Optimal Shrinkage Estimation of Fixed Effects in Linear Panel Data Models

Soonwoo Kwon[*]

September 8, 2020

Abstract

I develop an estimator method for fixed effects in linear panel data models. The estimator is optimal in mean squared error under minimal assumptions, within a class of shrinkage estimators. As a result, it (asymptotically) dominates the estimators that have been commonly used in applied research, and significantly so in many cases as demonstrated in the simulation studies. Furthermore, the fixed effects are allowed to change with time and be serially correlated. The estimator is derived by establishing new results for shrinkage estimation in the multivariate normal means model, with possibly serially correlated means. Using administrative data on New York City Public Schools, I measure teacher value-added using the proposed estimator, establishing the efficacy of the estimator.

[*] Yale University, Department of Economics, soonwoo.kwon@yale.edu.

1 Introduction

Fixed effects capture the unobserved individual level effect that can be arbitrarily correlated with the observed covariates. While its original purpose was to control for such effects to consistently estimate the parameter of interest, the fixed effects themselves are of primary interest in numerous contexts as well. For example, the teacher fixed effects in a value-added regression model have been interpreted as teacher quality, finding wide use in both academic research (Rockoff (2004), Rothstein (2010), and Chetty et al. (2014a)) and in school districts [1]. Other examples include firm effects in employee-employer matched data (Abowd and Kramarz (1999)), neighborhood effects in the study of intergenerational mobility (Chetty and Hendren (2018b)), and judge fixed-effects (Frandsen et al. (2019)).

One of the interesting features of fixed effects is that the dimension usually increases with the sample size, while the effective sample size one can use to estimate each component does not necessarily increase. This makes the natural estimator obtained by aggregating the residuals at the individual level noisy. For example, consider the case of estimating the teacher fixed effects of a large school district with a single year, or multiple but small years, of data. Here, the relevant asymptotic experiment is where the number of teachers grows. However, the number of students taught by each teacher is rather small and fixed in magnitude, making estimation of the teacher fixed-effects challenging.

Despite the noisiness of the aggregate residual estimator, many studies have been using the shrinkage estimators that are variants of the seminal estimator by James and Stein (1961) have been used in the literature. The James-Stein estimator is an estimator that shrinks the usual maximum likelihood estimator, y_i, for θ_i in the normal means model $y_i \overset{indep}{\sim} N(\theta_i, \sigma^2)$. The striking result is that such shrinkage dominates the

The econometrician observes (Y_{ijt}, X'_{ijt}) for $t = 1, \dots, T$, $j = 1, \dots J$, $i = 1, \dots, n_{jt}$, and considers the model

$$Y_{ijt} = X'_{ijt}\beta + \alpha_{jt} + \varepsilon_{ijt}. \tag{1}$$

The interpretation is that t is time, and we observe unit j for each time t. The index i does not necessarily indicate a unique identity. That is, for two different pairs (j,t) and (j',t'), the same subscript i may refer to different

[1] **Add reference**

units.[2] For example, in the teacher value-added model, t corresponds to school year, j to teacher and i to a student assigned to teacher j in school year t. In the neighborhood effects analysis such as in Chetty and Hendren (2018b), t corresponds to year, j to neighborhood and i to a resident of j in year t. The time-varying fixed effect α_{jt} is independent across j but is allowed to be serially correlated for a given j. The econometrician is interested in estimating the realized $\{\alpha_{jt}\}$. [**Talk about the constant term in more detail.**]

We assume $\varepsilon_{jt} = (\varepsilon_{1jt}, \ldots, \varepsilon_{n_{jt},jt})'$ is independent across (j,t) and strict exogeneity holds, i.e.,

$$E[\varepsilon_{jt}|X_{1jt}, \ldots, X_{n_{jt}jt}] = 0. \tag{2}$$

The aim of the paper is to estimate or predict, depending on whether it is treated fixed or random, α_{jt} which captures the (j,t) level heterogeneity.

2 Normal means model

For any random variable W_{ijt}, define $\overline{W}_{jt} = n_{jt}^{-1} \sum_{i=1}^{n_{jt}} W_{ijt}$ and $\widetilde{W}_{ijt} = W_{ijt} - \overline{W}_{jt}$. Consider the demeaned version of (1),

$$\widetilde{Y}_{ijt} = \widetilde{X}_{ijt}'\beta + \widetilde{\varepsilon}_{ijt}, \tag{3}$$

and notice that as long as $J \to \infty$ or $T \to \infty$, the OLS estimator $\widehat{\beta}$ is a consistent estimator of β. Throughout the paper, the asymptotic experiment we consider is as $J \to \infty$.

We have

$$\widehat{\alpha}_{jt} := \overline{Y}_{jt} - \overline{X}_{jt}'\widehat{\beta} = \overline{X}_{jt}'(\beta - \widehat{\beta}) + \alpha_{jt} + \overline{\varepsilon}_{jt} = \alpha_{jt} + \overline{\varepsilon}_{jt} + O_p(J^{-1/2}), \tag{4}$$

so that

$$\widehat{\alpha}_{jt} \to_d \alpha_{jt} + \overline{\varepsilon}_{jt}, \tag{5}$$

for all $j \leq J$ and $t \leq T$. In fact, under a mild boundedness condition on X_{ijt} that ensures $\sup_{j,t}\|\overline{X}_{jt}\| = O_p(1)$, such convergence is uniform over (j,t)

[2] An alternative assumption that will work as well is to assume that the data is a repeated cross-section across (j,t), but this is not necessarily the case for teacher value-added models.

because, by Cauchy-Schwarz,

$$P\left(\sup_{j,t}|\widehat{\alpha}_{jt}-\alpha_{jt}-\overline{\varepsilon}_{jt}|>\varepsilon\right)\leq P\left(\|\widehat{\beta}-\beta\|\sup_{j,t}\|\overline{X}_{jt}\|>\varepsilon\right), \tag{6}$$

and $\widehat{\beta}$ is a consistent estimator of β. Assuming independence between α_{jt} and $\overline{\varepsilon}_{jt}$ and $\overline{\varepsilon}_{jt}\sim N(0,\sigma^2_{jt})$, we have

$$\widehat{\alpha}_{jt}|\alpha_{jt}\sim N(\alpha_{jt},\sigma^2_{jt}), \tag{7}$$

approximately.

This shows the natural connection between the estimation of fixed effects in a linear panel model and the estimation of the means in the (heteroskedastic) normal means model. Note that even if we assume $Var(\varepsilon_{ijt})=\sigma^2$, we have $Var(\overline{\varepsilon}_{jt})=\sigma^2/n_{jt}$ so that the heteroskedasticity is still there due to the different cell sizes, n_{jt}. While I allow for general forms of heteroskedasticity, the variance terms are assumed to be known. This assumption is not crucial, as it can be replaced by a consistent estimator.

Using the more conventional notation in the literature, I consider the following normal means model

$$y_{jt}|\theta_{jt}\overset{indep}{\sim}N(\theta_{jt},\sigma^2_{jt}), \tag{8}$$

for $j=1,\dots,J$ and $t=1,\dots,T$. For any sequence $\{w_{jt}\}_{j=1,\dots,J,t=1,\dots,T}$, define $w_j:=(w_{j1},\dots,w_{jT})'$, $\overline{w}_j:=\frac{1}{T}\sum_{t=1}^{T}w_{jt}$ and $w:=(w_1',\dots,w_J')'$. Most of the results holds under a more general setting that allows for correlation within y_j even after conditioning on θ_j,

$$y_j|\theta_j\overset{indep}{\sim}N(\theta_j,M_j),$$

where M_j is positive definite matrix, but not necessarily diagonal. The parameter of interest is θ or some subvector of it.

I use $\|\cdot\|$ to denote the Euclidean norm when the argument is a vector and Frobenius norm if a matrix. For any matrix A, $(A)_{ij}$ denotes its (i,j) entry and $\sigma_k(A)$ its kth largest singular value. For two real symmetric matrices A and B, I write $A\geq B$ to denote that $A-B$ is positive semidefinite.

2.1 Risk criteria

The risk function I use to evaluate an estimator $\widehat{\theta}$ of θ is the compound (weighted) MSE,

$$R_W(\widehat{\theta}, \theta) = \tfrac{1}{J} E_\theta (\widehat{\theta} - \theta)' W (\widehat{\theta} - \theta), \tag{9}$$

where W is a positive-definite weight matrix. The expectation, E_θ, is understood to be evaluated at θ when θ is treated as fixed and conditional on θ when it is treated as random. For any $d \in \mathbb{R}^k$, let diag(d) denote the $k \times k$ diagonal matrix with diagonal elements d. Likewise, for a sequence of $T \times T$ matrices $D_1, \dots, D_J$, let diag$(D_1, \dots, D_J)$ denote the $JT \times JT$ block diagonal matrix that takes D_j as its jth diagonal block. Let $e_t \in \mathbb{R}^T$ be the tth elementary vector, and $e^t := (e_t', \dots, e_t')' \in \mathbb{R}^{TJ}$. If we take $W = diag(e^t)$, we have $R_W(\widehat{\theta}, \theta) = \frac{1}{J} \sum_{j=1}^{J} E_\theta (\widehat{\theta}_{jt} - \theta_{jt})^2$ so that we are interested only in the parameters that correspond to time t. Taking $W = \left(I_J \otimes \frac{1}{T} 1_T'\right)' \left(I_J \otimes \frac{1}{T} 1_T'\right)$, where 1_T is the vector of length T with all of its elements equal to 1, we have

$$R_W(\widehat{\theta}, \theta) = \tfrac{1}{J} \textstyle\sum_{j=1}^{J} E_\theta \left(\tfrac{1}{T} \sum_{t=1}^{T} \widehat{\theta}_{jt} - \overline{\theta}_j\right)^2, \tag{10}$$

so that now the interest is estimating the average of the fixed-effects over time.

We assume that $W = \mathrm{diag}(\widetilde{W}, \dots, \widetilde{W})$ where $\widetilde{W}$ is a $T \times T$ positive semidefinite matrix. Writing $\ell_j(\widehat{\theta}_j, \theta_j) := (\widehat{\theta}_j - \theta_j)' W_j (\widehat{\theta}_j - \theta_j)$, we have

$$R_W(\widehat{\theta}, \theta) = \tfrac{1}{J} \textstyle\sum_{j=1}^{J} E_\theta \ell_j(\widehat{\theta}_j, \theta_j), \tag{11}$$

so that the risk $R_W(\widehat{\theta}, \theta)$ can be written as the average of the individual level risks $E_\theta \ell_j(\widehat{\theta}_j, \theta_j)$ for $j \leq J$. All risk evaluations, and thus optimality results, are conditional on a realized sequence θ. From now on, I omit the subscript θ in E_θ for simplicity.

3 Previous work

When $\sigma_{jt}^2 = \sigma^2$ for all j and t and $W = I_{JT}$, the problem is exactly the one considered by the seminal work of Stein (1956) and James and Stein (1961). They introduced the shrinkage estimator and showed the usual MLE estimator is inadmissible when $JT > 3$. It was noted in later work by Efron and Morris (1973) that the same estimator can be derived from a hierarchical Bayes model where the hyperparameter is estimated using the marginal distribution of the data. One of the many striking features of this estimator is that it gives a smaller MSE for all possible true values of the parameter, not just the employed prior.

However, if one allows for heteroskedasticity, which is more likely to be the case in applications, then such strong optimality properties no longer hold.[3] In relatively recent work, Xie et al. (2012) derive optimal shrinkage estimators in a heteroskedastic normal means model. Xie et al. (2016) show that this is not specific to the case where the likelihood of the data is Gaussian, but applies to a larger class of distributions, including scale-location families and exponential families. Brown et al. (2018) show that a similar optimal estimator can be derived under a two-way normal means model.

Nonparametric approaches have been proposed by Brown and Greenshtein (2009), Jiang and Zhang (2009), Koenker and Mizera (2014a) and Fu et al. (2020). The work by Fu et al. (2020) is the only one that explicitly considers the heteroskedastic setting, though the other papers can also be applied to such settings with minor extensions. However, the optimality in such papers are usually in terms with respect to the unknown prior. Moreover, extending such methods to our context inevitably involves estimating nonparametric functions with T arguments, which is computationally undesirable even for moderately large T.

Kong et al. (2017) extend Xie et al. (2012) so that the correlation terms of the second stage model is not necessarily 0 but can be positive. All correlation terms are assumed to be equal.

For applications in economic research, Chetty et al. (2014a,b) and Chetty and Hendren (2018a,b) use empirical Bayes methods to estimate the non-time-varying fixed effects. Guarino et al. (2015) give a nice review of the methods used in the teacher value-added literature. Gilraine et al. (2019) provide a method of estimating the teacher value-added using the nonparametric approach by Koenker and Mizera (2014b). Bitler et al. (2019) show that the methods that have been used in the literature can be a bit misleading.

Liu et al. (2019) use empirical Bayes ideas in a panel data forecasting setting. Here, the fixed effects are estimated but the optimality is in terms of the forecast.

There is also a long literature on estimating the heterogeneity in panel data

[3] To be precise, the shrinkage estimator is still optimal under the normal prior, but no optimality claim can be made in a frequentist sense.

settings. The interest is in identification or estimating the distribution of the fixed-effects, rather than the entire vector of fixed-effects.

4 TVFE shrinkage esimator

Consider a second stage model

$$\theta_j \sim N(\mu, \Lambda),$$

where the location vector $\mu \in \mathbb{R}^T$ and the positive semidefinite $T \times T$ matrix Λ are hyperparameters to be tuned later. The restriction one imposes on Λ incorporates the prior knowledge on the underlying covariance structure. I denote by $\mathcal{L}$ a subset of $\mathcal{S}_T^+$, the set of positive definite $T \times T$ matrices, that the reflects the prior knowledge on Λ. For example, when θ_j is believed to be covariance stationary, $\mathcal{L}$ is 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. Taking $\Lambda = \lambda 1_T 1_T'$, brings us back to the $\theta_{jt} = \theta_j$ case. Unless otherwise specified, $\sup_\Lambda$ is understood to be taken over $\Lambda \in \mathcal{L}$. Write $M_j = \text{diag}(n_{j1}^{-1}, \dots, n_{jT}^{-1})$ and $M = \text{diag}(M_1, \dots, M_J)$. Here, n_{jt} does not have to denote the corresponding cell size, but just any positive number that reflects the heterogeneity. However, since the heterogeneity comes from the cell sizes in our setting, we keep the n_{jt} notation.

We can rewrite the model as

$$\begin{aligned} y|\theta &\sim N(\theta, M) \\ \theta &\sim N(\mu, (I_J \otimes \Lambda)). \end{aligned}$$

The marginal covariance of y is given as Σ, where

$$\Sigma = (I_J \otimes \Lambda) + M.$$

This implies

$$Cov(y_j, y_\ell) = \begin{cases} 0 & \text{if } j \neq \ell \\ (\Lambda + M_j) & \text{if } j = \ell, \end{cases} \tag{12}$$

which shows that all the correlation among y comes from Λ. However, note that the exact form depends on M_j.

