Optimal Shrinkage Estimation of Fixed Effects in Linear Panel Data Models

Soonwoo Kwon*

November 15, 2020
[Click for latest version.]

Abstract

Shrinkage methods are frequently used to estimate fixed effects. However, the risk properties of existing estimators are fragile to violations of the underlying distributional assumptions. I develop an estimator for the fixed effects that obtains the best possible mean squared error (MSE) within a class of shrinkage estimators. This class includes conventional estimators, and the optimality does not require distributional assumptions. Importantly, the fixed effects are allowed to vary with time and to be serially correlated, and the shrinkage optimally incorporates the underlying correlation structure in this case. In such a context, I also provide a method to forecast fixed effects one period ahead. A simulation study shows that the proposed estimator substantially reduces the MSE relative to conventional methods when the distributional assumptions of the conventional methods are violated, and loses very little when the assumptions are met. Using administrative data on the public schools of New York City, I estimate a teacher value-added model and show that the proposed estimator makes an empirically relevant difference. An optimized R package, `FEShR`, to implement the proposed method is provided.[1]

*Yale University, Department of Economics, `soonwoo.kwon@yale.edu`. I thank Donald Andrews and Timothy Armstrong for continuous support and helpful comments throughout the project. Xiaohong Chen and Yuichi Kitamura provided valuable feedback that improved many parts of the paper. I also thank Joseph Altonji, Ian Ball, Barbara Biasi, John Eric Humphries, Koohyun Kwon, Cormac O'Dea, Vitor Possebom, Nicholas Snashall-Woodhams, Suk Joon Son, Edward Vytlacil and Conor Walsh for helpful comments. The New York City Department of Education generously provided the data used in the empirical section.

[1] See `https://github.com/soonwookwon/FEShR` for details.

1 Introduction

Linear panel data models commonly include fixed effects to allow for unobserved heterogeneity. The fixed effects capture information about heterogeneity that is often empirically relevant, and thus fixed effects themselves are a parameter of interest in a number of studies. Following the work of Abowd et al. (1999), the literature on the analysis of wage differential factors has focused on employer (and employee) fixed effects in a linear panel specification with wages as the outcome. In the literature on teach-valued added (Rockoff, 2004; Rothstein, 2010; Chetty et al., 2014a), student test scores are regressed on student characteristics along with a teacher fixed effect, and this fixed effect is interpreted as a measure of teacher quality. Other examples include neighborhood effects in the study of intergenerational mobility studies (Chetty and Hendren, 2018), judge effects (Frandsen et al., 2019), school effects (Angrist et al., 2017), and hospital effects (Hull, 2020).[2]

Researchers typically estimate a large number of fixed effects, such as the number of firms in a given economy or the number of teachers in a school district. However, the effective sample size available for the estimation of each fixed effect is relatively small: a single employer can hire only so many employees, and a single teacher can teach only so many students. Formally, in the asymptotic experiment, the number of fixed effects often grows to infinity, but the sample size corresponding to each fixed effect remains finite. If the researcher uses the least squares estimator—the coefficient on the dummy variables for the fixed effect units under the corresponding OLS specification—one ends up with a large number of noisy estimates.

To resolve this problem, applied researchers have used estimators that shrink the least squares estimator using an Empirical Bayes (EB) method. Such EB estimators are derived under a hierarchical model. The model assumes that the true fixed effect is drawn from a normal distribution with unknown moments, which I refer to as the hyperparameters. The least squares estimator conditional on the true fixed effect is also assumed to follow a normal distribution, centered at the true fixed effect with known variance.[3] Under this model, the mean of the true fixed effect conditional on the least squares estimator (i.e., the posterior mean) provides a class of shrinkage estimators indexed by the hyperparameters. The hyperparameters determine the

[2]Some of the examples do not strictly fall into a linear panel data setting, but the idea is similar.

[3]In practice, the variance is unknown and a consistent estimator can be plugged in.

degree of shrinkage, where the least squares estimators with lower variances get less shrunk. The unknown hyperparameters are estimated using the marginal distribution of the least squares estimator implied by the hierarchical model, resulting in EB estimators. As the hyperparameters are tuned using the distributional assumptions made in the hierarchical model, the risk properties of EB estimators are inherently sensitive to such assumptions.

I provide an alternative shrinkage estimator with optimality properties that do not rely on such distributional assumptions. Moreover, I allow the fixed effects to vary with time and to be serially correlated within each unit. While it seems natural for the fixed effects do vary with time, allowing such time drift in the fixed effects makes the least squares estimator even noisier, which is possibly one reason such a specification has not been often used.[4] The proposed shrinkage method takes into account the underlying correlation structure, and pools the information across time in a way that minimizes the risk of the estimator. In particular, the EB estimator commonly used under the assumption of time-invariant fixed effects is a special case of the proposed estimator. With the proposed procedure, the data decides whether or not to use this estimator. In this context of time-varying fixed effects, I also provide an optimal forecast method to predict the fixed effect one period ahead.

The derivation of the proposed estimator starts from the same hierarchical model that the EB method employs. However, unlike in the EB approach, the model is used only to restrict the class of estimators to those defined by the posterior mean. Once the class of estimators is narrowed down, no further reference is made to any of the distributional assumptions imposed by the hierarchical model. The hyperparameters are chosen so that the corresponding estimator minimizes an estimate of the MSE, which is the risk criteria I use throughout the paper. I refer to this estimate of the risk as the unbiased risk estimate (URE) and the estimator resulting from choosing the hyperparameters by minizing the URE as the URE estimator.[5]

I show that the URE converges to the true loss in a suitable sense. This convergence between the risk estimate and the true risk translates to the (asymptotic) optimality of the URE estimator among the class of estimators in consideration. This class includes, for example, the conventional EB methods used in the literature. Also,

[4]The effective sample size available for estimation of a unit-time specific fixed effect is roughly $1/T$ of that for estimation of a unit specific fixed effect, where T is the number of time periods for which data is available.

[5]This terminology is adopted from, for example, Xie et al. (2012) and Brown et al. (2018).

while the least squares estimator does not belong to the class of estimators I consider, a simple approximation argument shows that the URE estimator weakly dominates this estimator as well. Hence, the URE estimator improves upon estimators used in the literature.

This optimality holds under only mild regularity conditions, and thus is robust to the distributional assumptions that the EB methods rely on. In particular, the normality assumptions on the true fixed effects and the least squares estimator are not required. The normality of the true fixed effect is usually difficult to justify. For the least squares estimator, normality could be plausible if the sample size available for a given fixed effect is large enough to make a central limit theorem argument. However, this is not typically the case. For example, in datasets used in the teacher value-added literature, there are many teachers that are linked to only around ten students.

Another implicit assumption made by EB methods is that the mean (i.e., the true fixed effect) and variance of the least squares estimator are independent. This rules out the existence of factors that affect both the true fixed effect and the variance of the least squares estimator, which can happen, for example, in the presence of conditional heteroskedasticity. Also, in the teacher value-added model, the variance term of the least squares estimator for the fixed effect of a teacher is inversely proportional to the number of the students the teacher has taught. Hence, if there is any relationship between the number of students in a teacher's class and the teacher's value-added, the EB assumption is violated.[6] Likewise, such assumption is violated in the employee-employer matched data setting if bigger firms pay higher wages. The URE estimator is robust to such violations.

To show the optimality of the URE estimator, I derive new results in a multivariate normal means setting, which has a natural connection with the least squares estimator. The normal means problem is the problem of estimating the mean vectors $\{\theta_j\}_{j=1}^J$ upon observing $y_j \overset{\text{indep}}{\sim} N(\theta_j, \Sigma_j)$ with $y_j \in \mathbf{R}^T$ and $j = 1, \ldots, J$. Under heteroskedasticity, in the sense that Σ_j varies with j, no estimator has been shown to be risk optimal (in a frequentist sense) unless $T = 1$, which has been dealt with by Xie et al. (2012). Allowing for $T > 1$ and general forms of Σ_j, I derive an estimator that obtains the best possible MSE within a certain class of estimators in this model.

[6]For example, EB assumptions are violated if better teachers are assigned more students or class size affects teaching effectiveness.

I emphasize that the application of this result is not restricted to the estimation of fixed effects. In any context with a large number of parameters to be estimated and an (approximately) unbiased estimator of these parameters, my method can be used to reduce the MSE.

Simulation results demonstrate the effectiveness of the URE estimator. While all theoretical results rely on asymptotic arguments, the URE estimator shows desirable risk properties even for moderate sample sizes. Across all scenarios, the MSE of the URE estimator is within 10% of the best possible MSE as long as the number of fixed effects to be estimated is greater than 600. Moreover, the results show that the risk reduction can be substantial compared to the EB methods when the distributional assumptions of the EB methods are not met. For some scenarios, this reduction is as large as 80%. This reduction comes at a relatively small price; even when the EB assumptions are met exactly, the risk of the URE estimator is at most 5% greater than that of the EB estimator.

I use the proposed method to estimate a teacher value-added model using administrative data on the public schools of New York City. The results emphasize that the choice of the estimator makes significant difference in policy-related empirical results, and that it is crucial to allow for the fixed effects to vary with time. I revisit the policy exercise of releasing the bottom 5% of the teachers according to the estimated fixed effects, and find that the composition of the released teachers changes by 25% by using the proposed method rather than the conventional methods, and by 75% by using the proposed forecast method.

Related literature. The literature on the normal means model starting from the seminal papers by Stein (1956) and James and Stein (1961) is abundant. However, (frequentist) risk properties of James-Stein type estimators in heteroskedastic normal means model has been considered only recently by Xie et al. (2012). Xie et al. (2012) consider the problem of estimating the univariate, heteroskedastic normal means model by minizing a risk estimate. Subsequently, Xie et al. (2016), Kong et al. (2017), Kou and Yang (2017), and Brown et al. (2018) use a similar approach in different settings, providing optimal shrinkage estimation methods. My paper provides an optimal shrinkage procedure for a normal means model that has not been considered in the literature, using similar URE methods introduced in such papers. Unlike the previous papers that require tuning at most two scalar hyperparameters, here the hyperparameter includes a $T \times T$ positive semidefinite matrix and possibly

an additional vector of length T. When $T = 1$, the proposed method reduces to the methods of Xie et al. (2012) and Kou and Yang (2017).

Recently, there has been increased interest in EB and shrinkage methods in econometrics. While the present paper does not use an EB approach in its strict sense, the class of estimators I consider is inspired by an EB setting. Also, the URE estimators fall into the category of shrinkage estimators, though it seems that this specific form has not been considered in the literature. Hansen (2016) provides a method to shrink maximum likelihood estimators to subspaces defined by nonlinear constraints and derive risk properties of the resulting estimator. In a related setting, Fessler and Kasy (2019) take an EB approach to effectively incorporate information implied by economic theory. Abadie and Kasy (2019) provide a method of choosing the tuning parameter for regularized estimation problems that gives desirable risk properties. Bonhomme and Weidner (2019) use EB methods to estimate population averages conditional on the given sample, and Liu et al. (2020) use nonparametric EB methods to provide forecasts for the outcome variable in a short panel setting. Armstrong et al. (2020) provide robust confidence intervals for EB estimators.

The literature on teacher value-added (Rockoff, 2004; Kane et al., 2008; Rothstein, 2010; Chetty et al., 2014a; see Koedel et al., 2015 for a recent review on the topic) has fruitfully employed EB shrinkage methods to estimate teacher fixed effects. My method can effectively estimate teacher value-added without resorting to restrictive distributional assumptions. Moreover, the value-added is allowed to vary with time. Chetty et al. (2014a) were the first to allow the teacher value-added to change with time.[7] The analysis by Bitler et al. (2019) suggests that it is important to allow for such time-drifts, and my empirical analysis adds evidence for this potential importance. Gilraine et al. (2020) proposes a nonparametric EB approach to estimate value-added by using the methods by Koenker and Mizera (2014). By taking this nonparametric EB approach, the normality assumption on the true fixed effect can be relaxed and a broader class of estimators is considered. This approach is complementary to the URE methods introduced here, which I discuss in more detail later.

Outline. Section 2 describes the linear panel data model and shows how the estimation of fixed effects is asymptotically equivalent to estimating the mean vector in a

[7]The estimator used by Chetty et al. (2014a) can be considered as a special case of the predictors introduced in Section 5.2, under the assumption of equal class sizes.

normal means problem. Section 3 defines the URE and the URE estimators obtained by minimizing the URE. Section 4 establishes the optimality of the URE estimators. Section 5 provides two methods to summarize the time trajectory of fixed effects. Section 6 demonstrates the efficacy of the URE estimators via a simulation study. In Section 7, I estimate a teacher value-added model using the proposed estimator. Proof of the main theorems are given in Appendix A.

Notation. Let $\{W_{ijt}\}$ be a real (either random or nonrandom) sequence, where the indices take values $j = 1, \dots J$, $t = 1, \dots, T$, and $i = 1, \dots, n_{jt}$ for any (j,t)-pair. The following vectors are defined by concatenating the sequence at different levels: $W_{jt} = (W_{1jt}, \dots, W_{n_{jt}jt})'$, $W_j = (W_{j1}', \dots, W_{jT}')'$, and $W = (W_1', \dots, W_J')'$. The (j,t)-level average is written $\overline{W}_{jt} = n_{jt}^{-1} \sum_{i=1}^{n_{jt}} W_{ijt}$ and the demeaned version of the sequence is defined $\widetilde{W}_{ijt} = W_{ijt} - \overline{W}_{jt}$.

Let $\|\cdot\|$ denote the Euclidean norm for both vectors and matrices (i.e., the Frobenius norm in the latter case). For any matrix A, $(A)_{ij}$ denotes its (i,j) entry and $\sigma_k(A)$ its kth largest singular value. By definition, $\sigma_1(A)$ is the operator norm of the matrix A. Likewise, $\lambda_k(A)$ denotes the kth largest eigenvalue of a square matrix A, so that $\sigma_k(A) = \lambda_k(A)$ for all k when A is positive semidefinite. Let $\kappa(A) = \sigma_1(A)/\sigma_k(A)$ be the condition number of any $k \times k$ matrix A. For two real symmetric matrices A and B, I write $A \geq B$ to denote that $A - B$ is positive semidefinite, with strict inequality meaning that $A - B$ is positive definite. For any $d \in \mathbf{R}^k$, let $\mathrm{diag}(d)$ denote the $k \times k$ diagonal matrix with diagonal elements d. The set of positive semidefinite $k \times k$ matrices is denoted by $\mathcal{S}_k^+$, and the $k \times k$ identity matrix is denoted by I_k.

2 Fixed effects and the normal means model

2.1 The linear panel data model

I consider the following linear panel data model,

$$Y_{ijt} = X_{ijt}'\beta + \alpha_{jt} + \varepsilon_{ijt}, \tag{1}$$

where $t = 1, \dots, T$, $j = 1, \dots J$, and for each (j,t) $i = 1, \dots, n_{jt}$. Here, $\{(Y_{ijt}, X_{ijt}')\}$ denotes the observed data, ε_{ijt} the idiosyncratic shock, and α_{jt} the time-varying fixed effect which is the object of interest. Typically, i is some individual level, j group

level, and t a time dimension. The time-varying fixed effect for j, $\alpha_j := (\alpha_{j1}, \dots \alpha_{jT})'$ is assumed to be independent across j but is allowed to be serially correlated. For the idiosyncratic error terms, assume $\overline{\varepsilon}_j = (\overline{\varepsilon}_{j1}, \dots, \overline{\varepsilon}_{jT})'$ is independent across j and with α_j, and denote its variance matrix by Σ_j. The variance matrix Σ_j is assumed known, and in practice a consistent estimator is plugged in under suitable conditions, which does not affect the asymptotic properties.

Remark 2.1 ($T = 1$)**.** For a special case, consider $T = 1$ and omit the time subscripts. Then, interpreting i as "time," the model simplifies to

$$Y_{ij} = X_{ij}'\beta + \alpha_j + \varepsilon_{ij},$$

which is the canonical panel data model indexed by individual and time. Hence, the setting in consideration includes this canonical panel model as a special case.

Example 2.1 (Teacher value-added)**.** In the teacher value-added model, j corresponds to teacher, t to school year, and i to a student assigned to teacher j in school year t. The outcome variable Y_{ijt} is a measure of student achievement (e.g., test score) and X_{ijt} is a vector of student characteristics. The fixed effect α_{jt} is the value-added of teacher j in year t, and is considered a measure of teacher quality. I use this model as a running example throughout the paper. For this example, I further assume that ε_{ijt} is i.i.d across all i, j, and t with variance σ_ε^2 so that $\Sigma_j = \sigma_\varepsilon^2 \text{diag}(1/n_{j1}, \dots, 1/n_{jT})$. See Koedel et al. (2015) for a recent review on the literature, including a discussion on specification issues.

There are numerous other examples that fall into this framework. In the widely used wage determination model first introduced by Abowd et al. (1999), j corresponds to employer, t to year, and i to employee. In this model, an employee fixed effect is typically included as well. The outcome variable is log wages and the employer fixed effect captures the wage differential due to the employer. In a different, but related setting used in the analysis of neighborhood effects on future economic outcomes by Chetty and Hendren (2018), j corresponds to either commuting zone or county. Here, the outcome variable is some measure of future economic outcome, and the fixed effect captures the effect of the neighborhood one resides in during her childhood to future economic outcome.

All asymptotic arguments are as $J \to \infty$ with T and n_{jt} fixed. This captures the common situation where the number of fixed effects to be estimated is large ($J \to \infty$),

with observations for each fixed effect unit being relatively small (n_{jt} remains fixed). In the teacher value-added model, this corresponds to the asymptotic experiment where the number of teachers grows to infinity, with the year dimensions and students per teacher fixed. I assume that a consistent estimator $\widehat{\beta}$ of β is readily available, which is easy to obtain under standard assumptions such as strict exogeneity (see, for example, Wooldridge (2010) for a textbook level discussion).

2.2 Connection with the normal means model

Let $\widehat{\alpha}_{jt}$ denote the least squares estimator for the fixed effects, which can be obtained by taking the coefficients of the (j,t)-level dummy variable[8] in the corresponding OLS specification:

$$\widehat{\alpha}_{jt} := \overline{Y}_{jt} - \overline{X}'_{jt}\widehat{\beta} = \overline{X}'_{jt}(\beta - \widehat{\beta}) + \alpha_{jt} + \overline{\varepsilon}_{jt} = \alpha_{jt} + \overline{\varepsilon}_{jt} + O_p(J^{-1/2}).$$

To see the connection between this estimator and the normal means model, note that, for any j, we have $\widehat{\alpha}_j \to_d \alpha_j + \overline{\varepsilon}_j$. Further assuming that $\overline{\varepsilon}_j$ follows a normal distribution (with variance matrix Σ_j), we have $(\alpha_j + \overline{\varepsilon}_j)\,|\alpha_j \sim N(\alpha_j, \Sigma_j)$ so that

$$\widehat{\alpha}_j|\alpha_j \sim N(\alpha_j, \Sigma_j), \tag{2}$$

approximately. I note that under a mild boundedness condition on X_{ijt} that ensures $\sup_j\|\overline{X}_j\| = O_p(1)$, such convergence is in fact uniform over j. This is because, by Cauchy-Schwarz, we have

$$\mathbf{P}\Big(\sup_j\|\widehat{\alpha}_j - \alpha_j - \overline{\varepsilon}_j\| > \varepsilon\Big) \le \mathbf{P}\Big(\|\widehat{\beta} - \beta\|\sup_j\|\overline{X}_j\| > \varepsilon\Big).$$

This shows the natural connection between the estimation of fixed effects in a linear panel data model and the estimation of the means in a (multivariate) normal means model. Note that even if ε_{ijt} is homoskedastic so that $\mathrm{var}(\varepsilon_{ijt}) = \sigma^2_\varepsilon$, the variance term of the aggregate term is $\mathrm{var}(\overline{\varepsilon}_{jt}) = \sigma^2_\varepsilon/n_{jt}$ so that heteroskedasticity is present due to the different cell sizes, n_{jt}. Under such heteroskedasticity, it has been noted by Xie et al. (2012) that EB methods do not enjoy the many risk properties

[8]When JT is very large, running an OLS regression with dummy variables is computationally inefficient, and thus standard statistics software that deal with large number of fixed effects do not estimate the fixed effects this way. However, I use this explanation due to its intuitiveness.

that they do under a homoskedastic setting, where the EB estimator is essentially the same as the James-Stein estimator.

Due to this connection, I now consider the problem of estimating the mean vectors under a multivariate normal means model. The problem is to estimate $\theta = (\theta_1', \ldots, \theta_J')'$ after observing the data $\{y_j\}_{j=1}^J$ where this is generated according to

$$y_j|\theta_j \overset{\text{indep}}{\sim} N(\theta_j, \Sigma_j), \tag{3}$$

for $j = 1, \ldots, J$ with $y_j, \theta_j \in \mathbf{R}^T$. The variance matrix $\Sigma_j \in \mathcal{S}_T^+$ is assumed to be known. This has the exact same structure as the asymptotic approximation of the least squares estimator as seen in (2), with the data y_j being the least squares estimator and θ_j the true fixed effect.

3 URE estimators

3.1 Class of shrinkage estimators

The URE estimators will be shown to be optimal within a class of shrinkage estimators, that nests commonly used estimators. The class of estimators corresponds to the Bayes estimators under a hierarchical model. The hierarchical model postulates a Gaussian model on the true mean vector (true fixed effect), which I refer to as a second level model, on top of the Gaussian model on the data (least squares estimator). I emphasize that both normality assumptions are used only to derive the class of estimators.

Consider the second level model

$$\theta_j \overset{\text{i.i.d.}}{\sim} N(\mu, \Lambda), \tag{4}$$

where the location vector $\mu \in \mathbf{R}^T$ and the variance matrix $\Lambda \in \mathcal{S}_T^+$ are unknown hyperparameters to be tuned. The restriction one imposes on Λ incorporates the prior knowledge on the underlying covariance structure. I denote by $\mathcal{L} \subset \mathcal{S}_T^+$ the set that reflects this prior knowledge. As a practical matter, this reduces the dimension of the optimization problem that one must solve to obtain the URE estimators. For example, when θ_j is believed to be covariance stationary, one can take $\mathcal{L}$ as the set of positive semidefinite Toeplitz matrices. This reduces the dimension of Λ to T from

$T(T+1)/2$ when Λ is left unrestricted.

The second level model (4), together with the normal means model (3), gives a hierarchical Bayes model. By standard calculations, the Bayes estimator of θ_j under this model is given as

$$\begin{aligned}\widehat{\theta}_j(\mu,\Lambda) := E[\theta_j|y] =& \mu + \Lambda(\Lambda+\Sigma_j)^{-1}(y_j-\mu) \\ =& \left(I_T - \Lambda(\Lambda+\Sigma_j)^{-1}\right)\mu + \Lambda(\Lambda+\Sigma_j)^{-1}y_j\end{aligned}$$

Analogous to the univariate case, I refer to $\Lambda(\Lambda+\Sigma_j)^{-1}$ as the shrinkage matrix. It can be shown that the largest singular value of the shrinkage matrix is less than 1, justifying the term "shrinkage." As in the univariate case, noisier observations get more severely shrunken in the sense that $\widetilde{\Sigma}_j \leq \Sigma_j$ implies $\sigma_t(\Lambda(\Lambda+\Sigma_j)^{-1}) \leq \sigma_t(\Lambda(\Lambda+\widetilde{\Sigma}_j)^{-1})$ for all $t = 1,\ldots,T$. The shrinkage occurs towards the mean of the second level model, μ. In the literature, this is frequently set to 0 after demeaning the least squares estimators, but this is not necessarily the best choice for URE estimators despite the demeaning. I come back to this issue later.

Example 3.1 (Independent case)**.** If $\Lambda = \lambda I_T$ and $\Sigma_j = \mathrm{diag}(\sigma_{j1}^2,\ldots,\sigma_{jT}^2)$, then the tth component of $\widehat{\theta}_j(\mu,\Lambda)$ is given as

$$\left(1 - \frac{\lambda}{\lambda+\sigma_{jt}^2}\right)\mu_t + \frac{\lambda}{\lambda+\sigma_{jt}^2}y_{jt},$$

which is the form of shrinkage estimators[9] used in the literature (Rockoff, 2004; Chetty and Hendren, 2018; etc.), with a specific choice of λ and μ. Moreover, when $\mu = 0$ and σ_{jt}^2 does not vary with j, an appropriate choice of λ in fact gives the James-Stein estimator.

Example 3.2 ($T = 2$)**.** To gain some intuition on how the correlation terms of the second level model affect the form of shrinkage, consider the case where $T = 2$ with $\Sigma_j = \left(\begin{smallmatrix}\sigma_{j1}^2 & 0\\ 0 & \sigma_{j2}^2\end{smallmatrix}\right)$ and $\Lambda = \left(\begin{smallmatrix}\lambda_1^2 & \lambda_1\lambda_2\rho\\ \lambda_1\lambda_2\rho & \lambda_2^2\end{smallmatrix}\right)$, and μ is set to 0. Write $y_j = (y_{j1}, y_{j2})'$ and

[9]More precisely, the estimators used in the literature take this form without the time-varying component, and thus everything is aggregated at the j level so that the subscript t disappears.

$\theta_j = (\theta_{j1}, \theta_{j2})'$. The estimator $\widehat{\theta}_j(0, \Lambda)$ can be explicitly calculated as

$$\begin{aligned}
&\Lambda(\Lambda + \Sigma_j)^{-1} y_j \\
=&\frac{1}{(\lambda_1^2 + \sigma_{j1}^2)(\lambda_2^2 + \sigma_{j2}^2) - \lambda_1^2\lambda_2^2\rho^2} \begin{pmatrix} \lambda_1^2(\lambda_2^2 + \sigma_{j2}^2) - \lambda_1^2\lambda_2^2\rho^2 & \lambda_1\lambda_2\rho\sigma_{j1}^2 \\ \lambda_1\lambda_2\rho\sigma_{j2}^2 & \lambda_2^2(\lambda_1^2 + \sigma_{j1}^2) - \lambda_1^2\lambda_2^2\rho^2 \end{pmatrix} y_j.
\end{aligned}$$

Hence, the estimator for θ_{j1} is

$$\frac{\lambda_1^2(\lambda_2^2 + \sigma_{j2}^2) - \lambda_1^2\lambda_2^2\rho^2}{(\lambda_1^2 + \sigma_{j1}^2)(\lambda_2^2 + \sigma_{j2}^2) - \lambda_1^2\lambda_2^2\rho^2} y_{j1} + \frac{\lambda_1\lambda_2\rho\sigma_{j1}^2}{(\lambda_1^2 + \sigma_{j1}^2)(\lambda_2^2 + \sigma_{j2}^2) - \lambda_1^2\lambda_2^2\rho^2} y_{j2}.$$

The coefficient on y_{j1} is positive and decreases in $|\rho|$; one uses less of the information from y_{j1} as the information from y_{j2} increases. The absolute value of the coefficient on y_{j2} increases with $|\rho|$, and thus using more of y_{j2} when there is more correlation. Both coefficients are smaller than 1 in magnitude. Furthermore, the Euclidean norm of the coefficients (as a vector in $\mathbf{R}^2$) is smaller than 1, showing that the estimator is indeed a shrinkage estimator.

Example 3.3 (Perfect correlation)**.** Let $\mathbf{1}_T$ denote the T-vector with elements all equal to 1. Consider the case where $\Lambda = \lambda \mathbf{1}_T \mathbf{1}_T'$, which is essentially assuming θ_{jt} is equal across t. Let $\Sigma_j = \sigma^2 \text{diag}(1/n_{j1}, \ldots, 1/n_{jT})$, which corresponds to the linear panel data model with idiosyncratic errors that are homoskedastic and uncorrelated across time. Denote the teacher-level sample size by $n_j = \sum_{t=1}^T n_{jt}$. In this context, the estimator $\widehat{\theta}(0, \Lambda)$ is given as

$$\begin{aligned}
&\lambda \mathbf{1}\mathbf{1}'(\Sigma_j^{-1} - \Sigma_j^{-1}\mathbf{1}(1/\lambda + \mathbf{1}'\Sigma_j^{-1}\mathbf{1})^{-1}\mathbf{1}'\Sigma_j^{-1}) y_j \\
=&\lambda \mathbf{1}\mathbf{1}' \left(\Sigma_j^{-1} - \frac{\Sigma_j^{-1}\mathbf{1}\mathbf{1}'\Sigma_j^{-1}}{1/\lambda + \sum_{t=1}^T n_{jt}/\sigma^2} \right) y_j \\
=&\mathbf{1} \frac{\lambda}{\sigma^2/n_j + \lambda} \left(\frac{1}{n_j} \sum_{t=1}^T n_{jt} y_{jt} \right),
\end{aligned}$$

where the first equality follows by the Woodbury matrix identity.

Note that $\frac{1}{n_j}\sum_{t=1}^T n_{jt} y_{jt}$ is a weighted mean of the least squares estimators of teacher j, and thus is essentially the least squares estimator for the teacher level fixed effect without time drift. This is exactly the estimator used in the majority of the

teacher value-added literature with an appropriate choice of λ.[10] In the proposed method, whether to use this estimator or not is determined in a data driven way, depending on whether this choice of Λ indeed minimizes the risk.

To better understand the operation the shrinkage matrix performs to the data, let UDU' denote the spectral decomposition of $\Sigma_j^{-1/2}\Lambda\Sigma_j^{-1/2}$, the signal-to-noise ratio matrix, with $D = \text{diag}(d_1, \ldots, d_T)$. For simplicity, consider the case with $\mu = 0$. It can be shown that

$$\widehat{\theta}_j(0, \Lambda) = \Lambda(\Lambda + \Sigma_j)^{-1} y_j = \Sigma_j^{1/2} UD(I_T + D)^{-1} U' \Sigma_j^{-1/2} y_j.$$

Here, the last $\Sigma_j^{-1/2}$ term simply standardizes the data, y_j, and the first $\Sigma_j^{1/2}$ term brings it data back to its original scale and direction. The $UD(I_T + D)^{-1}U'$ term captures the direction and degree of shrinkage. Specifically, U' rotates the standardized data $\Sigma_j^{-1/2} y_j$ in the direction of the eigenvectors of the signal-to-noise ratio matrix, $D(I_T + D)^{-1}$ shrinks this rotated data according to the eigenvalues of the eigenvalues of signal-to-noise ratio matrix, and finally U rotates the data back to its original axes. Note that $D(I_T + D)^{-1}$ is indeed a shrinkage term because $D(I_T + D)^{-1} = \text{diag}(d_1/(1 + d_1), \ldots, d_T/(1 + d_T))$ and $d_t/(1 + d_t) \in [0, 1)$ for all $t \leq T$. Since both U and $D(I_T + D)^{-1}$ depend on Λ, the choice of Λ determines the direction and magnitude of shrinkage. This is in contrast with the univariate case, where tuning λ just determines the magnititude of shrinkage.

One class of shrinkage estimators I consider is $\widehat{\theta}(\mu, \Lambda) := (\theta_1(\mu, \Lambda)', \ldots, \widehat{\theta}_j(\mu, \Lambda))'$, indexed by the hyperparameters (μ, Λ). This class of estimators includes the conventional EB methods, where one proceeds by substituting an "estimator" $(\widehat{\mu}^{\text{EB}}, \widehat{\Lambda}^{\text{EB}})$ for (μ, Λ). This is done by using the marginal distribution of the data implied by the hierarchical model, $y_j \overset{\text{indep}}{\sim} N(\mu, \Lambda + \Sigma_j)$, either by maximum likelihood or the method of moments. I denote the EB maximum likelihood estimator (EBMLE) by $\widehat{\theta}^{\text{EBMLE}} = \widehat{\theta}(\widehat{\mu}^{\text{EBMLE}}, \widehat{\Lambda}^{\text{EBMLE}})$ where $(\widehat{\mu}^{\text{EBMLE}}, \widetilde{\Lambda}^{\text{EBMLE}})$ is tuned by maximizing this marginal likelihood. I also consider another larger class of estimators where each y_j is shrunk toward a different location for each j, where this location depends on some auxiliary data. This extension is useful when one has additional covariates that can explain y_j. In the teacher value-added model, this is the case when teacher level

[10] Guarino et al. (2015) provides a review (and evaluation) of the shrinkage methods used in the literature.

covariates are available.

The risk of an estimator $\widehat{\theta}$ of θ is measured by the compound MSE,

$$R(\theta, \widehat{\theta}) = \tfrac{1}{J}\, \mathbf{E}_\theta(\widehat{\theta} - \theta)'(\widehat{\theta} - \theta),$$

where the term "compound" highlights the fact that risks across the independent experiments are aggregated.[11] While I consider only the unweighted case for expositional reasons, all results go through under the weighted compound MSE as long as the weights satisfy a mild boundedness condition, as shown in Appendix C. In Section 5.1, I consider a special case of such weights that has admit an intuitive interpretation. The expectation $\mathbf{E}_\theta$ is evaluated at θ, and the subscript θ is omitted unless ambiguous otherwise.

3.2 Risk estimate and URE estimators

Given the risk criterion, an optimal yet infeasible way to tune the hyperparameters is by minimizing the risk. Of course, this is infeasible because the risk function depends on the true mean vectors, which are unknown. I take an approach of estimating the risk, using Stein's unbiased risk estimate (SURE), and choosing the hyperparameters by minimizing this risk estimate. The idea of minimizing SURE to choose tuning parameters has been around since at least Li (1985). The approach has been taken recently in, for example, Xie et al. (2012, 2016), Kou and Yang (2017), Brown et al. (2018), and Abadie and Kasy (2019).

