Optimal Shrinkage Estimation of Fixed Effects in Linear Panel Data Models

Soonwoo Kwon

Brown University

Oct 29th, 2021

Fixed Effects in Linear Panel Data Models

- Linear panel data models are one of the most widely used econometric models.
- Frequently, fixed effects are added to allow for unobserved heterogeneity.
- Usual role is to control for unobserved heterogeneity.
- Fixed effects themselves are empirically relevant in many settings.
  - teacher effects in student achievement: Rockoff (2004), Chetty et al. (2014a)
  - neighborhood effects in future economic outcome: Chetty and Hendren (2018)
  - employer effects in wage determination: Abowd et al. (1999), Card et al. (2013)
- Rising interest due to better data (e.g., “matched” data).

Many Effects but Not Many Observations for Each Effect

- The number of fixed effects to be estimated is large.
  - $> 10^3$ teachers; > 2800 counties
- The sample available for each fixed effect is not necessarily large.
  - A single teacher can teach only so many students.
  - Technically, remains finite even as sample size $\to \infty$.
- Simply using the least squares estimator gives a long vector of noisy estimates.
  - "long" due to the first point, and
  - "noisy" due to the second.

What This Paper Does

- I propose an optimal estimator for the (full vector of) fixed effects.
  - Obtained by “shrinking” the least squares estimator.
- Obtains the best possible mean squared error within a class of estimators.
  - This class nests the estimators used in the literature.
- This optimality (“within a class”) does NOT require
  1. distributional assumptions on the true fixed effects,
  2. distributional assumptions on the idiosyncratic error terms, or
  3. independence between fixed effects and cell sizes.
- The fixed effects are allowed to vary with time, and to be serially correlated.
  - “Shrinkage” takes into account this serial correlation.
  - An optimal forecast method is also provided.

Quick Remark: Not Just About Fixed Effects

- Suppose we're interested in "individual (or group) effects" $\{\theta_j\}_{j=1}^J$ with J large.
  - School effects: Angrist et al. (2017)
  - Hospital quality: Hull (2020)
  - Insurance quality: Abaluck et al. (2020)

- Such effects do not have to come from a linear panel data model.
  - Dynamic/nonlinear panel data models.
  - "Grouped effects": Bonhomme and Manresa (2015)

- The methodology applies if we have an initial estimator $\widehat{\theta}_j$ such that $E\widehat{\theta}_j \approx \theta_j$.

- Fixed effects in linear panel models are the simplest possible example.
  - θ_j: fixed effect for j
  - $\widehat{\theta}_j$: least squares estimator of θ_j

- Now, back to fixed effects!

Running Example: Teacher Value-Added

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

- teacher $j = 1, \ldots, J$; student $i = 1, \ldots, n_j$
- s_{ij}: test **s**core
- X_{ij}: vector of student characteristics
- ε_{ij} : idiosyncratic error with known variance σ_ε^2
- α_j: fixed effect at teacher-year level
- $\widehat{\beta}$: "within" estimator of β.

The aim is to estimate $\{\alpha_j\}_{j=1}^J$.

Why? Provides a performance measure for teachers; widely used in education policy.

The Least Squares Estimator

- The least squares estimator (obtained by adding teacher dummies) is

$$\widehat{\alpha}_j := \frac{1}{n_j} \sum_{i=1}^{n_j} (y_{ij} - X_{ij}' \widehat{\beta}).$$

- Unbiased with variance (approximately) $\sigma_\varepsilon^2 / n_j$.
- Why not just use $\widehat{\alpha}_j$?

The Normal Means Model and Stein's Phenomenon

- Stein's phenomenon tells us we can do (much) better.
- Consider the following "normal means model,"

$$y_j \overset{indep}{\sim} N(\theta_j, \sigma_y^2)$$

  for $j = 1, \ldots, J$ with known σ_y^2.
- Stein's phenomenon: when $J \geq 3$, y is "inadmissible." (Stein, 1956)
  - James-Stein estimator: $\widehat{b}y$ where $\widehat{b}$ is determined by y and σ_y^2.
  - Typically, $\widehat{b} \in (0, 1)$ and thus the term "shrinkage."
- The least squares estimator approximately fits into this framework:

$$\widehat{\alpha}_j \overset{indep}{\sim} N(\alpha_j, \sigma_\varepsilon^2 / n_j).$$

