

Multiple Overimputation: A Unified Approach to Measurement Error and Missing Data*

Matthew Blackwell[†]

James Honaker[‡]

Gary King[§]

November 1, 2011

Abstract

Social scientists typically devote considerable effort to mitigating measurement error during data collection but then ignore the issue during data analysis. Although many statistical methods have been proposed for reducing measurement error-induced biases, few have been widely used because of implausible assumptions, high levels of model dependence, difficult computation, or inapplicability with multiple mismeasured variables. We develop an easy-to-use alternative without these problems; it generalizes the popular multiple imputation (MI) framework by treating missing data problems as a special case of extreme measurement error and corrects for both. Like MI, the proposed “multiple overimputation” (MO) framework is a simple two-step procedure. First, multiple (≈ 5) completed copies of the data set are created where cells measured without error are held constant, those missing are imputed from the distribution of predicted values, and cells (or entire variables) with measurement error are “overimputed,” that is imputed from the predictive distribution with observation-level priors defined by the mismeasured values and available external information, if any. In the second step, analysts can then run whatever statistical method they would have run on each of the overimputed data sets as if there had been no missingness or measurement error; the results are then combined via a simple averaging procedure. With this paper, we offer open source software that implements all the methods described herein.

*For helpful comments, discussions, and data we thank Gretchen Casper, Simone Dietrich, Justin Grimmer, Sunshine Hillygus, Burt Monroe, Adam Nye, Michael Peress, Eric Plutzer, MO Tavano, Shawn Treier, Joseph Wright and Chris Zorn.

[†]Doctoral Candidate, Department of Government, Harvard University, Institute for Quantitative Social Science, 1737 Cambridge Street, Cambridge, MA 02138 (mblackwell@iq.harvard.edu, <http://www.mattblackwell.org>)

[‡]Lecturer, The Pennsylvania State University, Department of Political Science, Pond Laboratory, University Park, PA 16802 (tercer@psu.edu, <http://www.personal.psu.edu/jah72/>)

[§]Albert J. Weatherhead III University Professor, Harvard University, Institute for Quantitative Social Science, 1737 Cambridge Street, Cambridge, MA 02138 (king@harvard.edu, <http://gking.harvard.edu>)

1 Introduction

Social scientists recognize the problem of measurement error in the context of data collection, but seem to ignore it when choosing statistical methods for the subsequent analyses. Some seem to believe that analyses of variables with measurement error will still be correct on average, but this is untrue; others act as if the attenuation that occurs in simple types of random measurement error with a single explanatory variable holds more generally, but this too is incorrect. More sophisticated application-specific methods for handling measurement error exist, but they are complicated to implement, require difficult-to-satisfy assumptions, and often lead to high levels of model dependence; few such methods apply when error is present in more than one variable or are used widely in applications, despite an active methodological literature. Unfortunately, the corrections used most often are the easiest to implement but typically have the strongest assumptions, which we discuss below (see [Stefanski \(2000\)](#) and [Guolo \(2008\)](#) for literature reviews). As with missing data problems a decade ago, many current empirical literatures could benefit from a comprehensive, easy-to-use approach.

We address this challenge by offering a unified approach to correcting for problems of measurement error and missing data in a single easy-to-use procedure. We do this by generalizing the multiple imputation (MI) framework designed for missing data ([Rubin, 1987](#); [King et al., 2001](#)) to broadly deal with measurement error as partially missing information and treat completely missing cell values as an extreme form of measurement error. The proposed generalization, which we call *multiple overimputation* (MO), enables researchers to treat cell values as either observed without (random) error, observed with error, or missing. We accomplish this by constructing prior distributions for individual cells (or entire variables) with means equal to the observed values, if any, and variance for the three data types set to zero, a (chosen or estimated) positive real number, or infinity, respectively.

Like MI, the easy-to-use MO procedure involves two steps. First, analysts use our software to create multiple (usually about five) data sets by drawing them from their posterior predictive distribution conditional on all available observation-level information. This procedure leaves the observed data constant across the data sets, imputes the missing values from their predictive posterior as usual under MI, and *overimputes*, that is, replaces or overwrites the values or variables measured with error with draws from their predictive posterior. Our basic approach to measurement error, which involves relatively minimal assumptions, allows for random measurement error in any number or combination of variables or cell values in a data set. With somewhat more specific assumptions, we also allow for measurement error that is heteroskedastic or correlated with other variables. As we show, the technique is relatively robust to violations of either set of assumptions and easy to apply.

An especially attractive advantage of MO (like MI) is the second step, which enables analysts to run whatever statistical procedure they would have run on the completed data sets, as if all the data had been correctly observed. A simple procedure is then used to average the results from the separate analyses. The combination of the two steps enables scholars to overimpute their data set once and to then set aside the problems of missing data and measurement error for all subsequent analyses. As a companion to this paper, we have modified a widely used MI software package to also perform MO ([Honaker, King and Blackwell, 2010](#)).

Section 2 describes our proposed MO framework, in the context of multiple variables measured with random error with a known, assumed, or completely unknown variance. There, we generalize the MI framework, prove that a fast existing algorithm can be used to create imputations for MO,

and offer Monte Carlo evidence that it works as designed. Section 3 goes further by deriving methods of estimating the (possibly heteroskedastic) measurement error variance so it need not be assumed. Section 4 generalizes our approach further still by allowing measurement error that is correlated with the true values of the variables. Section 5 then offers empirical illustrations.

2 A Multiple Overimputation Model

We conceptualize the linkage between measurement error and missing data in two equivalent ways. In one, measurement error is a specific type of missing data problem where observed proxy variables provide probabilistic prior information about the true unobserved cell values. In the other, missing cell values have an extreme form of measurement error where no available prior information exists. Either way, the two methodological problems go well together because variables (or cell values) measured with error fall logically between the extremes of observed without error and completely unobserved. This dual conceptualization also means that our MO approach to measurement error has all the advantages of MI in ease of use and robustness (Schafer, 1997; Freedman et al., 2008).

The validity of our approach is also easy to understand within this framework: Deleting cell values with measurement error and using MI introduces no biases, and running MI while also using observed cell values that are informative but measured with some error to help inform cell-level priors clearly improves efficiency and reduces model dependence. Adopting, instead, an application-specific approach will, under some conditions, perform better, but only with high costs in terms of designing and using specialized statistical models, analyses, and software.¹

2.1 The Foundation: A Multiple Imputation Model

MO builds on MI, which we now review. MI involves using a model to generate multiple imputations for each of the missing cell values (as predicted from all available information in the data set), separate analysis of each of the completed data sets without worry about missing data, and then the application of some easy rules for combining the separate results. The main computational difficulty comes in developing the imputation model.

For expository simplicity, consider a simple special case with only two variables, y_i and x_i ($i = 1, \dots, n$), where only x_i contains some missing values. These variables are not necessarily outcome and explanatory variables, as they can each play any role in the subsequent analysis model. Everything in this section generalizes to any number of variables and arbitrary patterns of missingness in any or all of the variables (Honaker and King, 2010).

We now write down a common model that could be used to apply to the data if they were complete, and then afterwards explain how to use it to impute any missing data scattered through the input variables. This model assumes that the joint distribution of y_i and x_i , $p(y_i, x_i | \mu, \Sigma)$, is multivariate normal:

$$(x_i, y_i) \sim \mathcal{N}(\mu, \Sigma), \quad \mu = (\mu_y, \mu_x), \quad \Sigma = \begin{pmatrix} \sigma_y^2 & \sigma_{xy} \\ \sigma_{xy} & \sigma_x^2 \end{pmatrix}, \quad (1)$$

where the elements of the mean vector μ and variance matrix Σ are constant over the observations. This model is deceptively simple but powerful: As there is no i subscript on the scalar means

¹Scholars who have made this connection before have focused almost exclusively on data with validation subsamples, which are relatively rare in the social sciences (Wang and Robins, 1998; Brownstone and Valletta, 1996; Cole, Chu and Greenland, 2006). For a related problem of “editing” data with suspicious cell values, Ghosh-Dastidar and Schafer (2003) develop a MI framework similar in spirit to ours, albeit with an implementation specific to their application.

μ_x and μ_y , it may appear as though only the marginal means are used to generate imputations. Yet, its joint distribution implies that a prediction is always based on a regression (the conditional expectation) of that one variable on *all* the others, with the population values of the coefficients in the regression a deterministic function of μ and Σ . This is extremely useful in missing data problems for predicting a missing value conditional on observed values. For instance, given model (1), the conditional expectation is $E[x_i|y_i] = \gamma_0 + \gamma_1(y_i - \mu_y)$, where $\gamma_0 = \mu_x$ and $\gamma_1 = \sigma_{xy}/\sigma_x$. Researchers have repeatedly demonstrated that this imputation model works as well as more complicated non-linear and non-normal alternatives even for ordinal or categorical variables, and even when more sophisticated models are preferred at the analysis stage (Schafer 1997 and citations in King et al. 2001).

Thus, to estimate the regression of each variable in turn on all the others, we only need to estimate μ and Σ . If there were no missing data, the results would be equivalent to running each of the separate regressions (y_i on x_i and x_i on y_i). But how can we run either of these regressions with arbitrary missing data? The trick, which we now explain, is to find a single set of estimates of μ and Σ from data with scattered missingness, and then to use these to deterministically compute the coefficients of all the separate regressions. To be more specific, the “complete-data” likelihood (i.e., still assuming no missing data) is simply the product of model (1) over the n observations:

$$\mathcal{L}(\theta|y, x) \propto \prod_i p(y_i, x_i|\theta) \tag{2}$$

$$= \prod_i p(x_i|y_i, \theta)p(y_i|\theta), \tag{3}$$

where $\theta = (\mu, \Sigma)$. (We use variables without an i subscript to denote the vector of observations, so $y = (y_1, \dots, y_n)$.) This likelihood is not usable as is because it is a function of the missing data, which we do not observe. Thus, we integrate out whatever missing values happen to exist for each observation to produce the actual (“observed-data”) likelihood:

$$\mathcal{L}(\theta|y, x^{\text{obs}}) \propto \prod_i \int p(x_i|y_i, \theta)p(y_i|\theta)dx^{\text{mis}} \tag{4}$$

$$= \prod_{i \in x^{\text{mis}}} p(y_i|\theta) \prod_{j \in x^{\text{obs}}} p(x_j|y_j, \theta)p(y_j|\theta), \tag{5}$$

where x^{obs} denotes the set of cell values in x that are observed and x^{mis} the set that are missing. That we can partition the complete data in this way is justified by the standard “missing at random” (MAR) assumption that the missing values may depend on observed values in the data matrix but not on unobservables (Schafer, 1997; Rubin, 1976). The key advantage of this expression is that it appropriately assumes that we only see what is actually observed, x^{obs} and y , but can still estimate μ and Σ .²

This result enables one to take a large data matrix with scattered missingness across any or all variables and impute missing values based on the regression of each variable on all of the others. The actual imputations used are based on the regression predicted values, their estimation uncertainty

²This observed-data likelihood is difficult to maximize directly in real data sets with arbitrary patterns of missingness. Fast algorithms to maximize it have been developed that use the relationship between (1), (4), and the implied regressions, using iterative techniques, including variants of Markov chain Monte Carlo, EM, or EM with bootstrapping.

(due to the fact that μ and Σ , and thus the calculated coefficients of the regression, are unknown), and the fundamental uncertainty (as represented in the multivariate normal in (1) or, equivalently, the regression error term from each conditional expectation).

MI works by imputing about five values for each missing cell entry (or more for data sets with unusually high missingness), creating “completed” data sets for each, running whatever analysis model we would have run on the each completed data set as if there were no missing values, and averaging the results using a simple set of rules. The assumption necessary for MI to work properly is that the missing cell values are textscmar. This is considerably less restrictive than, for example, the “missing completely at random” assumption required to avoid bias in listwise deletion. See the Appendix for a more formal treatment of the assumptions behind MO.

2.2 Incorporating Measurement Error

The measurement error literature uses a variety of assumptions that are, in different ways, more and also less restrictive than our approach. The “classical” error-in-variables model assumes the error is independent of the true value being measured. “Nondifferential” or “surrogate” error is that assumed independent of the dependent variable, conditional on both the true value being measured and any observed pre-treatment predictor variables. Other approaches use assumptions about exclusion restrictions or auxiliary information such as repeated measures. See [Imai and Yamamoto \(2010\)](#) for formal definitions and citations to the literature.

In our alternative approach, we marshal two distinct sources of information to overimpute cell values with measurement error. One could think of cell values with any positive level of measurement error as effectively missing values, and the observed cell value is useless information. In this situation, we can easily translate a measurement error problem into a missing data problem, for which the observed-data likelihood derived in Section 2.1 applies directly. The assumption required for this procedure is MAR, which is considerably less restrictive than the assumptions necessary for most prior approaches to dealing with measurement error, and, unlike most other measurement error approaches, it may be used for data sets with arbitrary patterns of measurement error (and missingness) in any (explanatory or dependent) variables.

Of course, if we think of observations measured with error as reasonable proxies for unobserved values, then treating them as missing will work but may discard valuable information. In fact, variables entirely measured with error may leave no information with which to make (over)imputations under this approach. Thus, we supplement the information that would come from treating cell values measured with error as completely unobserved, and its relatively minimal assumptions, with a second source of information — the proxy measurements themselves along with assumptions about the process by which the proxies are observed. This second source of information enables researchers to make somewhat stronger assumptions, when the measured proxies bear some relationship to the unobserved true values, in return for considerably more efficient estimates.

For expository clarity, we continue, without loss of generality, our simple two-variable example from the previous section. Thus, let y_i be a single fully observed cell value and x_i^* be a true but unobserved cell value (these variables may serve any role in a subsequent analysis model), with $(y_i, x_i^*) \sim \mathcal{N}(\mu, \Sigma)$ as above. To this, we add an observed w_i which is a proxy, measured with error, for x_i^* . For expository simplicity, we focus on the case with no (fully) missing values, which in this context would be unobserved cell values without corresponding proxy values.

With this setup, we describe the second source of information in our approach as coming from the specification of a specific probability density to represent the data generation process for the proxy w_i . This, of course, is an assumption and we allow a wide range of choices, subject to two

conditions, one technical and one substantive. First, the class of allowable data generation processes in our approach involves any probability density that possesses the property of *statistical duality*. This is a simple property (related to self-conjugacy in Bayesian analysis) possessed by a variety of distributions, such as normal, Laplace, Gamma, Inverse Gamma, Pareto, and others (Bityukov et al., 2006).³ (We use this property to ease implementation in Section 2.3.) Second, we require that the mean (or an additive function of the mean) of the distribution be the unobserved true cell value x_i^* , and that the parameters of the distribution are distinct from the complete-data parameters, θ , and are known or separately estimated.

