{\beta}_{1Nl}|^{\gamma-1}\text{sign}(\hat{\beta}_{1Nl})$. This can be rewritten as

$$-2\sum_{j=1}^{NT_N}\left(\epsilon_j - \mathbf{w}_j'(\hat{\beta}_{1N} - \beta_{10}) - \mathbf{z}_j'\hat{\beta}_{2N}\right)\mathbf{w}_j + \lambda_N\gamma\psi_N = 0 \Leftrightarrow$$

$$-2\sum_{j=1}^{NT_N}\epsilon_j\mathbf{w}_j + 2\sum_{j=1}^{NT_N}\mathbf{w}_j\mathbf{w}_j'(\hat{\beta}_{1N} - \beta_{10}) + 2\sum_{j=1}^{NT_N}\mathbf{z}_j'\hat{\beta}_{2N}\mathbf{w}_j + \lambda_N\gamma\psi_N = 0 \Leftrightarrow$$

$$\mathbf{\Sigma}_{1N}(\hat{\beta}_{1N} - \beta_{10}) = \frac{1}{NT_N}\sum_{j=1}^{NT_N}\epsilon_j\mathbf{w}_j - \frac{\lambda_N\gamma\psi_N}{2NT_N} - \frac{1}{NT_N}\sum_{j=1}^{NT_N}\mathbf{z}_j'\hat{\beta}_{2N}\mathbf{w}_j$$

Hence, for any $k \times 1$ vector α

$$(NT_N)^{1/2}\alpha'(\hat{\beta}_{1N} - \beta_{10}) = (NT_N)^{-1/2}\sum_{j=1}^{NT_N}\alpha'\mathbf{\Sigma}_{1N}^{-1}\epsilon_j\mathbf{w}_j$$

$$-(1/2)\gamma(NT_N)^{-1/2}\lambda_N\alpha'\mathbf{\Sigma}_{1N}^{-1}\psi_N - (NT_N)^{-1/2}\sum_{j=1}^{NT_N}\alpha'\mathbf{\Sigma}_{1N}^{-1}\mathbf{z}_j'\hat{\beta}_{2N}\mathbf{w}_j$$

Since $P(\hat{\beta}_{2N} = 0) \to 1$ the last term equals 0 with probability converging to 1. From the Cauchy-Schwarz inequality in $\mathbf{R}^k$ it follows that

$$|\alpha'\mathbf{\Sigma}_{1N}^{-1}\psi_N| \le ||\alpha'\Sigma_{1N}^{-1}||||\psi_N||$$

Since

$$||\alpha'\Sigma_{1N}^{-1}||^2 = \alpha'\mathbf{\Sigma}_{1N}^{-1}\mathbf{\Sigma}_{1N}^{-1}\alpha = \alpha'\mathbf{\Sigma}_{1N}^{-2}\alpha \le \tau_{1N}^{-2}||\alpha||^2 \le \tau_1^{-2}||\alpha||^2$$

by assumption (A2), we get with probability converging to one

$$\begin{aligned}|(1/2)(NT_N)^{-1/2}\gamma\lambda_N\alpha'\mathbf{\Sigma}_{1N}^{-1}\psi_N| &\le (1/2)(NT_N)^{-1/2}\gamma\lambda_N\tau_1^{-1}||\alpha||||\psi_N|| \\ &\le (1/2)(NT_N)^{-1/2}\gamma\lambda_N\tau_1^{-1}||\alpha||k^{1/2}(b_0/2)^{(\gamma-1)} \\ &= (1/2)\gamma\tau_1^{-1}||\alpha||(b_0/2)^{(\gamma-1)}T_N^{-1/2}N^{-1/2}\lambda_N k^{1/2} \to 0\end{aligned}$$

by assumption (A3). Hence,

$$(NT_N)^{1/2}\alpha'(\hat{\beta}_{1N} - \beta_{10}) \in (NT_N)^{-1/2}\sum_{j=1}^{NT_N}\alpha'\mathbf{\Sigma}_{1N}^{-1}\epsilon_j\mathbf{w}_j + o_p(1)$$

Since $s_N^{-1}=1/\sqrt{\sigma^2\alpha'\mathbf{\Sigma}_{1N}^{-1}\alpha}\leq \tau_2^{1/2}/(\sigma||\alpha||)$ by assumption (A2) it is also true that

$$(NT_N)^{1/2}s_N^{-1}\alpha'(\hat{\beta}_{1N}-\beta_{10})\in (NT_N)^{-1/2}s_N^{-1}\sum_{j=1}^{NT_N}\alpha'\mathbf{\Sigma}_{1N}^{-1}\epsilon_j\mathbf{w}_j+o_p(1) \tag{7.3}$$

