Which Value-Added Estimator Works Best and When?

Elizabeth Santorella

August 28, 2017

Value-added estimators have been extensively used to study teachers and other groups. These estimators describe how dispersed teachers (or others) are in their effects on an outcome: for example, variation in teacher quality contributes to about 1% of the variance in student test scores. Value-added modeling is also used by school districts to rank teachers and make firing decisions. Although a large volume of research has investigated whether and when the identification assumptions of value-added models hold (Rothstein, 2009, 2010; Koedel and Betts, 2011; Chetty *et al.*, 2014; Rothstein, 2017). the statistical properties of these estimators are less studied, especially in finite samples. For example, standard errors and hypothesis tests are often unavailable, and parameter estimates can be badly biased even when identified.

A common use of value-added modeling is to measure what portion of variance in outcomes is due to variation in teacher quality. This number is of interest because if teachers vary little in their quality, then attempts to hire and retain better teachers may have little effect on student achievement. Estimates vary: Kane and Staiger (2008), who experimentally validated their estimates, albeit with large standard errors, found that the standard deviation in teacher quality in Los Angeles was 10% of a standard deviation in test scores, while Buddin (2011) measured 27%. [1]

Individual teacher value-added scores are also used as right-hand-side variables. For example, regressing test scores on a past teacher's estimated value-added shows how quickly teacher effects on scores “fade out", and regressing long-term outcomes like income on estimated value-added, as in Chetty *et al.* (2014), translates variation in teacher effects on test scores into variation in teacher effects on income. When value-added scores are estimated with shrinkage, they can be right-hand side variables free of attenuation bias.

Value-added modeling is popular in school districts. In contrast with education researchers, who care about aggregate quantities, districts want to accurately evaluate individual teachers. Value-added modeling has real-world consequences: Teacher value-added scores make up 20% to 25% of a teacher's total evaluation score in New York State, 35% in Tennessee, and at least 50% in Florida. Scores are sometimes released to the public, as in New York City and Los Angeles (Santos (2012), Felch *et al.*). Publication of value-added scores has been especially heavily criticized, with researchers arguing that value-added

[1] What constitutes a large amount of dispersion in teacher quality is contentious. If the standard deviation of teacher quality is only 10% of the standard deviation of test scores, teachers contribute only 1% of variance. On the other hand, since teachers affect many students and have persistent effects on students' income and educational attainment, the value of improving a teacher's effectiveness by one standard deviation could be quite high (Chetty *et al.*, 2014).

scores are imprecisely measured, that uncertainty is not adequately disclosed, and that scores are very sensitive to methodological choices (Briggs and Domingue, 2011).

I think of a value-added model as one with the following properties: Each observation i corresponds to some individual $j(i)$. When forming the best linear predictor of an outcome given an indicator for $j(i)$ and covariates, the coefficient on the indicator is $\mu_{j(i)}$. These $\mu_{j(i)}$ have a causal interpretation: If a student's teacher j is experimentally replaced with teacher j', the student's outcome increases in expectation by $\mu_{j'} - \mu_j$. The μ are drawn identically and independently from the same distribution, $\mu_j \overset{iid}{\sim} F$, and the distribution F itself is of interest. The individual components of μ may also be of interest. High-dimensional covariates and few observations for each teacher are common complications, making it inadvisable to simply estimate μ through least squares. This setup lends itself naturally to an Empirical Bayes estimation procedure, which first estimates the distribution F and then forms a "posterior" estimate of μ. Empirical Bayes methods have been used to study teachers and in various other settings. For example, Ellison and Swanson (2016) study how much of the variation between schools in the fraction of high math achievers that are female is due to variation in schools. Feng and Jaravel (2016) study variation in patent examiners' propensity to grant patents and which patents benefit from being assigned to a lenient patent collector. Furthermore, many studies that do not rely explicitly on the teacher value-added literature share this literature's interest in estimating both individual effects and their distribution. For example, there is a wide literature in labor economics on estimating individual and firm effects (i.e. Abowd *et al.* (1999)). Recently, Barnett *et al.* (2017) studied "the extent to which individual physicians vary in opioid prescribing and the implications of that variation." Others have studied hospital effects on C-sections (Kozhimannil *et al.*, 2013) and variation in judge sentencing tendencies (Green and Winik, 2010).

In this project, I survey several popular value-added estimation procedures and study their statistical properties. I discuss conditions under which models are identified, clarify whether estimators are consistent or unbiased, and derive standard errors. [2] I also develop a maximum (quasi-)likelihood estimator. I investigate the bias and precision of different estimators in Monte Carlo data and check whether estimators give similar answers in real data. My focus is on the portion of variance that is due to variation in teacher quality, but I also discuss individual-specific estimates, and ask whether different procedures give highly correlated estimates of teacher effects, and whether a procedure can reliably identify teachers in the bottom 2%.

For clarity, I often use terminology relating to teachers and classrooms since value-added modeling is most used for studying teachers. However, these results extend readily to different settings.

This paper proceeds as follows. In Section 1, I develop a toy model to motivate why policymakers may care about the variance of teacher effects. In Section 2, I recap the historical development of the value-added literature and the settings in which value-added estimators have been used. Section 3 describes several estimators whose properties I develop and compare. Section 4 discusses the behavior of several procedures in Monte Carlo Data, and Section 5 compares the performance of these estimators on two real data sets: teachers' effects on test scores in New York City, and bureaucrats' effects on project completions in India. Section 6 concludes with recommendations about which estimator to use.

[2] I would also like to verify whether tests based on these standard errors are appropriately sized.

1 Toy Model, Motivation

Why should policymakers care about the variance of teacher effects? This section lays out a toy model in which the variance of teacher effects is a sufficient statistic for policy decisions. In particular, there should be more investment in teacher training when teachers vary more in their quality.

Assume that student i's academic ability a_i is a linear combination of teacher quality and other inputs x. Student i has teacher j, and teacher j has quality μ_j. Teacher quality is drawn from some distribution with mean $\bar{\mu}$ and variance σ_μ^2. In other words,

$$a_i = \mu_{j(i)} + x$$
$$\mu_j \sim \left[\bar{\mu}, \sigma_\mu^2\right]$$

A high σ_μ^2 is taken as an indication that interventions that improve teacher effectiveness may be worthwhile; if, on the other hand, teachers do not vary greatly in effectiveness, improving teacher quality will be difficult. To motivate this conclusion, imagine that a policymaker can invest either in increasing x, at constant marginal cost $C(x) = x$, or in improving teacher quality, at increasing marginal cost $C(\bar{\mu}_j) = \frac{C\bar{\mu}_j^2}{2\sigma_\mu}$: marginal cost is increasing in quality, but decreasing in the variance of quality [3]. If the policymaker solves

$$\max_{\bar{\mu}} \bar{\mu} + x \quad \text{subject to} \quad C(x) + C(\bar{\mu}) \leq M,$$

they will set $\bar{\mu}^* = \sigma_\mu / k$, assuming an interior solution, and invest $\sigma_\mu / 2k$ in teacher quality. Investing in teacher quality is more valuable when teacher quality is more variable, because high variation in teacher quality is an indication that the returns to training are high.

Other models also lead to the conclusion that σ_μ^2 is important. For example, when σ_μ^2 is high, deselecting low-quality teachers and replacing them with average teachers is more beneficial, as is trying to recruit above-average teachers.

2 Literature Review

> The extensive investigation of the contribution of teachers to student achievement produces two generally accepted results. First, there is substantial variation in teacher quality as measured by the value added to achievement or future academic attainment or earnings. Second, variables often used to determine entry into the profession and salaries, including post-graduate schooling, experience, and licensing examination scores, appear to explain little of the variation in teacher quality so measured, with the exception of early experience (Hanushek and Rivkin, 2010).

[3] A variety of teacher quality production functions can motivate this structure. For example, suppose teachers randomly receive training $T_j \sim N(0, 1)$ at constant marginal cost c. and quality is generated by $\mu_j = \sigma_\mu T_j$. Then spending $c\Delta$ to increase training by Δ increases mean teacher quality by $\sigma_\mu \Delta$.

Table 1: *Estimates of the variance of teacher effects, $\hat{\sigma}^2_\mu$, and forecast bias adapted from Table 6 of Kane and Staiger (2008). "1 - forecast bias" is the coefficient from regressing experimental test scores on non-experimentally estimated value-added scores. 95% confidence intervals are in brackets.*

	Math	Math	Reading	Reading
Student FEs?	N	Y	N	Y
Var (μ_j)	0.219	0.101	0.175	0.084
1 - forecast bias	0.905	1.859	1.089	2.144
	[0.552, 1.258]	[0.938, 2.780]	[0.523, 1.655]	[0.899, 3.389]

The earliest work on teacher quality noted that teacher output appeared unrelated to observable teacher characteristics other than experience and perhaps teacher test scores, and sometimes argued that variation in teacher quality is not an important determinant of differences in educational outcomes (Hanushek and Rivkin, 2010, 2006) [4]. However, later work has focused on "outcome-based" measures of teacher quality, treating quality as a latent variable to be estimated, and found that teachers explain about 1% to 3% of the variance in student outcomes Hanushek and Rivkin (2012).

The identification requirements of value-added models that treat teacher quality as a latent variable make such models controversial. These models typically involve a sorting on observables requirement: Any association between student attributes and teacher identities must be captured by variables included in the model. This requirement is necessary both for estimating the fraction of variance in student outcomes that is due to variation in teacher quality, and for evaluating individual teachers. Sorting on observables could be violated if, for example, students assigned to better teachers have parents who push them to study hard. More subtly, imagine that all teachers are identical, but some teachers are consistently assigned high- or low-achieving students; if student achievement can't be predicted well by observables, then these teachers will appear to be the cause of their students' achievement, and teacher quality may appear to vary even when it does not. The validity of the sorting on observables requirement has been contested (Rothstein, 2010). However, in this paper I focus on issues that can arise even when identification requirements are obeyed.

Several studies have addressed whether value-added scores are "forecast unbiased": that is, whether a teacher with a value-added score of $\hat{\mu}$ causally raises test scores by $\hat{\mu}$, in expectation. Unbiased estimates of the variance of teacher effects, σ^2_μ, are necessary for forecast-unbiased value-added scores, since value-added scores are a product of a mean residual and a shrinkage factor based on $\hat{\sigma}^2_\mu$. The literature has typically interpreted forecast bias as a sign of insufficient controls for student-teacher sorting, but it can also reflect bias in $\hat{\sigma}^2_\mu$, an issue I consider in this paper. Randomized and quasi-experimental analyses have somewhat ameliorated concerns that sorting on unobservables biases estimates of the variance of teacher quality upwards. Previous studies have generally concluded that value-added scores are close to forecast-unbiased, after converging on sets of specifications that tend to work well (Jacob, 2005; Kane and Staiger, 2008; Rothstein, 2009; Chetty *et al.*, 2014).

However, experimentally validated estimates tend to be smaller than other estimates,

[4] Briggs and Domingue (2011) finds that teachers' educational backgrounds do predict teacher effects

and methods of checking for bias are controversial. The only truly randomized assessment of value-added modeling comes from Kane and Staiger (2008), who estimated estimated individual value-added scores for teachers in Los Angeles, randomly assigned students to teachers in the next year, and confirmed that the previous value-added scores were an unbiased predictor of future student achievement. The results of Kane and Staiger (2008), reproduced in Table 1, show that a teacher one standard deviation above average improves math scores by 0.219 standard deviations in models that don't include student fixed effects and by 0.101 in models that do, with analogous estimates of 0.175 and 0.084 for reading; their experimental results seem to indicate that the former estimate is nearly unbiased and the latter significantly understates teachers' contributions to test score variance. However, they are unable to rule out large degrees of bias. Estimates of about 0.1 are relatively small for this literature. For example, Buddin (2011) also analyzed data from Los Angeles – the same district studied by Kane and Staiger (2008) – to generate value-added scores that were published in the LA Times Felch *et al.* and found that a teacher one standard deviation above average improves math test scores by 0.27 standard deviations. That is, Buddin (2011) find that teachers account for 7% of the variance in math test scores in Los Angeles, while according to Kane and Staiger (2008) they account for only 1%. Lacking experimental data, Chetty *et al.* (2014) introduce the use of "teacher switching quasi-experiments": they argue that teachers switch schools for exogenous reasons and that after switching schools, teachers' value-added will not be correlated with the ability of their current students. The quasi-experiments indicate that forecast bias is quite small: the coefficient from regressing changes in test scores with changes in value-added (with controls) is approximately 0.97 and at least 0.9. Rothstein (2017) replicates the quasi-experiments in North Carolina, questions the randomness of teacher transfers. He finds similar results when using the same specifications as Chetty, Friedman, and Rockoff, but a forecast bias of about 10% when using test score gains instead of levels as the dependent variable; he argues that this is because high value-added teachers tend to move to improving schools. On the other hand, Chetty *et al.* (2017) argue that Rothstein's specifications can generate bias, and show through simulation that it is possible to find similar results without violating identification assumptions.