The block diagonal structure of Σ gives

$$\Sigma^{-1} = (\text{diag}(\Lambda + M_1, \dots, \Lambda + M_J))^{-1} = \text{diag}((\Lambda + M_1)^{-1}, \dots, (\Lambda + M_J)^{-1})$$

so that $(I_J \otimes \Lambda)\Sigma^{-1} = \text{diag}(\Lambda(\Lambda + M_1)^{-1}, \dots, (\Lambda + M_J)^{-1})$. Standard calculation gives (see Lemma 2.1 of Brown et al. (2018), for example)

$$\begin{aligned} E[\theta_j|y] =& \mu + \Lambda(\Lambda + M_j)^{-1}(y_j - \mu) \\ =& \left(I_T - \Lambda(\Lambda + M_j)^{-1}\right)\mu + \Lambda(\Lambda + M_j)^{-1} y_j \end{aligned}$$

Analogous to the univariate case, I refer to $\Lambda(\Lambda + M_j)^{-1}$ as the shrinkage matrix. However, here the "shrinkage" involves rotation of the data as well. If $\Lambda = \text{diag}(\lambda, \dots, \lambda)$ and $M_j = \text{diag}(m_{j1}, \dots, m_{jT})$, then this becomes

$$E[\theta_{jt}|y] = \left(1 - \frac{\lambda}{\lambda + m_{jt}}\right)\mu_t + \frac{\lambda}{\lambda + m_{jt}} y_{jt},$$

which shows the direct connection to the univariate case.

To gain more insight of the shrinkage estimator, let UDU' be the spectral decomposition of $M_j^{-1/2}\Lambda M_j^{-1/2}$ with $D = \text{diag}(d_1, \dots, d_T)$. This gives

$$\begin{aligned} M_j(\Lambda + M_j)^{-1} =& M_j^{1/2}(I_T + M_j^{-1/2}\Lambda M_j^{-1/2})^{-1} M_j^{-1/2} \\ =& M_j^{1/2}(I_T + UDU')^{-1} M_j^{-1/2} \\ =& M_j^{1/2} U (I_T + D)^{-1} U' M_j^{-1/2}, \end{aligned} \tag{13}$$

and thus

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

Here, the last $M_j^{-1/2}$ term simply standardizes the data and the first $M_j^{1/2}$ term simply brings the data back to its original scale (and direction). The $UD(I_T + D)^{-1}U'$ term captures the shrinkage. More specifically, U' rotates the standardized data $M_j^{-1/2} y_j$, $D(I_T + D)^{-1}$ shrinks this rotated data appropriately, and finally U rotates the data back to its original axes. Note that $D(I_T + D)^{-1}$ is indeed a shrinkage because $D(I_T + D)^{-1} = \text{diag}(d_1/(1 + d_1), \dots, d_T/(1 + d_T))$ and $1/(1 + d_t) \in (0, 1]$ for all $t \leq T$. Since both U and $D(I_T + D)^{-1}$ depend on Λ, tuning Λ involves determining the direction and magnitude of shrinkage.

Empirical Bayes methods proceed by substituting a data driven estimator

$(\widehat{\mu}, \widehat{\Lambda})$, typically the maximum likelihood or method of moments estimator, for (μ, Λ). However, unlike under the independent, homoskedastic setting the resulting estimator is not known to have any optimality result in a frequentist sense of this estimator under our setting. [**It seems clear that it should be EB optimal in the sense of Morris (1983), but let's make sure.**]

I consider fixed T to reflect the relatively short time dimension of the panel data in most applied work. In the Gaussian case, allowing $T \to \infty$ does not make the theory any more difficult, and most of the results go through. In practice, however, the dimensionality of the hyperparameter grows as T gets larger, making the tuning step computationally much more demanding. Considering $T \to \infty$ can raise some interesting questions because in this case the number of hyperparameters also increases as the "sample size" increases, which is not the case in Xie et al. (2012, 2016) and Brown et al. (2018).

Define $\widehat{\theta}_j(\mu, \Lambda) = (I_T - \Lambda(\Lambda + M_j)^{-1})\mu + \Lambda(\Lambda + M_j)^{-1} y_j$ and $\widehat{\theta} := (\widehat{\theta}'_1, \ldots, \widehat{\theta}'_J)'$, where I omit the dependence on (μ, Λ) for simplicity. Rather than taking the empirical Bayes perspective, here we take an alternative route of estimating the risk, and then choosing the hyperparameters to minimize this risk estimate.

Inspired by SURE, consider the following unbiased risk estimate under the true θ, $\mathbf{URE}(\mu, \Lambda) = \frac{1}{J}\sum_{j=1}^{J} \mathbf{URE}_j(\mu, \Lambda)$, where

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

First, consider the case where $\mu = 0$, which is the case if the identification assumption for the estimation of the fixed effects are $\sum_{j=1}^{J} \mu_{jt} = 0$ for $t \leq T$ [**Make this more formal**]. In this case, I use $\mathbf{URE}_j(\Lambda)$ as a shorthand for $\mathbf{URE}_j(0, \Lambda)$, and $\mathbf{URE}(\Lambda)$ and $\widehat{\theta}(\Lambda)$ likewise. Some algebra (see Appendix A)

$$\begin{aligned}&\mathbf{URE}_j(\Lambda) - (\widehat{\theta}_j(\Lambda) - \theta_j)'(\widehat{\theta}_j(\Lambda) - \theta_j)\\ =&y_j' y_j - \theta_j'\theta_j - \mathrm{tr}(M_j) - 2\mathrm{tr}(\Lambda(\Lambda + M_j)^{-1}(y_j y_j' - \theta_j y_j' - M_j))\end{aligned} \tag{14}$$

Assumption 4.1 (Boundedness)**.**

(i) $\sup_j \|\theta_j\| < \infty$,

(ii) $0 < \inf_j \sigma_T(M_j) \leq \sup_j \sigma_1(M_j) < \infty$, and

(iii) $\sup_j E\|y_j\|^4 < \infty$.

Assumption 4.1 imposes rather mild boundedness conditions on the mean, variance, and fourth moment of the data. The supremums and infimums are taken over $j \in \{1, 2, \ldots\}$ though one only observes up to $j = 1, \ldots, J$. Note that (iii) imposes restriction on the true mean sequence $\{\theta_j\}_{j=1}^{\infty}$ as well because the expectation is evaluated under this true mean sequence. An implication of (ii) is that the condition number of M_j, $\kappa(M_j) := \sigma_1(M_j)/\sigma_T(M_j)$, is bounded above and away from zero.

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

$$\sup_{\Lambda \in \mathcal{S}_T^+} \left| \mathbf{URE}(\Lambda) - \ell(\theta, \widehat{\theta}(\Lambda)) \right| \xrightarrow{L_1} 0. \tag{15}$$

Remark 4.1. The theorem shows uniform convergence over the largest possible (hyper-)parameter space, $\mathcal{S}_T^+$, and thus over any $\mathcal{L} \subset \mathcal{S}_T^+$.

Remark 4.2. As is clear from the proof, the assumption I impose is not the weakest possible. However, even this stronger set of assumptions is not very restrictive and is easy to interpret.

We now define the oracle shrinkage estimator, which will serve as a benchmark. The oracle loss hyperparameter is

$$\widetilde{\Lambda}^{\mathrm{OL}} = \arg\min_{\Lambda} \ell(\theta, \widehat{\theta}(\Lambda))$$

and the corresponding estimator is $\widetilde{\theta}^{\mathrm{OL}} = \widehat{\theta}(\widetilde{\Lambda}^{\mathrm{OL}})$. The estimator depends on the unknown θ, and thus not a feasible estimator. By definition, no shrinkage estimator can have smaller risk than $\widetilde{\theta}^{\mathrm{OL}}$. Likewise, define the estimator that minimizes $\mathbf{URE}(\Lambda)$ as $\widehat{\theta}^{\mathrm{URE}}$ and the corresponding hyperparameter $\widehat{\Lambda}^{\mathrm{URE}}$.

Theorem 4.2. *Under Assumption 4.1, we have*

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

Proof. By definition, we have $\mathbf{URE}(\widehat{\Lambda}^{\mathrm{URE}}) \leq \mathbf{URE}(\widetilde{\Lambda}^{\mathrm{OL}})$, which gives

$$\begin{aligned} & P\left(\ell(\theta, \widehat{\theta}^{\mathrm{URE}}) \geq \ell(\theta, \widetilde{\theta}^{\mathrm{OL}}) + \varepsilon\right) \\ \leq & P\left(\ell(\theta, \widehat{\theta}^{\mathrm{URE}}) - \mathbf{URE}(\widehat{\Lambda}^{\mathrm{URE}}) \geq \ell(\theta, \widetilde{\theta}^{\mathrm{OL}}) - \mathbf{URE}(\widetilde{\Lambda}^{\mathrm{OL}}) + \varepsilon\right) \\ \leq & P\left(2 \sup_{\Lambda} |\ell(\theta, \widehat{\theta}(\Lambda)) - \mathbf{URE}(\Lambda)| \geq \varepsilon\right), \end{aligned}$$

and the term in the last line converges to zero due to Theorem 4.1 and the fact that L_1 convergence implies convergence in probability. □

Corollary 4.1. *Under Assumption 4.1, we have*

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

Proof. We have

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

Taking expectations and applying Theorem 4.1 gives the desired result. □

While there is no Λ that makes $\widehat{\theta}(\Lambda)=y$, which means that the "usual" maximum likelihood estimator is not included in the class of estimators we consider, the following argument shows that the proposed estimator weakly dominates the MLE as $J\to\infty$. Fix a (true) parameter sequence $\{\theta_j\}_{j=1}^{\infty}$ Since. $R(\theta,y)=E[(\theta-y)'(\theta-y)]=\mathrm{tr}(M)$, it suffices to show we can find an "estimator" $\widetilde{\theta}^{\mathrm{MLE}}$ such that

$$\lim_{J\to\infty}\left|R(\theta,\widetilde{\theta}^{\mathrm{MLE}})-\mathrm{tr}(M)\right|\leq\varepsilon.$$

I focus on estimators where $\Lambda=D(\lambda):=\mathrm{diag}(\lambda,\ldots,\lambda)$. For any fixed J, we have that

$$\lim_{\lambda\to\infty}R(\theta,\widehat{\theta}(D(\lambda)))=\mathrm{tr}(M),$$

and thus there exists λ_J such that

$$\left|R(\theta,\widehat{\theta}(D(\lambda_J)))-\mathrm{tr}(M)\right|\leq\frac{1}{J}.$$

Note that λ_J depends on the true parameter, and thus is not a feasible estimator. However, we know that $\widehat{\theta}(D(\lambda_J))$ has a larger risk than the oracle estimator as $J \to \infty$. Also, we have

$$\lim_{J\to\infty} \left| R(\theta, y) - R(\theta, \widehat{\theta}(D(\lambda_J))) \right| = 0.$$

4.1 Solving the minimization problem

Here, we try to characterize the minimization problem a little bit more. Note that the key is to derive the derivative of

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

with respect to Λ. Due to restriction that Λ is positive definite, we work with the Cholesky decomposition, $\Lambda = L'L$, where L is lower triangular with strictly positive diagonal elements. I first consider the case where μ is known.

4.1.1 First-order conditions ignoring the positive definite constraint

I derive the first-order conditions. (See 2.4.4 of the matrix cookbook). We have

$$\begin{aligned} &\nabla_\Lambda \mathrm{tr}((\Lambda + M_j)^{-1}M_j^2) \\ =&M_j^2 \nabla_\Lambda (\Lambda + M_j)^{-1} \\ =&-(\Lambda + M_j)^{-1}M_j^2(\Lambda + M_j)^{-1}. \end{aligned}$$

We also have

$$\begin{aligned} &\nabla_\Lambda (y_j - \mu)'[(\Lambda + M_j)^{-1}M_j^2(\Lambda + M_j)^{-1}](y_j - \mu) \\ =&\nabla_\Lambda \mathrm{tr}((y_j - \mu)'[(\Lambda + M_j)^{-1}M_j^2(\Lambda + M_j)^{-1}](y_j - \mu)) \\ =&\nabla_\Lambda \mathrm{tr}([(\Lambda + M_j)^{-1}M_j^2(\Lambda + M_j)^{-1}](y_j - \mu)(y_j - \mu)') \\ =&-2(\Lambda + M_j)^{-1}(y_j - \mu)(y_j - \mu)'(\Lambda + M_j)^{-1}M_j^2(\Lambda + M_j)^{-1}. \end{aligned}$$

Hence, the first order condition is given as

$$\sum_{j=1}^{J} \left[(\Lambda + M_j)^{-1}M_j^2(\Lambda + M_j)^{-1} - (\Lambda + M_j)^{-1}(y_j - \mu)(y_j - \mu)'(\Lambda + M_j)^{-1}M_j^2(\Lambda + M_j)^{-1}\right] = 0.$$

Equivalently, we can write this as

$$\sum_{j=1}^{J} \left[(I_T - (\Lambda + M_j)^{-1}(y_j - \mu)(y_j - \mu)')(\Lambda + M_j)^{-1} M_j^2 (\Lambda + M_j)^{-1} \right] = 0.$$

4.1.2 Comparison with the usual EBLME approach

For $y \sim N(\mu, (I_j \otimes \Lambda + M))$, the log-likelihood of the data is proportional to

$$\sum_{j=1}^{J} \left[-\frac{1}{2} \log|(\Lambda + M_j)| - \frac{1}{2}(y_j - \mu)'((\Lambda + M_j))^{-1}(y_j - \mu) \right]$$

so that Λ^{EBMLE} solves

$$\sum_{j=1}^{J} \left[(\Lambda + M_j)^{-1} - (\Lambda + M_j)^{-1}(y_j - \mu)(y_j - \mu)'(\Lambda + M_j)^{-1} \right] = 0.$$

4.2 Data-driven centering

We can also consider another tuning parameter μ that corresponds to the centering term, in which case the second level model becomes

$$\theta_j \sim N(\mu, \Lambda).$$

As in Brown et al. (2018), I restrict the (hyper)parameter space of μ to

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

where $q_{1-\tau}(\{w_j\}_{j=1}^{J})$ denotes the $1-\tau$ sample quantile (see, for example, Chapter 21 of van der Vaart (1998) for a formal definition) of $\{w_j\}_{j=1}^{J} \subset \mathbf{R}$. I write $\sup_\mu$ rather than $\sup_{\mu \in \mathcal{M}_t}$, but all supremums over μ are understood to be taken over $\mathcal{M}_J$. Additional to Theorem 4.1, we also need

$$\sup_{\mu,\Lambda} \left| \tfrac{1}{J} \textstyle\sum_{j=1}^{J} \mu'(\Lambda + M_j)^{-1} M_j (y_j - \theta_j) \right| \xrightarrow{L_1} 0,$$

for the method of choosing the hyperparameters by minimizing the $\mathbf{URE}(\mu, \Lambda)$ give risk properties comparable to that of the oracle.