Now, defining $\mathbf{W}_{i,N}=(\mathbf{w}_{i1},...,\mathbf{w}_{iT_N})'$ and $\epsilon_{i,N}=(\epsilon_{i1},...,\epsilon_{iNT_N})'$ one first notices that[12]

$$\frac{1}{NT_N}\sum_{j=1}^{NT_N}\mathbf{w}_j\mathbf{w}_j'\to\lim_{N\to\infty}\frac{1}{NT_N}\sum_{j=1}^{NT_N}E\left(\mathbf{w}_j\mathbf{w}_j'\right)=\lim_{N\to\infty}E\left(\frac{1}{T_N}\mathbf{W}_{1N}'\mathbf{W}_{1N}\right) \tag{7.4}$$

where the first limit is in probability. To see why (7.4) is true let Z_j be a fixed entry in $\mathbf{w}_j\mathbf{w}_j',\ j=1,...,NT_N$. Letting $\eta>0$ be arbitrary and using the Markov inequality

$$P\left(\left|\frac{1}{NT_N}\sum_{j=1}^{NT_N}Z_j-\frac{1}{NT_N}\sum_{j=1}^{NT_N}E\left(Z_j\right)\right|>\eta\right)\leq\frac{E\left(\frac{1}{NT_N}\sum_{j=1}^{NT_N}\left[Z_j-E\left(Z_j\right)\right]\right)^2}{\eta^2}$$

$$=\frac{E\left(\sum_{i=1}^{N}\sum_{t=1}^{T_N}\left[Z_{it}-E\left(Z_{it}\right)\right]\right)^2}{(NT_N\eta)^2}=\frac{\sum_{i=1}^{N}E\left(\sum_{t=1}^{T_N}\left[Z_{it}-E\left(Z_{it}\right)\right]\right)^2}{(NT_N\eta)^2}$$

$$=\frac{N\left(\sum_{t=1}^{T_N}Var(Z_{1t})+2\sum_{t=1}^{T_N}\sum_{s>t}^{T_N}Cov\left(Z_{1t},\ Z_{1s}\right)\right)}{(NT_N\eta)^2}$$

$$\leq\frac{N\left(T_N\max_{1\leq t\leq T_N}Var\left(Z_{1t}\right)+2T_N(T_N-1)/2\max_{1\leq t\leq T_N}Var\left(Z_{1t}\right)\right)}{(NT_N\eta)^2}$$

$$=\frac{\max_{1\leq t\leq T_N}Var\left(Z_{1t}\right)}{N\eta^2}\to 0$$

Hence,

$$(NT_N)^{-1/2}s_N^{-1}\sum_{j=1}^{NT_N}\alpha'\mathbf{\Sigma}_{1N}^{-1}\epsilon_j\mathbf{w}_j=\frac{(NT_N)^{-1/2}\sum_{j=1}^{NT_N}\alpha'\left(\frac{1}{NT_N}\sum_{j=1}^{NT_N}\mathbf{w}_j\mathbf{w}_j'\right)^{-1}\epsilon_j\mathbf{w}_j}{\sqrt{\sigma^2\alpha'\left(\frac{1}{NT_N}\sum_{j=1}^{NT_N}\mathbf{w}_j\mathbf{w}_j'\right)^{-1}\alpha}}$$

$$\in\frac{N^{-1/2}}{\sqrt{\sigma^2\alpha'\left(E\left[\frac{1}{T_N}\mathbf{W}_{1N}'\mathbf{W}_{1N}\right]\right)^{-1}\alpha}}\sum_{i=1}^{N}\alpha'\left(E\left[\frac{1}{T_N}\mathbf{W}_{1N}'\mathbf{W}_{1N}\right]\right)^{-1}T_N^{-1/2}\mathbf{W}_{iN}'\epsilon_{iN}+o_p(1)$$

Now,

$$E\left(\sum_{i=1}^{N}\alpha'\left(E\left[\frac{1}{T_N}\mathbf{W}_{1N}'\mathbf{W}_{1N}\right]\right)^{-1}T_N^{-1/2}\mathbf{W}_{iN}'\epsilon_{iN}\right)=0$$

[12] All limits are taken elementwise in the matrices.

by iterated expectations and

$$\begin{aligned} r_N^2 :&= E\left[\left(\sum_{i=1}^N \alpha'\left(E\left[\frac{1}{T_N}\mathbf{W}_{1N}'\mathbf{W}_{1N}\right]\right)^{-1} T_N^{-1/2}\mathbf{W}_{iN}'\epsilon_{iN}\right)^2\right] \\ &= \sum_{i=1}^N E\left(\alpha'\left(E\left[\frac{1}{T_N}\mathbf{W}_{1N}'\mathbf{W}_{1N}\right]\right)^{-1} T_N^{-1/2}\mathbf{W}_{iN}'\epsilon_{iN}\epsilon_{iN}'\mathbf{W}_{iN}T_N^{-1/2}\left(E\left[\frac{1}{T_N}\mathbf{W}_{1N}'\mathbf{W}_{1N}\right]\right)^{-1}\alpha\right) \\ &= \sigma^2 N\alpha'\left(E\left[\frac{1}{T_N}\mathbf{W}_{1N}'\mathbf{W}_{1N}\right]\right)^{-1}\alpha \end{aligned}$$