A simple special case of this data generation process is random normal measurement error, $w_i \sim \mathcal{N}(x_i^*, \sigma_u^2)$, with σ_u^2 set to a chosen or estimated value (we discuss interpretation and estimation of σ_u^2 in Section 3). Other special cases allow for heteroskedastic measurement error, such as might occur with GDP from a country where a government’s statistical office is professionalizing over time; mortality statistics from countries with and without death registration systems; or survey responses from a self-report vs elicited about that person from someone else in the same household. This approach can handle biased measurement error, where $E[w_i|x_i^*] = a_i + x_i^*$, so long as the bias, a_i , is known or estimable. For instance, if validation data is available, a researcher could estimate the bias of the measure or use a model to estimate how the offset changes with observed variables. From our perspective, a cell value (or variable) that doesn’t possess at least this minimally known set of relationships to its true value could more easily be considered a new observation of a different variable rather than a proxy for an unobserved one.⁴ When the bias is not known and cannot be estimated, we are left with a class of data generation processes (rather than a single one) for the proxy; this results under our procedure in a “robust Bayesian” *class* of posteriors (rather than a single Bayesian posterior), from which overimputations may be drawn (Berger, 1994; King and Zeng, 2002).

The result of these assumptions is a complete-data likelihood that can be used to encompass both methodological problems:

$$L(\theta, \sigma_u^2 | y, w, x^*) \propto \prod_i p(y_i, w_i, x_i^* | \theta, \sigma_u^2) \quad (6)$$

$$= \prod_i p(w_i | x_i^*, y_i, \theta, \sigma_u^2) p(x_i^* | y_i, \theta) p(y_i | \theta) \quad (7)$$

$$= \prod_i p(w_i | x_i^*, y_i, \sigma_u^2) p(x_i^* | y_i, \theta) p(y_i | \theta). \quad (8)$$

The first equality uses the rules of conditional probability; the key assumption (needed for the second equality) is expressed here by the density for w_i not depending on the parameters of the overall likelihood, $\theta = (\mu, \Sigma)$: $p(w_i | x_i^*, y_i, \theta, \sigma_u^2) = p(w_i | x_i^*, y_i, \sigma_u^2)$. Note that (8) is identical to the complete-data likelihood in MI model (3), with the additional factor, $p(w_i | x_i^*, y_i, \sigma_u^2)$, for the proxy’s data generation process. (To generate the observed-data likelihood in this case would of course require the analogous integration as in (4), which we omit here to save space. See Appendix A for a full description.)

³If a function $f(a, b)$ can be expressed as a family of probability densities for variable a given parameter b , $p(a|b)$, and a family of densities for variable b given parameter a , $p(b|a)$, so that $f(a, b) = p(a|b) = p(b|a)$, then $p(a|b)$ and $p(b|a)$ are said to be statistically dual.

⁴If the relationship between the underlying variable and its mismeasurement is completely unknown, a different approach may be required. For instance, structural equation modeling or factor analysis is sometimes appropriate if a large set of measures all capture some aspect of an unobserved concept.

We may sometimes wish to further simplify and assume normal error, $p(w_i|x_i^*, y_i, \theta, \sigma_u^2) = \mathcal{N}(x_i^*, \sigma_u^2)$, with a chosen or estimated value of the variance of the measurement error σ_u^2 . When σ_u^2 is small we have a reasonable precision in our estimate of the location of x_i^* . As the size of the measurement error grows, w_i reveals less information about the true value of x_i^* . Heuristically, as σ_u^2 becomes infinite, w_i tells us nothing, and we may as well discard it from the data set and treat it as missing. In this limiting case, where no information is directly observed about x_i^* , then $\lim_{\sigma_u^2 \rightarrow 0} p(w_i|x_i, \sigma_u^2)$ approaches a constant and the complete-data likelihood (8) becomes proportional to the model for missing data alone (3). This proves that the most commonly used model for missing data is a limiting special case of our approach.

2.3 Implementation

In a project designed for an unrelated purpose, [Honaker and King \(2010\)](#) propose a fast and computationally robust MI algorithm that allows for informative Bayesian priors on missing individual cell values. The algorithm is known as EMB, or EM with bootstrapping. They use this model to incorporate qualitative case-specific information about missing cells to improve imputations. To make it easy to implement our approach, we prove in [Appendix A](#) that the same algorithm can be used to estimate our model. The statistical duality property assumed there enables us to turn the data generation process for w_i into a prior on the unobserved value x_i^* , without changing the mathematical form of the density. For example, in the simple random normal error case, the data generation process for w_i is $\mathcal{N}(x_i^*, \sigma_u^2)$ but, using the property of statistical duality of the normal, this is equivalent to a prior density for the unobserved x_i^* , $\mathcal{N}(w_i, \sigma_u^2)$. This result shows that we can use the existing EMB algorithm.

This strategy also offers important intuitions: our approach can be interpreted as treating the proxy variables as informative, observation-level means (or functions of the means) in priors on the unobserved missing cell values. Our imputations of the missing values, then, will be precision-weighted combinations of the proxy variable and the predicted value from the conditional expectation (the regression of each variable on all others) using the missing data model. In addition, the parameters of this conditional expectation (computed from μ and Σ) are informed and updated by the priors on the individual cell values.

Under our overall approach, then, all cells in the data matrix with measurement error are replaced — overwritten in the data set, or *overimputed* in our terminology — with multiple overimputations that reflect our best guess and uncertainty in the location of the latent values of interest x_i^* . These overimputations include the information from our measurement error model, or equivalently the prior with mean set to the observed proxy variable measured with error, as well as all predictive information available in the observed variables in the data matrix. At the same time, all missing values are imputed. The same procedure is used to fill in multiple completed data sets; usually about five data sets is sufficient, but more may be necessary with large fractions of missing cells or high degrees of measurement error. Imputations and overimputations vary across the multiple completed data sets — with more variation when the predictive ability of the model is smaller and measurement error is greater — while correctly observed cell values remain constant.

Researchers create a collection of completed data sets once and then run as many analyses of these as desired. The same analysis model is applied to each of the completed (imputed and overimputed) data sets as if it were fully observed. A key point is that the analysis model need not be linear-normal even though the model for missing values and measurement error overimputation is ([Meng, 1994](#)). The researcher then applies the usual MI rules for combining these results (see [Appendix A](#)).

2.4 Monte Carlo Evidence

We now offer Monte Carlo evidence for MO, using a data generation process that would be difficult or impossible for most prior approaches to measurement error. We use two mismeasured variables, a non-normal dependent variable, scattered (but not completely random) missing data, and a nonlinear analysis model. The measurement error is independent random normal with variances that each account for 25% of the total variance for each proxy, meaning these are reasonably noisy measures. In doing so, we attempt to recreate a difficult but realistic political science data situation, with the addition of the true values so we can use them to validate the procedure.

In a real application, a researcher may only have a rough sense of the measurement error variances. We thus run our simulations assuming a range of levels for these variances, holding their true value fixed, to see how these differing assumptions affect estimation. (In the next section, we discuss how to interpret or estimate this variance.)

We generated proxies x and z for the true variables x^* and z^* , respectively, using a normal data generation process with the true variables as the mean and a variance equal to $\sigma_u^2 = \sigma_v^2 = 0.5$.⁵ At each combination of σ_u^2 and σ_v^2 , we calculate the mean square error (MSE) for the logit coefficients of the overimputed latent variables. We took the average MSE across these coefficients and present the results in Figure 1. On the left is the MSE surface with the error variances on the axes along the floor and MSE on vertical axis; the right graph shows the same information viewed from the top as a contour plot.

The figure shows that when we assume the absence of measurement error (i.e., $\sigma_u^2 = \sigma_v^2 = 0$), as most researchers do, we are left with high MSE values. As the assumed amount of measurement error grows, we see that the MO lowers the MSE smoothly. The MSE reaches a minimum at the true value of the measurement error variance (the gray dotted lines in the contour plot).⁶ Assuming values that are much too high also leads to larger MSEs, but the figure reveals one of the types of robustness of the MO procedure in that a large region exists where MSE is reduced relative to the naive model assuming no error, and so one need not know the measurement error variance except very generally. We discuss this issue further below.

Categorical variables measured with error While our imputation model assumes the data is drawn from a multivariate normal distribution, non-normal variables, such as categorical variables, can be included in the imputation and can even be overimputed for measurement error. It is well known in the multiple imputation literature that imputation via a normal model works well for categorical variables, and indeed as well as models designed especially for categorical variables and even when the analysis model is nonlinear (Schafer, 1997; Schafer and Olsen, 1998). Our own detailed simulations (not presented here) confirm that this standard result for MI also applies to MO.

⁵We let y_i , the dependent variable of the analysis model, follow a Bernoulli distribution with probability $\pi_i = 1/(1 + \exp(-X_i\beta))$, where $X_i = (x_i^*, z_i^*, s_i)'$ and $\beta = (-7, 1, 1, -1)$. We allow scattered missingness of a random 10% of the all cell values of y , x , and z when (the fully observed) s is greater than its mean. We created the true, latent data (x^*, z^*, s) by drawing from a multivariate normal with mean vector $(5, 3, 1)$ and covariance matrix $(1.5 \ 0.5 \ -0.2, 0.5 \ 1.5 \ -0.2, -0.2 \ -0.2 \ 0.5)$.

⁶In this simulation, the variance of the estimates is swamped by the squared bias of the estimates, so that any difference in the MSE is almost entirely due to bias, rather than efficiency. More succinctly, these plots are substantively similar if we replace MSE with bias.

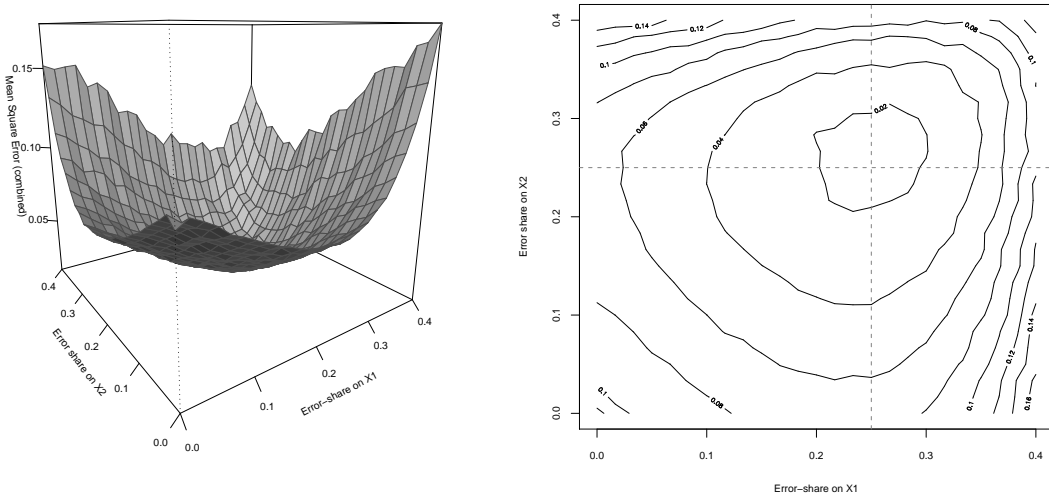


Figure 1: *On the left is a perspective plot of the mean square error of a logit analysis model estimates after multiple overimputation with various assumptions about the measurement error variance. The right shows the same information as a contour plot. Note that the axes here are the share of the observed variance due to measurement error which has a true value of 0.25, which is precisely where the MSE reaches a minimum.*

2.5 Comparison to Other Measurement Error Correction Techniques

Measurement error is a core threat to statistical analysis and many have proposed solutions to its problems. These solutions broadly fall into two camps: general-purpose methods and application-specific methods. General-purpose methods are easily implemented across a wide variety of models, while application-specific methods are closely tailored to a particular context. For more information about the various approaches to measurement error, see [Fuller \(1987\)](#) and [Carroll, Ruppert and Stefanski \(1995\)](#).

The first general-purpose method, *regression calibration* ([Carroll and Stefanski, 1990](#)), is similar in spirit to MO in that it replaces the observed, mismeasured variable with an estimate of the underlying unobserved variable and then performs the desired analysis on this “calibrated data.” In fact, one can think of MO as an combination of regression calibration and multiple imputation, two methods previously thought as distinct and in competition with one another ([Cole, Chu and Greenland, 2006](#)). As [White \(2006\)](#) points out, multiple imputation relies on validation data and ignores any replicate measures and regression calibration ignores any validation data completely. MO combines the best parts of each of these approaches by utilizing all information when it is available.

The easiest technique to implement is a simple method-of-moments estimator, which exploits the exact relationship between the biased estimates and the amount of measurement error present in the data. This estimator simply corrects a biased estimate of a linear regression coefficient by dividing it by the reliability ratio, $\sigma_{x^*}^2 / \sigma_w^2$. This technique depends heavily on the estimate of the measurement error variance and, in our simulations, has poor properties when this estimate is

incorrect. Further, the method-of-moment technique requires the analysis model to be linear.

Other general approaches to measurement error include simulation-extrapolation, or SIMEX, (Cook and Stefanski, 1994), and minimal-assumption bounds (Leamer, 1978; Klepper and Leamer, 1984; Black, Berger and Scott, 2000). These are both excellent approaches to measurement error, but they both have features that limit their general applicability. SIMEX is designed to simulate the effect of adding *additional* measurement error to a single mismeasured variable, then use these simulations to extrapolate back to the case with no measurement error. In situations with multiple mismeasured variables, SIMEX becomes harder to compute and more dependent on the extrapolation model. The minimal-assumption bounds are useful for specifying a range of parameter values consistent with a certain set of assumptions on the error model. Bounds typically require fewer assumptions than our multiple overimputation model, but obviously eliminate the possibility of point estimation. A comprehensive approach to measurement error could utilize minimal-assumption bounds with the overimputation bounds and point estimates below.

Structural equation modeling (SEM) attempts to solve the problem of measurement error in a different way.⁷ The goal of SEM is to find latent dimensions that could have generated a host of observed measures, while our goal is to rid a particular variable (or variables) of its measurement error. While discovering and measuring latent concepts is a useful and common task in political science, there are many cases in which we want to measure the effect of a specific variable and measurement error stands in the way. SEM would sweep that variable up into a larger construct and perhaps muddle the question at hand. Thus, MO is not so much a replacement for SEM, but rather an approach to a different set of substantive questions. Furthermore, MO is better equipped to handle gold-standard and validation data since it is unclear how to incorporate these into a structural equation modeling framework.

3 Specifying or Estimating the Measurement Error Variance

The measurement error variance is unidentified in our approach and all others, without some further data or assumptions (Stefanski, 2000). When little or no extra information is available, we show how to reparametrize σ_u^2 to a scale that is easier to understand and how we can provide bounds on the quantity of interest (Section 3.1). When replicated correlated proxies are available, we show how to estimate σ_u^2 directly (Section 3.2). And finally we show how to proceed when σ_u^2 varies over the data set or when gold standard observations are available (Section 3.3).

