aptara | AJPS | ajps˙447 | Dispatch: February 18, 2010 | CE: n/a
Journal | MSP No. | No. of pages: 21 | PE: Martha

What To Do about Missing Values in Time-Series Cross-Section Data

James Honaker The Pennsylvania State University
Gary King Harvard University

Applications of modern methods for analyzing data with missing values, based primarily on multiple imputation, have in the last half-decade become common in American politics and political behavior. Scholars in this subset of political science have thus increasingly avoided the biases and inefficiencies caused by ad hoc methods like listwise deletion and best guess imputation. However, researchers in much of comparative politics and international relations, and others with similar data, have been unable to do the same because the best available imputation methods work poorly with the time-series cross-section data structures common in these fields. We attempt to rectify this situation with three related developments. First, we build a multiple imputation model that allows smooth time trends, shifts across cross-sectional units, and correlations over time and space, resulting in far more accurate imputations. Second, we enable analysts to incorporate knowledge from area studies experts via priors on individual missing cell values, rather than on difficult-to-interpret model parameters. Third, because these tasks could not be accomplished within existing imputation algorithms, in that they cannot handle as many variables as needed even in the simpler cross-sectional data for which they were designed, we also develop a new algorithm that substantially expands the range of computationally feasible data types and sizes for which multiple imputation can be used. These developments also make it possible to implement the methods introduced here in freely available open source software that is considerably more reliable than existing algorithms.

We develop an approach to analyzing data with missing values that works well for large numbers of variables, as is common in American politics and political behavior; for cross-sectional, time series, or especially "time-series cross-section" (TSCS) data sets (i.e., those with T units for each of N cross-sectional entities such as countries, where often $T < N$), as is common in comparative politics and international relations; or for when qualitative knowledge exists about specific missing cell values. The new methods greatly increase the information researchers are able to extract from given amounts of data and is equivalent to having much larger numbers of observations available.

Our approach builds on the concept of "multiple imputation," a well-accepted and increasingly common approach to missing data problems in many fields. The idea is to extract relevant information from the observed portions of a data set via a statistical model, to impute multiple (around five) values for each missing cell, and to use these to construct multiple "completed" data sets. In each of these data sets, the observed values are the same, and the imputations vary depending on the estimated uncertainty in predicting each missing value. The great attraction of the procedure is that after imputation analysts can apply to each of the completed data sets whatever statistical method they would have used if there had been no missing values and then use a simple procedure to combine the results. Under normal circumstances, researchers can impute once and then analyze the imputed data sets as many times and for as many purposes as they wish. The task of running their analyses multiple times and combining results is routinely and transparently

James Honaker is a lecturer at The Pennsylvania State University (Department of Political Science, Pond Laboratory, University Park PA 16802; tercer@psu.edu). Gary King is Albert J. Weatherhead III University Professor, Harvard University (Institute for Quantitative Social Science, 1737 Cambridge Street, Harvard University, Cambridge MA 02138; http://GKing.Harvard.edu, King@Harvard.edu, (617) 495-2027).

We have written an easy-to-use software package that implements all the methods introduced in this article; it is called, "Amelia II: A Program for Missing Data" and is available at http://gking.harvard.edu/amelia. Our thanks to Neal Beck, Adam Berinsky, Matthew Blackwell, Jeff Lewis, Kevin Quinn, Don Rubin, Ken Scheve, and Jean Tomphie for helpful comments, the National Institutes of Aging (P01 AG17625-01), the National Science Foundation (SES-0318275, IIS-9874747, SES-0550873), and the Mexican Ministry of Health for research support.

American Journal of Political Science, Vol. 54, No. 2, April 2010, Pp. 561–581

©2010, Midwest Political Science Association ISSN 0092-5853

UNCORRECTED PROOF

handled by special purpose statistical analysis software. As a result, after careful imputation, analysts can ignore the missingness problem (King et al. 2001; Rubin 1987).