To obtain a risk estimate, consider the following unbiased risk estimate which is the SURE formula applied to the estimator $\widehat{\theta}(\mu, \Lambda)$,

$$\begin{aligned}&\mathrm{URE}(\mu, \Lambda)\\ :=&\frac{1}{J}\sum_{j=1}^{J}\left(\mathrm{tr}(\Sigma_j) - 2\,\mathrm{tr}((\Lambda+\Sigma_j)^{-1}\Sigma_j^2) + (y_j-\mu)'[(\Lambda+\Sigma_j)^{-1}\Sigma_j^2(\Lambda+\Sigma_j)^{-1}](y_j-\mu)\right),\end{aligned}$$

where I define the summand as $\mathrm{URE}_j(\mu, \Lambda)$. It is easy to show that $\mathbf{E}\,\mathrm{URE}(\mu, \Lambda) = R(\widehat{\theta}(\mu, \Lambda), \theta)$, and thus $\mathrm{URE}(\mu, \Lambda)$ is indeed an unbiased estimator of the true risk. While SURE applies to any estimator that takes the form $y_j + g(y_j)$ with g being

[11] This term originates from what Robbins (1951) referred to as the "compound statistical decision problem" in the context of a simple normal means problem.

weakly differentiable, normality of y_j is crucial for the unbiasedness to hold for all such estimators. However, the function g corresponding to $\widehat{\theta}(\mu, \Lambda)$ is in fact a simple affine function, so that the unbiasedness can be established by a simple bias-variance expansion. Accordingly, the unbiasedness of $\text{URE}(\mu, \Lambda)$ holds without any distributional assumptions on y_j, apart from the existence of second moments.

Clearly, unbiasedness itself will not guarantee that the estimator obtained by minimizing the risk estimate has good risk properties. In the next section, I show that $\text{URE}(\mu, \Lambda)$ is in fact uniformly close to the true risk, in the sense that minimizing this risk estimate is as good as minimizing the true loss, asymptotically.

I propose three shrinkage estimators that are closely related but differ in the location to which they shrink the data. The estimators are introduced in increasing degrees of freedom. All three estimators are obtained by minimizing a corresponding URE, and thus I refer to such estimators as URE estimators.

Grand mean. The first estimator, which is the simplest, takes $\mu = \overline{y}_J := \frac{1}{J}\sum_{j=1}^{J} y_j$ and thus shrinks the data toward the grand mean. The grand mean is an intuitive location to shrink to, and by fixing a value for μ this method effectively decreases the dimension of the hyperparameters. In the context of fixed effects, if the least squares estimators are demeaned for each time period, it follows that $\overline{y}_J = 0$. Hence, the estimator shrinks the data toward the origin. Most shrinkage estimators used in the teacher value-added literature shrink the least squares estimator toward the origin. Formally, this estimator is defined as $\widehat{\theta}^{\text{URE,m}} := \widehat{\theta}(\overline{y}_J, \widehat{\Lambda}^{\text{URE,m}})$, where

$$\widehat{\Lambda}^{\text{URE,m}} := \operatorname*{arg\,min}_{\Lambda \in \mathcal{L}} \text{URE}(\overline{y}_J, \Lambda).$$

General location. The second estimator leaves μ (almost) unrestricted and chooses the location by minimizing the URE. For theoretical reasons, it is not possible to allow for any $\mu \in \mathbf{R}^T$ as the centering location. The hyperparameter space for μ must be restricted so that a certain boundedness property holds. Following a similar idea used by Brown et al. (2018), I restrict μ to lie in

$$\mathcal{M}_J := \{\mu \in \mathbf{R} : |\mu_t| \leq q_{1-\tau}(\{|y_{jt}|\}_{j=1}^{J}) \text{ for } t = 1, \ldots, T\},$$

where $q_{1-\tau}(\{|y_{jt}|\}_{j=1}^{J})$ denotes the $1-\tau$ sample quantile of $\{|y_{jt}|\}_{j=1}^{J}$.[12] I recommend

[12] See, for example, Chapter 21 of van der Vaart (1998) for a formal definition.

choosing a small τ, such as $\tau = .01$. This restricts the centering term, component-wise, to be somewhere smaller than the 99 percentile of the data in terms of magnitude. I argue that this restriction is reasonable, because it seems rather hard to justify shrinking the data toward a point where there are almost no observations. In fact, this constraint is never binding in any of the simulation iterations reported in Section 6. This URE estimator that shrinks towards a general location, $\widehat{\theta}^{\text{URE},g}$, is defined as $\widehat{\theta}^{\text{URE},g} = \widehat{\theta}(\widehat{\mu}^{\text{URE}}, \widehat{\Lambda}^{\text{URE},g})$, where

$$(\widehat{\mu}^{\text{URE}}, \widehat{\Lambda}^{\text{URE},g}) := \underset{\mu \in \mathcal{M}_J, \Lambda \in \mathcal{L}}{\arg\min}\, \text{URE}(\mu, \Lambda).$$

Linear combination of covariates. The last estimator can be used in the presence of additional data, $Z_{jt} \in \mathbf{R}^k$ that is thought to explain θ_{jt}. In the linear panel data model, these are exactly the covariates that could not be included as explanatory variables because of the (j,t)-level fixed effects. Write $Z_j = (Z_{j1}, \ldots, Z_{jT})'$. I consider the estimator that shrinks the data toward $Z_j\gamma$,

$$\widehat{\theta}_j^{\text{cov}}(\gamma, \Lambda) := \left(I_T - \Lambda(\Lambda + \Sigma_j)^{-1}\right) Z_j\gamma + \Lambda(\Lambda + \Sigma_j)^{-1} y_j,$$

where now γ and Λ are hyperparameters to be tuned.[13] I denote by $\widehat{\theta}^{\text{cov}}(\gamma, \Lambda)$ the JT vector obtained by concatenating $\widehat{\theta}_j^{\text{cov}}(\gamma, \Lambda)$ for $j \leq J$. Define $\text{URE}_j^{\text{cov}}(\gamma, \Lambda) = \text{URE}_j(Z_j\gamma, \Lambda)$ and the compound risk estimate $\text{URE}^{\text{cov}}(\gamma, \Lambda) = \frac{1}{J}\sum_{j=1}^J \text{URE}_j^{\text{cov}}(\gamma, \Lambda)$. Then, the estimator is defined as $\widehat{\theta}^{\text{URE,cov}} = \widehat{\theta}^{\text{cov}}(\widehat{\gamma}^{\text{URE}}, \widehat{\Lambda}^{\text{URE,cov}})$, where

$$(\widehat{\gamma}^{\text{URE}}, \widehat{\Lambda}^{\text{URE,cov}}) := \underset{\gamma \in \Gamma_J, \Lambda \in \mathcal{L}}{\arg\min}\, \text{URE}^{\text{cov}}(\gamma, \Lambda).$$

Again, Γ_J is a hyperparemeter set that incorporates restrictions to ensure that URE approximates the true loss well:

$$\Gamma_J := \{\gamma \in \mathbf{R}^k : \|\gamma\| \leq B\|\widehat{\gamma}^{\text{OLS}}\|\},$$

where B is a large constant that does not depend on J, and $\widehat{\gamma}^{\text{OLS}}$ is the pooled OLS estimator obtained by regressing y_j on X_j, i.e., $\widehat{\gamma}^{\text{OLS}} = (\sum_{j=1}^J Z_j'Z_j)^{-1}\sum_{j=1}^J Z_j'y_j$. The idea is to include the intuitive OLS, and potentially coefficients with much larger

[13] This estimator is the Bayes estimator under a second level model where the θ_j is normally distributed with mean $Z_j\gamma$, i.e., $\theta_j|Z_j \sim N(Z_j\gamma, \Lambda)$.

magnitude as well. In simulations, I use $B = 10^3$ and this constraint never binds.

Example 3.4 (Teacher value-added)**.** In teacher value-added, teacher (or teacher-year) level covariates are frequently available. Such covariates cannot be used in the initial regression due to the inclusion of the teacher fixed effects. However, one can use such covariates to improve the precision of the teacher fixed-effect estimates. Frequently available teacher level covariates include, for example, gender, tenure, and union status of a teacher. Asymptotically, the inclusion of such covariates are guaranteed to improve the MSE. Furthermore, this improvement does not require that the true fixed effects are related with the covariates in a linear fashion. These last two points are made more clear in the next section.

While I only consider the simple case of shrinking toward a linear combination of the covariates where the linear combination is defined by the same coefficient γ for all time periods, the estimator can be extended in a straightforward manner to allow different coefficients $\gamma_t \in \mathbf{R}^k$ for each time period t. In this case, y_j is shrunk toward $(Z_{j1}'\gamma_1, \ldots, Z_{jT}'\gamma_T)'$. The optimality property to be shown in Section 4 can also be extended to this case with only minor modifications.

Another interesting extension of this estimator is to shrink toward a more general function of the covariates.[14] That is, for a function $m : \mathbf{R}^k \to \mathbf{R}$, one can consider shrinking to $(m(Z_{j1}), \ldots, m(Z_{jT}))'$. The linear case corresponds to the choice $m(z) = z'\gamma$. As long as the hyperparameter space for m is totally bounded and satisfies certain regularity conditions, the resulting URE estimator can be shown to optimal in this more flexible class of estimators as well. In practice, one chooses the hyperparameter by minimizing the URE over a sieve space that converges to the hyperparameter space as $J \to \infty$.

Remark 3.1 (Choice of the estimator)**.** Under some conditions, the three classes corresponding to $\widehat{\theta}^{\text{URE,m}}$, $\widehat{\theta}^{\text{URE,g}}$, and $\widehat{\theta}^{\text{URE,cov}}$ are nested. While $\overline{y}_J$ does not necessarily lie in $\mathcal{M}_J$, mild regularity conditions on the data ensure that this happens with probability approaching 1. Also, for an appropriate choice of the constant that defines Γ_J and including time dummies as covariates, the class of estimators that shrink toward a general location is nested by those that shrink toward $Z_j\gamma$. Hence, as $J \to \infty$, $\widehat{\theta}^{\text{URE,cov}}$ is guaranteed to have the smallest risk among the three, according to the

[14] Ignatiadis and Wager (2020) consider an estimator of the same form for the case where $T = 1$, but with a different focus.

optimality result given in the following section. However, this does not guarantee that this is the case in finite samples, and this estimator requires additional data. As a rule of thumb, I recommend using $\widehat{\theta}^{\mathrm{URE,cov}}$ if covariates are available, $\widehat{\theta}^{\mathrm{URE,g}}$ if covariates are unavailable and one has at least a moderate sample size (simulation results imply $J > 200$ is enough for $T = 4$), and $\widehat{\theta}^{\mathrm{URE,m}}$ otherwise.

3.3 Computation

The URE estimators involve solving a minimization problem over $\Lambda \in \mathcal{L}$, along with possibly an additional hyperparameter that governs the centering term. For concreteness, I consider the estimator $\widehat{\theta}^{\mathrm{URE,g}}$ and take $\mathcal{L} = \mathcal{S}_T^+$.

The minimization problem that must be solved is

$$\inf_{\Lambda \in \mathcal{S}_T^+, \mu \in \mathcal{M}_J} \mathrm{URE}(\mu, \Lambda).$$

For a fixed Λ, minimization with respect to μ is a quadratic programming program with bound constraints, which is a well understood problem with a number of efficient algorithms readily available. Hence, to utilize this quadratic structure with respect to μ, I profile out μ by solving this quadratic programming problem. Writing $\mu^*(\Lambda) := \inf_\mu \mathrm{URE}(\mu, \Lambda)$, the problem is to now solve

$$\inf_{\Lambda \in \mathcal{S}_T^+} \mathrm{URE}(\mu^*(\Lambda), \Lambda).$$

This is a nonconvex optimization problem with nonlinear constraints, where the nonlinearity of the constraints is due to the restriction that Λ is positive semidefinite. I transform this to a unconstrained problem by using the Cholesky decomposition by defining $f(L) := \mathrm{URE}(\mu^*(LL'), LL')$ and minimizing f over all lower triangle matrices L (i.e., over $\mathbf{R}^{T(T+1)/2}$).[15] Using Quasi-Newton methods such as the BFGS algorithm works well on this transformed problem, finding the minimum within reasonable time without being sensitive to the initial point.

Each evaluation of the objective function involves calculating the inverse of $(\Lambda + \Sigma_j)^{-1}$ for all j. That is, each evaluation involves inverting a $T \times T$ matrix J times,

[15] I note that this is not a common approach when optimizing over the positive definite cone, $\mathcal{S}_T^+$, possibly due to the fact that such transformation makes the problem "more nonlinear." Nonetheless, this approach works very well for the current problem.

which is unavoidable.[16] Since even the state-of-the-art algorithms have computational complexity around $O(T^{2.4})$ for inverting a $T \times T$ matrix, the computation burden increases quickly with T (and with J, of course, but to a much lesser extent). Nonetheless, the computation is not an issue for moderately large T. In the empirical example with around $J = 1,200$ and $T = 6$, calculating $\widehat{\theta}^{\text{URE,g}}$ takes around 100 seconds on a single core using the companion R package.

4 Optimality of the URE estimators

I show that the URE estimators defined in Section 3.2 asymptotically achieve the smallest possible asymptotic MSE among all estimators in the corresponding class. In particular, this shows that the URE estimators dominate the EB methods, which has been widely used in applied work. The main step in establishing such optimality is to show that the corresponding UREs are uniformly close to the true risk. Since we use an unbiased estimate of risk, this more or less boils down to a uniform law of large numbers (ULLN) argument. I first establish a simple high-level result for a generic URE estimator, and verify that the conditions for this high-level result holds for each of the estimators, under appropriate lower level conditions.

4.1 A generic result

Let $\psi \in \Psi$ denote a generic hyperparameter to be tuned, which can be Λ, (μ, Λ), or (γ, Λ) depending on the choice of the estimator, and let $\widehat{\theta}(\psi)$, indexed by ψ, denote the shrinkage estimators in consideration. The hyperparameter space Ψ is allowed to depend on the observations and to vary with J, as is the case for $\widehat{\theta}^{\text{URE,g}}$ and $\widehat{\theta}^{\text{URE,cov}}$. A generic URE estimator is defined as $\widehat{\theta}^{\text{URE}} = \widehat{\theta}(\widehat{\psi}^{\text{URE}})$, where $\widehat{\psi}^{\text{URE}}$ minimizes $\text{URE}(\psi)$. As a performance benchmark, consider the oracle loss "estimator," $\widetilde{\theta}^{\text{OL}} = \widehat{\theta}(\widetilde{\psi}^{\text{OL}})$, where

$$\widetilde{\psi}^{\text{OL}} = \underset{\psi \in \Psi}{\arg\min}\ \ell(\theta, \widehat{\theta}(\psi)).$$

Note that $\widetilde{\theta}^{\text{OL}}$ is not a feasible estimator because it depends on the true θ. However, it serves as a useful benchmark because it minimizes true loss, for any realized θ, and thus no estimator can have strictly smaller loss, and risk, than $\widetilde{\theta}^{\text{OL}}$. Hence, I refer to

[16] The companion R package, `FEShR`, efficiently implements all matrix inversions and loops in C++.

$R(\theta, \widehat{\theta}(\widetilde{\psi}^{\mathrm{OL}}))$ as the oracle risk.

The following simple lemma shows that if $\mathrm{URE}(\psi)$ is uniformly close to the true loss in L^1, then the URE estimator has asymptotic risk as good as the oracle.

Lemma 4.1. *Suppose* $\sup_{\psi \in \Psi} |\mathrm{URE}(\psi) - \ell(\theta, \widehat{\theta}(\psi))| \xrightarrow{L^1} 0$. *Then,*

$$\limsup_{J \to \infty} \left(R(\theta, \widehat{\theta}^{\mathrm{URE}}) - R(\theta, \widetilde{\theta}^{\mathrm{OL}}) \right) \leq 0. \tag{5}$$

Proof. By definition of $\widehat{\psi}^{\mathrm{URE}}$, we have $\mathrm{URE}(\widehat{\psi}^{\mathrm{URE}}) \leq \mathrm{URE}(\widetilde{\psi}^{\mathrm{OL}})$. This gives

$$\begin{aligned}
&\ell(\theta, \widehat{\theta}^{\mathrm{URE}}) - \ell(\theta, \widetilde{\theta}^{\mathrm{OL}}) \\
\leq &\left(\ell(\theta, \widehat{\theta}^{\mathrm{URE}}) - \mathrm{URE}(\widehat{\psi}^{\mathrm{URE}}) \right) + \left(\mathrm{URE}(\widetilde{\psi}^{\mathrm{OL}}) - \ell(\theta, \widetilde{\theta}^{\mathrm{OL}}) \right) \\
\leq &2 \sup_{\psi \in \Psi} |\ell(\theta, \theta(\psi)) - \mathrm{URE}(\psi)|.
\end{aligned}$$

Taking expectations and then taking $\limsup_{J \to \infty}$ on both sides, the result follows from the L^1 convergence condition. □

Remark 4.1. It is worth noting that, under a slightly weaker convergence condition that requires only convergence in probability, one obtains

$$\lim_{J \to \infty} \mathbf{P} \left(\ell(\theta, \widehat{\theta}^{\mathrm{URE}}) \geq \ell(\theta, \widetilde{\theta}^{\mathrm{OL}}) + \varepsilon \right) = 0,$$

which is in fact implied by (5). This result shows that loss of the URE estimator converges to the oracle loss, in probability.

Because the left-hand side of (5) cannot be strictly negative due to the definition of the oracle, this in fact shows that the asymptotic risk of the URE estimator is the same as the oracle under the given convergence assumption. Note that the optimality result is conditional on a true mean vector sequence $\{\theta_j\}_{j=1}^{\infty}$ such that the uniform L^1 convergence holds. When establishing the L^1 convergence for the specific estimators, I impose conditions on the data $\{y_j\}_{j=1}^{\infty}$ that ensure such convergence indeed holds.

4.2 Establishing optimality

All three estimators, $\widehat{\theta}^{\mathrm{URE,m}}$, $\widehat{\theta}^{\mathrm{URE,g}}$ and $\widehat{\theta}^{\mathrm{URE,cov}}$, take the form of

$$\widehat{\theta}_j(\mu_j, \Lambda) = \left(I_T - \Lambda(\Lambda + \Sigma_j)^{-1} \right) \mu_j + \Lambda(\Lambda + \Sigma_j)^{-1} y_j,$$

with different restrictions imposed on μ_j. Hence, the difference between the URE and the true loss in all three cases can be written as (see (12) in Appendix A.1 for the derivation)

$$\begin{aligned}&\frac{1}{J}\sum_{j=1}^{J}\Big(\mathrm{URE}_j(\mu_j,\Lambda)-(\widehat{\theta}_j(\mu_j,\Lambda)-\theta_j)'(\widehat{\theta}_j(\mu_j,\Lambda)-\theta_j)\Big)\\&=\Big(\mathrm{URE}(0,\Lambda)-\ell(\theta,\widehat{\theta}(0,L))\Big)-\frac{2}{J}\sum_{j=1}^{J}\left(\mu_j'(\Lambda+\Sigma_j)^{-1}\Sigma_j(y_j-\theta_j)\right).\end{aligned}\tag{6}$$

An application of the triangle inequality implies that it suffices to show the absolute values of the first and second terms of the right-hand side converge to zero. Then, it follows from Lemma 4.1 that the URE estimators obtain the oracle risk for each class. The first term, which is the difference between the URE and the loss for the estimator that shrinks toward zero, does not depend on μ_j and thus is common for all three estimators. I first show the convergence of this term, and then establish the convergence of the second term for each estimator.

The following assumption states that y_j's are independent, the fourth moment of y_j is bounded (uniformly over j), and that the smallest eigenvalue of the variance of y_j is bounded away from zero. I write $y_j \sim (\theta_j, \Sigma_j)$ to mean that y_j follows a distribution such that $\mathbf{E}\, y_j = \theta_j$ and $\mathrm{var}(y_j) = \Sigma_j$. The supremum $\sup_j$ is taken over all $j \geq 1$, and likewise for $\inf_j$. Hence, the assumption imposes conditions on the sequences $\{\mathbf{E}\|y_j\|\}_{j=1}^{\infty}$ and $\{\sigma_T(\Sigma_j)\}_{j=1}^{\infty}$.

Assumption 4.1 (Independent sampling and boundedness)**.** (i) $y_j \overset{\text{indep}}{\sim} (\theta_j, \Sigma_j)$, (ii) $\sup_j \mathbf{E}\|y_j\|^4 < \infty$ *and* (iii) $0 < \inf_j \sigma_T(\Sigma_j)$.

This assumption is maintained throughout the paper. Note that normality is not required. Accordingly, in the linear panel data model, the idiosyncratic terms are not required to follow a normal distribution. The independence assumption can be relaxed further provided a law of large number goes through for $y_j'y_j$. In the case where Σ_j is diagonal for all j, Assumption 4.1 (iii) boils down to assuming that $\mathrm{var}(y_{jt})$ is bounded away from zero over j and t. Also, in the case where $\Sigma_j = \Sigma$ for all j, the assumption trivially holds as long as Σ is invertible. I note that (ii) implies $\sup_j \|\theta_j\| < \infty$ and $\sup_j \mathrm{tr}(\Sigma_j) < \infty$ and that (ii) and (iii) together imply $\sup_j \kappa(\Sigma_j) < \infty$, which are implications repeatedly used in the proof.

Example 4.1 (Teacher value-added)**.** In the teacher value-added model, Assumption 4.1 holds if, for example, (a) $\underline{n} \leq n_{jt} \leq \overline{n}$ for all j, t, (b) the true fixed effects are uniformly bounded in magnitude, and (c) the idiosyncratic error term has bounded fourth moment. Since the class size of any teacher is clearly bounded, (a) is easily justified. Typically, the unit of teacher value-added is in standard deviation of the test score, and the test scores are bounded. Hence, as long as the standard deviation of the test scores for each year is strictly positive, which is always the case, (b) is satisfied. The existence of the fourth moment of the idiosyncratic error term, (c), is a mild regularity condition.

The following theorem shows that Assumption 4.1 is enough to ensure uniform convergence of the first term of (6).

Theorem 4.1 (Uniform convergence of URE$(0, \Lambda)$)**.** *Suppose Assumption 4.1 holds. Then,*

$$\sup_{\Lambda \in \mathcal{S}_T^+} \left| \mathrm{URE}(0, \Lambda) - \ell(\theta, \widehat{\theta}(0, \Lambda)) \right| \xrightarrow{L^1} 0. \tag{7}$$

This theorem establishes uniform convergence over the largest possible hyperparameter space for Λ, $\mathcal{S}_T^+$, and thus the convergence over any $\mathcal{L} \subset \mathcal{S}_T^+$ follows. Also, as is clear from the proof, Assumption 4.1 is stronger than necessary. However, I find this stronger set of assumptions not very restrictive with the advantage of being easy to interpret.[17] The proof essentially boils down to establishing a ULLN argument, and then verifying uniform integrability to strengthen the convergence mode from convergence in probability to convergence in L^1.[18] In fact, Theorem 4.1 implies the asymptotic optimality of the URE method when the centering parameter μ is taken to equal zero. With this result in hand, I now establish the optimality of each of the three URE estimators by showing that the second term of (6) converges to zero uniformly in L^1.

Define the oracle estimators of each class as $\widetilde{\theta}^{\mathrm{OL,m}}$, $\widetilde{\theta}^{\mathrm{OL,g}}$ and $\widetilde{\theta}^{\mathrm{OL,cov}}$, which are the estimators obtained by plugging in the oracle hyperparameters. Specifically, define

[17] For example, it suffices to assume that the average of $\mathbf{E}\|y_j\|^4$ satisfies a boundedness condition rather than the supremum over such quantities. Hence, the optimality results still hold if the data and the true mean vectors are indeed drawn from a normal distribution.

[18] The proof technique used in related papers such as Xie et al. (2012, 2016) and Kou and Yang (2017), of applying an equality due to Li (1986) followed by an application of Doob's martingale inequality do not go through here. The main reason is that the matrix hyperparameter Λ governs the direction of shrinkage as well as the magnitude, whereas there is only a scalar hyperparameter λ that determines the magnitude of shrinkage in such papers.

the optimal hyperparameters as

$$\begin{aligned}\widetilde{\Lambda}^{\mathrm{OL,m}} &:= \operatorname*{arg\,min}_{\Lambda \in \mathcal{S}_T^+} \ell(\theta, \widehat{\theta}(\overline{y}_J, \Lambda)), \\ (\mu^{\mathrm{OL}}, \widetilde{\Lambda}^{\mathrm{OL,g}}) &:= \operatorname*{arg\,min}_{(\mu,\Lambda) \in \mathcal{M}_J \times \mathcal{S}_T^+} \ell(\theta, \widehat{\theta}(\mu, \Lambda)), \text{ and} \\ (\widetilde{\gamma}^{\mathrm{OL}}, \widetilde{\Lambda}^{\mathrm{OL,cov}}) &:= \operatorname*{arg\,min}_{(\gamma,\Lambda) \in \Gamma_J \times \mathcal{S}_T^+} \ell(\theta, \widehat{\theta}^{\mathrm{cov}}(\gamma, \Lambda)).\end{aligned}$$

The, the corresponding estimators are defined as

$$\widetilde{\theta}^{\mathrm{OL,m}} := \widehat{\theta}(\overline{y}_J, \widetilde{\Lambda}^{\mathrm{OL,m}}),\ \widetilde{\theta}^{\mathrm{OL,g}} := \widehat{\theta}(\widetilde{\mu}^{\mathrm{OL}}, \widetilde{\Lambda}^{\mathrm{OL,g}}),\ \text{and}\ \widetilde{\theta}^{\mathrm{URE,cov}} := \widehat{\theta}^{\mathrm{cov}}(\widetilde{\gamma}^{\mathrm{OL}}, \widetilde{\Lambda}^{\mathrm{OL,cov}}).$$

Grand mean. This estimator shrinks the data toward $\overline{y}_J$, which corresponds to taking $\mu_j = \overline{y}_J$ in the last term of (6). Hence, the convergence result to be established is

$$\sup_{\Lambda \in \mathcal{S}_T^+} \left| \frac{1}{J} \sum\nolimits_{j=1}^{J} \overline{y}_J'(\Lambda + \Sigma_j)^{-1}\Sigma_j(y_j - \theta_j) \right| \xrightarrow{L^1} 0.$$

Two applications of the Cauchy-Schwarz inequality show that the expectation of the left-hand side is bounded by

$$\left(\mathbf{E}\|\overline{y}_J\|^2\right)^{1/2} \left(\mathbf{E} \sup_\Lambda \left\| \tfrac{1}{J} \textstyle\sum_{j=1}^{J} (\Lambda + \Sigma_j)^{-1}\Sigma_j(y_j - \theta_j) \right\|^2 \right)^{1/2}.$$

The limit supremum, as $J \to \infty$, of the first term is bounded by Assumption 4.1 (ii). For the second term, again a ULLN argument can be used to show that this converges to zero under Assumption 4.1.[19]

Therefore, under Assumption 4.1, it follows that

$$\limsup_{J \to \infty} \left(R(\theta, \widehat{\theta}^{\mathrm{URE,g}}) - R(\theta, \widetilde{\theta}^{\mathrm{OL,g}}) \right) = 0,$$

and thus the URE estimator obtains the best possible risk within the class. In particular, this class includes the EB methods that shrink to zero after demeaning the fixed effects, and thus establishes that this URE estimator dominates widely used estimators.

Furthermore, the URE estimator also can be shown to dominate the unbiased

[19] I show this in the proof of Theorem 4.2 because the same term appears there as well.

estimator, y, which corresponds to using the least squares estimators without any shrinkage in the context of fixed effects. This is the maximum likelihood estimator (MLE) when the distribution of y is assumed normal. With some abuse of terminology, I refer to this as the MLE even though we are not assuming normality. Because there is no $\Lambda \in \mathcal{S}_T^+$ such that $\widehat{\theta}(\overline{y}_J, \Lambda) = y$, the MLE y is not included in the class of estimators we consider. However, a simple approximation argument can be used to establish the dominance.

Suppose there is some sequence $\widetilde{\Lambda}_J^{\text{MLE}}$ such that $\widetilde{\theta}^{\text{MLE}} = \widehat{\theta}(\overline{y}_J, \widetilde{\Lambda}_J^{\text{MLE}})$ satisfies

$$\lim_{J\to\infty} \left| R(\theta, \widetilde{\theta}^{\text{MLE}}) - R(\theta, y) \right| = 0. \tag{8}$$

Then, it follows that

$$\begin{aligned} &\limsup_{J\to\infty} \left(R(\theta, \widehat{\theta}^{\text{URE}}) - R(\theta, y) \right) \\ \leq &\limsup_{J\to\infty} \left(R(\theta, \widehat{\theta}^{\text{URE}}) - R(\theta, \widetilde{\theta}^{\text{MLE}}) \right) + \limsup_{J\to\infty} \left(R(\theta, \widetilde{\theta}^{\text{MLE}}) - R(\theta, y) \right) \leq 0, \end{aligned}$$

where the last inequality follows due to the optimality of $\widehat{\theta}^{\text{URE}}$ and the assumption on $\widetilde{\theta}^{\text{MLE}}$. Hence, finding $\widetilde{\Lambda}_J^{\text{MLE}}$ that satisfies (8) is key to establishing that $\widehat{\theta}^{\text{URE,m}}$ weakly dominates y. Define $D(\lambda) := \text{diag}(\lambda, \ldots, \lambda)$, and note that for any fixed J, we have

$$\lim_{\lambda\to\infty} R(\theta, \widehat{\theta}(\overline{y}_J, D(\lambda))) = R(\theta, y).$$

Hence, there exists λ_J such that $\left| R(\theta, \widehat{\theta}(\overline{y}_J, D(\lambda_J))) - R(\theta, y) \right| \leq \frac{1}{J}$, and thus taking $\widetilde{\theta}^{\text{MLE}} = \widehat{\theta}(\overline{y}_J, D(\lambda_J))$ satisfies (8). This shows that shrinking the least squares estimator using the URE method cannot do worse than using the least squares estimator, which is a property that EB methods do not have.

General location. This estimator shrinks the data toward a general data-driven location μ, with the restriction that $\mu \in \mathcal{M}_J$. The convergence result to be established is

$$\sup_{(\mu,\Lambda)\in\mathcal{M}_J\times\mathcal{S}_T^+} \left| \frac{1}{J} \sum\nolimits_{j=1}^{J} \mu'(\Lambda + \Sigma_j)^{-1}\Sigma_j(y_j - \theta_j) \right| \xrightarrow{L^1} 0.$$

As in the shrinkage to the grand mean case, it follows from Cauchy-Schwarz inequality

that the expectation of the left-hand side is bounded by

$$\left(\mathbf{E}\sup_{\mu\in\mathcal{M}_J}\|\mu\|^2\right)^{1/2}\left(\mathbf{E}\sup_{\Lambda\in\mathcal{S}_T^+}\left\|\tfrac{1}{J}\sum_{j=1}^J(\Lambda+\Sigma_j)^{-1}\Sigma_j(y_j-\theta_j)\right\|^2\right)^{1/2}.$$

As mentioned earlier, it can be shown that the second term converges to zero by a ULLN argument under Assumption 4.1, which I show in the proof of Theorem 4.2.

To show that the term $\mathbf{E}\sup_{\mu\in\mathcal{M}_j}\|\mu\|^2$ is bounded, note that

$$\mathbf{E}\sup_{\mu\in\mathcal{M}_j}\|\mu\|^2=\sum_{t=1}^T\mathbf{E}\,q_{1-\tau}(\{y_{jt}^2\}_{j=1}^J)$$

by the definition of $\mathcal{M}_J$. Hence, it suffices to show $\mathbf{E}\,q_{1-\tau}(\{y_{jt}^2\}_{j=1}^J)=O(1)$ for each $t\leq T$. To control the sample quantile behavior of $\{y_{jt}^2\}_{j=1}^J$, I impose an additional condition. Write $\varepsilon_{jt}:=y_{jt}-\theta_{jt}$ so that $E\varepsilon_{jt}=0$ and $E\varepsilon_{jt}^2=\sigma_{jt}^2$, where σ_{jt}^2 denotes the tth diagonal entry of Σ_j. Note that $\overline{\sigma}_t^2:=\sup_j\sigma_{jt}^2<\infty$ by Assumption 4.1. I assume that the distribution of ε_{jt} belongs to a scale family with finite fourth moments.

Assumption 4.2 (Scale family)**.** *For each $t=1,\ldots,T$, we have $\varepsilon_{jt}/\sigma_{jt}\overset{i.i.d.}{\sim}F_t$, where F_t is a distribution function with finite fourth moments, for $j=1,\ldots,J$.*

Note that the assumption is notably weaker than requiring that the noise vectors ε_j for $j=1,\ldots,J$ belong to a multivariate scale family, which restricts the joint distribution across t in a much more stringent way. Here, I instead require that the error terms belong to a scale family only for each period. It can be shown that, by Assumption 4.1, the problem of bounding $\mathbf{E}\,q_{1-\tau}(\{y_{jt}^2\}_{j=1}^J)$ boils down to the problem of bounding $\mathbf{E}\,q_{1-\tau}(\{(\varepsilon_{jt}/\sigma_{jt})^2\}_{j=1}^J)$. Then, by Assumption 4.2, this simplifies to bounding the mean of the sample quantile of an i.i.d. sample. I use a result given by Okolewski and Rychlik (2001) to derive a bound on this quantity without having to further impose conditions on the distribution F_t.

Example 4.2 (Teacher value-added)**.** In the teacher value-added example, Assumption 4.2 is satisfied as long as the idiosyncratic error terms are i.i.d across j and have finite fourth moment. Hence, this assumption is almost always satisfied in teacher value-added models, or in linear panel data models in general.

The following theorem shows that $\mathrm{URE}(\mu,\Lambda)$ is close to the true loss uniformly over $(\mu,\Lambda)\in\mathcal{M}_J\times\mathcal{S}_T^+$ under this additional assumption.

Theorem 4.2 (Uniform convergence of URE(μ, Λ))**.** *Suppose Assumptions 4.1 and 4.2 hold. Then,*

$$\sup_{\mu \in \mathcal{M}_J, \Lambda \in \mathcal{S}_T^+} \left| \mathrm{URE}(\mu, \Lambda) - \ell(\theta, \widehat{\theta}(\mu, \Lambda)) \right| \xrightarrow{L^1} 0.$$

Again, by Lemma 4.1, this ensures that $\widehat{\theta}^{\mathrm{URE,g}}$ asymptotically obtains the oracle risk, as stated in the following corollary.