Finally, let $U_{iN} = \alpha'\left(E\left[\frac{1}{T_N}\mathbf{W}_{1N}'\mathbf{W}_{1N}\right]\right)^{-1} T_N^{-1/2}\mathbf{W}_{iN}'\epsilon_{iN}$. Since $E(U_{iN}^2) = \sigma^2\alpha'\left(E\left[\frac{1}{T_N}\mathbf{W}_{1N}'\mathbf{W}_{1N}\right]\right)^{-1}\alpha$ is convergent it is bounded. Furthermore,

$$\alpha' E\left(\frac{1}{T_N}\mathbf{W}_{1N}'\mathbf{W}_{1N}\right)\alpha = E\left(\alpha'\frac{1}{NT_N}\mathbf{W}_N'\mathbf{W}_N\alpha\right) \in \left[\alpha'\alpha\tau_1, \alpha'\alpha\tau_2\right]$$

and so the eigenvalues of $E\left[\frac{1}{T_N}\mathbf{W}_{1N}'\mathbf{W}_{1N}\right]$ are also contained in $[\tau_1,\tau_2]$. This implies $r_N^2 = \sigma^2 N\alpha'\left(E\left[\frac{1}{T_N}\mathbf{W}_{1N}'\mathbf{W}_{1N}\right]\right)^{-1}\alpha \geq \sigma^2 N\alpha'\alpha/\tau_2$. Hence, the Lindeberg condition is satisfied since for all $\delta > 0$

$$\lim_{N\to\infty}\left\{r_N^{-2}\sum_{i=1}^N \int_{\left\{|U_{i,N}|>\delta r_N\right\}} U_{iN}^2 dP\right\} \leq \lim_{N\to\infty}\frac{\tau_2}{\alpha'\alpha\sigma^2}\int_{\left\{|U_{1N}|>\delta\sqrt{\frac{\alpha'\alpha\sigma^2}{\tau_2}}\sqrt{N}\right\}} U_{1N}^2 dP = 0$$

since for any $\rho > 0$ (let $\delta\sqrt{\frac{\alpha'\alpha\sigma^2}{\tau_2}} = K$)

$$\lim_{N\to\infty} P\left(U_{1N}^2\mathbf{1}_{\left\{|U_{1N}|>K\sqrt{N}\right\}} > \rho\right) \leq \lim_{N\to\infty} P\left(\left\{|U_{1N}| > K\sqrt{N}\right\}\right) \leq \lim_{N\to\infty}\frac{E\left(U_{1N}^2\right)}{K^2 N} = 0$$

and $\left\{U_{1N}^2\right\}_{N=1}^\infty$ is uniformly integrable[13]. Hence,

$$\frac{N^{-1/2}}{\sqrt{\sigma^2\alpha'\left(E\left[\frac{1}{T_N}\mathbf{W}_{1N}'\mathbf{W}_{1N}\right]\right)^{-1}\alpha}}\sum_{i=1}^N \alpha'\left(E\left[\frac{1}{T_N}\mathbf{W}_{1N}'\mathbf{W}_{1N}\right]\right)^{-1} T_N^{-1/2}\mathbf{W}_{iN}'\epsilon_{iN} \xrightarrow{d} N(0,1)$$

And so by (7.3),

$$(NT_N)^{1/2}s_N^{-1}\alpha'(\hat{\beta}_{1N} - \beta_{10}) \xrightarrow{d} N(0,1)$$

[13] The uniform integrability of $\left\{U_{1N}^2\right\}_{N=1}^\infty$ implies that $\left\{U_{1N}^2\mathbf{1}_{\left\{|U_{1N}|>K\sqrt{N}\right\}}\right\}_{N=1}^\infty$ is uniformly integrable which is what is needed to utilize the well known fact that convergence in measure plus uniform integrability is equivalent to convergence in L^1.

or equivalently,

$$(NT_N)^{1/2}(\hat{\beta}_{1N}-\beta_{10}) \xrightarrow{d} N\left(0,\sigma^2\left(\lim_{N\to\infty} E\left[\frac{1}{T_N}\mathbf{W}'_{1N}\mathbf{W}_{1N}\right]\right)^{-1}\right) \tag{7.5}$$

□

Proof of Theorem 3. (i) If $T_N = T$ for a fixed T, $U_{1N} = U_1$ for all N where U_1 is defined in the obvious way, does not depend on N and belongs to L^2. Hence,

$$\lim_{K\to\infty}\sup_{1\leq N<\infty}\int_{\left\{|U_{1N}|>K\right\}} U_{1N}^2 dP = \lim_{K\to\infty}\int_{\left\{|U_1|>K\right\}} U_1^2 dP = 0$$

by Lebesgue's Dominated Convergence Theorem.

