Empirical Bayes

Empirical Bayes

Edward Ip
Geert Molenberghs

The empirical Bayes (EB) approach can be viewed from at least two perspectives: first, as a technical tool for borrowing information across individual cases for performing statistical inference; and second as a compromise between two statistical paradigm: Bayesian and frequentist. While the second topic is important for understanding EB's role within the inferential paradigms, it is too vast to do full justice to in this article. ~~We shall discuss EB's advantages and limitations after introducing it.~~

In education, the EB approach is particularly useful for enhancing the quality of educational measurements (student proficiency, teacher performance, school achievement) when multiple statistical units (schools, teachers, students) are measured. Let us illustrate how the EB procedure allows information to be borrowed across individual units for improving the quality of educational measurement. Suppose a state is interested in using a new portfolio assessment method for evaluating student performance on language ability. The goal is to evaluate the feasibility of replacing the traditional test with the new method. One is also interested in obtaining an accurate estimate of each school's performance so that they can design enhancement programs and allocate resources. The state administers the new assessment method to a sample of 100 students from each of 10 randomly selected schools. The minimum and maximum portfolio scores are 0 and 800, respectively. Table 1 displays the corresponding mean scores.

<Insert Table 1 here>

The question is: Are the mean scores good estimates for reporting school performance? Because of intra-school variability, a school's score depends upon the sample drawn: if a school's student population is heterogeneous, then the sample-to-sample variation may be large. Due to haphazard chance, a school may therefore appear not to be performing up to its true "potential;" the converse could also be true.

The primary idea behind EB is to use both local (sample mean score) and global (embedded in the distribution of the mean scores across schools) information to enhance the quality of the estimates. One can think of the 10 scores in Table 1 as a sample from the universe of the state's schools. The performance distribution of the state's schools can be used to inform the estimates for individual schools. It can be mathematically proven that, when both local and global data are used, the precision of the estimates for individual schools is better than when only local information is used, in terms of root mean square error.

EB normal-normal model

The EB setup is easily illustrated when both the global and local models are normally distributed: The global model states that the true mean scores θ_i**,** are sampled independently from a common underlying distribution, are not directly observed; and that they follow a normal distribution with mean μ and variance σ_θ^2:

$$\theta_i \sim N(\mu, \sigma_\theta^2)\,.\cancel{,} \qquad (1)$$

The above distribution is the~~the~~ so-called prior distribution. The local model allows for noise in measuring the score of the individual member and that the observed values of individual score y_i represent a realization of another normal distribution centered at θ_i :

$$y_i \mid \theta_i \sim N(\theta_i, \sigma_y^2)\,. \qquad (2)$$

<Insert Figure 1 near here>

This hierarchy is depicted in Figure 1. The normal prior distribution, ~~$g_\eta(\theta)$ (the normal distribution~~defined in (1)~~)~~ provides a model for global information, whereas the distribution $f_{\theta_i}(y)$ (the normal distribution in (2)) contains local information about the individual unit. This structure is rather general and can be applied to a variety of situations. For example, in the school performance example the observed values are collected at the student level. Suppose that each school contains a sample of *J* students; that the score of the j^{th} student from the i_{th} school is denoted by x_{ij} , $\boldsymbol{i = 1, \cdots, I,\ j = 1, \cdots, J}$; and that the student scores follow a distribution with mean θ_i and variance σ_x^2 . Then the mean score of each school is given by $y_i = (1/J)\sum_{j=1}^{J} x_{ij}$, and so:

$$y_i \mid \theta_i \sim N(\theta_i, \frac{\sigma_x^2}{J})\,. \qquad (3)$$

The structure in Figure 1 is of course not confined to normality. We shall see the some other examples later on.

The EB approach involves the following steps:

(S1). Estimate the global structure, using the marginal distribution of the observed data.

(S2). Estimate the value for each individual unit using global information from step S1, together with local information collected from each individual unit. This specific step of inference is through the posterior distribution of individual true (unobserved) measurement of each unit given the data.