Commonly used multiple imputation methods work well for up to 30–40 variables from sample surveys and other data with similar rectangular, nonhierarchical properties, such as from surveys in American politics or political behavior where it has become commonplace. However, these methods are especially poorly suited to data sets with many more variables or the types of data available in the fields of political science where missing values are most endemic and consequential, and where data structures differ markedly from independent draws from a given population, such as in comparative politics and international relations. Data from developing countries especially are notoriously incomplete and do not come close to fitting the assumptions of commonly used imputation models. Even in comparatively wealthy nations, important variables that are costly for countries to collect are not measured every year; common examples used in political science articles include infant mortality, life expectancy, income distribution, and the total burden of taxation.

When standard imputation models are applied to TSCS data in comparative and international relations, they often give absurd results, as when imputations in an otherwise smooth time series fall far from previous and subsequent observations, or when imputed values are highly implausible on the basis of genuine local knowledge. Experiments we have conducted where selected observed values are deleted and then imputed with standard methods produce highly uninformative imputations. Thus, most scholars in these fields eschew multiple imputation. For lack of a better procedure, researchers sometimes discard information by aggregating covariates into five-or ten-year averages, losing variation on the dependent variable within the averages (see for example, Iversen and Soskice 2006; Lake and Baum 2001; Moene and Wallerstein 2001; and Timmons 2005, respectively). Obviously this procedure can reduce the number of observations on the dependent variable by 80 or 90%, limits the complexity of possible functional forms estimated and number of control variables included, due to the restricted degrees of freedom, and can greatly affect empirical results—a point regularly discussed and lamented in the cited articles.

These and other authors also sometimes develop ad hoc approaches such as imputing some values with linear interpolation, means, or researchers' personal best guesses. These devices often rest on reasonable intuitions: many national measures change slowly over time, observations at the mean of the data do not affect inferences for some quantities of interest, and expert knowledge outside their quantitative data set can offer useful information. To put data in the form that their analysis software demands, they then apply listwise deletion to whatever observations remain incomplete. Although they will sometimes work in specific applications, a considerable body of statistical literature has convincingly demonstrated that these techniques routinely produce biased and inefficient inferences, standard errors, and confidence intervals, and they are almost uniformly dominated by appropriate multiple imputation-based approaches (Little and Rubin 2002).[1]

Applied researchers analyzing TSCS data must then choose between a statistically rigorous model of missingness, predicated on assumptions that are clearly incorrect for their data and which give implausible results, or ad hoc methods that are known not to work in general but which are based implicitly on assumptions that seem more reasonable. This problem is recognized in the comparative politics literature where scholars have begun to examine the effect of missing data on their empirical results. For example, Ross (2006) finds that the estimated relationship between democracy and infant mortality depends on the sample that remains after listwise deletion. Timmons (2005) shows that the relationship found between taxation and redistribution depends on the choice of taxation measure, but superior measures are subject to increased missingness and so not used by researchers. And Spence (2007) finds that Rodrik's (1998) results are dependent on the treatment of missing data.

We offer an approach here aimed at solving these problems. In addition, as a companion to this paper, we make available (at http://gking.harvard.edu/amelia)

[1]King et al. (2001) show that, with the average amount of missingness evident in political science articles, using listwise deletion under the most optimistic of assumptions causes estimates to be about a standard error farther from the truth than failing to control for variables with missingness. The strange assumptions that would make listwise deletion better than multiple imputation are roughly that we know enough about what generated our observed data to not trust them to impute the missing data, but we still somehow trust the data enough to use them for our subsequent analyses. For any one observation, the misspecification risk from using all the observed data and prior information to impute a few missing values will usually be considerably lower than the risk from inefficiency that will occur and selection bias that may occur when listwise deletion removes the dozens of more numerous observed cells. Application-specific approaches, such as models for censoring and truncation, can dominate general purpose multiple imputation algorithms, but they must be designed anew for each application type, are unavailable for problems with missingness scattered throughout an entire data matrix of dependent and explanatory variables, and tend to be highly model-dependent. Although these approaches will always have an important role to play in the political scientist's toolkit, since they can also be used together with multiple imputation, we focus here on more widely applicable, general purpose algorithms.

UNCORRECTED PROOF

an easy-to-use software package that implements all the methods discussed here. The software, called Amelia II: A Program for Missing Data, works within the R Project for Statistical Computing or optionally through a graphical user interface that requires no knowledge of R (Honaker, King, and Blackwell 2009). The package also includes detailed documentation on implementation details, how to use the method in real data, and a set of diagnostic routines that can help evaluate when the methods are applicable in a particular set of data. The nature of the algorithms and models developed here makes this software faster and more reliable than existing imputation packages (a point which statistical software reviews have already confirmed; see Horton and Kleinman 2007).