Corollary 4.1. *Under Assumptions 4.1 and 4.2,*

$$\limsup_{J \to \infty} \left(R(\theta, \widehat{\theta}(\widehat{\mu}^{\mathrm{URE}}, \widehat{\Lambda}^{\mathrm{URE,g}})) - R(\theta, \widetilde{\theta}^{\mathrm{OL,g}}) \right) \leq 0.$$

Under homoskedasticity (i.e., $\Sigma_j = \Sigma$ for all $j \leq J$), the optimal location parameter for both the URE estimator and the EBMLE estimator is the grand mean, so that $\widehat{\mu}^{\mathrm{URE}} = \widehat{\mu}^{\mathrm{EBMLE}} = \overline{y}_J$. This is possibly one reason why the grand mean has been frequently used as the centering location in applied work despite the heteroskedasticity.[20] However, under heteroskedasticity, which is frequently the case in the context of fixed effects, weighing the different observations according to the different variance matrices Σ_j (and the hyperparameter Λ) gives better risk properties. Hence, it is recommended that one uses $\widehat{\theta}^{\mathrm{URE,g}}$ rather than $\widehat{\theta}^{\mathrm{URE,m}}$ unless the sample size is relatively small, in which case the additional hyperparameters can result in overfitting.

Linear combination of covariates. This estimator shrinks each observation to a different location, $Z_j\gamma$, which depends on the covariate. By a similar calculation given in the case of shrinkage toward a general location, the key step in establishing optimality is to show

$$\left(\mathbf{E} \sup_{\gamma \in \Gamma_J} \|\gamma\|^2\right)^{1/2} \left(\mathbf{E} \sup_{\Lambda \in \mathcal{S}_T^+} \left\| \tfrac{1}{J} \textstyle\sum_{j=1}^J Z_j'(\Lambda + \Sigma_j)^{-1} \Sigma_j (y_j - \theta_j) \right\|^2 \right)^{1/2} \to 0.$$

Again, the strategy is to show that the first term is bounded and the second term converges to zero. Due to the presence of covariates, the second term is different from that of the previous estimators. Define $\varepsilon_j = y_j - \theta_j$. I make the following assumptions on the covariates.

[20] Another plausible explanation is that $\overline{y}_J$ is an EB method of moments estimator for μ, though one can obtain an alternative method of moments estimator with smaller variance by weighting appropriately.

Assumption 4.3 (Covariates)**.**

(i) $\{(y_j, Z_j)\}_{j=1}^J$ *is an independent sample with* $Z_j \overset{\text{i.i.d,}}{\sim} P_Z$,

(ii) $\sup_j \sigma_1(Z_j'Z_j) < \infty$ *a.s.*,

(iii) $\mathbf{E}[\varepsilon_j|Z_j] = 0$ *and* $\text{var}(\varepsilon_j|Z_j) = \Sigma_j$,

(iv) $\mu_{Z,2} := \mathbf{E}\, Z_j'Z_j$ *is nonsingular, and*

(v) $\sup_j \mathbf{E}\left[\|y_j\|^4|Z_j\right] < \infty$ *a.s.*

Again, the supremums are taken over all $j \geq 1$. The independent sampling assumption of (i) is standard. A sufficient condition for (ii) is that there exists some constant $\overline{C}_Z \in \mathbf{R}$ such that $\sup_{j,t}\|Z_{jt}\| < \overline{C}_Z < \infty$ almost surely, which amounts to assuming that the covariates are uniformly bounded. The first and second part of (iii) are exogeneity conditions for the first and second moments of the noise term, with respect to the covariates. The full rank condition given in (iv) is standard. The boundedness condition for the conditional expectation given in (v) is a conditional version of Assumption 4.1 (ii). Again, the boundedness conditions in (ii) and (v) can be relaxed to a boundedness condition on the averages of the given quantities.

Note that there is no assumption that states any linear relationship between the covariate matrix Z_j and the true mean θ_j and/or y_j. Hence, there is no such thing as "misspecification" as long as the exogeneity condition (ii) is met. Some specifications yield better risk properties than others, but as long as time dummies are included in the covariates with B being sufficiently large, any specification (choice of covariates) is guaranteed to improve upon $\widehat{\theta}^{\text{URE,g}}$ asymptotically.

Example 4.3 (Teacher value-added)**.** In the teacher value-added model, Z_{jt} corresponds to teacher-year level covariates. The results here show that, while such covariates could not have been included in the regression formula, such covariates can be used to obtain more accurate estimators of the fixed effects. The exogeneity condition, Assumption 4.3 (ii), is satisfied as long as the covariates are strictly exogenous with respect to the idiosyncratic error terms. This was in some sense already assumed because independence between the fixed effects and the idiosyncratic error terms was assumed.

Now, with some abuse of notation, I condition on a realization $\{Z_j\}_{j=1}^\infty$ and treat the covariates as fixed. I assume that this fixed sequence satisfies $\sup_j \sigma_1(Z_j'Z_j) < \infty$, $\frac{1}{J}\sum_{j=1}^J Z_j'Z_j \to \mu_{Z,2}$, and $\sup_j \mathbf{E}\left[\|y_j\|^4|Z_j\right] < \infty$ which holds for almost all real-

ized sequences due to Assumption 4.3(ii), (iv), and (v), and the strong law of large numbers. I directly impose these conditions on the fixed covariates in the following theorem, with the understanding that such conditions follow from Assumption 4.3. The following theorem shows that the URE is uniformly close to the true loss function over $(\gamma, \Lambda) \in \Gamma_J \times \mathcal{S}_T^+$ under these implied assumption on the covariates, along with the maintained Assumption 4.1.

Theorem 4.3 (Uniform convergence of $\mathrm{URE}^{\mathrm{cov}}(\gamma, \Lambda)$)**.** *Suppose* $\sup_j \sigma_1(Z_j'Z_j) < \infty$, $\lim_{J\to\infty} \frac{1}{J}\sum_{j=1}^J Z_j'Z_j = \mu_{Z,2}$*, and Assumption 4.1 holds. Then,*

$$\sup_{\gamma\in\Gamma_J, \Lambda\in\mathcal{S}_T^+} \left| \mathrm{URE}^{\mathrm{cov}}(\gamma, \Lambda) - \ell(\theta, \widehat{\theta}^{\mathrm{cov}}(\gamma, \Lambda)) \right| \xrightarrow{L^1} 0.$$

Again, invoking Lemma 4.1 gives the following corollary, which states the URE estimator obtains the oracle risk in this context as well.

Corollary 4.2. *Suppose* $\sup_j \sigma_1(Z_j'Z_j) < \infty$, $\lim_{J\to\infty} \frac{1}{J}\sum_{j=1}^J Z_j'Z_j = \mu_{Z,2}$*, and Assumption 4.1 holds. Then,*

$$\limsup_{J\to\infty} \left(R(\theta, \widehat{\theta}^{\mathrm{cov}}(\widehat{\gamma}^{\mathrm{URE}}, \widehat{\Lambda}^{\mathrm{URE,cov}})) - R(\theta, \widetilde{\theta}^{\mathrm{OL,cov}}) \right) \leq 0.$$

4.3 Discussion on the optimality results

The optimality of the URE estimators requires only mild conditions on the moments of the data, which is in contrast with the EB estimators that require stringent distributional assumptions to obtain optimality properties. Recall that y_j and θ_j correspond to the least squares estimator and the true fixed effect, respectively, in the context of fixed effects. The EB estimators are optimal in the sense of Robbins (1964)[21] when 1) the normality assumptions for both the least squares estimator and the true fixed effect hold and 2) the true fixed effect and variance of the least squares estimator are independent.[22]

The normality assumption on the true fixed effect is typically difficult to justify. Some evidence on the violation of such assumption in the context of teacher value-added is provided in Gilraine et al. (2020). The optimality results here are conditional

[21] That is, the estimator obtains the Bayes risk under the unknown moments of true fixed effects.

[22] The second assumption regarding the dependence between the mean and variance of the least squares estimator is more implicit, but can be seen from the fact that the model for the true fixed effect does not depend on the variance of the least squares estimator.

on a sequence of true mean vectors that is only required to satisfy a mild boundedness condition, and does not rely on such specific distributional assumptions on the true mean vector. The normality assumption on the least squares estimator can be less concerning because one can resort to a central limit theorem argument if the class size n_{jt} is somewhat large. However, this is not necessarily the case in many empirical contexts. For example, there are numerous classes with less than ten students in the data used in Section 7. The optimality result for the URE estimators imposes conditions on the moments of the least squares estimator, leaving the distribution unrestricted.

The independence assumption between the true fixed effect and the variance of the least squares estimator can be easily violated in empirical settings as well. Since the variance of the least squares estimator is inversely proportional to the cell size n_{jt}, the assumption is violated if the fixed effect is related with the cell size in some manner. For example, if teachers with higher value-added teach more students, or if the size of the class is related to teaching effectiveness, then such independence is unlikely to hold. Also, if the idiosyncratic error terms are conditionally heteroskedastic with respect to some observed covariates that are correlated with the fixed effects, such independence assumption is again violated. No assumption on the relationship between the mean θ_j and variance Σ_j is imposed in establishing the optimality of the URE estimators, and thus optimality is guaranteed whether or not the true fixed effect and the variance of the least squares estimators are independent.

It is worth mentioning that the nonparametric EB literature (Jiang and Zhang (2009), Brown and Greenshtein (2009), Koenker and Mizera (2014)) provides an alternative method to relax the normality assumption of the true mean, which has been adopted by Gilraine et al. (2020) to the teacher value-added setting. In the nonparametric EB setting, the distribution of the true fixed effect is remained unspecified except for certain regularity conditions. This allows for a significantly wider class of estimators than the class I consider. I view this approach as complementary to the URE approach for two main reasons. First, with time-varying fixed effects the nonparametric EB approach involves solving an optimization problem where the argument is a function of T variables, which makes computation very difficult for even moderate values of T. Moreover, the risk properties of the currently available nonparametric EB methods still rely on an independence assumption between the true fixed effect and the variance of the least square estimator and a normality assumption

on the least squares estimators.

Remark 4.2 (Unbalanced panel)**.** While it has been assumed that the given panel data is balanced at the (j,t)-level, this is rarely the case in empirical applications. For example, in teacher value-added, only some of the teachers are observed for the entire time span of the data and others appear only in some of the school years. The URE estimators and their optimality can be naturally extended to incorporate this unbalanced case. See Appendix B for details.

5 Summarizing the time trajectory

While the time-varying fixed effects gives more flexibility and contains more information, in some empirical contexts it is still desirable to have a scalar quantity for each j that summarizes the fixed effect for unit j. For example, in teacher value-added, a scalar that summarizes the value-added for each teacher is necessary to rank the teachers. To this end, I provide two methods that give a summary of the time trajectory of the fixed effects: estimating a weighted mean over time and forecasting the one-period-ahead fixed effects. Again, in the context of teacher value-added, the former provides a summary of a teacher's past performance, and the latter provides a prediction on how well a teacher is expected to do in the following year.

The estimators are derived using a similar idea as in the problem of estimating the full vectors: restrict the class of parameters using an appropriate model and tune the hyperparameters by minimizing a risk estimate. The hyperparmeters are tuned in a way that the MSE of the estimator for the weighted mean/or one-period-ahead fixed effects are optimal, rather then aiming for the MSE optimality for the problem of estimating the entire vector.

5.1 Estimating weighted means

The first, more simple way to summarize the time-varying fixed effects as a scalar is reporting a weighted mean of θ_j rather than the full vector. Let $w = (w_1, \ldots, w_T)' \in \mathbf{R}^T$ denote a weight vector such that $w_t \geq 0$ and $\sum_{t=1}^T w_t = 1$ which represents the weight that is of interest. That is, the parameter of interest is now $(w'\theta_1, \ldots, w'\theta_J)'$. Again, the class of estimators is restricted by postulating $\theta_j \overset{\text{i.i.d.}}{\sim} N(\mu, \Lambda)$ on top of a normality assumption on y_j. The posterior mean of $w'\theta_j$ under this model is given

as $\mathbf{E}[w'\theta_j|y] = w'\widehat{\theta}_j(\mu,\Lambda)$, which is simply a weighted version of $\widehat{\theta}_j(\mu,\Lambda)$. The loss function for this class of estimators is given as

$$\begin{aligned}&\frac{1}{J}\sum_{j=1}^{J}(w'\widehat{\theta}_j(\mu,\Lambda)-w'\theta_j)'(w'\widehat{\theta}_j(\mu,\Lambda)-w'\theta_j)\\ =&\frac{1}{J}\sum_{j=1}^{J}(\widehat{\theta}_j(\mu,\Lambda)-\theta_j)'ww'(\widehat{\theta}_j(\mu,\Lambda)-\theta_j),\end{aligned}$$

which is simply weighted versions of the loss function used for the estimation of full vectors with weight matrix $W := ww'$. For example, when the interest is in the simple average over time, one can take $w = \mathbf{1}_T$.

Appendix C provides a URE, $\text{URE}^W(\mu,\Lambda)$, for this weighted loss function (and for more general weighted loss functions) and some details on how the URE estimator derived by minimizing this URE obtains the oracle risk under the class of estimators. I note that the tuning parameter that is optimal for the full vector estimation is not necessarily optimal for the estimation of weighted means.

5.2 Forecasting θ_{T+1}

Another succinct summary of the time trajectory is the forecast for the fixed effects of period $T+1$. This forecasting problem is of independent interest as well. The problem is to predict $\theta_{T+1} = (\theta_{1,T+1},\dots\theta_{J,T+1})'$ with only the T period data in hand. The approach is similar to the URE approach taken for estimation problem. I derive a class of predictors using a hierarchical model, and tune the hyperparameters by minimizing a unbiased prediction error estimate (UPE). For this reason, the resulting forecasts are referred to as UPE forecasts.

Consider the second level model $\theta_j \sim N(0,\Lambda)$ centered at zero. Here, I consider the case where the fixed effects are demeaned so that the fixed effects are assumed be drawn from a distribution centered at zero. Write the block matrices of the tuning parameter Λ and the variance matrix Σ_j as

$$\Lambda = \begin{pmatrix}\Lambda_{-T} & \Lambda_{T,-T}\\ \Lambda'_{T,-T} & \lambda_T\end{pmatrix}, \Sigma_j = \begin{pmatrix}\Sigma_{j,-T} & \Sigma_{j,T,-T}\\ \Sigma'_{j,T,-T} & \Sigma_{j,T}\end{pmatrix} = \begin{pmatrix}\Sigma_{j,1} & \Sigma'_{j,1,-1}\\ \Sigma_{j,1,-1} & \Sigma_{j,-1}\end{pmatrix}$$

where Λ_{-T}, $\Sigma_{j,-T}$ and $\Sigma_{j,-1}$ are $(T-1)\times(T-1)$ matrices. From the property

of positive semidefinite matrices, Λ is positive semidefinite if only if Λ_T is positive semidifinite and $\Lambda_{-T} \geq \frac{1}{\lambda_T}\Lambda_{T,-T}\Lambda_{T,-T}'$.

Here, the hyperparameter space is restricted to some bounded set $\mathcal{L} \subset \mathcal{S}_T^+$. A recommended choice of $\mathcal{L}$ is to take

$$\mathcal{L} := \left\{\Lambda \in \mathcal{S}_T^+ : \sigma_1(\Lambda) \leq K\sigma_1\left(\frac{1}{J}\sum\nolimits_{j=1}^{J} y_j y_j'\right)\right\}$$

for some large number K that does not depend on J. The motivation is that, under the hierarchical model, $\frac{1}{J}\sum_{j=1}^{J}\mathbf{E}\, y_j y_j' = \Lambda + \frac{1}{J}\sum_{j=1}^{J}(\theta_j\theta_j' + \Sigma_j)$, and thus $\frac{1}{J}\sum_{j=1}^{J} y_j y_j'$ gives a sense of the scale of Λ. By multiplying by a large number K, the bound is made less restrictive. In the empirical application, I use $K = 100$ and this constraint does not bind.

The aim is to tune the hyperparameter in a way that it minimizes prediction error of predicting $\theta_{T+1} := (\theta_{1,T+1}, \ldots, \theta_{J,T+1})'$. However, the difficulty here is that an unbiased estimator of this prediction error is unavailable because we do not observe anything for period $T+1$. Hence, the strategy is to tune the hyperparameters by considering the problem of predicting $\theta_T = (\theta_{1T}, \ldots, \theta_{JT})'$ using only the first $T-1$ periods of data. Then, under a suitable stationarity assumption, I extrapolate and use this hyperparameter to predict θ_{T+1}.

First, consider the problem of forecasting θ_T with only the data from the first $T-1$ periods. Define $y_{j,-t} = (y_{j1}, \ldots, y_{j,t-1}, y_{j,t+1}, \ldots y_{j,T})'$ and $y_{-t} = (y_{1,-T}', \ldots, y_{J,-T}')'$ to be the vectors y_j and y, respectively, with the observations corresponding to period t removed. The class of estimators I consider is again the posterior mean implied by the hierarchical model,

$$\mathbf{E}[\theta_{jT}|y_{-T}] = \Lambda_{T,-T}'(\Lambda_{-T} + \Sigma_{j,-T})^{-1} y_{j,-T}.$$

Define $B(\Lambda, \Sigma_{-T}) = (\Lambda_{-T} + \Sigma_{-T})^{-1}\Lambda_{T,-T}$. The performance criterion is the mean prediction error, $\mathbf{E}\,\mathrm{PE}(\Lambda; T)$, where

$$\mathrm{PE}(\Lambda; T) := \tfrac{1}{J}\textstyle\sum_{j=1}^{J}(B(\Lambda, \Sigma_{j,-T})' y_{j,-T} - \theta_{jT})^2.$$

Ideally one would choose Λ to minimize $\mathrm{PE}(\Lambda; T)$. However, this prediction error depends on the true parameters, and thus this strategy is infeasible. Again, I derive an estimator of the prediction error and choose Λ by minimizing this. Some algebra

shows (see the first couple paragraphs of Appendix A.4)

$$\begin{aligned}&\mathbf{E}[(B(\Lambda,\Sigma_{j,-T})'y_{j,-T}-\theta_{jT})^2]\\=&\mathbf{E}[(B(\Lambda,\Sigma_{j,-T})'y_{j,-T}-y_{jT})^2]-\Sigma_{jT}+2B(\Lambda,\Sigma_{j,-T})'\Sigma_{j,T,-T}.\end{aligned}$$

Therefore, an unbiased estimator of the mean prediction error is given as

$$\mathrm{UPE}(\Lambda)=\frac{1}{J}\sum_{j=1}^{J}\left((B(\Lambda,\Sigma_{j,-T})'y_{j,-T}-y_{jT})^2-\Sigma_{jT}+2B(\Lambda,\Sigma_{j,-T})'\Sigma_{j,T,-T}\right).$$

Define $\widehat{\Lambda}^{\mathrm{UPE}}$ as the Λ that minimizes $\mathrm{URE}(\Lambda)$. The proposed estimator for $\theta_{j,T+1}$ is $B(\widehat{\Lambda}^{\mathrm{UPE}},\Sigma_{j,-T})'y_{j,-1}$.

Remark 5.1 (Estimator of Chetty et al. (2014a))**.** Here, I have been considering predicting θ_T with the first the observations from the first $T-1$ periods. However, more generally, one can also consider predicting θ_t with all observations except for the period t observation. If $\Sigma_j=\Sigma$ with Σ being diagonal, it can be shown that the Λ minimizes $\mathrm{UPE}(\Lambda)$ gives $B(\Lambda,\Sigma_{-t})=\widehat{\beta}^{\mathrm{OLS},t}$, which is the OLS estimator of regressing y_{jt} on $y_{j,-t}$. Hence, one estimates θ_{jt} with $y_{j,-t}'\widehat{\beta}^{\mathrm{OLS},t}$, which is exactly the estimator used in Chetty et al. (2014a).

Recall that the goal is to forecast θ_{T+1}, not θ_T. Hence, the goal is to show that $\mathrm{UPE}(\Lambda)$ is a good estimator of the prediction error for the problem of predicting θ_{T+1},

$$\mathrm{PE}(\Lambda;T+1)=\frac{1}{J}\sum_{j=1}^{J}(B(\Lambda,M_{j,-1})'y_{j,-1}-\theta_{j,T+1})^2.$$

By essentially the same argument made by Lemma 4.1, if

$$\sup_{\Lambda\in\mathcal{L}}|\mathrm{UPE}(\Lambda)-\mathrm{PE}(\Lambda;T+1)|\xrightarrow{L^1}0, \tag{9}$$

the mean prediction error of the estimator obtained by minimizing $\mathrm{UPE}(\Lambda)$ obtains the oracle mean prediction error, which is the mean prediction error of the estimator that minimizes $\mathrm{PE}(\Lambda;T+1)$.

For (9) to hold, a suitable stationarity assumption is necessary. To this end, suppose that $\left\{((\theta_j',\theta_{j,T+1})',\Sigma_j)\right\}_{j=1}^{\infty}$ is a random sample drawn from the density

$f_{(\theta',\theta_{T+1})',\Sigma}$. First, since I now treat the mean vector and variance matrices as random, I impose the following assumption to ensure that the distribution of the mean and variance parameters is consistent with Assumption 4.1. Let f_Σ denote the marginal density of Σ_j and $\text{supp}(f_\Sigma)$ the support of f_Σ.

Assumption 5.1 (Assumption 4.1 with random parameters)**.**
(i) $y_j|\theta_j, \Sigma_j \overset{\text{indep}}{\sim} (\theta_j, \Sigma_j)$,
(ii) $\sup_j \mathbf{E}[\|y_j\|^4|\theta_j, \Sigma_j] < \infty$, *and*
(iii) $\text{supp}(f_\Sigma) \subset \{\Sigma \in \mathcal{S}_T^+ : \sigma_T(\Sigma) > \underline{\sigma}_\Sigma\}$ *for some* $\underline{\sigma}_\Sigma > 0$.

To state the stationarity assumption, let $f_{\theta,\Sigma_{-T}}$ and $f_{(\theta'_{-1},\theta_{T+1})',\Sigma_{-1}}$ denote the marginal densities that correspond to $(\theta_j, \Sigma_{j,-T})$ and $((\theta'_{j,-1}, \theta_{j,T+1})', \Sigma_{j,-1})$, respectively. The following assumption states that the distributions of $(\theta_j, \Sigma_{j,-T})$ and $((\theta'_{j,-1}, \theta_{j,T+1})', \Sigma_{j,-1})$ are the same.

Assumption 5.2 (Stationarity)**.** $f_{\theta,\Sigma_{-T}} = f_{(\theta'_{-1},\theta_{T+1})',\Sigma_{-1}}$.

I emphasize that this assumption does not imply that the observations are mean stationary or variance stationary. Moreoever, this stationarity assumption on the joint distribution of the mean and variance does not impose any restriction on the dependence structure between the two, and thus the corresponding optimality result does not require independence of the mean and variance of y_j.

The following theorem shows that these two assumptions are enough to ensure that (9) holds almost surely, where the almost sureness is with respect to the randomness of the sequence $\left\{((\theta'_j, \theta_{j,T+1})', \Sigma_j)\right\}_{j=1}^\infty$.

Theorem 5.1. *Under Assumptions 5.1 and 5.2,*

$$\sup_{\Lambda \in \mathcal{L}} |\text{UPE}(\Lambda) - \text{PE}(\Lambda; T+1)| \xrightarrow{L^1} 0,$$

almost surely.

6 Simulation results

I carry out a simulation study to assess the finite sample performance of the URE estimators. I focus on experimenting the performance of the "shrinking to a general

location" estimator $\widehat{\theta}^{\text{URE,g}}$ with $T = 4$ and $\tau = .05$. The simulation study implies four main takeaways.

First, the MSE of the URE estimator gets close to the oracle risk with moderately large sample sizes. Across all data generating processes (DGPs) I have considered, the MSE of the URE estimator was less than 10% greater than the oracle risk as long as the sample size J is greater than 600. This shows that while all optimality results are only asymptotic, the sample size required to reach the oracle is not very large.

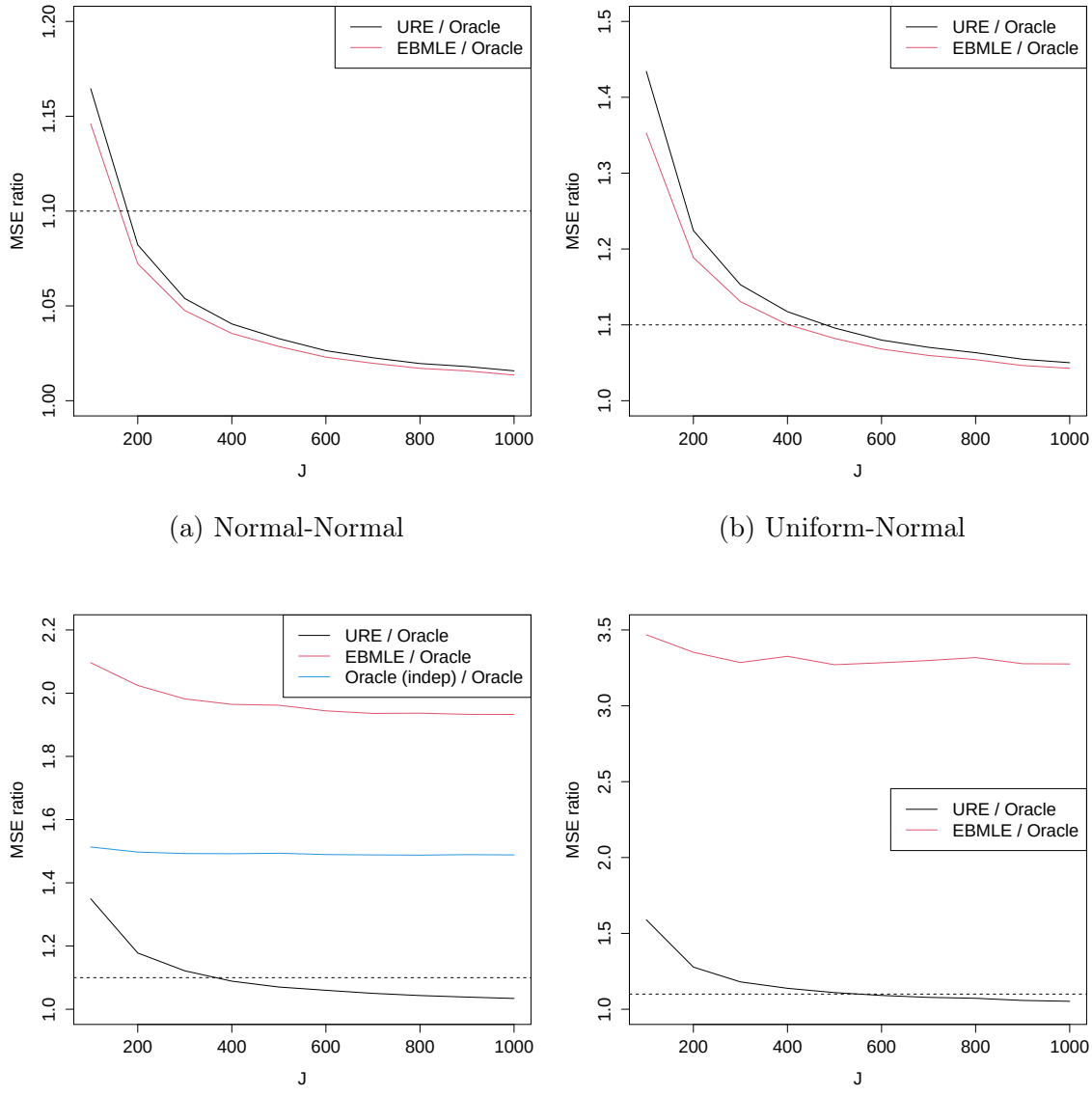
Second, there are numerous scenarios were the URE estimator shows significantly better performance than the EBMLE. This is largely expected, but still the magnitude is somewhat surprising, because there are cases where the URE estimator reduces the MSE of the EBMLE by more than 80%. Such improvements are most largest when there is a dependence structure between θ_j and Σ_j.

Third, the URE estimators perform almost as well as the EBMLE when the DGP satisfies or is close to satisfying the EB assumption. Even for such DGPs the risk of the URE is less than 5% greater even for small sample sizes such as $J = 100$. This is reassuring, because a concern about robust methods such as the URE estimator is that they may sacrifice performance too much under more typical assumptions for the sake of guarding against violations of such typical assumptions (the EB distributional assumptions in this case). The numerical results show that this is not the case.

Lastly, the URE estimator dominates the MLE, y, across all scenarios by a significant margin. The MLE is a useful benchmark, because regardless of the DGP, it is still an unbiased estimator of θ and a very natural one as well. However, simulation results show that it is almost always a good idea to use the URE estimator over the MLE, when the aim is to minimize MSE.

Figure 1 shows the simulation results for the four main scenarios.In the first Normal-Normal scenario, the true mean vectors are drawn from a normal distribution, $\theta_j \overset{\text{i.i.d.}}{\sim} N(0, I_T)$, the variance matrix from a Wishart distribution, $\Sigma_j = \frac{1}{30}\widetilde{\Sigma}_j$ where $\widetilde{\Sigma}_j \overset{\text{i.i.d.}}{\sim} \text{Wishart}(\Sigma_0, 30)$ with

$$\Sigma_0 := \begin{pmatrix} 1 & .75 & .5 & .25 \\ & 1 & .75 & .5 \\ & & 1 & .75 \\ & & & 1 \end{pmatrix},$$



(a) Normal-Normal

(b) Uniform-Normal

(c) Normal-Normal with group structure

(d) Conditional Heteroskedasticity

Figure 1: Simulation results for the four main scenarios. Each plot has a black line and red lines, which correspond to the MSE of the URE and EBMLE, respectively, as a fraction of the oracle MSE. The x-axis is the sample size from $J = 100$ to 1000. The dotted line is a horizontal line at 1.1 plotted to see when the MSE of the URE gets within 10% of the oracle.

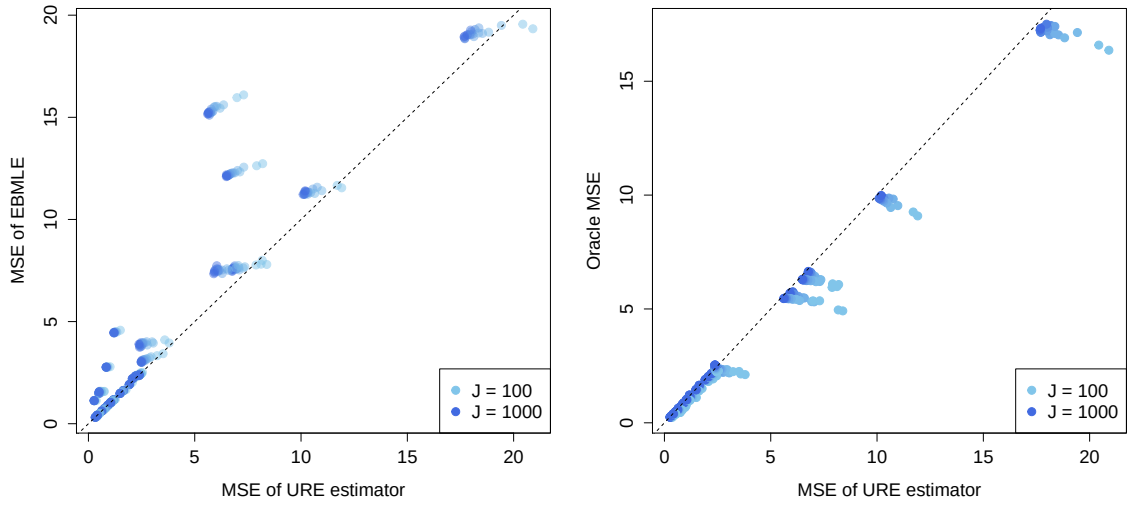
and $y_j \overset{\text{indep}}{\sim} N(\theta_j, \Sigma_j)$. This is a scenario where the distributional assumptions for EB are exactly met. As expected, EBMLE performs well, getting within 10% of the oracle with sample size as small as $J = 100$. The URE estimator shows good performance as well, with the difference in MSE with EBMLE being within 2% across all sample sizes.

The DGP of the Uniform-Normal scenario is the same as the Normal-Normal scenario except that $\theta_{jt} \sim \text{Unif}[0, .5t]$, drawn independently across both j and t. This DGP slightly violates the EB assumptions because the mean parameters are drawn from a uniform distribution rather than a normal distribution, but otherwise satisfies the distributional assumptions imposed by EB. The result is similar to the Uniform-Normal case, with the EBMLE performing very well, and the URE estimator showing a very slightly higher MSE than the EBMLE.

The third DGP is similar to the Normal-Normal scenario except that there is a group structure and the mean vectors are serially correlated. Specifically, half of the sample is drawn from the same DGP as in the Normal-Normal scenario, and the remaining half is drawn from a similar Normal-Normal scenario but with higher variance and greater mean. Here, we can think of the DGP as giving a small dependence structure on the mean and variance through the different groups, and thus the EB assumption is violated. Here, the URE estimator still performs well, getting within 10% of the oracle as soon as $J = 400$. However, the EBMLE shows MSE significantly higher then the URE, and has about twice the MSE when $J = 10^3$. The blue line corresponds to the oracle risk when the correlation structure is ignored and thus restricts the hyperprameter space to $\mathcal{L} = \{\text{diag}(\lambda_1, \ldots, \lambda_4) : \lambda_t \geq 0\}$.[23] Ignoring the possible correlation inflates the MSE by around 50%, showing the importance of taking such information into account.

In the last DGP, there are covariates $X_{jt} \in \mathbf{R}^2$ drawn from a uniform distribution that affects both θ_j and Σ_j. Hence, here the mean and variance are dependent through the covariates, which again violates the EB assumption. The mean and variance are set as $\theta_{jt} = X_{jt}'\beta + U_{jt}$ and $\Sigma_j = D_j \Sigma_0 D_j$ with $U_{jt} \overset{\text{i.i.d.}}{\sim} \text{Unif}[0, .3]$ and $D_j = \text{diag}(X_{j1}'\gamma, \ldots, X_{jT}'\gamma)$. Here, the URE estimator still performs fairly well, with MSE not exceeding the oracle by more than 60% for even smaller sample sizes and getting within 10% of the oracle when $J = 600$. However, the EBMLE shows poor

[23] The mean vectors in the other scenarios are independent across time, and thus this blue line is not included in the corresponding plots.