Empirical Bayes Interpretation of the James-Stein Estimator

- Note the heteroskedasticity in $\widehat{\alpha}_j \overset{indep}{\sim} N(\alpha_j, \sigma_\varepsilon^2/n_j)$.
- James-Stein is under homoskedasticity, so not directly applicable.
- However, there is an EB interpretation of the James-Stein estimator.
- Consider the following hierarchical model,

$$y_j|\theta_j \overset{indep}{\sim} N(\theta_j, \sigma_y^2), \quad \theta_j \overset{i.i.d.}{\sim} N(0, \sigma_\theta^2).$$

- The optimal estimator is given as $\mathbf{E}[\theta_j|y_j] = \frac{\sigma_\theta^2}{\sigma_\theta^2+\sigma_y^2} y_j$.
- Since σ_θ^2 is unknown, one plugs in an estimate of σ_θ^2.
  - This estimation uses the marginal distribution $y_j \overset{indep}{\sim} N(0, \sigma_\theta^2 + \sigma_y^2)$.
  - Plugging in an unbiased estimate gives James-Stein.

Common Practice: Empirical Bayes (EB)

- On top of $\widehat{\alpha}_j|\alpha_j \sim N(\alpha_j, \sigma_\varepsilon^2/n_j)$, further assume

$$\alpha_j \overset{i.i.d.}{\sim} N(0, \sigma_\alpha^2).$$

- The Empirical Bayes estimator under this setting is given as

$$\mathbf{E}[\alpha_j|\widehat{\alpha}_j] = \frac{\sigma_\alpha^2}{\sigma_\alpha^2 + \sigma_\varepsilon^2/n_j}\widehat{\alpha}_j,$$

with an estimator $\widehat{\sigma}_\alpha^2$ in place of the unknown σ_α^2.

- Almost all papers use either this or the least squares estimator.

Limitations of the Empirical Bayes Estimator

- Risk properties are sensitive to "Empirical Bayes assumptions":
  1. Normality of the true fixed effect: $\alpha_j \sim N(0, \sigma_\alpha^2)$
  2. Normality of the least squares estimator: $\widehat{\alpha}_j | \alpha_j \sim N(\alpha_j, \sigma_\varepsilon^2 / n_j)$
  3. Independence between mean and variance (implicit but important!): "$\alpha_j \perp n_j$"

- Why? σ_α^2 is estimated using these assumptions.

- If $n_j = n_0$ for all j, EB is still "optimal."
  - ...but this is never the case!

Proposed Method: Minimizing a Risk Estimate

- Restrict the class of estimators to those defined by the "conditional expectation"

$$\mathbf{E}[\alpha_j | \widehat{\alpha}_j] = \frac{\sigma_\alpha^2}{\sigma_\alpha^2 + \sigma_\varepsilon^2 / n_j} \widehat{\alpha}_j,$$

- Choose σ_α^2 by directly minimizing the mean squared error.
  - But, we don't know the mean squared error! (it depends on α_j)
- Minimize an estimate of the mean squared error instead.
  - Li (1986), Xie et al. (2012)
- The resulting estimator obtains the best possible mean squared error.
- This optimality does not require the Empirical Bayes assumptions.
  - Essentially, bounded fourth moments of the least squares estimator is enough.

Does it Work in Practice?

- Simulations results are very encouraging.
- Large reduction in mean squared error when EB assumptions are violated.
- Loses little ($< 6\%$) even when such assumptions are met.
  - Robustness comes at a negligible cost
- Empirical exercise shows that it makes a meaningful difference as well.
  - Has real impact: releases different teachers
  - In a good way: releases "worse" teachers
- Implementation is easy with the provided R package.
  - `FEShR` (**F**ixed **E**ffects **Sh**rinkage estimation with **R**)

Outline

1. Fixed effects and normal means problem
2. URE estimators
3. Optimality results
4. Simulation study
5. Empirical application

Linear Panel Data Model

$$s_{ijt} = X_{ijt}'\beta + \alpha_{jt} + \varepsilon_{ijt},$$

- $j = 1, \dots, J$; $t = 1, \dots, T$; $i = 1, \dots, n_{jt}$
- $X_{ijt} \in \mathbf{R}^{d_X}$: observed covariates
- α_{jt}: fixed-effect of the pair (j, t).
- ε_{ijt} : idiosyncratic error independent across i and j.