3.1 Interpretation through Reparametrization and Bounding

Section 2.4 shows that using the true measurement error variance σ_u^2 with MO will greatly reduce the bias and MSE relative to the usual procedure of making believe measurement error does not exist (which we refer to as the “denial” estimator). Moreover, in the simulation presented there (and in others we have run), the researcher needs only have a general sense of the value of these variances to greatly decrease the bias of the estimates. Of course, knowing the value of σ_u^2 (or σ_u) is not always obvious, especially on its given scale. In this section, we deal with this problem by reparameterizing it into a more intuitive quantity and then putting bounds on the ultimate quantity of interest.

The alternative parametrization we have found useful is the *proportion of the proxy variable’s observed variance due to measurement error*, which we denote $\rho = \frac{\sigma_u^2}{\sigma_x^2 + \sigma_u^2} = \frac{\sigma_u^2}{\sigma_w^2}$, where σ_w^2 , the

⁷Lee (2007) covers a number of Bayesian approaches to structural equation modeling, including some that take into consideration missing data.

variance of our proxy. This is easy to calculate directly if the proxy is observed for an entire variable (or at least more than one cell value). Thus, if we know the extent of the measurement error, we can create an estimated version of $\hat{\sigma}_u^2 = \rho \hat{\sigma}_w^2$ and substitute it for σ_u^2 in the complete-data likelihood (8).

In Figure 2, we present Monte Carlo simulations of how our method works when we alter our assumptions on the scale of ρ rather than σ_u^2 .⁸ More importantly, it shows how providing little or no information about the measurement error can bound the quantities of interest. Leamer (1978, pp. 238–243) showed that we can use a series of reverse regressions in order to bound the true coefficient without making any assumptions about the amount of measurement error. We compare these “minimal-assumption” bounds to the more model-based multiple overimputation bounds. The vertical axis in the left panel is the value of the coefficient of a regression of the overimputed w on y . The orange points and vertical lines are the estimates and 95% confidence intervals from overimputation as we change our assumption about ρ on the horizontal axis.

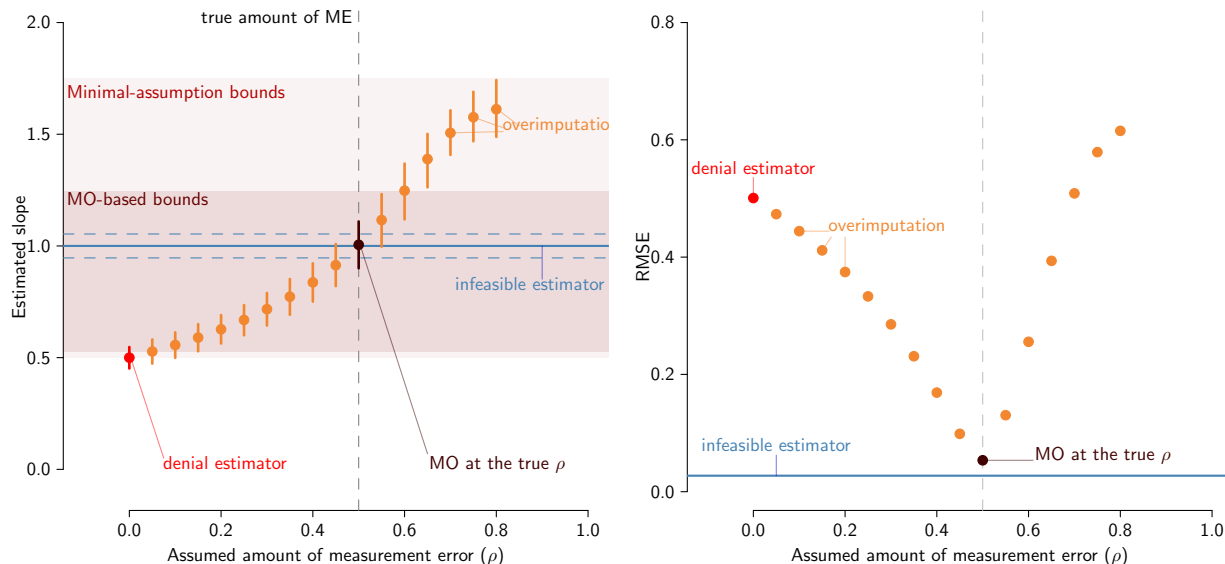


Figure 2: *Simulation results using the denial estimator (that assumes no measurement error, in red), the complete-data, infeasible estimator (in blue), and the MO estimator (in orange), with varying assumptions about the degree of measurement error. The MO estimator at the correct value of ρ is in dark red. The left panel shows estimates of the coefficients of interest along with confidence bands. In the background, the light tan area shows the minimal-assumption bounds and the dark tan region gives bounds assuming $\rho \in [0.05, 0.6]$. The right panel shows RMSE for the same range of estimates.*

We can see that the denial estimator, which treats w as if it were perfectly measured (in red), severely underestimates the effect calculated from the complete data (solid blue horizontal line), as

⁸For these simulations, we have $y_i = \beta x_i + \epsilon_i$ with $\beta = 1$, $\epsilon_i \sim \mathcal{N}(0, 1.5^2)$, $x_i^* \sim \mathcal{N}(5, 1)$, and $\sigma_u^2 = 1$. Thus, we have $\rho = 0.5$. We used sample sizes of 1,000 and 10,000 simulations

we might expect from the standard attenuation result. As we assume higher levels of ρ with MO, our estimates move smoothly toward the correct inference, hitting it right when ρ reaches its true value (denoted by the vertical dashed line). Increasing ρ after this point leads to overcorrections, but one needs to have a very bad estimate of ρ to make things worse than the denial estimator. The root mean square error leads to a similar conclusion and is thus also minimized at the correct value of ρ .

A crucial feature of MO is that it can be informative even if one has highly limited knowledge of the degree of measurement error. To illustrate this, the left panel of Figure 2 offers two sets of bounds on the quantity of interest, each based on different assumptions about ρ . We use the reverse regression technique of Leamer (1978) to generate minimal-assumption bounds, which make no assumptions about ρ (the mean of these bounds are in light tan). In practice, it would be hard to justify using a variable with over 80% measurement error, but even in this extreme situation the bounds on the quantity of interest do convey a great deal of information. They indicate, for example, that the denial estimator is an underestimate of the quantity of interest and almost surely within approximately the range [0.5,1.75]. Note that all of our MO estimates are within these bounds. In simulation in which we lowered the true ρ , we have found that even dramatic overestimates of ρ still lead to MO estimates that obey these bounds.⁹

Alternatively, we might consider making a more informative (and reasonable) assumption about ρ . Suppose that we know that there is some positive measurement error, but that less than 70% of the observed variance is due to measurement error. These are informative assumptions about ρ and allow MO to estimate bounds on the estimated coefficient. The result is that the bounds shrink (in dark tan, marked “MO-based”) closer around the truth. MO thus helps researchers learn about how various assumptions about measurement error affect their estimates.¹⁰ These bounds depend on the imputation model, but because MO allows for arbitrary patterns of mismeasurement, this bounding procedure is extraordinarily flexible. The MO-based bounding approach to measurement error shifts the burden from choosing the correct share of measurement error to choosing a range of plausible shares. Researchers may feel comfortable assuming away higher values of ρ since we may legitimately consider a variable with, say, 80% measurement error as a different variable entirely. The lower bound on ρ can often be close to 0 in order to allow for small amounts of measurement error.¹¹

This figure also highlights the dangers of incorrectly specifying ρ . As we assume that more of the proxy is measurement error, we eventually overshoot the true coefficient and begin to see increased MSE. Note, though, that there is again considerable robustness to incorrectly specifying the prior in this case. Any positive value ρ does better than the naive estimator until we assume that almost 70% of the proxy variance is due to error. This result will vary, of course, with the true degree of measurement error and the model under study.

⁹More generally, simulations run at various values of the true ρ lead to the same qualitative results as presented here. Underestimates of ρ lead to underestimates of the true slope and overestimate of ρ lead to overestimates of the true slope.

¹⁰If we use MO at all levels of ρ to generate the most assumption-free MO-based bounds possible, the bounds largely agree with the minimal-assumptions bounds.

¹¹These simulations also point to a use of MO as tool for sensitivity analysis. MO not only provides bounds on the quantities of interest, but can provide what the estimated quantity of interest would be under various assumptions about the amount of measurement error.

3.2 Estimation with Multiple Proxies

When multiple proxies (or “repeated measures”) of the same true variable are available, we can use relationships among them to provide point estimates of the required variances, and to set the priors in MO. For example, suppose for the same true variable x^* we have two unbiased proxies with normal errors that are independent after conditioning on x^* :

$$w_1 = x^* + u : u \sim N(0, \sigma_u^2), \quad w_2 = ax^* + b + v : v \sim N(0, (c\sigma_u)^2) \quad (9)$$

where a, b, c are unknown parameters, that rescale the additional proxy measure to a different range, mean, and different degree of measurement error. The covariances and correlations between these proxies can be solved as $E[\text{cov}(w_1, w_2)] = a \text{var}(x^*)$ and $E[\text{cor}(w_1, w_2)] = \gamma \text{var}(x^*)/\text{var}(w_1)$, where a is one of the scale parameters above, and γ is a ratio:

$$\gamma^2 = a^2 \frac{\text{var}(w_1)}{\text{var}(w_2)} = \frac{\text{var}(x^*) + \text{var}(u)}{\text{var}(x^*) + (c^2/a^2)\text{var}(u)} \quad (10)$$

If the measurement error is uncorrelated with x^* the variances decompose as $\sigma_u^2 = \sigma_{w_1}^2 - \sigma_{x^*}^2$. This leads to two feasible estimates of the error variances for setting priors. First:

$$s^2(u) = \text{var}(w_1) - \text{cov}(w_1, w_2) = \text{var}(w_1) - \text{var}(x^*) a \quad (11)$$

which is exactly correct when $a=1$, that is, when w_2 is on the same scale (with possibly differing intercept) as w_1 . Similarly,

$$s^2(u) = \text{var}(w_1)(1 - \text{cor}(w_1, w_2)) = \text{var}(w_1) - \text{var}(x^*) \gamma \quad (12)$$

which is exactly correct when $c = a \Leftrightarrow \gamma = 1$, that is, the second proxy has the same relative proportion of error as the original proxy.

3.3 Estimation with Heteroskedastic Measurement Error

In some applications, the amount of measurement error may vary across observations. Although most corrections in the literature ignore this possibility, it is easy to include in the MO framework, and doing so often makes estimation easier. To include this information, merely add a subscript i to the variance of the measurement error: $p(w_i|x_i^*, \sigma_{ui}^2) = \mathcal{N}(w_i|x_i^*, \sigma_{ui}^2)$. We consider two examples.

First, suppose the data include some observations measured with error and some without error. That is, for fully observed data points, let

$w_i = x_i^*$, or equivalently $\sigma_{ui}^2 = 0$. This implies that $p(w_i|x_i^*)$ drops out of the complete-data likelihood and x_i^* becomes an observed cell. Then the imputation model would only overimpute cell values measured with error and leave the “gold-standard” observations as is. If the other observations have a common error variance, σ_u^2 , then we can easily estimate this quantity, since the variance of the gold-standard observations is σ_x^2 and the mismeasured observations have variance $\sigma_x^2 + \sigma_u^2$. This leads to the feasible estimator,

$$\hat{\sigma}_u^2 = \hat{\sigma}_{mm}^2 - \hat{\sigma}_{gs}^2, \quad (13)$$

where $\hat{\sigma}_{mm}^2$ is the estimated variance of the mismeasured observations and $\hat{\sigma}_{gs}^2$ is the estimated variance of the gold-standard observations.¹²

¹²This logic assumes that the gold-standard observations are a random sample of the observations. When this assumption is implausible, we can use the reparameterization approach of Section 3.1.

As second special case of heteroskedastic measurement error, MO can handle situations where the variance is a linear function of another variable. That is, when $\sigma_{ui}^2 = rZ_i$, where Z_i is variable and r is the proportional constant relating the variable to the error variance. If we know r (or we can estimate it through variance function approaches), then we can easily incorporate this into the prior above using $p(w_i|x_i^*, r, Z_i) \sim \mathcal{N}(w_i|x_i^*, rZ_i)$.

4 Correlated Proxies

In this section and the next, we show how MO is robust to data problems that may occur in a large number of settings and applications. We show here how MO is robust to theoretical measurement dilemmas that occur regularly in political science data. In the sequel, we show more pragmatic robustness to a number of measurement applications in real data.

Until now we have assumed that measurement error is independent of all other variables. We now show how to relax this assumption. Many common techniques for treating measurement error make this strong assumption and are not robust when it is violated. For example, probably the most commonly implemented measurement error model (in the rare cases that a correction is attempted at all) is the classic errors-in-variables (EIV) model. We thus first briefly describe the EIV model to illustrate the strong assumptions required. The EIV model is also a natural point of comparison to MO, since both can be thought of as replacing mismeasured observations with predictions from auxiliary models.

4.1 The Foundation: The Errors-in-variables Model

As before, assume y_i and x_i^* are jointly normal with parameters as in (1). Suppose instead of x^* we have a set of proxy variables which are measures of x^* with some additional normally distributed random noise:

$$w_{i1} = x_i^* + u_i, \quad u_i \sim \mathcal{N}(0, \sigma_u^2); \quad (14)$$

$$w_{i2} = x_i^* + v_i, \quad v_i \sim \mathcal{N}(0, \sigma_v^2); \quad (15)$$

ordered such that $\sigma_u^2 < \sigma_v^2$, making w_1 the superior of the two proxies as it has less noise.

Suppose the true relationship is $y_i = \alpha x_i^* + \epsilon_{i1}$, and we instead use the best available proxy and estimate $y_i = \beta w_{i1} + \epsilon_{i2} = \beta(x_i^* + u_i) + \epsilon_{i2}$. We then get some degree of attenuation $0 < \beta < \alpha$ since the coefficient on u_i should be zero. This attenuation is shown in one example in the right of Figure 3 where the relationship between y and w_1 shown in red is weaker than the true relationship with x^* estimated in the left graph and copied in black on the right.

In this simple example we can calculate the expectation of this attenuation. The coefficient on w_{i1} will be

$$\mathbb{E}[\hat{\beta}_1] = \mathbb{E}\left[\frac{\sum_i (x_i^* + u_i - (\overline{x^* + u}))(y_i - \bar{y})}{\sum_i (x_i^* + u_i - (\overline{x^* + u}))^2}\right] = \frac{\sum_i (x_i^* - \bar{x}^*)(y_i - \bar{y})}{\sum_i (x_i^* - \bar{x}^*)^2 + \sigma_u^2}, \quad (16)$$

where $\overline{x^* + u}$ and \bar{x}^* are the sample means of w_1 and x^* , respectively. The last term in the denominator, σ_u^2 , causes this attenuation. If the variance of the measurement error is zero the term drops out and we get the correct estimate. As the measurement error increases, the ratio tends to zero.