Despite uncertainty about how to test identification restrictions, most researchers agree that in large samples and with controls for past student test scores, value-added models can accurately estimate the variance in teacher quality. (Useful reviews are given by Koedel *et al.* (2015), Hanushek and Rivkin (2010), and Staiger and Rockoff (2010).) By contrast, using value-added models to assess individual teachers remains controversial Koedel *et al.* (2015). Briggs and Domingue (2011), for example, re-analyze data from Buddin (2011), whose results were published in the LA Times, and find that with richer controls, individual teachers' value-added scores shift dramatically. Staiger and Rockoff (2010) state that value-added scores have a reliability of 30% to 50%.

Statistically, the value-added literature has been influenced by the literature on Empirical Bayes, hierarchical linear models, and correlated random effects.

In summary, two well-studied areas are whether the identification requirements of value-added models are obeyed and how confidently these models can evaluate individual teachers. However, there has been little work on how well value-added models can evaluate individual teachers. There has also been relatively little work on how value-added procedures behave in finite samples and how to quantify uncertainty in structural parameters.

Table 2: *Comparison of estimators.*

	MLE	*Kane and Staiger*	*Mod-KS*	*Fessler and Kasy*
Consistent under baseline model	Y	N	Y	Y
Consistent under baseline + no sorting	Y	Y	Y	Y
Unbiased under baseline + no sorting	?	N	N	N
Closed-form solution	N	Y	Y	N
Closed-form standard errors on hyperparameters	Y	Y	Y	N

3 Estimators

In this section, I lay out a statistical model and discuss estimation of that model via maximum likelihood. I then discuss other value-added estimation procedures used in the literature. I treat the model in this section as a baseline and explain how it can be tweaked to match other authors' models, and under what circumstances different estimation procedures yield the same results. Observations are at the student level. Student i has classroom $c(i)$, $j(i)$, test score y_i, and covariates $\boldsymbol{x}_i$. [5]

Data is drawn from some distribution $\mathcal{D}$. I describe the model in terms of best linear predictors. The model's *parameters* are teacher effects and error terms, and *hyperparameters* are best linear predictor coefficients on covariates and variances of teacher effects and error terms. Asymptotics are as the number of teachers approaches infinity, so we will be able to consistently estimate hyperparameters but not parameters.

To begin defining best linear predictors, stack all of the data from teacher j, who has n_s students, into a vector $\boldsymbol{y}_j \in \mathcal{R}^{n_s}$, a matrix $\boldsymbol{x}_j \in \mathcal{R}^{n_s \times k}$, and mean covariates $\bar{\bar{x}}_j \in \mathcal{R}^k$. [6] $\boldsymbol{y}_j$ and $\boldsymbol{x}_j$ both have one row for each student. Also define a variable s_j that encapsulates the configuration of students to classrooms: For example, s_j tells how many students are in each classroom, and whether any two students are in the same classroom.

The best linear predictor of test scores given covariates and configuration is

$$E^*_{\mathcal{D}}\left[\boldsymbol{y}_j | \mu_j, \boldsymbol{x}_j, s_j\right] = \alpha + \mu_j + \boldsymbol{x}_j \boldsymbol{\beta}, \tag{1}$$

The teacher effect, μ_j, is teacher j's *value-added*, her causal effect on the outcome of interest. The best linear predictor of teacher effects given covariates is

$$E^*_{\mathcal{D}}\left[\mu_j | \bar{\bar{x}}_j, s_j\right] = \bar{\bar{x}}_j^T \boldsymbol{\lambda}. \tag{2}$$

[5] I use bolded letters (i.e. $\boldsymbol{x}$) to represent vectors, and bolded and italicized letters (i.e. $\boldsymbol{x}$) to represent matrices.

[6] Maximum likelihood estimation is greatly simplified if $\bar{x}_j$ is a precision-weighted mean, in a way that will be made clear.

$\boldsymbol{\lambda}$ is a vector governing the association of covariates with teacher quality. It could capture teacher-specific characteristics – for example, more experienced teachers are better – or reflect sorting – for example, teachers of honors classes may be better.

Combining Equations 1 and 2, $E^*_{\mathcal{D}}\left[\boldsymbol{y}_j|\boldsymbol{x}_j, \bar{\bar{x}}_j, s_j\right] = \alpha + \boldsymbol{x}_j\boldsymbol{\beta} + \bar{\bar{x}}_j^T\boldsymbol{\lambda}$. We can define errors $\tilde{\mu}_j$ and $\boldsymbol{\nu}_j$ with

$$\begin{aligned} \mu_j &\equiv \bar{\bar{x}}_j\boldsymbol{\lambda} + \tilde{\mu}_j, \quad \tilde{\mu}_j \perp \bar{\bar{x}}_j \\ \boldsymbol{y}_j &\equiv \boldsymbol{x}_j\boldsymbol{\beta} + \bar{\bar{x}}_j^T\boldsymbol{\lambda} + \boldsymbol{\nu}_j, \quad \boldsymbol{\nu}_j \perp \boldsymbol{x}_j, \bar{\bar{x}}_j \end{aligned} \tag{3}$$

In order to ascribe a casual interpretation to hyperparameter estimates, we need sorting on observables. First, we need that variation in teacher effects that cannot be captured by covariates must be orthogonal to non-teacher shocks to test scores: $\tilde{\mu}_j \perp \left(\boldsymbol{\nu}_j - \tilde{\mu}_j\right)$. Second, more subtly, we need unobservable shocks to test scores to be *independent* of *assignments* to teachers: $\left(\boldsymbol{\nu}_j - \tilde{\mu}_j\right) \perp\!\!\!\perp s_j|\boldsymbol{x}_j, \bar{\bar{x}}_j$. To see why this second restriction is necessary, imagine that all teachers are identical – $\mu_j = 0 \quad \forall j$ – but some teachers are consistently assigned students with high values of $\boldsymbol{\nu}_j$. In that case, some teachers will consistently appear to have students that over- or under-perform what would be expected from their covariates, making it appear that teachers vary in their quality when they actually do not.

In order to make this model estimable via maximum likelihood, we need several more assumptions. First, $\boldsymbol{\beta}$ must correspond to an unrestricted linear predictor. That is, define the best linear predictor $\boldsymbol{\pi}$, so that

$$E^*_{\mathcal{D}}\left[\boldsymbol{y}_j|I_n \otimes \text{vec}(\boldsymbol{x}_j), \bar{\bar{x}}_j\right] = \left(I_n \otimes \text{vec}(\boldsymbol{x}_j)\right)\boldsymbol{\pi} + \bar{\bar{x}}\boldsymbol{\lambda}.$$

We need that $\left(I_n \otimes \text{vec}(\boldsymbol{x}_j)\right)\boldsymbol{\pi} = \boldsymbol{x}_j\boldsymbol{\beta}$. Finally, let's put more structure on the covariance of errors. Define $E\left[\boldsymbol{\nu}_j\boldsymbol{\nu}_j^T\ s_j\right] \equiv \Sigma_j$. Denote hyperparameters $\eta = \left(\alpha, \boldsymbol{\beta}, \boldsymbol{\lambda}, \sigma^2_\mu, \sigma^2_\theta, \sigma^2_\varepsilon\right)$.

$$\begin{aligned} &\Sigma\left(\eta, s_j\right)_{i,i'} = \sigma^2_\mu + \sigma^2_\theta + \sigma^2_\varepsilon \quad \text{when } i = i' \\ &\Sigma\left(\eta, s_j\right)_{i,i'} = \sigma^2_\mu + \sigma^2_\theta \quad \text{when } i \neq i' \text{ but } i \text{ and } i' \text{ are in the same class} \\ &\Sigma\left(\eta, s_j\right)_{i,i'} = \sigma^2_\mu \quad \text{when } i \text{ and } i' \text{ are not in the same class} \end{aligned}$$

$\text{Var}\left(\mu_j\right) = \text{Var}\left(\bar{\bar{x}}_j^T\boldsymbol{\lambda}\right) + \sigma^2_\mu$ is the amount of variance in y contributed by teachers; when teacher effects have a large variance, teachers are an important determinant of y. When σ^2_μ is large, there are large differences in teacher quality that are not predictable from observables. When variance in $\bar{\bar{x}}\boldsymbol{\lambda}$ is large, there are large differences in teacher quality that are predictable by observables. σ^2_θ and σ^2_ϵ are the shares of variance from classroom-level shocks and individual-specific shocks.

3.1 Maximum (Quasi-)Likelihood

No model like the one above has, to my knowledge, been estimated via maximum likelihood, but rather with GMM-like "moment-matching" procedures, as discussed at length below.

To generate a likelihood function, we must assume a functional form for the distributions of $\boldsymbol{y}_j$ and μ_j. Appendix A shows that quasi-likelihood based on normality delivers consistent estimates of η, even when the true distribution $\mathcal{D}$ does not have normal disturbances. Consider the model

$$\boldsymbol{y}_j|\boldsymbol{x}_j, \bar{\bar{x}}_j, s_j \sim N\left(\boldsymbol{x}_j\boldsymbol{\beta} + \bar{\bar{x}}_j^T\boldsymbol{\lambda}, \Sigma\left(\theta, s_j\right)\right)$$

with the corresponding likelihood function $f\left(\boldsymbol{y}_j, \boldsymbol{x}_j, \bar{\bar{x}}_j, s_j; \theta\right)$. Appendix A proves that $\eta = \arg\max_\theta \mathbb{E}_{\mathcal{D}} \log f\left(\boldsymbol{y}_j|\boldsymbol{x}_j, \bar{\bar{x}}_j, s_j; \theta\right)$. Appendix **??** derives a relatively simple closed-form solution for the likelihood.

Although the maximum likelihood estimands have relatively simple formulas, they do require defining several more terms. Define classroom means $\bar{y}_c$ and $\bar{\boldsymbol{x}}_c$, and deviations from classroom means $\tilde{y}_i$ and $\tilde{x}_i$. Denote the precision of the mean classroom error

$$\frac{1}{h_c} \equiv \mathrm{Var}\left(\bar{y}_c - \bar{\boldsymbol{x}}_c\boldsymbol{\beta} - \mu_{j(c)}\right) = \sigma_\theta^2 + \sigma_\varepsilon^2/n_c. \tag{4}$$

Then use precision weights to create teacher-level means of y and x:

$$\bar{\bar{y}}_j \equiv \frac{\sum_{c:j(c)=j} h_c \bar{y}_c}{\sum_{c:j(c)=j} h_c} \qquad \bar{\bar{x}}_j \equiv \frac{\sum_{c:j(c)=j} h_c \bar{\boldsymbol{x}}_c}{\sum_{c:j(c)=j} h_c} \tag{5}$$

and define classroom deviations from teacher means $\tilde{\bar{y}}_c = \bar{y}_c - \bar{\bar{y}}_{j(c)}$ and $\tilde{\bar{x}}_c = \bar{\boldsymbol{x}}_c - \bar{\bar{x}}_{j(c)}$. One expression for the likelihood is

$$\begin{aligned} f\left(\boldsymbol{y}_j|\boldsymbol{x}_j, s_j; \alpha, \boldsymbol{\lambda}, \boldsymbol{\beta}, \sigma_\mu^2, \sigma_\theta^2, \sigma_\varepsilon^2\right) =& g\left(\alpha, \boldsymbol{\lambda}, \boldsymbol{\beta}, \sigma_\mu^2, \sigma_\theta^2, \sigma_\varepsilon^2\right) \\ & \int_\mu \phi\left(\mu - \bar{\bar{x}}_j^T\boldsymbol{\lambda}; \sigma_\mu^2\right) \phi\left(\bar{\bar{y}}_j - \bar{\bar{x}}_j^T\boldsymbol{\beta} - \alpha - \mu; \frac{1}{\sum_c h_c}\right) \mathrm{d}\mu \end{aligned} \tag{6}$$

The integral in Equation 6 has a Bayesian interpretation and yields an Empirical Bayes posterior: Teacher effects are drawn $\mu_j \sim N\left(\bar{\bar{x}}_j^T\boldsymbol{\lambda}, \sigma_\mu^2\right)$, and test scores are drawn $\bar{\bar{y}}_j \sim N\left(\mu_j + \bar{\bar{x}}_j^T\boldsymbol{\beta}, \frac{1}{\sum_c h_c}\right)$, so the Empirical Bayes posterior of teacher j's value-added is

$$\mu_j \sim N\left(\frac{\sigma_\mu^2}{\sigma_\mu^2 + 1/\sum h_c}\left(\bar{\bar{y}}_j - \bar{\bar{x}}_j^T\boldsymbol{\beta} - \alpha\right) + \frac{1/\sum h_c}{\sigma_\mu^2 + 1/\sum h_c}\bar{\bar{x}}_j^T\boldsymbol{\lambda}, \left(\frac{1}{\sigma_\mu^2} + \sum h_c\right)^{-1}\right) \tag{7}$$

Up to an additive constant, the log-likelihood of the data is

$$\begin{aligned} 2\mathrm{LL}\left(\sigma_\mu^2, \sigma_\theta^2, \sigma_\varepsilon^2, \boldsymbol{\beta}, \boldsymbol{\lambda}; \boldsymbol{y}_j, \boldsymbol{x}_j\right) \propto & -\left(N_{\text{students}} - N_{\text{classes}}\right)\log(\sigma_\varepsilon^2) + \sum_c \log h_c - \sum_j \log\left(\sum_{c:j(c)=j} h_c\right) \\ & - \sum_j \log\left(\sigma_\mu^2 + \frac{1}{\sum_{c:j(c)=j} h_c}\right) - \frac{1}{\sigma_\varepsilon^2}\sum_i \left(\tilde{y}_i - \tilde{x}_i\boldsymbol{\beta}\right)^2 \\ & - \sum_c h_c\left(\tilde{\bar{y}}_c - \tilde{\bar{x}}_c\boldsymbol{\beta}\right)^2 - \sum_j \frac{1}{\sigma_\mu^2 + \frac{1}{\sum_{c:j(c)=j} h_c}}\left(\bar{\bar{y}}_j - \bar{\bar{x}}_j^T\left(\boldsymbol{\beta} + \boldsymbol{\lambda}\right)\right)^2 \end{aligned} \tag{8}$$

This equation can easily be optimized numerically. The software package available at http://www.github.com/esantorella/tva iterates between estimating $\hat{\boldsymbol{\beta}}$, $\hat{\boldsymbol{\lambda}}$, and $\hat{\alpha}$, which have closed-form solutions in terms of other hyperparameters, and estimating σ_μ^2, σ_θ^2, and σ_ε^2 using L-BFGS.