To elaborate on (S1) and (S2), use notation of (1) and (2). First, consider the case in which σ_θ^2 and σ_y^2 are both known, with ~~then~~ only μ unknown and estimated in (S1). Because ~~the marginal distributions of~~ y_i are independently normally distributed with mean μ and variance $\sigma_\theta^2 + \sigma_y^2$, the maximum-likelihood estimate for μ is the sample mean:

$$\hat{\mu} = \bar{y} = \frac{1}{I}\sum_{i=1}^{I} y_i \ , \qquad (4)$$

(S2) involves estimating the true but unobserved mean score for each school. The typical EB estimate for a two-level model of this kind would be the posterior mean of the unknown true mean score given the observed data and the estimated global structure:

$$\hat{\theta}_i^{EB} = E(\theta_i \mid y_i, \mu) \, . \qquad (5)$$

From normal theory, the conditional distributions of θ_i given y_i are independently normally distributed:

$$\theta_{\ i} \mid y_i \sim N(B\mu + (1-B)y_i, (1-B)\sigma_y^2) \, , \qquad (6)$$

where $$B = \frac{\sigma_y^2}{\sigma_y^2 + \sigma_\theta^2} \, .$$

Therefore, the EB estimate $\hat{\theta}_i^{EB}$ is given by

$$\hat{\theta}_i^{EB} = B\bar{y} + (1-B)y_i \, , \qquad (7)$$

from “plugging in” the estimated parameter for the global structure—i.e., replacing μ with $\hat{\mu} = \bar{y}$. The uncertainty associated with the EB estimate $\hat{\theta}_i^{EB}$ is based on the variance $V(\theta_i \mid y_i, \mu)$ of the posterior distribution, with 95% confidence interval:

$$(\hat{\theta}_i^{EB} - 1.96 \times \sqrt{(1-B)}\sigma_y, \hat{\theta}_i^{EB} + 1.96 \times \sqrt{(1-B)}\sigma_y). \qquad (8)$$

Note that the EB approach, in contrast with general Bayesian approaches, always directly estimates the global structure from empirical data and then uses plug-in estimates of the global structure for estimating individual locality. The E in EB refers to this plug-in principle. To further illustrate the plug-in principle, retain the normal-normal case with the parameter σ_θ^2 also unknown. The marginal maximum likelihood estimate for the pair (μ, σ_θ^2) is given by

$$\hat{\mu} = \bar{y} = \frac{1}{I}\sum_{i=1}^{I} y_i \ , \qquad (9a)$$

$$\hat{\sigma}_\theta^2 = \max(s^2 - \sigma_y^2, 0) \, , \qquad (9b)$$

where $s^2 = (1/I)\sum_{i=1}^{I}(y_i - \bar{y})^2$ is the sample variance of the observed y_i. Following the plug-in principle, the EB estimate $\hat{\theta}_i^{EB}$ for an individual school is now

$$\hat{\theta}_i^{EB} = \hat{B}\bar{y} + (1-\hat{B})y_i \, , \qquad (10)$$

where $\hat{B} = \sigma_y^2 / (\sigma_y^2 + \hat{\sigma}_\theta^2),$ in which $\hat{\sigma}_\theta^2$ is plugged in from (9b). The confidence intervals take form (8), although now using $\hat{B}$ rather than B. In general, the EB-based confidence interval would be too narrow because it does not account for the uncertainty associated with the estimation of $\hat{B}$. There are different correction methods, including jackknife and bootstrap. The reading list includes articles that cover methods not requiring a fully Bayesian specification. We shall describe an alternative approach in the section on the EB application to the National Assessment of Educational Progress.

Example

The values in Table 1 were simulated from a two-level process with $\mu = 500$, $\sigma_\theta = 20$, $\sigma_x = 100$, $I = 10$, and $J = 100$. If we assume σ_x to be known, then the variance of the mean score $\sigma_y^2 = 100^2/100 = 100$. Thus, the estimates for (μ, σ_θ^2) are given by (9) and applied to Table 1, which results in $\hat{\mu} = 495.8$, $\hat{\sigma}_\theta = 15.3$, and $\hat{B}$ =0.3. Figure 2 compares the EB estimates and the values of the mean scores, which are ML estimates based only on data from one specific school.