The coefficients in the EIV approach can be estimated either directly or in two stages. A two-stage estimation procedure is the common framework to build intuition about the model and the role of the additional proxy measure. In this approach, we first obtain estimates of x^* from the relationship between the w 's since they only share x^* in common, $\hat{w}_{i1} = \hat{\gamma} w_{i2}$, and then use these

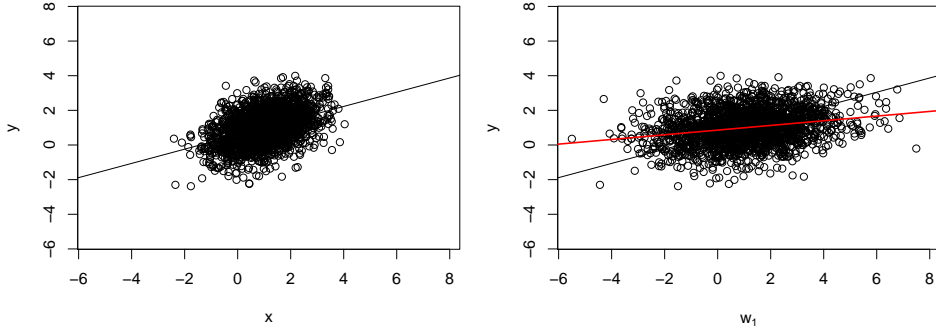


Figure 3: *On the left we see the true relationship between y and the latent x^* . When the mismeasured proxy w_1 is used instead, the estimated relationship (shown in red) is attenuated compared to the true relationship (shown in black in both graphs).*

predictions to estimate $y_i = \delta \hat{w}_{i1} + \epsilon_{i3}$, where now $\hat{\delta}$ is an unbiased estimate of α . The relationship between the two proxy variables is shown in the left of Figure 4, and the relationship between the first stage predicted values of w_1 and y is shown in green in the right figure. This coincides almost exactly with the true relationship still shown in black in this figure.

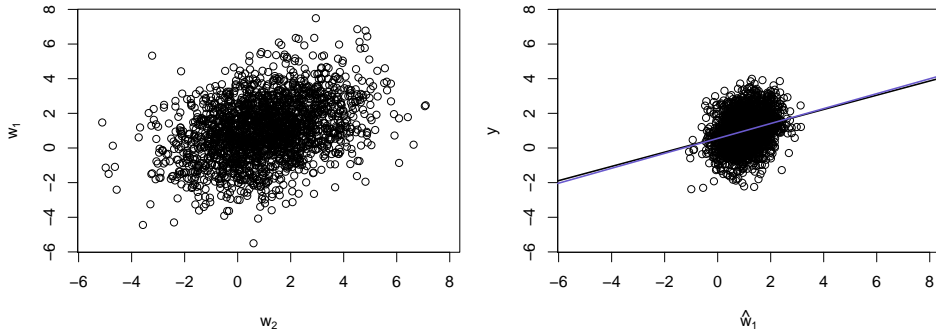


Figure 4: *The relationship between two mismeasured proxy variables (left), and the relationship between the predicted values from this model and y (right). The relationship here, shown in green, recovers the true relationship, shown in black.*

In Figure 5 we illustrate how the EIV model performs in data that meet its assumptions. The black distributions represent the distribution of coefficients estimated when the latent data x^* is available in a simulated data set of size 200.¹³ The naive regressions that do not account for measurement error are shown in red in both graphs. The coefficient on w_1 is attenuated towards zero (bottom panel). The estimated constant term is biased upwards to compensate (top panel). In each simulated data set, we use the EIV model (in green), and see that the distribution of estimated parameters using the proxies resembles the distribution using the latent data, although with slightly greater variance. Thus there is some small efficiency loss, but the EIV model clearly recovers unbiased estimates when its assumptions are met.

¹³In these simulations, $n = 200$, $(x^*, y) \sim \mathcal{N}(\mu, \Sigma)$, $\mu = (1, 1)$, $\Sigma = \begin{pmatrix} 1 & 0.4 \\ 0.4 & 1 \end{pmatrix}$, $\sigma_u^2 = 0.5$, $\sigma_v^2 = 0.5$.

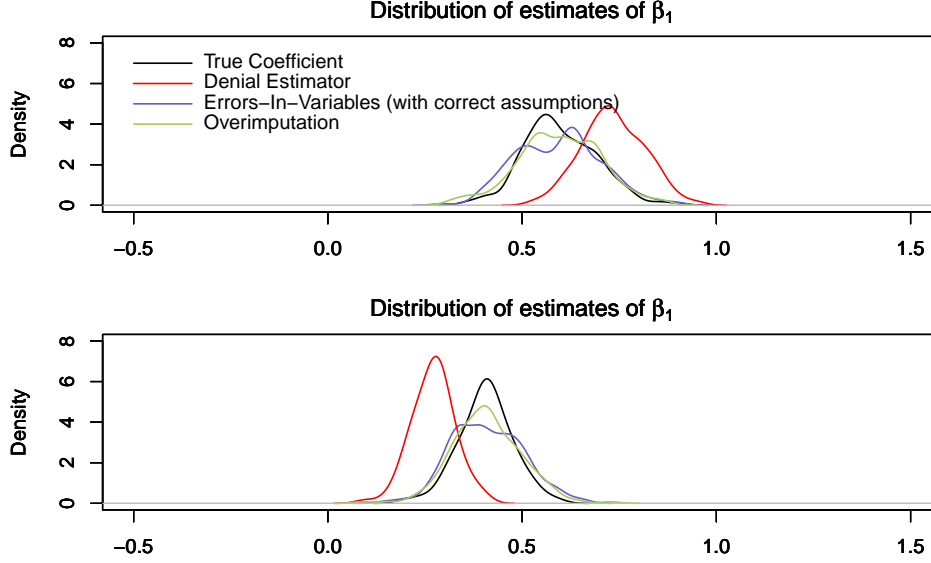


Figure 5: Coefficients estimated from variables with measurement error (shown in red) attenuate the effect of the independent variable towards zero, and also bias the constant in compensation. The estimates recovered from the EIV model (in green) recover the true distribution, but are of course less efficient (slightly higher variance) than the original latent data (in black).

We also run the MO on the same simulated data sets in which we ran the EIV model. The distribution of coefficients (which we present below) recovers the distribution that would have been estimated if the latent data had been available. Thus, in the simple setting where the assumptions of the EIV model are met, our approach performs equivalently.

4.2 Robustness to Violating Assumptions

If we think of the coefficient on x^* as the ratio of $\text{cov}(x^*, y)$ to $\text{var}(x^*)$, then the attenuation in equation (16) is being driven by the fact that $\text{var}(w_1) > \text{var}(x^*)$ because of the added measurement error. Therefore $\text{var}(w_1)$ is not a good estimate of $\text{var}(x^*)$, even though $\text{cov}(w_1, y)$ is a good measure of $\text{cov}(x^*, y)$. With this in mind, the numerically simpler—but equivalent—one stage approach to the errors-in-variables model has a useful intuition. We substitute $\text{cov}(w_1, w_2)$ as an estimate of $\text{var}(x^*)$ because w_1, w_2 only covary through x^* . Thus we have as our estimate of the relationship:¹⁴

$$\hat{\delta} = \frac{\sum_i (w_{i1} - \bar{w}_1)(y_i - \bar{y})}{\sum_i (w_{i1} - \bar{w}_1)(w_{i2} - \bar{w}_2)} = \frac{\sum_i (x_i^* - \bar{x}^*)(y_i - \bar{y}) + u_i(y_i - \bar{y})}{\sum_i (x_i^* - \bar{x}^*)^2 + u_i(x_i^* - \bar{x}^*) + v_i(x_i^* - \bar{x}^*) + u_i v_i}. \quad (17)$$

In order to recover the true relationship between x^* and y we need the last term in the numerator and the last three in the denominator to drop out of equation (17). To obtain a consistent estimate, then, EIV requires: (1) $E(u_i \cdot y_i) = 0$, (2) $E(u_i \cdot x_i^*) = 0$ and $E(v_i \cdot x_i^*) = 0$, and (3) $E(u_i \cdot v_i) = 0$. Indeed, when these conditions are not met the resulting bias in the EIV correction can easily be

¹⁴In a multivariate setting this becomes $\hat{\delta} = (W_1' W_2)^{-1} W_1' Y$ where W_j is the set of regressors using the j -th proxy measure for x^* .

larger than the original bias caused by measurement error. However, as we now show in the following three subsections, MO is robust to violations of all but the last condition.

4.2.1 Measurement error correlated with y

The first of the conditions for EIV to work is that the measurement error is unrelated to the observed dependent variable. As an example of this problem, we might think that infant mortality is related to international aid because donors want to reduce child deaths. If countries receiving aid are intentionally underreporting infant mortality, to try to convince donors the aid is working, then the measurement error in infant mortality is negatively correlated with the dependent variable, foreign aid. If instead countries searching for aid are intentionally overreporting infant mortality as a stimulus for receiving aid, then measurement error is positively correlated with the dependent variable. Both scenarios are conceivable. This problem with the errors-in-variables approach is well known, because the errors-in-variables model has an instrumental variables framework, and this is equivalent to the problem of the instrument being exogenous of y in the more common usage of instrumental variables as a treatment for endogeneity.

In Figure 6(a) we demonstrate this bias with simulated data.¹⁵ The violet densities show the distribution of parameter estimates when there is negative correlation of 0.1 (dashed) and 0.3 (solid) between the measurement error and the dependent variable. In the latter case the bias in the correction has exceeded the original bias from measurement error, still depicted in red. The blue densities show that positive correlation of the errors create bias of similar magnitude in the opposite direction. Again, the size of the bias can be greater than that originally produced by the measurement error we were attempting to correct. Moreover, the common belief with measurement issues is that any resulting bias attenuates the coefficients so that estimates are at least conservative, however, here we see that the bias in the error-in-variables approach can actually exaggerate the magnitude of the effect.

We now analyze the same simulated data sets with MO. To apply the MO model, we estimate the measurement error variance from the correlation between the two proxies and leave the mean set to the better proxy. As Figure 6(b) indicates, MO recovers the distribution of coefficients for each of the data generation processes: The green line represents the distribution when there is no correlation. The violet line represents the distribution when there is positive correlation. The blue line (barely visible under the other two) represents the distribution with negative correlation. All three distributions are close to each other and close to the true distribution in black using the latent data.

4.2.2 Measurement error correlated with x^*

The second requirement of the EIV model is that the measurement error is independent of the latent variable. If, for example, we believe that income is poorly measured, and wealthier respondents feel pressure to underreport their income while poorer respondents feel pressure to overreport, then the measurement error can be correlated with the latent variable.

In Figure 7(a) we demonstrate the bias this produces in EIV. Here, the error in w_2 is correlated with the latent x^* .¹⁶ The biases are in the opposite directions as when the correlation is with y ,

¹⁵In these simulations, similar to previous, $n = 200$, $(x^*, y, u, v) \sim \mathcal{N}(\mu, \Sigma)$, $\mu = (1, 1, 0, 0)$, $\Sigma = \begin{pmatrix} 1 & 0.4 & 0 & 0 \\ 0.4 & 1 & 0 & 0 \\ \rho & 0 & \sigma_u^2 & 0 \\ 0 & 0 & \rho & \sigma_v^2 \end{pmatrix}$, $\sigma_u^2 = 0.5$, $\sigma_v^2 = 0.5$. Thus, the measurement errors are drawn at the same time as x^* and y with mean zero. While ρ allows the error, v , to covary with y , and across the simulations it is set as one of $\rho \in \{-0.3, -0.1, 0.1, 0.3\}$. The observed mismeasured variables are constructed as $w_1 = x^* + u$, $w_2 = x^* + v$.

¹⁶Similar to the construction of the last simulations, we set $n = 200$, $(x^*, y, u, v) \sim \mathcal{N}(\mu, \Sigma)$, $\mu = (1, 1, 0, 0)$,

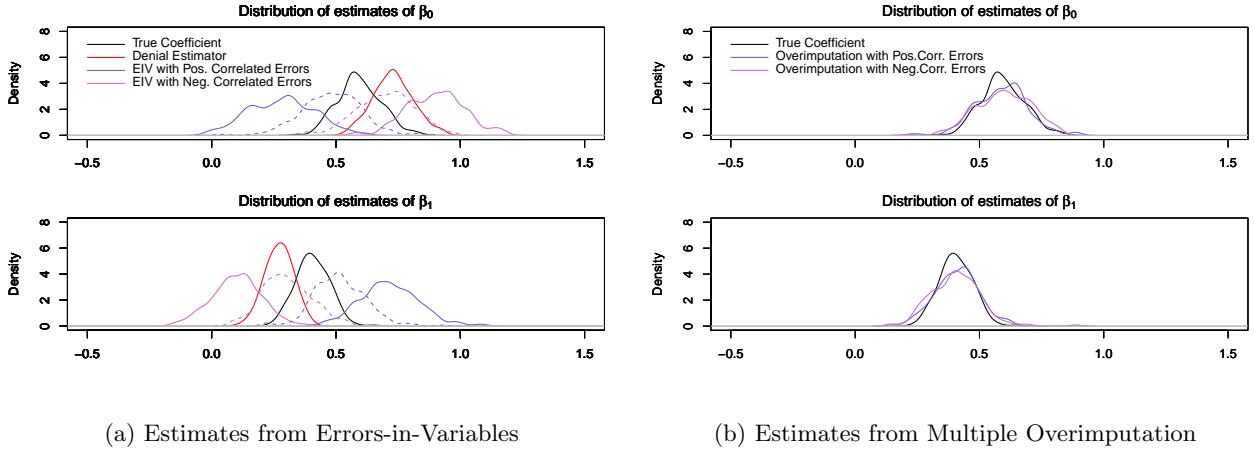


Figure 6: *With data generated so that proxy variables are correlated with the dependent variable, EIV (left graphs) gives biased estimates whereas MO (right graphs) gives robust, unbiased estimates.*

although lesser in magnitude. Errors positively correlated with x^* lead to attenuated coefficients, and negatively correlated errors lead to overstated coefficients, as shown by the blue and violet distributions in Figure 7(a), respectively. Dashed lines are the result of small levels of correlations (± 0.1) and the solid lines a greater degree (± 0.3).