Write $\varepsilon_{jt} := y_{jt} - \theta_{jt}$ so that $E\varepsilon_{jt} = 0$ and $E\varepsilon_{jt}^2 = M_{jt}$, where M_{jt} denotes the tth diagonal entry of M_j. Note that $\overline{M}_t := \sup_j M_{jt} < \infty$ by Assumption 4.1. I assume that the distribution of ε_{jt} belongs to a scale family with finite fourth moments. Define the standardized noise term as $\eta_{jt} := \varepsilon_{jt}/M_{jt}^{1/2}$, and let F_t for $t = 1, \ldots, T$ be distribution functions with finite fourth moments.

Assumption 4.2 (Scale family)**.** *For each $t = 1, \ldots, T$, we have $\eta_{jt} \overset{i.i.d.}{\sim} F_t$ for $j = 1, \ldots, J$.*

The following theorem shows that $\mathbf{URE}(\mu, \Lambda)$ is uniformly close to the true loss.

Theorem 4.3 (Uniform convergence of $\mathbf{URE}(\mu, \Lambda)$)**.** *Suppose Assumption 4.1 and 4.2 hold. Then,*

$$\sup_{\mu \in \mathcal{M}_J, \Lambda \in \mathcal{S}_T^+} \left| \mathbf{URE}(\mu, \Lambda) - \ell(\theta, \widehat{\theta}(\mu, \Lambda)) \right| \xrightarrow{L_1} 0. \tag{16}$$

4.3 $\widetilde{W} \neq I$

In this case, we have

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

from which it will follow Similar results as before hold, under essentially the same conditions. **[Find applications where a specific time period is of interest?]**

4.4 Semiparametric shrinkage

For the univariate model where $T = 1$, we have

$$\widehat{\theta}(\Lambda) = \frac{\Lambda}{\Lambda + M_j} y_j,$$

so the estimator shrinks those observations observed with more noise more to zero. Also, the shrinkage factor lies in $[0, 1]$. Motivated by such observations, Xie et al. (2012) consider a class of semiparametric shrinkage estimators,

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

where $b(\cdot)$ is a weakly decreasing function taking values in $[0,1]$. Likewise, I can make the shrinkage part more flexible. Recall that

$$\Lambda(\Lambda+M_j)^{-1}=I_T-M_j(\Lambda+M_j)^{-1}=M_j^{1/2}U_{\Lambda,j}D_{\Lambda,j}(I_T+D_{\Lambda,j})^{-1}U_{\Lambda,j}'M_j^{-1/2}.$$

A semiparametric version of this is given as

$$M_j^{1/2}U_{\Lambda,j}\mathrm{diag}(b(d_{\Lambda,j,1}),\ldots,b(d_{\Lambda,jT}))U_{\Lambda,j}'M_j^{-1/2},$$

where $b(\cdot)$ is an increasing function that is Lipschitz continuous with Lipchitz constant $C>1$ taking values in $[0,1]$. Let $S_T^{+,M}$ denote the set of those matrices $S\in S_T^+$ with $\sigma_1(S)\leq M$, and define $\mathcal{L}(S_T^{+,M},S_T^+)$ as the set of Lipschitz functions from $B: S_T^{+,M}\to S_T^+$ with Lipschitz constant $C>1$, and uniformly bounded by 1. Note that $\mathcal{L}(S_T^{+,M},S_T^+)$ is totally bounded. Also, the usual shrinkage function

$$B(M_j^{-1/2}\Lambda M_j^{-1/2})=(I+M_j^{-1/2}\Lambda M_j^{-1/2})^{-1}$$

is a special case. Moreover, we have

$$\begin{aligned}&\|B(M_j^{-1/2}\Lambda M_j^{-1/2})-\widetilde{B}(M_j^{-1/2}\widetilde{\Lambda}M_j^{-1/2})\|\\ \leq&\|B(M_j^{-1/2}\Lambda M_j^{-1/2})-B(M_j^{-1/2}\widetilde{\Lambda}M_j^{-1/2})\|+\|B(M_j^{-1/2}\widetilde{\Lambda}M_j^{-1/2})-\widetilde{B}(M_j^{-1/2}\widetilde{\Lambda}M_j^{-1/2})\|\\ \leq&C\|M_j^{-1/2}\|^2\|\Lambda-\widetilde{\Lambda}\|+T\|B-\widetilde{B}\|,\end{aligned}$$

which can be used to establish uniform convergence. However, this involves fitting a function with $T(T+1)/2$ variables and Λ.

4.5 (j,t) level covariates

Suppose there are (j,t) level covariates that are thought to be explain the fixed-effects. Let $Z_{jt}=(Z_{jt1},\ldots,Z_{jtK})'\in\mathbf{R}^K$ be the vector of such covariates, and write $Z_j=(Z_{j1},\ldots,Z_{jT})'$. I assume that $\{(y_j,Z_j)\}_{j=1}^J$ is an independent sample with the Z_j's being identically distributed as well. One way to incorporate such information is to postulate a second level model where the expectation of the estimated fixed effects for j lies in the column space of Z_j,

$$\theta_j|Z_j\sim N(Z_j\gamma,\Lambda),$$

[**Make the dependence structure more clear**] where we now choose parameters γ and Λ to minimize the **URE**. Under this second level model, the posterior mean of θ_j is given as

$$\theta_j^{\mathrm{cov}}(\gamma, \Lambda) = \left(I_T - \Lambda(\Lambda + M_j)^{-1}\right) Z_j \gamma + \Lambda(\Lambda + M_j)^{-1} y_j,$$

and thus I consider the class of estimators of such form. I impose the following restrictions on the covariates.

Assumption 4.3 (Covariates)**.**
(i) $\sup_j \sigma_1(Z_j' Z_j) < \infty$ *a.s.*,
(ii) $E[\varepsilon_j | Z_j] = 0$ *and* $\operatorname{var}(\varepsilon_j | Z_j) = M_j$ *for* $t = 1, \ldots, T$, *and*
(iii) $EZ_j' Z_j$ *is nonsingular.*

A sufficient condition for (i) is that there exists some constant $\overline{Z} \in \mathbf{R}$ such that $\sup_{j,t,k} |Z_{jtk}| < \overline{Z} < \infty$ almost surely, which amounts to assuming that the covariates are uniformly bounded. The first and second part of (ii) are standard exogeneity conditions for the first and second moments of the noise term. The full rank condition given in the first part of (iii) is standard. From now on, with some abuse of notation, I implicitly condition on $(Z_j)_{j=1}^{\infty}$ and treat the covariates as fixed. I assume that this (fixed) sequence satisfies $\sup_j \sigma_1(Z_j) < \infty$ and $\frac{1}{J} \sum_{j=1}^{J} Z_j' Z_j \to EZ_j' Z_j$, which holds for almost all realized sequences due to (i), (iii), and the strong law of large numbers.

Define the unbiased risk estimate for the jth component as $\mathbf{URE}_j^{\mathrm{cov}}(\gamma, \Lambda) = \mathbf{URE}_j(Z_j \gamma, \Lambda)$, and the compound risk as $\mathbf{URE}^{\mathrm{cov}}(\gamma, \Lambda) = \frac{1}{J} \sum_{j=1}^{J} \mathbf{URE}_j^{\mathrm{cov}}(\gamma, \Lambda)$. I restrict the (hyper)parameter space of γ to

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

where B is a large constant that does not depend on J, and $\widehat{\gamma}^{\mathrm{OLS}}$ is the pooled OLS estimator of regressing y_j on X_j, i.e., $\widehat{\gamma}^{\mathrm{OLS}} = (\sum_{j=1}^{J} Z_j' Z_j)^{-1} \sum_{j=1}^{J} Z_j' y_j$. The idea is to keep the parameter space bounded, yet making sure it includes the natural OLS estimator and sufficiently flexible. The following theorem shows that the provided risk estimate is uniformly close to the true loss function over the given parameter space.

Theorem 4.4 (Uniform convergence of $\mathbf{URE}^{\mathrm{cov}}(\gamma, \Lambda)$)**.** *Suppose Assumption*

4.1 and 4.3 hold. Then,

$$\sup_{\gamma\in\Gamma_J,\Lambda\in\mathcal{S}_T^+}\left|\mathbf{URE}^{\text{cov}}(\gamma,\Lambda)-\ell(\theta,\widehat{\theta}^{\text{cov}}(\gamma,\Lambda))\right|\overset{L_1}{\to}0. \tag{17}$$

The expression in (17), can be bounded by[4]

$$\sup_\Lambda\left|\mathbf{URE}(\Lambda)-\ell(\theta,\widehat{\theta}(\Lambda))\right|+\sup_{\gamma\in\Gamma_J}\left|\tfrac{1}{J}\textstyle\sum_{j=1}^J(Z_j\gamma)'(\Lambda+M_j)^{-1}M_j(y_j-\theta_j)\right|.$$

Theorem 4.1 covers the first term in the right-hand side, and thus it suffices to show that the second term converges to 0 in L_1.

4.6 Forecasting θ_{T+1}

Now, I turn to the problem of forecasting θ_{T+1} given data for the first T periods. Again, we restrict the class of estimators to those that take the form of the posterior mean under the second level model. Note that we have

$$\begin{pmatrix} y_j \\ \theta_j \end{pmatrix}\sim N\left(0,\begin{pmatrix} M_j+\Lambda & \Lambda \\ \Lambda & \Lambda \end{pmatrix}\right),$$

with $\{(y_j',\theta_j')'\}_{j=1}^J$ being an independent sequence of random vectors. Write the tuning parameter Λ and the variance matrix M_j as

$$\Lambda=\begin{pmatrix} \Lambda_{-T} & \Lambda_{T,-T} \\ \Lambda_{T,-T}' & \lambda_T \end{pmatrix}, M_j=\begin{pmatrix} M_{j,-T} & M_{j,T,-T} \\ M_{j,T,-T}' & M_{j,T} \end{pmatrix}=\begin{pmatrix} M_{j,1} & M_{j,1,-1}' \\ M_{j,1,-1} & M_{j,-1} \end{pmatrix}$$

where Λ_{-T}, $M_{j,-T}$ and $M_{j,-1}$ are $(T-1)\times(T-1)$ matrices. From the property of positive semidefinite matrices, we know that 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, I impose the restriction that the hyperparameter space $\mathcal{L}\subset S_T^+$ is bounded. A recommended choice of $\mathcal{L}$ is to take $\mathcal{L}:=\{\Lambda\in S_T^+:\sigma_1(\Lambda)\leq K\sigma_1(\frac{1}{J}\sum_{j=1}^J y_jy_j')\}$ for some large number K.

Write $y_{j,-T}=(y_{j1},\ldots,y_{j,T-1})'$ and $y_{-T}=(y_{1,-T}',\ldots,y_{J,-T}')'$ which are y_j and y, respectively, with the observations corresponding to period T removed. Likewise, $y_{j,-1}$ and y_{-1} are y_j and y with the first period observations removed, respectively. The class of estimators I consider is

[4] This follows from (24) of Appendix A.

$$E[\theta_{jT}|y_{-T}] = \Lambda'_{T,-T}(\Lambda_{-T} + M_{j,-T})^{-1} y_{j,-T},$$

which is the posterior mean of θ_{jT} given only the observations from the first $T-1$ periods. As in the estimation problem, we want to choose Λ so that we minimize the prediction error

$$\tfrac{1}{J}\textstyle\sum_{j=1}^{J} E(\Lambda'_{T,-T}(\Lambda_{-T} + M_{j,-T})^{-1} y_{j,-T} - \theta_{jT})^2.$$

Again, this depends on the true parameters, and thus we use an estimate of this risk instead. Define $B(\Lambda, M_{-T}) = (\Lambda_{-T} + M_{-T})^{-1}\Lambda_{T,-T}$, and note that

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

The cross term can be written as

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

It follows that

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

When M_j is diagonal, the lest term on the right-hand side is zero, and thus the mean prediction error we wish to minimize becomes

$$\frac{1}{J}\sum_{j=1}^{J}\left(E[(B(\Lambda, M_{j,-T})' y_{j,-T} - y_{jT})^2] - M_{j,T}\right).$$

An unbiased estimator for this prediction error is given as

$$\frac{1}{J}\sum_{j=1}^{J}\left((B(\Lambda, M_{j,-T})' y_{j,-T} - y_{jT})^2 - M_{j,T}\right). \tag{18}$$

The difference between the true forecasting error and the unbiased estimate of it for the jth component is given as

$$\left((B(\Lambda, M_{j,-T})'y_{j,-T} - y_{jT})^2 - M_{j,T}\right) \Lambda'_{T,-T}(\Lambda_{-T} + M_{j,-T})^{-1}y_{j,-T} - \theta_{jT})^2.$$

We want to show that

$$\left|\tfrac{1}{J}\textstyle\sum_{j=1}^{J}\left((B(\Lambda, M_{j,-T})'y_{j,-T} - y_{jT})^2 - M_{j,T}\right) - \tfrac{1}{J}\textstyle\sum_{j=1}^{J}(B(\Lambda, M_{j,-1})'y_{j,-1} - \theta_{j,T+1})^2\right|$$

converges to zero uniformly over Λ in L_1. First, I prove the result under the following high-level condition on $\{((\theta'_j, \theta_{j,T+1})', M_j)\}_{j=1}^{\infty}$.