(a) URE vs EBMLE

(b) URE vs Oracle

Figure 2: Results from all scenarios. The two plots aggregate the simulation results from all DGPs considered. The left figure is a scatter plot of the MSE of the URE estimator against that of the EBMLE. The color of the dots show the sample size, from $J = 100$ to $J = 10^3$ in increments of 100, with lighter indicating smaller sample size. The right figure is a scatter plot of the MSE of the URE estimator against the oracle MSE.

performance. It shows MSE twice as large as the MSE of the URE estimator for $J = 100$, and is three times larger for larger J. Moreover, if the estimator $\widehat{\theta}^{\text{URE,cov}}$ is used to incorporate the covariates, the risk can be reduced by more than 60% compared to $\widehat{\theta}^{\text{URE,g}}$.

Finally, I show two more plots that gather the results from all DGPs I have considered. Figure 2a is a scatter plot of the MSE of the URE estimator against the MSE of the EBMLE. The plot shows that while there are some cases where the URE estimator shows slightly higher MSE when the sample size is smaller, the difference vanishes as the sample size gets larger. The majority of the dots lie on top of the 45-degree line, implying that the MSE of the URE estimator is smaller across most scenarios. Figure 2b is a similar scatter plot but now the y-axis is the oracle risk rather then the EBMLE risk. By definition of the oracle risk, there can be no dots on top of the 45-degree line. While the lighter blue dots are sometimes a little bit away from the 45-degree line, as the sample size grows larger (i.e., the dots get darker) they

get close to the 45-degree line. In particular, all dots that correspond to $J = 10^3$ are positioned fairly close to the 45-degree line, showing that the MSE of the URE gets close to the oracle across all scenarios.

7 An application to teacher value-added

In this section, I use the proposed methods to estimate the teacher effects on student achievement (i.e., teacher value-added) in the public schools of New York City (NYC). The literature on teacher value-added have used shrinkage methods extensively due to the nature of the data in this setting–presence of many teachers but only a moderate number of students per teacher–that leads to noisy measures of teacher fixed effects. For a thorough review on the topic, readers are referred to Koedel et al. (2015). I show that allowing value-added to vary with time and using the URE estimators (and forecasts) give significantly different empirical results compared to the conventional approach.

7.1 Baseline model and data

I use a standard teacher value-added model specified as the following simple linear panel data model introduced in (1):

$$y_{ijt} = X'_{ijt}\beta + \alpha_{jt} + \varepsilon_{ijt}, \tag{10}$$

where y_{ijt} is the (standardized) test score in either english language arts (ELA) or math and $X_{ijt} \in \mathbf{R}^{10}$ is a vector of student characteristics that includes: previous year's test score, gender, ethnicity, special education status (SWD), english language learner status (ELL), and eligibility for free or reduced price lunch (FL). The results are not sensitive to which covariates are added and/or interacted with other covariates as long as previous year's test scores are included. The only main difference from the standard models in the literature is the additional t subscript on the teacher fixed effect α_{jt}, which allows the teacher fixed effects to vary with time. The idiosyncratic error term ε_{ijt} is i.i.d. across i, j, and t with variance σ^2, and is independent with all other times on the right-hand side of (10).

To estimate the value-added model, I use administrative data on all public schools of NYC between academic years 2012/2013 and 2018/2019. The data for the 2012/2013

academic year is used only to extract the information on the students' test score for the previous year, and thus I have $T = 6$. Importantly, the data includes information on, among others, student-teacher linkage. As in Bitler et al. (2019), attention is restricted to 4th and 5th grade students because they are required to take the ELA (and math) test, and it is easy to link a single teacher to each student for elementary school students. I carry out the analysis using ELA test scores, but using the math scores gives similar results. Finally, I restrict the sample to those students whose ELA teachers were present in all six years of the data. The final data includes $J = 1,185$ teachers and $174,239$ student-year observations.

Following standard practice in the literature, the coefficient vector is estimated using a fixed effects estimator, with the only difference being the level at which the fixed effects are specified. The signs and magnitudes of each component of the estimate are in line with the results found in the literature (e.g., Koedel et al. (2015) and Bitler et al. (2019)). The estimation results are reported in Table 1 of Appendix D.1.

7.2 Some observations from the least squares estimator

The least squares estimator for the fixed effect, $\widehat{\alpha}_{jt}$, is the mean of the residuals corresponding to teacher j and year t. I estimate the variance of the least squares estimator by $\widehat{\sigma}^2/n_{jt}$ where $\widehat{\sigma}^2$ is the usual estimator for the variance term obtained by dividing the sum of squared residuals by the appropriate degrees of freedom, with a precise definition given in Appendix D.1. In the vast majority of the literature, teacher value-added is assumed to be time-invariant, and the least squares estimator for teacher j in this context can be written as $\widehat{\alpha}_{j0} = \frac{1}{n_j}\sum_{t=1}^{T} n_{jt}\widehat{\alpha}_{jt}$.[24] I make two preliminary observations regarding the least squares estimators that illustrate 1) there is significant time variation in the fixed effects and that 2) EB methods are unlikely to be optimal in the present setting.

The variation of the least squares estimators within teacher is large, hinting that value-added may vary significantly with time. To see this, I decompose the total variation of the least squares estimator as the variation of within teacher and across

[24]Strictly speaking, $\widehat{\alpha}_j$ is not the least squares estimator the literature has been using, because β here is estimated with fixed effects specified at the teacher-year level, not at the usual teacher level. The results presented in this section is not sensitive to this difference with the added advantage of less notation.

teachers:

$$\frac{1}{JT}\sum_{j=1}^{J}\sum_{t=1}^{T}(\widehat{\alpha}_{jt}-\overline{\widehat{\alpha}})^2=\frac{1}{J}\sum_{j=1}^{J}\left(\frac{1}{T}\sum_{t=1}^{T}(\widehat{\alpha}_{jt}-\overline{\widehat{\alpha}}_{j\cdot})^2\right)+\frac{1}{J}\sum_{j=1}^{J}(\overline{\widehat{\alpha}}_{j\cdot}-\overline{\widehat{\alpha}})^2 \tag{11}$$

where $\overline{\widehat{\alpha}}_{j\cdot}:=\frac{1}{T}\sum_{t=1}^{T}\widehat{\alpha}_{jt}$ and $\overline{\widehat{\alpha}}=\frac{1}{JT}\sum_{j=1}^{J}\sum_{t=1}^{T}\widehat{\alpha}_{jt}$ is the average of the least squares estimator at the teacher level and across all teachers, respectively. The first term on the left-hand side can be interpreted as the average variation across time and the second term as the variation across teachers. Calculations show that the average variation across time accounts for about 51% of the total variation. This implies there may be significant time variation in the fixed effects, and thus allowing for the value-added to vary with time can be a more reasonable specification.[25]

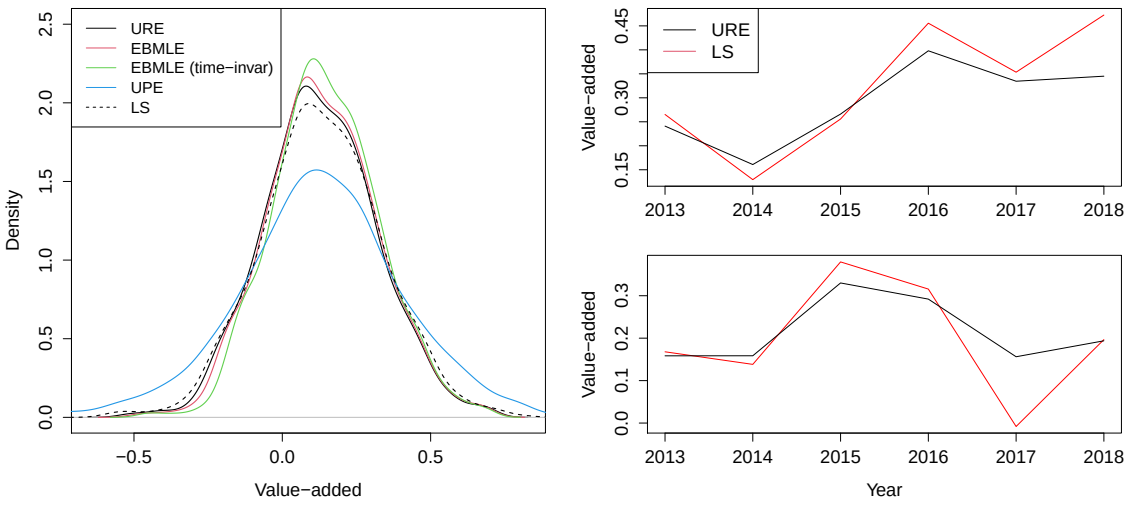
The average number of students per teacher in a single year is around 24.5, with standard deviation approximately as large as 11.7. This large variation in the number of students per teacher translates to a large degree of heteroskedasticity of the least squares estimators, which is one of the reasons that EB methods can be suboptimal (in frequentist sense). Moreover, an OLS regression of the least squares estimator ($\widehat{\alpha}_{jt}$) on the corresponding cell size (n_{jt}) show that there is a significant positive relation between the two variables. This implies that the there may be a dependence structure between the variance of the least squares estimator and the true fixed effect, which is another potential violation of the EB assumptions.

7.3 Estimation results and policy exercise

Figure 3a shows the distribution of teacher value-added estimates using four different estimators: the conventional estimator (EBMLE that assumes that value-added does not vary with time; green), the EBMLE (red) and URE (black) estimators under time-varying value-added, and the optimal UPE forecast (blue) based on the UPE.[26] For the URE and EBMLE estimators under time-varying fixed effects, the average over time within a teacher is used as a summary of the teacher's value-added, and the density plot is for this average rather than the estimate for each time period. Compared to the least squares estimator (black dashed line), the density of the three

[25] To my knowledge, the paper by Chetty et al. (2014a) is the only one to allow for time-varying teacher value-added. The recent analysis by Bitler et al. (2019) implies that allowing for time-variation can be important.

[26] The definition of each estimator is given in Appendix D.2.



(a) Density plots for shrinkage estimates (b) Shrinkage patterns for sample "teachers"

Figure 3: Shrinkage results. The plot on the left shows the density of the four different estimators discussed. The one on the right shows the time trajectory of the average of value-added estimates for a group of teachers.

estimators excluding the UPE forecast are all more concentrated at the mode, due to the shrinkage. The density plots show that there is a notable difference between the conventional method and the estimators that allow for time drifts. Not allowing for time-varying makes the estimates even more concentrated, and thus the conventional method gives a distribution more concentrated at the mode.

Moreover, the forecasts generated by minimizing the UPE are considerably more disperse than any other estimators. This is expected because unlike the other estimators, there is no averaging step in for the forecasts. Note that under the assumption of the time-invariant fixed effects, the distribution for the forecasts is necessarily the same as the distribution of fixed effect estimates, because there is no difference between a teacher's current or future fixed effect. However, the significant difference between the blue line and the other lines show that predicting a teacher's future value-added by only considering past value-added can be misleading.

Figure 3b shows how the URE estimator shrinks the least squares estimator.[27]

[27]Due to data confidentiality issues, both the least squares estimator and the URE estimator are averages across a number of teachers. However, the shrinkage pattern is the same for individual teachers.

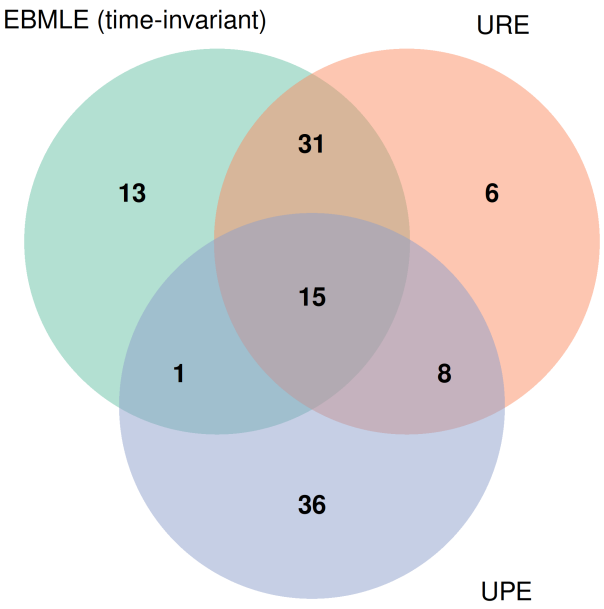


Figure 4: Composition of the bottom 5% teachers under different estimators.

While I use the estimator that shrinks toward a general location, the optimal general location turns out to be close to zero, and thus the URE estimator can be thought to shrink the least squares toward an imaginary horizontal line at zero. As is clear from the plots, the URE estimator does not necessarily shrink each component to zero, but shrinks a smoothed version of the trajectory toward zero.[28] The optimal tuning parameter $\widehat{\Lambda}^{\text{URE}}$ has positive off-diagonal terms, which is in line with positive serial correlation of the true fixed effects. Hence, the estimates for those years with more extreme values gets shrink towards the common trend making the entire trajectory smoother. For example, the least squares estimate for 2016 in the upper plot gets decreased to while that for 2014 gets increased. In contrast, if one does not take into consideration the possible serial correlation, then the URE estimator shrinks the least squares estimator toward zero for each time period, resulting in potential over-shrinkage. This demonstrates the importance of allowing for serial correlation.

A common policy exercise in the literature (Hanushek (2011), Chetty et al. (2014b), and Gilraine et al. (2020)) is to replace the teachers in the bottom 5% in the value-added distribution with an average teacher. I revisit this policy exercise with a focus on how the composition of the bottom 5% teachers changes depending on the choice of

[28]Nonetheless, the URE estimator is still a shrinkage estimator in the sense that the Euclidean norm of the estimator is smaller than the least squares estimator.

the estimator. Figure 4 shows the Venn diagram of the sets of the 60 teachers released under three different choices of estimators: the conventional time-invariant EBMLE, URE, and UPE forecasts. By using the URE estimator instead of the conventional estimator, the composition of the released teachers change by around 24% (14 teachers). Hence, allowing teacher value-added to vary with time significantly changes the composition of the group of teachers to be released. In contrast, Gilraine et al. (2020) find that using a flexible nonparametric EB method (under the assumption of time-invariant value-added) have little effect in the composition of the released teachers. Hence, allowing time drifts indeed seem to be the driving factor of such change.

In policy settings where future performance of the teachers is more relevant than the past performance, it is natural to base the decision on forecasts. For example, if the interest is in maximizing student outcome in the following year, forecasts for the next period teacher value-added is more informative than a summary of past performance. When the value-added is allowed to vary with time, one can use the optimal UPE forecasts in such context. On the other hand, if one specifies value-added to be time-invariant, past and future value-added are the same by definition, and thus will release the bottom 5% according to the conventional estimator. The Venn diagram shows that whether the fixed effects are allowed to vary with time or not changes the composition of the bottom 5% teachers dramatically, with only 16 teachers (approximately 26%) belonging to this group under both estimators.

8 Conclusion

I develop new shrinkage estimators for the fixed effects in linear panel data models–the URE estimators. The fixed effects are allowed to vary with time and to be serially correlated. The estimators are obtained by shrinking the least squares estimators, where the direction and magnitude of shrinkage is determined by minimizing an estimate of the risk. They are shown to (asymptotically) dominate conventional estimators under mild regularity conditions, and does not rely on strong distributional assumptions as conventional methods do.

While I focus on estimating the fixed effects in a linear panel setting, I emphasize that the URE method can be applied to any setting where the empirical researcher has an approximately unbiased estimator for individual/group-level effects. Such examples include Angrist et al. (2017) and Hull (2020). While the models considered

therein are not strictly linear panel models, the researchers derive an unbiased estimator for the group (school or hospital) effects. Such group effects can be shrunk by the methods introduced here rather then using EB methods to guard against stronger distributional assumptions.

A natural direction for future work is to make the class of estimators wider without losing tractability in terms of both theory and computation. In Appendix E, I show how one can extend the semiparametric shrinkage idea of Xie et al. (2012) to this setting. While the theory is straightforward for this extension, computation is extremely difficult and thus additional restrictions are necessary. Another possible method is to consider the class of estimators implied by the nonparametric EB setting, and taking an URE approach to tune the unknown (nonparametric) distribution of the true fixed effect. This an open problem that is yet to be solved even in the case where $T = 1$. Such extensions will make the optimality result to hold over a significantly wider class of estimators, further improving the risk property of the URE estimators.

References

Abadie, A. and M. Kasy (2019): "Choosing Among Regularized Estimators in Empirical Economics: The Risk of Machine Learning," *The Review of Economics and Statistics*, 101, 743–762.

Abowd, J. M., F. Kramarz, and D. N. Margolis (1999): "High Wage Workers and High Wage Firms," *Econometrica*, 67, 251–333.

Andrews, D. W. (1992): "Generic Uniform Convergence," *Econometric Theory*, 8, 241–257.

Angrist, J. D., P. D. Hull, P. A. Pathak, and C. R. Walters (2017): "Leveraging Lotteries for School Value-Added: Testing and Estimation," *The Quarterly Journal of Economics*, 132, 871–919.

Armstrong, T. B., M. Kolesár, and M. Plagborg-Møller (2020): "Robust Empirical Bayes Confidence Intervals," *arXiv:2004.03448 [econ, stat]*.

Bitler, M., S. Corcoran, T. Domina, and E. Penner (2019): "Teacher Effects on Student Achievement and Height: A Cautionary Tale," Tech. Rep. w26480, National Bureau of Economic Research, Cambridge, MA.

Bonhomme, S. and M. Weidner (2019): "Posterior Average Effects," *arXiv:1906.06360 [econ, stat]*.

Brown, L. D. and E. Greenshtein (2009): "Nonparametric Empirical Bayes and Compound Decision Approaches to Estimation of a High-Dimensional Vector of Normal Means," *The Annals of Statistics*, 37, 1685–1704.

Brown, L. D., G. Mukherjee, and A. Weinstein (2018): "Empirical Bayes Estimates for a Two-Way Cross-Classified Model," *The Annals of Statistics*, 46, 1693–1720.

Chetty, R., J. N. Friedman, and J. E. Rockoff (2014a): "Measuring the Impacts of Teachers I: Evaluating Bias in Teacher Value-Added Estimates," *American Economic Review*, 104, 2593–2632.

——— (2014b): "Measuring the Impacts of Teachers II: Teacher Value-Added and Student Outcomes in Adulthood," *American Economic Review*, 104, 2633–2679.

Chetty, R. and N. Hendren (2018): "The Impacts of Neighborhoods on Intergenerational Mobility II: County-Level Estimates," *The Quarterly Journal of Economics*, 133, 1163–1228.

Fessler, P. and M. Kasy (2019): "How to Use Economic Theory to Improve Estimators: Shrinking Toward Theoretical Restrictions," *The Review of Economics and Statistics*, 101, 681–698.

Frandsen, B., L. Lefgren, and E. Leslie (2019): "Judging Judge Fixed Effects," Tech. Rep. w25528, National Bureau of Economic Research, Cambridge, MA.

Gilraine, M., J. Gu, and R. McMillan (2020): "A New Method for Estimating Teacher Value-Added," Tech. Rep. w27094, National Bureau of Economic Research, Cambridge, MA.

Guarino, C. M., M. Maxfield, M. D. Reckase, P. N. Thompson, and J. M. Wooldridge (2015): "An Evaluation of Empirical Bayes's Estimation of Value-Added Teacher Performance Measures," *Journal of Educational and Behavioral Statistics*, 40, 190–222.

Hansen, B. E. (2016): "Efficient Shrinkage in Parametric Models," *Journal of Econometrics*, 190, 115–132.

Hanushek, E. A. (2011): "The Economic Value of Higher Teacher Quality," *Economics of Education Review*, 30, 466–479.

Horn, R. A. and C. R. Johnson (1990): *Matrix Analysis*, Cambridge University Press.

Hull, P. (2020): "Estimating Hospital Quality with Quasi-Experimental Data," *working paper*.

Ignatiadis, N. and S. Wager (2020): "Covariate-Powered Empirical Bayes Estimation," *arXiv:1906.01611 [stat]*.

James, W. and C. Stein (1961): "Estimation with Quadratic Loss," in *Proceedings of the Fourth Berkeley Symposium on Mathematical Statistics and Probability, Volume 1: Contributions to the Theory of Statistics*, The Regents of the University of California.

Jiang, W. and C.-H. Zhang (2009): "General Maximum Likelihood Empirical Bayes Estimation of Normal Means," *The Annals of Statistics*, 37, 1647–1684.

Kane, T. J., J. E. Rockoff, and D. O. Staiger (2008): "What Does Certification Tell Us about Teacher Effectiveness? Evidence from New York City," *Economics of Education Review*, 27, 615–631.

Koedel, C., K. Mihaly, and J. E. Rockoff (2015): "Value-Added Modeling: A Review," *Economics of Education Review*, 47, 180–195.

Koenker, R. and I. Mizera (2014): "Convex Optimization, Shape Constraints, Compound Decisions, and Empirical Bayes Rules," *Journal of the American Statistical Association*, 109, 674–685.

Kong, X., Z. Liu, P. Zhao, and W. Zhou (2017): "SURE Estimates under Dependence and Heteroscedasticity," *Journal of Multivariate Analysis*, 161, 1–11.

Kou, S. C. and J. J. Yang (2017): "Optimal Shrinkage Estimation in Heteroscedastic Hierarchical Linear Models," in *Big and Complex Data Analysis*, ed. by S. E. Ahmed, Cham: Springer International Publishing, 249–284.

Li, K.-C. (1985): "From Stein's Unbiased Risk Estimates to the Method of Generalized Cross Validation," *The Annals of Statistics*, 13, 1352–1377.

——— (1986): "Asymptotic Optimality of C_L and Generalized Cross-Validation in Ridge Regression with Application to Spline Smoothing," *The Annals of Statistics*, 14, 1101–1112.

Liu, L., H. R. Moon, and F. Schorfheide (2020): "Forecasting With Dynamic Panel Data Models," *Econometrica*, 88, 171–201.

Okolewski, A. and T. Rychlik (2001): "Sharp Distribution-Free Bounds on the Bias in Estimating Quantiles via Order Statistics," *Statistics & Probability Letters*, 52, 207–213.

Robbins, H. (1951): "Asymptotically Subminimax Solutions of Compound Statistical Decision Problems," in *Proceedings of the Second Berkeley Symposium on Mathematical Statistics and Probability*, The Regents of the University of California.

——— (1964): "The Empirical Bayes Approach to Statistical Decision Problems," *Annals of Mathematical Statistics*, 35, 1–20.

Rockoff, J. E. (2004): "The Impact of Individual Teachers on Student Achievement: Evidence from Panel Data," *American Economic Review*, 94, 247–252.

Rothstein, J. (2010): "Teacher Quality in Educational Production: Tracking, Decay, and Student Achievement," *The Quarterly Journal of Economics*, 125, 175–214.

Stein, C. (1956): "Inadmissibility of the Usual Estimator for the Mean of a Multivariate Normal Distribution," in *Proceedings of the Third Berkeley Symposium on Mathematical Statistics and Probability, Volume 1: Contributions to the Theory of Statistics*, The Regents of the University of California.

van der Vaart, A. W. (1998): *Asymptotic Statistics*, Cambridge Series in Statistical and Probabilistic Mathematics, Cambridge: Cambridge University Press.

Wooldridge, J. M. (2010): *Econometric Analysis of Cross Section and Panel Data*, MIT Press.

Xie, X., S. C. Kou, and L. Brown (2016): "Optimal Shrinkage Estimation of Mean Parameters in Family of Distributions with Quadratic Variance," *The Annals of Statistics*, 44, 564–597.

Xie, X., S. C. Kou, and L. D. Brown (2012): "SURE Estimates for a Heteroscedastic Hierarchical Model," *Journal of the American Statistical Association*, 107, 1465–1479.

Appendix A Proof of main theorems

A.1 Proof of Theorem 4.1

The difference between the URE and the loss is given as

$$\begin{aligned}
&\mathrm{URE}(\mu,\Lambda)-\ell(\theta,\widehat{\theta}(\mu,\Lambda))\\
=&\frac{1}{J}\sum_{j=1}^{J}\left(\mathrm{URE}_j(\mu,\Lambda)-(\widehat{\theta}_j(\mu,\Lambda)-\theta_j)'(\widehat{\theta}_j(\mu,\Lambda)-\theta_j)\right).
\end{aligned}$$

Expanding the summand of the second line gives

$$\begin{aligned}
&\mathrm{URE}_j(\mu,\Lambda)-(\widehat{\theta}_j(\mu,\Lambda)-\theta_j)'(\widehat{\theta}_j(\mu,\Lambda)-\theta_j)\\
=&\operatorname{tr}(\Sigma_j)-2\operatorname{tr}((\Lambda+\Sigma_j)^{-1}\Sigma_j^2)\\
&+(y_j-\mu)'[(\Lambda+\Sigma_j)^{-1}\Sigma_j^2(\Lambda+\Sigma_j)^{-1}](y_j-\mu)\\
&-(y_j-\theta_j-\Sigma_j(\Lambda+\Sigma_j)^{-1}(y_j-\mu))'(y_j-\theta_j-\Sigma_j(\Lambda+\Sigma_j)^{-1}(y_j-\mu))\\
=&\operatorname{tr}(\Sigma_j)-2\operatorname{tr}((\Lambda+\Sigma_j)^{-1}\Sigma_j^2)-(y_j-\theta_j)'(y_j-\theta_j)\\
&+2(y_j-\mu)'(\Lambda+\Sigma_j)^{-1}\Sigma_j(y_j-\theta_j)\\
=&y_j'y_j-\theta_j'\theta_j-\operatorname{tr}(\Sigma_j)-2\operatorname{tr}(\Lambda(\Lambda+\Sigma_j)^{-1}(y_jy_j'-y_j\theta_j'-\Sigma_j))\\
&-2\mu'(\Lambda+\Sigma_j)^{-1}\Sigma_j(y_j-\theta_j).
\end{aligned}\tag{12}$$

Taking $\mu=0$, I obtain

$$\begin{aligned}
&\mathrm{URE}_j(\Lambda)-(\widehat{\theta}_j(\Lambda)-\theta_j)'(\widehat{\theta}_j(\Lambda)-\theta_j)\\
=&y_j'y_j-\theta_j'\theta_j-\operatorname{tr}(\Sigma_j)-2\operatorname{tr}(\Lambda(\Lambda+\Sigma_j)^{-1}(y_jy_j'-y_j\theta_j'-\Sigma_j)),
\end{aligned}$$

where, for simplicity, I write $\mathrm{URE}_j(\Lambda)$ as a shorthand for $\mathrm{URE}_j(0,\Lambda)$, and likewise for $\mathrm{URE}(\Lambda)$ and $\widehat{\theta}(\Lambda)$. It follows that

$$\begin{aligned}
\sup_{\Lambda}\left|\mathrm{URE}(\Lambda)-\ell(\theta,\widehat{\theta}(\Lambda))\right|=&\sup_{\Lambda}\left|\frac{1}{J}\sum\nolimits_{j=1}^{J}(\mathrm{URE}_j(\Lambda)-\ell_j(\theta_j,\widehat{\theta}_j(\Lambda)))\right|\\
\le&\left|\frac{1}{J}\sum\nolimits_{j=1}^{J}(y_j'y_j-\theta_j'\theta_j-\operatorname{tr}(\Sigma_j))\right|\\
&+\sup_{\Lambda}\left|\frac{1}{J}\sum\nolimits_{j=1}^{J}\operatorname{tr}(\Lambda(\Lambda+\Sigma_j)^{-1}(y_jy_j'-y_j\theta_j'-\Sigma_j))\right|,
\end{aligned}$$

where the inequality follows from the triangle inequality. I show that each of the two terms in the last expression converges to zero in L^1.

For the first term, because $\mathbf{E}\, y_j' y_j = \theta_j'\theta_j + \mathrm{tr}(\Sigma_j)$ for all $j \leq J$ and y_j's are independent, we have

$$\begin{aligned}\mathbf{E}\left(\frac{1}{J}\sum_{j=1}^{J}(y_j'y_j - \theta_j'\theta_j - \mathrm{tr}(\Sigma_j))\right)^2 &= \frac{1}{J^2}\sum_{j=1}^{J}\mathbf{E}(y_j'y_j - \theta_j'\theta_j - \mathrm{tr}(\Sigma_j))^2 \\ &= \frac{1}{J^2}\sum_{j=1}^{J}\mathrm{var}(y_j'y_j).\end{aligned}$$

Therefore, if $\lim \frac{1}{J^2}\sum_{j=1}^{J}\mathrm{var}(y_j'y_j) = 0$, then this term converges to zero in L^2 and thus in L^1. Assumption 4.1(ii) ensures that this is the case.

For the second term, note that

$$\begin{aligned}&\sup_{\Lambda}\left|\frac{1}{J}\sum_{j=1}^{J}\mathrm{tr}(\Lambda(\Lambda+\Sigma_j)^{-1}(y_jy_j' - y_j\theta_j' - \Sigma_j))\right| \\ =&\sup_{\Lambda}\left|\frac{1}{J}\sum_{j=1}^{J}\mathrm{tr}((I-\Sigma_j(\Lambda+\Sigma_j)^{-1})(y_jy_j' - y_j\theta_j' - \Sigma_j))\right| \\ \leq&\left|\frac{1}{J}\sum_{j=1}^{J}(y_j'y_j - \theta_j'y_j - \mathrm{tr}(\Sigma_j))\right| + \sup_{\Lambda}\left|\frac{1}{J}\sum_{j=1}^{J}\mathrm{tr}(\Sigma_j(\Lambda+\Sigma_j)^{-1}(y_jy_j' - y_j\theta_j' - \Sigma_j))\right| \\ =:&\,(\mathrm{I})_J + (\mathrm{II})_J.\end{aligned}$$

To show that $(\mathrm{I})_J \xrightarrow{L^1} 0$, I again show L^2 convergence. Because $\mathbf{E}(y_j'y_j - \theta_j'y_j) = \mathrm{tr}(\Sigma_j)$ for all $j \leq J$ and y_j's are independent, it follows that

$$\begin{aligned}\mathbf{E}\left(\frac{1}{J}\sum_{j=1}^{J}(y_j'y_j - \theta_j'y_j - \mathrm{tr}(\Sigma_j))\right)^2 =&\frac{1}{J^2}\sum_{j=1}^{J}\mathbf{E}(y_j'y_j - \theta_j'y_j - \mathrm{tr}(\Sigma_j))^2 \\ =&\frac{1}{J^2}\sum_{j=1}^{J}\mathrm{var}(y_j'y_j - \theta_j'y_j).\end{aligned}$$

Hence, it suffices to establish that $\lim_{J\to\infty}\frac{1}{J^2}\sum_{j=1}^{J}\mathrm{var}(y_j'y_j - \theta_j'y_j) = 0$. The summand

is bounded by

$$\operatorname{var}(y_j'y_j - \theta_j'y_j) \leq 2\operatorname{var}(y_j'y_j) + 2\theta_j'\Sigma_j\theta_j \leq 2\operatorname{var}(y_j'y_j) + 2\operatorname{tr}(\Sigma_j)\|\theta_j\|_\infty^2.$$

Hence, if $\limsup_{J\to\infty}\frac{1}{J}\sum_{j=1}^{J}(\operatorname{var}(y_j'y_j)+\operatorname{tr}(\Sigma_j)\|\theta_j\|_\infty^2) < \infty$ it follows that $(\mathrm{I})_J \xrightarrow{L^2} 0$. A sufficient condition for this to hold is that $\sup_j \operatorname{var}(y_j'y_j)$, $\sup_j \operatorname{tr}(\Sigma_j)$, and $\sup_j\|\theta_j\|_\infty^2$ are all finite, which is true by Assumption 4.1 (ii).

To show that $(\mathrm{II})_J \xrightarrow{L^1} 0$, define the random function $G_J(\Lambda)$ as

$$G_J(\Lambda) = \frac{1}{J}\sum_{j=1}^{J}\operatorname{tr}(\Sigma_j(\Lambda+\Sigma_j)^{-1}(y_jy_j' - y_j\theta_j' - \Sigma_j))$$

so that the aim is to show $\sup_\Lambda|G_J(\Lambda)| \xrightarrow{L^1} 0$. I use the fact that convergence in probability and a uniform integrability condition imply convergence in L^1. That is, I show $\sup_\Lambda|G_J(\Lambda)| \xrightarrow{p} 0$ and that $\{\sup_\Lambda|G_J(\Lambda)|\}_{J\geq 1}$ is uniformly integrable.