The solutions for $\hat{\boldsymbol{\beta}}$, $\hat{\boldsymbol{\lambda}}$, and $\hat{\alpha}$ are intuitive. After concentrating out $\hat{\boldsymbol{\lambda}}$, $\hat{\boldsymbol{\beta}}$ attempts to jointly minimize students' deviations from the classroom mean and classrooms' deviations from the teacher mean:

$$\hat{\boldsymbol{\beta}} = \arg\min_b \frac{1}{\hat{\sigma}_\varepsilon^2} \sum_i \left(\tilde{y}_i - \tilde{x}_i^T b \right)^2 + \sum_j \sum_{c:j(c)=j} \hat{h}_c \left(\tilde{\bar{y}}_c - \tilde{\bar{x}}_c b \right)^2 \tag{9}$$

$\hat{\boldsymbol{\lambda}}$ and $\hat{\alpha}$ are given by weighted least squares, and minimize differences between teachers that can't be explained by differences within teachers:

$$\hat{\boldsymbol{\lambda}}, a = \arg\min_{\boldsymbol{\ell},a} \sum_j \left(\frac{1}{\sum_c \hat{h}_c} + \hat{\sigma}_\mu^2 \right)^{-1} \left(\bar{\bar{y}}_j - \bar{\bar{x}}_j^T \hat{\boldsymbol{\beta}} - \bar{\bar{x}}_j^T \boldsymbol{\ell} - \alpha \right)^2 \tag{10}$$

When there are no classroom-level shocks, $\hat{\boldsymbol{\beta}}$ and $\hat{\boldsymbol{\lambda}}$ coincide with the estimands from a correlated random effects framework. When $\hat{\sigma}_\theta^2 = 0$, precisions h_c are proportional to the number of students in the class, so each observation is given equal weight. Equation 9 collapses to

$$\hat{\boldsymbol{\beta}} = \arg\min_{\boldsymbol{b}} \sum_i \left(y_i - \bar{\bar{y}}_{j(i)} - \left(\boldsymbol{x}_i - \bar{\bar{x}}_{j(i)} \right)^T \boldsymbol{b} \right)^2,$$

and Equation 10 becomes

$$\hat{\boldsymbol{\beta}} + \hat{\boldsymbol{\lambda}} = \arg\min_{\boldsymbol{c}} \sum_j \left(\bar{\bar{y}}_j - \bar{\bar{x}}_j^T \boldsymbol{c} \right)^2.$$

These equations yield the same coefficients as running the regression (Chamberlain (1984), Chamberlain (1982))

$$y_i = x_i^T \boldsymbol{\beta} + \bar{\bar{x}}_{j(i)}^T \boldsymbol{\lambda} + \varepsilon_i.$$

3.1.1 Inference

Since this is MLE, asymptotic inference is (conceptually) easy. Appendix **??** gives the first derivative of the likelihood function. I would like to implement bootstrap inference as in Chamberlain (2013).

3.1.2 Variance of teacher effects

The quantity of interest is

$$\mathrm{Var}(\mu_j) = \mathrm{Var}\left(\bar{\bar{x}}_j^T \boldsymbol{\lambda} \right) + \sigma_\mu^2$$

An obvious estimator is the sample analog:

$$\begin{aligned}\widehat{\mathrm{Var}}(\mu|\hat{\boldsymbol{\lambda}}) &= \widehat{\mathrm{Var}}\left(\bar{\bar{x}}_j^T\hat{\boldsymbol{\lambda}}|\hat{\boldsymbol{\lambda}}\right) + \hat{\sigma}_\mu^2 \\ &= \frac{1}{n}\sum_j\left(\bar{\bar{x}}_j^T\hat{\boldsymbol{\lambda}}\right)^2 - \left(\frac{1}{n}\sum_j\bar{\bar{x}}_j^T\hat{\boldsymbol{\lambda}}\right)^2 + \hat{\sigma}_\mu^2\end{aligned}$$

However, the sample analog is biased upwards. $\mathbb{E}\left[\mathrm{Var}\left(\bar{\bar{x}}_j^T\hat{\boldsymbol{\lambda}}|\hat{\boldsymbol{\lambda}}\right)\right] > \mathbb{E}\left[\mathrm{Var}\left(\bar{\bar{x}}_j^T\boldsymbol{\lambda}\right)\right]$, for a clear reason: estimation error in $\hat{\boldsymbol{\lambda}}$ will tend to make this quantity larger. Imagine that $\boldsymbol{\lambda} = 0$: $\hat{\boldsymbol{\lambda}}$ will not be zero, so there will appear to be some correlation between teacher effects and covariates when there is not. Specifically, the sample analog is biased upwards by

$$\begin{aligned}\mathbb{E}\left[\mathrm{Var}\left(\bar{\bar{x}}^T\hat{\boldsymbol{\lambda}}|\hat{\boldsymbol{\lambda}}\right)\right] - \mathrm{Var}\left(\bar{\bar{x}}^T\boldsymbol{\lambda}\right) &= \mathrm{Var}\left(\bar{\bar{x}}^T\hat{\boldsymbol{\lambda}}\right) - \mathrm{Var}\left(\mathbb{E}\left[\bar{\bar{x}}^T\hat{\boldsymbol{\lambda}}|\hat{\boldsymbol{\lambda}}\right]\right) - \mathrm{Var}\left(\bar{\bar{x}}^T\boldsymbol{\lambda}\right) \\ &= \mathrm{Var}\left(\bar{\bar{x}}^T\hat{\boldsymbol{\lambda}}\right) - \mathrm{Var}\left(\mathbb{E}\left[\bar{\bar{x}}^T\right]\hat{\boldsymbol{\lambda}}\right) - \mathrm{Var}\left(\bar{\bar{x}}^T\boldsymbol{\lambda}\right) \\ &= \mathbb{E}\left[\mathrm{Var}\left(\bar{\bar{x}}^T\hat{\boldsymbol{\lambda}}|\bar{\bar{x}}\right)\right] + \mathrm{Var}\left(\mathbb{E}\left[\bar{\bar{x}}^T\hat{\boldsymbol{\lambda}}|\bar{\bar{x}}\right]\right) - \mathbb{E}\left[\bar{\bar{x}}\right]^T\mathrm{Cov}(\hat{\boldsymbol{\lambda}})\,\mathbb{E}\left[\bar{\bar{x}}\right] - \mathrm{Var}\left(\bar{\bar{x}}^T\boldsymbol{\lambda}\right) \\ &= \mathbb{E}\left[\bar{\bar{x}}^T\mathrm{Cov}\left(\hat{\boldsymbol{\lambda}}\right)\bar{\bar{x}}\right] - \mathbb{E}\left[\bar{\bar{x}}\right]^T\mathrm{Cov}(\hat{\boldsymbol{\lambda}})\,\mathbb{E}\left[\bar{\bar{x}}\right] \\ &= \mathbb{E}\left[\left(\bar{\bar{x}} - \mathbb{E}\,\bar{\bar{x}}\right)^T\mathrm{Cov}\left(\hat{\boldsymbol{\lambda}}\right)\left(\bar{\bar{x}} - \mathbb{E}\,\bar{\bar{x}}\right)\right].\end{aligned}$$

Therefore, a bias-corrected estimator of the variance of teacher effects is

$$\widehat{\mathrm{Var}}\left(\mu_j\right) = \frac{1}{n}\sum_j\left(\bar{\bar{x}}_j^T\boldsymbol{\lambda}\right)^2 - \left(\frac{1}{n}\sum_j\bar{\bar{x}}_j^T\boldsymbol{\lambda}\right)^2 - \frac{1}{n}\sum_j\left(\bar{\bar{x}}_j^T - \frac{1}{J}\sum_k\bar{\bar{x}}_k\right)^T\hat{\Sigma}\left(\bar{\bar{x}}_j - \frac{1}{J}\sum_k\bar{\bar{x}}_k\right) + \hat{\sigma}_\mu^2.$$

where $\hat{\Sigma}$ is the asymptotic variance of $\hat{\boldsymbol{\lambda}}$.

3.2 Empirical Bayes estimator from Kane and Staiger (2008)

Kane and Staiger (2008) develop a model that other value-added papers use as a baseline, such as Chetty *et al.* (2014). Guarino *et al.* (2014) and others note that this estimator is not consistent when teacher effects are correlated with covariates. As discussed in Section 2, Kane and Staiger (2008) experimentally validated value-added scores and did not reject the hypothesis that the scores were forecast-unbiased. However, their estimates also suggest that, when controlling for student fixed effects, too value-added scores actually understate the magnitude of teacher effects. Section 3.2.4 demonstrates that this estimator is asymptotically downward biased, and 3.3 discusses amending estimation to follow a fixed effects rather than random effects assumption.

The model used by Kane and Staiger (2008) is identical to that used in Section 3.1, with one exception: in assuming that teacher effects are uncorrelated with covariates, Kane and Staiger impose $\boldsymbol{\lambda} = 0$. Their model maintains the restriction that there is no sorting on unobservables.

3.2.1 Estimation

Kane and Staiger (2008)'s estimation procedure, like many other Empirical Bayes procedures and like the maximum likelihood procedure above, proceeds in two phases. First, we estimate the hyperparameters of the model: $\boldsymbol{\beta}$, σ_μ^2, σ_θ^2, and σ_ε^2. Then we estimate the parameters of the model, each teacher's value of μ_j, using the distribution described by the previously-estimated hyperparameters as a prior.

The first stage, estimation of hyperparameters, itself comprises two steps. First, we estimate $\hat{\boldsymbol{\beta}}$, then we use $\hat{\boldsymbol{\beta}}$ to generate residuals. Next, we use a "moment-matching" procedure to estimate the variances σ_μ^2, σ_θ^2, and σ_ϵ^2 based on variances and covariances of residuals. In more detail:

$\hat{\boldsymbol{\beta}}$ is estimated by regressing outcomes y_i on covariates $\boldsymbol{x}_i$. This gives a consistent and unbiased estimate of $\boldsymbol{\beta}$ if and only if teacher effects are uncorrelated with covariates; otherwise, this estimate will suffer from omitted variable bias:

$$\begin{aligned} \hat{\boldsymbol{\beta}} &= \boldsymbol{\beta} + \left(\sum \boldsymbol{x}_i \boldsymbol{x}_i^T\right)^{-1} \sum \boldsymbol{x}_i \left(\mu_{j(i)} + \nu_i\right) \\ \mathbb{E}\left[\hat{\boldsymbol{\beta}}\right] &= \boldsymbol{\beta} + \mathbb{E}\left[\boldsymbol{x}_i \boldsymbol{x}_i^T\right]^{-1} \mathbb{E}\left[\boldsymbol{x}_i \mu_{j(i)}\right] \end{aligned}$$

(The estimator presented in the next section explores estimating $\boldsymbol{\beta}$ in the presence of teacher fixed effects, which corrects this omitted variable bias.)