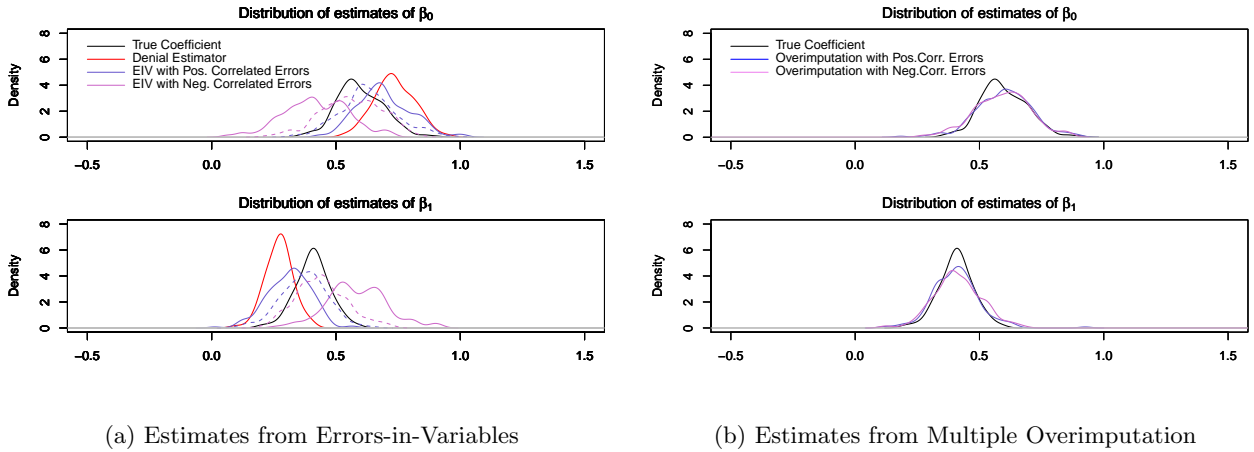


Figure 7: *Here we show the estimates when the error in the instrument w_2 is correlated with the latent variable x^* . Positive (blue distributions) correlation leads to attenuated estimated effects in the errors-in-variables framework, and negative (violet) correlation exaggerates the effect, as shown in the left. The MO estimates show no bias.*

$\Sigma = (1 \ 0.4 \ 0 \ \rho, 0.4 \ 1 \ 0 \ 0, 0 \ 0 \ \sigma_u^2 \ 0, \rho \ 0 \ 0 \ \sigma_v^2)$, $\sigma_u^2 = 0.5$, $\sigma_v^2 = 0.5$ and sequencing $\rho \in \{-0.3, -0.1, 0.1, 0.3\}$ across sets of simulations.

The coefficients resulting from MO, with measurement error variance estimated from the correlation between the proxies, are contrasted in Figure 7(b). All the distributions recover the same parameters. Because they sit on top of each other, only the simulations with the greatest correlation (± 0.3) are shown. For both parameters, and for both positive and negative correlation, the MO estimates reveal no bias.

4.2.3 Measurement errors that covary across proxies

The final condition requires the errors in the proxies be uncorrelated. If all the alternate measures of the latent variable have the same error process then the additional measures provide no additional information. For example, if we believe GDP is poorly measured, it is not enough to find two alternate measures of GDP; we also need to know that those sources are not making the same errors in their assumptions, propagating the same errors from the same raw sources, or contaminating each other’s measure by each making sure their estimates are in line with other published estimates. To the extent the errors in the alternate measures are correlated, then σ_{uv} will attenuate the estimate in the same fashion as σ_u^2 did originally.

Thus, we now simulate data where the measurement errors across alternate proxies are correlated.¹⁷ Figure 8(a) shows positively (negatively) correlated errors lead to bias in the EIV estimates that are in the same (opposite) direction as the original measurement error. Intuitively, if the errors are perfectly correlated, both the original proxy, and the alternate proxy would be the exact same variable, and thus all of the original measurement error would return. Importantly, what we see is that this is a limitation of the data that MO cannot overcome when cell level priors are directly created from the observed data. As alternate proxies contain correlated errors, identifying the amount of the variance in the proxies by the correlation of the measures is misleading. Positive or negative correlation in the measurement errors leads respectively to under or over estimation of the amount of measurement error in the data, directly biasing results as in EIV. When cell priors are set by the use of auxiliary proxies, our method continues to require the measurement errors (although not the indicates themselves of course) be uncorrelated across alternate measures, so that it is possible to consistently estimate the degree of measurement error present in the data.

Even in this most difficult of settings, MO remains robust. In another set of simulations, we compare how various estimators perform when both proxies are correlated with y . Allowing these simulations to vary the amount of correlation gives an indication of how various estimators perform in this difficult situation.¹⁸ Figure 9 shows that MO outperforms EIV at every level of this correlation. When the dependence between the error and y is weak, MO almost matches its zero-correlation minimum. Thus, MO appears to be robust to even moderate violations of the these assumptions, especially when compared with other measurement error approaches. Interestingly, the denial estimator can perform better than all estimators under certain conditions, yet these conditions depend heavily on the parameters of the data. If we change the effect of x^* on y from negative to positive, the performance of the denial estimator reverses itself. Since we obviously have little knowledge about all of these parameters *a priori*, the denial estimator is of little use.

¹⁷Here we set $n = 200$, $(x^*, y, u, v) \sim \mathcal{N}(\mu, \Sigma)$, $\mu = (1, 1, 0, 0)$, $\Sigma = (1 \ 0.4 \ 0 \ 0, 0.4 \ 1 \ 0 \ 0, 0 \ 0 \ \sigma_u^2 \ \rho, 0 \ 0 \ \rho \ \sigma_v^2)$, $\sigma_u^2 = 0.5$, $\sigma_v^2 = 0.5$ and sequencing $\rho \in \{-0.3, -0.1, 0.1, 0.3\}$ across sets of simulations.

¹⁸These simulations follow the pattern above except they include a perfectly measured covariate, z , which determines which observations are selected for mismeasurement. Thus, we have $(x^*, y, z, u, v) \sim \mathcal{N}(\mu, \Sigma)$, with $\mu = (1, 1, -1, 0, 0)$ and $\Sigma = (1 \ \sigma_{xy} \ -0.4 \ 0 \ 0, \sigma_{xy} \ 1 \ -0.2 \ \rho\sigma_u \ \rho\sigma_v, -0.4 \ -0.2 \ 1 \ 0 \ 0 \ 0, \rho\sigma_u \ 0 \ \sigma_u^2 \ 0, 0 \ \rho\sigma_v \ 0 \ 0 \ \sigma_v^2)$ with $\sigma_u^2 = 0.5$ and $\sigma_v^2 = 0.75$. We ran simulations at both $\sigma_{xy} = 0.4$ and $\sigma_{xy} = -0.4$. Each observation had probability $\pi_i = (1 + e^{3.5+2z})^{-1}$, which has a mean of 0.25. We used the multiple proxies approach to estimating the measurement

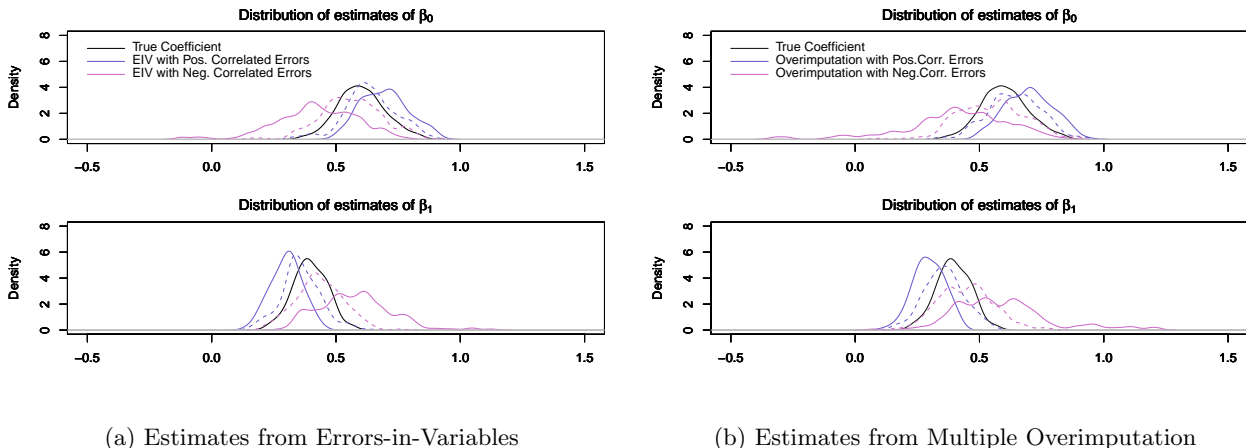


Figure 8: *With data generated so that proxy variables have measurement error correlated with each other (so that new information is not available with measures) both EIV (left graphs) and MO (right graphs) gives biased estimates.*

Since there are gold-standard data in these simulations, we can also investigate the performance of simply discarding the mismeasured data and running MI. As expected, MI is unaffected by the degree of correlation since it disregards the correlated proxies. Yet these proxies have *some* information when the correlation is around zero and, due to this, MO outperforms MI in this region. As the correlation increases, though, it becomes clear that simply imputing the mismeasured cells has more desirable properties. Of course, with such high correlation, we might wonder if these are actually proxies in our data or simply new variables.

These simulations give key insights into how we should handle data measured with error. MO is appropriate when we have a variable that we can reasonably describe as a proxy—that is, having roughly uncorrelated, mean-zero error. Even if these assumptions fail to hold exactly, MO retains its desirable properties. In situations where we suspect that the measurement error on all of our proxies has moderate correlation with other variables in the data, it may be wiser to treat the mismeasurement as missingness and use multiple imputation. Of course, this approach assumes there exist gold-standard data, which may be scarce.

5 Empirical Applications of Overimputation

We offer three separate illustrations of the use of multiple overimputation.

5.1 Unemployment and Presidential Approval

To first show a practical example of the differences between our MO solution and the more common errors-in-variables (EIV) approach, we construct a measurement error process from a natural source of existing data.

It is often the case, particularly in yearly-aggregated cross-national data, that key independent variables are not measured or available at the correct point in time the model requires. Some economic and demographic statistics are only collected at intervals, sometimes as rarely as once

error. For EIV, we use applied the model as if the entire variable were mismeasured.

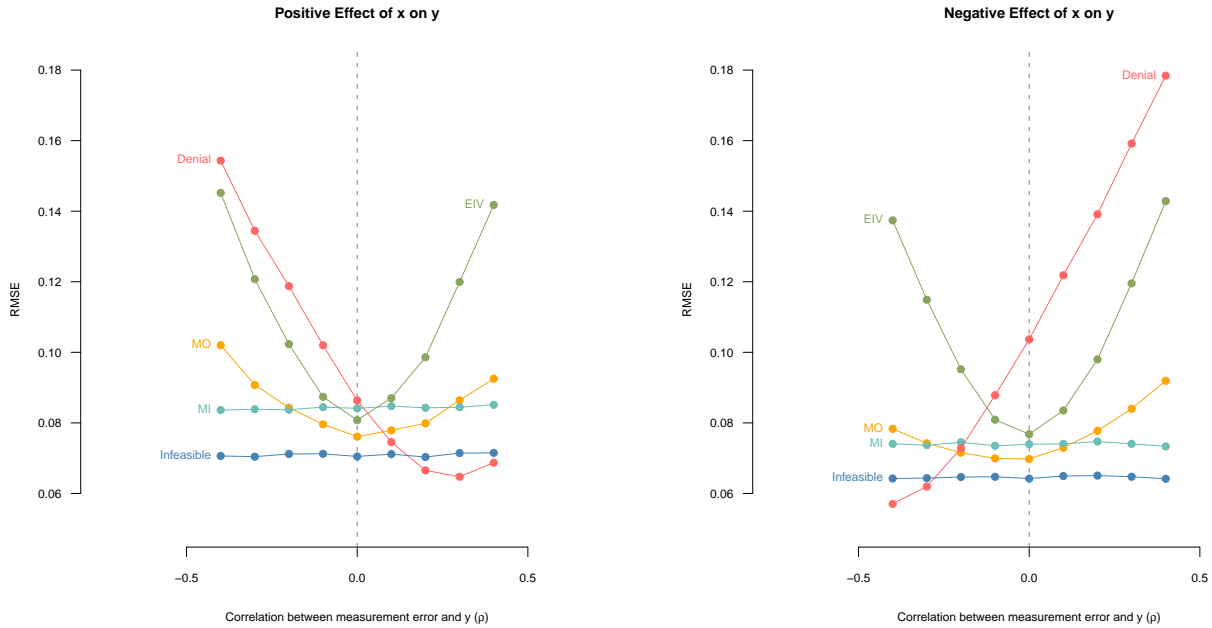


Figure 9: *Root mean squared error for various estimators with data generated so that each proxy variable has measurement error correlated with the dependent variable. On the left, x^* has a positive relationship with y and on the right, it has a negative effect. Note that both EIV (green) and MO (orange) perform worse as the correlation moves away from zero, but MO always performs better. The denial estimator can actually perform well in certain situations, yet this depends heavily on the direction of the relationship. Both the infeasible estimator and MI are unaffected by the amount of correlation.*

every five or ten years. The available data, measured at the wrong period in time, is often used as a reasonable proxy for the variable’s value in the desired point in time, with the understanding that there is measurement error which increases the more distant the available data is from the analyst’s desired time period.

We mimic this process in actual data by intentionally selecting a covariate at increasing distance in time from the correct location, as a natural demonstration of our method in real data. In our example, we are interested in the relationship between the level of unemployment and the level of Presidential approval in the US, for which there is rich data of both series over time.¹⁹

We assume that the correct relationship is approximately contemporaneous. That is, the current level of unemployment is directly related to the President’s approval rating. Unemployment moves over time, so the further in time our measure of unemployment is from the present moment, the weaker the proxy for the present level of unemployment, and the more the measurement error in the available data. We iteratively consider repeated models where the measurement of unemployment

¹⁹Monthly national unemployment is taken from the Bureau of Labor Statistics, labor force series. Presidential approval is from the Gallup historical series, aggregated to the monthly level. We use data from 1971 to 2011. We use the last three years of each four-year Presidential term of office, to avoid approval levels within the “honeymoon” period, without adding controls into the model. We added a monthly indicator for cumulative time in office, but this only slightly strengthened these results, and so we leave the presentation as the simplest, bivariate relationship.

we use grows one additional month further from the present time.

The EIV approach relies on multiple proxies. To naturally create two proxies with increasing levels of measurement error, we use a measure of unemployment k -months before the dependent variable, and k -months after. That is, if we are attempting to explain current approval, we assume that the unemployment k months in the past (the k -lag) and k months in the future (the k -lead) are proxies for the current level of unemployment, which we assume is unavailable to our analyst. As k increases, the measures of unemployment may have drifted increasingly far from the present unemployment level, so both proxies employed have increased measurement error. We use these same two proxies in each of our MO models (as previously described in sections 2.4 and 3.2).