Examples:

- Employee-employer matched data models:
  - employer j, year t, employee i
- Neighborhood effects:
  - county j, year t, resident i

From Fixed Effects to the Normal Means Model

- Let $\widehat{\beta}$ be a consistent estimator of β, as $J \to \infty$.
- For any W_{ijt}, define $\overline{W}_{jt} = \frac{1}{n_{jt}} \sum_{i=1}^{n_{jt}} W_{ijt}$.
- Write $\alpha_j := (\alpha_{j1}, \ldots, \alpha_{jT})'$.
- Least squares estimator $\widehat{\alpha}_{jt} := \bar{s}_{jt} - \overline{X}_{jt}' \widehat{\beta}$ satisfies

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

  approximately, under $\bar{\varepsilon}_j \sim N(0, \Sigma_j)$.
- Σ_j is not necessarily diagonal, but assumed to be known.
- Heteroskedasticity rises even when ε_{ijt} is homoskedastic due to different cell sizes.
- The problem is approximately equivalent to esimating $\{\theta_j\}_{j=1}^J$ under

$$y_j \overset{indep}{\sim} N(\theta_j, \Sigma_j).$$

Multivariate Normal Means Model

- I now focus on the multivariate normal means problem

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

  where $y_j, \theta_j \in \mathbf{R}^T$ and Σ_j is a known $T \times T$ positive definite matrix.
- Consider a second level model,

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

  where the μ and $\Lambda \in \mathcal{L}$ are unknown hyperparameters.
- $\mathcal{L} \subset S_T^+$ reflects the prior knowledge of the variance structure of θ_j.

Class of Shrinkage Estimators

- The posterior mean of θ_j under this normal-normal hierarchical model is

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

- This is the class of estimators I consider.
  - All optimality results are within such class.
  - Includes conventional estimators.
  - Remains to "tune" μ and Λ (in an optimal way!)
- "Shrinks" the least squares estimator y_j towards μ.
- Now, let's forget about all the distributional assumptions and only assume

$$y_j \overset{\text{indep}}{\sim} (\theta_j, \Sigma_j).$$

Example: No Serial Correlation

- $\mathrm{diag}(d_1, \ldots, d_T) :=$ diagonal matrix with tth diagonal entry d_t.
- Take $\mu = 0$, $\Lambda = \mathrm{diag}(\lambda, \ldots, \lambda)$ and $\Sigma_j = \mathrm{diag}(\sigma_{j1}^2, \ldots, \sigma_{jT}^2)$.
- Then, we have
$$\widehat{\theta}_{jt}(\mu, \Lambda) = \frac{\lambda}{\lambda + \sigma_{jt}^2} y_{jt}.$$
- This is exactly what we had in the univariate case.

Example: $T = 2$

- Consider the case where $\mu = 0$,

$$\Sigma_j = \begin{pmatrix} \sigma_{j1}^2 & 0 \\ 0 & \sigma_{j2}^2 \end{pmatrix} \text{ and } \Lambda = \begin{pmatrix} \lambda_1^2 & \lambda_1\lambda_2\rho \\ \lambda_1\lambda_2\rho & \lambda_2^2 \end{pmatrix}.$$

- Write $y_j = (y_{j1}, y_{j2})'$ and $\theta_j = (\theta_{j1}, \theta_{j2})'$.

- The estimator for θ_{j1} is

$$\underbrace{\frac{\lambda_1^2(\lambda_2^2 + \sigma_{j2}^2) - \lambda_1^2\lambda_2^2\rho^2}{(\lambda_1^2 + \sigma_{j1}^2)(\lambda_2^2 + \sigma_{j2}^2) - \lambda_1^2\lambda_2^2\rho^2}}_{\text{Positive and decreases in } |\rho|} y_{j1} + \underbrace{\frac{\lambda_1\lambda_2\rho\sigma_{j1}^2}{(\lambda_1^2 + \sigma_{j1}^2)(\lambda_2^2 + \sigma_{j2}^2) - \lambda_1^2\lambda_2^2\rho^2}}_{\text{Absolute value increases in } |\rho|} y_{j2}.$$