I show $\sup_\Lambda|G_J(\Lambda)| \xrightarrow{p} 0$ by using the results given by Andrews (1992).The results therein require a totally bounded parameter space. However, the parameter space in consideration, $\mathcal{S}_T^+$, does not satisfy this requirement. This can be dealt with by an appropriate reparametrization. Let $\underline{\sigma}_\Sigma = \inf_j \sigma_T(\Sigma_j)$ denote the infimum of the smallest eigenvalues of Σ_j's for $j \geq 1$, which is bounded away from zero by assumption. Consider the transformation defined by $h(\Lambda) = (\underline{\sigma}_\Sigma I_T + \Lambda)^{-1}$, and write the image of such transformation as $\widetilde{\mathcal{L}} := \{h(\Lambda) : \Lambda \in S_T^+\}$. Note that $h : \mathcal{S}_T^+ \to \widetilde{\mathcal{L}}$ is one-to-one and onto, with its inverse given as $h^{-1}(\widetilde{\Lambda}) = \widetilde{\Lambda}^{-1} - \underline{\sigma}_\Sigma I_T$. For $\widetilde{\Lambda} \in \widetilde{\mathcal{L}}$, define $\widetilde{G}_J := G_J \circ h^{-1}$ so that

$$\sup_{\Lambda\in\mathcal{S}_T^+}|G_J(\Lambda)| = \sup_{\Lambda\in\mathcal{S}_T^+}|G_J(h^{-1}(h(\Lambda)))| = \sup_{\widetilde{\Lambda}\in\widetilde{\mathcal{L}}}|G_J(h^{-1}(\widetilde{\Lambda}))| = \sup_{\widetilde{\Lambda}\in\widetilde{\mathcal{L}}}|\widetilde{G}_J(\widetilde{\Lambda})|.$$

Hence, showing $\sup_\Lambda|G_J(\Lambda)| \xrightarrow{p} 0$ is equivalent to $\sup_{\widetilde{\Lambda}\in\widetilde{\mathcal{L}}}|\widetilde{G}_J(\widetilde{\Lambda})| \xrightarrow{p} 0$. Let $\mathcal{S}_T$ denote the set of all real positive $T \times T$ matrices. While the choice of metric is irrelevant, equip $\mathcal{S}_T$ with the metric d induced by the Frobenius norm for concreteness. Note that $\widetilde{\mathcal{L}} \subset S_T$. I show that the (reparametrized) parameter space $\widetilde{\mathcal{L}}$ is indeed totally bounded. For any $\widetilde{\Lambda} \in \widetilde{\mathcal{L}}$, I have $0 \leq \widetilde{\Lambda} \leq \underline{\sigma}_\Sigma^{-1} I_T$ so that $\sigma_1(\widetilde{\Lambda}) \leq \underline{\sigma}_\Sigma^{-1}$. Moreover, since the largest singular value equals the operator norm and all norms on S_T are equivalent, this shows that $\widetilde{\mathcal{L}}$ is bounded, and thus totally bounded because

$\widetilde{\mathcal{L}}$ can be seen as a subset of the Euclidean space with dimension T^2.

It remains to show that a) $|\widetilde{G}_J(\widetilde{\Lambda})| \xrightarrow{p} 0$ for all $\widetilde{\Lambda} \in \widetilde{\mathcal{L}}$ and b) $\widetilde{G}_J(\widetilde{\Lambda})$ is stochastically equicontinuous. For a), we can show $|G_J(\Lambda)| \xrightarrow{p} 0$ for all $\Lambda \in \mathcal{S}_T^+$ instead because for any $\widetilde{\Lambda} \in \widetilde{\mathcal{L}}$, there exists $\Lambda \in \mathcal{S}_T^+$ such that $G_J(\Lambda) = \widetilde{G}_J(\widetilde{\Lambda})$. Now, note that

$$\begin{aligned}&\mathbf{E}\,\mathrm{tr}(\Sigma_j(\Lambda+\Sigma_j)^{-1}(y_jy_j' - y_j\theta_j' - \Sigma_j))\\ =\,&\mathrm{tr}(\Sigma_j(\Lambda+\Sigma_j)^{-1}\,\mathbf{E}(y_jy_j' - y_j\theta_j' - \Sigma_j)) = 0,\end{aligned}$$

and y_j's are independent. This gives

$$\mathbf{E}\,G_J(\Lambda)^2 = \frac{1}{J^2}\sum_{j=1}^{J}\mathbf{E}\,\mathrm{tr}(\Sigma_j(\Lambda+\Sigma_j)^{-1}(y_jy_j' - y_j\theta_j' - \Sigma_j))^2$$

I give a bound on $|\mathrm{tr}(\Sigma_j(\Lambda+\Sigma_j)^{-1}(y_jy_j' - y_j\theta_j' - \Sigma_j))|$. Let UDU' denote the spectral decomposition of $\Sigma_j^{-1/2}\Lambda\Sigma_j^{-1/2}$ with $D = \mathrm{diag}(d_1, \ldots, d_T)$. Then, I have

$$\Sigma_j(\Lambda+\Sigma_j)^{-1} = \Sigma_j^{1/2}U(I_T+D)^{-1}U'\Sigma_j^{-1/2}.$$

It follows that

$$\begin{aligned}&\mathrm{tr}(\Sigma_j(\Lambda+\Sigma_j)^{-1}(y_jy_j' - y_j\theta_j' - \Sigma_j))\\ =\,&\mathrm{tr}(\Sigma_j^{1/2}U(I_T+D)^{-1}U'\Sigma_j^{-1/2}(y_jy_j' - y_j\theta_j' - \Sigma_j))\\ =\,&\mathrm{tr}((I_T+D)^{-1}U'\Sigma_j^{-1/2}(y_jy_j' - \theta_jy_j' - \Sigma_j)\Sigma_j^{1/2}U).\end{aligned} \tag{13}$$

Write $H_j = \Sigma_j^{-1/2}(y_jy_j' - y_j\theta_j' - \Sigma_j)\Sigma_j^{1/2}$, and observe that

$$\begin{aligned}&\left|\mathrm{tr}((I_T+D)^{-1}U'\Sigma_j^{-1/2}(y_jy_j' - y_j\theta_j' - \Sigma_j)\Sigma_j^{1/2}U)\right|\\ =\,&\left|\sum_{t=1}^{T}\frac{1}{1+d_t}(U'H_jU)_{tt}\right|\\ \le\,&\sum_{t=1}^{T}\frac{1}{1+d_t}\left|(U'H_jU)_{tt}\right|\\ \le\,&\sum_{t=1}^{T}\left|(U'H_jU)_{tt}\right|,\end{aligned} \tag{14}$$

where the last inequality holds because $0 \le 1/(1+d_t) \le 1$. Let U_t denote the tth

column of the orthogonal matrix U. I have

$$|(U'H_jU)_{tt}| = |U_t'H_jU_t| \leq \|H_jU_t\| \leq \sup_{U\in\mathbb{R}^T, \|U\|=1} \|H_jU\| = \sigma_1(H_j),$$

where the first inequality follows from Cauchy-Schwarz, and the last equality from the fact that the operator norm of a matrix is equal to its largest singular value.

Combining these results gives

$$\mathbf{E}\, G_J(\Lambda)^2 \leq \frac{T^2}{J^2} \sum_{j=1}^{J} \mathbf{E}\ \sigma_1(H_j)^2.$$

Now, to derive a bound for $\sigma_1(H_j)$, observe that

$$\begin{aligned}
\sigma_1(H_j) =& \sigma_1(\Sigma_j^{-1/2}(y_jy_j' - y_j\theta_j' - \Sigma_j)\Sigma_j^{1/2}) \\
\leq& \sigma_1(\Sigma_j^{-1/2})\sigma_1(y_jy_j' - y_j\theta_j' - \Sigma_j)\sigma_1(\Sigma_j^{1/2}) \\
\leq& \kappa(\Sigma_j)^{1/2}\sigma_1(y_jy_j' - y_j\theta_j' - \Sigma_j).
\end{aligned}$$

Since the largest singular value of a matrix is bounded by its Frobenius norm, it follows that

$$\begin{aligned}
&\sigma_1(y_jy_j' - y_j\theta_j' - \Sigma_j)^2 \\
\leq& \operatorname{tr}((y_jy_j' - y_j\theta_j' - \Sigma_j)'(y_jy_j' - y_j\theta_j' - \Sigma_j)) \\
=& (y_j'y_j)^2 + \theta_j'y_j\theta_j'y_j + \sigma^4 \operatorname{tr}(\Sigma_j)^2 - 2y_j'y_jy_j'\theta_j - 2y_j'\Sigma_jy_j + 2\theta_j'\Sigma_jy_j.
\end{aligned}$$

Taking expectations yields

$$\begin{aligned}
&\mathbf{E}(y_j'y_j)^2 + \theta_j'\theta_j\, \mathbf{E}\, y_j'y_j + \sigma^4 \operatorname{tr}(\Sigma_j)^2 - 2\, \mathbf{E}\, y_j'y_jy_j'\theta_j - 2\, \mathbf{E}\, y_j'\Sigma_jy_j + 2\theta_j'\Sigma_j\, \mathbf{E}\, y_j \\
=& \operatorname{var}(y_j'y_j) + (\theta_j'\theta_j + \operatorname{tr}(\Sigma_j))^2 + \theta_j'\theta_j(\theta_j'\theta_j + \operatorname{tr}(\Sigma_j)) \\
&+ \sigma^4 \operatorname{tr}(\Sigma_j)^2 - 2\theta_j'\, \mathbf{E}(y_jy_j'y_j) - 2\theta_j'\Sigma_j\theta_j - 2\sigma^4 \operatorname{tr}(\Sigma_j)^2 + 2\theta_j'\Sigma_j\theta_j \\
=& \operatorname{var}(y_j'y_j) + 2(\theta_j'\theta_j)^2 + 3\theta_j'\theta_j \operatorname{tr}(\Sigma_j) - 2\theta_j'\, \mathbf{E}(y_jy_j'y_j) \\
\leq& \operatorname{var}(y_j'y_j) + 2\|\theta_j\|^4 + 3\|\theta_j\|^2 \operatorname{tr}(\Sigma_j) + 2\|\theta_j\|\, \mathbf{E}\|y_j\|^3.
\end{aligned}$$

This shows that if

$$\limsup_{J\to\infty}\frac{1}{J}\sum_{j=1}^{J}\kappa(\Sigma_j)\left(\operatorname{var}(y_j'y_j)+2\|\theta_j\|^4+3\|\theta_j\|^2\operatorname{tr}(\Sigma_j)+2\|\theta_j\|\,\mathbf{E}(\|y_j\|^3)\right)<\infty, \tag{15}$$

then $|G_J(\Lambda)|\to 0$ in L^2, and thus in probability. Hence, if $\sup_j \sigma_1(\Sigma_j)/\sigma_T(\Sigma_j)$, $\sup_j \operatorname{var}(y_j'y_j)$, $\sup_j|\theta_j|$, and $\sup_j \operatorname{tr}(\Sigma_j)$ are bounded, the result holds. Note that this true by Assumption 4.1.

It remains to show that $\widetilde{G}_J(\widetilde{\Lambda})$ is stochastically equicontinuous. I do this by showing that $\widetilde{G}_J(\widetilde{\Lambda})$ satisfies a Lipschitz condition as in Assumption SE-1 of Andrews (1992). Specifically, I show that $|\widetilde{G}_J(\widetilde{\Lambda})-\widetilde{G}_J(\widetilde{\Lambda}^\dagger)|\le B_J\|\widetilde{\Lambda}-\widetilde{\Lambda}^\dagger\|$ for all $\widetilde{\Lambda},\widetilde{\Lambda}^\dagger\in\widetilde{\mathcal{L}}$ with $B_J=O_p(1)$. Let $\widetilde{\Lambda},\widetilde{\Lambda}^\dagger\in\widetilde{\mathcal{L}}$ be arbitrarily taken. First, I show that $\widetilde{\mathcal{L}}$ is convex. Take any $\alpha\in[0,1]$. Note that $\alpha\widetilde{\Lambda}+(1-\alpha)\widetilde{\Lambda}^\dagger$ is nonsingular because $\widetilde{\Lambda}$ and $\widetilde{\Lambda}^\dagger$ are positive definite and the space of positive definite matrices is convex. Then, for $\Lambda_\alpha=(\alpha\widetilde{\Lambda}+(1-\alpha)\widetilde{\Lambda}^\dagger)^{-1}-\underline{\sigma}_\Sigma I_T$, we have $h(\Lambda_\alpha)=\alpha\widetilde{\Lambda}+(1-\alpha)\widetilde{\Lambda}^\dagger$, which shows that $\alpha\widetilde{\Lambda}+(1-\alpha)\widetilde{\Lambda}^\dagger\in\widetilde{\mathcal{L}}$. The mean value theorem gives

$$\widetilde{G}_J(\widetilde{\Lambda})-\widetilde{G}_J(\widetilde{\Lambda}^\dagger)=\nabla\widetilde{G}_J(\widetilde{\Lambda}^\alpha)\cdot\operatorname{vec}(\widetilde{\Lambda}-\widetilde{\Lambda}^\dagger),$$

where $\nabla\widetilde{G}_J(\widetilde{\Lambda}):=\frac{\partial}{\partial\operatorname{vec}(\widetilde{\Lambda})}\widetilde{G}_J(\widetilde{\Lambda})$, and $\widetilde{\Lambda}^\alpha:=\alpha\widetilde{\Lambda}+(1-\alpha)\widetilde{\Lambda}^\dagger$ for some $\alpha\in[0,1]$. This implies, by Cauchy-Schwarz,

$$|\widetilde{G}_J(\widetilde{\Lambda})-\widetilde{G}_J(\widetilde{\Lambda}^\dagger)|\le\|\nabla\widetilde{G}_J(\widetilde{\Lambda}^\alpha)\|\|\widetilde{\Lambda}-\widetilde{\Lambda}^\dagger\|, \tag{16}$$

where I use the fact that the Frobenius norm of a matrix and the Euclidean norm of the vectorized version of it are the same. Note that $\|\nabla\widetilde{G}_J(\widetilde{\Lambda})\|=\|\frac{\partial}{\partial\widetilde{\Lambda}}\widetilde{G}_J(\widetilde{\Lambda})\|$ by definition of the Frobenius norm.

By the formula for the derivative of a matrix inverse and the derivative of a trace, and the chain rule for matrix derivatives, I have

$$\begin{aligned}&\frac{\partial}{\partial\widetilde{\Lambda}}G_J(\widetilde{\Lambda})\\ =&\frac{1}{J}\sum_{j=1}^{J}\widetilde{\Lambda}^{-1}(\widetilde{\Lambda}^{-1}-\underline{\sigma}_\Sigma I_T+\Sigma_j)^{-1}\Sigma_j(y_jy_j'-y_j\theta_j'-\Sigma_j)(\widetilde{\Lambda}^{-1}-\underline{\sigma}_\Sigma I_T+\Sigma_j)^{-1}\widetilde{\Lambda}^{-1}.\end{aligned}$$

Write the summand in the second line as $g_j(\widetilde{\Lambda})$. I first derive a bound on $\sigma_1(g_j(\widetilde{\Lambda}))$,

and use this to bound $\|\frac{\partial}{\partial\widetilde{\Lambda}}G_J(\widetilde{\Lambda})\|$ by using the fact that

$$\left\|\frac{\partial}{\partial\widetilde{\Lambda}}G_J(\widetilde{\Lambda})\right\| \leq T^{1/2}\sigma_1\left(\frac{\partial}{\partial\widetilde{\Lambda}}G_J(\widetilde{\Lambda})\right) \leq \frac{1}{J}\sum_{j=1}^{J}\sigma_1(g_j(\widetilde{\Lambda})).$$

Since the operator norm is submultiplicative, it follows that

$$\sigma_1(g_j(\widetilde{\Lambda})) \leq \sigma_1(\widetilde{\Lambda}^{-1}(\widetilde{\Lambda}^{-1} - \underline{\sigma}_\Sigma I_T + \Sigma_j)^{-1})^2\sigma_1(\Sigma_j(y_jy_j' - y_j\theta_j' - \Sigma_j)).$$

I proceed by bounding the two singular values that appear on the right hand side. For the first term, note that

$$\begin{aligned}
&\sigma_1(\widetilde{\Lambda}^{-1}(\widetilde{\Lambda}^{-1} - \underline{\sigma}_\Sigma I_T + \Sigma_j)^{-1})^2 \\
=&\sigma_1(\widetilde{\Lambda}^{-1}(\widetilde{\Lambda}^{-1} - \underline{\sigma}_\Sigma I_T + \Sigma_j)^{-2}\widetilde{\Lambda}^{-1}) \\
=&\sigma_1((I + \widetilde{\Lambda}^{1/2}(\Sigma_j - \underline{\sigma}_\Sigma I_T)\widetilde{\Lambda}^{1/2})^{-2}) \\
\leq&1,
\end{aligned} \tag{17}$$

where the first equality uses the fact that $\sigma_1(A)^2 = \sigma_1(AA') = \sigma_1(A'A)$ for any matrix A. The last inequality follows because $\widetilde{\Lambda}^{1/2}(\Sigma_j - \underline{\sigma}_\Sigma I_T)\widetilde{\Lambda}^{1/2}$ is positive semidefinite so that $0 \leq (I + \widetilde{\Lambda}^{1/2}(\Sigma_j - \underline{\sigma}_\Sigma I_T)\widetilde{\Lambda}^{1/2})^{-2} \leq I_T$, and $A \leq B$ implies $\sigma_1(A) \leq \sigma_1(B)$ for any two positive semidefinite matrices A and B. A bound on $\sigma_1(\Sigma_j(y_jy_j' - y_j\theta_j' - \Sigma_j))$ is given by

$$\sigma_1(\Sigma_j(y_jy_j' - y_j\theta_j' - \Sigma_j)) \leq \sigma_1(\Sigma_j)(y_j'y_j + (y_j'y_j)^{1/2}(\theta_j'\theta_j)^{1/2} + \sigma_1(\Sigma_j)).$$

Combining these results, I obtain

$$\begin{aligned}
\sup_{\widetilde{\Lambda}\in\widetilde{\mathcal{L}}}\left\|\frac{\partial}{\partial\widetilde{\Lambda}}G_J(\widetilde{\Lambda})\right\| \leq&\frac{1}{J}\sum_{j=1}^{J}\sigma_1(\Sigma_j)(y_j'y_j + \|y_j\|\|\theta_j\| + \sigma_1(\Sigma_j)) \\
=&\frac{1}{J}\sum_{j=1}^{J}\sigma_1(\Sigma_j)\left(y_j'y_j + \|y_j\|\|\theta_j\| - \mathbf{E}(y_j'y_j + \|y_j\|\|\theta_j\|)\right) \\
&+\frac{1}{J}\sum_{j=1}^{J}\left(\mathbf{E}(y_j'y_j + \|y_j\|\|\theta_j\|) + \sigma_1(\Sigma_j)\right).
\end{aligned}$$

The term in the second line is $o_p(1)$ because

$$\sup_j \operatorname{var}(\sigma_1(\Sigma_j)(y_j'y_j + \|y_j\|\|\theta_j\|)) \leq 2\sup_j \sigma_1^2(\Sigma_j)(\mathbf{E}\|y_j\|^4 + \|\theta_j\|^2\,\mathbf{E}\|y_j\|^2) < \infty.$$

The term in the last line is bounded as $J \to \infty$ because the summand is bounded uniformly over j. This shows that $B_J := \sup_{\widetilde{\Lambda}\in\widetilde{\mathcal{L}}}\|\frac{\partial}{\partial\widetilde{\Lambda}}G_J(\widetilde{\Lambda})\| = O_p(1)$. Combining this with (16) gives

$$|\widetilde{G}_J(\widetilde{\Lambda}) - \widetilde{G}_J(\widetilde{\Lambda}^\dagger)| \leq B_J\|\widetilde{\Lambda} - \widetilde{\Lambda}^\dagger\|,$$

for all $\Lambda, \widetilde{\Lambda} \in \widetilde{\mathcal{L}}$ and $B_J = O_p(1)$, which establishes the desired Lipschitz condition. This completes the proof for $\sup_{\Lambda\in\mathcal{S}_T^+}|G_J(\Lambda)| \xrightarrow{p} 0$.

Now, to strengthen the convergence in probability to convergence in L^1, I show that $\{\sup_\Lambda|G_J(\Lambda)|\}_{J\leq 1}$ is uniformly integrable. A bound on $\sup_\Lambda|G_J(\Lambda)|$ is given by

$$\begin{aligned}
\sup_\Lambda|G_J(\Lambda)| &= \sup_\Lambda\left|\tfrac{1}{J}\textstyle\sum_{j=1}^J \operatorname{tr}(\Sigma_j(\Lambda+\Sigma_j)^{-1}(y_jy_j' - y_j\theta_j' - \Sigma_j))\right| \\
&\leq \frac{1}{J}\sum_{j=1}^J \sup_\Lambda\left|\operatorname{tr}(\Sigma_j(\Lambda+\Sigma_j)^{-1}(y_jy_j' - y_j\theta_j' - \Sigma_j))\right| \\
&\leq \frac{T}{J}\sum_{j=1}^J \sigma_1(H_j) \\
&\leq \frac{T}{J}\sum_{j=1}^J \kappa(\Sigma_j)(y_j'y_j + \|\theta_j\|\|y_j\| + \sigma_1(\Sigma_j))
\end{aligned}$$

where the last inequality follows from (13) and (14). Let $\overline{G}_J$ denote the expression in the last line, and suppose that $\limsup_{J\to\infty}\mathbf{E}\,\overline{G}_J^2 < \infty$, which I verify below. Then, I have $\sup_J \mathbf{E}(\sup_\Lambda|G_J(\Lambda)|)^2 < \infty$, from which the uniform integrability follows. It remains only to show that $\limsup_{J\to\infty}\mathbf{E}\,\overline{G}_J^2 < \infty$. By Cauchy-Schwarz, it follows that

$$\mathbf{E}\;\overline{G}_J^2 \leq \frac{T^2}{J}\sum_{j=1}^J \mathbf{E}\left(\kappa(\Sigma_j)(y_j'y_j + \|\theta_j\|\|y_j\| + \sigma_1(\Sigma_j))\right)^2,$$

and the term in the summand is uniformly bounded over $j \geq 1$. This establishes $\limsup_{J\to\infty}\mathbf{E}\;\overline{G}_J^2 < \infty$, and thus that $\{\sup_\Lambda|G_J(\Lambda)|\}_{J\leq 1}$ is uniformly integrable. This concludes the proof.

A.2 Proof of Theorem 4.2

All supremums over μ are understood to be taken over $\mathcal{M}_J$, though for simplicity I write $\sup_\mu$. Observe that, by (12),

$$\begin{aligned}
&\sup_{\mu,\Lambda}\left|\mathrm{URE}(\mu,\Lambda)-\ell(\theta,\widehat{\theta}(\mu,\Lambda))\right| \\
\leq&\sup_{\Lambda}\left|\mathrm{URE}(\Lambda)-\ell(\theta,\widehat{\theta}(\Lambda))\right|+\sup_{\mu,\Lambda}\left|\frac{1}{J}\sum_{j=1}^{J}\mu'(\Lambda+\Sigma_j)^{-1}\Sigma_j(y_j-\theta_j)\right|.
\end{aligned}$$

Since Theorem 4.1 shows that the first term on the right-hand side converges to zero in L^1, it now remains to show

$$\sup_{\mu,\Lambda}\left|\frac{1}{J}\sum_{j=1}^{J}\mu'(\Lambda+\Sigma_j)^{-1}\Sigma_j(y_j-\theta_j)\right|\xrightarrow{L^1}0. \tag{18}$$

By two applications of Cauchy-Schwarz, we have

$$\begin{aligned}
&\mathbf{E}\sup_{\mu,\Lambda}\left|\frac{1}{J}\sum_{j=1}^{J}\mu'(\Lambda+\Sigma_j)^{-1}\Sigma_j(y_j-\theta_j)\right| \\
\leq&\ \mathbf{E}\sup_{\mu}\|\mu\|\cdot\sup_{\Lambda}\left\|\frac{1}{J}\sum_{j=1}^{J}(\Lambda+\Sigma_j)^{-1}\Sigma_j(y_j-\theta_j)\right\| \\
\leq&\left(\mathbf{E}\sup_{\mu}\|\mu\|^2\right)^{1/2}\left(\mathbf{E}\sup_{\Lambda}\left\|\frac{1}{J}\sum_{j=1}^{J}(\Lambda+\Sigma_j)^{-1}\Sigma_j(y_j-\theta_j)\right\|^2\right)^{1/2}.
\end{aligned} \tag{19}$$

I show $\limsup_{J\to\infty}\mathbf{E}\sup_\mu\|\mu\|^2<\infty$ and

$$\lim_{J\to\infty}\mathbf{E}\sup_{\Lambda}\left\|\frac{1}{J}\sum_{j=1}^{J}(\Lambda+\Sigma_j)^{-1}\Sigma_j(y_j-\theta_j)\right\|^2=0,$$

from which then (18) will follow.

Write $H_J(\Lambda):=\|\frac{1}{J}\sum_{j=1}^{J}(\Lambda+\Sigma_j)^{-1}\Sigma_j(y_j-\theta_j)\|$. As in the proof for Theorem 4.1, I show (a) $\sup_\Lambda H_J(\Lambda)\xrightarrow{p}0$ and (b) $\sup_J\mathbf{E}\left(\sup_\Lambda H_J(\Lambda)\right)^{2+\delta}<\infty$ for some $\delta>0$. Since (b) is a sufficient condition for $\{\sup_\Lambda H_J(\Lambda)^2\}_{J\geq1}$ being uniformly integrable, (a) and (b) together imply $\sup_\Lambda H_J(\Lambda)\xrightarrow{L^2}0$.

I show $\sup_\Lambda H_J(\Lambda) \xrightarrow{p} 0$ by again using a ULLN argument as in Andrews (1992). First, to show $H_J(\Lambda) \xrightarrow{p} 0$, it is enough to show $\mathbf{E}\, H_J(\Lambda)^2 \to 0$. Note that

$$\begin{aligned}
\mathbf{E}\, H_J(\Lambda)^2 &= \mathbf{E} \left\| \frac{1}{J} \sum_{j=1}^J (\Lambda + \Sigma_j)^{-1} \Sigma_j (y_j - \theta_j) \right\|^2 \\
&= \frac{1}{J^2}\, \mathbf{E} \left(\sum_{\ell=1}^J (\Lambda + \Sigma_\ell)^{-1} \Sigma_\ell (y_\ell - \theta_\ell) \right)' \left(\sum_{j=1}^J (\Lambda + \Sigma_j)^{-1} \Sigma_j (y_j - \theta_j) \right) \\
&= \frac{1}{J^2}\, \mathbf{E} \left(\sum_{j=1}^J (y_j - \theta_j)' \Sigma_j (\Lambda + \Sigma_j)^{-2} \Sigma_j (y_j - \theta_j) \right) \\
&\leq \frac{1}{J^2} \sum_{j=1}^J \operatorname{tr}(\Sigma_j (\Lambda + \Sigma_j)^{-2} \Sigma_j \Sigma_j) \\
&\leq \frac{1}{J^2} \sum_{j=1}^J \operatorname{tr}(\Sigma_j),
\end{aligned}$$

where the last inequality follows from von Neumann's trace inequality and the fact that $\sigma_1(\Sigma_j(\Lambda + \Sigma_j)^{-2}\Sigma_j) \leq 1$. Moreover, by Assumption 4.1, we have $\sup_j \operatorname{tr}(\Sigma_j) \leq T \sup_j \sigma_1(\Sigma_j) < \infty$, which implies $\frac{1}{J^2} \sum_{j=1}^J \operatorname{tr}(\Sigma_j) \to 0$. This establishes that $H_J(\Lambda)$ converges to zero in L^2, and thus in probability.

To show that this convergence is uniform over $\Lambda \in \mathcal{S}_T^+$, by a similar argument as in the proof of Theorem 4.1, it suffices to show that $\widetilde{H}_J := H_J \circ h^{-1}$ satisfies a Lipschitz condition, i.e.,

$$|\widetilde{H}_J(\widetilde{\Lambda}) - \widetilde{H}_J(\widetilde{\Lambda}^\dagger)| \leq B_{H,J} \|\widetilde{\Lambda} - \widetilde{\Lambda}^\dagger\| \tag{20}$$

for all $\widetilde{\Lambda}, \widetilde{\Lambda}^\dagger \in \widetilde{\mathcal{L}}$, where $B_{H,J} = O_p(1)$. Define $\widetilde{A}_j = \widetilde{\Lambda}^{-1} + (\Sigma_j - \underline{\sigma}_\Sigma I_T)$ and $\widetilde{A}_j^\dagger$ likewise with $\widetilde{\Lambda}$ replaced with $\widetilde{\Lambda}^\dagger$. Observe that

$$\begin{aligned}
|\widetilde{H}_J(\widetilde{\Lambda}) - \widetilde{H}_J(\widetilde{\Lambda}^\dagger)| =& \left| \| \tfrac{1}{J} \textstyle\sum_{j=1}^J \widetilde{A}_j^{-1} \Sigma_j (y_j - \theta_j) \| - \| \tfrac{1}{J} \textstyle\sum_{j=1}^J \widetilde{A}_j^{\dagger -1} \Sigma_j (y_j - \theta_j) \| \right| \\
\leq& \| \tfrac{1}{J} \textstyle\sum_{j=1}^J (\widetilde{A}_j^{-1} - \widetilde{A}_j^{\dagger -1}) \Sigma_j (y_j - \theta_j) \| \\
\leq& \tfrac{1}{J} \textstyle\sum_{j=1}^J \sigma_1(\widetilde{A}_j^{-1} - \widetilde{A}_j^{\dagger -1}) \| \Sigma_j (y_j - \theta_j) \|,
\end{aligned} \tag{21}$$

where the first inequality follows from the reverse triangle inequality and the second

by the triangle inequality and the definition of the operator norm. We have

$$\begin{aligned}\widetilde{A}_j^{-1} - \widetilde{A}_j^{\dagger -1} =& \widetilde{A}_j^{\dagger -1}(\widetilde{A}_j^{\dagger} - \widetilde{A}_j)\widetilde{A}_j^{-1} \\ =& \widetilde{A}_j^{\dagger -1}(\widetilde{\Lambda}^{\dagger -1} - \widetilde{\Lambda}^{-1})\widetilde{A}_j^{-1} \\ =& \widetilde{A}_j^{\dagger -1}\widetilde{\Lambda}^{-1}(\widetilde{\Lambda} - \widetilde{\Lambda}^{\dagger})\widetilde{\Lambda}^{\dagger -1}\widetilde{A}_j^{-1},\end{aligned} \tag{22}$$

which implies

$$\begin{aligned}\sigma_1(\widetilde{A}_j^{-1} - \widetilde{A}_j^{\dagger -1}) \leq& \sigma_1(\widetilde{A}_j^{\dagger -1}\widetilde{\Lambda}^{-1})\sigma_1(\widetilde{\Lambda} - \widetilde{\Lambda}^{\dagger})\sigma_1(\widetilde{\Lambda}^{\dagger -1}\widetilde{A}_j^{-1}) \\ \leq& (\sup_{\widetilde{\Lambda}\in\widetilde{\mathcal{L}}} \sigma_1((\widetilde{\Lambda}^{-1} + (\Sigma_j - \underline{\sigma}_{\Sigma} I_T))^{-1}\widetilde{\Lambda}^{-1})^2)\sigma_1(\widetilde{\Lambda} - \widetilde{\Lambda}^{\dagger}),\end{aligned} \tag{23}$$

where the first inequality follows from (22) and the fact that the operator norm is submultiplicative and the second inequality from the fact that $\sigma_1(C) = \sigma_1(C')$ for any matrix C. Furthermore, we have shown in (17) that $\sigma_1((\widetilde{\Lambda}^{-1} + (\Sigma_j - \underline{\sigma}_{\Sigma} I_T))^{-1}\widetilde{\Lambda}^{-1})^2$ is bounded above by 1. Hence, I obtain

$$\sigma_1(\widetilde{A}_j^{-1} - \widetilde{A}_j^{\dagger -1}) \leq \sigma_1(\widetilde{\Lambda} - \widetilde{\Lambda}^{\dagger}) \leq \|\widetilde{\Lambda} - \widetilde{\Lambda}^{\dagger}\|,$$

where the last inequality follows from the fact that the operator norm of a matrix is less than or equal to its Frobenius norm.

Plugging this bound into (21), it follows that

$$|\widetilde{H}_J(\widetilde{\Lambda}) - \widetilde{H}_J(\widetilde{\Lambda}^{\dagger})| \leq \left(\frac{1}{J}\sum_{j=1}^{J} \|\Sigma_j(y_j - \theta_j)\|\right) \|\widetilde{\Lambda} - \widetilde{\Lambda}^{\dagger}\|.$$

Therefore, it remains to show $\frac{1}{J}\sum_{j=1}^{J}\|\Sigma_j(y_j - \theta_j)\| = O_p(1)$ to establish (20). Observe that

$$\begin{aligned}&\frac{1}{J}\sum_{j=1}^{J}\|\Sigma_j(y_j - \theta_j)\| \\ =& \frac{1}{J}\sum_{j=1}^{J}(\|\Sigma_j(y_j - \theta_j)\| - \mathbf{E}\|\Sigma_j(y_j - \theta_j)\|) + \frac{1}{J}\sum_{j=1}^{J}\mathbf{E}\|\Sigma_j(y_j - \theta_j)\|.\end{aligned}$$

Since $\sup_j \operatorname{var}(\|\Sigma_j(y_j - \theta_j)\|) \leq \sup_j E\|\Sigma_j(y_j - \theta_j)\|^2 = \sup_j \operatorname{tr}(\Sigma_j)^3 < \infty$ the first term converges to zero in probability by an application of Chebyshev's inequality. Also, because $\sup_j \mathbf{E}\|\Sigma_j(y_j - \theta_j)| \leq \sup_j \sigma_1(\Sigma_j)(\mathbf{E}\|y_j\| + \|\theta_j\|) < \infty$ by Assumption

4.1, the second term is $O(1)$. This establishes $\frac{1}{J}\sum_{j=1}^{J}\|\Sigma_j(y_j-\theta_j)\| = O_p(1)$, and thus $\sup_\Lambda H_J(\Lambda) \xrightarrow{p} 0$.