<Insert Figure 2 about here>

The EB estimate in (10) is a weighted linear combination of two sources of information: globally from $\bar{y}$ and locally from y_i. Effectively, the EB approach amounts to a shrinkage effect that pulls all the ML estimates toward the overall mean. The strength of the shrinkage depends on the relative heterogeneity of data at the global and local levels. If the true mean scores are relatively homogeneous (small σ_θ^2, large σ_y^2), then the factor B should be close to 1, implying that the shrinkage effect would be large, and vice versa. The shrinkage factor B in this case is analogous to the intra-class correlation in classical test theory—the ratio of the between-student variation to the sum of the between-student variation and the between-school variation.

The EB estimate satisfies many desirable properties within the decision-theoretic framework. The square loss function is often used to quantify the risk for choosing one decision (estimator) over another. If the true value of the target parameter is θ, estimated by $\hat{\theta}$, then the squared loss function is $L(\theta, \hat{\theta}) = (\theta - \hat{\theta})^2$. In the normal-normal setting, if our interest is in the value of θ_i for each individual unit, it can be proven that under squared loss the EB estimator is superior to any other estimator, including those that are based on the observed data y_i alone. When the prior distribution for θ is known, then the EB estimator is ideal for any symmetric loss function. When the global structure is not known or the setup is more general, various versions of EB estimators have been proposed. Many authors have searched for optimality under various scenarios (Stein, Robbins, Maritz, Morris, Efron, Louis, among others).

The structure in Figure 1 can be further generalized in several directions. We will now give examples with different parametric choices for local and global levels.

EB gamma-Poisson model

Consider data collected over spelling errors made by school children. In a dictation test, school children are given a fixed number of *J* words. The number of errors that the *i*th child made can be modeled as a Poisson process with rate parameter θ_i, assumed to vary across individual children (Van Duijn and Bockenholt, 1995). It is mathematical convenient to assume that the θ_i follow a gamma distribution, that is,~~.~~ t~~T~~he global structure for spelling error rate parameters can be specified through the following two-parameter prior distribution:

$$p_{Gamma}(\theta_i \mid \alpha, \beta) = \frac{\theta_i^{\alpha-1}\beta^{\alpha}e^{-\beta\theta_i}}{\Gamma(\alpha)}, \quad \theta_i, \alpha, \beta > 0, \tag{11}$$

in which α and β are, respectively, the shape and rate (inverse scale) parameters, and Γ is the gamma function. If α is a positive integer, then $\Gamma(\alpha) = (\alpha-1)! = (\alpha-1)(\alpha-2)\cdots 3\times 2\times 1$. Conditional on θ_i, the total number of spelling errors for child *i* is given by the Poisson count model:

$$P(Y_i = y_i \mid \theta_i) = \frac{e^{\theta_i}\theta_i^{y_i}}{y_i\,!}, \tag{12}$$

where $y_i = 0, 1, 2, \cdots$. The marginal distribution for Y_i then is a negative binomial distribution:

$$P(Y_i = y_i) = (\frac{\beta}{\beta+1})^{\alpha} \frac{\Gamma(\alpha+y_i)}{\Gamma(\alpha)} \frac{1}{y_i\,!} \frac{1}{(\beta+1)^{y_i}}. \tag{13}$$

The standard method for estimating (α, β) is maximum likelihood, with the likelihood function assembled from contributions (13). It is convenient to reparameterize by the mean of the gamma distribution $\mu = \alpha / \beta$. The MLE of μ is $\bar{y} = (1/I)\sum_{i=1}^{I} y_i$, but there is no closed-form solution for α. One can also use the method of moments for the global structure, i.e., matching the moments of the marginal distributions with empirical moments and then solving for the required parameters. This produces (Maritz, 1969):

$$\hat{\mu} = \bar{y}, \tag{14a}$$

$$\hat{\beta} = \frac{\bar{y}^2}{(s^2 - \bar{y})}, \quad \text{if} \quad \mathrm{s}^2 > \bar{y}, \tag{14b}$$

where $s^2 = \frac{1}{I}\sum_{i=1}^{I}(y_i - \bar{y})^2$. If $s^2 < \bar{y}$, then the estimated prior distribution is taken to be degenerate at $\theta = \bar{y}$.