In order to estimate σ_μ^2, use the following procedure. Let $C(j)$ denote the set of classes taught by teacher j. σ_μ^2 is the average product of mean residuals in pairs of classes taught by the same teacher: [7]

$$\hat{\sigma}_\mu^2 = \frac{2}{\sum_j |C(j)| \left(|C(j)| - 1\right)} \sum_j \sum_{c, c' \in C(j)} \left(\bar{y}_c - \bar{\boldsymbol{x}}_c^T \hat{\boldsymbol{\beta}}\right) \left(\bar{y}_{c'} - \bar{\boldsymbol{x}}_{c'}^T \hat{\boldsymbol{\beta}}\right) \tag{11}$$

To estimate $\hat{\sigma}_\theta^2$ and $\hat{\sigma}_\varepsilon^2$, we use similar "moment-matching" ideas. Since ε is responsible for within-classroom variation in $\tilde{y}$, σ_ε^2 is the mean variance of $\tilde{y}_i$ within a classroom:

$$\hat{\sigma}_\varepsilon^2 = \frac{1}{N_{\text{students}} - N_{\text{classes}}} \sum_j \left(\tilde{y}_i - \tilde{\boldsymbol{x}}_i^T \hat{\boldsymbol{\beta}}\right)^2$$

$\hat{\sigma}_\theta^2$ is chosen to explain the variance in y_i that is not explained by μ, ε, or $\hat{\boldsymbol{\beta}}$:

$$\hat{\sigma}_\theta^2 = \widehat{\text{Var}}(y_i - \boldsymbol{x}_i^T \hat{\boldsymbol{\beta}}) - \hat{\sigma}_\mu^2 - \hat{\sigma}_\varepsilon^2.$$

3.2.2 Inference

We can reformulate this "moment-matching" procedure as the solution to a set of moment functions. After setting up a moment function, we can estimate the asymptotic distribution of the hyperparameters either through the Bayesian Bootstrap, as in Chamberlain (2013), or through the Generalized Method of Moments.

[7] Kane and Staiger use a different version of this procedure, using only sequential pairs of classrooms.

Denote the parameters $\eta = (\boldsymbol{\beta}, \sigma^2_\mu, \sigma^2_\varepsilon, \sigma^2_\theta)$. Letting $S(j)$ be the set of students with teacher j, $S(j) = \{i : j(i) = j\}$, the moment function, which is at the teacher level, is

$$g_j(\eta) = \begin{pmatrix} \sum_{i \in S(j)} x_i(y_i - x_i^T \hat{\boldsymbol{\beta}}) \\ \sum_{c,c' \in C(j)} \left(\bar{y}_c - \bar{\boldsymbol{x}}_c^T \hat{\boldsymbol{\beta}}\right) \left(\bar{y}_{c'} - \bar{\boldsymbol{x}}_{c'}^T \hat{\boldsymbol{\beta}}\right) - \hat{\sigma}^2_\mu \left(\frac{|C(j)|(|C(j)|-1)}{2}\right) \\ \sum_{i \in S(j)} \left(\tilde{y}_i - \tilde{\boldsymbol{x}}_i^T \hat{\boldsymbol{\beta}}\right)^2 - \hat{\sigma}^2_\varepsilon \left(|S(j)| - |C(j)|\right) \\ \sum_{i \in S(j)} \left(y_i - \boldsymbol{x}_i^T \hat{\boldsymbol{\beta}}\right)^2 - |S(j)| \left(\hat{\sigma}^2_\mu - \hat{\sigma}^2_\theta - \hat{\sigma}^2_\varepsilon\right) \end{pmatrix}$$

To take the n^{th} Bayesian Bootstrap draw, draw weights $\omega^n \in \mathbb{R}^{\text{Nteachers}}$ according to $\omega^n \sim \text{Dirichlet}\,(1, 1, \ldots, 1)$ (Rubin, 1981). Bootstrap draws of hyperparameters become

$$\begin{aligned} \hat{\beta}^n &= \left(\sum_i \omega^n_{j(i)} x_i x_i^T\right)^{-1} \sum_i \omega^n_{j(i)} x_i y_i \\ \hat{\sigma}_\mu^{2(n)} &= \frac{2}{\sum_j \omega^n_j |C(j)| \left(|C(j)| - 1\right)} \sum_j \sum_{c,c' \in C(j)} \omega^n_j \left(\bar{y}_c - \bar{\boldsymbol{x}}_c^T \hat{\boldsymbol{\beta}}\right) \left(\bar{y}_{c'} - \bar{\boldsymbol{x}}_{c'}^T \hat{\boldsymbol{\beta}}^n\right) \\ \hat{\sigma}_\varepsilon^{2(n)} &= \frac{1}{\sum_j \omega^n_j \left(|S(j)| - |C(j)|\right)} \sum_i \omega^n_{j(i)} \left(\tilde{y}_i - \tilde{x}_i^T \hat{\beta}^n\right)^2 \\ \hat{\sigma}_\theta^{2(n)} &= \frac{1}{\sum_j \omega^n_j |S(j)|} \sum_i \omega^n_{j(i)} \left(y_i - x_i^T \hat{\beta}^n\right)^2 - \hat{\sigma}_\mu^{2(n)} - \hat{\sigma}_\varepsilon^{2(n)} \end{aligned}$$

Estimating standard errors through GMM or bootstrapping is computationally expensive, and GMM standard errors are only asymptotically valid. In the case where each teacher teaches the same number of classes, a quicker, finite-sample-valid lower bound on the variance of $\hat{\sigma}^2_\mu$ is available:

Lemma 3.1.

$$\operatorname{Var}\left(\hat{\sigma}^2_\mu\right) \geq \frac{1}{N_{\text{teachers}}} \mathbb{E}\left[\operatorname{Var}\left(\frac{1}{|C(j)| \left(|C(j)| - 1\right)} \sum_{c,c' \in C(j)} \left(\bar{y}_c - \bar{\boldsymbol{x}}_c^T \hat{\beta}\right) \left(\bar{y}_{c'} - \bar{\boldsymbol{x}}_{c'}^T \hat{\beta}\right)\right)\right] \tag{12}$$

Proof.

$$\begin{aligned} \operatorname{Var}\left(\hat{\sigma}^2_\mu\right) &= \mathbb{E}\left[\operatorname{Var}\left(\hat{\sigma}^2_\mu | \hat{\boldsymbol{\beta}}\right)\right] + \operatorname{Var} \mathbb{E}\left[\hat{\sigma}^2_\mu | \hat{\boldsymbol{\beta}}\right] \\ &\geq \mathbb{E}\left[\operatorname{Var}\left(\hat{\sigma}^2_\mu | \hat{\boldsymbol{\beta}}\right)\right] \\ &= \mathbb{E}\left[\operatorname{Var}\left(\frac{1}{N_{\text{teachers}}} \sum_j \frac{1}{|C(j)| \left(|C(j)| - 1\right)} \sum_{c,c' \in C(j)} \left(\bar{y}_c - \bar{\boldsymbol{x}}_c^T \hat{\beta}\right) \left(\bar{y}_{c'} - \bar{\boldsymbol{x}}_{c'}^T \hat{\beta}\right) \middle| \hat{\boldsymbol{\beta}}\right)\right] \end{aligned}$$

Conditional on $\hat{\boldsymbol{\beta}}$, $(\bar{y}_{A(j)} - \bar{x}_{A(j)}^T \hat{\boldsymbol{\beta}})(\bar{y}_{B(j)} - \bar{x}_{B(j)}^T \hat{\boldsymbol{\beta}})$ are independently and identically distributed across teachers, so

$$
\begin{aligned}
&\mathbb{E}\left[\operatorname{Var}\left(\frac{1}{N_{\text{teachers}}}\sum_j \frac{1}{|C(j)|\left(|C(j)|-1\right)}\sum_{c,c'\in C(j)}\left(\bar{y}_c-\bar{\boldsymbol{x}}_c^T\hat{\beta}\right)\left(\bar{y}_{c'}-\bar{\boldsymbol{x}}_{c'}^T\hat{\beta}\right)\middle|\,\hat{\boldsymbol{\beta}}\right)\right]\\
&=\frac{1}{N_{\text{teachers}}}\mathbb{E}\left[\operatorname{Var}\left(\frac{1}{|C(j)|\left(|C(j)|-1\right)}\sum_{c,c'\in C(j)}\left(\bar{y}_c-\bar{\boldsymbol{x}}_c^T\hat{\beta}\right)\left(\bar{y}_{c'}-\bar{\boldsymbol{x}}_{c'}^T\hat{\beta}\right)|\hat{\boldsymbol{\beta}}\right)\right]
\end{aligned}
$$

□

Equation 12 can be replaced by its sample analog.

3.2.3 Individual Teacher Effects

While $\bar{y}_j-\bar{x}_j^T\hat{\beta}$ is an unbiased estimate of μ_j, Kane and Staiger use shrinkage to produce a best linear predictor of μ_j. First, generate the precision $h_c=\operatorname{Var}(\bar{y}_c-\bar{\boldsymbol{x}}_c^T\boldsymbol{\beta})^{-1}$ of each mean classroom residual; these are the same precisions used for maximum likelihood in Equation 4. Then construct a precision-weighted mean using h_c and multiply it by shrinkage factor ρ_j use linear shrinkage ρ_j and precisions h_c to generate a mean squared error-minimizing estimate of μ_j:

$$
\begin{aligned}
\hat{\mu}_j &= \rho_j\frac{\sum_c h_{c:j(c)=j}\left(\bar{y}_c-\bar{\boldsymbol{x}}_c^T\boldsymbol{\beta}\right)}{\sum_c h_c}\\
\rho_j &= \arg\min_{\rho}\mathbb{E}\left[\left(\rho\frac{\sum_c h_c\bar{y}_c}{\sum_c h_c}-\mu_j\right)^2\right]=\frac{\hat{\sigma}_\mu^2}{\hat{\sigma}_\mu^2+\frac{1}{\sum_c h_c}}
\end{aligned}
\tag{13}
$$

Kane and Staiger note that when μ, θ, and ε are normally distributed, Equation 13 has a Bayesian interpretation. The estimated variances $\hat{\sigma}_\mu^2$, $\hat{\sigma}_\theta^2$, and $\hat{\sigma}_\varepsilon^2$ are treated as a prior and observed test scores as data to create Empirical Bayes maximum a posteriori estimates of teacher effects, which shrink mean residuals towards zero.

Equation 13 equals Equation 7, from maximum likelihood, when $\hat{\boldsymbol{\lambda}}=0$: Conditional on hyperparameter estimates, both procedures deliver the same estimated individual teacher effects. However, even with the imposition of $\boldsymbol{\lambda}=0$, the procedures will generally not estimate the same hyperparameters. When estimating $\hat{\beta}$, the Kane and Staiger procedure implicitly gives each observation equal weight, while maximum likelihood uses precision weighting to put relatively more weight on smaller classes. Maximum likelihood uses precision weights to give relatively Even with the imposition of $\boldsymbol{\lambda}=0$, the estimates of $\hat{\beta}$ will not generally be the same unless σ_θ^2.

3.2.4 Inconsistency and bias under misspecification

Consistency and bias of $\hat{\sigma}_\mu^2$

As discussed extensively in Guarino *et al.* (2014) and mentioned in Chetty *et al.* (2014), the Kane and Staiger estimator will only be valid if there is no correlation between observable student characteristics and teacher value-added. Their work demonstrates that $\hat{\boldsymbol{\beta}}$ will be biased when estimated in a regression that omits teacher fixed effects; this project demonstrates that omitted variable bias in $\hat{\boldsymbol{\beta}}$ leads to an asymptotic *negative* bias in $\hat{\sigma}_\mu^2$. Intuitively, variation in teacher effects that is correlated with student characteristics is incorrectly attributed to the student characteristics.

Consistency of $\hat{\boldsymbol{\beta}}$, independence of errors between teachers, and finite second moments are jointly sufficient for consistency of $\hat{\sigma}_\mu^2$. Unbiasedness of $\hat{\boldsymbol{\beta}}$ is necessary for unbiasedness of $\hat{\sigma}_\mu^2$.