Assumption 4.4 (High-level stationarity)**.**

$\left|\tfrac{1}{J}\sum_{j=1}^{J}\left(B(\Lambda, M_{j,-T})'M_{j,-T}B(\Lambda, M_{j,-T}) - B(\Lambda, M_{j,-1})'M_{j,-1}B(\Lambda, M_{j,-1})\right)\right|$, *and*
$\left|\tfrac{1}{J}\sum_{j=1}^{J}\left((B(\Lambda, M_{j,-T})'\theta_{j,-T} - \theta_{j,T})^2 - (B(\Lambda, M_{j,-1})'\theta_{j,-1} - \theta_{j,T+1})^2\right)\right|$

converge to zero uniformly over Λ*.*

Theorem 4.5.

$$\sup_{\Lambda}\left|\tfrac{1}{J}\textstyle\sum_{j=1}^{J}\left((B(\Lambda, M_{j,-T})'y_{j,-T} - y_{jT})^2 - M_{j,T}\right) - \tfrac{1}{J}\textstyle\sum_{j=1}^{J}(B(\Lambda, M_{j,-1})'y_{j,-1} - \theta_{j,T+1})^2\right| \xrightarrow{L_1} 0.$$

While the tuning parameter Λ is required to be positive semidefinite, I relax this and only require that Λ_{-T} is positive semidefinite. This makes the problem more tractable while allowing for a larger class of estimators.[**Add a line explaining that I am not actually relaxing anything. The optimal $\Lambda_{T,-T}$ given below can be chosen by taking λ_T to be small enough.**] We want to choose $\Lambda_{T,-T}$ and Λ_{-T} to minimize

$$\frac{1}{J}\sum_{j=1}^{J}(E[(\Lambda'_{T,-T}(\Lambda_{-T} + M_{j,-T})^{-1}y_{j,-T} - y_{jT})^2] - M_{j,T}). \qquad (19)$$

Define $\widetilde{y}^{\Lambda}_{j,-T} = (\Lambda_{-T} + M_{j,-T})^{-1}y_{j,-T}$, $\widetilde{Y}_{\Lambda,-T} = (\widetilde{y}^{\Lambda}_{1,-T}, \ldots \widetilde{y}^{\Lambda}_{J,-T})'$, and $y_T = (y_{1T}, \ldots, y_{JT})'$. Let $\widehat{\beta}(\widetilde{Y}_{\Lambda,-T}, y_T)$ denote the OLS "estimator" of regressing y_{jT} on $\widetilde{y}^{\Lambda}_{j,-T}$, and $\beta^*(\widetilde{Y}_{\Lambda,-T}, y_T)$ its population version:

$$\widehat{\beta}(\widetilde{Y}_{\Lambda,-T}, y_T) = (\widetilde{Y}'_{\Lambda,-T}\widetilde{Y}_{\Lambda,-T})^{-1}\widetilde{Y}'_{\Lambda,-T}y_T \text{ and}$$
$$\beta^*(\widetilde{Y}_{\Lambda,-T}, y_T) = (E\widetilde{Y}'_{\Lambda,-T}\widetilde{Y}_{\Lambda,-T})^{-1}E\widetilde{Y}'_{\Lambda,-T}y_T.$$

Then, it is clear that the optimal $\Lambda_{T,-T}$'s for (18) and (19), respectively, are given as $\widehat{\beta}(\widetilde{Y}_{\Lambda,-T}, y_T)$ and $\beta^*(\widetilde{Y}_{\Lambda,-T}, y_T)$.

Hence, the problem now boils down to choosing Λ_{-T} that minimizes

$$\frac{1}{J}\sum_{j=1}^{J}(y'_{j,-T}(\Lambda_{-T} + M_{j,-T})^{-1}\widehat{\beta}(\widetilde{Y}_{\Lambda,-T}, y_T) - y_{jT})^2. \tag{20}$$

The question is, under what condition is minimizing (20) equivalent to minimizing

$$\frac{1}{J}\sum_{j=1}^{J}(y'_{j,-1}(\Lambda_{-T} + M_{j,-1})^{-1}\widehat{\beta}(\widetilde{Y}_{\Lambda,-1}, y_{T+1}) - y_{j,T+1})^2. \tag{21}$$

Let $P(\widetilde{Y}_{\Lambda,-T})$ and $P(\widetilde{Y}_{\Lambda,-1})$ be the projection matrix corresponding to the column spaces of $\widetilde{Y}_{\Lambda,-T}$ and $\widetilde{Y}_{\Lambda,-1}$, respectively. The difference of (20) and (21) can be written as

$$\begin{aligned}
&\tfrac{1}{J}\textstyle\sum_{j=1}^{J}\left((y_T^{\Lambda\prime}\widehat{\beta}_T - y_{jT})^2 - (y_{j,-1}^{\Lambda}{}'\widehat{\beta}_{T+1} - y_{j,T+1})^2\right)\\
=&\tfrac{1}{J}\left(y_T{}'(I - P(\widetilde{Y}_{\Lambda,-T}))y_T - y_{T+1}{}'(I - P(\widetilde{Y}_{\Lambda,-1}))y_{T+1}\right)\\
=&\tfrac{1}{J}\textstyle\sum_{j=1}^{J}(y_{jT}^2 - y_{j,T+1}^2) - \tfrac{1}{J}(y_T{}'P(\widetilde{Y}_{\Lambda,-T})y_T - y_{T+1}{}'P(\widetilde{Y}_{\Lambda,-1})y_{T+1}).
\end{aligned}$$

The first term of the last expression can be shown to shrink to zero in probability if $\frac{1}{J}\sum((\theta_{jT}^2 + M_{jT}) - (\theta_{j,T+1}^2 + M_{j,T+1})) \to 0$, with some additional regularity conditions already assumed above.

Assumption 4.5. $\frac{1}{J}\sum(\theta_{jT}^2 - \theta_{j,T+1}^2) \to 0$ *and* $\frac{1}{J}\sum(M_{jT} - M_{j,T+1}) \to 0$.

The first statement holds almost surely if the θ_j's are assumed to be drawn from a covariance stationary matrix. **[If we model the "unstationarity" we can relax this but, of course, this comes with a price of additional modeling.]**

For the second term, we have

$$\begin{aligned}
&\tfrac{1}{J}y_T{}'P(\widetilde{Y}_{\Lambda,-T})y_T\\
=&\tfrac{1}{J}y_T'\widetilde{Y}_{\Lambda,-T}(\widetilde{Y}'_{\Lambda,-T}\widetilde{Y}_{\Lambda,-T})^{-1}\widetilde{Y}'_{\Lambda,-T}y_T\\
=&\tfrac{1}{J}y_T'\widetilde{Y}_{\Lambda,-T}(\tfrac{1}{J}\widetilde{Y}'_{\Lambda,-T}\widetilde{Y}_{\Lambda,-T})^{-1}\tfrac{1}{J}\widetilde{Y}'_{\Lambda,-T}y_T.
\end{aligned}$$

I derive a probability limit for $\frac{1}{J}\widetilde{Y}'_{\Lambda,-T}y_T$ and $\frac{1}{J}\widetilde{Y}'_{\Lambda,-T}\widetilde{Y}_{\Lambda,-T}$. We have

$$\begin{aligned}&\tfrac{1}{J}\widetilde{Y}'_{\Lambda,-T}y_T\\ =&\tfrac{1}{J}\textstyle\sum_{j=1}^{J}(\Lambda_{-T}+M_{j,-T})^{-1}y_{j,-T}y_{jT}\\ =&\tfrac{1}{J}\textstyle\sum_{j=1}^{J}(\Lambda_{-T}+M_{j,-T})^{-1}(y_{j,-T}y_{jT}-\theta_{j,-T}\theta_{j,T})+\tfrac{1}{J}\textstyle\sum_{j=1}^{J}(\Lambda_{-T}+M_{j,-T})^{-1}\theta_{j,-T}\theta_{j,T}.\end{aligned}$$

I show that the first term in the last line vanishes to zero in probability. Let $A_{j,k}$ denote the tth column of $(\Lambda_{-T}+M_{j,-T})^{-1}$. It is enough to show that

$$E\left(\tfrac{1}{J}\textstyle\sum_{j=1}^{J}A'_{j,t}(y_{j,-T}y_{jT}-\theta_{j,-T}\theta_{j,T})\right)^2\to 0$$

for all $t\leq T-1$, which holds if and only if

$$\textstyle\sum_{t=1}^{T-1}E\left(\tfrac{1}{J}\textstyle\sum_{j=1}^{J}A'_{j,t}(y_{j,-T}y_{jT}-\theta_{j,-T}\theta_{j,T})\right)^2\to 0.$$

Observe that

$$\begin{aligned}&E(\tfrac{1}{J}\textstyle\sum_{j=1}^{J}A'_{j,t}(y_{j,-T}y_{jT}-\theta_{j,-T}\theta_{j,T}))^2\\ =&\tfrac{1}{J^2}E\textstyle\sum_{j=1}^{J}(A'_{j,t}(y_{j,-T}y_{jT}-\theta_{j,-T}\theta_{j,T}))^2\\ =&\tfrac{1}{J^2}E\textstyle\sum_{j=1}^{J}A'_{j,t}\mathrm{Var}(y_{j,-T}y_{jT})A_{j,t}.\end{aligned}$$

It follows that

$$\begin{aligned}&\textstyle\sum_{t=1}^{T-1}E\left(\tfrac{1}{J}\textstyle\sum_{j=1}^{J}A'_{j,t}(y_{j,-T}y_{jT}-\theta_{j,-T}\theta_{j,T})\right)^2\\ =&\tfrac{1}{J^2}\textstyle\sum_{j=1}^{J}\mathrm{tr}\left((\Lambda_{-T}+M_{j,-T})^{-1}\mathrm{Var}(y_{j,-T}y_{jT})(\Lambda_{-T}+M_{j,-T})^{-1}\right).\end{aligned}$$

Under fairly weak conditions such as $\mathrm{Var}(y_{j,-T}y_{jT})$ is bounded and $\lambda_{T-1}(M_{j,-T})$ is bounded away from zero, this goes to 0.

Likewise, it is straightforward to show that

$$\tfrac{1}{J}\widetilde{Y}'_{\Lambda,-T}\widetilde{Y}_{\Lambda,-T}=\tfrac{1}{J}\textstyle\sum_{j=1}^{J}(\Lambda_{-T}+M_{j,-T})^{-1}(\theta_{j,-T}\theta'_{j,-T}+M_{j,-T})(\Lambda_{-T}+M_{j,-T})^{-1}+o_p(1).$$

Hence, if we can show

$$\begin{aligned}&\lim_{J\to\infty}\frac{1}{J}\sum_{j=1}^{J}(\Lambda_{-T}+M_{j,-T})^{-1}(\theta_{j,-T}\theta_{j,-T}'+M_{j,-T})(\Lambda_{-T}+M_{j,-T})^{-1}\\ =&\lim_{J\to\infty}\frac{1}{J}\sum_{j=1}^{J}(\Lambda_{-T}+M_{j,-1})^{-1}(\theta_{j,-1}\theta_{j,-1}'+M_{j,-1})(\Lambda_{-T}+M_{j,-1})^{-1}\text{ and}\\ &\lim_{J\to\infty}\frac{1}{J}\sum_{j=1}^{J}(\Lambda_{-T}+M_{j,-T})^{-1}\theta_{j,-T}\theta_{j,T}=\lim_{J\to\infty}\frac{1}{J}\sum_{j=1}^{J}(\Lambda_{-T}+M_{j,-1})^{-1}\theta_{j,-1}\theta_{j,T+1},\end{aligned}\tag{22}$$

uniformly over Λ, minimizing (20) and (21) are asymptotically equivalent. The first equality is implied by the following two inequalities,

$$\begin{aligned}&\tfrac{1}{J}\textstyle\sum_{j=1}^{J}(\Lambda_{-T}+M_{j,-T})^{-1}\theta_{j,-T}\theta_{j,-T}'(\Lambda_{-T}+M_{j,-T})^{-1}\\ =&\tfrac{1}{J}\textstyle\sum_{j=1}^{J}(\Lambda_{-T}+M_{j,-1})^{-1}\theta_{j,-1}\theta_{j,-1}'(\Lambda_{-T}+M_{j,-1})^{-1}\text{ and}\\ &\tfrac{1}{J}\textstyle\sum_{j=1}^{J}(\Lambda_{-T}+M_{j,-T})^{-1}M_{j,-T}(\Lambda_{-T}+M_{j,-T})^{-1}\\ =&\tfrac{1}{J}\textstyle\sum_{j=1}^{J}(\Lambda_{-T}+M_{j,-1})^{-1}M_{j,-1}(\Lambda_{-T}+M_{j,-1})^{-1}.\end{aligned}\tag{23}$$

[What if I use $M_{j,-1}$ in place of $M_{j,-T}$ so I at least have the scale right? This should work, but I should compare the conditions to see what makes more sense.]

Assumption 4.6 (Stationarity)**.** ***[Here,*** $\theta_j \in \mathbb{R}^{T+1}$ ***and*** M_j ***is a*** $(T+1)\times(T+1)$ ***matrix - terrible notation!]*** *$(M_j,\theta_j) \overset{i.i.d.}{\sim} (M,\theta)$ where $(M_{-1},\theta_{-1}) \overset{d}{=} (M_{-T},\theta_{-T})$.*

Under Assumption 4.6, both Assumption 3.1 and the high-level condition (22) holds almost surely by the strong law of large numbers, given mild regularity conditions. To show that the convergence is uniform over Λ, I show that the function that corresponds to the summand belongs to the Glivenko-Cantelli class. I also need is $\inf_j \lambda_T(M_j) > \varepsilon$ almost surely for some $\varepsilon > 0$.

More questions

[Can the "correlated mean and variance" case be extended to this context?]
[What is the class of "weights" I can allow for the method to remain valid.]

[What is the class of estimators that such precise estimation of risk is possible? See Section 7 of Chapter 4, TPE.]
[Is there any advantage of taking the positive part in this context as well?]
[Additional covariates (i.e., teacher characteristics) can be interesting.]
[Take a closer look of the last paragraph on p.282 of TPE, and exercise 7.16, which references Morris (1983a).)]
[Exercise 7.17 derives an EB for the "general case" - see of there's anything useful.]
[Exercise 7.18 mentions minimax properties for the general estimator.]
[It seems that what we need is the additiveness of the fixed effects but not necessarily the other parts of the regression function.]
[An extension of Brown et al. (2018) may be possible by allowing a covariance term between α_i and β_j. However, here it is unclear why would one not just to each vector at a time. That probably is the correct way?]
[How about focusing on the last period prediction? Will the inclusion of data from previous years help?]
[Blockwise estimators?]
[Teacher fixed effects interacted with other student characteristics?]
[Using the tweedie formula, and then restricting f? This seems infeasible; even with $T > 3$ the kernel estimator should be painfully inefficient...]
[How about restricting the prior to belong to a location-scale family; or just assuming it belongs to a scale family may considerably simply things]
[Seems that I can generalize this to the exponential family]
[What about jointly drawing (θ_j, σ_j)?]
[The point is that, under heteroskedasticity, the equivalence between the compound decision problem and the empirical Bayes setting (under separable rules) does not hold, and thus the optimality result in the sense of Jiang and Zhang (2009) is not feasible. However, maybe we can ask the question of in what cases does the GMLEB obtain the "oracle?"] [I think I should plug Q back in again so that I can

allow for estimation for subperiods and/or averages.]

[As Efron and Morris (1973) - see Section 9 - and Xie et al. (2012) mention, it seems that]

[Can we play around with Q to get the desired result?]

[Section 3.3 of Fourdrinier et al. (2018) seems useful. Chapter 6 also gives some interesting results. Chapter 8 gives some results for loss estimation, which can also be useful.]

References

Abowd, J. M. and F. Kramarz (1999). Chapter 40 The analysis of labor markets using matched employer-employee data. In *Handbook of Labor Economics*, Volume 3, pp. 2629–2710. Elsevier.