The EB estimate for θ_i is the posterior mean, with plug-in parameters given by (14). The posterior distribution for the gamma-Poisson model is also a gamma distribution with shape and rate parameters $(\alpha + y_i)$ and $(1+\beta)$. Because this posterior gamma distribution has mean $(\alpha + y_i)/(1+\beta)$, the EB estimate follows as:

$$\hat{\theta}_i^{EB} = \hat{B}\hat{\mu} + (1-\hat{B})y_i = \hat{B}\bar{y} + (1-\hat{B})y_i, \tag{15}$$

where $\hat{B} = \hat{\beta}/(1+\hat{\beta})$. The confidence interval for $\hat{\theta}_i^{EB}$ is based on the posterior variance

$$V(\theta_i \mid y_i, \mu, \beta) = (\mu\beta + y_i)(1+\beta)^2. \tag{16}$$

Again, a plug-in estimate for $(\hat{\mu}, \hat{\beta})$ is used .

EB beta-binomial model

The beta-binomial model has been used in the study of criterion-referenced testing for various purposes such as the determination of test length (Novick and Lewis, 1974) and cutoff scores (Huynh, 1977). A criterion-referenced test consists of a sample of *n* items drawn from a domain of items, developed from a framework of learning objectives. The number correct for the *i*th student is used to infer the student's proficiency, assumed unobservable. Denote the proficiency of the *i*th student by θ_i . Then, (1) states that θ_i , $i = 1, \cdots, I$, varies across students, following a beta distribution, and (2) states that, given proficiency θ_i, the number of correctly answered questions y_i follows a binomial distribution with parameters (n, θ_i), where *n* is the total number of questions on the test. The beta-binomial model can be expressed by the following two equations:

Comment [s1]: Rephrase. (1) is a normal distribution. So it cannot state that theta_i have beta distribution.

$$p_{Beta}(\theta_i \mid \alpha, \beta) = \frac{\theta_i^{\alpha-1}(1-\theta_i)^{\beta-1}}{\mathrm{Beta}(\alpha, \beta)}, \qquad 0 \le \theta_i \le 1, \tag{17}$$

where $\mathrm{Beta}(\alpha, \beta) = \frac{\Gamma(\alpha)\Gamma(\beta)}{\Gamma(\alpha+\beta)}$ is the beta function; and

$$P(Y_i = y_i \mid \theta_i) = \binom{n}{y_i} \theta_i^{y_i} (1-\theta_i)^{I-y_i}, \tag{18}$$

$y_i = 0, 1, \cdots, n$.

The marginal distribution of y_i is the so-called beta-binomial density:

$$P(Y_i = y_i) = \frac{1}{\mathrm{Beta}(\alpha, \beta)} \binom{n}{y_i} \frac{\Gamma(\alpha + y_i)\Gamma(\beta + I - y_i)}{\Gamma(\alpha + \beta + I)}. \tag{19}$$

To estimate the global-structure parameters, we resort to the method of moments, because ML does not generally lead to a closed form. Much as in the case of gamma-Poisson, we benefit from a reparameterization from (α, β) to (μ, λ), where $\mu = \alpha/(\alpha+\beta)$, and $\lambda = \alpha + \beta$. The parameter μ is the mean of the beta distribution, whereas the parameter λ suggests an effective sample size for the prior distribution, decreasing with variance. If *V* denotes the variance of the beta distribution, then $V = \mu(1-\mu)/(\lambda+1)$. Note that if a binary variable Y has mean μ, then with a sample size *n* the standard error for the ML estimate for μ is $V = \mu(1-\mu)/n$. Hence, $\lambda + 1$ plays the role of sample size. The moment estimates for (μ, λ) are:

$$\hat{\mu} = \frac{\sum_{i=1}^{I} y_i}{nI}, \text{ and} \tag{20a}$$

$$\hat{\lambda} = \frac{\hat{\mu}(1-\hat{\mu}) - s^2}{s^2 - \frac{\hat{\mu}(1-\hat{\mu})}{I}}, \tag{20b}$$

where $s^2 = \frac{1}{I}\sum_{i=1}^{I}(\frac{y_i}{n} - \hat{\mu})^2$.