We estimate the relationship between unemployment and Presidential approval using our MO framework, and the common EIV approach, while using pairs of proxies that are from 1 to 12 months away from the present. We also estimate the relationship between approval and all individual lags and leads of unemployment; these give us all the possible denial estimators, with all the available proxies. In figure 10, these coefficients from the denial estimators, are shown in red, where the red bar represents the 95 percent confidence interval for the coefficient and the center point the estimated value. The x -axis measures how many months in time the covariate used in the model is from the month of the dependent variable. Positive values of x use proxies that are measured later than desired, negative values are measured too far in the past. The correct, contemporaneous relationship between unemployment and approval is in the center of this series (when x is 0) marked in black.

The EIV estimates are shown in blue. We see that with increased measurement error in the available proxies, the EIV estimates rapidly deteriorate. When the proxies for current unemployment are four months from the value of the dependent variable, the EIV estimates of the relationship are 1.40 times the true value, that is, they are biased by 40 percent. At six months the confidence interval no longer contains the true value and the bias is 98 percent. With unemployment measured at a one year gap, EIV returns an estimate 6.5 times the correct value. The MO estimates, however, are comparatively robust across these proxies. The confidence intervals expand gradually as the proxies contain less information and more measurement error. The bias is always moderate, between +16% and -12% and always clearly superior to the denial estimator, until the proxies are fully twelve months distant from the dependent variable. Finally, at one year's distance, the MO estimates are biased by 46 percent, while the denial estimator is biased at -48 percent.

A partial explanation can be understood from our previous results. In periods where unemployment trends upwards (or downwards) the k -month lag and the k -month lead of unemployment will generally have opposite signed measurement error. So the measurement errors in the proxies will be negatively correlated. We saw in figure 9 that this is a problem for both models, but that MO is much more robust to this violation than the EIV model.

We could do better than shown; we do not propose that this is the best possible model for covariates that are mismeasured in time. Adding other covariates into the imputation model could increase the efficiency of the overimputations. Averaging the two proxies would give an interpolation that might be a superior proxy to those used, and we demonstrate an application of averaging across proxies in MO in section 5.3. Moreover, in many applications, if there is periodic missingness over time in a variable, the best approach might be to impute all the missing values in the series with an imputation model built for time-series cross-sectional data, such as developed in Honaker and King (2010); this reinforces the main thesis of our argument, that measurement error and missing data are fundamentally the same problem. Rather, what we have shown in this

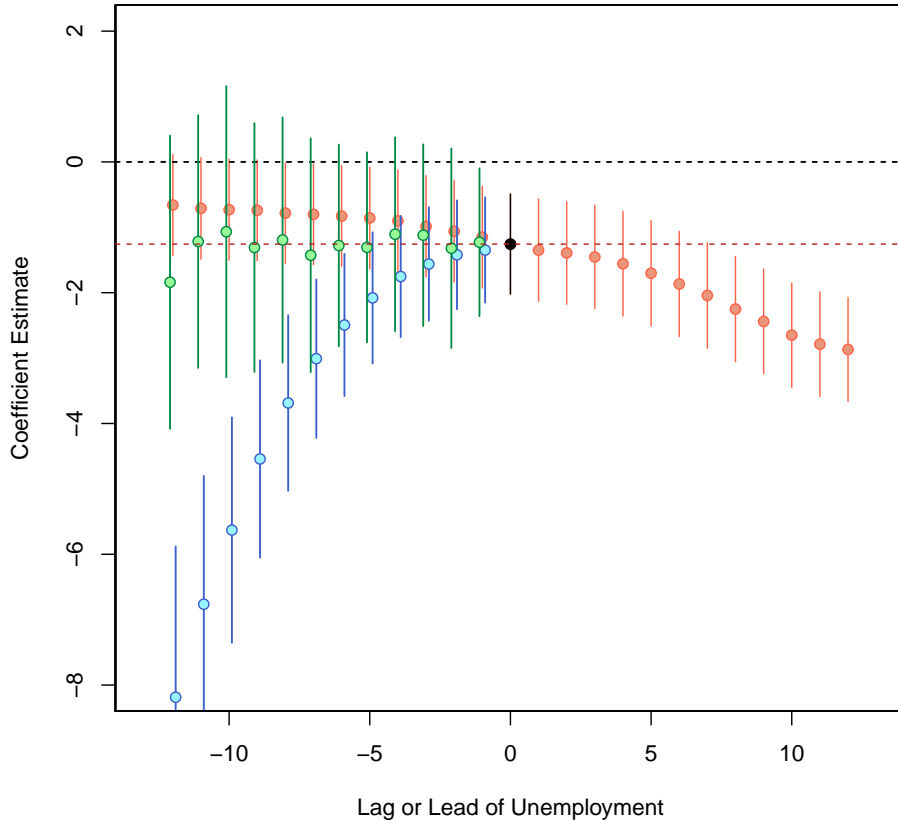


Figure 10: *An experiment in measurement error, in the estimation of the relationship between unemployment and Presidential approval, whose true, contemporaneous value is shown in black. The blue confidence intervals represent EIV estimates of this relationship using proxies of unemployment measured increasingly distant in time. The EIV estimates fail quickly as the proxies move away from month zero. The green estimates show the robust MO estimates of the relationship. These are consistently superior to the red estimates which show the denial estimators using the unemployment rates mismeasured in time, ignoring the measurement error.*

example, is that in naturally occurring data, in a simple research question, where we can witness and control a measurement error process, the most commonly used model for measurement error fails catastrophically, and our framework is highly robust to even a difficult situation with proxies with negatively correlated errors.

5.2 Social Ties and Opinion Formation

Having looked at an example where other measurement error methodologies are available, we turn to a conceptually simple example that poses a number of difficult methodological hazards. We examine here the small area estimation challenges faced in the work of [Huckfeldt, Plutzer and Sprague \(1993\)](#). The authors are interested in the social ties that shape attitudes on abortion.

In particular they are interested in contrasting how differing networks and contexts, such as the neighborhood one lives in, and the church you participate in, shape political attitudes.

Seventeen neighbourhoods were chosen in South Bend, Indiana, and 1500 individuals randomly sampled across these neighborhoods. This particular analysis is restricted to the set of people who stated they belonged to a church and could name it. The question of interest is what shapes abortion opinions, the individual level variables common in random survey designs (income, education, party identification), or the social experiences and opinions of the groups and contexts the respondent participates in. Abortion attitudes are measured by a six point scale summing how many times you respond that abortion should be legal in a set of six scenarios.

The key variable explaining abortion opinion is how liberal or conservative are the attitudes toward abortion at the church or parish to which you belong. This is measured by averaging over the abortion attitudes of *all the other people in the survey* who state they go to the same named church or parish as you mention. Obviously, in a random sample, even geographically localized, this is going to be an average over a small number of respondents. The median number is 6.²⁰ The number tends to be smaller among Protestants who have typically smaller congregations than Catholics who participate in generally larger parishes. In either case, the church positions are measured with a high degree of measurement error because the sample size within any church is small. This is a classic “small area estimation” problem. Here we know the sample size, mean and standard deviation of the sampled opinions from within any parish that lead to the construction of each observation of this variable.

This is an example of a variable with measurement error, where there are no other proxies available, but we can analytically calculate the observation level priors. For any individual, i , if c_i is the set of n_i respondents who belong to i 's church (not including i), the priors are given by:

$$p(w_i|x_i^*) = \mathcal{N}(\bar{c}_i, sd(c_i)/\sqrt{n_i}) \quad (18)$$

where the $sd(c_i)$ can be calculated directly as the standard deviation within a group if n_i is generally large, or we can estimate this with the within-group variance, across all groups, as $1/n\sqrt{\sum_i(w_{ij} - \bar{w}_j)^2}$.

This is clearly a case where the measurement error is heteroskedastic; different respondents will have different numbers of fellow parishioners included in the survey. Moreover this degree of measurement error is not itself random as Catholics—who tend to have more conservative attitudes towards abortion—are from generally larger parishes, thus their church attitude will be measured with less error than Protestants who will have greater measurement error in their church attitude while being more liberal. The direction of the measurement error is still random, but the variance in the measurement error is correlated with the dependent variable. Furthermore while we have focused on the measurement error in the church attitude variable, the authors are interested in distinguishing the socializing forces of church and community, and the same small area estimation problem applies to measuring the average abortion position of the community a respondent lives in. Obviously though, the sample size within any of the 17 neighborhoods is much larger than for the parishes and thus the degree of measurement error is smaller in this variable.²¹ Finally, as it

²⁰The mean is 10.2 with an interquartile range of 3 to 20.

²¹Within parishes, the median sample size is 6, and only 6 percent of observations have at least thirty observed responses to the abortion scale among fellow congregants in their parish. Thus we use the small sample, within-group estimate for the standard deviations, pooling variance across parishes. Within neighborhoods, however, the median sample size is 47, fully 95 percent of observations have thirty or more respondents in their neighborhood, and so we estimate the standard deviation in each neighborhood directly from only the observations in that neighborhood.

is survey data, there is a variety of missing data across the variables due to nonresponse. Despite all these complicating factors this is a set up well suited to our method. The priors are analytically tractable, the heterogeneous nature of the measurement error poses no problems because we set priors individually for every cell, and measurement error across different variables poses no problems because the strength of the MI framework is handling different patterns of missingness.²²

	Naive Regression Model	MO Measurement Only	MO Measurement and Missingness
Constant	3.38** (1.12)	-0.39 (2.09)	-1.68 (1.89)
Education	0.17** (0.04)	0.15** (0.04)	0.14** (0.04)
Income	-0.05 (0.05)	-0.04 (0.05)	-0.00 (0.05)
Party ID	-0.10* (0.04)	-0.11* (0.04)	-0.08* (0.04)
Church Attendance	-0.57** (0.07)	-0.56** (0.07)	-0.51** (0.06)
Mean Neighborhood Attitude	0.11 (0.21)	0.84 (0.55)	0.99* (0.48)
Mean Parish Attitude	0.13° (0.07)	0.43* (0.19)	0.48** (0.18)
Catholic	-0.48* (0.27)	-0.23 (0.23)	-0.02 (0.21)
n	357	521	772

** : $p < 0.01$, * : $p < 0.05$, ° : $p < 0.10$

Table 1: *Mean Parish Attitudes are estimated by the average of across those other respondents in the survey who attend the same church. These “small area estimates” with small sample size and large standard errors have an analytically calculable measurement error. Without accounting for measurement error there is no discernable effect (column 1) but after applying MO (column 2) to correct for measurement error, we see that the average opinion in a respondent’s congregation predicts their own attitude towards abortion.*

We replicate the final model in table 2 of [Huckfeldt, Plutzer and Sprague \(1993\)](#). Our table 1 shows the results of the naive regression subject to measurement error in the first column. Parish attitudes have no effect on the abortion opinions of churchgoers, but individual-level variables, such as education and party identification and the frequency with which the respondent attends church predict abortion attitudes. The act of going to church seems to decrease the degree of support for legalized abortion, but the beliefs of the fellow congregants in that church have no social effect or

²²For additional work on small area estimation from an multiple overimputation framework, see [Honaker and Plutzer \(2011\)](#). In particular, there are additional possibly efficiency gains from treating the errors within individuals in the same church or community as correlated, as well as bringing in auxiliary Census data, and this work shows to approach this with two levels of imputations at both the individual and aggregated level.

pressure. Interestingly, Catholics appear to be different from non-Catholics, with around a half point less support for abortion on a six point scale.

The second column applies our model for measurement error, determining the observation-level priors for neighborhood and parish attitudes analytically as a function of the sample of respondents in that neighborhood and parish. Only the complete observations are used in column two, so differences with the original model are due to corrections of the measurement error in the small area estimates. We see now the effect of social ties. Respondents that go to churches where the support for legal abortion is higher, themselves have greater support for legal abortion. This may be because abortion is a moral issue that can be shaped in the church context and influenced by coreligionists, or this maybe a form of self selection of church attendance to churches that agree on the abortion issue. With either interpretation, this tie between the attitudes in the network of the respondent’s church and the respondent’s own personal attitude disappears due to measurement error caused by the inevitable small samples of parishioners in any individual church.

Of course our MO approach can simultaneously correct for missing data also, and multiple imputation of non-response increases by one half the number of observations available in this regression.²³ Most of the same results remain, while the standard errors shrink due to the increase in sample size. Similar to the parish variable, local neighborhood attitudes are now statistically significant at the ninety-five percent level. The one variable that changes noticeably is the dummy variable for Catholics which is halved in effect and no longer statistically significant once we correct for measurement error, and the rest of the effect disappears when we impute missing data.²⁴ In all, MO strengthens the author’s findings, finds support for their theories in this particular model where previously there was no result, and aligns this regression with the other models presented in their work.

5.3 The Effect of Political Preferences on Vote Choice

Ansolabehere, Rodden and Snyder (2008) show that the causal effect of opinions about economic policy on vote choice is much stronger than previously estimated (but consistent with what one would expect) via a simple alternative method of removing measurement error: averaging many multiple measures of the same concept. Although the data requirements make approach only occasionally applicable, it is powerful, when possible, and instructive. They consider $K = 34$ survey items $\{w_1, w_2, \dots, w_K\}$, all taken to be imperfect indicators of an unobserved variable, x , and assume common measurement error variance σ_x^2 . That is, $w_{ik} = x_i + u_{ik}$ for each i , where $E[u_{ik}] = 0$ and $E[u_{ik}^2] = \sigma_k^2$. While any individual measure has variance $\sigma_x^2 + \sigma_k^2$, the average of the measures, $\bar{w}_i = \frac{1}{K} \sum_{k=1}^K w_{ik}$ has variance $\sigma_x^2 + \bar{\sigma}^2/K$, where $\bar{\sigma}^2 = \frac{1}{K} \sum_{k=1}^K \sigma_k^2$ is the average measurement error variance among the items. If all of the measures have similar amounts of measurement error, then the average of the items will have far lower levels of measurement error than any single item. Furthermore, the effect of measurement error will decrease as K increases.

We now show that in the more usual situation where researchers have access to one or only a few measures of their key concepts, MO can still recover reliable estimates because it makes more efficient use of the data and available assumptions. It also provides more information by

²³Forty-seven percent of this missingness is due to respondents who answer some, but not all, of the abortion scenarios that constitute the abortion scale. Knowing the pattern of answers to the other completed abortion questions, as well as the other control variables in the model, help predict these missing responses.

²⁴Catholics are still less likely to support abortion (a mean support of 3.1 compared to 3.7 for non-Catholics), but this difference is explained by variables controlled for in the model such as individual demographics and the social ties of Catholic churches which have lower mean parish attitudes than non-Catholic churches.

enabling one to avoid the assumption that all available measures are indicators of exactly the same underlying concept.

To illustrate these features, we reanalyze [Ansolabehere, Rodden and Snyder \(2008\)](#) with their data from the American National Election Study. Using their general approach, we find that a one standard deviation increase in economic conservatism leads to an 0.24 increase in the probability of voting for Bob Dole.