Andrews, D. W. (1992). Generic Uniform Convergence. *Econometric Theory 8*(2), 241–257.

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

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*(4), 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*(4), 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*(9), 2593–2632.

Chetty, R., J. N. Friedman, and J. E. Rockoff (2014b). Measuring the Impacts of Teachers II: Teacher Value-Added and Student Outcomes in Adulthood. *American Economic Review 104*(9), 2633–2679.

Chetty, R. and N. Hendren (2018a). The Impacts of Neighborhoods on Intergenerational Mobility I: Childhood Exposure Effects*. *The Quarterly Journal of Economics 133*(3), 1107–1162.

Chetty, R. and N. Hendren (2018b). The Impacts of Neighborhoods on Intergenerational Mobility II: County-Level Estimates*. *The Quarterly Journal of Economics 133*(3), 1163–1228.

Efron, B. and C. Morris (1973). Stein's Estimation Rule and Its Competitors–An Empirical Bayes Approach. pp. 15.

Fourdrinier, D., W. E. Strawderman, and M. T. Wells (2018). *Shrinkage Estimation*. Springer Series in Statistics. Cham: Springer International Publishing.

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

Fu, L. J., W. Sun, and G. M. James (2020). Nonparametric Empirical Bayes Estimation on Heterogeneous Data. *arXiv:2002.12586 [stat]*.

Gilraine, M., J. Gu, and R. McMillan (2019). A New Method for Estimating Teacher Value-Added.

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*(2), 190–222.

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

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*(4), 1647–1684.

Koenker, R. and I. Mizera (2014a). Convex Optimization, Shape Constraints, Compound Decisions, and Empirical Bayes Rules. *Journal of the American Statistical Association 109*(506), 674–685.

Koenker, R. and I. Mizera (2014b). Convex Optimization, Shape Constraints, Compound Decisions, and Empirical Bayes Rules. *Journal of the American Statistical Association 109*(506), 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.

Liu, L., H. R. Moon, and F. Schorfheide (2019). Forecasting with Dynamic Panel Data Models. pp. 92.

Mirsky, L. (1975). A trace inequality of John von Neumann. *Monatshefte für Mathematik 79*(4), 303–306.

Morris, C. N. (1983). Parametric Empirical Bayes Inference: Theory and Applications. *Journal of the American Statistical Association 78*(381), 47–55.

Okolewski, A. and T. Rychlik (2001). Sharp distribution-free bounds on the bias in estimating quantiles via order statistics. *Statistics & Probability Letters 52*(2), 207–213.

Rockoff, J. E. (2004). The Impact of Individual Teachers on Student Achievement: Evidence from Panel Data. *American Economic Review 94*(2), 247–252.

Rothstein, J. (2010). Teacher Quality in Educational Production: Tracking, Decay, and Student Achievement. *The Quarterly Journal of Economics 125*(1), 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.

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*(2), 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*(500), 1465–1479.

Appendix A Proof of Theorem 4.1

The difference between the risk estimate and the loss is given as

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

Expanding the summand of the last expression gives

$$\begin{aligned}
&\mathbf{URE}_j(\mu,\Lambda) - (\widehat{\theta}_j(\mu,\Lambda)-\theta_j)'(\widehat{\theta}_j(\mu,\Lambda)-\theta_j) \\
=&\mathrm{tr}(M_j) - 2\mathrm{tr}((\Lambda+M_j)^{-1}M_j^2) \\
&+ (y_j-\mu)'[(\Lambda+M_j)^{-1}M_j^2(\Lambda+M_j)^{-1}](y_j-\mu) \\
&- (y_j-\theta_j-M_j(\Lambda+M_j)^{-1}(y_j-\mu))'(y_j-\theta_j-M_j(\Lambda+M_j)^{-1}(y_j-\mu)) \\
=&\mathrm{tr}(M_j) - 2\mathrm{tr}((\Lambda+M_j)^{-1}M_j^2) - (y_j-\theta_j)'(y_j-\theta_j) \\
&+ 2(y_j-\mu)'(\Lambda+M_j)^{-1}M_j(y_j-\theta_j) \\
=&y_j'y_j - \theta_j'\theta_j - \mathrm{tr}(M_j) - 2\mathrm{tr}(\Lambda(\Lambda+M_j)^{-1}(y_jy_j'-\theta_jy_j'-M_j)) \\
&- 2\mu'(\Lambda+M_j)^{-1}M_j(y_j-\theta_j).
\end{aligned} \tag{24}$$

Take $\mu=0$, to obtain (14). We have

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

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 $Ey_j'y_j = \theta_j'\theta_j + \mathrm{tr}(M_j)$ for all $j \le J$ and y_j's are independent, we have

$$E\left(\frac{1}{J}\sum_{j=1}^{J}(y_j'y_j-\theta_j'\theta_j-\mathrm{tr}(M_j))\right)^2 = \frac{1}{J^2}\sum_{j=1}^{J}E(y_j'y_j-\theta_j'\theta_j-\mathrm{tr}(M_j))^2$$
$$= \frac{1}{J^2}\sum_{j=1}^{J}\mathrm{Var}(y_j'y_j).$$

Therefore, if $\lim \frac{1}{J^2}\sum_{j=1}^{J}\mathrm{Var}(y_j'y_j)=0$, then $(\mathrm{I})_J$ 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, we have

$$\sup_{\Lambda}\left|\frac{1}{J}\sum_{j=1}^{J}\mathrm{tr}(\Lambda(\Lambda+M_j)^{-1}(y_jy_j'-\theta_jy_j'-M_j))\right|$$
$$=\sup_{\Lambda}\left|\frac{1}{J}\sum_{j=1}^{J}\mathrm{tr}((I-M_j(\Lambda+M_j)^{-1})(y_jy_j'-\theta_jy_j'-M_j))\right|$$
$$\le\left|\frac{1}{J}\sum_{j=1}^{J}(y_j'y_j-\theta_j'y_j-\mathrm{tr}(M_j))\right|+\sup_{\Lambda}\left|\frac{1}{J}\sum_{j=1}^{J}\mathrm{tr}(M_j(\Lambda+M_j)^{-1}(y_jy_j'-\theta_jy_j'-M_j))\right|$$
$$=: (\mathrm{I})_J+(\mathrm{II})_J.$$

To show that $(\mathrm{I})_J \overset{L_1}{\to} 0$, we again show L_2 convergence. Because $E(y_j'y_j-\theta_j'y_j)=\mathrm{tr}(M_j)$ for all $j\le J$ and y_j's are independent, we have

$$E\left(\frac{1}{J}\sum_{j=1}^{J}(y_j'y_j-\theta_j'y_j-\mathrm{tr}(M_j))\right)^2 =\frac{1}{J^2}\sum_{j=1}^{J}E(y_j'y_j-\theta_j'y_j-\mathrm{tr}(M_j))^2$$
$$=\frac{1}{J^2}\sum_{j=1}^{J}\mathrm{var}(y_j'y_j-\theta_j'y_j).$$

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

$$\mathrm{Var}(y_j'y_j-\theta_j'y_j)\le 2\mathrm{Var}(y_j'y_j)+2\theta_j'M_j\theta_j\le 2\mathrm{Var}(y_j'y_j)+2\mathrm{tr}(M_j)\|\theta_j\|_\infty^2.$$

Hence, if $\limsup_{J\to\infty}\frac{1}{J}\sum_{j=1}^{J}(\mathrm{Var}(y_j'y_j)+\mathrm{tr}(M_j)\|\theta_j\|_\infty^2)<\infty$ it follows that $(\mathrm{I})_J\overset{L_2}{\to}0$. A simple sufficient condition is that $\sup_j\mathrm{Var}(y_j'y_j)$, $\sup_j\mathrm{tr}(M_j)$, and

$\sup_j \|\theta_j\|_\infty^2$ are all finite, which is the case by Assumption 4.1.

To show that $(\mathrm{II})_J \xrightarrow{L_1} 0$, define the random function $G_j(\Lambda)$ as

$$G_J(\Lambda) = \frac{1}{J} \sum_{j=1}^{J} \mathrm{tr}(M_j(\Lambda + M_j)^{-1}(y_j y_j' - \theta_j y_j' - M_j)).$$

so that the aim is to show $\sup_\Lambda |G_J(\Lambda)| \xrightarrow{L_1} 0$. I establish this by showing $\sup_\Lambda |G_J(\Lambda)| \xrightarrow{p} 0$ 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 all require a totally bounded parameter space. However, the parameter space in consideration, S_T^+, does not satisfy this requirement. I overcome this by taking an appropriate reparametrization. Let $\underline{\sigma}_M = \inf_j \sigma_T(M_j)$ denote the infimum of the smallest eigenvalues of M_j's for $j \geq 1$, which is bounded away from zero. Consider the transformation defined by $h(\Lambda) = (\underline{\sigma}_M I_T + \Lambda)^{-1}$, and write the image of such transformation as $\widetilde{\mathcal{L}} := \{h(\Lambda) : \Lambda \in S_T^+\}$. Note that $h : S_T^+ \to \widetilde{\mathcal{L}}$ is one-to-one and onto, with its inverse given as $h^{-1}(\widetilde{\Lambda}) = \widetilde{\Lambda}^{-1} - \underline{\sigma}_M I_T$. For $\widetilde{\Lambda} \in \widetilde{\mathcal{L}}$, define $\widetilde{G}_J := G_J \circ h^{-1}$ so that

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

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 S_T denote the set of all real positive $T \times T$ matrices. While the choice of metric is not important, I equip S_T with the metric induced by the Frobenius norm for concreteness. Note that $\widetilde{\mathcal{L}} \subset S_T$ In the latter formation, the parameter space $\widetilde{\Lambda}$ is indeed totally bounded. To see this, let the metric d (on S_T) induced by the Frobenius norm. For any $\widetilde{\Lambda} \in \widetilde{\mathcal{L}}$, we have $0 \leq \widetilde{\Lambda} \leq 1$ so that $\sigma_1(\widetilde{L}) \leq 1$, where $\sigma_1(A)$ denotes the largest singular value of a matrix A. 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 $\Lambda \in S_T^+$ instead because for any $\widetilde{\Lambda} \in \widetilde{\mathcal{L}}$, there exists $\Lambda \in S_T^+$ such that $G_J(\Lambda) = \widetilde{G}_J(\widetilde{\Lambda})$.

Now, note that

$$\begin{aligned}&E\mathrm{tr}(M_j(\Lambda+M_j)^{-1}(y_jy_j'-\theta_jy_j'-M_j))\\=&\,\mathrm{tr}(M_j(\Lambda+M_j)^{-1}E(y_jy_j'-\theta_jy_j'-M_j))=0,\end{aligned}$$

and y_j's are independent. This gives

$$E\left[G_J(\Lambda)^2\right]=\frac{1}{J^2}\sum_{j=1}^{J}E\mathrm{tr}(M_j(\Lambda+M_j)^{-1}(y_jy_j'-\theta_jy_j'-M_j))^2$$

I give a bound on $|\mathrm{tr}(M_j(\Lambda+M_j)^{-1}(y_jy_j'-\theta_jy_j'-M_j))|$. It follows from (13) that

$$\begin{aligned}&\mathrm{tr}(M_j(\Lambda+M_j)^{-1}(y_jy_j'-\theta_jy_j'-M_j))\\=&\mathrm{tr}(M_j^{1/2}U(I_T+D)^{-1}U'M_j^{-1/2}(y_jy_j'-\theta_jy_j'-M_j))\\=&\mathrm{tr}((I_T+D)^{-1}U'M_j^{-1/2}(y_jy_j'-\theta_jy_j'-M_j)M_j^{1/2}U).\end{aligned}\tag{25}$$

[It seems that this part of the proof can be greatly simplified, by using the fact that

$$\begin{aligned}&|\mathrm{tr}(M_j(\Lambda+M_j)^{-1}(y_jy_j'-\theta_jy_j'-M_j))|\\\leq&\sum_{t=1}^{T}\sigma_t(y_jy_j'-\theta_jy_j'-M_j)\\\leq&T\sigma_1(y_jy_j'-\theta_jy_j'-M_j)\end{aligned}$$

which follows by Mirsky (1975), and then proceeding as usual.] Write $H_j=M_j^{-1/2}(y_jy_j'-\theta_jy_j'-M_j)M_j^{1/2}$, and observe that

$$\begin{aligned}&\left|\mathrm{tr}((I_T+D)^{-1}U'M_j^{-1/2}(y_jy_j'-\theta_jy_j'-M_j)M_j^{1/2}U)\right|\\=&\left|\sum_{t=1}^{T}\frac{1}{1+d_t}(U'H_jU)_{tt}\right|\\\leq&\sum_{t=1}^{T}\frac{1}{1+d_t}\left|(U'H_jU)_{tt}\right|\\\leq&\sum_{t=1}^{T}\left|(U'H_jU)_{tt}\right|,\end{aligned}\tag{26}$$

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

tth column of the orthogonal matrix U. We have

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

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, we have

$$E\left[G_J(\Lambda)^2\right] \leq \frac{T^2}{J^2}\sum_{j=1}^{J} E[\sigma_1(H_j)^2].$$

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

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

Since the largest singular value of a matrix is bounded by its Frobenius norm, we have

$$\begin{aligned}&\sigma_1(y_jy_j' - \theta_jy_j' - M_j)^2\\ \leq& \text{tr}((y_jy_j' - \theta_jy_j' - M_j)'(y_jy_j' - \theta_jy_j' - M_j))\\ =&(y_j'y_j)^2 + \theta_j'\theta_jy_j'y_j + \sigma^4\text{tr}(M_j)^2 - 2y_j'y_jy_j'\theta_j - 2y_j'M_jy_j + 2\theta_j'M_jy_j.\end{aligned}$$

Taking expectations, we obtain

$$\begin{aligned}&E(y_j'y_j)^2 + \theta_j'\theta_jEy_j'y_j + \sigma^4\text{tr}(M_j)^2 - 2Ey_j'y_jy_j'\theta_j - 2Ey_j'M_jy_j + 2\theta_j'M_jEy_j\\ =&var(y_j'y_j) + (\theta_j'\theta_j + \text{tr}(M_j))^2 + \theta_j'\theta_j(\theta_j'\theta_j + \text{tr}(M_j))\\ &+ \sigma^4\text{tr}(M_j)^2 - 2\theta_j'E(y_jy_j'y_j) - 2\theta_j'M_j\theta_j - 2\sigma^4\text{tr}(M_j)^2 + 2\theta_j'M_j\theta_j\\ =&var(y_j'y_j) + (\theta_j'\theta_j)^2 + 2\theta_j'\theta_j\text{tr}(M_j) + \sigma^4\text{tr}(M_j)^2 + (\theta_j'\theta_j)^2 + \theta_j'\theta_j\text{tr}(M_j)\\ &+ \sigma^4\text{tr}(M_j)^2 - 2\theta_j'E(y_jy_j'y_j) - 2\theta_j'M_j\theta_j - 2\sigma^4\text{tr}(M_j)^2 + 2\theta_j'M_j\theta_j\\ =&var(y_j'y_j) + 2(\theta_j'\theta_j)^2 + 3\theta_j'\theta_j\text{tr}(M_j) - 2\theta_j'E(y_jy_j'y_j)\\ \leq&var(y_j'y_j) + 2\|\theta_j\|^4 + 3\|\theta_j\|^2\text{tr}(M_j) + 2\|\theta_j\|E\|y_j\|^3.\end{aligned}$$

This shows that if

$$\limsup_{J\to\infty}\frac{1}{J}\sum_{j=1}^{J}\kappa(M_j)\left(var(y_j'y_j)+2\|\theta_j\|^4+3\|\theta_j\|^2\mathrm{tr}(M_j)+2\|\theta_j\|E(\|y_j\|^3)\right)<\infty, \tag{27}$$

then $|G_J(\Lambda)|\to 0$ in L_2, and thus in probability. For example, if $\sup_j \sigma_1(M_j)/\sigma_T(M_j)$, $\sup_j \mathrm{var}(y_j'y_j)$, $\sup_j|\theta_j|$, and $\sup_j \mathrm{tr}(M_j)$ are bounded, the result holds.