Multiple Imputation Model

Most common methods of statistical analysis require rectangular data sets with no missing values, but data sets from the real political world resemble a slice of swiss cheese with scattered missingness throughout. Considerable information exists in partially observed observations about the relationships between the variables, but listwise deletion discards all this information. Sometimes this is the majority of the information in the original data set.[2]

Continuing the analogy, what most researchers try to do is to fill in the holes in the cheese with various types of guesses or statistical estimates. However, unless one is able to fill in the holes with the true values of the data that are missing (in which case there would be no missing data), we are left with "single imputations" which cause statistical analysis software to think the data has more observations than were actually observed and to exaggerate the confidence you have in your results by biasing standard errors and confidence intervals.

That is, if you fill the holes in the cheese with peanut butter, you should not pretend to have more cheese! Analysis would be most convenient for most computer programs if we could melt down the cheese and reform it into a smaller rectangle with no holes, adding no new information, and thus not tricking our computer program into thinking there exists more data than there really is. Doing the equivalent, by filling in observations and then deleting some rows from the data matrix, is too difficult to do properly; and although methods of analysis adapted to the swiss cheese in its original form exist (e.g., Heckman 1990; King et al. 2004), they are mostly not available for missing data scattered across both dependent and explantory variables.

Instead, what multiple imputation does is to fill in the holes in the data using a predictive model that incorporates all available information in the observed data together along with any prior knowledge. Separate "completed" data sets are created where the observed data remains the same, but the missing values are "filled in" with different imputations. The "best guess" or expected value for any missing value is the mean of the imputed values across these data sets; however, the uncertainty in the predictive model (which single imputation methods fail to account for) is represented by the variation across the multiple imputations for each missing value. Importantly, this removes the overconfidence that would result from a standard analysis of any one completed data set, by incorporating into the standard errors of our ultimate quantity of interest the variation across our estimates from each completed data set. In this way, multiple imputation properly represents all information in a data set in a format more convenient for our standard statistical methods, does not make up any data, and gives accurate estimates of the uncertainty of any resulting inferences.

We now describe the predictive model used most often to generate multiple imputations. Let D denote a vector of p variables that includes all dependent and explanatory variables to be used in subsequent analyses, and any other variables that might predict the missing values. Imputation models are predictive and not causal and so variables that are posttreatment, endogenously determined, or measures of the same quantity as others can all be helpful to include as long as they have some predictive content. In particular, including the dependent variable to impute missingness in an explanatory variable induces no endogeneity bias, and randomly imputing an explanatory variable creates no attenuation bias, because the imputed values are drawn from the observed data posterior. The imputations are a convenience for the analyst because they rectangularize the data set, but they add nothing to the likelihood and so represent no new information even though they enable the analyst to avoid listwise deleting any unit that is not fully observed on all variables.

We partition D into its observed and missing elements, respectively: $D = \{D^{\text{obs}}, D^{\text{mis}}\}$. We also define a

[2] If archaeologists threw away every piece of evidence, every tablet, every piece of pottery that was incomplete we would have entire cultures that disappeared from the historical record. We would no longer have the *Epic of Gilgamesh*, or any of the writings of Sappho. It is a ridiculous proposition because we can take all the partial sources, all the information in each fragment, and build them together to reconstruct much of the complete picture without any invention. Careful models for missingness allow us to do the same with our own fragmentary sources of data.

UNCORRECTED PROOF

missingness indicator matrix M (with the same dimensions as D) such that each element is a 1 if the corresponding element of D is missing and 0 if observed. The usual assumption in multiple imputation models is that the data are *missing at random* (MAR), which means that M can be predicted by D^{obs} but not (after controlling for D^{obs}) D^{mis}, or more formally $p(M \mid D) = p(M \mid D^{\text{obs}})$. MAR is related to the assumptions of ignorability, nonconfounding, or the absence of omitted variable bias that are standard in most analysis models. MAR is much safer than the more restrictive *missing completely at random* (MCAR) assumption which is required for listwise deletion, where missingness patterns must be unrelated to observed or missing values: $P(M \mid D) = P(M)$. MCAR would be appropriate if coin flips determined missingness, whereas MAR would be better if missingness might also be related to other variables, such as mortality data not being available during wartime. An MAR assumption can be wrong, but it would by definition be impossible to know on the basis of the data alone, and so all existing general purpose imputation models assume it. The key to improving multiple imputation model is including more information in the model so that the stringency of the ignorability assumption is lessened.