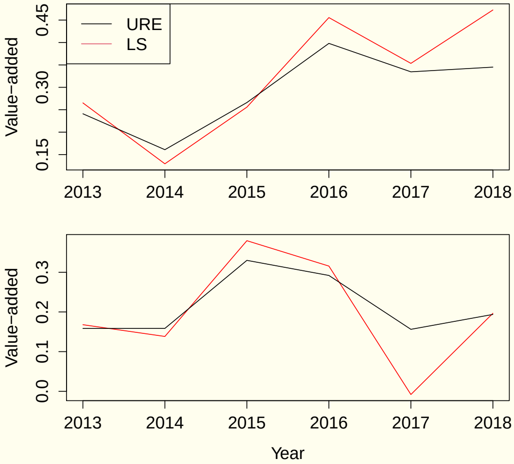
Example: Perfect Correlation

- Take $\mu = 0$, $\Lambda = \lambda \mathbf{1}_T \mathbf{1}_T'$ and $\Sigma_j = \sigma^2 \mathrm{diag}(1/n_{j1}, \ldots, 1/n_{jT})$
- This corresponds to the case where the fixed effects are time-invariant.
- Denote the teacher-level sample size by $n_j = \sum_{t=1}^{T} n_{jt}$.
- The estimator for θ_{jt} is given as

$$\frac{\lambda}{\sigma^2/n_j + \lambda} \left(\frac{1}{n_j} \sum_{t=1}^{T} n_{jt} y_{jt} \right).$$

- Note that $\frac{1}{n_j} \sum_{t=1}^{T} n_{jt} y_{jt}$ is the teacher-level least squares estimator.
  - Nests the conventional estimator used under the time-invariant case.

An Illustration: Shrinkage Pattern for Sample Teachers



What Does the Shrinkage Matrix Do?

- Let UDU' denote the spectral decomposition of $\Sigma_j^{-1/2}\Lambda\Sigma_j^{-1/2}$, which is the signal-to-noise ratio matrix.
- We can show

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

- "standardize - rotate - shrink - rotate back - destandardize"
- Λ is involved in both the degree and direction of shrinkage.
  - For the univariate case, we had $\frac{\lambda}{\lambda+\sigma_j^2}y_j$

Performance Criteria: Compound Mean Sqaured Error

- It remains to tune (μ, Λ) in

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

- Write $\theta = (\theta_1', \ldots, \theta_J')'$ and likewise for $\widehat{\theta}$.

- The risk function I use is the compound MSE,

$$\begin{aligned} R(\widehat{\theta}, \theta) =& \tfrac{1}{J} E(\widehat{\theta} - \theta)'(\widehat{\theta} - \theta) \\ =& \frac{1}{J} \sum_{j=1}^{J} E(\widehat{\theta}_j - \theta_j)'(\widehat{\theta}_j - \theta_j). \end{aligned}$$

- θ is treated as fixed.

Tuning the Hyperparameters

- Ideally, one would choose the hyperparemeters to minimize the true loss,

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

- This is infeasible because it requires knowledge of true θ.
- I choose the hyperparameters by minimizing a risk estimate instead.
  - I call the resulting estimators as URE estimators.
- **EB approach**: "estimate" (μ,Λ) using the marginal distribution,

$$y_j \overset{indep}{\sim} N(\mu, \Sigma_j + \Lambda).$$

- Let $(\widehat{\mu}^{\mathrm{EB}}, \widehat{\Lambda}^{\mathrm{EB}})$ denote (μ,Λ) that maximizes this marginal likelihood.

The Risk Estimate

- Define $\textbf{URE}(\mu, \Lambda) = \frac{1}{J}\sum_{j=1}^{J} \textbf{URE}_j(\mu, \Lambda)$ with

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

- This is by Stein's unbiased risk estimate (SURE).
  - Unbiased risk estimates for estimators of the form $y_j + g(y_j)$ where y_j is normal.
- y_j need not be normal because the estimator is linear in y_j.

- **Want to show**: minimizing $\mathrm{URE}(\mu, \Lambda)$ is as good as minimizing $\ell(\widehat{\theta}(\mu, \Lambda), \theta)$.
- **What were we trying to do?**: tune (μ, Λ) to estimate θ!