The posterior distribution $\theta \mid y_i, \mu, \lambda$ is also a beta distribution with α, β replaced by $\alpha + y_i$ and $\beta + n - y_i$, respectively. Therefore, the mean of the posterior distribution is $\alpha + y_i / (\alpha + \beta + n)$, and after plugging in the estimates from above the EB estimate for individual θ_i ~~takes a revealing form~~is estimated as:

$$\hat{\theta}_i^{EB} = \hat{B}\hat{\mu} + (1 - \hat{B})\frac{y_i}{n}, \qquad (21)$$

where $\hat{B} = \hat{\lambda}/(\hat{\lambda} + n)$. The shrinkage factor $\hat{B}$ is a function of the relative strength of the sample sizes from global and local sources, with the global source providing an effective sample size $\hat{\lambda}$. When sample sizes vary across units, so does shrinkage. The confidence interval for $\hat{\theta}_i^{EB}$ ~~in~~ is based on the estimated posterior variance:

$$V(\theta_i \mid y_i, \hat{\mu}, \hat{\lambda}) = \frac{(y_i - \hat{\mu}\hat{\lambda})(n - y_i + \hat{\lambda} - \hat{\mu}\hat{\lambda})}{(n + \hat{\lambda})^2 (n + \hat{\lambda} + 1)}. \qquad (22)$$

Generalization and extensions

These ~~EB~~ instances of the EB method are special cases in which the prior and conditional distributions form conjugate pairs, leading to mathematically tractible marginal distributions. Furthermore, these are examples of the parametric EB approach. We will examine Figure 1 for extension of the approach. Starting from the top, the prior density $g_\eta(\theta)$ that provides the global information assumes a specific parametric form with hyperparameter η. In the normal-normal case, $g_\eta(\theta)$ is the normal distribution, with η assumed known in the first and $\eta = \mu$ in the second case. The conditional distribution for y_i, is often assumed to follow a separate parametric function $f(y \mid \theta)$. Typically, EB-based inferences target the unobserved individual θ_i, to which end there are the following steps:

(1) Form the marginal distribution of the observed data

$$p(\boldsymbol{y}) = \int g_\eta(\boldsymbol{\theta}) f(\boldsymbol{y} \mid \boldsymbol{\theta}) d\boldsymbol{\theta}, \qquad (23)$$

where $g_\eta(\boldsymbol{\theta}) = g_\eta(\theta_1) \times \cdots \times g_\eta(\theta_I)$, and $f(\boldsymbol{y} \mid \boldsymbol{\theta}) = f(y_1 \mid \theta_1) \times \cdots \times f(y_I \mid \theta_I)$.

(2) Using (23), estimate the hyperparameter η using, for example, maximum likelihood; leading to $\hat{\eta}$.

(3) Form the posterior distribution $p(\theta_i \mid y_i, \eta)$ using Bayes' theorem, and plug $\hat{\eta}$ into the posterior:

$$p(\theta_i \mid y_i, \eta) = \frac{p(y_i \mid \theta_i) g_\eta(\theta_i)}{\int p(y_i \mid \theta_i) g_\eta(\theta_i) d\theta_i}. \qquad (24)$$

(4) Form the EB estimate using step (3): $\hat{\theta}_i^{EB} = E(\theta_i \mid y_i, \hat{\eta})$. (5) Form confidence intervals of $\hat{\theta}_i^{EB}$, using variance function $V(\theta_i \mid y_i, \hat{\eta})$.