An approach that has become standard for the widest range of uses is based on the assumption that D is multivariate normal, $D \sim N(\mu, \Sigma)$, an implication of which is that each variable is a linear function of all others. Although this is an approximation, and one not usually appropriate for analysis models, scholars have shown that for imputation it usually works as well as more complicated alternatives designed specially for categorical or mixed data (Schafer 1997; Schafer and Olsen 1998). All the innovations in this paper would easily apply to these more complicated alternative models, but we focus on the simpler normal case here. Furthermore, as long as the imputation model contains at least as much information as the variables in the analysis model no biases are generated by introducing more complicated models (Meng 1994). In fact, the two-step nature of multiple imputation has two advantages over "optimal" one-step approaches. First, including variables or information in the imputation model not needed in the analysis model can make estimates even more efficient than a one-step model, a property known as "super-efficiency." And second, the two-step approach is much less model-dependent because no matter how badly specified the imputation model is, it can only affect the cell values that are missing.

Once m imputations are created for each missing value, we construct m completed data sets and run whatever procedure we would have run if all our data had been observed on each. From each analysis, a quantity of interest is computed (a descriptive feature, causal effect, prediction, counterfactual evaluation, etc.) and the results are combined. The combination can follow Rubin's (1987) original rules, which involve averaging the point estimates and using an analogous but slightly more involved procedure for the standard errors, or more simply by taking $1/m$ of the total required simulations of the quantities of interest from each of the m analyses and summarizing the set of simulations as is now common practice with single models (e.g., King, Tomz, and Wittenberg 2000).

Computational Difficulties and Bootstrapping Solutions

A key computational difficulty in implementing the normal multiple imputation algorithm is taking random draws of μ and Σ from their posterior densities in order to represent the estimation uncertainty in the problem. One reason this is hard is that the $p(p+3)/2$ elements of μ and Σ increase rapidly with the number of variables p. So for example a problem with only 40 variables has 860 parameters and drawing a set of these parameters at random requires inverting an 860×860 variance matrix containing 370,230 unique elements.

Only two statistically appropriate algorithms are widely used to take these draws. The first proposed is the imputation-posterior (IP) approach, which is a Markov-chain, Monte Carlo-based method that takes both expertise to use and considerable computational time. The expectation maximization importance sampling (EMis) algorithm is faster than IP, requires less expertise, and gives virtually the same answers. See King et al. (2001) for details of the algorithms and citations to those who contributed to their development. Both EMis and IP have been used to impute many thousands of data sets, but all software implementations have well-known problems with large data sets and TSCS designs, creating unacceptably long run-times or software crashes.

We approach the problem of sampling μ and Σ by mixing theories of inference. We continue to use Bayesian analysis for other all parts of the imputation process and to replace the complicated process of drawing μ and Σ from their posterior density with a bootstrapping algorithm. Creative applications of bootstrapping have been developed for several application-specific missing data problems (Efron 1994; Lahlrl 2003; Rubin 1994; Rubin and Schenker 1986; Shao and Sitter 1996), but to our knowledge the technique has not been used to develop and implement a general purpose multiple imputation algorithm.

UNCORRECTED PROOF

The result is conceptually simple and easy to implement. Whereas EMis and especially IP are elaborate algorithms, requiring hundreds of lines of computer code to implement, bootstrapping can be implemented in just a few lines. Moreover, the variance matrix of μ and Σ need not be estimated, importance sampling need not be conducted and evaluated (as in EMis), and Markov chains need not be burnt in and checked for convergence (as in IP). Although imputing much more than about 40 variables is difficult or impossible with current implementations of IP and EMis, we have successfully imputed real data sets with up to 240 variables and 32,000 observations; the size of problems this new algorithm can handle appears to be constrained only by available memory. We believe it will accommodate the vast majority of applied problems in the social sciences.