Now, to show that $\sup_\Lambda H_J(\Lambda)$ converges to zero in L^2, it is enough to show that $\{\sup_\Lambda H_J(\Lambda)^2\}_{J\leq 1}$ is uniformly integrable. A sufficient condition for this to hold is

$$\sup_J \mathbf{E}\sup_\Lambda H_J(\Lambda)^{2+\delta} < \infty,$$

for some $\delta > 0$. First, I derive an upper bound of $H_J(\Lambda)$,

$$\begin{aligned} H_J(\Lambda) =& \left\| \frac{1}{J}\sum_{j=1}^{J}(\Lambda+\Sigma_j)^{-1}\Sigma_j(y_j-\theta_j) \right\| \\ \leq& \frac{1}{J}\sum_{j=1}^{J}\sigma_1((\Lambda+\Sigma_j)^{-1}\Sigma_j))\|y_j-\theta_j\| \\ \leq& \frac{1}{J}\sum_{j=1}^{J}\|y_j-\theta_j\|, \end{aligned}$$

where the first inequality follows from the triangle inequality and the definition of the operator norm, and the second inequality follows because

$$\sigma_1((\Lambda+\Sigma_j)^{-1}\Sigma_j)^2 = \sigma_1(\Sigma_j(\Lambda+\Sigma_j)^{-2}\Sigma_j) = \sigma_1(I_T + \Sigma_j^{-1/2}\Lambda\Sigma_j^{-1/2}) \leq 1.$$

Therefore, I have

$$\begin{aligned} \sup_\Lambda H_J(\Lambda)^{2+\delta} &\leq \left(\frac{1}{J}\sum_{j=1}^{J}\|y_j-\theta_j\|\right)^{2+\delta} \\ &\leq \frac{1}{J}\sum_{j=1}^{J}\|y_j-\theta_j\|^{2+\delta} \\ &\leq \frac{1}{J}\sum_{j=1}^{J}2^{1+\delta}(\|y_j\|^{2+\delta}+\|\theta_j\|^{2+\delta}), \end{aligned} \tag{24}$$

where the second inequality follows from Jensen's inequality, and the last inequality follows from the triangle inequality and the fact that $(a+b)^p \leq 2^{p-1}(a^p+b^p)$ for any

$a, b \geq 0$ and $p \geq 1$. Taking expectations shows that, for any $\delta \in [0, 2]$,

$$\limsup_J \mathbf{E} \sup_\Lambda H_J(\Lambda)^{2+\delta} < \infty,$$

and thus $\sup_J \mathbf{E} \sup_\Lambda H_J(\Lambda)^{2+\delta} < \infty$. This concludes the proof for $\sup_\Lambda H_J(\Lambda) \xrightarrow{L^2} 0$.

It remains to show $\limsup_{J\to\infty} \mathbf{E} \sup_\mu \|\mu\|^2 < \infty$. I have

$$\sup_\mu \|\mu\|^2 = \sup_\mu \sum_{t=1}^T \mu_t^2 = \sum_{t=1}^T q_{1-\tau}^2(\{|y_{jt}|\}_{j=1}^J),$$

and thus it suffices to show that $\limsup_{J\to\infty} Eq_{Jt}^2 < \infty$ for $t = 1, \ldots, T$. Observe that

$$\begin{aligned} q_{1-\tau}^2(\{|y_{jt}|\}_{j=1}^J) = \; & q_{1-\tau}(\{y_{jt}^2\}_{j=1}^J) \leq q_{1-\tau}(\{2\theta_{jt}^2 + 2\varepsilon_{jt}^2\}_{j=1}^J) \\ & \leq 2q_{1-\tau/2}(\{\theta_{jt}^2\}_{j=1}^J) + 2q_{1-\tau/2}(\{\varepsilon_{jt}^2\}_{j=1}^J), \end{aligned}$$

where the last inequality follows from a property of a quantile that the $1-\tau$ quantile of the sum of two random variables are bounded by the sum of the $1-\tau/2$ quantiles of those two random variables. It follows that $q_{1-\tau/2}(\{\theta_{jt}^2\}_{j=1}^J) < \sup_j \theta_{jt}^2 < \infty$, and thus it suffices to show that $\limsup_{J\to\infty} Eq_{1-\tau/2}(\{\varepsilon_{jt}^2\}_{j=1}^J) < \infty$. I have

$$q_{1-\tau/2}(\{\varepsilon_{jt}^2\}_{j=1}^J) = q_{1-\tau/2}(\{\sigma_{jt}^2\eta_{jt}^2\}_{j=1}^J) \leq q_{1-\tau/2}(\{\overline{\sigma}_t^2\eta_{jt}^2\}_{j=1}^J) = \overline{\sigma}_t^2 q_{1-\tau/2}(\{\eta_{jt}^2\}_{j=1}^J),$$

where the first equality holds by Assumption 4.2 and the inequality holds because replacing σ_{jt}^2 by $\overline{\sigma}_t^2$ makes all the sample points larger, and thus the sample quantile larger. Define $\underline{\tau} = 2(1 - \lceil J(1-\tau/2) \rceil / J)$, which is the largest $\underline{\tau} \leq \tau$ such that $J(1-\underline{\tau}/2)$ is an integer. By a result on the bias of sample quantiles given by Okolewski and Rychlik (2001), it follows that

$$\sup_J \mathbf{E}\; q_{1-\underline{\tau}/2}(\{\eta_{jt}^2\}_{j=1}^J) \leq \left(\frac{\operatorname{var}(\eta_{jt}^2)}{(1-\underline{\tau}/2)\underline{\tau}/2} \right)^{1/2} + F_t^{-1}(1-\underline{\tau}/2) < \infty,$$

and $q_{1-\tau/2}(\{\eta_{jt}^2\}_{j=1}^J) \leq q_{1-\underline{\tau}/2}(\{\eta_{jt}^2\}_{j=1}^J)$ because $\underline{\tau} \leq \tau$. This establishes that $\mathbf{E} \sup_\mu \|\mu\|^2 < \infty$, which concludes the proof.

A.3 Proof of Theorem 4.3

By essentially the same calculations given in (19), it is enough to show

$$\limsup_{J\to\infty} \mathbf{E} \sup_{\gamma\in\Gamma_J} \|\gamma\|^2 < \infty, \text{ and}$$
$$\lim_{J\to\infty} \mathbf{E} \sup_{\Lambda} \left\| \frac{1}{J} \sum_{j=1}^{J} Z_j'(\Lambda + \Sigma_j)^{-1}\Sigma_j(y_j - \theta_j) \right\|^2 = 0,$$

The first line is equivalent to showing

$$\limsup_{J\to\infty} \mathbf{E}(\textstyle\sum_{j=1}^J y_j'Z_j)(\sum_{j=1}^J Z_j'Z_j)^{-2}(\sum_{j=1}^J Z_j'y_j) < \infty. \tag{25}$$

Simple calculations show that

$$\begin{aligned}
&\mathbf{E}(\textstyle\sum_{j=1}^J y_j'Z_j)(\sum_{j=1}^J Z_j'Z_j)^{-2}(\sum_{j=1}^J Z_j'y_j) \\
=&\mathbf{E}(\textstyle\sum_{j=1}^J (\varepsilon_j' + \theta_j')Z_j)(\sum_{j=1}^J Z_j'Z_j)^{-2}(\sum_{j=1}^J Z_j'(\theta_j + \varepsilon_j)) \\
=&(\textstyle\sum_{j=1}^J \theta_j'Z_j)(\sum_{j=1}^J Z_j'Z_j)^{-2}(\sum_{j=1}^J Z_j'\theta_j) + \mathbf{E}(\sum_{j=1}^J \varepsilon_j'Z_j)(\sum_{j=1}^J Z_j'Z_j)^{-2}(\sum_{j=1}^J Z_j'\varepsilon_j),
\end{aligned}$$

where the last equality follows because the "cross terms" are zero due to the conditional mean independence assumption. I show that the first and second term of the last line is $O(1)$ and $o(1)$, respectively, which in turn will imply (25). Note that

$$\begin{aligned}
&\|(\textstyle\sum_{j=1}^J \theta_j'Z_j)(\sum_{j=1}^J Z_j'Z_j)^{-2}(\sum_{j=1}^J Z_j'\theta_j)\| \\
\leq&\sigma_1((\tfrac{1}{J}\textstyle\sum_{j=1}^J Z_j'Z_j)^{-2})\|\tfrac{1}{J}\textstyle\sum_{j=1}^J Z_j'\theta_j\|^2 \\
\leq&\sigma_1((\tfrac{1}{J}\textstyle\sum_{j=1}^J Z_j'Z_j)^{-2})(\tfrac{1}{J}\textstyle\sum_{j=1}^J \sigma_1(Z_j)\|\theta_j\|)^2,
\end{aligned}$$

with $\sigma_1((\frac{1}{J}\sum_{j=1}^J Z_j'Z_j)^{-2}) \to \sigma_1((\mathbf{E}\, Z_j'Z_j)^{-2})$ and $\limsup_{J\to\infty} \frac{1}{J}\sum_{j=1}^J \sigma_1(Z_j)\|\theta_j\| < \infty$. This shows that $\limsup_{J\to\infty}(\sum_{j=1}^J \theta_j'Z_j)(\sum_{j=1}^J Z_j'Z_j)^{-2}(\sum_{j=1}^J Z_j'\theta_j) < \infty$.

It remains to show $E(\sum_{j=1}^J \varepsilon_j'Z_j)(\sum_{j=1}^J Z_j'Z_j)^{-2}(\sum_{j=1}^J Z_j'\varepsilon_j) \to 0$. Because

$$\begin{aligned}
&(\textstyle\sum_{j=1}^J \varepsilon_j'Z_j)(\sum_{j=1}^J Z_j'Z_j)^{-2}(\sum_{j=1}^J Z_j'\varepsilon_j) \\
=&\operatorname{tr}((\textstyle\sum_{j=1}^J Z_j'Z_j)^{-2}(\sum_{j=1}^J Z_j'\varepsilon_j)(\sum_{j=1}^J \varepsilon_j'Z_j)) \\
=&\operatorname{tr}((\textstyle\sum_{j=1}^J Z_j'Z_j)^{-2}(\sum_{j=1}^J \sum_{\ell=1}^J Z_j'\varepsilon_j\varepsilon_\ell'Z_\ell)),
\end{aligned}$$

it follows that

$$\begin{aligned}
&\mathbf{E}(\textstyle\sum_{j=1}^J \varepsilon_j' Z_j)(\sum_{j=1}^J Z_j' Z_j)^{-2}(\sum_{j=1}^J Z_j' \varepsilon_j) \\
&= \operatorname{tr}((\textstyle\sum_{j=1}^J Z_j' Z_j)^{-2}(\sum_{j=1}^J \sum_{\ell=1}^J Z_j' \,\mathbf{E}[\varepsilon_j \varepsilon_\ell'] Z_\ell)) \\
&= \operatorname{tr}((\textstyle\sum_{j=1}^J Z_j' Z_j)^{-2}(\sum_{j=1}^J Z_j' \Sigma_j Z_j)) \\
&= \tfrac{1}{J} \operatorname{tr}((\tfrac{1}{J}\textstyle\sum_{j=1}^J Z_j' Z_j)^{-2}(\tfrac{1}{J}\sum_{j=1}^J Z_j' \Sigma_j Z_j))
\end{aligned}$$

Again, note that $(\frac{1}{J}\sum_{j=1}^J Z_j' Z_j)^{-2} \to (\mu_{Z,2})^{-2}$, and

$$\limsup_{J\to\infty} \left\| \frac{1}{J} \sum_{j=1}^J Z_j' \Sigma_j Z_j \right\| \leq \limsup_{J\to\infty} \frac{1}{J} \sum_{j=1}^J \sigma_1(Z_j)^2 \sigma_1(\Sigma_j) < \infty.$$

This shows that $\frac{1}{J}\operatorname{tr}((\frac{1}{J}\sum_{j=1}^J Z_j' Z_j)^{-2}(\frac{1}{J}\sum_{j=1}^J Z_j' \Sigma_j Z_j)) \to 0$, which concludes the proof for (25).

To show $\lim_{J\to\infty} \mathbf{E} \sup_\Lambda \|\frac{1}{J}\sum_{j=1}^J Z_j'(\Lambda+\Sigma_j)^{-1}\Sigma_j(y_j-\theta_j)\|^2 = 0$, I follow the lines of argument given in the proof of Theorem 4.2 carefully. The main difference is that now the summand is multiplied by Z_j'. Write $H_{Z,J}(\Lambda) := \|\frac{1}{J}\sum_{j=1}^J Z_j'(\Lambda+\Sigma_j)^{-1}\Sigma_j(y_j-\theta_j)\|$. First, I show $EH_{Z,J}(\Lambda)^2 \to 0$, which implies $H_{Z,J}(\Lambda) \xrightarrow{p} 0$. Write $\overline{\sigma}_Z := \sup_j \sigma_1(Z_j)$, and note that $\sup_j \sigma_1(Z_j Z_j') = \sup_j \sigma_1(Z_j' Z_j) = \overline{\sigma}_Z^2$. I have

$$\begin{aligned}
EH_{Z,J}(\Lambda)^2 &= \mathbf{E}\|\tfrac{1}{J}\textstyle\sum_{j=1}^J Z_j'(\Lambda+\Sigma_j)^{-1}\Sigma_j(y_j-\theta_j)\|^2 \\
&= \tfrac{1}{J^2}\,\mathbf{E}(\textstyle\sum_{\ell=1}^J Z_\ell'(\Lambda+\Sigma_\ell)^{-1}\Sigma_\ell(y_\ell-\theta_\ell))'(\sum_{j=1}^J Z_j'(\Lambda+\Sigma_j)^{-1}\Sigma_j(y_j-\theta_j)) \\
&= \tfrac{1}{J^2}\,\mathbf{E}(\textstyle\sum_{j=1}^J (y_j-\theta_j)'\Sigma_j(\Lambda+\Sigma_j)^{-1} Z_j Z_j'(\Lambda+\Sigma_j)^{-1}\Sigma_j(y_j-\theta_j)) \\
&\leq \tfrac{1}{J^2}\textstyle\sum_{j=1}^J \operatorname{tr}(\Sigma_j(\Lambda+\Sigma_j)^{-1} Z_j Z_j'(\Lambda+\Sigma_j)^{-1}\Sigma_j\Sigma_j) \\
&\leq \tfrac{1}{J^2}\textstyle\sum_{j=1}^J \sigma_1(\Sigma_j(\Lambda+\Sigma_j)^{-1} Z_j Z_j'(\Lambda+\Sigma_j)^{-1}\Sigma_j)\operatorname{tr}(\Sigma_j) \\
&\leq \tfrac{1}{J^2}\textstyle\sum_{j=1}^J \overline{\sigma}_Z^2 \operatorname{tr}(\Sigma_j),
\end{aligned}$$

where the second inequality follows from von Neumann's trace inequality and the last equality from the fact that the operator norm is submultiplicative and the bound $\sigma_1(\Sigma_j(\Lambda+\Sigma_j)^{-2}\Sigma_j) \leq 1$. Since $\sup_j \operatorname{tr}(\Sigma_j) \leq T \sup_j \sigma_1(\Sigma_j) < \infty$, I conclude that $\frac{1}{J^2}\sum_{j=1}^J \overline{\sigma}_Z^2 \operatorname{tr}(\Sigma_j) \to 0$. This establishes that $H_{Z,J}(\Lambda)$ converges to zero in L^2, and thus in probability.

To show that this convergence is uniform over $\Lambda \in \mathcal{S}_T^+$, by a similar argument as in the proof of Theorem 4.1, it suffices to show that $\widetilde{H}_{Z,J} := H_{Z,J} \circ h^{-1}$ satisfies a

Lipschitz condition, i.e.,

$$|\widetilde{H}_{Z,J}(\widetilde{\Lambda}) - \widetilde{H}_{Z,J}(\widetilde{\Lambda}^\dagger)| \leq B^Z_{H,J}\|\widetilde{\Lambda} - \widetilde{\Lambda}^\dagger\| \tag{26}$$

for all $\widetilde{\Lambda}, \widetilde{\Lambda}^\dagger \in \widetilde{\mathcal{L}}$, where $B^Z_{H,J} = O_p(1)$. Define $\widetilde{A}_j = \widetilde{\Lambda}^{-1} + (\Sigma_j - \underline{\sigma}_\Sigma I_T)$ and $\widetilde{A}^\dagger_j$ likewise with $\widetilde{\Lambda}$ replaced with $\widetilde{\Lambda}^\dagger$. Observe that

$$\begin{aligned}
&|\widetilde{H}_{Z,J}(\widetilde{\Lambda}) - \widetilde{H}_{Z,J}(\widetilde{\Lambda}^\dagger)| \\
=&\left|\left\|\frac{1}{J}\sum_{j=1}^J Z_j'\widetilde{A}_j^{-1}\Sigma_j(y_j - \theta_j)\right\| - \left\|\frac{1}{J}\sum_{j=1}^J Z_j'\widetilde{A}_j^{\dagger-1}\Sigma_j(y_j - \theta_j)\right\|\right| \\
\leq&\left\|\frac{1}{J}\sum_{j=1}^J Z_j'(\widetilde{A}_j^{-1} - \widetilde{A}_j^{\dagger-1})\Sigma_j(y_j - \theta_j)\right\| \\
\leq&\frac{1}{J}\sum_{j=1}^J \sigma_1(Z_j)\sigma_1(\widetilde{A}_j^{-1} - \widetilde{A}_j^{\dagger-1})\left\|\Sigma_j(y_j - \theta_j)\right\|,
\end{aligned} \tag{27}$$

where the first inequality follows from the reverse triangle inequality and the second by the triangle inequality and the definition of the operator norm.

In the proof of Theorem 4.2, I showed that

$$\sigma_1(\widetilde{A}_j^{-1} - \widetilde{A}_j^{\dagger-1}) \leq \|\widetilde{\Lambda} - \widetilde{\Lambda}^\dagger\|.$$

Plugging this bound into (27), I obtain

$$|\widetilde{H}_{Z,J}(\widetilde{\Lambda}) - \widetilde{H}_{Z,J}(\widetilde{\Lambda}^\dagger)| \leq (\tfrac{1}{J}\textstyle\sum_{j=1}^J \sigma_1(Z_j)\|\Sigma_j(y_j - \theta_j)\|)\|\widetilde{\Lambda} - \widetilde{\Lambda}^\dagger\|.$$

Furthermore, I have

$$\sum_{j=1}^J \sigma_1(Z_j)\|\Sigma_j(y_j - \theta_j)\| \leq \overline{\sigma}_Z \sum_{j=1}^J \|\Sigma_j(y_j - \theta_j)\|,$$

and I have already shown $\frac{1}{J}\sum_{j=1}^J \|\Sigma_j(y_j - \theta_j)\| = O_p(1)$ in the proof of Theorem 4.2. This establishes (26), and thus $\sup_\Lambda H_{Z,J}(\Lambda) \xrightarrow{p} 0$.

Now, to show that $\sup_\Lambda H_{Z,J}(\Lambda)$ converges to zero in L^2, it is enough to show that

$\{\sup_\Lambda H_{Z,J}(\Lambda)^2\}_{J\leq 1}$ is uniformly integrable. A sufficient condition for this is

$$\sup_J \mathbf{E} \sup_\Lambda H_{Z,J}(\Lambda)^{2+\delta} < \infty,$$

for some $\delta > 0$. An upper bound of $H_{Z,J}(\Lambda)$ is given by

$$\begin{aligned} H_{Z,J}(\Lambda) =& \left\| \frac{1}{J} \sum_{j=1}^{J} Z_j'(\Lambda + \Sigma_j)^{-1}\Sigma_j(y_j - \theta_j) \right\| \\ \leq& \frac{1}{J} \sum_{j=1}^{J} \sigma_1(Z_j)\sigma_1((\Lambda + \Sigma_j)^{-1}\Sigma_j))\|y_j - \theta_j\| \\ \leq& \overline{\sigma}_Z \frac{1}{J} \sum_{j=1}^{J} \|y_j - \theta_j\|, \end{aligned}$$

where the first inequality follows from the triangle inequality and the definition of the operator norm, and the second inequality follows because $\sigma_1((\Lambda + \Sigma_j)^{-1}\Sigma_j) \leq 1$. Therefore, following (24), I have

$$\sup_\Lambda H_{Z,J}(\Lambda)^{2+\delta} \leq \overline{\sigma}_Z^{2+\delta} \frac{1}{J} \sum_{j=1}^{J} 2^{1+\delta}(\|y_j\|^{2+\delta} + \|\theta_j\|^{2+\delta}),$$

Taking expectations, we obtain

$$\limsup_J \mathbf{E} \sup_\Lambda H_J(\Lambda)^{2+\delta} < \infty,$$

for any $\delta \in [0, 2]$, and thus $\sup_J \mathbf{E} \sup_\Lambda H_{Z,J}(\Lambda)^{2+\delta} < \infty$. This concludes the proof for $\sup_\Lambda H_{Z,J}(\Lambda) \xrightarrow{L^2} 0$.

A.4 Proof of Theorem 5.1

I first give details on the derivation of the UPE. Note that

$$\begin{aligned} &\mathbf{E}(B(\Lambda, \Sigma_{j,-T})'y_{j,-T} - \theta_{jT})^2 \\ =& \mathbf{E}(B(\Lambda, \Sigma_{j,-T})'y_{j,-T} - y_{jT})^2 + \mathbf{E}(y_{jT} - \theta_{jT})^2 \\ &- 2\,\mathbf{E}[(y_{jT} - B(\Lambda, \Sigma_{j,-T})'y_{j,-T})(y_{jT} - \theta_{jT})]. \end{aligned}$$

The cross term can be written as

$$\begin{aligned}
&\mathbf{E}[(y_{jT} - B(\Lambda, \Sigma_{j,-T})' y_{j,-T})(y_{jT} - \theta_{jT})] \\
=&\mathbf{E}[(y_{jT} - \theta_{jT} - B(\Lambda, \Sigma_{j,-T})'(y_{j,-T} - \theta_{j,-T}) + \theta_{jT} - B(\Lambda, \Sigma_{j,-T})'\theta_{j,-T})(y_{jT} - \theta_{jT})] \\
=&\Sigma_{j,T} - B(\Lambda, \Sigma_{j,-T})'\Sigma_{j,T,-T}.
\end{aligned}$$

Hence, it follows that

$$\begin{aligned}
&\mathbf{E}(B(\Lambda, \Sigma_{j,-T})' y_{j,-T} - \theta_{jT})^2 \\
=&\mathbf{E}(B(\Lambda, \Sigma_{j,-T})' y_{j,-T} - y_{jT})^2 - \Sigma_{jT} + 2B(\Lambda, \Sigma_{j,-T})'\Sigma_{j,T,-T},
\end{aligned}$$

which shows the UPE is indeed unbiased.

Now, I prove Theorem 5.1. Consider the following bound,

$$\begin{aligned}
&\left| \frac{1}{J}\sum_{j=1}^{J} \left((B(\Lambda, \Sigma_{j,-T})' y_{j,-T} - y_{jT})^2 - \Sigma_{j,T} \right) - \frac{1}{J}\sum_{j=1}^{J} (B(\Lambda, \Sigma_{j,-1})' y_{j,-1} - \theta_{j,T+1})^2 \right| \\
\leq &\left| \frac{1}{J}\sum_{j=1}^{J} \left((B(\Lambda, \Sigma_{j,-T})' y_{j,-T} - y_{jT})^2 - \Sigma_{j,T} \right) - \frac{1}{J}\sum_{j=1}^{J} (B(\Lambda, \Sigma_{j,-T})' y_{j,-T} - \theta_{j,T})^2 \right| \\
&+ \left| \frac{1}{J}\sum_{j=1}^{J} (B(\Lambda, \Sigma_{j,-T})' y_{j,-T} - \theta_{j,T})^2 - \frac{1}{J}\sum_{j=1}^{J} (B(\Lambda, \Sigma_{j,-1})' y_{j,-1} - \theta_{j,T+1})^2 \right|, \qquad (28)
\end{aligned}$$

which is by the triangle inequality.

Further calculation gives

$$\begin{aligned}
&\frac{1}{J}\sum_{j=1}^{J} (B(\Lambda, \Sigma_{j,-T})' y_{j,-T} - \theta_{j,T})^2 \\
=&\frac{1}{J}\sum_{j=1}^{J} (B(\Lambda, \Sigma_{j,-T})' y_{j,-T} - y_{j,T} + y_{jT} - \theta_{j,T})^2 \\
=&\frac{1}{J}\sum_{j=1}^{J} ((B(\Lambda, \Sigma_{j,-T})' y_{j,-T} - y_{j,T})^2 + (y_{jT} - \theta_{j,T})^2) \\
&- 2\frac{1}{J}\sum_{j=1}^{J} (y_{j,T} - B(\Lambda, \Sigma_{j,-T})' y_{j,-T})(y_{jT} - \theta_{j,T}).
\end{aligned}$$

The cross term can be decomposed as

$$
\begin{aligned}
&\frac{1}{J}\sum_{j=1}^{J}(y_{j,T} - B(\Lambda,\Sigma_{j,-T})'y_{j,-T})(y_{jT}-\theta_{j,T}) \\
=&\frac{1}{J}\sum_{j=1}^{J}(y_{jT} - \theta_{jT} - B(\Lambda,\Sigma_{j,-T})'(y_{j,-T}-\theta_{j,-T}) + \theta_{jT} - B(\Lambda,\Sigma_{j,-T})'\theta_{j,-T})(y_{jT}-\theta_{jT}) \\
=&\frac{1}{J}\sum_{j=1}^{J}(y_{jT}-\theta_{jT})^2 - \frac{1}{J}\sum_{j=1}^{J}B(\Lambda,\Sigma_{j,-T})'(y_{j,-T}-\theta_{j,-T})(y_{jT}-\theta_{jT}) \\
&+\frac{1}{J}\sum_{j=1}^{J}(\theta_{jT} - B(\Lambda,\Sigma_{j,-T})'\theta_{j,-T})(y_{jT}-\theta_{jT}).
\end{aligned}
$$

Plugging this into the first term of right-hand side in (28), it follows that

$$
\begin{aligned}
&\left|\frac{1}{J}\sum_{j=1}^{J}\left((B(\Lambda,\Sigma_{j,-T})'y_{j,-T}-y_{jT})^2 - \Sigma_{j,T}\right) - \frac{1}{J}\sum_{j=1}^{J}(B(\Lambda,\Sigma_{j,-T})'y_{j,-T}-\theta_{j,T})^2\right| \\
\leq&\left|\frac{1}{J}\sum_{j=1}^{J}\left((y_{jT}-\theta_{jT})^2 - \Sigma_{j,T}\right)\right| \\
&+\left|\frac{2}{J}\sum_{j=1}^{J}B(\Lambda,\Sigma_{j,-T})'((y_{j,-T}-\theta_{j,-T})(y_{jT}-\theta_{jT}) - \Sigma_{j,T,-T})\right| \\
&+\left|\frac{2}{J}\sum_{j=1}^{J}\theta_{jT}(y_{jT}-\theta_{jT})\right| + \left|\frac{2}{J}\sum_{j=1}^{J}B(\Lambda,\Sigma_{j,-T})'\theta_{j,-T}(y_{jT}-\theta_{jT})\right| \\
:=&(\mathrm{I})_J + (\mathrm{II})_J + (\mathrm{III})_J + (\mathrm{IV})_J.
\end{aligned}
$$

The aim is to show that each of the four terms in the last line converges to 0 in L^1, uniformly over $\Lambda \in \mathcal{L}$. In fact, I show uniformity over $(\Lambda_{T,-T}, \Lambda_{-T}) \in \overline{\mathcal{L}} := \overline{\mathcal{L}}_{T,-T} \times \overline{\mathcal{L}}_{-T}$, where

$$
\begin{aligned}
&\overline{\mathcal{L}}_{T,-T} = \{\Lambda_{T,-T} \in \mathbf{R}^{T-1} : \|\Lambda_{T,-T}\| \leq K_{T,-T}\}, \text{ and} \\
&\overline{\mathcal{L}}_{-T} = \{\Lambda_{-T} \in S^+_{T-1} : \|\Lambda_{-T}\| \leq K_{-T}\}.
\end{aligned}
$$

Here, $K_{T,-T}$ and K_{-T} are positive numbers large enough so that $\{\Lambda_{T,-T} : \Lambda \in \mathcal{L}\} \subset \overline{\mathcal{L}}_{T,-T}$ and $\{\Lambda_{-T} : \Lambda \in \mathcal{L}\} \subset \overline{\mathcal{L}}_{-T}$, which exist due to the fact that $\mathcal{L}$ is bounded. Note that $\Lambda \in \mathcal{L}$ implies $(\Lambda_{T,-T}, \Lambda_T) \in \overline{\mathcal{L}}$, and thus establishing convergence uniformly over

the latter is sufficient. Note that

$$\begin{aligned}\sup_{(\Lambda_{T,-T},\Lambda_T)\in\overline{\mathcal{L}}}\|B(\Lambda,\Sigma_{j,-T})\| \leq & \sup_{(\Lambda_{T,-T},\Lambda_T)\in\overline{\mathcal{L}}}\|\Lambda_{T,-T}\|\,\sigma_1((\Lambda_{-T}+\Sigma_{j,-T})^{-1}) \\ \leq & K_{T,-T}\sigma_{T-1}^{-1}(\Sigma_{j,-T}) \\ \leq & K_{T,-T}\sigma_T^{-1}(\Sigma_j),\end{aligned} \tag{29}$$

where the last line follows because the relationship between eigenvalues of a matrix and the eigenvalues of its principal submatrices (see, for example, Theorem 4.3.15 of Horn and Johnson (1990)). In some of the derivations later on, it is useful to make clear that $B(\Lambda,\Sigma_{j,-T})$ depends on Λ only through $(\Lambda_{T,-T},\Lambda_{-T})$. When this fact has be highlighted, I write $B(\Lambda_{T,-T},\Lambda_{-T},\Sigma_{j,-T}) := B(\Lambda,\Sigma_{j,-T})$. Now, I condition all random quantities on a sequence $\{((\theta_j',\theta_{j,T+1})',\Sigma_j)\}_{j=1}^{\infty}$. Note that by Assumption 5.1, now I can assume that Assumption 4.1 holds.

To show that $(\mathrm{I})_J$ converges to zero in L^2, and thus in L^1, note that

$$\mathbf{E}\left|\frac{1}{J}\sum_{j=1}^{J}\left((y_{jT}-\theta_{jT})^2-\Sigma_{j,T}\right)\right|^2 \leq \frac{1}{J^2}\sum_{j=1}^{J}8(\mathbf{E}\,y_{jT}^4+\theta_{jT}^4).$$

The summand in the last line is uniformly bounded over j, which establishes the convergence.

Similarly, $(\mathrm{III})_J \xrightarrow{L^2} 0$ can be easily shown by noting that

$$\mathbf{E}\left|\frac{2}{J}\sum_{j=1}^{J}\theta_{jT}(y_{jT}-\theta_{jT})\right|^2 \leq \frac{4}{J^2}\sum_{j=1}^{J}\theta_{jT}^2\Sigma_{jT},$$

and the summand of the right-hand side is bounded uniformly over j.