Several observations can be made. First, if the distributions $g_\eta(\theta)$ and $f(y \mid \theta)$ do not form a conjugate pair, the solutions for $\hat{\eta}$ and $\hat{\theta}_i^{EB}$ may require numerical procedures, as will their corresponding precision measures. The expectation-maximization (EM) algorithm (Demspter, Laird, and Rubin 1977) is often used for estimating the hyperparameters. Second, from a Bayesian perspective the empirical Bayes approach has stopped short of accommodating uncertainty associated with hyperparameter estimation. We shall discuss this issue within our next example. Third, the parametric EB approach regarding $g_\eta(\theta)$ and $f(y \mid \theta)$ can be extended by approached $g_\eta(\theta)$ nonparametrically instead. The pioneering work of Robbins (1955) established a theoretical foundation. In educational measurement, however, a more common extension is by incorporating regression models into either the prior $g_\eta(\theta)$, the conditional $f(y \mid \theta)$, or both. This will be exemplified in the next section.

EB analysis for the National Assessment of Education Progress

This example is chosen for two reasons. First, NAEP is a highly visible educational assessment tool in the United States, and reports are of great public interest because they are often cited to support specific educational and political agendas. Second and more technical, the statistical models used (Beaton and Johnson 1992) exhibit rich features, useful for illustrating both the power and limitations of the EB approach (cf. Scott and Ip 2002).

The NAEP survey measures the academic performance of U.S. students cross-sectionally and over time. Mandated by the U.S. Congress and funded by the federal government, NAEP reports academic achievements and identifies differences in performance between student population subgroups categorized by demographic and other contextual variables (e.g., time spent watching television). Unlike individual achievement tests (e.g. Scholastic Aptitude Test, SAT), NAEP reports on overall performance of subgroups or aggregates of students. While NAEP calculates individual proficiency scores and associated student sampling weights using draws from students' estimated ability distributions, it does not report individual student performance, this being prohibited by law. The NAEP approach for analyzing student data consists of two interlinked statistical models. The first captures local information about a student's academic performance based on his/her responses to cognitive items on a given subject matter. The second model creates a global structure, allowing sharing of information across students and schools. Denote by $\boldsymbol{\theta}_i = (\theta_{i1}, \cdots, \theta_{iS})$ the subscale proficiency level vector for student *i*, by $\boldsymbol{x}_i$ a vector of contextual and background variables, and by y_{ij} the response to the j^{th} item (correct = 1, incorrect = 0). For reading assessment, there are three subscales ($S = 3$): reading for literary experience, for information, and to perform a task. The global model states that:

$$\boldsymbol{\theta}_i \sim \text{MVN}(\beta^T \boldsymbol{x}_i, \Sigma), \qquad (25)$$

where β groups regression coefficients and Σ is a variance-covariance matrix. Such a specification allows the different subscales to be correlated. The local model for individual students is based on the item-response model (IRT, Lord 1980). For dichotomous items, the ~~following~~ three-parameter logistic IRT model given below is used for scaling student responses:

$$f(y_{ij} \mid \theta_{is(j)}) = P(Y_{ij} = 1 \mid \theta_{is(j)}) = c_j + \frac{(1 - c_j)}{1 + \exp(-1.7a_j(\theta_{is(j)} - b_j))}, \tag{26}$$

where $\theta_{is(j)}$ is the proficiency for subscale *s* to which item *j* belongs, $j = 1, \cdots, J_i$ and a_j, b_j, c_j are the item discrimination, difficulty, and guessing parameters, respectively. Factor 1.7 in (26) is a historical artifact relating the logit and probit links. To focus on the EB structure, we make the following simplifications: all items are considered dichotomous; their item parameters are a priori determined through a standard IRT-based calibration procedure and will be treated as known; and the student sampling weights are uniform.

Comment [s2]: NAEP has polytomous items---so you cannot just consider then to be dichotomous. Do you mean "we only analyze the items that are dichotomous"?

The EB approach is based on (25) and (26) and hyperparameter η contains β and Σ. Function $f(y_i \mid \theta_i)$ is now multivariate because y_i is the vector of the binary responses J_{is} for sudent *i* at the items of subscale *s*. Assuming local (conditional) independence, i.e., that the item responses for each subscale are independent given proficiency level θ_{is}, then $f(y_i \mid \theta_{is}) = f(y_{i1} \mid \theta_{is}) \times \cdots \times f(y_{iJ_{is}} \mid \theta_{is})$. Further assuming that the responses between subscales are independent, the overall response function therefore is the product over all subscales. Because of multiple test forms, the number of items answered per subscale is student-dependent. The likelihood is based on these observed only. This approach is valid assuming that the "missing" responses are missing at random (Little and Rubin 1987).