Obtaining the Oracle Risk

- For simplicity, let's consider $\mu = 0$, so Λ is the only tuning parameter.
- Let $\widehat{\theta}^{\mathrm{URE}} := \widehat{\theta}(0, \widehat{\Lambda}^{\mathrm{URE}})$ and $\widetilde{\theta}^{oracle} := \widehat{\theta}(0, \widetilde{\Lambda}^{oracle})$, where

$$\widehat{\Lambda}^{\mathrm{URE}} = \arg\min_{\Lambda} \mathrm{URE}(0, \Lambda) \text{ and } \widetilde{\Lambda}^{oracle} := \arg\min_{\Lambda} \ell(\widehat{\theta}(0, \Lambda), \theta).$$

- $R(\widetilde{\theta}^{oracle}, \theta)$: "oracle risk"
- We want to show that the risk of $\widehat{\theta}^{\mathrm{URE}}$ obtains the oracle risk.
- For this to be true, $\mathrm{URE}(0, \Lambda)$ has to be a good estimate of the loss.

Obtaining the Oracle Risk - Uniform Loss Estimation

- For now, assume the following key condition holds

$$E\left[\sup_{\Lambda\in\mathcal{S}_T^+}\left|\textbf{URE}(\Lambda)-\ell(\widehat{\theta}(\Lambda),\theta)\right|\right]\to 0.$$

- Note that $\textbf{URE}(\widehat{\Lambda}^{\mathrm{URE}})\leq\textbf{URE}(\widetilde{\Lambda}^{oracle})$ by definition, so that

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

- Taking expectations and then $\limsup_{J\to\infty}$, we have

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

Obtaining the Oracle Risk

- Hence, $\widehat{\theta}^{\mathrm{URE}}$ is asymptotically as good as any estimator taking the form,

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

  which includes, for example, EB estimators such as $\widehat{\theta}(\widehat{\Lambda}^{\mathrm{EBMLE}})$.

- The least squares estimator, y, does not belong to this class of estimators.

- However, since $\widehat{\theta}(\Lambda) \to y$ as "$\Lambda \to \infty$", a simple approximation argument shows that $\widehat{\theta}^{\mathrm{URE}}$ cannot do worse than y.

Uniform Convergence of $\mathrm{URE}(\Lambda)$ - Assumptions

Therefore, the aim is to establish

$$E\left[\sup_{\Lambda \in \mathcal{S}_T^+} \left| \mathbf{URE}(\Lambda) - \ell(\widehat{\theta}(\Lambda), \theta) \right| \right] \to 0.$$

Consider the following assumption.

Assumption 1 (Boundedness)

(i) $\sup_j E\|y_j\|^4 < \infty$ *and* (ii) $0 < \inf_j \sigma_T(\Sigma_j)$.

(i) $\approx$ "bounded fourth moments"
(ii) $\approx$ "bounded cell size"

Remark. Both conditions are stronger than necessary.

Uniform Convergence of $\mathrm{URE}(\Lambda)$ - Result

Theorem 1 (Uniform convergence of $\mathrm{URE}(\Lambda)$)

Under Assumption 1,

$$E\left[\sup_{\Lambda \in \mathcal{S}_T^+} \left| \mathbf{URE}(\Lambda) - \ell(\widehat{\theta}(\Lambda), \theta) \right| \right] \to 0.$$

- The optimality requires only Assumption 1.
- The URE approach gives us some guard against "misspecification."
- The proof technique differs from related papers.
  - Main reason: the shrinkage occurs after rotating the data.

Uniform Convergence of $\mathrm{URE}(\Lambda)$ - Sketch of Proof

Some algebra shows

$$\begin{aligned}&\mathbf{URE}(\mu,\Lambda)-\ell(\theta,\widehat{\theta}(\mu,\Lambda))\\=&\frac{1}{J}\sum_{j=1}^{J}(y_j'y_j-\theta_j'\theta_j-\mathrm{tr}(\Sigma_j))-\frac{2}{J}\sum_{j=1}^{J}\mathrm{tr}(\Lambda(\Lambda+\Sigma_j)^{-1}(y_jy_j'-\theta_jy_j'-\Sigma_j)).\end{aligned}$$

It suffices to show uniform L_1 convergence of the absolute values of these two terms.

The first term is immediate by Chebyshev's inequality.

The second term can be shown to $\to 0$ via a uniform LLN argument.