Specifically, our algorithm draws m samples of size n with replacement from the data D.[3] In each sample, we run the highly reliable and fast EM algorithm to produce point estimates of μ and Σ (see the appendix for a description). Then for each set of estimates, we use the original sample units to impute the missing observations in their original positions. The result is m multiply imputed data sets that can be used for subsequent analyses.

Since our use of bootstrapping meets standard regularity conditions, the bootstrapped estimates of μ and Σ have the right properties to be used in place of draws from the posterior. The two are very close empirically in large samples (Efron 1994). In addition, bootstrapping has better lower order asymptotics than the parametric approaches IP and EMis implement and, just as symmetry-inducing transformations (like $\ln(\sigma^2)$ in regression problems) makes the asymptotics kick in faster in likelihood models. It may then be that our approach will more faithfully represent the underlying sampling density in smaller samples than the standard approaches, but this should be verified in future research.[4]

The already fast speed of our algorithm can be increased by approximately m $*$ 100% because our algorithm has the property that computer scientists call "embarrassingly parallel," which means that it is easy to segment the computation into separate, parallel processes with no dependence among them until the end. In a parallel environment, our algorithm would literally finish before IP begins (i.e., after starting values are computed, which are typically done with EM), and about at the point where EMis would be able to begin to utilize the parallel environment.

We now replicate the "MAR-1" Monte Carlo experiment in King et al. (2001, 61), which has 500 observations and about 78% of the rows fully observed. This simulation was developed to show the near equivalence of results from EMis and IP, and we use it here to demonstrate that those results are also essentially equivalent to our new bootstrapped-based EM algorithm. Figure 1 plots the estimated posterior distribution of three parameters for our approach (labeled EMB), IP/EMis (for which only one line was plotted because they were so close), the complete data with the true values included, and listwise deletion. For all three graphs in the figure, one for each parameter, IP, EMis, and EMB all give approximately the same result. The distribution for the true data is also almost the same, but slightly more peaked (i.e., with smaller variance), as should be the case since the simulated observed data without missingness has more information. IP has a smaller variance than EMB for two of the parameters and larger for one; since EMB is more robust to distributional and small sample problems, it may well be more accurate here but in any event they are very close in this example. The (red) listwise deletion density is clearly biased away from the true density with the wrong sign, and much larger variance.

Trends in Time, Shifts in Space

The commonly used normal imputation model assumes that the missing values are linear functions of other variables' observed values, observations are independent conditional on the remaining observed values, and all the observations are exchangable in that the data are not organized in hierarchical structures. These assumptions have

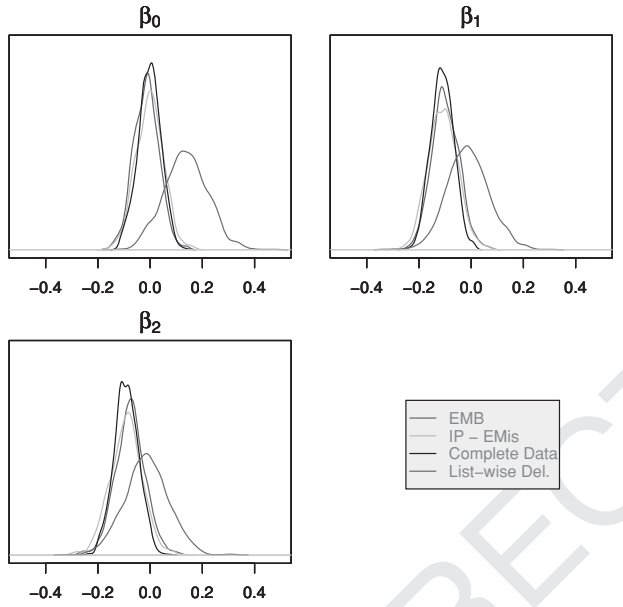
[3] This basic version of the bootstrap algorithm is appropriate when sufficient covariates are included (especially as described in the fourth section) to make the observations conditionally independent. Although we have implemented more sophisticated bootstrap algorithms for when conditional independence cannot be accomplished by adding covariates (Horowitz 2001), we have thus far not found them necessary in practice.