To show that $\sup_{\overline{\mathcal{L}}}(\mathrm{II})_J \xrightarrow{L^1} 0$ and $\sup_{\overline{\mathcal{L}}}(\mathrm{IV})_J \xrightarrow{L^1} 0$, I again use a result by Andrews (1992), which will establish convergence in probability, and then show a uniform integrability condition to show that convergence holds in L^1 as well. Here, I write $\sup_{\overline{\mathcal{L}}}$ as a shorthand for $\sup_{(\Lambda_{T,-T},\Lambda_{-T})\in\overline{\mathcal{L}}}$. I start with $(\mathrm{II})_J$. For pointwise convergence

(in L^2), note that

$$\begin{aligned}
&\mathbf{E}\left|\frac{2}{J}\sum_{j=1}^{J} B(\Lambda,\Sigma_{j,-T})'((y_{j,-T}-\theta_{j,-T})(y_{jT}-\theta_{jT})-\Sigma_{j,T,-T})\right|^2 \\
=&\frac{4}{J^2}\sum_{j=1}^{J}\operatorname{tr}(B(\Lambda,\Sigma_{j,-T})B(\Lambda,\Sigma_{j,-T})'\operatorname{var}((y_{j,-T}-\theta_{j,-T})(y_{jT}-\theta_{jT})) \\
\leq&\frac{4}{J^2}\sum_{j=1}^{J}\sigma_1(B(\Lambda,\Sigma_{j,-T})B(\Lambda,\Sigma_{j,-T})')\operatorname{tr}(\operatorname{var}((y_{j,-T}-\theta_{j,-T})(y_{jT}-\theta_{jT}))) \\
\leq&\frac{4}{J^2}\sum_{j=1}^{J}K_{T,-T}^2\sigma_T^{-2}(\Sigma_j)\operatorname{tr}(\operatorname{var}((y_{j,-T}-\theta_{j,-T})(y_{jT}-\theta_{jT}))),
\end{aligned}$$

where the second inequality follows from von Neumann's trace inequality and the fact that $\sigma_1(xx') = \sigma_1(x'x) = \|x\|^2$ for any $x \in \mathrm{R}^{T-1}$, and the last inequality from (29). Moreover, I have

$$\begin{aligned}
&\operatorname{tr}(\operatorname{var}((y_{j,-T}-\theta_{j,-T})(y_{jT}-\theta_{jT}))) \\
\leq&\sum_{t=1}^{T-1}\mathbf{E}(y_{jt}-\theta_{jt})^2(y_{jT}-\theta_{jT})^2 \\
\leq&\sum_{t=1}^{T-1}(\mathbf{E}(y_{jt}-\theta_{jt})^4(y_{jT}-\theta_{jT})^4)^{1/2},
\end{aligned}$$

where the second inequality is by Cauchy-Schwarz. Note that the term in the last line is bounded uniformly over j, which establishes $(\mathrm{II})_\mathrm{J} \xrightarrow{L^2} 0$. It remains to establish a Lipschitz condition. Define

$$G_J(\Lambda_{T,-T},\Lambda_{-T}) = \frac{2}{J}\sum_{j=1}^{J} B(\Lambda,\Sigma_{j,-T})'((y_{j,-T}-\theta_{j,-T})(y_{jT}-\theta_{jT})-\Sigma_{j,T,-T}).$$

I show that $G_J(\Lambda_{T,-T},\Lambda_{-T})$ is Lipschitz in $\Lambda_{T,-T}$ and Λ_{-T}, respectively, with Lipschitz constants bounded in probability that do not depend on the other parameter held fixed, which will establish that $G_J(\Lambda_{T,-T},\Lambda_{-T})$ is Lipschitz with respect to

$(\Lambda_{T,-T}, \Lambda_{-T})$. Note that, for any $\Lambda_{T,-T}, \widetilde{\Lambda}_{T,-T} \in \overline{\mathcal{L}}_{T,-T}$,

$$\begin{aligned} &\|B(\Lambda_{T,-T}, \Lambda_{-T}, \Sigma_{j,-T}) - B(\widetilde{\Lambda}_{T,-T}, \Lambda_{-T}, \Sigma_{j,-T})\| \\ \leq &\|(\Lambda_{T,-T} - \widetilde{\Lambda}_{T,-T})'(\Sigma_{j,-T} + \Lambda_{-T})^{-1}\| \\ \leq &\|\Lambda_{T,-T} - \widetilde{\Lambda}_{T,-T}\| \sigma_1((\Sigma_{j,-T} + \Lambda_{-T})^{-1}) \\ \leq &\underline{\sigma}_{\Sigma}^{-1} \|\Lambda_{T,-T} - \widetilde{\Lambda}_{T,-T}\|. \end{aligned} \tag{30}$$

Also, for any $\Lambda_{-T}, \widetilde{\Lambda}_{-T} \in \overline{\mathcal{L}}_{-T}$, we have

$$\begin{aligned} &\|B(\Lambda_{T,-T}, \Lambda_{-T}, \Sigma_{j,-T}) - B(\Lambda_{T,-T}, \widetilde{\Lambda}_{-T}, \Sigma_{j,-T})\| \\ \leq &\|\Lambda'_{T,-T}((\Sigma_{j,-T} + \Lambda_{-T})^{-1} - (\Sigma_{j,-T} + \widetilde{\Lambda}_{-T})^{-1})\| \\ \leq &K_{T,-T} \sigma_1((\Sigma_{j,-T} + \Lambda_{-T})^{-1} - (\Sigma_{j,-T} + \widetilde{\Lambda}_{-T})^{-1}) \end{aligned}$$

To derive a bound for $\sigma_1((\Sigma_{j,-T} + \Lambda_{-T})^{-1} - (\Sigma_{j,-T} + \widetilde{\Lambda}_{-T})^{-1})$, note that

$$\begin{aligned} &(\Sigma_{j,-T} + \Lambda_{-T})^{-1} - (\Sigma_{j,-T} + \widetilde{\Lambda}_{-T})^{-1} \\ \leq &(\Sigma_{j,-T} + \widetilde{\Lambda}_{-T})^{-1}((\Sigma_{j,-T} + \widetilde{\Lambda}_{-T}) - (\Sigma_{j,-T} + \Lambda_{-T}))(\Sigma_{j,-T} + \Lambda_{-T})^{-1} \\ = &(\Sigma_{j,-T} + \widetilde{\Lambda}_{-T})^{-1}(\widetilde{\Lambda}_{-T} - \Lambda_{-T})(\Sigma_{j,-T} + \Lambda_{-T})^{-1}. \end{aligned}$$

This implies $\sigma_1((\Sigma_{j,-T} + \Lambda_{-T})^{-1} - (\Sigma_{j,-T} + \widetilde{\Lambda}_{-T})^{-1}) \leq \underline{\sigma}_{\Sigma}^{-2} \|\Lambda_{-T} - \widetilde{\Lambda}_{-T}\|$, which in turn implies the following Lipschitz condition,

$$\|B(\Lambda_{T,-T}, \Lambda_{-T}, \Sigma_{j,-T}) - B(\Lambda_{T,-T}, \widetilde{\Lambda}_{-T}, \Sigma_{j,-T})\| \leq \underline{\sigma}_{\Sigma}^{-2} \|\Lambda_{-T} - \widetilde{\Lambda}_{-T}\|. \tag{31}$$

Now, combining (30) and (31), we have for any $(\Lambda_{T,-T}, \Lambda_{-T}), (\widetilde{\Lambda}_{T,-T}, \widetilde{\Lambda}_{-T}) \in \overline{\mathcal{L}}$,

$$\begin{aligned} &\|B(\Lambda_{T,-T}, \Lambda_{-T}, \Sigma_{j,-T}) - B(\widetilde{\Lambda}_{T,-T}, \widetilde{\Lambda}_{-T}, \Sigma_{j,-T})\| \\ \leq &\|B(\Lambda_{T,-T}, \Lambda_{-T}, \Sigma_{j,-T}) - B(\widetilde{\Lambda}_{T,-T}, \Lambda_{-T}, \Sigma_{j,-T})\| \\ &+ \|B(\widetilde{\Lambda}_{T,-T}, \Lambda_{-T}, \Sigma_{j,-T}) - B(\widetilde{\Lambda}_{T,-T}, \widetilde{\Lambda}_{-T}, \Sigma_{j,-T})\| \\ \leq &\underline{\sigma}_{\Sigma}^{-1} \|\Lambda_{T,-T} - \widetilde{\Lambda}_{T,-T}\| + \underline{\sigma}_{\Sigma}^{-2} \|\Lambda_{-T} - \widetilde{\Lambda}_{-T}\| \\ \leq &(\underline{\sigma}_{\Sigma}^{-1} \vee \underline{\sigma}_{\Sigma}^{-2})(\|\Lambda_{T,-T} - \widetilde{\Lambda}_{T,-T}\| + \|\Lambda_{-T} - \widetilde{\Lambda}_{-T}\|). \end{aligned} \tag{32}$$

Because $\|(\Lambda_{T,-T}, \Lambda_{-T})\| := \|\Lambda_{T,-T}\| + \|\Lambda_{-T}\|$ defines a norm on the product space $\overline{\mathcal{L}}$, this shows that $B(\cdot, \cdot, \Sigma_{j,-T})$ is Lipshitz on $\overline{\mathcal{L}}$.

It follows that

$$
\begin{aligned}
&|G_J(\Lambda_{T,-T}, \Lambda_{-T}) - G_J(\widetilde{\Lambda}_{T,-T}, \widetilde{\Lambda}_{-T})| \\
\leq &\frac{2}{J}\sum_{j=1}^{J} |(B(\Lambda, \Sigma_{j,-T}) - B(\widetilde{\Lambda}, \Sigma_{j,-T}))'((y_{j,-T} - \theta_{j,-T})(y_{jT} - \theta_{jT}) - \Sigma_{j,T,-T})| \\
\leq &\frac{2}{J}\sum_{j=1}^{J} \|B(\Lambda, \Sigma_{j,-T}) - B(\widetilde{\Lambda}, \Sigma_{j,-T})\| \|(y_{j,-T} - \theta_{j,-T})(y_{jT} - \theta_{jT}) - \Sigma_{j,T,-T})\| \\
\leq &((\underline{\sigma}_{\Sigma}^{-1} \vee \underline{\sigma}_{\Sigma}^{-2})\frac{2}{J}\sum_{j=1}^{J} \|(y_{j,-T} - \theta_{j,-T})(y_{jT} - \theta_{jT}) - \Sigma_{j,T,-T}\|) \|(\Lambda_{T,-T}, \Lambda_{-T}) - (\widetilde{\Lambda}_{T,-T}, \widetilde{\Lambda}_{-T})\|.
\end{aligned}
$$

Hence, now it suffices to show

$$
\frac{1}{J}\sum_{j=1}^{J} \|(y_{j,-T} - \theta_{j,-T})(y_{jT} - \theta_{jT}) - \Sigma_{j,T,-T}\| = O_p(1). \tag{33}
$$

A bound for the left-hand side is given by

$$
\begin{aligned}
&\frac{1}{J}\sum_{j=1}^{J} \|(y_{j,-T} - \theta_{j,-T})(y_{jT} - \theta_{jT}) - \Sigma_{j,T,-T}\| \\
\leq &\frac{1}{J}\sum_{j=1}^{J} \|(y_{j,-T} - \theta_{j,-T})(y_{jT} - \theta_{jT})\| + \frac{1}{J}\sum_{j=1}^{J} \|\Sigma_{j,T,-T}\| \\
= &\frac{1}{J}\sum_{j=1}^{J} (\|(y_{j,-T} - \theta_{j,-T})(y_{jT} - \theta_{jT})\| - \mathbf{E}\|(y_{j,-T} - \theta_{j,-T})(y_{jT} - \theta_{jT})\|) \\
&+ \frac{1}{J}\sum_{j=1}^{J} (\mathbf{E}\|(y_{j,-T} - \theta_{j,-T})(y_{jT} - \theta_{jT})\| + \|\Sigma_{j,T,-T}\|) \\
= &(\mathrm{A})_J + (\mathrm{B})_J.
\end{aligned}
$$

I show that $(\mathrm{A})_J = o_p(1)$ and $(\mathrm{B})_J = O(1)$, from which (33) will follow.

To show $(\mathrm{A})_J \xrightarrow{p} 0$, it suffices to show that the variance of the summand is bounded

over j, since then it converges to zero in L^2. Observe that

$$\begin{aligned}
&\operatorname{var}(\|(y_{j,-T}-\theta_{j,-T})(y_{jT}-\theta_{jT})\|)\\
\leq&\mathbf{E}\|(y_{j,-T}-\theta_{j,-T})(y_{jT}-\theta_{jT})\|^2\\
\leq&4\sum_{t=1}^{T-1}\mathbf{E}(|y_{jT}|^2+|\theta_{jT}|^2)(|y_{j,t}|^2+|\theta_{j,t}|^2) \qquad (34)\\
\leq&4\sum_{t=1}^{T-1}((\mathbf{E}|y_{jT}|^4\,\mathbf{E}|y_{j,t}|^4)^{1/2}+|\theta_{jT}|^2\,\mathbf{E}|y_{j,t}|^2+|\theta_{j,t}|^2\,\mathbf{E}|y_{jT}|^2+|\theta_{jT}|^2|\theta_{j,t}|^2),
\end{aligned}$$

where the last inequality follows by Cauchy-Schwarz. The expression in the last line is bounded uniformly over j, and thus the variance term is as well. Because $\mathbf{E}\|(y_{j,-T}-\theta_{j,-T})(y_{jT}-\theta_{jT})\| \leq (\mathbf{E}\|(y_{j,-T}-\theta_{j,-T})(y_{jT}-\theta_{jT})\|^2)^{1/2}$ by Jensen's inequality, this also establishes $\limsup_{J\to\infty}(\mathrm{B})_J < \infty$. This concludes the proof for $\sup_{\overline{\mathcal{L}}}|G_J(\Lambda_{T,-T},\Lambda_{-T})| \xrightarrow{p} 0$.

Now, I show that the convergence is in fact in L^1 by establishing uniform integrability of $\sup_{\overline{\mathcal{L}}}|G_J(\Lambda_{T,-T},\Lambda_{-T})|$. To this end, I verify a sufficient condition,

$$\sup_j \mathbf{E}\left(\sup_{\overline{\mathcal{L}}}|G_J(\Lambda_{T,-T},\Lambda_{-T})|\right)^2 < \infty. \qquad (35)$$

First, I bound $|G_J(\Lambda_{T,-T},\Lambda_{-T})|$. Note that

$$\begin{aligned}
&|G_J(\Lambda_{T,-T},\Lambda_{-T})|\\
\leq&\frac{2}{J}\sum_{j=1}^{J}\|B(\Lambda,\Sigma_{j,-T})\|\|(y_{j,-T}-\theta_{j,-T})(y_{jT}-\theta_{jT})-\Sigma_{j,T,-T})\|\\
\leq&K_{T,-T}\underline{\sigma}_M^{-1}\frac{2}{J}\sum_{j=1}^{J}\|(y_{j,-T}-\theta_{j,-T})(y_{jT}-\theta_{jT})-\Sigma_{j,T,-T})\|,
\end{aligned}$$

where the first inequality follows from the triangle inequality and Cauchy-Schwarz

and the second inequality by (29). Hence, it follows that

$$\begin{aligned}
&\mathbf{E}\sup_{\overline{\mathcal{L}}}|G_J(\Lambda_{T,-T},\Lambda_{-T})|^2 \\
\leq& K_{T,-T}^2\underline{\sigma}_M^{-2}\,\mathbf{E}\left(\frac{2}{J}\sum_{j=1}^{J}\|(y_{j,-T}-\theta_{j,-T})(y_{jT}-\theta_{jT})-\Sigma_{j,T,-T}\|\right)^2 \\
\leq& K_{T,-T}^2\underline{\sigma}_M^{-2}\frac{4}{J}\sum_{j=1}^{J}\mathbf{E}\|(y_{j,-T}-\theta_{j,-T})(y_{jT}-\theta_{jT})-\Sigma_{j,T,-T}\|^2 \\
\leq& K_{T,-T}^2\underline{\sigma}_M^{-2}\frac{8}{J}\sum_{j=1}^{J}(\mathbf{E}\|(y_{j,-T}-\theta_{j,-T})(y_{jT}-\theta_{jT})\|^2+\|\Sigma_{j,T,-T}\|^2),
\end{aligned}$$

where the second inequality follows from Cauchy-Schwarz. Since I have shown that the summand in the last line is bounded over j in (34), we have (35). We conclude that $\sup_{\overline{\mathcal{L}}}|G_J(\Lambda_{T,-T},\Lambda_{-T})| \xrightarrow{L^1} 0$.

I follow these same steps for $(\text{IV})_J$. I define

$$H_J(\Lambda_{T,-T},\Lambda_{-T},\Sigma_{j,-T}) := \frac{2}{J}\sum_{j=1}^{J}B(\Lambda,\Sigma_{j,-T})'\theta_{j,-T}(y_{jT}-\theta_{jT}).$$

For pointwise convergence, note that

$$\begin{aligned}
\text{var}(B(\Lambda,\Sigma_{j,-T})'\theta_{j,-T}y_{jT}) =& (B(\Lambda,\Sigma_{j,-T})'\theta_{j,-T})^2\Sigma_{jT} \\
\leq& \|B(\Lambda,\Sigma_{j,-T})\|^2\|\theta_{j,-T}\|^2\Sigma_{jT} \\
\leq& K_{T,-T}^2\underline{\sigma}_M^{-2}\|\theta_{j,-T}\|^2\Sigma_{jT},
\end{aligned}$$

where the first inequality follows by Cauchy-Schwarz and the second inequality by (29). The expression in the last line is bounded over j, and thus $H_J(\Lambda_{T,-T},\Lambda_{-T},\Sigma_{j,-T})$ converges to zero in L^2.

Now, I show that $H_J(\Lambda_{T,-T}, \Lambda_{-T}, \Sigma_{j,-T})$ satisfies a Lipschitz condition. I have

$$\begin{aligned}
&|H_J(\Lambda_{T,-T}, \Lambda_{-T}) - H_J(\widetilde{\Lambda}_{T,-T}, \widetilde{\Lambda}_{-T})| \\
\leq &\frac{2}{J}\sum_{j=1}^{J}|(B(\Lambda, \Sigma_{j,-T}) - B(\widetilde{\Lambda}, \Sigma_{j,-T}))'\theta_{j,-T}(y_{j,T} - \theta_{j,T})| \\
\leq &\frac{2}{J}\sum_{j=1}^{J}\|B(\Lambda, \Sigma_{j,-T}) - B(\widetilde{\Lambda}, \Sigma_{j,-T})\|\|\theta_{j,-T}\||y_{jT} - \theta_{jT}| \\
\leq &((\underline{\sigma}_{\Sigma}^{-1} \vee \underline{\sigma}_{\Sigma}^{-2})\frac{2}{J}\sum_{j=1}^{J}\|\theta_{j,-T}\||y_{jT} - \theta_{jT}|,
\end{aligned}$$

where the second inequality is by Cauchy-Schwarz and the third inequality follows from (32). The fact that $\frac{2}{J}\sum_{j=1}^{J}\|\theta_{j,-T}\||y_{jT} - \theta_{jT}| = O_p(1)$ follows from similar, but simpler, steps we have taken to show (33). This implies $\sup_{\mathcal{L}}|H_J(\Lambda_{T,-T}, \Lambda_{-T})| \xrightarrow{p} 0$. Again, following the same arguments we have used to show (35), we can easily show that $\sup_{\mathcal{L}}|H_J(\Lambda_{T,-T}, \Lambda_{-T})|$ is uniformly integrable, from which it follows that $\sup_{\mathcal{L}}|H_J(\Lambda_{T,-T}, \Lambda_{-T})| \xrightarrow{L^1} 0$. This concludes the proof for the first term of the right-hand side of (28) converging to zero in L^1.

For the second term of the right-hand side of (28), note that

$$\begin{aligned}
&(B(\Lambda, \Sigma_{j,-T})'y_{j,-T} - \theta_{j,T})^2 \\
=&(B(\Lambda, \Sigma_{j,-T})'y_{j,-T} - B(\Lambda, \Sigma_{j,-T})'\theta_{j,-T} + B(\Lambda, \Sigma_{j,-T})'\theta_{j,-T} - \theta_{j,T})^2 \\
=&(B(\Lambda, \Sigma_{j,-T})'y_{j,-T} - B(\Lambda, \Sigma_{j,-T})'\theta_{j,-T})^2 + (B(\Lambda, \Sigma_{j,-T})'\theta_{j,-T} - \theta_{j,T})^2 \\
&+ 2(B(\Lambda, \Sigma_{j,-T})'y_{j,-T} - B(\Lambda, \Sigma_{j,-T})'\theta_{j,-T})(B(\Lambda, \Sigma_{j,-T})'\theta_{j,-T} - \theta_{j,T}).
\end{aligned}$$

Furthermore, I have

$$\begin{aligned}
&\mathbf{E}(B(\Lambda, \Sigma_{j,-T})'y_{j,-T} - B(\Lambda, \Sigma_{j,-T})'\theta_{j,-T})^2 \\
=&\operatorname{var}(B(\Lambda, \Sigma_{j,-T})'y_{j,-T}) \\
=&B(\Lambda, \Sigma_{j,-T})'\Sigma_{j,-T}B(\Lambda, \Sigma_{j,-T}).
\end{aligned}$$

Hence, it follows that

$$
\begin{aligned}
&\left|\tfrac{1}{J}\textstyle\sum_{j=1}^{J}(B(\Lambda,\Sigma_{j,-T})'y_{j,-T}-\theta_{j,T})^2-\tfrac{1}{J}\sum_{j=1}^{J}(B(\Lambda,\Sigma_{j,-1})'y_{j,-1}-\theta_{j,T+1})^2\right| \\
\leq&\left|\tfrac{1}{J}\textstyle\sum_{j=1}^{J}\left((B(\Lambda,\Sigma_{j,-T})'(y_{j,-T}-\theta_{j,-T}))^2-B(\Lambda,\Sigma_{j,-T})'\Sigma_{j,-T}B(\Lambda,\Sigma_{j,-T})\right)\right| \\
&+2\left|\tfrac{1}{J}\textstyle\sum_{j=1}^{J}(B(\Lambda,\Sigma_{j,-T})'(y_{j,-T}-\theta_{j,-T}))(B(\Lambda,\Sigma_{j,-T})'\theta_{j,-T}-\theta_{j,T})\right| \\
&+\left|\tfrac{1}{J}\textstyle\sum_{j=1}^{J}\left((B(\Lambda,\Sigma_{j,-1})'(y_{j,-1}-\theta_{j,-1}))^2-B(\Lambda,\Sigma_{j,-1})'\Sigma_{j,-1}B(\Lambda,\Sigma_{j,-1})\right)\right| \\
&+2\left|\tfrac{1}{J}\textstyle\sum_{j=1}^{J}(B(\Lambda,\Sigma_{j,-1})'(y_{j,-1}-\theta_{j,-1}))(B(\Lambda,\Sigma_{j,-1})'\theta_{j,-1}-\theta_{j,-1})\right| \\
&+\left|\tfrac{1}{J}\textstyle\sum_{j=1}^{J}\left(B(\Lambda,\Sigma_{j,-T})'\Sigma_{j,-T}B(\Lambda,\Sigma_{j,-T})-B(\Lambda,\Sigma_{j,-1})'\Sigma_{j,-1}B(\Lambda,\Sigma_{j,-1})\right)\right| \\
&+\left|\tfrac{1}{J}\textstyle\sum_{j=1}^{J}\left((B(\Lambda,\Sigma_{j,-T})'\theta_{j,-T}-\theta_{j,T})^2-(B(\Lambda,\Sigma_{j,-1})'\theta_{j,-1}-\theta_{j,T+1})^2\right)\right| \\
=&(\mathrm{I})_J+(\mathrm{II})_J+(\mathrm{III})_J+(\mathrm{IV})_J+(\mathrm{V})_J+(\mathrm{VI})_J.
\end{aligned}
$$

I show that each of the six terms converges to zero uniformly over $\mathcal{L}$ in the L^1 sense. The proof for the first four terms are extremely similar. Hence, I provide a proof for only $(\mathrm{I})_J$, and a sketch for the other three terms. Note that the terms $(\mathrm{V})_J$ and $(\mathrm{VI})_J$ are nonrandom.

Note that the summand in $(\mathrm{I})_J$ can be written as

$$
\mathrm{tr}(B(\Lambda,\Sigma_{j,-T})B(\Lambda,\Sigma_{j,-T})'((y_{j,-T}-\theta_{j,-T})(y_{j,-T}-\theta_{j,-T})'-\Sigma_{j,-T})),
$$

which has mean zero. Hence, if the expectation of the square of this term is bounded over j, then $(\mathrm{I})_J \xrightarrow{L^2} 0$. We have

$$
\begin{aligned}
&|\mathrm{tr}(B(\Lambda,\Sigma_{j,-T})B(\Lambda,\Sigma_{j,-T})'((y_{j,-T}-\theta_{j,-T})(y_{j,-T}-\theta_{j,-T})'-\Sigma_{j,-T}))| \\
\leq&|\mathrm{tr}(B(\Lambda,\Sigma_{j,-T})B(\Lambda,\Sigma_{j,-T})'(y_{j,-T}-\theta_{j,-T})(y_{j,-T}-\theta_{j,-T})')| \\
&+|\mathrm{tr}(B(\Lambda,\Sigma_{j,-T})B(\Lambda,\Sigma_{j,-T})'\Sigma_{j,-T})| \\
\leq&\|B(\Lambda,\Sigma_{j,-T})\|^2(\|y_{j,-T}-\theta_{j,-T}\|^2+\mathrm{tr}(\Sigma_{j,-T})),
\end{aligned}
$$

where the last inequality follows from von Neumann's trace inequality and the equivalence between the largest singular value of the outer product of a vector and its

squared L^2 norm. It follows that

$$\begin{aligned}
&\mathbf{E}\ \mathrm{tr}(B(\Lambda,\Sigma_{j,-T})B(\Lambda,\Sigma_{j,-T})'((y_{j,-T}-\theta_{j,-T})(y_{j,-T}-\theta_{j,-T})'-\Sigma_{j,-T}))^2\\
\leq&\mathbf{E}\,\|B(\Lambda,\Sigma_{j,-T})\|^4(\|y_{j,-T}-\theta_{j,-T}\|^2+\mathrm{tr}(\Sigma_{j,-T}))^2\\
\leq&\mathbf{E}\ K_{T,-T}^4\underline{\sigma}_\Sigma^{-4}(8\|y_{j,-T}\|^4+8\|\theta_{j,-T}\|^4+\mathrm{tr}(\Sigma_{j,-T})^2+4(\|y_{j,-T}\|^2+\|\theta_{j,-T}\|^2)\,\mathrm{tr}(\Sigma_{j,-T})),
\end{aligned}$$

where the term in the last line is bounded over j. This shows that $(\mathrm{I})_J \xrightarrow{L^2} 0$.

Now, to obtain a uniform convergence result, write

$$\begin{aligned}
&G_{I,J}(\Lambda_{T,-T},\Lambda_{-T})\\
=&\frac{1}{J}\sum_{j=1}^{J}\mathrm{tr}(B(\Lambda,\Sigma_{j,-T})B(\Lambda,\Sigma_{j,-T})'((y_{j,-T}-\theta_{j,-T})(y_{j,-T}-\theta_{j,-T})'-\Sigma_{j,-T})).
\end{aligned}$$

For any two $x,\widetilde{x}\in\mathrm{R}^{T-1}$, we have

$$\|xx'-\widetilde{x}\widetilde{x}'\|\leq\|x-\widetilde{x}\|(\|x\|+\|\widetilde{x}\|),$$

where the inequality holds by adding and subtracting $x\widetilde{x}'$, applying the triangle inequality, and then Cauchy-Schwarz. This, combined with (29) and (32), gives

$$\begin{aligned}
&\|B(\Lambda,\Sigma_{j,-T})B(\Lambda,\Sigma_{j,-T})'-B(\widetilde{\Lambda},\Sigma_{j,-T})B(\widetilde{\Lambda},\Sigma_{j,-T})'\|\\
\leq&2\underline{\sigma}_\Sigma^{-1}K_{T,-T}\|B(\Lambda,\Sigma_{j,-T})-B(\widetilde{\Lambda},\Sigma_{j,-T})\|\\
\leq&2\underline{\sigma}_\Sigma^{-1}(\underline{\sigma}_\Sigma^{-1}\vee\underline{\sigma}_\Sigma^{-2})K_{T,-T}\|(\Lambda_{T,-T},\Lambda_{-T})-(\widetilde{\Lambda}_{T,-T},\widetilde{\Lambda}_{-T})\|,
\end{aligned}\tag{36}$$

which shows that $B(\Lambda,\Sigma_{j,-T})B(\Lambda,\Sigma_{j,-T})$ is Lipschitz. This will translate into a Lipschitz condition on $G_{I,J}(\Lambda_{T,-T},\Lambda_{-T})$. For simplicity, I write

$$B^2(\Lambda,\Sigma_{j,-T})=B(\Lambda,\Sigma_{j,-T})B(\Lambda,\Sigma_{j,-T})'.$$

Observe that

$$
\begin{aligned}
&|G_{I,J}(\Lambda_{T,-T},\Lambda_{-T}) - G_{I,J}(\widetilde{\Lambda}_{T,-T},\widetilde{\Lambda}_{-T})| \\
\leq &\left|\frac{1}{J}\sum_{j=1}^{J}\operatorname{tr}((B^2(\Lambda,\Sigma_{j,-T}) - B^2(\widetilde{\Lambda},\Sigma_{j,-T}))'(y_{j,-T}-\theta_{j,-T})(y_{j,-T}-\theta_{j,-T})')\right| \\
&+\left|\frac{1}{J}\sum_{j=1}^{J}\operatorname{tr}((B^2(\Lambda,\Sigma_{j,-T}) - B^2(\widetilde{\Lambda},\Sigma_{j,-T}))'\Sigma_{j,-T})\right| \\
\leq &\frac{1}{J}\sum_{j=1}^{J}\sigma_1(B^2(\Lambda,\Sigma_{j,-T}) - B^2(\widetilde{\Lambda},\Sigma_{j,-T}))\operatorname{tr}((y_{j,-T}-\theta_{j,-T})(y_{j,-T}-\theta_{j,-T})') \\
&+\frac{1}{J}\sum_{j=1}^{J}\sigma_1(B^2(\Lambda,\Sigma_{j,-T}) - B^2(\widetilde{\Lambda},\Sigma_{j,-T}))\operatorname{tr}(\Sigma_{j,-T}),
\end{aligned}
$$

where the second inequality follows from the triangle inequality, von Neumann's trace inequality, and the fact that the sum of the eigenvalues of asymmetric matrix equals its trace. Now, using the fact that the operator norm is bounded by the Frobenius norm, we obtain

$$
\begin{aligned}
&|G_{I,J}(\Lambda_{T,-T},\Lambda_{-T}) - G_{I,J}(\widetilde{\Lambda}_{T,-T},\widetilde{\Lambda}_{-T})| \\
=&2\underline{\sigma}_{\Sigma}^{-1}(\underline{\sigma}_{\Sigma}^{-1}\vee\underline{\sigma}_{\Sigma}^{-2})K_{T,-T}B_J\|(\Lambda_{T,-T},\Lambda_{-T}) - (\widetilde{\Lambda}_{T,-T},\widetilde{\Lambda}_{-T})\|,
\end{aligned}
$$

where $B_J = \frac{1}{J}\sum_{j=1}^{J}\operatorname{tr}((y_{j,-T}-\theta_{j,-T})(y_{j,-T}-\theta_{j,-T})' + \Sigma_{j,-T})$, which is $O_p(1)$ by the law of large numbers. This establishes that $\sup_{\overline{\mathcal{L}}}|G_{I,J}(\Lambda_{T,-T},\Lambda_{-T})| \xrightarrow{p} 0$. Again, the mode of convergence can be strengthened to L^1 by verifying a uniform integrability conditions. To this end, note that the summand in the definition of $G_{I,J}(\Lambda_{T,-T},\Lambda_{-T})$ can be bounded by

$$
\begin{aligned}
&|\operatorname{tr}(B(\Lambda,\Sigma_{j,-T})B(\Lambda,\Sigma_{j,-T})'((y_{j,-T}-\theta_{j,-T})(y_{j,-T}-\theta_{j,-T})' - \Sigma_{j,-T}))| \\
\leq &K_{T,-T}^2\underline{\sigma}_{\Sigma}^{-2}\operatorname{tr}((y_{j,-T}-\theta_{j,-T})(y_{j,-T}-\theta_{j,-T})' + \Sigma_{j,-T})),
\end{aligned}
$$

which follows by the same steps used when showing the Lipschitz condition. Since the expectation of the square of the right-hand side is bounded uniformly over j, it follows that $\sup_j \mathbf{E}\sup_{\overline{\mathcal{L}}}|G_{I,J}(\Lambda_{T,-T},\Lambda_{-T})|^2 < \infty$. This concludes the proof for $(\mathrm{I})_J$, and the exact same steps with "$-T$ replaced with -1" also shows that $(\mathrm{III})_J$ converges

to zero uniformly over $\overline{\mathcal{L}}$ in L^1.

For $(\text{II})_J$, note that

$$\begin{aligned}
&\left|\frac{1}{J}\sum_{j=1}^{J}(B(\Lambda,\Sigma_{j,-T})'(y_{j,-T}-\theta_{j,-T}))(B(\Lambda,\Sigma_{j,-T})'\theta_{j,-T}-\theta_{j,T})\right| \\
\leq &\left|\frac{1}{J}\sum_{j=1}^{J}\theta_{j,-T}'B(\Lambda,\Sigma_{j,-T})B(\Lambda,\Sigma_{j,-T})'(y_{j,-T}-\theta_{j,-T})\right| \\
&+\left|\frac{1}{J}\sum_{j=1}^{J}\theta_{j,T}B(\Lambda,\Sigma_{j,-T})'(y_{j,-T}-\theta_{j,-T})\right|.
\end{aligned}$$

Note that the summand of the first term on the right-hand side can be written as

$$\text{tr}(B(\Lambda,\Sigma_{j,-T})B(\Lambda,\Sigma_{j,-T})'(y_{j,-T}-\theta_{j,-T})\theta_{j,-T}'),$$

which is very similar to the summand of $(\text{I})_J$. The same steps used there go through without any added difficulty. The second term is even simpler, and extremely similar to $(\text{IV})_J$ above in the decomposition of the first term on the right-hand side of (28). The same lines of argument used to establish convergence of such term can be used here to show the desired convergence result. Note that none of the convergence results depend on the choice of sequence $\{((\theta_j',\theta_{j,T+1})',\Sigma_j)\}_{j=1}^{\infty}$ under Assumption 5.1.

Now, it remains to show that $(\text{V})_J$ and $(\text{VI})_J$ converges to zero uniformly over $\overline{\mathcal{L}}$, for almost all sequences $\{((\theta_j',\theta_{j,T+1})',\Sigma_j)\}_{j=1}^{\infty}$. Here, it is convenient to treat $\{((\theta_j',\theta_{j,T+1})',\Sigma_j)\}_{j=1}^{\infty}$. I denote by $\mathbf{E}_f$ the expectation with respect to the random sequence $\{((\theta_j',\theta_{j,T+1})',\Sigma_j)\}_{j=1}^{\infty}$. All almost sure assertions in the remainder of the proof is with respect to the randomness of $\{((\theta_j',\theta_{j,T+1})',\Sigma_j)\}_{j=1}^{\infty}$. It follows that

$$\begin{aligned}
&\left|\frac{1}{J}\sum_{j=1}^{J}\left(B(\Lambda,\Sigma_{j,-T})'\Sigma_{j,-T}B(\Lambda,\Sigma_{j,-T})-B(\Lambda,\Sigma_{j,-1})'\Sigma_{j,-1}B(\Lambda,\Sigma_{j,-1})\right)\right| \\
\leq &\left|\frac{1}{J}\sum_{j=1}^{J}\left(B(\Lambda,\Sigma_{j,-T})'\Sigma_{j,-T}B(\Lambda,\Sigma_{j,-T})-\mathbf{E}_f\,B(\Lambda,\Sigma_{j,-T})'\Sigma_{j,-T}B(\Lambda,\Sigma_{j,-T})\right)\right| \\
&+\left|\frac{1}{J}\sum_{j=1}^{J}\left(B(\Lambda,\Sigma_{j,-1})'\Sigma_{j,-1}B(\Lambda,\Sigma_{j,-1})-\mathbf{E}_f\,B(\Lambda,\Sigma_{j,-1})'\Sigma_{j,-1}B(\Lambda,\Sigma_{j,-1})\right)\right|,
\end{aligned}$$

where in the inequality I use Assumption 5.2 and use the fact that the expectations

are equal. I show that the first term on the right-hand side converges to 0 almost surely, uniformly over $\overline{\mathcal{L}}$. The same result can be shown for the second term using the exact same argument. Define

$$\begin{aligned}
&G_{\mathrm{V},J}(\Lambda_{T,-T}, \Lambda_{-T}) \\
=&\frac{1}{J}\sum_{j=1}^{J} \left(B(\Lambda, \Sigma_{j,-T})'\Sigma_{j,-T}B(\Lambda, \Sigma_{j,-T}) - \mathbf{E}_f\, B(\Lambda, \Sigma_{j,-T})'\Sigma_{j,-T}B(\Lambda, \Sigma_{j,-T})\right).
\end{aligned}$$

Note that since $\mathbf{E}_f\, \sigma_1(\Sigma_j)$ exists, we have

$$\mathbf{E}_f\, B(\Lambda, \Sigma_{j,-T})'\Sigma_{j,-T}B(\Lambda, \Sigma_{j,-T}) \leq K_{T,-T}^{-2}\underline{\sigma}_{\Sigma}^{-2}\, \mathbf{E}_f\, \sigma_1(\Sigma_j) < \infty.$$

Hence, by the strong law of large numbers, we have $G_{\mathrm{V},J}(\Lambda_{T,-T}, \Lambda_{-T}) \to 0$ almost surely. For uniformity over $\overline{\mathcal{L}}$, again I verify a Lipschitz condition for $G_{\mathrm{V},J}(\Lambda_{T,-T}, \Lambda_{-T})$. We have

$$\begin{aligned}
&\left|G_{\mathrm{V},J}(\Lambda_{T,-T}, \Lambda_{-T}) - G_{\mathrm{V},J}(\widetilde{\Lambda}_{T,-T}, \widetilde{\Lambda}_{-T})\right| \\
\leq&\frac{1}{J}\sum_{j=1}^{J}|\mathrm{tr}((B^2(\Lambda, \Sigma_{j,-T}) - B^2(\widetilde{\Lambda}, \Sigma_{j,-T}))\Sigma_{j,-T})| \\
&+ \frac{1}{J}\sum_{j=1}^{J}\mathbf{E}_f|\mathrm{tr}((B^2(\Lambda, \Sigma_{j,-T}) - B^2(\widetilde{\Lambda}, \Sigma_{j,-T}))\Sigma_{j,-T})| \\
\leq&2\underline{\sigma}_{\Sigma}^{-1}(\underline{\sigma}_{\Sigma}^{-1} \vee \underline{\sigma}_{\Sigma}^{-2})K_{T,-T}B_J\|(\Lambda_{T,-T}, \Lambda_{-T}) - (\widetilde{\Lambda}_{T,-T}, \widetilde{\Lambda}_{-T})\|
\end{aligned}$$

where the first inequality follows from multiple applications of the triangle inequality, and the second inequality follows with $B_J = \frac{1}{J}\sum_{j=1}^{J}(\mathrm{tr}(\Sigma_{j,-T}) + E\,\mathrm{tr}(\Sigma_{j,-T}))$ from von Neumann's trace inequality and (36). Let $\overset{a.s.}{\to}$ denote almost sure convergence with respect to the density $f_{(\theta',\theta_{T+1})',M}$. By the strong law of large numbers, it follows that $B_J \overset{a.s.}{\to} 2\,\mathbf{E}_f\,\mathrm{tr}(\Sigma_{j,-t})$. Hence, by Lemma 1 of Andrews (1992), I conclude that $\sup_{\overline{\mathcal{L}}}|G_{\mathrm{V},J}(\Lambda_{T,-T}, \Lambda_{-T})| \overset{a.s.}{\to} 0$.