- Here, the parameter space S_T^+ is not totally bounded.
- Reparametrize as $\widetilde{\Lambda}=(\underline{\sigma}_\Sigma I_T+\Lambda)^{-1}$, and establish a Lipschitz condition with respect to $\widetilde{\Lambda}$, where $\underline{\sigma}_\Sigma := \inf_j \sigma_T(\Sigma_j)$.
- Here, it is crucial that $\underline{\sigma}_\Sigma > 0$.

Useful Variation: Covariates

- Suppose there are (j, t)-level covariates, supposedly related to the fixed effects.
  - These variables cannot be included in the original regression.
- $Z_{jt} = (Z_{jt1}, \ldots, Z_{jtK})' \in \mathbf{R}^K$: vector of such covariates, and $Z_j = (Z_{j1}, \ldots, Z_{jT})'$.
- $\{(y_j, Z_j)\}_{j=1}^J$: an independent sample with the Z_j's being identically distributed.
- To incorporate such information, postulate a second level model,

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

- Under this second level model, the posterior mean of θ_j is given as

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

Covariates - Assumptions

Assumption 2 (Covariates)

The covariates are bounded, exogenous, and of full rank.

- Misspecification "doesn't matter"
  - Somewhat analogous to obtaining better R^2 with more regressors.
  - Nonparametric specifications can give smaller mean squared error.
- Assumptions 1 & 2 ensure that the oracle risk is obtained within the class

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

- A nonparametric version is also possible.

Forecasting θ_{T+1}

- Another succinct summary of the time trajectory of the effects.
- The forecasting problem is of independent interest.
- The main idea is to consider "forecasting" θ_T, and then extrapolating to $T+1$.
  - Tune Λ by minimizing the "URE" corresponding to $E[\theta_T | y_{-T}]$
  - Use this Λ to forecast θ_{T+1} using y_{-1}
  - A stationarity condition is key in the extrapolation step.

Simulation Results

I focus on experimenting the performance of $\widehat{\theta}(\widehat{\mu}^{\mathrm{URE}}, \widehat{\Lambda}^{\mathrm{URE}})$ with $T = 4$.

Four main takeaways: URE estimators

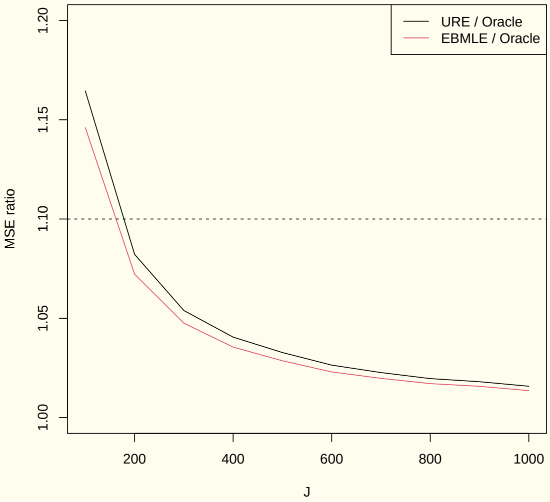
1. perform as well as the EBMLE under EB assumptions,
2. can be much better than the EBMLE when EB assumptions are violated
3. obtain the oracle risk reasonably quickly, and
   - For $T = 4$, the risk gets within 10% of the oracle when $J = 500$
4. dominates the least squares estimators by a significant margin.

Normal-Normal

- $\theta_j \overset{i.i.d.}{\sim} N(0, I_T)$
- $\Sigma_j \sim$ Wishart, centered at

$$\begin{pmatrix} 1 & .75 & .5 & .25 \\ & 1 & .75 & .5 \\ & & 1 & .75 \\ & & & 1 \end{pmatrix}$$

- $y_j \overset{indep}{\sim} N(\theta_j, \Sigma_j)$

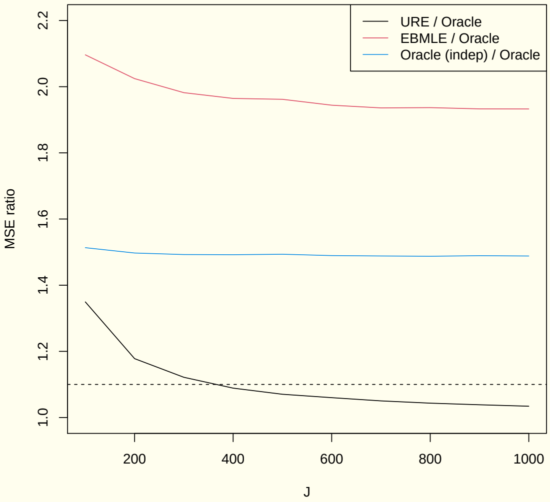


Normal-Normal with Group Structure

Similar to normal-normal, but one group has

- standard dev twice as large
- a smaller mean
  - direction does not matter

Mean vector is serially correlated.