Since we know the probability limit of $\hat{\boldsymbol{\beta}}$, and $\hat{\sigma}_\mu^2$ is a smooth function of $\hat{\boldsymbol{\beta}}$, we also know the probability limit of $\hat{\sigma}_\mu^2$. Note that Equation 11 gives

$$\begin{aligned}\hat{\sigma}_\mu^2 = \frac{1}{N_{\text{teachers}}} \sum_j \frac{2}{|C(j)|\,|C(j)-1|} \sum_{c,c' \in C(j)} & \left(\mu_j + \theta_c + \bar{\varepsilon}_c + (\boldsymbol{\beta} - \hat{\boldsymbol{\beta}})^T \bar{\boldsymbol{x}}_c\right) \\ & \left(\mu_j + \theta_{c'} + \bar{\varepsilon}_{c'} + (\boldsymbol{\beta} - \hat{\boldsymbol{\beta}})^T \bar{\boldsymbol{x}}_{c'}\right).\end{aligned}$$

Also,

$$\begin{aligned}\boldsymbol{\beta} \rightarrow_p & \boldsymbol{\beta} + \mathbb{E}\left[x_i x_i^T\right]^{-1} \mathbb{E}\left[x_i \mu_{j(i)}\right] \\ = & \boldsymbol{\beta} + \mathbb{E}\left[x_i x_i^T\right]^{-1} \mathbb{E}\left[\bar{\bar{x}}_j \bar{\bar{x}}_j^T \boldsymbol{\lambda}\right]\end{aligned}$$

Therefore, letting c and c' represent a randomly chosen pair of classes taught by the same teacher, $\hat{\sigma}_\mu^2$ converges to

$$\begin{aligned}\hat{\sigma}_\mu^2 \rightarrow_p & \mathbb{E}\left[\mu_j^2\right] - 2\,\mathbb{E}\left[(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}) \bar{\bar{x}}_j \bar{\bar{x}}_j^T\right] \boldsymbol{\lambda} + \mathbb{E}\left[(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta})^T \bar{x}_c \bar{x}_c^T (\hat{\boldsymbol{\beta}} - \boldsymbol{\beta})\right] \\ = & \operatorname{Var}\left(\mu_j\right) + \boldsymbol{\lambda}^T \mathbb{E}\left[\bar{\bar{x}}_j \bar{\bar{x}}_j^T\right] \left(\mathbb{E}\left[x_i x_i^T\right]^{-1} \mathbb{E}\left[\bar{x}_c \bar{x}_{c'}^T\right] \mathbb{E}\left[x_i x_i^T\right]^{-1} - 2\,\mathbb{E}\left[x_i x_i^T\right]^{-1}\right) \mathbb{E}\left[\bar{\bar{x}}_j \bar{\bar{x}}_j^T\right] \boldsymbol{\lambda}\end{aligned} \tag{14}$$

Equation 14 shows that if there is no sorting on observables (so $\boldsymbol{\lambda} = 0$), then $\hat{\sigma}_\mu^2 \rightarrow_p \sigma_\mu^2$. Equation 14 also allows us to bound the asymptotic bias by considering two extremes: the case in which average student characteristics are perfectly correlated across different classrooms taught by the same teacher, so that $\mathbb{E}\left[\bar{x}_c \bar{x}_{c'}^T\right] = \mathbb{E}\left[\bar{x}_c \bar{x}_{c'}^T\right]$, and the case in which average characteristics are perfectly anti-correlated. Appendix C shows that

$$-b^T\,\mathbb{E}\left[x_i x_i^T\right] b \leq b\,\mathbb{E}\left[\bar{x}_c \bar{x}_{c'}^T\right] b \leq b^T\,\mathbb{E}\left[x_i x_i^T\right] b \quad \forall b \in \mathcal{R}^k.$$

When the first constraint binds, students are identical within a classroom, and the characteristics of two classrooms taught by the same teacher are perfectly anti-correlated. When the second constraint binds, all students taught by the same teacher are identical. Plugging these bounds into Equation 14,

$$-3 \leq \frac{\operatorname{plim}\left(\hat{\sigma}_\mu^2\right) - \operatorname{Var}\left(\mu_j\right)}{\boldsymbol{\lambda}^T\,\mathbb{E}\left[\bar{\bar{x}}_j \bar{\bar{x}}_j^T\right] \mathbb{E}\left[x_i x_i^T\right]^{-1} \mathbb{E}\left[\bar{\bar{x}}_j \bar{\bar{x}}_j^T\right] \lambda} \leq -1 \tag{15}$$

Equation 15 makes several facts apparent. The estimator is always negatively biased in asymptopia, and bias is more severe when sorting is strong. This happens when$\boldsymbol{\lambda}$ is large in magnitude. The asymptotic bias becomes smaller when classrooms taught by the same teacher are very similar to each other.

3.3 Modified version of above estimator: Modified-KS

As discussed above, omitted variables bias hyperparameter estimates in the Kane and Staiger estimator. Chetty *et al.* (2014) suggest remedying this by including teacher fixed effects when residualizing. That is, we obtain $\hat{\boldsymbol{\beta}}$ as the coefficient on x_i in a regression of outcomes on x_i and teacher fixed effects [8]. In a similar spirit, Guarino *et al.* (2014) discuss a similar issue in the context of a slightly different value-added procedure from that of Kane and Staiger (2008): the "mixed model" of Ballou *et al.* (2004), which differs from the model of Kane and Staiger (2008) in that it does not explicitly model classroom effects (θ_c). This model assumes that teacher effects are uncorrelated with student covariates, and performs poorly when that assumption is false. Guarino *et al.* (2014) explain that "estimators that include the teacher assignment indicators along with the covariates in a multiple regression analysis" perform better. In this section and in the simulations in Section **??**, I show that including fixed effects when estimating $\hat{\boldsymbol{\beta}}$ improves the *asymptotic* performance of the estimator when teacher effects are correlated with observables, but that including teacher fixed effects generates an incidental parameter problem that can create upward biases in finite samples.

Equation 14 makes the improved asymptotic performance of the estimator clear: This procedure yields a consistent estimate of $\boldsymbol{\beta}$, and a consistent estimate of $\hat{\boldsymbol{\beta}}$ leads to yields a consistent estimate of $\hat{\sigma}_\mu^2$. However, $\hat{\sigma}_\mu^2$ is still biased in finite samples:

$$\begin{aligned}\mathbb{E}\,\widehat{\text{Var}}(\mu_j) - \text{Var}(\mu_j) =& \frac{1}{N_{\text{teachers}}} \sum_j \frac{2}{|C(j)|\,|C(j)-1|} \sum_{c,c' \in C(j)} \mathbb{E}\left[\left(\hat{\beta}-\beta\right)^T \bar{x}_c \bar{x}_{c'}^T \left(\hat{\beta}-\beta\right)\right] \\ &- \frac{2}{N_{\text{teachers}}} \sum_j \frac{2}{|C(j)|\,|C(j)-1|} \sum_{c,c' \in C(j)} \mathbb{E}\left[\nu_c \left(\hat{\beta}-\beta\right)^T\right] \bar{x}_{c'}\end{aligned}$$

The sign of the bias is in general ambiguous, but seems to be positive in simulations, both with simulated outcomes and with permutation tests. The first term will be positive if student characteristics are sufficiently correlated across classrooms taught by the same teacher. This term will disappear quickly as the number of teachers increases.

3.3.1 Inference

Inference is the same as in the Kane and Staiger estimator, except that the first component of the moment condition changes to reflect that $\hat{\boldsymbol{\beta}}$ is now estimated off of within-teacher

[8] Chetty *et al.* (2014) use an estimator much more complicated than the Kane and Staiger estimator; they model the "drift" in teacher value-added across years. In this section, I use their modification to the estimation of $\hat{\beta}$ but do not study the rest of their model.

variation.

$$g_j(\eta) = \begin{pmatrix} \sum_{i \in S(j)} \left(x_i - \frac{1}{|S(j)|} \sum_{i' \in S(j)} \right) \left(y_i - x_i^T \hat{\boldsymbol{\beta}} - \frac{1}{|S(j)|} \sum_{i' \in S(j)} \left(y_{i'} - \boldsymbol{x}_{i'}^T \hat{\boldsymbol{\beta}} \right) \right) \\ \sum_{c,c' \in C(j)} \left(\bar{y}_c - \bar{\boldsymbol{x}}_c^T \hat{\boldsymbol{\beta}} \right) \left(\bar{y}_{c'} - \bar{\boldsymbol{x}}_{c'}^T \hat{\boldsymbol{\beta}} \right) - \hat{\sigma}_\mu^2 \left(\frac{|C(j)|(|C(j)|-1)}{2} \right) \\ \sum_{i:j(i)=j} \left(\tilde{y}_i - \tilde{\boldsymbol{x}}_i^T \hat{\boldsymbol{\beta}} \right)^2 - \hat{\sigma}_\varepsilon^2 \left(|S(j)| - |C(j)| \right) \\ \sum_{i:j(i)=j} \left(y_i - \boldsymbol{x}_i^T \hat{\boldsymbol{\beta}} \right)^2 - |S(j)| \left(\hat{\sigma}_\mu^2 - \hat{\sigma}_\theta^2 - \hat{\sigma}_\varepsilon^2 \right) \end{pmatrix}$$

3.4 Estimator derived from Fessler and Kasy (2017)

The difficulties with the previous two estimators stem from their failure to incorporate error from sampling variation in $\hat{\boldsymbol{\beta}}$ into estimation of $\hat{\sigma}_\mu^2$. The Kane and Staiger estimator attributes any correlation between student characteristics and teacher quality to the students, under-valuing the role of teachers, while the modified version tends to over-emphasize the role of teachers due to attenuation bias in the controls. Those estimators proceed sequentially, first estimating $\hat{\boldsymbol{\beta}}$, then assuming that $\hat{\boldsymbol{\beta}} = \boldsymbol{\beta}$ in future analysis. The general Empirical Bayes method developed in Fessler and Kasy (2017) explicitly accounts for sampling variation, and therefore might be expected to perform better.

In this method, as applied to estimating teacher effects, a preliminary estimator of teacher effects and of $\hat{\boldsymbol{\beta}}$ is constructed. Since variation in preliminary estimates of $\hat{\mu}$ between different teachers comes from both sampling error and true variation in teacher quality, we can back out the part of the variance that is due to true differences in teacher quality, as long as we can estimate sampling variation.

In the terminology of Fessler and Kasy (2017), the model is characterized by hyperparameters, which describe the entire distribution, and parameters, such as individual teacher effects. The parameters, η, are first estimated by a "preliminary, unrestricted estimator," and then the hyperparameters θ are backed out of the preliminary estimates of η. Finally, the hyperparameters are used to shrink the preliminary estimator towards a mean squared error-reducing Empirical Bayes estimate of the parameters. In concrete terms: The parameters are teacher effects μ, coefficients on covariates $\boldsymbol{\beta}$, and the variance V of the preliminary estimator $\begin{pmatrix} \hat{\mu}^p \\ \hat{\boldsymbol{\beta}}^p \end{pmatrix}$. The hyperparameters are V, $\boldsymbol{\beta}$, the mean of teacher effects m, and the variance of teacher effects σ_μ^2. We assume that the preliminary estimator $\hat{\mu}^p$ is normally distributed conditional on the true μ, and the teacher effects are iid. That is,

$$\begin{aligned} \text{Parameters} \quad & \eta = (\mu, \boldsymbol{\beta}, \gamma, V) \\ \text{Hyperparameters} \quad & \theta = (\sigma_\mu^2, \boldsymbol{\beta}, \gamma, V) \\ \text{Distribution of preliminary estimator} \quad & \left. \begin{pmatrix} \hat{\mu}^p \\ \hat{\boldsymbol{\beta}}^p \end{pmatrix} \right| \eta \sim N\left(\begin{pmatrix} \mu \\ \boldsymbol{\beta} \end{pmatrix}, V \right) \\ \text{Distribution of parameters} \quad & \mu_j | \theta \sim^{\text{iid}} N(m, \sigma_\mu^2) \end{aligned}$$

Combining the last two equations gives the distribution of the preliminary estimator given hyperparameters, finally relating observables to the quantities we are interested in:

$$\left(\begin{array}{c} \hat{\mu}^p \\ \hat{\boldsymbol{\beta}}^p \end{array} \right) \Bigg| \theta \sim N\left(\left(\begin{array}{c} z\gamma \\ \boldsymbol{\beta} \end{array} \right), \Sigma(\sigma_\mu^2, \hat{V}) \right) \tag{16}$$

where

$$\Sigma(\sigma_\mu^2, \hat{V}) = \left(\begin{array}{cc} \sigma_\mu^2 I & 0 \\ 0 & 0 \end{array} \right) + \hat{V}$$

To obtain preliminary estimates $\hat{\mu}^p$, $\hat{\boldsymbol{\beta}}^p$, and $\hat{V}$, we regress outcomes on teacher dummies and covariates. $\hat{V}$ should be a covariance matrix clustered at the classroom level, to reflect the model's assumptions that ν_i may be correlated within but not between classrooms. Then we estimate Equation 16 using maximum likelihood. That is,

$$\hat{\sigma}_\mu^2, \hat{\gamma}, \hat{\boldsymbol{\beta}} = \underset{\sigma_\mu^2, \gamma, \boldsymbol{\beta}}{\arg\min} \, \mathrm{LL}(\sigma_\mu^2, \gamma, \boldsymbol{\beta} | \hat{\mu}^p, \hat{\boldsymbol{\beta}}^p, \hat{V})$$

$$\mathrm{LL}(\sigma_\mu^2, \gamma, \boldsymbol{\beta} | \hat{\mu}^p, \hat{\boldsymbol{\beta}}^p, \hat{V}) = \log(\det\Sigma(\sigma_\mu^2, \hat{V})) + \left(\begin{array}{c} \hat{\mu}^p - z\gamma \\ \hat{\boldsymbol{\beta}}^p - \boldsymbol{\beta} \end{array} \right)^T \Sigma(\hat{V}, \sigma_\mu^2)^{-1} \left(\begin{array}{c} \hat{\mu}^p - z\gamma \\ \hat{\boldsymbol{\beta}}^p - \boldsymbol{\beta} \end{array} \right) \tag{17}$$

Methods for solving Equation 17 are given in Appendix D.