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 $\Lambda_\alpha=(\alpha\widetilde{\Lambda}+(1-\alpha)\widetilde{\Lambda}^\dagger)^{-1}-\underline{\sigma}_M I_T$, we have $h(\Lambda_\alpha)=\alpha\widetilde{\Lambda}+(1-\alpha)\widetilde{\Lambda}^\dagger$, which shows that $\alpha\widetilde{\Lambda}+(1-\alpha)\widetilde{\Lambda}^\dagger\in\widetilde{\mathcal{L}}$. Now, by the mean value theorem, we have

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

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

$$|\widetilde{G}_J(\widetilde{\Lambda})-\widetilde{G}_J(\widetilde{\Lambda}^\dagger)|\le\|\nabla\widetilde{G}_J(\widetilde{\Lambda}^\alpha)\|\|\widetilde{\Lambda}-\widetilde{\Lambda}^\dagger\|, \tag{28}$$

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 we also have $\|\nabla\widetilde{G}_J(\widetilde{\Lambda})\|=\|\frac{\partial}{\partial\widetilde{\Lambda}}\widetilde{G}_J(\widetilde{\Lambda})\|$.

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

$$\begin{aligned}&\tfrac{\partial}{\partial\widetilde{\Lambda}}G_J(\widetilde{\Lambda})\\ =&\tfrac{1}{J}\textstyle\sum_{j=1}^{J}\widetilde{\Lambda}^{-1}(\widetilde{\Lambda}^{-1}-\underline{\sigma}_M I_T+M_j)^{-1}M_j(y_jy_j'-y_j\theta_j'-M_j)(\widetilde{\Lambda}^{-1}-\underline{\sigma}_M I_T+M_j)^{-1}\widetilde{\Lambda}^{-1}.\end{aligned}$$

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

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

Since the operator norm is submultiplicative, we have

$$\sigma_1(g_j(\widetilde{\Lambda})) \leq \sigma_1(\widetilde{\Lambda}^{-1}(\widetilde{\Lambda}^{-1} - \underline{\sigma}_M I_T + M_j)^{-1})^2 \sigma_1(M_j(y_j y_j' - y_j\theta_j' - M_j)).$$

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

$$\begin{aligned}
&\sigma_1(\widetilde{\Lambda}^{-1}(\widetilde{\Lambda}^{-1} - \underline{\sigma}_M I_T + M_j)^{-1})^2 \\
=&\sigma_1(\widetilde{\Lambda}^{-1}(\widetilde{\Lambda}^{-1} - \underline{\sigma}_M I_T + M_j)^{-2}\widetilde{\Lambda}^{-1}) \\
=&\sigma_1((I + \widetilde{\Lambda}^{1/2}(M_j - \underline{\sigma}_M I_T)\widetilde{\Lambda}^{1/2})^{-2}) \\
\leq&1,
\end{aligned} \tag{29}$$

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 $\widetilde{\Lambda}^{1/2}(M_j - \underline{\sigma}_M I_T)\widetilde{\Lambda}^{1/2}$ is positive semidefinite so that $0 \leq (I + \widetilde{\Lambda}^{1/2}(M_j - \underline{\sigma}_M I_T)\widetilde{\Lambda}^{1/2})^{-2} \leq I_T$, and $A \leq B$ implies $\sigma_1(A) \leq \sigma_1(B)$ for any two positive semidefinite matrices A and B. We can bound $\sigma_1(M_j(y_j y_j' - y_j\theta_j' - M_j))$ simply by

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

Combining these results, we have

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

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

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

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 (28), we have

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

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 $\sup_{\Lambda \in S_T^+} |G_J(\Lambda)| \xrightarrow{p} 0$.

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

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

where the last inequality follows from (25) and (26). Let $\overline{G}_J$ denote the expression in the last line, and suppose that $\limsup_{J \to \infty} E\overline{G}_J^2 < \infty$, which I verify below. Then, we have $\sup_J E(\sup_\Lambda |G_J(\Lambda)|)^2 < \infty$, from which the uniform integrability follows. It remains only to show that $\limsup_{J \to \infty} E\overline{G}_J^2 < \infty$. By Cauchy-Schwarz, we have

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

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

Appendix B Proof of Theorem 4.3

Observe that, by (24),

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

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

$$\sup_{\mu,\Lambda} \left| \tfrac{1}{J} \textstyle\sum_{j=1}^J \mu'(\Lambda + M_j)^{-1} M_j (y_j - \theta_j) \right| \xrightarrow{L_1} 0. \tag{30}$$

By two applications of Cauchy-Schwarz, we have

$$\begin{aligned}
&E\sup_{\mu,\Lambda}\left|\tfrac{1}{J}\textstyle\sum_{j=1}^{J}\mu'(\Lambda+M_j)^{-1}M_j(y_j-\theta_j)\right|\\
\le{}& E\sup_{\mu}\|\mu\|\cdot\sup_{\Lambda}\left\|\tfrac{1}{J}\textstyle\sum_{j=1}^{J}(\Lambda+M_j)^{-1}M_j(y_j-\theta_j)\right\| \qquad (31)\\
\le{}& \left(E\sup_{\mu}\|\mu\|^2\right)^{1/2}\left(E\sup_{\Lambda}\left\|\tfrac{1}{J}\textstyle\sum_{j=1}^{J}(\Lambda+M_j)^{-1}M_j(y_j-\theta_j)\right\|^2\right)^{1/2}.
\end{aligned}$$

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

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

from which then (30) will follow.

Write $H_J(\Lambda):=\|\frac{1}{J}\sum_{j=1}^{J}(\Lambda+M_j)^{-1}M_j(y_j-\theta_j)\|$. As in the proof for Theorem 4.1, I show (a) $\sup_{\Lambda}H_J(\Lambda)\xrightarrow{p}0$ and (b) $\sup_J E(\sup_{\Lambda}H_J(\Lambda))^{2+\delta}<\infty$ for some $\delta>0$. Since (b) is a sufficient condition for $\left\{\sup_{\Lambda}H_J(\Lambda)^2\right\}_{J\ge1}$ 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 uniform law of large numbers argument as in Andrews (1992). First, to show $H_J(\Lambda)\xrightarrow{p}0$, it is enough to show $EH_J(\Lambda)^2\to0$. We have

$$\begin{aligned}
EH_J(\Lambda)^2 &= E\|\tfrac{1}{J}\textstyle\sum_{j=1}^{J}(\Lambda+M_j)^{-1}M_j(y_j-\theta_j)\|^2\\
&= \tfrac{1}{J^2}E(\textstyle\sum_{\ell=1}^{J}(\Lambda+M_\ell)^{-1}M_\ell(y_\ell-\theta_\ell))'(\textstyle\sum_{j=1}^{J}(\Lambda+M_j)^{-1}M_j(y_j-\theta_j))\\
&= \tfrac{1}{J^2}E(\textstyle\sum_{j=1}^{J}(y_j-\theta_j)'M_j(\Lambda+M_j)^{-2}M_j(y_j-\theta_j))\\
&\le \tfrac{1}{J^2}\textstyle\sum_{j=1}^{J}\operatorname{tr}(M_j(\Lambda+M_j)^{-2}M_jM_j)\\
&\le \tfrac{1}{J^2}\textstyle\sum_{j=1}^{J}\operatorname{tr}(M_j),
\end{aligned}$$

where the last inequality follows from von Neumann's trace inequality and the fact that $\sigma_1(M_j(\Lambda+M_j)^{-2}M_j)\le1$. Moreover, we have $\sup_j\operatorname{tr}(M_j)\le T\sup_j\sigma_1(M_j)<\infty$, which implies $\frac{1}{J^2}\sum_{j=1}^{J}\operatorname{tr}(M_j)\to0$. This establishes that $H_J(\Lambda)$ converges to zero in L_2, and thus in probability.

To show that this convergence is uniform over $\Lambda\in S_T^+$, 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.,

$$|\widetilde{H}_J(\widetilde{\Lambda})-\widetilde{H}_J(\widetilde{\Lambda}^\dagger)|\le B_{H,J}\|\widetilde{\Lambda}-\widetilde{\Lambda}^\dagger\| \qquad (32)$$

for all $\widetilde{\Lambda},\widetilde{\Lambda}^\dagger\in\widetilde{\mathcal{L}}$, where $B_{H,J}=O_p(1)$. Define $\widetilde{A}_j=\widetilde{\Lambda}^{-1}+(M_j-\underline{\sigma}_MI_T)$ and

$\widetilde{A}_j^{\dagger}$ likewise with $\widetilde{\Lambda}$ replaced with $\widetilde{\Lambda}^{\dagger}$. Observe that

$$\begin{aligned} |\widetilde{H}_J(\widetilde{\Lambda}) - \widetilde{H}_J(\widetilde{\Lambda}^{\dagger})| =& |\|\tfrac{1}{J}\textstyle\sum_{j=1}^{J} \widetilde{A}_j^{-1} M_j(y_j - \theta_j)\| - \|\tfrac{1}{J}\textstyle\sum_{j=1}^{J} \widetilde{A}_j^{\dagger -1} M_j(y_j - \theta_j)\|| \\ \le& \|\tfrac{1}{J}\textstyle\sum_{j=1}^{J}(\widetilde{A}_j^{-1} - \widetilde{A}_j^{\dagger -1}) M_j(y_j - \theta_j)\| \\ \le& \tfrac{1}{J}\textstyle\sum_{j=1}^{J} \sigma_1(\widetilde{A}_j^{-1} - \widetilde{A}_j^{\dagger -1}) \|M_j(y_j - \theta_j)\|, \end{aligned} \tag{33}$$

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

$$\begin{aligned} \widetilde{A}_j^{-1} - \widetilde{A}_j^{\dagger -1} =& \widetilde{A}_j^{\dagger -1}(\widetilde{A}_j^{\dagger} - \widetilde{A}_j)\widetilde{A}_j^{-1} \\ =& \widetilde{A}_j^{\dagger -1}(\widetilde{\Lambda}^{\dagger -1} - \widetilde{\Lambda}^{-1})\widetilde{A}_j^{-1} \\ =& \widetilde{A}_j^{\dagger -1}\widetilde{\Lambda}^{-1}(\widetilde{\Lambda} - \widetilde{\Lambda}^{\dagger})\widetilde{\Lambda}^{\dagger -1}\widetilde{A}_j^{-1}, \end{aligned} \tag{34}$$

which implies

$$\begin{aligned} \sigma_1(\widetilde{A}_j^{-1} - \widetilde{A}_j^{\dagger -1}) \le& \sigma_1(\widetilde{A}_j^{\dagger -1}\widetilde{\Lambda}^{-1})\sigma_1(\widetilde{\Lambda} - \widetilde{\Lambda}^{\dagger})\sigma_1(\widetilde{\Lambda}^{\dagger -1}\widetilde{A}_j^{-1}) \\ \le& (\textstyle\sup_{\widetilde{\Lambda} \in \widetilde{\mathcal{L}}} \sigma_1((\widetilde{\Lambda}^{-1} + (M_j - \underline{\sigma}_M I_T))^{-1}\widetilde{\Lambda}^{-1})^2)\sigma_1(\widetilde{\Lambda} - \widetilde{\Lambda}^{\dagger}), \end{aligned} \tag{35}$$

where the first inequality follows from (34) 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 (29) that $\sigma_1((\widetilde{\Lambda}^{-1} + (M_j - \underline{\sigma}_M I_T))^{-1}\widetilde{\Lambda}^{-1})^2$ is bounded above by 1. Hence, we have

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

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 (33), we have

$$|\widetilde{H}_J(\widetilde{\Lambda}) - \widetilde{H}_J(\widetilde{\Lambda}^{\dagger})| \le (\tfrac{1}{J}\textstyle\sum_{j=1}^{J}\|M_j(y_j - \theta_j)\|)\|\widetilde{\Lambda} - \widetilde{\Lambda}^{\dagger}\|.$$

Therefore, it remains to show $\frac{1}{J}\sum_{j=1}^{J}\|M_j(y_j - \theta_j)\| = O_p(1)$ to establish (32).

Observe that

$$\begin{aligned}&\tfrac{1}{J}\textstyle\sum_{j=1}^{J}\|M_j(y_j-\theta_j)\|\\ =&\tfrac{1}{J}\textstyle\sum_{j=1}^{J}(\|M_j(y_j-\theta_j)\|-E\|M_j(y_j-\theta_j)\|)+\tfrac{1}{J}\textstyle\sum_{j=1}^{J}E\|M_j(y_j-\theta_j)\|.\end{aligned}$$

Since $\sup_j \operatorname{var}(\|M_j(y_j-\theta_j)\|)\leq \sup_j E\|M_j(y_j-\theta_j)\|^2=\sup_j \operatorname{tr}(M_j)^3<\infty$ the first term converges to zero in probability by an application of Chebyshev's inequality. Also, because $\sup_j E\|M_j(y_j-\theta_j)|\leq \sup_j \sigma_1(M_j)(E\|y_j\|+\|\theta_j\|)<\infty$, the second term is $O(1)$. This establishes $\frac{1}{J}\sum_{j=1}^{J}\|M_j(y_j-\theta_j)\|=O_p(1)$, 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 $\left\{\sup_\Lambda H_J(\Lambda)^2\right\}_{J\leq 1}$ is uniformly integrable. A sufficient condition for this is

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

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

$$\begin{aligned}H_J(\Lambda)=&\|\tfrac{1}{J}\textstyle\sum_{j=1}^{J}(\Lambda+M_j)^{-1}M_j(y_j-\theta_j)\|\\ \leq&\tfrac{1}{J}\textstyle\sum_{j=1}^{J}\sigma_1((\Lambda+M_j)^{-1}M_j))\|y_j-\theta_j\|\\ \leq&\tfrac{1}{J}\textstyle\sum_{j=1}^{J}\|y_j-\theta_j\|,\end{aligned}$$

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

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

Therefore, we have

$$\begin{aligned}\sup_\Lambda H_J(\Lambda)^{2+\delta}\leq&\left(\tfrac{1}{J}\textstyle\sum_{j=1}^{J}\|y_j-\theta_j\|\right)^{2+\delta}\\ \leq&\tfrac{1}{J}\textstyle\sum_{j=1}^{J}\|y_j-\theta_j\|^{2+\delta}\\ \leq&\tfrac{1}{J}\textstyle\sum_{j=1}^{J}2^{1+\delta}(\|y_j\|^{2+\delta}+\|\theta_j\|^{2+\delta}),\end{aligned}\tag{36}$$

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 $\delta\in[0,2]$,

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

and thus $\sup_J E\sup_\Lambda H_J(\Lambda)^{2+\delta} < \infty$. This concludes the proof for $\sup_\Lambda H_J(\Lambda) \overset{L_2}{\to} 0$.