We then perform MO using only two of the thirty-four variables. To avoid cherry picking results, we reran the analysis using all possible subsets of two variables chosen from the available 34. For each of these pairs, we overimputed the first variable, using the second as a proxy (see Section 3.2). We then estimate the effect of that overimputed variable on voting for Bob Dole using a probit model. We compare this method with simply taking the pairwise averages and using them as the measure of economic policy preferences. These approaches mimic a common situation in political science when researchers have access to relatively few variables.

Figure 11 shows the relationship between the two estimates. Each column represents the average of the estimated effects for one measure, averaged across all its pairs. Note that for every variable, MO estimates a larger effect than does averaging, as can be seen by the positive slope of every line. The “gold-standard” estimate suggested by [Ansolabehere, Rodden and Snyder \(2008\)](#) is well above the any of the pairwise averaging estimates, but it lies firmly in the middle of the pairwise MO estimates. This striking result shows that MO makes more efficient use of the available data to correct for measurement error.

While the average results of the pairwise MO align with the thirty-four measure gold-standard, there is considerable variance among the individual measures. This is in part due to a fundamental difference between MO and averaging (or more general scale construction techniques like factor analysis). MO corrects measurement error on a given variable instead of constructing a new measure of an underlying concept. This often valuable result allows us to investigate how the estimated effect of economic preferences varies across the choice of measure. With pairwise MO, we find that classic economic ideology items regarding the size of government and its role in the economy have a much larger estimated effect on vote choice than questions on welfare policy, equal opportunity, and poor people — all of which were treated the same under averaging. It is interesting to note that correcting the classic questions on economic policy lead to even higher estimates than implied by the gold-standard. Furthermore, the lowest estimated effects come from variables that relate to views of the poor and their benefits from the government, which in part may proxy for other issues such as racial politics.

As [Ansolabehere, Rodden and Snyder \(2008\)](#) point out, averaging is a “tried and true” method for alleviating measurement error and it works very well when a battery of questions exists for a given concept. When, as usual, less information is available, MO may be able to extract more information from the available data.

6 What Can Go Wrong?

MO’s 2-step estimation procedure makes it, like MI, highly robust to misspecification, especially compared to structural equation-like approaches. However, like any statistical procedure, using it inappropriately can lead to incorrect inferences. Inappropriate uses include the following. First, using MO, or any measurement error correction procedure, to deal with very small degrees of measurement error may reduce bias at the expense of a larger increase in variability. Given the likely high levels of measurement error in political science variables, this is a concern, but will not

Pairwise Multiple Overimputation

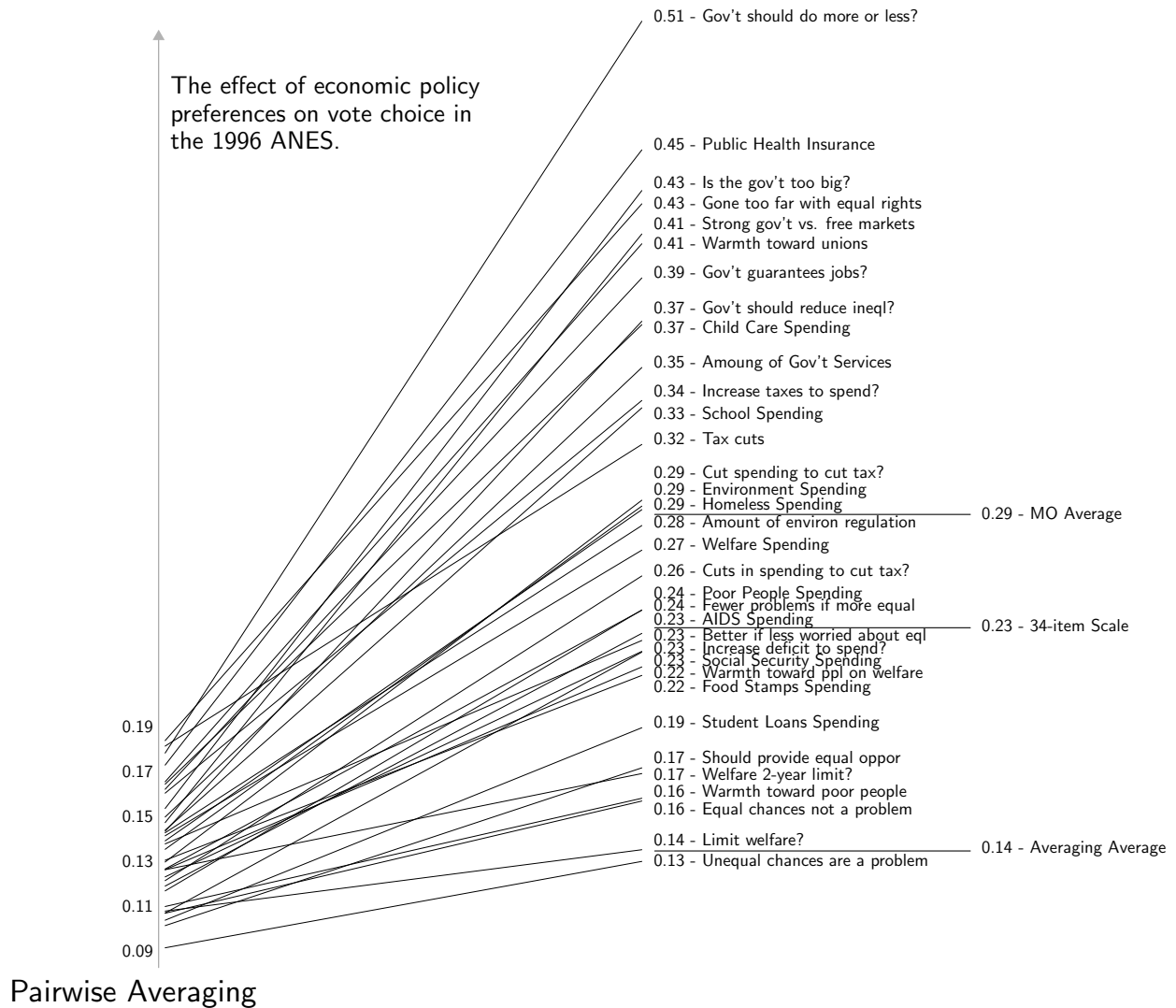


Figure 11: *The lines connect estimates from averaging across all pairwise estimates containing the specified variable (left) and estimates from multiple overimputation (right). MO estimates a higher average effect, and one that is closer to the “gold-standard” 34-item scale in each case. Furthermore, MO finds higher estimated effects for classic economic ideology questions and lower effects for questions on welfare and economic opportunity.*

normally be much of an issue. Second, overestimating the amount of measurement error in an MO application can lead to incorrect inferences, but these inferences will typically remain within the

minimal-assumption bounds, and so users should be sure to consult the bounds as a check. Further, MO handles these situations better than, say, method-of-moments estimators (Fuller, 1987) even in simple cases.²⁵ Third, violations of the key assumptions about measurement error, especially MAR assumptions, can lead to bias and, like assumptions about omitted variable bias or ignorability, are not normally testable without additional data. Sensitivity tests could be conducted however. Finally with these qualifications, there are conditions under which simple techniques like listwise deletion or ignoring the problem altogether will be preferred over MO, but these conditions normally make it highly unlikely that one would continue to trust the data for subsequent analyses (King et al., 2001).

7 Conclusion

Measurement error is a common, and commonly ignored, problem in the social sciences. Few of the methods proposed for it have been widely used, largely because of implausible assumptions, high levels of model dependence, difficult computation, and inapplicability with multiple mismeasured variables.

Here, we generalize the multiple imputation framework to handle observed data measured with error. Our multiple overimputation (MO) generalization overwrites observed but mismeasured observations with a distribution of values reflecting the best guess and uncertainty in the latent variable. Our conceptualization of the problem is that missing values are merely an extreme form of measurement error, and in fact an easy case to address with standard imputation methods because there is so little to condition on in the model. However, correctly implementing the multiple imputation framework to also handle “partially missing” data, via informative observation-level priors derived from the mismeasured data, allows us to unify the treatment of all levels of measurement error including the case of completely missing values.

This approach makes feasible rigorous treatment of measurement error across multiple covariates, with heteroskedastic errors, and in the presence of violations of assumptions necessary for common measurement treatments, such as the errors-in-variables model. The model works in survey data and time series, cross-sectional data, and with priors on individual missing cell values or those measured with error. With MO, Scholars can preprocess their data to account for measurement error and missing data, and then use the overimputed data sets our software produces with whatever model they would have used without it, ignoring the measurement issues. These advances, along with the more application-specific techniques of Imai and Yamamoto (2010) and Katz and Katz (2010), represent important steps for the correction of measurement error in the social sciences.

The advances described here can be implemented when the degree of measurement error can be analytically determined from known sample properties, estimated with additional proxies, or even when it can only be bounded by the analyst. However, looking forward, often the original creators of new measures are in the best position to know the degree of measurement error present (for example, through measures of intercoder reliability, comparison to gold-standard validation checks, or other internal knowledge) and we would encourage those who create data to include their estimates of variable or cell-level measurement error as important auxiliary information, much as sampling frame weights are considered essential in the survey literatures. Now that easy-to-use procedures exist for analyzing these data, we hope this information will be made more widely available and used.

²⁵In the simulations of Section 3.1, we find that a method-of-moments estimator can have up to 188 times higher squared bias without any offsetting increase in efficiency.

A Multiple Overimputation for Missing Data and Measurement Error

Here we introduce a general MO model and a specific EM algorithm implementation to this end. We also show that it is equivalent to MI with observation-level priors as introduced by [Honaker and King \(2010\)](#). We also offer more general notation than that in the text.

A.1 General Framework

Consider a data set with independent and identically distributed random vectors $x_i = (x_{i1}, \dots, x_{ip})$ with $i \in \{1, \dots, N\}$. We are interested in the distribution of x_i , yet we only observe a distorted version of it, y_i . Let θ refer to the unknown parameters of the ideal data and γ refer to those of the error distribution. Thus, we have distributions $p(x_i|\theta)$ and $p(y_i|x_i, \gamma)$. As with MI, our goal is to produce copies of the ideal data, x_i , based on the observed data y_i .

We define $e_i = (e_{i1}, \dots, e_{ip})$ to be a vector of error indicators. The typical element e_{ij} takes a value of 1 to indicate that variable j on observation i is measured with error so that we observe a proxy, $y_{ij} = w_{ij}$ instead of x_{ij} . Similarly, we define m_i to be a vector of missingness indicators. When m_{ij} takes the value 1, then y_{ij} is missing. If both $m_{ij} = 0$ and $e_{ij} = 0$, then the observation is perfectly measured and $y_{ij} = x_{ij}$. Let m_i and e_i have a joint distribution $p(m_i, e_i|y_i, x_i, \phi)$, whose parameters ϕ are distinct from θ and γ .

With these definitions in hand, we can decompose each observation into various subsets. Let x_i^{obs} be all the perfectly measured values, so that $x_i^{\text{obs}} = \{x_{ij}; e_{ij} = m_{ij} = 0\}$. We also have x_i^{mis} , which are the variables that are missing in observation i : $x_i^{\text{mis}} = \{x_{ij}; m_{ij} = 1\}$. Finally, we must define those variables that are measured with error. Let x_i^{err} be the unobserved, latent variables and w_i be their observed proxies: $w_i = \{w_{ij}; e_{ij} = 1\}$, $x_i^{\text{err}} = \{x_{ij}; e_{ij} = 1\}$. Thus, the observed data for any unit will be $y_i = (x_i^{\text{obs}}, w_i)$ and the ideal data would be $x_i = (x_i^{\text{obs}}, x_i^{\text{err}}, x_i^{\text{mis}})$. Note that while the dimensions of x_i and y_i are fixed, the dimensions of both w_i and x_i^{obs} can change from unit to unit.

We can write the observed-data probability density function for unit i as

$$p(y_i, m_i, e_i|\theta, \gamma, \phi) = \int \int p(x_i|\theta)p(w_i|x_i, \gamma)p(m_i, e_i|y_i, x_i, \phi)dx_i^{\text{err}}dx_i^{\text{mis}}. \quad (19)$$

We make the assumption that the data is mismeasured at random (MMAR), which states that the mismeasurement and missingness processes do not depend on the unobserved data.²⁶ Formally, we state MMAR as $p(m_i, e_i|y_i, x_i, \phi) = p(m_i, e_i|y_i, \phi)$. With this assumption in hand, we can rewrite (19) as $p(y_i, m_i, e_i|\theta, \gamma, \phi) = p(m_i, e_i|y_i, \phi)p(y_i|\theta, \gamma)$, and since we are primarily interested in inferences on θ , the first term becomes part of the proportionality constant and we are left with the observed-data distribution

$$p(y_i|\theta, \gamma) = \int \int p(x_i|\theta)p(w_i|x_i, \gamma)dx_i^{\text{err}}dx_i^{\text{mis}}. \quad (20)$$

Taking a Bayesian point of view, we can combine this with a prior on (θ, γ) giving us a posterior, $p(\theta, \gamma|y_i)$.

²⁶This is an augmented version of the *missing at random* (MAR) assumption ([Rubin, 1976](#)). MMAR would be violated if the presence of measurement error depended on the value of the latent variable itself. Since we have mismeasured proxies included in y_i , the dependence would have to be after controlling for the proxies. The most likely violation of this assumption would be if follow-up data were collected on certain observations that were different on some unmeasured covariate.