3.4.1 Individual teacher effects

To get Empirical Bayes shrinkage estimates of individual-specific results, assume that the parameters in θ have been estimated correctly and treat them as a prior. Let $S = \hat{V}^{-1}$.

$$\begin{aligned} \hat{\mu}^p \left| (\mu, \hat{\boldsymbol{\beta}}^p - \boldsymbol{\beta}, V) \right. &\sim N\left(\mu + S_{11}^{-1} S_{12} \left(\hat{\boldsymbol{\beta}}^p - \boldsymbol{\beta} \right), S_{11}^{-1} \right) \\ \mu \left| \left(z\gamma, \sigma_\mu^2 \right) \right. &\sim N\left(z\gamma, \sigma_\mu^2 I \right) \end{aligned}$$

Therefore,

$$\mu | \hat{\mu}, \hat{\boldsymbol{\beta}} - \boldsymbol{\beta} \sim N\left(\left(S_{11} + \frac{1}{\sigma_\mu^2} I \right)^{-1} \left(S_{11} \left(\hat{\mu} - \epsilon \right) + \frac{1}{\sigma_\mu^2} z\gamma \right), \left(S_{11} + \frac{1}{\sigma_\mu^2} I \right)^{-1} \right),$$

where

$$\begin{aligned} \epsilon &= \left(\hat{V}^{-1} \right)_{11}^{-1} \hat{V}_{12}^{-1} \left(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta} \right) \\ &= \hat{V}_{12} \hat{V}_{22}^{-1} \left(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta} \right). \end{aligned}$$

To find Empirical Bayes estimates of $\hat{\mu}$, substitute in the maximum likelihood estimates of $\boldsymbol{\beta}$, σ_μ^2, and m.

4 Simulation Experiments

Goal for this section:

- Describe two datasets: Teachers in NYC (from **??**) and bureaucrats in India (from **??**)
- Re-do Figure 1 using real covariates and assignments from New York data, but simulated teacher effects and outcomes. Currently, Figure 1 uses entirely invented data.
- Create a table version of Figure 1 showing bias and MSE for each estimator as a function of the correlation between covariates and outcomes
- Next, using same simulation procedure, turn attention to individual estimates: Show correlation between real and simulated effects, Spearman rank correlation, MSE, and percent of teachers who are miscategorized when bottom 2% are labeled.
- Repeat all of these experiments with India data, and maybe show how results change as sample size changes.

I ran simulations in which one hundred teachers are randomly assigned to one hundred classrooms of ten students each. Then, in each of nine succeeding periods, they are randomly reassigned to one hundred classrooms each consisting of ten previously unseen students. This creates a dataset with 10,000 observations. Students have one observable characteristic, x, which has mean zero and variance one. Teacher effects are correlated with their students' average value of x with a correlation of ρ, and teacher effects have a variance of 1. Outcome y_i is generated by

$$y_i = \mu_{j(i)} + x_i + \varepsilon_i,$$

where ε is normally distributed with mean zero and variance 1.

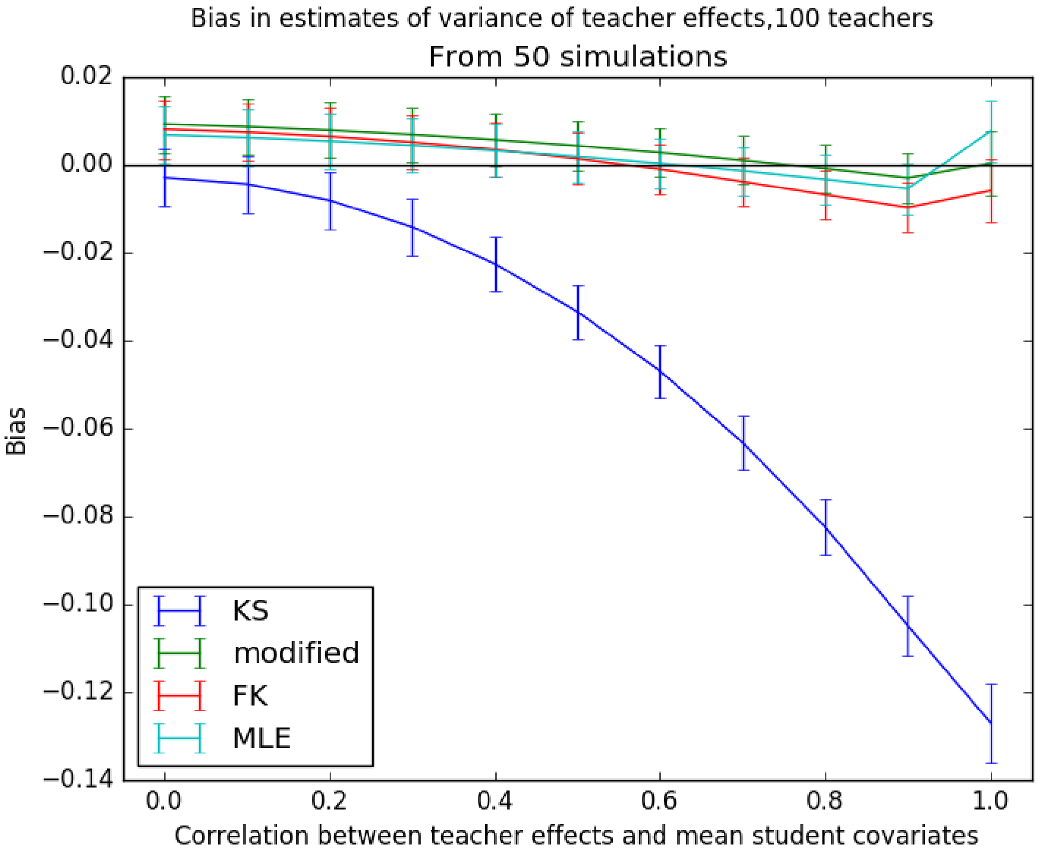
Students are randomly assigned to one of ten values of a categorical variable. This variable does not affect outcomes, but is included as a control in estimation.

4.1 Do Estimated Individual Effects Reflect True Effects?

4.1.1 Correlation between estimated and true teacher effects

	Truth	Kane and Staiger	Mod-KS	A
Truth	1.0			
Kane and Staiger	0.355	1.0		
Mod-KS	0.4	0.994	1.0	
A	0.413	0.969	0.977	1.0

Figure 1: *This simulation illustrates that the Kane and Staiger estimator is biased downwards when teacher effects are correlated with student characteristics. The modified version of this estimator has an upward bias that does not depend on the correlation of teacher effects with student characteristics.*



4.1.2 Accuracy of identifying bottom 2%

“Ineffective" bureaucrats have *estimated* VA scores in the bottom 2%. Effective bureaucrats have a *true* value-added in the top 50%.

	% of “ineffective" actually effective	Average VA of “ineffective"	MSE
Truth	0%	-2.29	0.00
Kane and Staiger	29%	-0.63	0.89
Mod-KS	25%	-0.74	0.86
A	23%	-0.86	1.02

5 Comparison on real data

Goal for this section:

- VAM parameter estimates from New York data
- VAM parameter estimates from India data
- Show how results change as sample size changes
- Describe results from **??** on randomization inference: Permutation tests demonstrate bias under the null hypothesis and provide p-values.
- Do what extent to different estimators agree on value-added scores?

Comparison of point estimates from different estimators on the dataset of officers in the Indian Administrative Service described in **??**.

Table 3 shows that value-added scores from all estimators are highly correlated.

Table 3: *Comparison of point estimates from different estimators.*

label	Kane and Staiger	Modified KS	Fessler and Kasy
Variance Components			
Teacher Effect (σ_μ^2)	-0.268%	25.682%	0.704%
Classroom Effect (σ_θ^2)	0.716%	4.228%	
Individual Shock (σ_ϵ^2)	33.105%	31.47%	
σ_μ		50.677%	8.388%

6 Conclusion

The main considerations governing choice of a value-added estimator are efficiency, computational resource needs, and whether individual value-added scores or just hyperparameter estimates are desired. For a practitioner who cares only about ranking teachers and does not care about the magnitude of each teacher's score, the choice of estimator can be governed mainly by computational considerations since, as Table 3 shows, correlations between value-added scores from different estimators are quite high. However, if a cardinal interpretation of value-added scores is desired, it becomes important to recover the right hyperparameters in order to impose the proper degree of shrinkage.

The Kane and Staiger estimator is the least computationally intensive, but comes with the most stringent identification requirements. The slowest step is regressing outcomes on covariates, which is $\mathcal{O}(NK^2)$ with N observations and K covariates. This algorithm then works with residuals, performing several quick $\mathcal{O}(N)$ computations. However, this estimator is only consistent when teachers are as good as randomly assigned. When teacher effects are correlated with student characteristics, teacher effects will be mistakenly attributed to student characteristics, biasing σ_μ^2 downwards and shrinking teacher effects closer towards zero. (When teacher effects are not correlated with student characteristics, this estimator will be the most efficient.)

The “modified-KS" estimator is more computationally intensive but allows for teacher effects to be correlated with student characteristics. It adds teacher fixed effects to the least squares computation. Using an exact least squares solver this increases asymptotic run-time to $\mathcal{O}(N(K + N_{\text{teachers}})^2)$, but with an iterative sparse least squares solver the impact of adding teacher fixed effects may be much less.

Maximum Likelihood and the procedure derived from Fessler and Kasy are both much slower due to the need for numerical optimization but seem to be less biased and do a better job assigning value-added scores, both in cardinal and ordinal terms.

References

Abowd, J. M., Kramarz, F. and Margolis, D. N. (1999). High Wage Workers and High Wage Firms. *Econometrica*, **67** (2), 251–333.

Ballou, D., Sanders, W. and Wright, P. (2004). Controlling for student background in value-added assessment of teachers. *Journal of educational and behavioral statistics*, **29** (1), 37–65.

Barnett, M. L., Olenski, A. R. and Jena, A. B. (2017). Opioid-Prescribing Patterns of Emergency Physicians and Risk of Long-Term Use. *New England Journal of Medicine*, **376** (7), 663–673.

Briggs, D. and Domingue, B. (2011). Due diligence and the evaluation of teachers: A review of the value-added analysis underlying the effectiveness rankings of Los Angeles Unified School District teachers by the. *National Education Policy Center*.

Buddin, R. (2011). Measuring teacher and school effectiveness at improving student achievement in Los Angeles elementary schools.

Chamberlain, G. (1982). Panel Data. In *Handbook of Econometrics*, vol. II, Elsevier Science Publishers BV, pp. 1248–1313.

— (1984). Multivariate Regression Models for Panel Data. *Journal of Econometrics*, **18 (1982)**, 5–46.

Chamberlain, G. E. (2013). Predictive effects of teachers and schools on test scores, college attendance, and earnings. *Proceedings of the National Academy of Sciences*, **110** (43), 17176–17182.

Chetty, R., Friedman, J. N. and Rockoff, J. E. (2014). Measuring the Impacts of Teachers I: Evaluating Bias in Teacher Value-Added Estimates. *American Economic Review*, **104** (9), 2593–2632.

—, — and — (2017). Measuring the Impacts of Teachers: Reply. *American Economic Review*, **107** (6), 1685–1717.

Ellison, G. and Swanson, A. (2016). Do Schools Matter for High Math Achievement? Evidence from the American Mathematics Competitions. *American Economic Review*, **106** (6), 1244–1277.

Felch, J., Ferrell, S., Garvey, M., Lauder, T. S., Lauter, D., Marquis, J., Pesce, A., Poindexter, S., Schwencke, K., Shuster, B., Song, J. and Smith, D. (). Los Angeles Teacher Ratings. *Los Angeles Times*.

Feng, J. and Jaravel, X. (2016). Who Feeds the Trolls? Patent Trolls and the Patent Examination Process.

Fessler, P. and Kasy, M. (2017). *How to use economic theory to improve estimators*. Tech. rep., Harvard University OpenScholar.

Green, D. P. and Winik, D. (2010). Using Random Judge Assignments to Estimate the Effects of Incarceration and Probation on Recidivism Among Drug Offenders*. *Criminology*, **48** (2), 357–387.

Guarino, C., Maxfield, M., Reckase, M., Thompson, P. and Wooldridge, J. (2014). An Evaluation of Empirical Bayes' Estimation of Value-Added Teacher Performance Measures.

Hanushek, E. A. and Rivkin, S. G. (2006). Chapter 18 Teacher Quality. In *Handbook of the Economics of Education*, vol. 2, Elsevier, pp. 1051–1078, dOI: 10.1016/S1574-0692(06)02018-6.