<Insert Figure 3 about here>

The parameters β and Σ are estimated by marginal maximum likelihood, the core of which is an EM algorithm (Mislevy, Johnson, and Muraki 1992). A standard EB approach would use the plug-in estimate for the regression coefficients β on students' background variables and Σ, and for the variance-covariance matrix for the subscales, to produce the expected value of student subscale proficiency θ_{is} given the observed responses and the plug-in estimates.

Bayesian shrinkage for θ_{is} tends to pull the estimate of proficiency purely based on an individual's responses toward the mean of the student's subgroup, defined by the levels of $\boldsymbol{x}_i$. Beaton and Johnson (1990) and Mislevy (1991) found asymptotic bias in statistics involving background variables that are not conditioned on, the magnitude of which relates to the extent to which responses may account for the unobserved proficiency θ_{is} and the degree to which the unconditioned background variables are explained by their counterparts in the model. Thus, the EB estimate with conditioning variables tends to mitigate potential bias~~, estimated to be approximately~~by about 10% across many NAEP analyses using the unconditional model (Mislevy 1991).

Figure 3 depicts the effect of Bayesian shrinkage. Here, we apply standard IRT procedures to the 1996 Long Term Trend Reading Data, based on responses to 22 items on a single scale for N = 918 students and assuming that the prior distribution is standard normal. The S-shaped curve shows that the absolute values of the EB estimates are smaller than their ML counterparts and that the shrinkage effect is larger for values that are farther away from the overall mean (zero).

<Insert Figure 4 near here>

While the EB estimate using the conditioning variables limits bias and generally outperforms ML, the measures for uncertainty associated with it (e.g., standard error), are underestimated, especially when based on the assumption that the plug-in estimates contain no measurement errors. There are many correction procedures, at individual and at population levels. NAEP adopted a methodology, called plausible value, based on multiple imputation (Rubin, 1987). The plausible-value method is implemented through several steps (Johnson, Mislevy and Thomas 1996):

(P1) Draw a value of (β, Σ) from a normal approximation to the posterior $p(\beta, \Sigma \mid \boldsymbol{y}, \boldsymbol{x})$, denoted by $(\tilde{\beta}^{(1)}, \tilde{\Sigma}^{(1)})$ [1];

(P2) Based on this, compute the mean and variance-covariance matrix for the posterior distribution of the proficiencies vector $p(\boldsymbol{\theta}_i \mid \boldsymbol{y}, \boldsymbol{x}, \tilde{\beta}^{(1)}, \tilde{\Sigma}^{(1)})$;

(P3) Draw a value for $\boldsymbol{\theta}_i^{(1)}$ using a multivariate normal approximation of $p(\boldsymbol{\theta}_i \mid \boldsymbol{y}, \boldsymbol{x}, \tilde{\beta}^{(1)}, \tilde{\Sigma}^{(1)})$, with mean and variance-covariance calculated from (P2);

(P4) Repeat steps (P1) through (P3) *M* times.

In NAEP, *M*=5. The total sampling variance for the proficiency estimate, or of any statistic based on the posterior $p(\boldsymbol{\theta}_i \mid \boldsymbol{y}, \boldsymbol{x}, \beta, \Sigma)$, is given by the sum of the average sampling variance over the *M* sets of plausible values $\boldsymbol{\theta}_i^{(m)}$, $m = 1, \cdots, M$, and the variance among the *M* estimates. The former, the so called within-imputation variance, is meant to be an approximation to the posterior variance $V(\boldsymbol{\theta}_i \mid \boldsymbol{y}, \boldsymbol{x}, \beta, \Sigma)$. The latter, the between-imputation variance is designed as a correction for the uncertainty due to not directly observing the θ_{is} and is:

$$G = \sum_{m=1}^{M} \frac{(\theta_{is}^{(m)} - \bar{\theta}_{is})^2}{M - 1}, \qquad (27)$$

where $\bar{\theta}_{is}$ is the mean of $\theta_{is}^{(m)}$. The final estimate of the variance of $\hat{\boldsymbol{\theta}}_i^{EB}$ is:

$$V(\hat{\theta}_i^{EB}) = \hat{V}(\boldsymbol{\theta}_i \mid \boldsymbol{y}, \boldsymbol{x}, \beta, \Sigma) + (1 + M^{-1})G. \qquad (28)$$