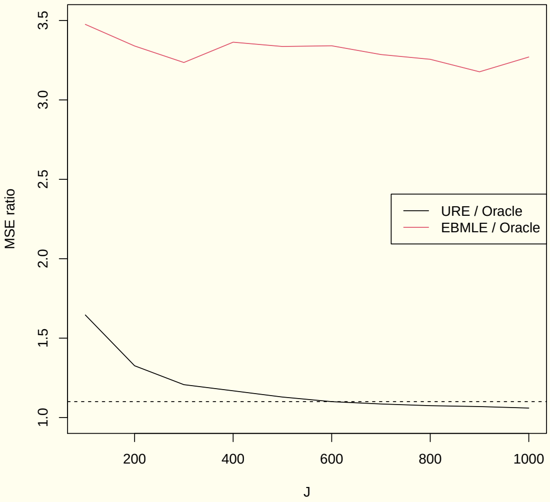


Conditional Heteroskedasticity

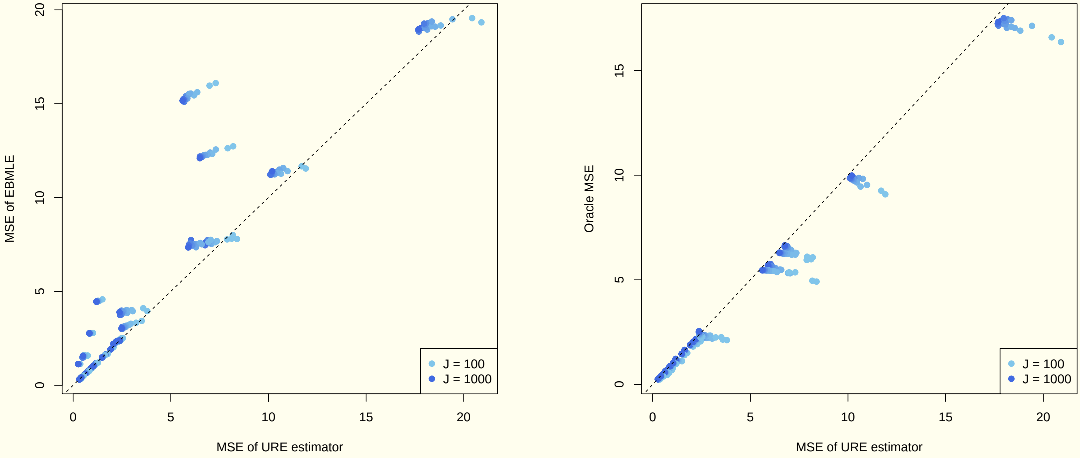
- $y_j \overset{indep}{\sim} N(\theta_j, \Sigma_j)$.
- $X_{jt} \in \mathbf{R}^2$ drawn from a uniform distribution.
- $\theta_{jt} = X'_{jt}\beta + \mathrm{Unif}[0, .3]$
- $\Sigma_j = D_j \Sigma D_j$ with

$$D_j = diag(X'_{j1}\gamma, \ldots, X'_{jT}\gamma)$$

and $\Sigma =$ a correlation matrix.



All Scenarios: URE vs EBMLE & URE vs Oracle



Empirical Application: Teacher Value-Added

- Administrative data on all public schools of NYC from 12/13 to 18/19 on
  - Student biographical data and test score (state-wide ELA test)
  - Student-teacher linkage
- Restrict attention to 4th and 5th grade students
  - Easier to match each student with single teacher.
  - ELA tests are for students in grades 3-8.
- $T = 6$ and $J = 1185$.
  - 12/13 data is used only to get test scores from previous year
  - Focus on teachers that are present in all six years
- Around 170k student-year observations
  - Average # of students per teacher (across all 6 years) ≈ 147.
  - Standard dev of # of students per teacher ≈ 55.