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

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

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

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

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. We have $q_{1-\tau/2}(\{\theta_{jt}^2\}_{j=1}^J) < \sup_j \theta_{jt}^2 < \infty$, and thus is suffices to show that $\limsup_{J\to\infty} Eq_{1-\tau/2}(\{\varepsilon_{jt}^2\}_{j=1}^J) < \infty$. It follows that

$$q_{1-\tau/2}(\{\varepsilon_{jt}^2\}_{j=1}^J) = q_{1-\tau/2}(\{M_{jt}\eta_{jt}^2\}_{j=1}^J) \le q_{1-\tau/2}(\{\overline{M}_t\eta_{jt}^2\}_{j=1}^J) = \overline{M}_t q_{1-\tau/2}(\{\eta_{jt}^2\}_{j=1}^J),$$

where the first equality holds by Assumption 4.2 and the inequality holds because replacing M_{jt} by $\overline{M}_t$ makes all the sample points larger, and thus the sample quantile larger. By a result on the bias of sample quantiles given by Okolewski and Rychlik (2001), we have

$$\sup_J Eq_{1-\tau/2}(\{\eta_{jt}^2\}_{j=1}^J) \le \left(\frac{\operatorname{var}(\eta_{jt}^2)}{(1-\tau/2)\tau/2}\right)^{1/2} + F_t^{-1}(1-\tau/2) < \infty.$$

This establishes that $E\sup_\mu \|\mu\|^2 < \infty$, which concludes the proof.

Appendix C Proof of Theorem 4.4

By essentially the same calculations given in (31), it is enough to show

$$\begin{aligned} &\limsup_{J\to\infty} E\sup\nolimits_{\gamma\in\Gamma_J} \|\gamma\|^2 < \infty, \text{ and} \\ &\lim_{J\to\infty} E\sup\nolimits_\Lambda \left\| \tfrac{1}{J}\textstyle\sum_{j=1}^J Z_j'(\Lambda + M_j)^{-1}M_j(y_j - \theta_j)\right\|^2 = 0, \end{aligned}$$

The first line is equivalent to showing

$$\limsup_{J\to\infty} E(\textstyle\sum_{j=1}^J y_j' Z_j)(\sum_{j=1}^J Z_j' Z_j)^{-2}(\sum_{j=1}^J Z_j' y_j) < \infty. \tag{37}$$

Simple calculations show that

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

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 (37). Note that

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

with $\sigma_1((\frac{1}{J}\sum_{j=1}^J Z_j' Z_j)^{-2}) \to \sigma_1((EZ_j' Z_j)^{-2})$ 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

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

it follows that

$$\begin{aligned}
&E(\textstyle\sum_{j=1}^J \varepsilon_j' Z_j)(\sum_{j=1}^J Z_j' Z_j)^{-2}(\sum_{j=1}^J Z_j' \varepsilon_j)\\
=&\mathrm{tr}((\textstyle\sum_{j=1}^J Z_j' Z_j)^{-2}(\sum_{j=1}^J \sum_{\ell=1}^J Z_j' E[\varepsilon_j \varepsilon_\ell'] Z_\ell))\\
=&\mathrm{tr}((\textstyle\sum_{j=1}^J Z_j' Z_j)^{-2}(\sum_{j=1}^J Z_j' M_j Z_j))\\
=&\textstyle\frac{1}{J}\mathrm{tr}((\frac{1}{J}\sum_{j=1}^J Z_j' Z_j)^{-2}(\frac{1}{J}\sum_{j=1}^J Z_j' M_j Z_j))
\end{aligned}$$

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

$$\limsup\nolimits_{J\to\infty}\|\textstyle\frac{1}{J}\sum_{j=1}^J Z_j' M_j Z_j\| \le \limsup_{J\to\infty} \frac{1}{J}\sum_{j=1}^J \sigma_1(Z_j)^2 \sigma_1(M_j) < \infty.$$

This shows that $\frac{1}{J}\mathrm{tr}((\frac{1}{J}\sum_{j=1}^J Z_j'Z_j)^{-2}(\frac{1}{J}\sum_{j=1}^J Z_j'M_jZ_j)) \to 0$, which concludes the proof for (37).

To show $\lim_{J\to\infty} E\sup_\Lambda \|\frac{1}{J}\sum_{j=1}^J Z_j'(\Lambda + M_j)^{-1}M_j(y_j - \theta_j)\|^2 = 0$, I follow the lines of argument given in the proof of Theorem 4.3 carefully. The main difference is that now we have a Z_j' multiplied in the summand. Write $H_{Z,J}(\Lambda) := \|\frac{1}{J}\sum_{j=1}^J Z_j'(\Lambda + M_j)^{-1}M_j(y_j - \theta_j)\|$. 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 $\sup_j \sigma_1(Z_jZ_j') = \sup_j \sigma_1(Z_j'Z_j) = \overline{\sigma}_Z^2$ We have

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

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 $\sigma_1(M_j(\Lambda+M_j)^{-2}M_j) \leq 1$. Since we have $\sup_j \mathrm{tr}(M_j) \leq T\sup_j \sigma_1(M_j) < \infty$, we conclude that $\frac{1}{J^2}\sum_{j=1}^J \overline{\sigma}_Z^2\mathrm{tr}(M_j) \to 0$. 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 $\Lambda \in S_T^+$, 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.,

$$
|\widetilde{H}_{Z,J}(\widetilde{\Lambda}) - \widetilde{H}_{Z,J}(\widetilde{\Lambda}^\dagger)| \leq B_{H,J}^Z\|\widetilde{\Lambda} - \widetilde{\Lambda}^\dagger\| \tag{38}
$$

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

$$
\begin{aligned}
&|\widetilde{H}_{Z,J}(\widetilde{\Lambda}) - \widetilde{H}_{Z,J}(\widetilde{\Lambda}^\dagger)| \\
=&|\|\tfrac{1}{J}\textstyle\sum_{j=1}^J Z_j'\widetilde{A}_j^{-1}M_j(y_j - \theta_j)\| - \|\tfrac{1}{J}\sum_{j=1}^J Z_j'\widetilde{A}_j^{\dagger-1}M_j(y_j - \theta_j)\|| \\
\leq&\|\tfrac{1}{J}\textstyle\sum_{j=1}^J Z_j'(\widetilde{A}_j^{-1} - \widetilde{A}_j^{\dagger-1})M_j(y_j - \theta_j)\| \\
\leq&\tfrac{1}{J}\textstyle\sum_{j=1}^J \sigma_1(Z_j)\sigma_1(\widetilde{A}_j^{-1} - \widetilde{A}_j^{\dagger-1})\|M_j(y_j - \theta_j)\|,
\end{aligned}
\tag{39}
$$

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 4.3, we have shown that

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

Plugging this bound into (39), we have

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

Furthermore we have

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

and we have already shown $\frac{1}{J}\sum_{j=1}^J \|M_j(y_j - \theta_j)\| = O_p(1)$ in the proof of Theorem 4.3. This establishes (38), 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 $\left\{\sup_\Lambda H_{Z,J}(\Lambda)^2\right\}_{J\leq 1}$ is uniformly integrable. A sufficient condition for this is

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

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

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

where the first inequality follows from the triangle inequality and the definition of the operator norm, and the second inequality follows because $\sigma_1((\Lambda + M_j)^{-1}M_j) \leq 1$. Therefore, following (36), we have

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

Taking expectations, we obtain

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

for any $\delta \in [0, 2]$, and thus $\sup_J E \sup_\Lambda H_{Z,J}(\Lambda)^{2+\delta} < \infty$. This concludes the proof for $\sup_\Lambda H_{Z,J}(\Lambda) \xrightarrow{L_2} 0$.

Appendix D Proof of Theorem 4.5

We use the following decomposition

$$\begin{aligned}
&\left|\tfrac{1}{J}\textstyle\sum_{j=1}^{J}\left((B(\Lambda, M_{j,-T})'y_{j,-T} - y_{jT})^2 - M_{j,T}\right) - \tfrac{1}{J}\textstyle\sum_{j=1}^{J}(B(\Lambda, M_{j,-1})'y_{j,-1} - \theta_{j,T+1})^2\right| \\
\leq&\left|\tfrac{1}{J}\textstyle\sum_{j=1}^{J}\left((B(\Lambda, M_{j,-T})'y_{j,-T} - y_{jT})^2 - M_{j,T}\right) - \tfrac{1}{J}\textstyle\sum_{j=1}^{J}(B(\Lambda, M_{j,-T})'y_{j,-T} - \theta_{j,T})^2\right| \\
&+\left|\tfrac{1}{J}\textstyle\sum_{j=1}^{J}(B(\Lambda, M_{j,-T})'y_{j,-T} - \theta_{j,T})^2 - \tfrac{1}{J}\textstyle\sum_{j=1}^{J}(B(\Lambda, M_{j,-1})'y_{j,-1} - \theta_{j,T+1})^2\right|.
\end{aligned} \tag{40}$$

We have

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

We can further decompose the cross term as

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

Plugging this into the first term of right-hand side in (40), we have

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

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

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

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

$$\begin{aligned}\sup_{(\Lambda_{T,-T},\Lambda_T)\in\overline{\mathcal{L}}} \|B(\Lambda, M_{j,-T})\| \le& \sup_{(\Lambda_{T,-T},\Lambda_T)\in\overline{\mathcal{L}}} \|\Lambda_{T,-T}\|\, \sigma_1((\Lambda_{-T} + M_{j,-T})^{-1})\\ \le& K_{T,-T}\sigma^{-1}_{T-1}(M_{j,-T})\\ \le& K_{T,-T}\sigma^{-1}_{T}(M_j),\end{aligned} \tag{41}$$

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 $B(\Lambda, M_{j,-T})$ depends on Λ only through $(\Lambda_{T,-T}, \Lambda_{-T})$. In such case I write $B(\Lambda_{T,-T}, \Lambda_{-T}, M_{j,-T}) := B(\Lambda, M_{j,-T})$.

The fact that $(\mathrm{I})_J$ converges to zero in L_2, and thus in L_1, follows from

$$\begin{aligned}&E\left|\tfrac{1}{J}\textstyle\sum_{j=1}^{J}\left((y_{jT}-\theta_{jT})^2 - M_{j,T}\right)\right|^2\\ =&\tfrac{1}{J^2}\textstyle\sum_{j=1}^{J} E\left((y_{jT}-\theta_{jT})^2 - M_{j,T}\right)^2\\ \le&\tfrac{1}{J^2}\textstyle\sum_{j=1}^{J} E(y_{jT}-\theta_{jT})^4\\ \le&\tfrac{1}{J^2}\textstyle\sum_{j=1}^{J} 8(Ey^4_{jT}+\theta^4_{jT}),\end{aligned}$$

and the summand in the last line is uniformly bounded over j.

Similarly, $(\mathrm{III})_J \overset{L_2}{\to} 0$ can be easily shown by noting that

$$E\left|\tfrac{2}{J}\textstyle\sum_{j=1}^{J}\theta_{jT}(y_{jT}-\theta_{jT})\right|^2 \le \tfrac{4}{J^2}\textstyle\sum_{j=1}^{J}\theta^2_{jT}M_{jT},$$

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

To show that $\sup_{\overline{\mathcal{L}}}(\mathrm{II})_J \overset{L_1}{\to} 0$ and $\sup_{\overline{\mathcal{L}}}(\mathrm{IV})_J \overset{L_1}{\to} 0$, 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

$$\begin{aligned}
&E\left|\tfrac{2}{J}\textstyle\sum_{j=1}^{J} B(\Lambda, M_{j,-T})'((y_{j,-T}-\theta_{j,-T})(y_{jT}-\theta_{jT})-M_{j,T,-T})\right|^2\\
=&\tfrac{4}{J^2}\textstyle\sum_{j=1}^{J} \mathrm{tr}(B(\Lambda, M_{j,-T})B(\Lambda, M_{j,-T})'\mathrm{var}((y_{j,-T}-\theta_{j,-T})(y_{jT}-\theta_{jT}))\\
\leq&\tfrac{4}{J^2}\textstyle\sum_{j=1}^{J} \sigma_1(B(\Lambda, M_{j,-T})B(\Lambda, M_{j,-T})')\mathrm{tr}(\mathrm{var}((y_{j,-T}-\theta_{j,-T})(y_{jT}-\theta_{jT})))\\
\leq&\tfrac{4}{J^2}\textstyle\sum_{j=1}^{J} K_{T,-T}^2\sigma_T^{-2}(M_j)\mathrm{tr}(\mathrm{var}((y_{j,-T}-\theta_{j,-T})(y_{jT}-\theta_{jT}))),
\end{aligned}$$

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 (41). Moreover, we have

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

where the second inequality is by Cauchy-Schwarz. Note that the term in the last line is bounded uniformly over j, which establishes $(\mathrm{II})_{\mathrm{J}}\overset{L_2}{\to}0$. It remains to establish a Lipschitz condition. Write

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

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 and 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 $\Lambda_{T,-T},\widetilde{\Lambda}_{T,-T}\in\overline{\mathcal{L}}_{T,-T}$,

$$\begin{aligned}
&\|B(\Lambda_{T,-T},\Lambda_{-T},M_{j,-T})-B(\widetilde{\Lambda}_{T,-T},\Lambda_{-T},M_{j,-T})\|\\
\leq&\|(\Lambda_{T,-T}-\widetilde{\Lambda}_{T,-T})'(M_{j,-T}+\Lambda_{-T})^{-1}\|\\
\leq&\|\Lambda_{T,-T}-\widetilde{\Lambda}_{T,-T}\|\sigma_1((M_{j,-T}+\Lambda_{-T})^{-1})\\
\leq&\underline{\sigma}_M^{-1}\|\Lambda_{T,-T}-\widetilde{\Lambda}_{T,-T}\|.
\end{aligned} \tag{42}$$