[4] Extreme situations, such as small data sets with bootstrapped samples that happen to have constant values or collinearity, should not be dropped (or uncertainty estimates will be too small), but is easily avoided via the traditional use of empirical (or "ridge") priors (Schafer 1997, 155).

The usual applications of bootstrapping outside the imputation context requires hundreds of draws, whereas multiple imputation only requires five or so. The difference has to do with the amount of missing information. In the usual applications, 100% of the parameters of interest are missing, whereas for imputation, the fraction of cells in a data matrix that are missing is normally considerably less than half. For problems with much larger fractions of missing information, m will need to be larger than five but rarely anywhere near as large as would be required for the usual applications of bootstrapping. The size of m is easy to determine by merely creating additional imputed data sets and seeing whether inferences change.

UNCORRECTED PROOF

FIGURE 1 Histograms Representing Posterior Densities from Monte Carlo Simulated Data ($n = 500$ and About 78% of the Units Fully Observed), via Three Algorithms and the Complete (Normally Unobserved) Data



IP and EMis, and our algorithm (EMB) are very close in all three graphs, whereas listwise deletion is notably biased with higher variance.

proven to be reasonable for survey data, but they clearly do not work for TSCS data. In this section and the next, we take advantage of these discrepancies to improve imputations by adapting the standard imputation model, with our new algorithm, to reflect the special nature of these data. Most critically in TSCS data, we need to recognize the tendency of variables to move smoothly over time, to jump sharply between some cross-sectional units like countries, to jump less or be similar between some countries in close proximity, and for time series patterns to differ across many countries.[5] We discuss smoothness over time and shifts across countries in this section and then consider issues of prior information, nonignorability, and spatial correlation in the next.

Many time-series variables, such as GDP, human capital, mortality, etc., change relatively smoothly over time. If an observation in the middle of a time series is missing, then the true value often will not deviate far from a smooth trend plotted through the data. The smooth trend need not be linear, and so the imputation technique of linear interpolation, even if modified to represent uncertainty appropriately, may not work. Moreover, sharp deviations from a smooth trend may be caused by other variables, such as a civil war. This same war might also explain why the observation is missing. Such deviates will sometimes make linear interpolation badly biased, even when accurate imputations can still be constructed based on predictions using other variables in the data set (such as the observed intensity of violence in the country).