(ii) By the Cauchy-Schwarz inequality

$$|U_{1N}| \leq \left\|\alpha'\left(E\left[\frac{1}{T_N}\mathbf{W}'_{1N}\mathbf{W}_{1N}\right]\right)^{-1}\right\|\left\|T_N^{-1/2}\mathbf{W}'_{1N}\epsilon_{1N}\right\|$$

Since $\left\|\alpha'\left(E\left[\frac{1}{T_N}\mathbf{W}'_{1N}\mathbf{W}_{1N}\right]\right)^{-1}\right\|$ is convergent it is bounded by a constant C. Hence,

$$U_{1N}^2 \leq C^2\left\|T_N^{-1/2}\mathbf{W}'_{1N}\epsilon_{1N}\right\|^2 = C^2\sum_{i=1}^{k}\frac{1}{T_N}\sum_{j=1}^{T_N}\left(\mathbf{W}_{1N}^{ji}\epsilon_{1N}^{j}\right)^2$$

where $\mathbf{W}_{1N}^{ji}$ is jth row in the ith column of $\mathbf{W}_{1N}$. Since the rows of $\mathbf{W}_{1N}$ are identically distributed and $\mathbf{W}_{1N}^{ji} \perp\!\!\!\perp \epsilon_{1N}^{j}$ one has (calculate the characteristic functions and notice they are identical) $\mathbf{W}_{1N}^{ji}\epsilon_{1N}^{j} \sim \mathbf{W}_{1N}^{1i}\epsilon_{1N}^{1}$ where $\mathbf{W}_{1N}^{1i}\epsilon_{1N}^{1} \sim Z_i$ for some $Z_i \in L^2$. Hence, $\left\{\left\{\left(\mathbf{W}_{1N}^{ji}\epsilon_{1N}^{j}\right)^2\right\}_{j=1}^{T_N}\right\}_{N=1}^{\infty}$ is uniformly integrable for all $1 \leq i \leq k$ since

$$\begin{aligned}
&\lim_{K\to\infty}\sup_{1\leq N<\infty}\sup_{1\leq j\leq T_N}\int_{\left\{\left(\mathbf{W}_{1N}^{ji}\epsilon_{1N}^{j}\right)^2>K\right\}}\left(\mathbf{W}_{1N}^{ji}\epsilon_{1N}^{j}\right)^2 dP\\
&= \lim_{K\to\infty}\sup_{1\leq N<\infty}\sup_{1\leq j\leq T_N}\int_{\{x^2>K\}} x^2 dP_{\mathbf{W}_{1N}^{ji}\epsilon_{1N}^{j}}(x)\\
&= \lim_{K\to\infty}\sup_{1\leq N<\infty}\sup_{1\leq j\leq T_N}\int_{\{x^2>K\}} x^2 dP_{Z_i}(x)\\
&= \lim_{K\to\infty}\int_{\{x^2>K\}} x^2 dP_{Z_i}(x) = 0
\end{aligned}$$

by Lebesgue's Dominated Convergence Theorem. By Hoffmann-Jørgensen (1994) (page 338)[14] this implies that $\left\{\frac{1}{T_N}\sum_{j=1}^{T_N}\left(\mathbf{W}_{1N}^{ji}\epsilon_{1N}^{j}\right)^2\right\}_{N=1}^{\infty}$ is uniformly integrable

[14] The partial averages of a uniformly integrable sequence are themselves uniformly integrable.

which in turn implies that $\left\{\sum_{i=1}^{k}\frac{1}{T_N}\sum_{j=1}^{T_N}\left(\mathbf{W}_{1N}^{ji}\epsilon_{1N}^{j}\right)^2\right\}_{N=1}^{\infty}$ is uniformly integrable by Hoffmann-Jørgensen (1994) (page 337)[15]. Since $\left\{C^2\sum_{i=1}^{k}\frac{1}{T_N}\sum_{j=1}^{T_N}\left(\mathbf{W}_{1N}^{ji}\epsilon_{1N}^{j}\right)^2\right\}_{N=1}^{\infty}$ dominates $\left\{U_{1N}^2\right\}$ this yields the desired result.

(iii) If $\mathbf{W}_{1N}$ and ϵ_{1N} are uniformly bounded U_{1N} has moments of any order and so $\left\{U_{1N}^2\right\}_{N=1}^{\infty}$ is uniformly integrable. □

Proof of Corollary 1. For fixed T_N (7.5) reads

$$N^{1/2}(\hat{\beta}_{1N}-\beta_{10})\xrightarrow{d}N\left(0,\sigma^2\left(E\left[\mathbf{W}_1'\mathbf{W}_1\right]\right)^{-1}\right)$$

where absence of subscript N indicates that the matrices no longer depend on T_N. In the fixed effects setting $\mathbf{W}_1=\left(\mathbf{D}\mathbf{D}'\right)^{-1/2}\mathbf{D}\tilde{\mathbf{W}}_1$ and so

$$\begin{aligned}E\left(\mathbf{W}_1'\mathbf{W}_1\right)&=E\left(\tilde{\mathbf{W}}_1'\mathbf{D}'\left(\mathbf{D}\mathbf{D}'\right)^{-1/2}\left(\mathbf{D}\mathbf{D}'\right)^{-1/2}\mathbf{D}\tilde{\mathbf{W}}_1\right)\\&=E\left(\tilde{\mathbf{W}}_1'\mathbf{D}'\left(\mathbf{D}\mathbf{D}'\right)^{-1}\mathbf{D}\tilde{\mathbf{W}}_1\right)=E\left(\ddot{\tilde{\mathbf{W}}}_1'\ddot{\tilde{\mathbf{W}}}_1\right)\end{aligned}$$

where the last inequality used that $\mathbf{D}'\left(\mathbf{D}\mathbf{D}'\right)^{-1}\mathbf{D}$ is symmetric and idempotent and that premultiplication of it corresponds to columnwise demeaning.