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

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

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

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

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

$$\|B(\Lambda_{T,-T}, \Lambda_{-T}, M_{j,-T}) - B(\Lambda_{T,-T}, \widetilde{\Lambda}_{-T}, M_{j,-T})\| \leq \underline{\sigma}_M^{-2}\|\Lambda_{-T} - \widetilde{\Lambda}_{-T}\|. \quad (43)$$

Now, combining (42) and (43), we have for any $(\Lambda_{T,-T}, \Lambda_{-T}), (\widetilde{\Lambda}_{T,-T}, \widetilde{\Lambda}_{-T}) \in \overline{\mathcal{L}}$,

$$\begin{aligned} &\|B(\Lambda_{T,-T}, \Lambda_{-T}, M_{j,-T}) - B(\widetilde{\Lambda}_{T,-T}, \widetilde{\Lambda}_{-T}, M_{j,-T})\| \\ \leq &\|B(\Lambda_{T,-T}, \Lambda_{-T}, M_{j,-T}) - B(\widetilde{\Lambda}_{T,-T}, \Lambda_{-T}, M_{j,-T})\| \\ &+ \|B(\widetilde{\Lambda}_{T,-T}, \Lambda_{-T}, M_{j,-T}) - B(\widetilde{\Lambda}_{T,-T}, \widetilde{\Lambda}_{-T}, M_{j,-T})\| \quad (44) \\ \leq &\underline{\sigma}_M^{-1}\|\Lambda_{T,-T} - \widetilde{\Lambda}_{T,-T}\| + \underline{\sigma}_M^{-2}\|\Lambda_{-T} - \widetilde{\Lambda}_{-T}\| \\ \leq &(\underline{\sigma}_M^{-1} \vee \underline{\sigma}_M^{-2})(\|\Lambda_{T,-T} - \widetilde{\Lambda}_{T,-T}\| + \|\Lambda_{-T} - \widetilde{\Lambda}_{-T}\|). \end{aligned}$$

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

It follows that

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

and thus now it suffices to show

$$\tfrac{1}{J}\textstyle\sum_{j=1}^{J}\|(y_{j,-T} - \theta_{j,-T})(y_{jT} - \theta_{jT}) - M_{j,T,-T}\| = O_p(1). \quad (45)$$

We can bound the left-hand side by

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

I show that $(\mathrm{A})_J = o_p(1)$ and $(\mathrm{B})_J = O(1)$, from which (45) 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

$$\begin{aligned}
&\mathrm{var}(\|(y_{j,-T}-\theta_{j,-T})(y_{jT}-\theta_{jT})\|)\\
\leq&E\|(y_{j,-T}-\theta_{j,-T})(y_{jT}-\theta_{jT})\|^2\\
\leq&4\textstyle\sum_{t=1}^{T-1}E(|y_{jT}|^2+|\theta_{jT}|^2)(|y_{j,t}|^2+|\theta_{j,t}|^2)\\
\leq&4\textstyle\sum_{t=1}^{T-1}((E|y_{jT}|^4E|y_{j,t}|^4)^{1/2}+|\theta_{jT}|^2E|y_{j,t}|^2+|\theta_{j,t}|^2E|y_{jT}|^2+|\theta_{jT}|^2|\theta_{j,t}|^2),
\end{aligned} \tag{46}$$

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 $E\|(y_{j,-T}-\theta_{j,-T})(y_{jT}-\theta_{jT})\| \leq (E\|(y_{j,-T}-\theta_{j,-T})(y_{jT}-\theta_{jT})\|^2)^{1/2}$, 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,

$$\sup_j E(\sup_{\overline{\mathcal{L}}}|G_J(\Lambda_{T,-T},\Lambda_{-T})|)^2 < \infty. \tag{47}$$

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

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

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

Schwarz and the second inequality by (41). Hence, it follows that

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

where the second inequality follows from Cauchy-Schwarz. Since I have shown that the summand in the last line is bounded over j in (46), we have (47). We conclude that $\sup_{\overline{\mathcal{L}}}|G_J(\Lambda_{T,-T},\Lambda_{-T})|\overset{L_1}{\to}0$.

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

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

For pointwise convergence, note that

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

where the first inequality follows by Cauchy-Schwarz and the second inequality by (41). The expression in the last line is bounded over j, and thus $H_J(\Lambda_{T,-T},\Lambda_{-T},M_{j,-T})$ converges to zero in L_2. Now, I show that $H_J(\Lambda_{T,-T},\Lambda_{-T},M_{j,-T})$ satisfies a Lipschitz condition. We have

$$\begin{aligned}
&|H_J(\Lambda_{T,-T},\Lambda_{-T})-H_J(\widetilde{\Lambda}_{T,-T},\widetilde{\Lambda}_{-T})|\\
\le&\tfrac{2}{J}\textstyle\sum_{j=1}^{J}|(B(\Lambda,M_{j,-T})-B(\widetilde{\Lambda},M_{j,-T}))'\theta_{j,-T}(y_{j,T}-\theta_{j,T})|\\
\le&\tfrac{2}{J}\textstyle\sum_{j=1}^{J}\|B(\Lambda,M_{j,-T})-B(\widetilde{\Lambda},M_{j,-T})\|\|\theta_{j,-T}\||y_{jT}-\theta_{jT}|\\
\le&((\underline{\sigma}_M^{-1}\vee\underline{\sigma}_M^{-2})\tfrac{2}{J}\textstyle\sum_{j=1}^{J}\|\theta_{j,-T}\||y_{jT}-\theta_{jT}|,
\end{aligned}$$

where the second inequality is by Cauchy-Schwarz and the third inequality follows from (44). 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 (45). This implies $\sup_{\mathcal{L}}|H_J(\Lambda_{T,-T},\Lambda_{-T})| \overset{p}{\to} 0$. Again, following the same arguments we have used to show (47), we can easily show that $\sup_{\mathcal{L}}|H_J(\Lambda_{T,-T},\Lambda_{-T})|$ is uniformly integrable, from which it follows that $\sup_{\mathcal{L}}|H_J(\Lambda_{T,-T},\Lambda_{-T})| \overset{L_1}{\to} 0$. This concludes the proof for the first term of the right-hand side of (40) converging to zero in L_1.

For the second term of the right-hand side of (40), note that

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

Furthermore, we have

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

Hence, we have

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

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 $(\mathrm{I})_{\mathrm{J}}$, and a sketch for the other three terms. Note that the terms $(\mathrm{V})_J$ and $(\mathrm{VI})_J$ are nonrandom. Here, I provide conditions on the hypothetical "sampling distribution" on (θ_j, M_j) that ensure $(\mathrm{V})_J$ and $(\mathrm{VI})_J$ converge to zero uniformly over $\mathcal{L}$ for almost all sequences $\{(\theta_j, M_j)\}_{j=1}^{\infty}$. To this end, I consider the following assumption. Suppose that $\left\{((\theta_j', \theta_{j,T+1})', M_j)\right\}_{j=1}^{\infty}$ is a realization of a random sample drawn from a density $f_{(\theta',\theta_{T+1})',M}$. I assume that such marginal density is consistent with Assumption 4.1(ii) in the sense that the support of f_M, the marginal density of M_j, is a subset of $\left\{M \in S_T^+ : \sigma_T(M) > \underline{\sigma}_M\right\}$ for some positive number $\underline{\sigma}_M$. I also assume that $E_f\sigma_1(M_j) < \infty$ and $E_f\theta_{jt}^2 < \infty$ for $t = 1, \ldots, T+1$. Let $f_{\theta,M_{-T}}$ and $f_{(\theta'_{-1},\theta_{T+1})',M_{-1}}$ denote the marginal densities that correspond to $(\theta_j, M_{j,-T})$ and $((\theta'_{j,-1}, \theta_{j,T+1})', M_{j,-1})$, respectively.

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

I emphasize that this assumption does not imply that the observations are mean stationary or covariance stationary.

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

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

which has mean zero. Hence, if the expectation of the square of this term is bounded over j, then $(\mathrm{I})_J \overset{L_2}{\to} 0$. We have

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

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

$$\begin{aligned}
&E\mathrm{tr}(B(\Lambda, M_{j,-T})B(\Lambda, M_{j,-T})'((y_{j,-T} - \theta_{j,-T})(y_{j,-T} - \theta_{j,-T})' - M_{j,-T}))^2 \\
\le&E\|B(\Lambda, M_{j,-T})\|^4(\|y_{j,-T} - \theta_{j,-T}\|^2 + \mathrm{tr}(M_{j,-T}))^2 \\
\le&EK_{T,-T}^4\underline{\sigma}_M^{-4}(8\|y_{j,-T}\|^4 + 8\|\theta_{j,-T}\|^4 + \mathrm{tr}(M_{j,-T})^2 + 4(\|y_{j,-T}\|^2 + \|\theta_{j,-T}\|^2)\mathrm{tr}(M_{j,-T})),
\end{aligned}$$

where the term in the last line is bounded over j. This shows that $(\mathrm{I})_J \overset{L_2}{\to} 0$.

Now, to obtain a uniform convergence result, write

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

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 (41) and (44), gives

$$\begin{aligned}&\|B(\Lambda, M_{j,-T})B(\Lambda, M_{j,-T})' - B(\widetilde{\Lambda}, M_{j,-T})B(\widetilde{\Lambda}, M_{j,-T})'\|\\ \leq&2\underline{\sigma}_M^{-1}K_{T,-T}\|B(\Lambda, M_{j,-T}) - B(\widetilde{\Lambda}, M_{j,-T})\|\\ \leq&2\underline{\sigma}_M^{-1}(\underline{\sigma}_M^{-1}\vee\underline{\sigma}_M^{-2})K_{T,-T}\|(\Lambda_{T,-T}, \Lambda_{-T}) - (\widetilde{\Lambda}_{T,-T}, \widetilde{\Lambda}_{-T})\|,\end{aligned}\tag{48}$$

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

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

Observe that

$$\begin{aligned}&|G_{I,J}(\Lambda_{T,-T}, \Lambda_{-T}) - G_{I,J}(\widetilde{\Lambda}_{T,-T}, \widetilde{\Lambda}_{-T})|\\ \leq&|\tfrac{1}{J}\textstyle\sum_{j=1}^{J}\mathrm{tr}((B^2(\Lambda, M_{j,-T}) - B^2(\widetilde{\Lambda}, M_{j,-T}))'(y_{j,-T}-\theta_{j,-T})(y_{j,-T}-\theta_{j,-T})')|\\ &+|\tfrac{1}{J}\textstyle\sum_{j=1}^{J}\mathrm{tr}((B^2(\Lambda, M_{j,-T}) - B^2(\widetilde{\Lambda}, M_{j,-T}))'M_{j,-T})|\\ \leq&\tfrac{1}{J}\textstyle\sum_{j=1}^{J}\sigma_1(B^2(\Lambda, M_{j,-T}) - B^2(\widetilde{\Lambda}, M_{j,-T}))\mathrm{tr}((y_{j,-T}-\theta_{j,-T})(y_{j,-T}-\theta_{j,-T})')\\ &+\tfrac{1}{J}\textstyle\sum_{j=1}^{J}\sigma_1(B^2(\Lambda, M_{j,-T}) - B^2(\widetilde{\Lambda}, M_{j,-T}))\mathrm{tr}(M_{j,-T}),\end{aligned}$$

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

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

where $B_J = \frac{1}{J}\sum_{j=1}^{J}\mathrm{tr}((y_{j,-T}-\theta_{j,-T})(y_{j,-T}-\theta_{j,-T})' + M_{j,-T})$, 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

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

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 $\sup_j E\sup_{\overline{\mathcal{L}}}|G_{I,J}(\Lambda_{T,-T},\Lambda_{-T})|^2 < \infty$. 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 $(\mathrm{II})_J$, note that

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

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

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

which is very similar to the summand of $(\mathrm{I})_J$. The same steps used there go through without any added difficulty. The second term is even simpler, and extremely similar to $(\mathrm{IV})_J$ above in the decomposition of the first term on the right-hand side of (40). The same lines of argument used to establish convergence of such term can be used here to show the desired convergence result.

Now, it remains to show that $(\mathrm{V})_J$ and $(\mathrm{VI})_J$ converges to zero uniformly over $\overline{\mathcal{L}}$, almost surely, under the assumption that such sequence is a random sample from the density $f_{(\theta',\theta_{T+1})',M}$. Hence, here I treat the sequence

$\{((\theta_j', \theta_{j,T+1})', M_j)\}_{j=1}^{\infty}$ as random and aim to show a almost sure result.

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

where in the inequality I use Assumption D.1 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

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

Note that since $E_f \sigma_1(M_j)$ exists, we have

$$
E_f B(\Lambda, M_{j,-T})' M_{j,-T} B(\Lambda, M_{j,-T}) \leq K_{T,-T}^{-2} \underline{\sigma}_M^{-2} E_f \sigma_1(M_j) < \infty.
$$

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

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

where the first inequality follows from multiple applications of the triangle inequality, and the second inequality follows with $B_J = \frac{1}{J}\sum_{j=1}^{J}(\mathrm{tr}(M_{j,-T}) +$

$E\mathrm{tr}(M_{j,-T})$) from von Neumann's trace inequality and (48). 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 $B_J \overset{a.s.}{\to} 2E_f\mathrm{tr}(M_{j,-t})$. 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 $(\mathrm{VI})_\mathrm{J}$, the triangle inequality gives

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

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

$$\begin{aligned}&G_{\mathrm{VI},J}(\Lambda_{T,-T},\Lambda_{-T})\\ =&\frac{1}{J}\sum_{j=1}^{J}\left((B(\Lambda,M_{j,-T})'\theta_{j,-T}-\theta_{j,T})^2-E_f(B(\Lambda,M_{j,-T})'\theta_{j,-T}-\theta_{j,T})^2\right).\end{aligned}$$

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

$$\begin{aligned}&E_f(B(\Lambda,M_{j,-T})'\theta_{j,-T}-\theta_{j,T})^2\\ \leq&2E_f\mathrm{tr}(B^2(\Lambda,M_{j,-T})\theta_{j,-T}\theta_{j,-T}')+E_f\theta_{jT}^2\\ \leq&2K_{T,-T}^2\underline{\sigma}_M^{-2}\textstyle\sum_{t=1}^{T-1}E_f\theta_{jt}^2+E_f\theta_{jT}^2<\infty,\end{aligned}$$

where the second inequality follows because

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

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

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

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

and thus, by (44) and (48),

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

Likewise, we have

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

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

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

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