For $(\text{VI})_{\text{J}}$, the triangle inequality gives

$$\begin{aligned}
&\left|\frac{1}{J}\sum_{j=1}^{J}\left((B(\Lambda,\Sigma_{j,-T})'\theta_{j,-T}-\theta_{j,T})^2-(B(\Lambda,\Sigma_{j,-1})'\theta_{j,-1}-\theta_{j,T+1})^2\right)\right| \\
\leq&\left|\frac{1}{J}\sum_{j=1}^{J}\left((B(\Lambda,\Sigma_{j,-T})'\theta_{j,-T}-\theta_{j,T})^2-\mathbf{E}_f(B(\Lambda,\Sigma_{j,-T})'\theta_{j,-T}-\theta_{j,T})^2\right)\right| \\
&+\left|\frac{1}{J}\sum_{j=1}^{J}\left((B(\Lambda,\Sigma_{j,-1})'\theta_{j,-1}-\theta_{j,T+1})^2-\mathbf{E}_f(B(\Lambda,\Sigma_{j,-1})'\theta_{j,-1}-\theta_{j,T+1})^2\right)\right|.
\end{aligned}$$

Again, I show that the desired convergence result only for the first term since the result for the second term will follow from the exact same steps. Define

$$\begin{aligned}
&G_{\text{VI},J}(\Lambda_{T,-T},\Lambda_{-T}) \\
=&\frac{1}{J}\sum_{j=1}^{J}\left((B(\Lambda,\Sigma_{j,-T})'\theta_{j,-T}-\theta_{j,T})^2-\mathbf{E}_f(B(\Lambda,\Sigma_{j,-T})'\theta_{j,-T}-\theta_{j,T})^2\right).
\end{aligned}$$

To show $G_{\text{VI},J}(\Lambda_{T,-T},\Lambda_{-T})\xrightarrow{a.s.}0$, note that

$$\begin{aligned}
&\mathbf{E}_f(B(\Lambda,\Sigma_{j,-T})'\theta_{j,-T}-\theta_{j,T})^2 \\
\leq&2\,\mathbf{E}_f\,\text{tr}(B^2(\Lambda,\Sigma_{j,-T})\theta_{j,-T}\theta_{j,-T}')+\mathbf{E}_f\,\theta_{jT}^2 \\
\leq&2K_{T,-T}^2\underline{\sigma}_{\Sigma}^{-2}\textstyle\sum_{t=1}^{T-1}\mathbf{E}_f\,\theta_{jt}^2+\mathbf{E}_f\,\theta_{jT}^2<\infty,
\end{aligned}$$

where the second inequality follows because

$$\text{tr}(B^2(\Lambda,\Sigma_{j,-T})\theta_{j,-T}\theta_{j,-T}')\leq\sigma_1(B^2(\Lambda,\Sigma_{j,-T}))\,\text{tr}(\theta_{j,-T}\theta_{j,-T}')$$

due to von Neumann's trace inequality. Hence, by the strong law of large numbers, we have $G_{\text{VI},J}(\Lambda_{T,-T},\Lambda_{-T})\xrightarrow{a.s.}0$.

Once again, I verify a Lipschitz condition to show that this convergence is in fact uniform over $\overline{\mathcal{L}}$. Note that

$$\begin{aligned}
&(B(\Lambda,\Sigma_{j,-T})'\theta_{j,-T}-\theta_{j,T})^2-(B(\widetilde{\Lambda},\Sigma_{j,-T})'\theta_{j,-T}-\theta_{j,T})^2 \\
=&\,\text{tr}((B^2(\Lambda,\Sigma_{j,-T})-B^2(\widetilde{\Lambda},\Sigma_{j,-T}))\theta_{j,-T}\theta_{j,-T}') \\
&-2(B(\Lambda,\Sigma_{j,-T})-B(\Lambda,\Sigma_{j,-T}))'\theta_{j,-T}\theta_{jT},
\end{aligned}$$

and thus, by (32) and (36),

$$\begin{aligned}
&|(B(\Lambda,\Sigma_{j,-T})'\theta_{j,-T}-\theta_{j,T})^2-(B(\widetilde{\Lambda},\Sigma_{j,-T})'\theta_{j,-T}-\theta_{j,T})^2| \\
=&|\mathrm{tr}((B^2(\Lambda,\Sigma_{j,-T})-B^2(\widetilde{\Lambda},\Sigma_{j,-T}))\theta_{j,-T}\theta_{j,-T}')| \\
&+2|(B(\Lambda,\Sigma_{j,-T})-B(\Lambda,\Sigma_{j,-T}))'\theta_{j,-T}\theta_{jT}| \\
\leq&2\underline{\sigma}_\Sigma^{-1}(\underline{\sigma}_\Sigma^{-1}\vee\underline{\sigma}_\Sigma^{-2})K_{T,-T}\|\theta_{j,-T}\|^2\|(\Lambda_{T,-T},\Lambda_{-T})-(\widetilde{\Lambda}_{T,-T},\widetilde{\Lambda}_{-T})\| \\
&+2(\underline{\sigma}_\Sigma^{-1}\vee\underline{\sigma}_\Sigma^{-2})\|\theta_{j,-T}\theta_{jT}\|\|(\Lambda_{T,-T},\Lambda_{-T})-(\widetilde{\Lambda}_{T,-T},\widetilde{\Lambda}_{-T})\| \\
:=&B_j\|(\Lambda_{T,-T},\Lambda_{-T})-(\widetilde{\Lambda}_{T,-T},\widetilde{\Lambda}_{-T})\|
\end{aligned}$$

Likewise, we have

$$\begin{aligned}
&|\mathbf{E}_f(B(\Lambda,\Sigma_{j,-T})'\theta_{j,-T}-\theta_{j,T})^2-\mathbf{E}_f(B(\widetilde{\Lambda},\Sigma_{j,-T})'\theta_{j,-T}-\theta_{j,T})^2| \\
\leq&2\underline{\sigma}_\Sigma^{-1}(\underline{\sigma}_\Sigma^{-1}\vee\underline{\sigma}_\Sigma^{-2})K_{T,-T}\,\mathbf{E}_f\|\theta_{j,-T}\|^2\|(\Lambda_{T,-T},\Lambda_{-T})-(\widetilde{\Lambda}_{T,-T},\widetilde{\Lambda}_{-T})\| \\
&+2(\underline{\sigma}_\Sigma^{-1}\vee\underline{\sigma}_\Sigma^{-2})\|\mathbf{E}_f\,\theta_{j,-T}\theta_{jT}\|\|(\Lambda_{T,-T},\Lambda_{-T})-(\widetilde{\Lambda}_{T,-T},\widetilde{\Lambda}_{-T})\| \\
:=&B_{E_f}\|(\Lambda_{T,-T},\Lambda_{-T})-(\widetilde{\Lambda}_{T,-T},\widetilde{\Lambda}_{-T})\|.
\end{aligned}$$

Combining the two inequalities, we have

$$\begin{aligned}
&|G_{\mathrm{VI},J}(\Lambda_{T,-T},\Lambda_{-T})-G_{\mathrm{VI},J}(\widetilde{\Lambda}_{T,-T},\widetilde{\Lambda}_{-T})| \\
\leq&\left(\frac{1}{J}\sum_{j=1}^J B_j+B_{E_f}\right)\|(\Lambda_{T,-T},\Lambda_{-T})-(\widetilde{\Lambda}_{T,-T},\widetilde{\Lambda}_{-T})\|
\end{aligned}$$

Hence, it suffices to show that $\frac{1}{J}\sum_{j=1}^J B_j \overset{a.s.}{\to} B$ for some fixed B. By the strong law of large numbers, we have

$$\begin{aligned}
&\frac{1}{J}\sum_{j=1}^J\|\theta_{j,-T}\|^2\overset{a.s.}{\to}\mathbf{E}\|\theta_{j,-T}\|^2<\infty \\
&\frac{1}{J}\sum_{j=1}^J\|\theta_{j,-T}\theta_{jT}\|\overset{a.s.}{\to}\mathbf{E}\|\theta_{j,-T}\theta_{jT}\|\leq(\mathbf{E}\,\theta_{j,T}^2\,\mathbf{E}\|\theta_{j,-T}\|^2)^{\frac{1}{2}}<\infty,
\end{aligned}$$

which establishes that $\frac{1}{J}\sum_{j=1}^J B_j$ indeed converges almost surely to a finite value, which concludes the proof.

Appendix B Unbalanced panels

In practice, it is rarely the case that the empirical researcher has a balanced panel. This corresponds to the case where for each j, one observes only a subvector of y_j. Note that the full vector, y_j, is now in some sense hypothetical, but it is convenient to consider that one observes a subvector of this full vector. I show that the URE approach remains valid in the case of unbalanced panels, with minor adjustments.

Let $t_1^j < \cdots < t_{o_j}^j$ denote the time periods t for which observations for j exist, where $1 \leq o_j \leq T$. Let O_j denote the $o_j \times T$ matrix that picks out only the observed periods, $O_j = (e_{t_1^j}, \ldots, e_{t_{o_j}^j})'$ with $e_\ell \in \mathbf{R}^T$ denoting the ℓth standard basis vector. I define the subvector or submatrix corresponding to the observed periods of y_j, θ_j, Σ_j, μ and Λ as $y_j^o = O_j y_j$, $\theta_j^o = O_j \theta_j$, $\Sigma_j^o = O_j \Sigma_j O_J'$, $\Lambda_j^o = O_j \Lambda O_j'$ and $\mu_j^o = O_j \mu$. Again, consider the second level model $\theta_j \overset{\text{i.i.d.}}{\sim} N(\mu, \Lambda)$.

The aim is to estimate $\theta^o := (\theta_1^{o\prime}, \ldots, \theta_J^{o\prime})'$, and thus a natural class of estimators I consider is given by the posterior mean

$$\widehat{\theta}_j^o(\mu, \Lambda) = E[\theta_j^o | y_j^o] = \Sigma_j^o(\Lambda_j^o + \Sigma_j^o)^{-1}\mu_j^o + \Lambda_j^o(\Lambda_j^o + \Sigma_j^o)^{-1} y_j^o.$$

The loss function is modified so that it takes into account the different number of observations for each j:

$$\ell^o(\widehat{\theta}^o(\mu, \Lambda), \theta^o) := \frac{1}{J}\sum_{j=1}^{J} \frac{1}{o_j} \|\widehat{\theta}_j^o(\mu, \Lambda) - \theta_j^o\|^2,$$

where I write the summand as $\ell_j^o(\widehat{\theta}_j^o(\mu, \Lambda), \theta_j^o)$. Note that in the balanced case of $o_j = T$ for all $j = 1, \ldots, J$, this coincides with the loss function we have been using but scaled by $1/T$. The scaling by $1/o_j$ ensures that each (j, t) component is weighted equally across all j and t.

An unbiased risk estimate of $\widehat{\theta}_j^o(\mu, \Sigma)$ is given by $\mathbf{URE}^o(\mu, \Lambda) = \frac{1}{J}\sum_{j=1}^{J} \mathbf{URE}_j^o(\mu, \Lambda)$, where

$$\begin{aligned}&\mathbf{URE}_j^o(\mu, \Lambda)\\ =&\frac{1}{o_j}\left(\operatorname{tr}(\Sigma_j^o) - 2\operatorname{tr}((\Lambda_j^o + \Sigma_j^o)^{-1}\Sigma_j^{o2}) + (y_j^o - \mu_j^o)'[(\Lambda_j^o + \Sigma_j^o)^{-1}\Sigma_j^{o2}(\Lambda_j^o + \Sigma_j^o)^{-1}](y_j^o - \mu^o)\right).\end{aligned}$$

Theorem B.1 (Uniform convergence of $\mathbf{URE}^o(\mu, \Lambda)$)**.** *Suppose Assumptions 4.1 and*

4.2 hold. Then,

$$\sup_{(\mu,\Lambda)\in\mathcal{M}_J\times\mathcal{S}_T^+} |\mathbf{URE}^o(\mu,\Lambda) - \ell^o(\widehat{\theta}^o(\mu,\Lambda),\theta^o)| \xrightarrow{L^1} 0. \tag{37}$$

As a consequence, under the same assumptions we required for the balanced the case, the URE estimator obtains the oracle risk in the unbalanced case as well.

Proof. The difference between the risk estimate and the loss function (for j) is

$$\begin{aligned}
&|\mathbf{URE}^o(\mu,\Lambda) - \ell^o(\widehat{\theta}^o(\mu,\Lambda),\theta^o)| \\
=&\left|\frac{1}{J}\sum_{j=1}^{J}\frac{1}{o_j}\left(y_j^{o\prime}y_j^o - \theta_j^{o\prime}\theta_j^o + \mathrm{tr}(\Sigma_j^o)\right)\right| + \left|\frac{1}{J}\sum_{j=1}^{J}\frac{2}{o_j}\mathrm{tr}(\Lambda_j^o(\Lambda_j^o+\Sigma_j^o)^{-1}(y_j^oy_j^{o\prime} - \theta_j^oy_j^{o\prime} - \Sigma_j^o))\right| \\
&+\left|\frac{1}{J}\sum_{j=1}^{J}\frac{2}{o_j}\mu_j^{o\prime}(\Lambda_j^o+\Sigma_j^o)^{-1}\Sigma_j^o(y_j^o-\theta_j^o)\right| \\
=&(I)_J + (II)_J + (III)_J,
\end{aligned}$$

which follows from the same steps as in (12) and the triangle inequality. Hence it suffices to show that each of the three terms converges to zero in L^1, uniformly over $(\mu,\Lambda)\in\mathcal{M}_J\times\mathcal{S}_T^+$.

The proof is a minor modification of the proofs for Theorems 4.1 and 4.3, and thus I only point out the modifications that must be made. The convergence of the first term can be shown, by noting that

$$E\left(\frac{1}{o_j}\left(y_j^{o\prime}y_j^o - \theta_j^{o\prime}\theta_j^o + \mathrm{tr}(\Sigma_j^o)\right)\right)^2 \le E\left(y_j^{o\prime}y_j^o - \theta_j^{o\prime}\theta_j^o + \mathrm{tr}(\Sigma_j^o)\right)^2,$$

and the right-hand side is bounded uniformly over j for the same reason that this term without the o superscripts is bounded.

For $(II)_J$, the main step is to show that the derivative of

$$\mathrm{tr}(\Sigma_j^o(\Lambda_j^o+\Sigma_j^o)^{-1}(y_j^oy_j^{o\prime} - \theta_j^oy_j^{o\prime} - \Sigma_j^o))$$

with respect to $\widetilde{\Lambda} = (\underline{\sigma}_\Sigma I_T + \Lambda)^{-1}$ is bounded uniformly over j. Under such reparametrization, we have $\Lambda_j^o = O_j\widetilde{\Lambda}O_j' - \underline{\sigma}_\Sigma I_{o_j}$. Define $\widetilde{\Lambda}_j^o = O_j\widetilde{L}O_j'$. From similar calculations in

the proof of Theorem 4.1 and the chain rule for matrix derivatives, I obtain

$$\begin{aligned}&\frac{\partial}{\partial\widetilde{\Lambda}}\operatorname{tr}(\Sigma_j^o(\Lambda_j^o+\Sigma_j^o)^{-1}(y_j^o y_j^{o\prime}-\theta_j^o y_j^{o\prime}-\Sigma_j^o))\\ =&O_j'\widetilde{\Lambda}^{o-1}(\widetilde{\Lambda}^{o-1}-\underline{\sigma}_\Sigma I_{o_j}+\Sigma_j^o)^{-1}\Sigma_j^o(y_j^o y_j^{o\prime}-y_j^o\theta_j^{o\prime}-\Sigma_j^o)(\widetilde{\Lambda}^{o-1}-\underline{\sigma}_\Sigma I_{o_j}+\Sigma_j^o)^{-1}\widetilde{\Lambda}^{o-1}O_j.\end{aligned}$$

Observe that the norm of the last expression is the same with the norm of the same expression with the O_j' term at the beginning and the O_j at the end removed. Hence, the norm of this term can be bounded using the exact same arguments given in the paragraph containing (17).

To show the convergence of $(III)_J$, observe that

$$\left|\frac{1}{J}\sum_{j=1}^{J}\frac{2}{o_j}\mu_j^{o\prime}(\Lambda_j^o+\Sigma_j^o)^{-1}\Sigma_j^o(y_j^o-\theta_j^o)\right|\leq\|\mu\|\left\|\frac{1}{J}\sum_{j=1}^{J}\frac{2}{o_j}O_j'(\Lambda_j^o+\Sigma_j^o)^{-1}\Sigma_j^o(y_j^o-\theta_j^o)\right\|,$$

by Cauchy-Schwarz. Taking the expectation of the supremum of the right-hand side over $(\mu,\Lambda)\in\mathcal{M}_J\times\mathcal{S}_T^+$, and then applying the Cauchy-Scwharz inequality again, it follows that it is sufficient to show

$$\begin{aligned}&\sup_J E\sup_{\mu\in\mathcal{M}_J}\|\mu\|^2<\infty\text{ and}\\ &E\sup_{\Lambda\in\mathcal{S}_T^+}\left\|\frac{1}{J}\sum\nolimits_{j=1}^{J}\frac{2}{o_j}O_j'(\Lambda_j^o+\Sigma_j^o)^{-1}\Sigma_j^o(y_j^o-\theta_j^o)\right\|^2\to 0.\end{aligned}$$

The first line has already been established in the proof of Theorem 4.2. The second line can be shown by a similar derivative calculation to that used in establishing the convergence of $(II)_J$ and, again, the same lines of argument used in the proof of Theorem 4.2. □

Appendix C Weighted MSE

Here, I consider the case where the loss function is weighted in the sense

$$R(\theta,\widehat{\theta})=\tfrac{1}{J}\,\mathbf{E}_\theta(\widehat{\theta}-\theta)'W(\widehat{\theta}-\theta),$$

where W is a positive semidefinite $T \times T$ matrix. While I assume that the weight is the same for each j, all results in this section go though if one allows a different weight W_j for each j as long as $\sup_j \sigma_1(W_j) < \infty$.

The corresponding risk estimate is given as

$$\begin{aligned}
&\mathbf{URE}_j^W(\mu, \Lambda) \\
=&\operatorname{tr}(W\Sigma_j) - 2\operatorname{tr}((\Lambda + \Sigma_j)^{-1}\Sigma_j W \Sigma_j) \\
&+ (y_j - \mu)'[(\Lambda + \Sigma_j)^{-1}\Sigma_j W \Sigma_j (\Lambda + \Sigma_j)^{-1}](y_j - \mu).
\end{aligned}$$

It is straightforward to see that analogous versions of Theorems 4.1, 4.2, and 4.3 go through for any positive semidefinite weight matrix W, which implies that minimizing the risk estimate obtains the oracle under weighted losses as well. To see this, note that the difference between the risk estimate and the loss is given as

$$\begin{aligned}
&\mathbf{URE}_j^W(\mu, \Lambda) - (\widehat{\theta}_j(\mu, \Lambda) - \theta_j)' W (\widehat{\theta}_j(\mu, \Lambda) - \theta_j) \\
=&\operatorname{tr}(W\Sigma_j) - 2\operatorname{tr}((\Lambda + \Sigma_j)^{-1}\Sigma_j W \Sigma_j) \\
&+ (y_j - \mu)'[(\Lambda + \Sigma_j)^{-1}\Sigma_j W \Sigma_j (\Lambda + \Sigma_j)^{-1}](y_j - \mu) \\
&- (y_j - \theta_j - \Sigma_j(\Lambda + \Sigma_j)^{-1}(y_j - \mu))' W (y_j - \theta_j - \Sigma_j(\Lambda + \Sigma_j)^{-1}(y_j - \mu)) \\
=&\operatorname{tr}(W\Sigma_j) - 2\operatorname{tr}((\Lambda + \Sigma_j)^{-1}\Sigma_j W \Sigma_j) - (y_j - \theta_j)' W (y_j - \theta_j) \\
&+ 2(y_j - \mu)'(\Lambda + \Sigma_j)^{-1}\Sigma_j W (y_j - \theta_j) \\
=&y_j' W y_j - \theta_j' W \theta_j - \operatorname{tr}(W\Sigma_j) - 2\operatorname{tr}(W\Lambda(\Lambda + \Sigma_j)^{-1}(y_j y_j' - y_j \theta_j' - \Sigma_j)) \\
&- 2\mu'(\Lambda + \Sigma_j)^{-1}\Sigma_j W (y_j - \theta_j).
\end{aligned} \tag{38}$$

Since W does not vary with j, the exact same proofs given for the theorems under $W = I_T$ all go through without any additional assumptions. Since the proof is essentially just a repetition of the provided proofs with additional $\sigma_1(W)$ terms appearing in numerous places, I omit the proof for the weighted case.

To see why considering weighted loss functions can be interesting, let Q denote any $R \times T$ matrix, and suppose that the interest is in estimating the linear combinations $\{Q\theta_j\}_{j=1}^J$ rather than the original vector of the true means. Under the second level model of $\theta_j \overset{\text{i.i.d.}}{\sim} N(\mu, \Lambda)$, the posterior mean of the parameter of interest is given as

$Q\widehat{\theta}_j(\mu,\Lambda)$ and the resulting loss is

$$\begin{aligned}&\frac{1}{J}\sum_{j=1}^{J}(Q\widehat{\theta}_j(\mu,\Lambda)-Q\theta_j)'(Q\widehat{\theta}_j(\mu,\Lambda)-Q\theta_j)\\ =&\frac{1}{J}\sum_{j=1}^{J}(\widehat{\theta}_j(\mu,\Lambda)-\theta_j)'Q'Q(\widehat{\theta}_j(\mu,\Lambda)-\theta_j),\end{aligned}$$

which is the original loss function with weight matrix $W = Q'Q$. Hence, the weighted loss function arises naturally whenever a linear function of the parameter is of interest.

Weighted loss in the unbalanced case. Suppose the empirical researcher is interested in a linear combination of the true mean vector under the unbalanced case. In such a scenario, the weights should be adjusted so that it reflects the missing cells. To see this, consider the case where $Q = \frac{1}{T}\mathbf{1}_T'$ so that the interest is in the time average of the mean for each unit j. Consider the extreme case where there is only one observation available, at period 1, for j. Then, if one does not wish to distinguish between different teachers, it seems reasonable that the parameter of interest in this case should be θ_{j1} rather than $\frac{1}{T}\theta_{j1}$. However, if the researcher mechanically takes QO_j', then she will end up with this latter term. When Q is a row vector with only nonnegative entries, a natural way to resolve this is to rescale QO_j so that the sum of its entries is equal to the sum of the entries of Q by defining $Q_j^o = \frac{Q\mathbf{1}_T}{\mathbf{1}_{o_j}'QO_j'}QO_j'$. Note that $\frac{Q\mathbf{1}_T}{\mathbf{1}_{o_j}'QO_j'} = T/o_j$ when $Q = \frac{1}{T}\mathbf{1}_T'$, which gives the desired weighting.

In the general case where Q is a $R \times T$ matrix with positive entries, the same can be achieved by scaling QO_j' so that the sum of all the entries are equal the sum of entries of Q. Accordingly, define the scaled version as $Q_j^o = \frac{\mathbf{1}_R'Q\mathbf{1}_T}{\mathbf{1}_R'QO_j'\mathbf{1}_{o_j}}QO_j'$ and the corresponding weight matrix $W_j^o = Q_j^{o\prime}Q_j^o$. In this case, the risk estimate to be minimized is given as $\mathbf{URE}^{o,W}(\mu,\Lambda) = \frac{1}{J}\sum_{j=1}^{J}\mathbf{URE}_j^{o,W}(\mu,\Lambda)$, where

$$\begin{aligned}\mathbf{URE}_j^{o,W}(\mu,\Lambda) :=& \operatorname{tr}(W_j^o\Sigma_j^o) - 2\operatorname{tr}((\Lambda_j^o+\Sigma_j^o)^{-1}\Sigma_j^oW_j^o\Sigma_j^o)\\ &+ (y_j^o-\mu_j^o)'[(\Lambda_j^o+\Sigma_j^o)^{-1}\Sigma_j^oW_j^o\Sigma_j^o(\Lambda_j^o+\Sigma_j^o)^{-1}](y_j^o-\mu^o).\end{aligned}$$

Again, this risk estimate can be shown to converge uniformly to the corresponding

loss

$$\frac{1}{J}\sum_{j=1}^{J}(\widehat{\theta}_j^o(\mu,\Lambda)-\theta_j^o)'W_j^o(\widehat{\theta}_j^o(\mu,\Lambda)-\theta_j^o).$$

Appendix D Details for the empirical exercise

D.1 Estimation of β and σ^2

The coefficient vector on the observables, β, is estimated by OLS on the following demeaned version of the regression formula,

$$\widetilde{y}_{ijt}=\widetilde{X}_{ijt}'\beta+\widetilde{\varepsilon}_{ijt},$$

so that $\widehat{\beta}=(\sum_{j=1}^{J}\sum_{t=1}^{T}\sum_{i=1}^{n_{jt}}\widetilde{X}_{ijt}\widetilde{X}_{ijt}')^{-1}(\sum_{j=1}^{J}\sum_{t=1}^{T}\sum_{i=1}^{n_{jt}}\widetilde{X}_{ijt}\widetilde{y}_{ijt})$. The coefficient estimates are reported in Table 1. The variance of the idiosyncratic term σ is estimated by dividing the sum of squared residuals by the appropriate degrees of freedom,

$$\widehat{\sigma}:=\frac{1}{\sum_{j=1}^{J}\sum_{t=1}^{T}(n_{jt}-1)-10}\sum_{j=1}^{J}\sum_{t=1}^{T}\sum_{i=1}^{n_{jt}}(\widetilde{y}_{ijt}-\widetilde{X}_{ijt}'\widehat{\beta})^2$$

D.2 Definition of the estimators for fixed effects

The least squares estimator for the time-varying fixed effect α_{jt} is given as

$$\widehat{\alpha}_{jt}=\frac{1}{n_{jt}}\sum_{i=1}^{n_{jt}}\left(\widetilde{y}_{ijt}-\widetilde{X}_{ijt}'\widehat{\beta}\right),$$

and for the time-invariant case, $\widehat{\alpha}_{j0}=\frac{1}{n_j}\sum_{t=1}^{T}n_{jt}\widehat{\alpha}_{jt}$ with $n_j=\sum_{t=1}^{T}n_{jt}$.

The EBMLE for the time invariant case (i.e. the "conventional estimator") is defined as $\widehat{\alpha}_{j0}^{\text{EBMLE}}=\frac{\widehat{\sigma}/n_j}{\widehat{\sigma}/n_j+\widehat{\lambda}^{\text{EBMLE}}}\widehat{\mu}^{\text{EBMLE}}+\frac{\widehat{\lambda}^{\text{EBMLE}}}{\widehat{\sigma}/n_j+\widehat{\lambda}^{\text{EBMLE}}}\widehat{\alpha}_{j0}$, where $(\widehat{\mu}^{\text{EBMLE}},\widehat{\lambda}^{\text{EBMLE}})$ is obtained by maximizing the likelihood of $\widehat{\alpha}_{j0}$ given by $\widehat{\alpha}_{j0}\overset{\text{indep}}{\sim}N\left(\mu,\frac{\widehat{\sigma}}{n_j}+\lambda\right)$.

To describe the estimators for the time variant case, write $\widehat{\Sigma}_j=\text{diag}(\widehat{\sigma}/n_{j1},\ldots,\widehat{\sigma}/n_{jT})$. The EBMLE and URE estimators for α_j in this case is given by

$$\Sigma_j(\widehat{\Lambda}+\Sigma_j)^{-1}\widehat{\mu}+\widehat{\Lambda}(\widehat{\Lambda}+\Sigma_j)^{-1}\widehat{\alpha}_j$$

Table 1: Parameter estimates for the baseline value-added model (10).

	Dependent variable:	
	ELA score	Math score
Score from previous year	0.629^{***}	0.707^{***}
	(0.002)	(0.002)
Male	-0.069^{***}	0.022^{***}
	(0.003)	(0.002)
Black	-0.090^{***}	-0.112^{***}
	(0.006)	(0.005)
Hispanic	-0.060^{***}	-0.065^{***}
	(0.005)	(0.004)
Asian	0.071^{***}	0.122^{***}
	(0.005)	(0.004)
Multi-Racial	0.016	0.009
	(0.014)	(0.012)
Native American	-0.023	-0.018
	(0.015)	(0.013)
ELL	-0.180^{***}	-0.111^{***}
	(0.006)	(0.004)
SWD	-0.263^{***}	-0.207^{***}
	(0.005)	(0.004)
FL	-0.046^{***}	-0.042^{***}
	(0.003)	(0.003)
Observations	174,239	195,792

Note: *p<0.1; **p<0.05; ***p<0.01

where $(\widehat{\mu}, \widehat{\Lambda})$ is chosen by maximizing the likelihood implied by $\widehat{\alpha}_j \overset{\text{indep}}{\sim} N(\mu, \Lambda + \Sigma_j)$ for EBMLE, and by minimizing $\text{URE}(\mu, \Lambda)$ for the URE estimator with $\widehat{\alpha}_j$ playing the role of y_j in the definition of $\text{URE}(\mu, \Lambda)$.

Appendix E Semiparametric shrinkage

Here, I illustrate how the semiparametric shrinkage idea by Xie et al. (2012) can be extended to this setting. I consider the simple shrinkage estimator that shrinks to the origin. For the univariate model where $T = 1$, the shrinkage estimator that shrinks to the origin can be written as

$$\widehat{\theta}(0, \Lambda) = \frac{\Lambda}{\Lambda + \Sigma_j} y_j.$$

Hence, the estimator is obtained by multiplying a shrinkage factor to the observation. This shinkage factor lies in $[0, 1]$ and shrinks the decreases in Σ_j. Motivated by such observation, Xie et al. (2012) consider a class of semiparametric shrinkage estimators,

$$\widehat{\theta}^b(0, \Lambda) = b(\Sigma_j) y_j,$$

where $b(\cdot)$ is a weakly decreasing function taking values in $[0, 1]$.

To extend this idea to the multivariate setting, recall that the shrinkage matrix is given as

$$\Lambda(\Lambda + \Sigma_j)^{-1} = \Lambda^{\frac{1}{2}}(I + \Lambda^{-\frac{1}{2}}\Sigma_j\Lambda^{-\frac{1}{2}})^{-1}\Lambda^{-\frac{1}{2}}.$$

Replacing the middle term $(I + \Lambda^{-\frac{1}{2}}\Sigma_j\Lambda^{-\frac{1}{2}})^{-1}$ by $B(\Lambda^{-\frac{1}{2}}\Sigma_j\Lambda^{-\frac{1}{2}})$ where $B : \mathcal{S}_T^+ \to \mathcal{S}_T^+$ is decreasing (with respect to the partial ordering $\leq$) and $\sigma_1(B(\cdot)) \leq 1$ is a direct extension of the univariate case.

However, here I assume an additional Lipschitz condition, with known finite Lipschitz constant, on the function $B(\cdot)$ because 1) the partial ordering may end up imposing no restriction at all resulting in severe overfitting and 2) to invoke a uniform convergence result, it is convenient to have a totally bounded parameter space, which can be obtained by giving a bound on the Lipschitz constant of $B(\cdot)$. Under this bound and Assumption 4.1, the URE method (and the corresponding optimality) can be extended to this wider class of estimators in a straightforward manner. Note that the part of the shrinkage matrix we are relaxing, $(I + \Lambda^{-\frac{1}{2}}\Sigma_j\Lambda^{-\frac{1}{2}})^{-1}$, is

Lipschitz as well, so the parametric estimator considered in the main text is indeed nested in this semiparametric class. However, computation is extremely difficult; the optimization problem that must be solved is with respect to a $T(T+1)/2$ matrix and a function of $T(T+1)/2$ variables.