— and — (2010). Generalizations about using value-added measures of teacher quality. *The American Economic Review*, **100** (2), 267–271.

— and — (2012). The distribution of teacher quality and implications for policy. *Annu. Rev. Econ.*, **4** (1), 131–157.

Jacob, B. A. (2005). Accountability, incentives and behavior: the impact of high-stakes testing in the Chicago Public Schools. *Journal of Public Economics*, **89** (5-6), 761–796.

Kane, T. and Staiger, D. (2008). Estimating teacher impacts on student achievement: An experimental evaluation. *NBER Working Paper*.

Koedel, C. and Betts, J. R. (2011). Does student sorting invalidate value-added models of teacher effectiveness? An extended analysis of the Rothstein critique. *Association for Education Finance and Policy*, **6** (1), 18–42.

—, Mihaly, K. and Rockoff, J. E. (2015). Value-added modeling: A review. *Economics of Education Review*, **47**, 180–195.

Kozhimannil, K. B., Law, M. R. and Virnig, B. A. (2013). Cesarean Delivery Rates Vary Tenfold Among US Hospitals; Reducing Variation May Address Quality And Cost Issues. *Health Affairs*, **32** (3), 527–535.

Rothstein, J. (2009). Student sorting and bias in value-added estimation: Selection on observables and unobservables. *American Education Finance Association*, **4** (4), 537–571.

— (2010). Teacher Quality in Educational Production: Tracking, Decay, and Student Achievement. *The Quarterly Journal of Economics*, **125** (1), 175–214.

— (2017). Measuring the Impacts of Teachers: Comment. *American Economic Review*, **107** (6), 1656–1684.

Rubin, D. B. (1981). The Bayesian Bootstrap. *The Annals of Statistics*, **9** (1), 130–134.

Santos, F. (2012). City Teacher Data Reports Are Released. *WNYC*.

Staiger, D. O. and Rockoff, J. E. (2010). Searching for Effective Teachers with Imperfect Information. *Journal of Economic Perspectives*, **24** (3), 97–118.

A Maximum Quasi-Likelihood Robustly Estimates Variances

If we assume the model of Section 3, in which data is drawn from some distribution $\mathcal{D}$ and we do not assume a functional form for the distributions of μ, θ, and ε, then quasi-likelihood based on normality delivers consistent estimates of hyper-parameters $\eta = \left(\sigma_\mu^2, \sigma_\theta^2, \sigma_\varepsilon^2, \beta, \lambda, \alpha\right)$. To show this, I will derive a likelihood function in an alternate manner, show that it is maximized at η (CITATION), and then show that is is the same as the likelihood function presented in 3.1.

Stack all of teacher j's data into a vector y_j, a matrix x_j, and a precision-weighted mean of covariates $\bar{\bar{x}}_j$. We can restate the model of **??** in terms of best linear predictors:

$$\begin{aligned} E^*_{\mathcal{D}}\left[y_j | x_j, \mu_j\right] &= x_j^T \beta + \mu_j \\ E^*_{\mathcal{D}}\left[\mu_j | \bar{\bar{x}}_j\right] &= \bar{\bar{x}}_j^T \lambda. \end{aligned}$$

Therefore, $E^*_{\mathcal{D}}\left[y_j | x_j, \bar{\bar{x}}_j\right] = x_j^T \beta + \bar{\bar{x}}_j^T \lambda$ and $y_j \equiv x_j^T \eta + \bar{\bar{x}}_j^T \lambda + \nu_j$, where $\nu_j \perp x_j, \bar{\bar{x}}_j$. The sorting on observables requirement becomes $\mu_j \perp \nu_j - \mu_j$.

β corresponds to an unrestricted linear predictor. That is, define the best linear predictor π, so that

$$\mathrm{E}_{\mathcal{D}}\left[y_j | I_N \otimes \operatorname{vec}(x_j), \bar{\bar{x}}_j\right] = \left(I_n \otimes \operatorname{vec}(x_j)\right)\pi + \bar{\bar{x}}_j^T \lambda;$$

then $\left(I_n \otimes \operatorname{vec}(x_j)\right)\pi = x_j^T \beta$.

We can also define $E^*\left[\nu_j \nu_j^T\right] = \Sigma_j$. Let's put more structure on Σ_j, as a function of hyperparameters η and of s_j, which captures the structure of assignments of students to classrooms. Specifically, assume that

$$\begin{aligned} \Sigma\left(\eta, s_j\right)_{i,k} &= \sigma_\mu^2 + \sigma_\theta^2 + \sigma_\varepsilon^2 \quad \text{when} i = k \\ \Sigma\left(\eta, s_j\right)_{i,k} &= \sigma_\mu^2 + \sigma_\theta^2 \quad \text{when } i \neq k \text{ but } i \text{ and } k \text{ are in the same class} \\ \Sigma\left(\eta, s_j\right)_{i,k} &= \sigma_\mu^2 \quad \text{otherwise} \end{aligned}$$

Quasi-likelihood based on normality delivers consistent estimates of η. Consider the normal model

$$y_j | x_j, \bar{\bar{x}}_j \sim N\left(x_j \beta + \bar{\bar{x}}_j \lambda, \Sigma(\theta, s_j)\right),$$

with the corresponding likelihood function $f\left(y_j | x_j, \bar{\bar{x}}_j, s_j; \theta\right)$.

Lemma A.1.

$$\eta = \arg\max_\theta \mathrm{E}_{\mathcal{D}} \log f\left(y_j | x_j, \bar{\bar{x}}_j, s_j; \theta\right)$$

Proof.

$$\begin{aligned} \mathrm{E}_{\mathcal{D}} \log f(y_j | x_j, \bar{\bar{x}}_j; \theta) = &-\frac{1}{2} \log \det \Sigma(\theta, s_j) \\ &-\frac{1}{2} \mathrm{E}_{\mathcal{D}}\left[\left(y_j - x_j^T b - \bar{\bar{x}}_j^T \ell\right)^T \Sigma(\theta, s_j)^{-1}\left(y_j - x_j^T b - \bar{\bar{x}}_j^T \ell\right) | x_j, \bar{\bar{x}}_j, s_j\right] \end{aligned} \tag{18}$$

Since β corresponds to an unrestricted linear predictor, the values of β and λ that maximize Equation 18 do not depend on Σ, so

$$\arg\max_{b,\ell} \mathrm{E}_{\mathcal{D}} \log f(y_j|x_j, \bar{\bar{x}}_j; \sigma_\mu^2, \sigma_\theta^2, \sigma_\varepsilon^2, b, \ell) = \beta, \lambda$$

After plugging in $b = \beta$ and $\ell = \lambda$, we can rewrite Equation 18 as Therefore,

$$\begin{aligned}
&\mathrm{E}_{\mathcal{D}} \log f(y_j|x_j, \bar{\bar{x}}_j; \theta, \beta, \lambda) \\
&\quad = -\frac{1}{2} \log \det\Sigma(\theta, s_j) - \frac{1}{2}\mathrm{E}_{\mathcal{D}} \left[\left(y_j - x_j^T\beta - \bar{\bar{x}}_j^T\lambda \right)^T \Sigma(\theta, s_j)^{-1} \left(y_j - x_j^T\beta - \bar{\bar{x}}_j^T\lambda \right) |x_j, \bar{\bar{x}}_j, s_j \right] \\
&\quad = -\frac{1}{2} \log \det\Sigma(\theta, s_j) - \frac{1}{2}\mathrm{E}_{\mathcal{D}} \left[\mathrm{trace} \left(y_j - x_j^T\beta - \bar{\bar{x}}_j^T\lambda \right)^T \Sigma(\theta, s_j)^{-1} \left(y_j - x_j^T\beta - \bar{\bar{x}}_j^T\lambda \right) |x_j, \bar{\bar{x}}_j, s_j \right] \\
&\quad = -\frac{1}{2} \log \det\Sigma(\theta, s_j) - \frac{1}{2}\mathrm{trace} \left(\Sigma(\theta, s_j)^{-1} \mathrm{E}_{\mathcal{D}} \left[\left(y_j - x_j^T\beta - \bar{\bar{x}}_j^T\lambda \right)^T \left(y_j - x_j^T\beta - \bar{\bar{x}}_j^T\lambda \right) \right] \right] \right) \\
&\quad = -\frac{1}{2} \log \det\Sigma(\theta, s_j) - \frac{1}{2}\mathrm{trace} \left(\Sigma(\theta, s_j)^{-1} \Sigma(\eta, s_j) \right)
\end{aligned}$$

$$\begin{aligned}
\arg\max_\theta \mathrm{E}_{\mathcal{D}} \log f(y_j|x_j, \bar{\bar{x}}_j; \theta, \beta, \lambda) &= \arg\max_\theta - \log \det\Sigma(\theta, s_j) - \mathrm{trace} \left(\Sigma(\theta, s_j)^{-1} \Sigma(\eta, s_j \right) \\
&= \arg\max_\theta \log\det \left(\Sigma(\theta, s_j)^{-1} \Sigma(\eta, s_j) \right) - \mathrm{trace} \left(\Sigma(\theta, s_j)^{-1} \Sigma(\eta, s_j \right)
\end{aligned}$$

Let $\Sigma(\eta, s_j)^{1/2}$ be the symmetric, positive definite square root of the symmetric, positive definite matrix $\Sigma(\eta, x_j)$, and let λ_i be the eigenvalues of $\Sigma(\eta, s_j)^{1/2}\Sigma(\theta, s_j)^{-1}\Sigma(\eta, s_j)^{1/2}$. Since $\Sigma(\eta, s_j)^{1/2}\Sigma(\theta, s_j)^{-1}\Sigma(\eta, s_j)^{1/2}$ is positive definite, all of its eigenvalues are positive.

Then

$$\begin{aligned}
&\log\det \left(\Sigma(\theta, s_j)^{-1} \Sigma(\eta, s_j) \right) - \mathrm{trace} \left(\Sigma(\theta, s_j)^{-1} \Sigma(\eta, s_j \right) \\
&\qquad = \log\det \left(\Sigma(\eta, s_j)^{1/2} \Sigma(\theta, s_j)^{-1} \Sigma(\eta, s_j)^{1/2} \right) - \mathrm{trace} \left(\Sigma(\eta, s_j)^{1/2} \Sigma(\theta, s_j)^{-1} \Sigma(\eta, s_j)^{1/2} \right) \\
&\qquad = \log \Pi_i \lambda_i - \sum_i \lambda_i \\
&\qquad = \sum_i \left(\log(\lambda_i) - \lambda_i \right) \qquad (19)
\end{aligned}$$

Equation 19 is maximized when all $\lambda_i = 1$, which occurs when $\Sigma(\theta, s_j) = \Sigma(\eta, s_j)$. As long as the teacher teaches multiple classes and at least one class has multiple students, the only value that solves this equation is $\theta = \eta$.

□

Next, let's show that the likelihood in Equation 18 is equivalent to the likelihood in Equation **??**. First, note that through repeated applications of the Sherman-Morrision identity, we can invert $\Sigma(\eta, s_j)$. Let ℓ_c be a vector whose element i is 1 if student i is in classroom c and zero otherwise.

Define

$$\Sigma_C^{-1} = \frac{1}{\sigma_\varepsilon^2} I - \frac{1}{\sigma_\varepsilon^2} \sum_{c=1}^{C} \frac{\ell_c \ell_c^T}{n_c + \sigma_\varepsilon^2 / \sigma_\theta^2}$$

$$\Sigma(\eta, s_j)^{-1} = \frac{1}{\sigma_\varepsilon^2} I - \frac{\sigma_\theta^2}{\sigma_\varepsilon^2} \sum_c h_c \ell_c \ell_c^T / n_c - \frac{\sigma_\mu^2 \Sigma_C^{-1} \ell \ell^T \Sigma_C^{-1}}{1 + \frac{n\sigma_\mu^2}{\sigma_\varepsilon^2} - \frac{\sigma_\mu^2 \sigma_\theta^2}{\sigma_\varepsilon^2} \sum_c n_c h_c}$$

Equation 18 becomes

$$\begin{aligned} w =& \frac{1}{\sigma_\varepsilon^2} \sum_i \left(y_{ic} - x_{ic}^T \beta - \bar{\bar{x}}_{ic}^T \lambda - \alpha \right)^2 \\ &- \frac{\sigma_\theta^2}{\sigma_\varepsilon^2} \sum_c \frac{h_c}{n_c} \sum_i \left(y_{ic} - x_{ic}^T \beta - \bar{\bar{x}}_{j(i,c)}^T \lambda - \alpha \right)^2 - \sigma_\mu^2 k + \dots \end{aligned}$$

I haven't been able to analytically prove that the two equations are the same, and the determinant of Σ is a real sticking point, but I did numerically confirm that they are equivalent.

B MLE

How to get simple formula for the likelihood: tedious math here, could be copied from another document.

Also, analytic formula for gradient.