Teacher Value-Added: Model Specification

$$s_{ijt} = X'_{ijt}\beta + \alpha_{jt} + \varepsilon_{ijt},$$

- s_{ijt}: standardized ELA test score
- X_{ijt} includes student characteristics standard in the literature:
  - previous year's ELA test score
  - gender, ethnicity
  - special education status (SWD)
  - english language learner (ELL) status
  - eligibility for free/reduced price lunch (FL)
- ε_{ijt} : idiosyncratic error, i.i.d across i, j, t, with variance σ^2.

Parameter Estimates

	ELA
last year's score	0.629***
Male	−0.069***
Asian	0.133
Black	−0.028
Hispanic	0.002
Multi-Racial	0.078
Native American	0.039
White	0.062
ELL	−0.180***
SWD	−0.263***
FL	−0.046***

URE Estimate

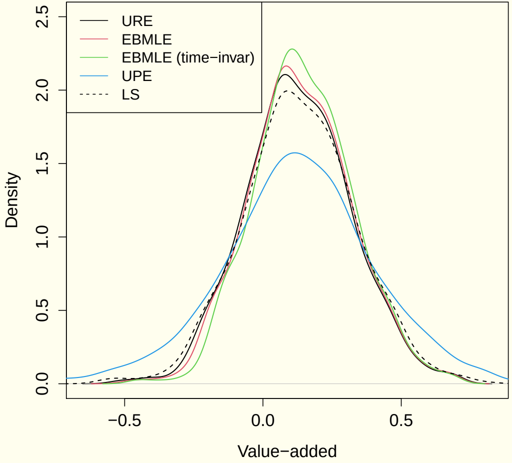
- Least sqaures estimators show huge variation over time
  - Average of standard deviation across time $\approx .21$
  - Standard deviation of estimates for teacher level fixed effects $\approx .20$.

- The optimal $\widehat{\Lambda}^{\mathrm{URE}}$ I obtain is

$$
\begin{pmatrix}
1 & 0.623 & 0.416 & 0.434 & 0.275 & 0.328 \\
 & 1 & 0.456 & 0.493 & 0.344 & 0.406 \\
 & & 1 & 0.387 & 0.499 & 0.563 \\
 & & & 1 & 0.323 & 0.321 \\
 & & & & 1 & 0.569 \\
 & & & & & 1
\end{pmatrix}
$$

- Computation takes about three minutes without any parallelization.
  - Main computation burden comes from repeated inversion of $T \times T$ matrices.

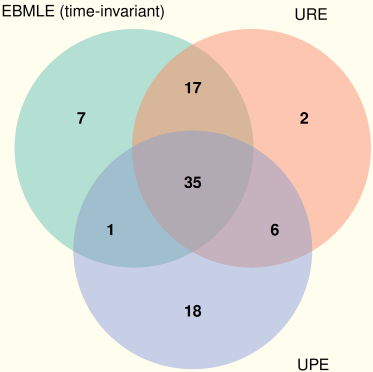
Comparison with the Conventional Estimates



Policy Exercise: Releasing the Bottom 5% Teachers

- Popular policy exercise (or thought experiment) in the literature.
  - Hanushek (2011), Chetty et al. (2014b), Gilraine et al. (2020).
- Release the bottom 5% of the teachers according to the value-added estimates.
  - Related: "retention policy" focuses on the top 5%
- A policy context where forecasts are arguably more relevant.
- Does the choice of estimator make a difference?
- Keep the last year as the "out-of-sample" observations.
  - Treat the least squares estimator for last years as true value-added

Composition of Released Teachers



- **EBMLE (time-invariant)**: "conventional"
  URE: time average of proposed estimator
  UPE: optimal forecasts
- Releases significantly different teachers.
- Similar findings for the top 5%.

This Change is in the "Correct Direction"

- Calculate the average value-added of the released teachers
  - Use the out-of-sample observations
- Average value-added of released teachers:

"conventional" $\rightarrow$ URE $\rightarrow$ optimal forecasts

-.22	**-.25**	**-.26**

- Again, similar findings for the top 5%.

Conclusion

- I propose an estimator for the fixed effects in a linear panel data model, that is optimal within a class of shrinkage estimators.
- Main idea: restrict the class of estimators by using a normal-normal hierarchical model, and choose the tuning parameters by minimizing a risk estimate.
- Not limited to fixed effects in linear panel data models!