The proof of part (ii) is similar using $\mathbf{W}_1=\Omega^{-1/2}\tilde{\mathbf{W}}_1$. □

Next we turn to the properties of the Marginal Bridge estimator.

Lemma 4. *For q even and any $w_N>0$,*

$$P\left(w_N>\max_{1\le k\le m}\left|\sum_{j=1}^{NT}x_{jk}\epsilon_j\right|\right)\ge 1-\frac{Km(NT_N)^{q/2}}{w_N^q}$$

Proof. By the Markov inequality,

$$\begin{aligned}P\left(w_N>\max_{1\le k\le m}\left|\sum_{j=1}^{NT}x_{jk}\epsilon_j\right|\right)&=1-P\left(\max_{1\le k\le m}\left|\sum_{j=1}^{NT_N}x_{jk}\epsilon_j\right|\ge w_N\right)\\&\ge 1-\frac{E\left(\max_{1\le k\le m}\left|\sum_{j=1}^{NT_N}x_{jk}\epsilon_j\right|^q\right)}{w_N^q}\end{aligned}$$

Since for any sequence of random variables $\{Z_k\}_{k=1}^{m}\subseteq L^q$

$$E\left(\max_{1\le k\le m}Z_k^q\right)\le E\left(\sum_{k=1}^{m}Z_k^q\right)\le m\max_{1\le k\le m}E\left(Z_k^q\right)$$

[15] Finite sums of uniformly integrable sequences are themselves uniformly integrable.

it suffices to show that $E\left(\sum_{j=1}^{NT_N} x_{jk}\epsilon_j\right)^q \leq K(NT_N)^{q/2}$ for all $k \in \{1, ..., m\}$. By the multinomial theorem

$$E\left(\sum_{j=1}^{NT_N} x_{jk}\epsilon_j\right)^q = E\left(\sum_{|\alpha|=q}\binom{q}{\alpha}\prod_{j=1}^{NT_N}(x_{jk}\epsilon_j)^{\alpha_j}\right)$$

where we have used the multi-index convention $\alpha = (\alpha_1, ..., \alpha_{NT})$ and $|\alpha| = q$ means $\sum_{j=1}^{NT_N}\alpha_j = q$. The sum is taken over all such vectors α whose entries add up to q. Since $(\mathbf{X}_{iN}, \epsilon_{iN})$ is i.i.d. and $\mathbf{X}_{iN} \perp\!\!\!\perp \epsilon_{iN}$ it follows that $(\mathbf{X}_{1N}, \epsilon_{1N}, ..., \mathbf{X}_{NN}, \epsilon_{NN})$ are independent (calculate the characteristic function and observe that it factorizes). This implies $\mathbf{X}_N \perp\!\!\!\perp \epsilon_N$. Finally, it is seen that $\epsilon_{11}, ..., \epsilon_{NT_N}$ are independent and (again calculate the characteristic function and observe that it factorizes by using $(\mathbf{X}_{iN}, \epsilon_{iN})$ is i.i.d and the that uncorrelatedness implies independence for gaussian variables.). From these observations it follows that all summands including an $\alpha_j = 1$ for some $j = 1, ..., NT_N$ equal 0 since $E(\epsilon_j) = 0$ for all $j = 1, ..., NT_N$. It remains to be shown that no other term is of greater order than $O\left((NT_N)^{q/2}\right)$. So assume $\min_{1\leq j\leq NT_N}\alpha_j \geq 2$ [16]. Consider the set A of all vectors α with s entries different from 0 and larger than or equal to 2 with non-zero values $a_1, ..., a_s$ and $a_1 + ... + a_s = q$.