Hence, this approach aims at conducting proper inference for student proficiency at both the individual and population levels by enhancing the EB approach so that it approximates a fully Bayesian approach. It is argued that the fully Bayesian approach takes into account the various sources of uncertainty, including those derived from using a plug-in estimate for (β, Σ).

Formatted: Indent: First line: 1,27 cm

[1] The operational NAEP procedure keeps Σ fixed and draws only for β.

For illustration, we analyze the 1998 eighth-grade national reading assessment (N = 11,051) and Table 2 provides the mean and the within- and between-imputation variances; *M* = 100. For more details and ramifications see Scott and Ip (2002). Here, the between- and within-imputation variances are comparable, implying that the second source of variability increases the standard error by about a factor $\sqrt{2} = 1.4$, relative to EB.

<Insert Table 2 near here>

Other educational applications and summary

The EB approach relies on a two-level formulation of how data across statistical units arise, and offers a rigorous theoretical framework for using global information gleaned across individual units for informing local estimates, which are only based on responses from a specific individual. In contrast to a fully Bayesian approach, EB emphasizes estimating prior distributions from the data. From a modeling perspective, the EB approach is amenable for (non)linear multi-level modeling, especially when inference targets unobservables such as proficiencies or teaching skills.

The range of EB applications is wide. Braun (2006) applied the approach to analyze important performance predictors on the Graduate Record Examination (GRE) across various college departments. He used a normal-normal model in which the student's first-year grade point average is treated as the response *y*, the student's GRE scores as predictors, and regression coefficients as unobservables of interest. The global structure is a regression model incorporating department-level covariates. Meta-analysis (see the article on that topic in this volume) for educational studies also make use of EB; information is combined across studies, and the effect size of a specific classroom intervention from study *i* can be conceptualized as an unobserved variable θ_i. Information is borrowed across studies through the specification of a global data-generating structure $g_\eta(\theta)$ (Hedges 1987; Raudenbush and Bryk 1985).

Comment [s3]: This terminology is a bit confusing. Change it to something simpler.

For evaluating teacher and school effectiveness, EB was applied to value-added models—a collection of models that attempts to delineate, from a return-on-investment perspective, the effectiveness of school systems and personnel based on the complex interactions among student characteristics, school effects, community characteristics, school district policies, etc. Information is borrowed from teacher assessment results for inferences on a school system or employee, and regression models are incorporated at global and local levels to control for interaction effects. McCaffrey et al. (2004) review valued-added models and related EB methodologies. Multi-level growth-curve modeling for assessing educational-intervention effects (Plewis 2000; Pituch 2001) is another versatile application area. Typically, a growth curve from an individual student contains only few data points, but strength can be borrowed across students to stabilize individual growth curve estimates. This conventionally results in individual curves shrunken toward a smoother, population-based growth version. Another classic EB example is in the study of school effects through hierarchical linear modeling (Raudenbush 1986). The author sets up a multi-level hierarchical model with covariates at individual and school levels.

Comment [s4]: How can regression models control for interaction effects? Explain

The EB approach capitalizes on using information from higher-level statistical units to enhance estimates at individual level, without subjective input on prior distributions. From a Bayesian perspective, the EB approach is regarded as an approximation to a fully Bayesian approach. The general notion of borrowing information across statistical units is powerful and extends readily to many innovative applications. A recent example is in simultaneous multiple hypothesis testing (Efron, 2004). Unsurprisingly, EB has been regarded as one of the most important advances in the field of statistics since World War II (Efron, 2007).