Analyzing the ideal data x_i would be much easier than y_i since the mismeasured and missing data contribute to likelihood in complicated ways. Thus, MO seeks to form a series of complete, ideal data sets: $x_{i(1)}, x_{i(2)}, \dots, x_{i(m)}$. Each of these overimputed data sets is of the form $x_{i(k)} = (x_i^{\text{obs}}, x_{i(k)}^{\text{err}}, x_{i(k)}^{\text{mis}})$, so that the perfectly measured data is constant across the overimputations. We refer to this as overimputation because we replace observed data w_i with draws from an imputation model for x_i^{err} . To form these overimputations, we take draws from the posterior predictive distribution of the unobserved data:

$$(x_{i(k)}^{\text{err}}, x_{i(k)}^{\text{mis}}) \sim p(x_{i(k)}^{\text{err}}, x_{i(k)}^{\text{mis}} | y_i) = \int p(x_{i(k)}^{\text{err}}, x_{i(k)}^{\text{mis}} | y_i, \theta, \gamma) p(\theta, \gamma | y_i) d\theta d\gamma. \quad (21)$$

Once we have these m overimputations, we can simply run m separate analyses on each data set and combine them using straightforward rules. Consider some quantity of interest, Q . Let q_1, \dots, q_m denote the separate estimates of Q which come from applying the same analysis model to each of the overimputed data sets. The overall point estimate \bar{q} of Q is simply the average $\bar{q} = \frac{1}{m} \sum_{j=1}^m q_j$. As shown by Rubin (1978), the variance of the multiple overimputation point estimate is the average of the estimated variances from within each completed data set, plus the sample variance in the point estimates across the data sets (multiplied by a factor that corrects for bias because $m < \infty$): $\bar{s}^2 = \frac{1}{m} \sum_{j=1}^m s_j^2 + S_q^2(1 + 1/m)$, where s_j is the standard error of the estimate of q_j from the analysis of data set j and $S_q^2 = \sum_{j=1}^m (q_j - \bar{q})^2 / (m - 1)$.²⁷

A.2 A Modified-EM Approach to Multiple Overimputation

The last formulation of (21) hints at one way to draw multiple imputations: (1) draw $(\theta_{(i)}, \gamma_{(i)})$ from its posterior $p(\theta, \gamma | y_i)$, then (2) draw $(x_{i(k)}^{\text{err}}, x_{i(k)}^{\text{mis}})$ from $p(x_{i(k)}^{\text{err}}, x_{i(k)}^{\text{mis}} | y_i, \theta_{(i)}, \gamma_{(i)})$. Usually these procedures are implemented with either data augmentation (that is, Gibbs sampling) or the expectation-maximization (EM) algorithm combined with an additional sampling step. We focus here on how our method works in the EM algorithm, since these two approaches are closely linked and often lead to similar inferences (Schafer, 1997; King et al., 2001; Honaker and King, 2010). EM consists of two steps: the expectation step, when we use the current guess of the parameters to fill in the missing data, and the maximization step, when we use the observed data and our current guess of the missing data to estimate the complete-data parameters. These two steps iterate until the parameters estimates converge.

If the mismeasured cells were in fact missing, we could easily apply a typical EM algorithm for missing data. In this case, though, the observed proxies, w_i , give us observation-level information about x_i^{err} . The EM algorithm usually incorporates prior beliefs about the parameters in the M-step, which is convenient when our prior beliefs are on the parameters of the data (μ, Σ) . Here our information is about the location of a missing value, not about the parameters themselves.

We therefore include this information in the expectation- or E-step of the EM algorithm. This step calculates the expected value of the complete-data sufficient statistics over the full conditional distribution of the missing data. That is, it finds $E(T(x_i) | y_i, \theta^{(t)}, \gamma)$, where $\theta^{(t)}$ is the current guess of the complete-data parameters. In our model, we adjust the E-step to incorporate the measurement error distribution as implied by the observed-data likelihood, (20). Using this likelihood, the

²⁷A second procedure for combining estimates is useful when simulating quantities of interest, as in King, Tomz and Wittenberg (2000) and Imai, King and Lau (2008). To draw m simulations of the quantity of interest, we merely draw $1/m$ of the needed simulations from each of the overimputed data sets.

modified E-step calculates

$$E(T(x_i)|y_i, \theta^{(t)}, \gamma) = \int \int T(x_i) \underbrace{p(x_i^{\text{err}}, x_i^{\text{mis}}|x_i^{\text{obs}}, \theta^{(t)})}_{\text{imputation}} \underbrace{p(w_i|x_i, \gamma)}_{\text{mismeasurement}} dx_i^{\text{err}} dx_i^{\text{mis}}, \quad (22)$$

where in typical missing data applications of EM, the mismeasurement term would be absent. The imputation part of the expectation draws information from a regression of the missing data on the observed data, while the mismeasurement part draws information from the proxy.²⁸ Thus, both sources of information help estimate the true sufficient statistics of the latent, ideal data. The M-step proceeds as usual, finding the parameters that were most likely to have give rise to the estimated sufficient statistics. Note that we could incorporate this alteration to the full conditional posterior into an MCMC approach, though instead of averaging across the distribution, a Gibbs sampler would take a draw from it.

A.3 A Multiple Overimputation Model for Normal Data

In the above description of the model, we have left the distributions unspecified. To implement the model, we must provide additional information. We assume that the complete, ideal data (x_i) is multivariate normal with mean μ and covariance Σ , so that $\theta = (\mu, \Sigma)$. This implies that any conditional distribution of the ideal is also normal.

The above measurement error distribution is in its most general form, a function of the entire ideal data vector (x_i) and some parameters, γ . As noted by [Stefanski \(2000\)](#), all approaches to correcting measurement error must include additional information about this distribution. We assume that $w_{ij} \stackrel{\text{ind}}{\sim} \mathcal{N}(x_{ij}, \lambda_{ij}^2)$ for each proxy $w_{ij} \in w_i$ and each unit i , where the measurement error variance λ_{ij}^2 is known or estimable using techniques from Section 3. Our assumption corresponds to that of classical measurement error, yet our modified EM algorithm can handle more general cases than this. If the measurement error is known to be biased or dependent upon another variable, we can simply adjust the cell-level means above and proceed as usual. Essentially, one must have knowledge of *how* the variable was mismeasured. The simulation results in Section 4.2 further indicate that MO is robust to these assumptions in certain situations.

With the measurement error model above, the normality of the data makes the calculation of the sufficient statistics straightforward. To ease exposition, we assume that there are no missing values, so that $x_i^{\text{mis}} = \emptyset$. With only measurement error, the E-step becomes

$$E(T(x_i)|y_i, \theta^{(t)}) = \int T(x_i) p(x_i^{\text{err}}|x_i^{\text{obs}}, \theta^{(t)}) \prod_{w_{ij} \in w_i} p(w_{ij}|x_{ij}, \lambda_{ij}^2) dx_i^{\text{err}}, \quad (23)$$

where $T(x_i)$ is the set of sufficient statistics for the multivariate normal. In a slight abuse of notation, we can gather the independent measurement error distributions, w_i , into a multivariate normal with mean x_i^{err} and covariance matrix $\Lambda_i = \lambda_i^2 I$, where $\lambda_i^2 = \{\lambda_{ij}^2; e_{ij} = 1\}$ and I is the identity matrix with dimension equal to $\sum_j e_{ij}$.

In order to calculate the expectation in (23), we must know the full conditional distribution, which is $p(x_i^{\text{err}}|y_i, \theta, \lambda_i^2) \propto p(x_i^{\text{err}}|x_i^{\text{obs}}, \theta)p(w_i|x_i^{\text{err}}, \lambda_i^2)$. Note that each of the distributions is (possibly multivariate) normal, with $x_i^{\text{err}}|x_i^{\text{obs}}, \theta \sim \mathcal{N}(\mu_{e|o}, \Sigma_{e|o})$ and $w_i|x_i^{\text{err}}, \lambda_i^2 \sim \mathcal{N}(x_i^{\text{err}}, \Lambda_i)$, where $(\mu_{e|o}, \Sigma_{e|o})$ are deterministic functions of θ and x_i^{obs} . This distribution amounts to the regression of

²⁸Note that we treat γ as fixed since, in our implementation, it is known or estimable. One could extend these methods to simultaneously estimate γ , though this would require additional information.

x_i^{err} on x_i^{obs} . If the values were simply missing, rather than measured with error, then the E-step would simply take the expectations with respect to this conditional expectation. With measurement error, we must combine these two sources of information. Using standard results on the normal distribution, we can write the full conditional as

$$(x_i^{\text{err}} | y_i, \theta^{(t)}, \lambda_i^2) \sim \mathcal{N}(\mu^*, \Sigma^*), \quad \Sigma^* = (\Lambda_i^{-1} + \Sigma_{e|o}^{-1})^{-1}, \quad \mu^* = \Sigma^*(\Lambda_i^{-1} w_i + \Sigma_{e|o}^{-1} \mu_{e|o}). \quad (24)$$

We simply change our E-step to calculate this expectation for each cell measured with error and proceed with the M-step as usual.²⁹ Note that while we assume that the measurement errors on different variables are independent, one could incorporate dependence into Λ_i . The result in (24) is identical to the results in the appendix of [Honaker and King \(2010\)](#), when we set a prior distribution for x_i^{err} that is normal with mean w_i and variance Λ_i . See their paper for additional implementation details.

References

- Ansolabehere, Stephen, Jonathan Rodden and James M. Snyder. 2008. “The Strength of Issues: Using Multiple Measures to Gauge Preference Stability, Ideological Constraint, and Issue Voting.” *American Political Science Review* 102(02):215–232.
- Berger, James. 1994. “An Overview of Robust Bayesian Analysis (With Discussion).” *Test* 3:5–124.
- Bitjukov, SI, VV Smirnova, NV Krasnikov and VA Taperechkina. 2006. Statistically dual distributions in statistical inference. In *Statistical Problems in Particle Physics, Astrophysics and Cosmology: proceedings of PHYSTAT05, Oxford, UK, 12-15 September 2005*. pp. 102–105. <http://arxiv.org/abs/math/0411462v2>.
- Black, Dan A., Mark C. Berger and Frank A. Scott. 2000. “Bounding Parameter Estimates with Nonclassical Measurement Error.” *Journal of the American Statistical Association* 95(451):739–748.
URL: <http://www.jstor.org/stable/2669454>
- Brownstone, David and Robert G. Valletta. 1996. “Modeling Earnings Measurement Error: A Multiple Imputation Approach.” *Review of Economics and Statistics* 78(4):705–717.
- Carroll, Raymond J. and Leonard A. Stefanski. 1990. “Approximate Quasi-likelihood Estimation in Models With Surrogate Predictors.” *Journal of the American Statistical Association* 85(411):652–663.
URL: <http://www.jstor.org/stable/2290000>
- Carroll, R.J., D. Ruppert and L.A. Stefanski. 1995. *Measurement error in nonlinear models*. Vol. 63 Chapman & Hall/CRC.
- Cole, Stephen R, Haitao Chu and Sander Greenland. 2006. “Multiple-imputation for measurement-error correction.” *International Journal of Epidemiology* 35(4):1074–81.

²⁹If there are missing values in unit i , we need to alter the definitions of Λ_i^{-1} and w_i to be 0 for the entries corresponding to the missing variables.

- Cook, J. and L. Stefanski. 1994. "Simulation-extrapolation estimation in parametric measurement error models." *Journal of the American Statistical Association* 89:1314–1328.
- Freedman, Laurence S, Douglas Midthune, Raymond J Carroll and Victor Kipnis. 2008. "A comparison of regression calibration, moment reconstruction and imputation for adjusting for covariate measurement error in regression." *Stat Med* 27(25):5195–216.
- Fuller, Wayne A. 1987. *Measurement error models*. Wiley New York.
- Ghosh-Dastidar, B. and J.L. Schafer. 2003. "Multiple edit/multiple imputation for multivariate continuous data." *Journal of the American Statistical Association* 98(464):807–817.
- Guolo, Annamaria. 2008. "Robust techniques for measurement error correction: a review." *Statistical Methods in Medical Research* 17(6):555–80.
- Honaker, James and Eric Plutzer. 2011. "Small Area Estimation with Multiple Overimputation." Paper presented at the Midwest Political Science Association, Chicago.
- Honaker, James and Gary King. 2010. "What to do About Missing Values in Time Series Cross-Section Data." *American Journal of Political Science* 54(2, April):561–581. <http://gking.harvard.edu/files/abs/pr-abs.shtml>.
- Honaker, James, Gary King and Matthew Blackwell. 2010. "Amelia II: A Program for Missing Data." <http://gking.harvard.edu/amelia>.
- Huckfeldt, Robert, Eric Plutzer and John Sprague. 1993. "Alternative Contexts of Political Behavior: Churches, Neighborhoods, and Individuals." *Journal of Politics* 55(2, May):365–381.
- Imai, Kosuke, Gary King and Olivia Lau. 2008. "Toward A Common Framework for Statistical Analysis and Development." *Journal of Computational Graphics and Statistics* 17(4):1–22. <http://gking.harvard.edu/files/abs/z-abs.shtml>.
- Imai, Kosuke and Teppei Yamamoto. 2010. "Causal Inference with Differential Measurement Error: Nonparametric Identification and Sensitivity Analysis." *American Journal of Political Science* 54(2, April):543–560.
- Katz, Jonathan N. and Gabriel Katz. 2010. "Correcting for Survey Misreports Using Auxiliary Information with an Application to Estimating Turnout." *American Journal of Political Science* 54(3):815–835.
- King, Gary, James Honaker, Anne Joseph and Kenneth Scheve. 2001. "Analyzing Incomplete Political Science Data: An Alternative Algorithm for Multiple Imputation." *American Political Science Review* 95(1, March):49–69. <http://gking.harvard.edu/files/abs/evil-abs.shtml>.
- King, Gary and Langche Zeng. 2002. "Estimating Risk and Rate Levels, Ratios, and Differences in Case-Control Studies." *Statistics in Medicine* 21:1409–1427. <http://gking.harvard.edu/files/abs/1s-abs.shtml>.
- King, Gary, Michael Tomz and Jason Wittenberg. 2000. "Making the Most of Statistical Analyses: Improving Interpretation and Presentation." *American Journal of Political Science* 44(2, April):341–355. <http://gking.harvard.edu/files/abs/making-abs.shtml>.

- Klepper, Steven and Edward E. Leamer. 1984. "Consistent Sets of Estimates for Regressions with Errors in All Variables." *Econometrica* 52(1):163–184.
URL: <http://www.jstor.org/stable/1911466>
- Leamer, Edward. 1978. *Specification Searches*. New York: Wiley.
- Lee, Sik-Yum. 2007. *Structural equation modeling: A Bayesian approach*. Vol. 680 John Wiley & Sons Inc.
- Meng, Xiao-Li. 1994. "Multiple-Imputation Inferences with Uncongenial Sources of input." *Statistical Science* 9(4):538–573.
- Rubin, Donald. 1976. "Inference and Missing Data." *Biometrika* 63:581–592.
- Rubin, Donald B. 1978. "Bayesian inference for causal effects: The role of randomization." *The Annals of Statistics* 6:34–58.
- Rubin, Donald B. 1987. *Multiple Imputation for Nonresponse in Surveys*. New York: John Wiley.
- Schafer, Joseph L. 1997. *Analysis of incomplete multivariate data*. London: Chapman & Hall.
- Schafer, Joseph L. and Maren K. Olsen. 1998. "Multiple Imputation for multivariate Missing-Data Problems: A Data Analyst's Perspective." *Multivariate Behavioral Research* 33(4):545–571.
- Stefanski, L. A. 2000. "Measurement Error Models." *Journal of the American Statistical Association* 95(452):1353–1358.
- Wang, Naisyin and James Robins. 1998. "Large-sample theory for parametric multiple imputation procedures." *Biometrika* 85:935–948.
- White, Ian R. 2006. "Commentary: Dealing with measurement error: multiple imputation or regression calibration?" *International Journal of Epidemiology* 35(4):1081–1082.
URL: <http://ije.oxfordjournals.org/content/35/4/1081.short>