$$\begin{aligned}
&E\left(\sum_{\alpha\in A}\prod_{j=1}^{NT_N}(x_{jk}\epsilon_j)^{\alpha_j}\right)\\
&= E\left(\sum_{j_1=1}^{NT_N}\sum_{j_2\neq j_1}^{NT_N}\cdots\sum_{j_s\notin\{j_1,...,j_{s-1}\}}^{NT_N}(x_{j_1k}\epsilon_{j_1})^{a_1}(x_{j_2k}\epsilon_{j_2})^{a_2}\cdot...\cdot(x_{j_sk}\epsilon_{j_s})^{a_s}\right)\\
&= KE\left(\sum_{j_1=1}^{NT_N}\sum_{j_2\neq j_1}^{NT_N}\cdots\sum_{j_s\notin\{j_1,...,j_{s-1}\}}^{NT_N}(x_{j_1k})^{a_1}(x_{j_2k})^{a_2}\cdot...\cdot(x_{j_sk})^{a_s}\right)\\
&\leq K(NT_N)^{q/2}E\left(\sum_{j_1=1}^{NT_N}\sum_{j_2\neq j_1}^{NT_N}\cdots\sum_{j_s\notin\{j_1,...,j_{s-1}\}}^{NT_N}\left[\frac{x_{j_1k}^2}{NT_N}\right]^{a_1/2}\left[\frac{x_{j_2k}^2}{NT_N}\right]^{a_2/2}\cdot...\cdot\left[\frac{x_{j_sk}^2}{NT_N}\right]^{a_s/2}\right)\\
&= K(NT_N)^{q/2}E\left(\sum_{j_1=1}^{NT_N}\left[\frac{x_{j_1k}^2}{NT_N}\right]^{a_1/2}\sum_{j_2\neq j_1}^{NT_N}\left[\frac{x_{j_2k}^2}{NT_N}\right]^{a_2/2}\cdot...\cdot\sum_{j_s\notin\{j_1,...,j_{s-1}\}}^{NT_N}\left[\frac{x_{j_sk}^2}{NT_N}\right]^{a_s/2}\right)\\
&\leq K(NT_N)^{q/2}
\end{aligned}$$

where the second equality used $\mathbf{X}_N \perp\!\!\!\perp \epsilon_N$ and $K := E(\epsilon_1^{a_1})\cdot...\cdot E(\epsilon_1^{a_s})$. The first estimate follows from the nonnegativity of all terms on the right hand side. The last estimate uses that for any subset S of $\{1, ..., NT_N\}$, $\sum_{j\in S}\left[x_{jk}^2/(NT_N)\right]^{a_i/2} \leq \sum_{j=1}^{NT_N}\left[x_{jk}^2/(NT_N)\right]^{a_i/2} \leq \sum_{j=1}^{NT_N}\left[x_{jk}^2/(NT_N)\right] = 1$ since $\left[x_{jk}^2/(NT_N)\right] \leq 1$ and

[16] Here we follow the convention of setting $t^0 = 1$ even when $t = 0$ and so we only have to consider the case $\min_{1\leq j\leq NT_N}\alpha_j \geq 2$.

so $\left[x_{jk}^2/(NT_N)\right]^{a_i/2} \leq \left[x_{jk}^2/(NT_N)\right]$ for $a_i \geq 2$. Since s was arbitrary and there are only finitely many terms of the above kind this establishes the lemma.

□

Proof of Theorem 4. Let $\epsilon_N = (\epsilon_1, ..., \epsilon_{NT_N})'$ and recall $\xi_{Nk} = 1/\left(NT_N\right)\sum_{j=1}^{NT_N} \mathbf{w}_j'\beta_{10}x_{jk}$. Then,

$$\begin{aligned} U_N(\beta) &= \sum_{k=1}^{p_N}\sum_{j=1}^{NT_N}\left(y_j - x_{jk}\beta_k\right)^2 + \lambda_N\sum_{k=1}^{p_N}|\beta_k|^\gamma \\ &= \sum_{k=1}^{p_N}\left[\sum_{j=1}^{NT_N} y_j^2 + NT_N\beta_k^2 - 2\sum_{j=1}^{NT_N} y_j x_{jk}\beta_k\right] + \lambda_N\sum_{k=1}^{p_N}|\beta_k|^\gamma \\ &= \sum_{k=1}^{p_N}\left[\sum_{j=1}^{NT_N}\epsilon_j^2 + \sum_{j=1}^{NT_N}(\mathbf{w}_j'\beta_{01})^2 + 2\sum_{j=1}^{NT_N}(\mathbf{w}_j'\beta_{01})\epsilon_j + NT_N\beta_k^2 - 2\sum_{j=1}^{NT_N}(\epsilon_j + \mathbf{w}_j'\beta_{01})x_{jk}\beta_k\right] + \lambda_N\sum_{k=1}^{p_N}|\beta_k|^\gamma \\ &= \sum_{k=1}^{p_N}\left[\sum_{j=1}^{NT_N}\epsilon_j^2 + \sum_{j=1}^{NT_N}(\mathbf{w}_j'\beta_{01})^2 + 2\sum_{j=1}^{NT_N}(\mathbf{w}_j'\beta_{01})\epsilon_j + NT_N\beta_k^2 - 2(\epsilon_N'\mathbf{x}_k + NT_N\xi_{Nk})\beta_k\right] + \lambda_N\sum_{k=1}^{p_N}|\beta_k|^\gamma \end{aligned}$$