C Bounding the Asymptotic Bias in the Kane and Staiger Procedure

We want to show that

$$-b^T \, \mathbb{E}\left[x_{ic} x_{ic}^T\right] b \le b \, \mathbb{E}\left[\bar{x}_{j1} \bar{x}_{j2}^T\right] b \le b^T \, \mathbb{E}\left[x_{ic} x_{ic}^T\right] b \quad \forall b \in \mathcal{R}^k. \tag{20}$$

To begine, note that since $\bar{x}_{j1}$ and $\bar{x}_{j2}$ are exchangeable, $\mathbb{E}\left[\bar{x}_{j1} \bar{x}_{j1}^T\right] - \mathbb{E}\left[\bar{x}_{j1} \bar{x}_{j2}^T\right]$ is positive semidefinite:

$$\mathbb{E}\left[\bar{x}_{j1} \bar{x}_{j1}^T\right] - \mathbb{E}\left[\bar{x}_{j1} \bar{x}_{j2}^T\right] = \frac{1}{2} \, \mathbb{E}\left[\left(\bar{x}_{j1} - \bar{x}_{j2}\right)\left(\bar{x}_{j1} - \bar{x}_{j2}\right)^T\right].$$

Also, $\mathbb{E}\left[x_{ic} x_{ic}^T\right] - \mathbb{E}\left[\bar{x}_{j(i,c)1} \bar{x}_{j(i,c)1}^T\right]$ is positive semidefinite, since (letting x_{ic}^1 be a student drawn from classroom 1 of teacher $j(i,c)$):

$$\begin{aligned} \mathbb{E}\left[x_{ic} x_{ic}^T\right] - \mathbb{E}\left[\bar{x}_{j(i,c)1} \bar{x}_{j(i,c)1}^T\right] &= \mathbb{E}\left[x_{ic}^1 x_{ic}^{1T}\right] - 2\,\mathbb{E}\left[x_{ic}^1 \bar{x}_{j(i,c)1}\right] + \mathbb{E}\left[\bar{x}_{j(i,c)1} \bar{x}_{j(i,c)}^T\right] \\ &= \mathbb{E}\left[\left(x_{ic}^1 - \bar{x}_{j(i,c)1}\right)\left(x_{ic}^1 - \bar{x}_{j(i,c)1}\right)^T\right] \end{aligned}$$

Therefore, for any vector-valued b of the appropriate dimension,

$$b^T \left(\mathbb{E}\left[\bar{x}_{j1} \bar{x}_{j1}^T\right] - \mathbb{E}\left[\bar{x}_{j1} \bar{x}_{j2}^T\right] \right) b \ge 0 \tag{21}$$

and

$$b^T \left(\mathbb{E}\left[x_{ic}x_{ic}^T\right] - E\left[\bar{x}_{j(i,c)1}\bar{x}_{j(i,c)1}^T\right]\right) b \geq 0. \tag{22}$$

Combining Equations 21 and 22,

$$b^T\, \mathbb{E}\left[\bar{x}_{j(i,c)1}\bar{x}_{j(i,c)2}^T\right] b \leq b^T\, \mathbb{E}\left[x_{ic}x_{ic}^T\right] \quad \forall b \in \mathcal{R}^k.$$

When this constraint binds, students are identical within a classroom, and classroom averages are the same for all classes taught by the same teacher. In the opposite case, when classrooms are perfectly anticorrelated, then

$$b^T\, \mathbb{E}\left[\bar{x}_{j(i,c)1}\bar{x}_{j(i,c)2}^T\right] b \geq -b^T\, \mathbb{E}\left[x_{ic}x_{ic}^T\right] b \quad \forall b \in \mathcal{R}^K.$$

D Solving Equation 17

The log-likelihood from Equation 17 is

$$\mathrm{LL}(\sigma_\mu^2, \gamma, \beta | \hat{\mu}^p, \hat{\beta}^p, \hat{V}) = \log(\det\Sigma(\sigma_\mu^2, \hat{V})) + \begin{pmatrix} \hat{\mu}^p - z^T\gamma \\ \hat{\beta}^p - \beta \end{pmatrix}^T \Sigma(\sigma_\mu^2, \hat{V})^{-1} \begin{pmatrix} \hat{\mu}^p - z^T\gamma \\ \hat{\beta}^p - \beta \end{pmatrix}$$

Although Equation 17 can be easily solved using a black-box numerical optimization routine, it can be solved significantly faster with analytic simplifications and by providing a gradient and Hessian. It is especially helpful to decompose Equation 17 into a part that is not a function of the hyperparameters σ_μ^2, γ, or β, and a part that is a function of hyperparameters but is not computationally intensive.

D.1 Concentrating out parameters

We can solve analytically for β and γ as a function of σ_μ^2. First, combine γ and β into one parameter by stacking: $b^T = \left(\gamma^T, \beta^T\right)$, and let $R = \begin{pmatrix} z^T & 0 \\ 0 & I \end{pmatrix}$. Then

$$Rb = \begin{pmatrix} z^T & 0 \\ 0 & I \end{pmatrix} \begin{pmatrix} \gamma \\ \beta \end{pmatrix} = \begin{pmatrix} z^T\gamma \\ \beta \end{pmatrix}$$

Then we can rewrite the log-likelihood as

$$\mathrm{LL}(\sigma_\mu^2, b | \hat{\mu}^p, \hat{\beta}^p, \hat{V}) = \log(\det\Sigma(\sigma_\mu^2, \hat{V})) + \left(\begin{pmatrix} \hat{\mu}^p \\ \hat{\beta}^p \end{pmatrix} - Rb\right)^T \Sigma(\sigma_\mu^2, \hat{V})^{-1} \left(\begin{pmatrix} \hat{\mu}^p \\ \hat{\beta}^p \end{pmatrix} - Rb\right)^T$$

Then the optimal b as a function of Σ is

$$b(\sigma_\mu^2) = \left(R^T\Sigma^{-1}R\right)^{-1} R^T\Sigma^{-1} \begin{pmatrix} \hat{\mu}^p \\ \hat{\gamma}^p \end{pmatrix}$$

and

$$\begin{aligned}\left(\begin{pmatrix} \hat{\mu}^p \\ \hat{\beta}^p \end{pmatrix} - Rb(\sigma_\mu^2)\right) &= \left(I - R\left(R^T\Sigma^{-1}R\right)^{-1}R^T\Sigma^{-1}\right)\begin{pmatrix} \hat{\mu}^p \\ \hat{\gamma}^p \end{pmatrix} \\ &\equiv \left(I - P_R(\sigma_\mu^2)\right)\begin{pmatrix} \hat{\mu}^p \\ \hat{\gamma}^p \end{pmatrix}\end{aligned}$$

Substituting this back into the log-likelihood gives an expression that is a function only of σ_μ^2:

$$\mathrm{LL}(\sigma_\mu^2, b(\sigma_\mu^2)|\hat{\mu}^p, \hat{\beta}^p, \hat{V}) = \log\det\Sigma(\sigma_\mu^2) + \begin{pmatrix} \hat{\mu}^p \\ \hat{\gamma}^p \end{pmatrix}^T \left(I - P_R(\sigma_\mu^2)\right)^T \Sigma(\sigma_\mu^2)^{-1}\left(I - P_R(\sigma_\mu^2)\right)\begin{pmatrix} \hat{\mu}^p \\ \hat{\gamma}^p \end{pmatrix}$$

D.2 Simplifying the determinant and its Derivatives

We can rewrite the determinant as a function of the eigenvalues of the Schur complement of $\hat{V}$ and σ_μ^2. Although an eigendecomposition is expensive, it only needs to be done once to estimate the determinant of Σ at any value of σ_μ^2. Let the eigenvalues of $\hat{V}_{11} - \hat{V}_{12}\hat{V}_{22}^{-1}\hat{V}_{12}$ be λ_i.

$$\begin{aligned}\det\Sigma &= \det\left(\Sigma_{11} - \Sigma_{12}\Sigma_{22}^{-1}\Sigma_{12}\right)\det\left(\Sigma_{22}\right) \\ &= \det\left(\sigma_\mu^2 I + \hat{V}_{11} - \hat{V}_{12}\hat{V}_{22}^{-1}\hat{V}_{12}\right)\det\left(\hat{V}_{22}\right) \\ &= \Pi_i\left(\lambda_i + \sigma_\mu^2\right)\det\left(\hat{V}_{22}\right) \\ \log\det\Sigma &= \sum_i \log(\lambda_i + \sigma_\mu^2) + \log\det\hat{V}_{22}\end{aligned}$$

So the log-likelihood becomes

$$\mathrm{LL}(\sigma_\mu^2, b(\sigma_\mu^2)|\hat{\mu}^p, \hat{\beta}^p, \hat{V}) = \sum_i \log(\lambda_i + \sigma_\mu^2) + \begin{pmatrix} \hat{\mu}^p \\ \hat{\gamma}^p \end{pmatrix}^T \left(I - P_R(\sigma_\mu^2)\right)^T \Sigma(\sigma_\mu^2)^{-1}\left(I - P_R(\sigma_\mu^2)\right)\begin{pmatrix} \hat{\mu}^p \\ \hat{\gamma}^p \end{pmatrix}$$

Also note that

$$\begin{aligned}\frac{\partial \log\det\Sigma}{\partial\sigma_\mu^2} &= \sum_i \frac{1}{\lambda_i + \sigma_\mu^2} \\ \frac{\partial^2 \log\det\Sigma}{\partial(\sigma_\mu^2)^2} &= -\sum_i \frac{1}{(\lambda_i + \sigma_\mu^2)^2}\end{aligned}$$

D.3 Gradient

To compute the gradient, we will need the derivative of Σ^{-1}:

$$\begin{aligned}\frac{\partial\Sigma^{-1}}{\partial\sigma_\mu^2} &= -\Sigma^{-1}\frac{\partial\Sigma}{\partial\sigma_\mu^2}\Sigma^{-1} \\ &= -\Sigma^{-1}\begin{pmatrix} I & 0 \\ 0 & 0 \end{pmatrix}\Sigma^{-1}\end{aligned}$$

Also, derivative of the projection matrix with respect to σ_μ^2:

$$\begin{aligned}
\frac{\partial P_R}{\partial \sigma_\mu^2} &= R\left(R^T\Sigma^{-1}R\right)^{-1}R^T\frac{\partial \Sigma^{-1}}{\partial \sigma_\mu^2}\left(I-P_R\right)\\
&= -R\left(R^T\Sigma^{-1}R\right)^{-1}R^T\Sigma^{-1}\begin{pmatrix} I & 0 \\ 0 & 0 \end{pmatrix}\Sigma^{-1}\left(I-P_R\right)\\
&= -P_R\Sigma^{-1}\left(I-P_R\right)
\end{aligned}$$

Then the derivative of the log-likelihood is

$$\frac{\partial \mathrm{LL}}{\partial \sigma_\mu^2} = \sum_i \frac{1}{\lambda_i+\sigma_\mu^2} + \hat{b}^T\left[-\frac{\partial P_R}{\partial \sigma_\mu^2}^T\Sigma^{-1}(I-P_R) - (I-P_R)^T\Sigma^{-1}\frac{\partial P_R}{\partial \sigma_\mu^2} + (I-P_R)^T\frac{\partial \Sigma^{-1}}{\partial \sigma_\mu^2}(I-P_R)\right]\hat{b}$$

$$\begin{aligned}
-\frac{\partial P_R}{\partial \sigma_\mu^2}^T\Sigma^{-1}(I-P_R) &= \left(I-P_R\right)^T\Sigma^{-1}\begin{pmatrix} I & 0 \\ 0 & 0 \end{pmatrix}P_R^T\Sigma^{-1}\left(I-P_R\right)\\
-(I-P_R)^T\Sigma^{-1}\frac{\partial P_R}{\partial \sigma_\mu^2}^T &= \left(I-P_R\right)^T\Sigma^{-1}P_R\begin{pmatrix} I & 0 \\ 0 & 0 \end{pmatrix}\Sigma^{-1}\left(I-P_R\right)\\
(I-P_R)^T\frac{\partial \Sigma^{-1}}{\partial \sigma_\mu^2}(I-P_R) &= -(I-P_R)^T\Sigma^{-1}\begin{pmatrix} I & 0 \\ 0 & 0 \end{pmatrix}\Sigma^{-1}(I-P_R)
\end{aligned}$$

$$\frac{\partial \mathrm{LL}}{\partial \sigma_\mu^2} = \sum_i \frac{1}{\lambda_i+\sigma_\mu^2} + \hat{b}^T\left(I-P_R\right)^T\Sigma^{-1}\left[\begin{pmatrix} I & 0 \\ 0 & 0 \end{pmatrix}P_R^T + P_R\begin{pmatrix} I & 0 \\ 0 & 0 \end{pmatrix} - \begin{pmatrix} I & 0 \\ 0 & 0 \end{pmatrix}\right]\Sigma^{-1}\left(I-P_R\right)\hat{b}$$