So minimizing $U_N(\beta)$ is equivalent to minimizing $\sum_{k=1}^{p_N}\left[NT_N\beta_k^2 - 2(\epsilon_N'\mathbf{x}_k + NT_N\xi_{Nk})\beta_k + \lambda_N|\beta_k|^\gamma\right]$. Since $0 < \gamma < 1$ it follows from Lemma A of Knight and Fu (2000) that $\beta_k = 0$ if and only if

$$\lambda_N/(NT_N) > |\epsilon_N'\mathbf{x}_k/(NT_N) + \xi_{Nk}|^{2-\gamma}c_\gamma$$

where $c_\gamma = \left(\frac{2}{2-\gamma}\right)\left(\frac{2(2-\gamma)}{2-\gamma}\right)^{1-\gamma}$. Defining $w_N = c_\gamma^{-1/(2-\gamma)}(\lambda_N/(NT_N)^{\gamma/2})^{1/(2-\gamma)}$ the above inequality is equivalent to

$$w_N > (NT_N)^{-1/2}|\epsilon_N'\mathbf{x}_k + (NT_N)\xi_{Nk}|$$

So to prove the theorem it is enough to show

$$P\left(w_N > (NT_N)^{-1/2}\max_{k\in J_N}\left|\epsilon_N'\mathbf{x}_k + (NT_N)\xi_{Nk}\right|\right) \to 1 \tag{7.6}$$

and

$$P\left(w_N > (NT_N)^{-1/2}\min_{k\in K_N}\left|\epsilon_N'\mathbf{x}_k + (NT_N)\xi_{Nk}\right|\right) \to 0 \tag{7.7}$$

We first prove (7.6). On $A_N = \left\{\frac{\left|\sum_{j=1}^{NT_N} x_{jk}x_{jl}\right|}{(NT_N)^{1/2}} \leq c_0,\ k \in K_N, l \in J_N\right\}$ and under assumption (B6)

$$\begin{aligned} &\max_{l\in J_N}(NT_N)^{1/2}|\xi_{Nl}| = \max_{l\in J_N}\frac{1}{(NT_N)^{1/2}}\left|\sum_{j=1}^{NT_N}\sum_{k=1}^{k_N} x_{jk}\beta_{10k}x_{jl}\right| \\ &= \max_{l\in J_N}\frac{1}{(NT_N)^{1/2}}\left|\sum_{k=1}^{k_N}\beta_{10k}\sum_{j=1}^{NT_N} x_{jk}x_{jl}\right| \leq \max_{l\in J_N}\sum_{k=1}^{k_N}|\beta_{10k}|\left|\frac{1}{(NT_N)^{1/2}}\sum_{j=1}^{NT_N} x_{jk}x_{jl}\right| \\ &\leq b_1c_0k_N \end{aligned}$$

Hence,

$$\begin{aligned}
&P\left(w_N > (NT_N)^{-1/2}\max_{k\in J_N}|\epsilon_N'\mathbf{x}_k + (NT_N)\xi_{Nk}|\right)\\
&\geq P\left(w_N > (NT_N)^{-1/2}\max_{k\in J_N}|\epsilon_N'\mathbf{x}_k| + (NT_N)^{1/2}\max_{k\in J_N}|\xi_{Nk}|,\ A_N\right)\\
&\geq P\left(w_N > (NT_N)^{-1/2}\max_{k\in J_N}|\epsilon_N'\mathbf{x}_k| + b_1c_0k_N,\ A_N\right)\\
&\geq P\left(w_N > (NT_N)^{-1/2}\max_{k\in J_N}|\epsilon_N'\mathbf{x}_k| + b_1c_0k_N\right) + P(A_N) - 1
\end{aligned}$$

where the last estimate follows from the inclusion-exclusion principle. By Lemma 4 for every even q,

$$\begin{aligned}
&P\left((NT_N)^{1/2}\left(w_N - b_1c_0k_N\right) > \max_{k\in J_N}|\epsilon_N'\mathbf{x}_k|\right) \geq 1 - \frac{Km_N(NT_N)^{q/2}}{\left((NT_N)^{1/2}[w_N - b_1c_0k_N]\right)^q}\\
&= 1 - \frac{Km_N/w_N^q}{\left(1 - b_1c_0k_N/w_N\right)^q}
\end{aligned}$$

Furthermore, by assumption (B4) we may choose q sufficiently large such that

$$m_N/w_N^q \in O(1)\frac{m_N}{\left(\lambda_N(NT_N)^{-\gamma/2}\right)^{q/(2-\gamma)}} \to 0$$

and by assumption (B3)

$$b_1c_0k_N/w_N \in O(1)\frac{k_N}{\left(\lambda_N(NT_N)^{-\gamma/2}\right)^{1/(2-\gamma)}} \to 0$$

Finally, $P(A_N)$ can be made arbitrarily close to 1 by assumption (B5) which establishes (7.6). Next we verify (7.7).

$$\begin{aligned}
P\left(w_N > (NT_N)^{-1/2}\min_{k\in K_N}|\epsilon_N'\mathbf{x}_k + (NT_N)\xi_{Nk}|\right) &= P\left(\bigcup_{k\in K_N}\left\{w_N > |(NT_N)^{-1/2}\epsilon_N'\mathbf{x}_k + (NT_N)^{1/2}\xi_{Nk}|\right\}\right)\\
&\leq \sum_{k\in K_N} P\left(\left\{w_N > |(NT_N)^{-1/2}\epsilon_N'\mathbf{x}_k + (NT_N)^{1/2}\xi_{Nk}|\right\}\right)
\end{aligned} \tag{7.8}$$

Since $\min_{k\in K_N}|\xi_{Nk}| \geq \xi_0 > 0$ by assumption (B1) we may write,

$$\begin{aligned}
&P\left(\left\{w_N > |(NT_N)^{-1/2}\epsilon_N'\mathbf{x}_k + (NT_N)^{1/2}\xi_{Nk}|\right\}\right) \leq P\left(\left\{w_N > (NT_N)^{1/2}|\xi_{Nk}| - (NT_N)^{-1/2}|\epsilon_N'\mathbf{x}_k|\right\}\right)\\
&\leq P\left(\left\{w_N > (NT_N)^{1/2}\xi_0 - (NT_N)^{-1/2}|\epsilon_N'\mathbf{x}_k|\right\}\right) = P\left(\left\{(NT_N)^{-1/2}|\epsilon_N'\mathbf{x}_k| > (NT_N)^{1/2}\xi_0 - w_N\right\}\right)\\
&= 1 - P\left(\left\{(NT_N)^{1/2}\xi_0 - w_N \geq (NT_N)^{-1/2}|\epsilon_N'\mathbf{x}_k|\right\}\right) \leq 1 - P\left(\left\{(NT_N)^{1/2}\xi_0 - w_N > (NT_N)^{-1/2}|\epsilon_N'\mathbf{x}_k|\right\}\right)
\end{aligned}$$

By Lemma 4,

$$P\left(\left\{(NT_N)^{1/2}\left((NT_N)^{1/2}\xi_0 - w_N\right) > |\epsilon_N'\mathbf{x}_k|\right\}\right) \geq 1 - \frac{K(NT_N)^{q/2}}{\left((NT_N)^{1/2}[(NT_N)^{1/2}\xi_0 - w_N]\right)^q} \tag{7.9}$$

$$= 1 - \frac{K}{((NT_N)^{1/2}\xi_0 - w_N)^q} \tag{7.10}$$

And so for all $k \in K_N$

$$P\left(\left\{w_N > |(NT_N)^{-1/2}\epsilon_N'\mathbf{x}_k + (NT_N)^{1/2}\xi_{Nk}|\right\}\right) \leq \frac{K}{((NT_N)^{1/2}\xi_0 - w_N)^q}$$

Inserting this into (7.8) yields

$$P\left(w_N > (NT_N)^{-1/2}\min_{k\in k_N}|\epsilon_N'\mathbf{x}_k + (NT_N)\xi_{Nk}|\right) \leq \frac{Kk_N}{((NT_N)^{1/2}\xi_0 - w_N)^q}$$

$$= \frac{Kk_N/(NT_N)^{q/2}}{\left(\xi_0 - w_N/(NT_N)^{1/2}\right)^q}$$

By assumption (B2),

$$\frac{w_N}{(NT_N)^{1/2}} \in O(1)\left(\frac{\lambda_N(NT_N)^{-\gamma/2}}{(NT_N)^{(2-\gamma)/2}}\right)^{1/(2-\gamma)} = O(1)\left(\lambda_N/(NT_N)\right)^{1/(2-\gamma)} \subseteq o(1)$$

Furthermore, for any $q \geq 1$, $k_N/(NT_N)^{q/2} \to 0$ by (B3). Hence,

$$P\left(w_N > (NT_N)^{-1/2}\min_{k\in K_N}|\epsilon_N'\mathbf{x}_k + (NT_N)\xi_{Nk}|\right) \to 0.$$

□

References

Arellano, M. (2003). *Panel data econometrics.* Oxford University Press, USA.

Candes, E. and T. Tao (2007). The Dantzig selector: statistical estimation when p is much larger than n. *Annals of Statistics 35*(6), 2313–2351.

Fan, J. and J. Lv (2008). Sure independence screening for ultrahigh dimensional feature space. *Journal of the Royal Statistical Society: Series B (Statistical Methodology) 70*(5), 849–911.

Hoffmann-Jørgensen, J. (1994). Probability with a View Toward Statistics, Volume 1.

Huang, J., J. Horowitz, and S. Ma (2008). Asymptotic properties of bridge estimators in sparse high-dimensional regression models. *Annals of Statistics 36*(2), 587–613.

Hunter, D. and R. Li (2005). Variable selection using MM algorithms. *Annals of statistics 33*(4), 1617.

Knight, K. and W. Fu (2000). Asymptotics for lasso-type estimators. *Annals of Statistics*, 1356–1378.

Meinshausen, N. and B. Yu (2009). Lasso-type recovery of sparse representations for high-dimensional data. *Annals of Statistics 37*(1), 246–270.

Stoyanov, J. (1997). Counterexamples in probability. *Chichester-New York.*

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

Van der Vaart, A. and J. Wellner (1996). *Weak convergence and empirical processes.* Springer Verlag.

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