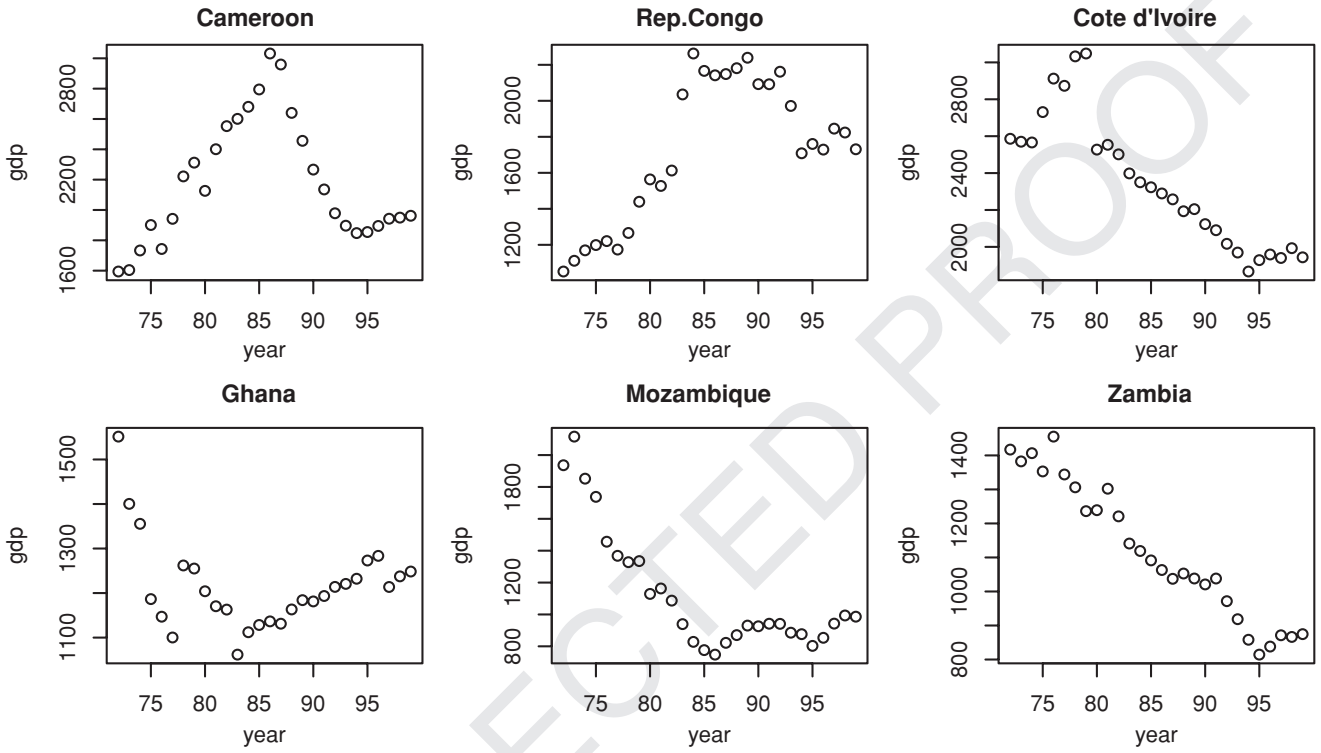
We include the information that some variables tend to have smooth trends over time in our imputation model by supplementing the data set to be imputed with smooth basis functions, constructed prior to running the imputation algorithm. These basis functions can be created via polynomials, LOESS, splines, wavelets, or other approaches, most of which have arbitrary approximation capabilities for any functional form. If many basis functions are needed, one approach would be to create basis functions for each variable within a country and to use the first few principal components of the whole set of these variables, run separately by country or interacted with country indicators. In contrast to direct interpolation, including basis functions in the imputation model will increase the smoothness of the imputations *only* if the observed data are well predicted by the basis functions conditional on other variables, and even then the predictive capacity of other variables in the model may cause deviations from smoothness if the evidence supports it.

Including q-order polynomials are the easiest to construct but may not work as well as the other choices. (In addition to being relatively rigid, polynomials work better for interpolation than extrapolation, and so missing values at the end of a series will have larger confidence intervals, but the degree of model dependence may be even larger (King and Zeng 2006).) Since trends over time in one unit may not be related to other units, when using this option we also include interactions of the polynomials with the cross-sectional unit. When the polynomial of time is simply zero-order, this becomes a model of

[5]The closest the statistical literature on missing data has come to tackling TSCS data would seem to be "repeated measures" designs, where clinical patients are observed over a small number of irregularly spaced time intervals (Little 1995; Molenberghs and Verbeke 2005). Mmissingness occurs principally in the dependent variable (the patient's response to treatment) and largely due to attrition, leading to monotone missingness patterns. As attrition is often due to a poor response to treatment, MAR is usually implausible and so missingness models are necessarily assumption-dependent (Davey, Shanahan, and Schafer 2001; Kaciroti et al. 2008). Since in typical TSCS applications, missingness is present in all variables, and time series are longer, direct application of these models is infeasible or inadequate. Researchers with data sets closer to this framework, particularly with such nonignorable missingness mechanisms, may find them more useful.

UNCORRECTED PROOF

FIGURE 2 Time Series of GDP in Six African Nations with Diverse Trends and Levels



UNCORRECTED PROOF

"fixed effects," and so this approach (or the other more sophisticated approaches) can also deal with shifts across cross-sections. As q increases, the time pattern will fit better to the observed data. With k cross-sections, a q-order polynomial will require adding $((q+1) \times k) - 1$ variables to the imputation model. As an illustration, below, we estimate a cubic polynomial for six countries and thus add $((3+1) \times 6) - 1 = 23$ fully observed covariates. For variables that are either central to our subsequent analysis or for which the time series process is important, we also recommend including lags of that variable. Since this is a predictive model, we can also include leads of the same variable as well, using the future to predict the past. Given the size of most data sets, this strategy would be difficult or impossible with IP or EMis, but our EMB algorithm, which works with much larger numbers of variables, makes this strategy feasible and easy to implement.

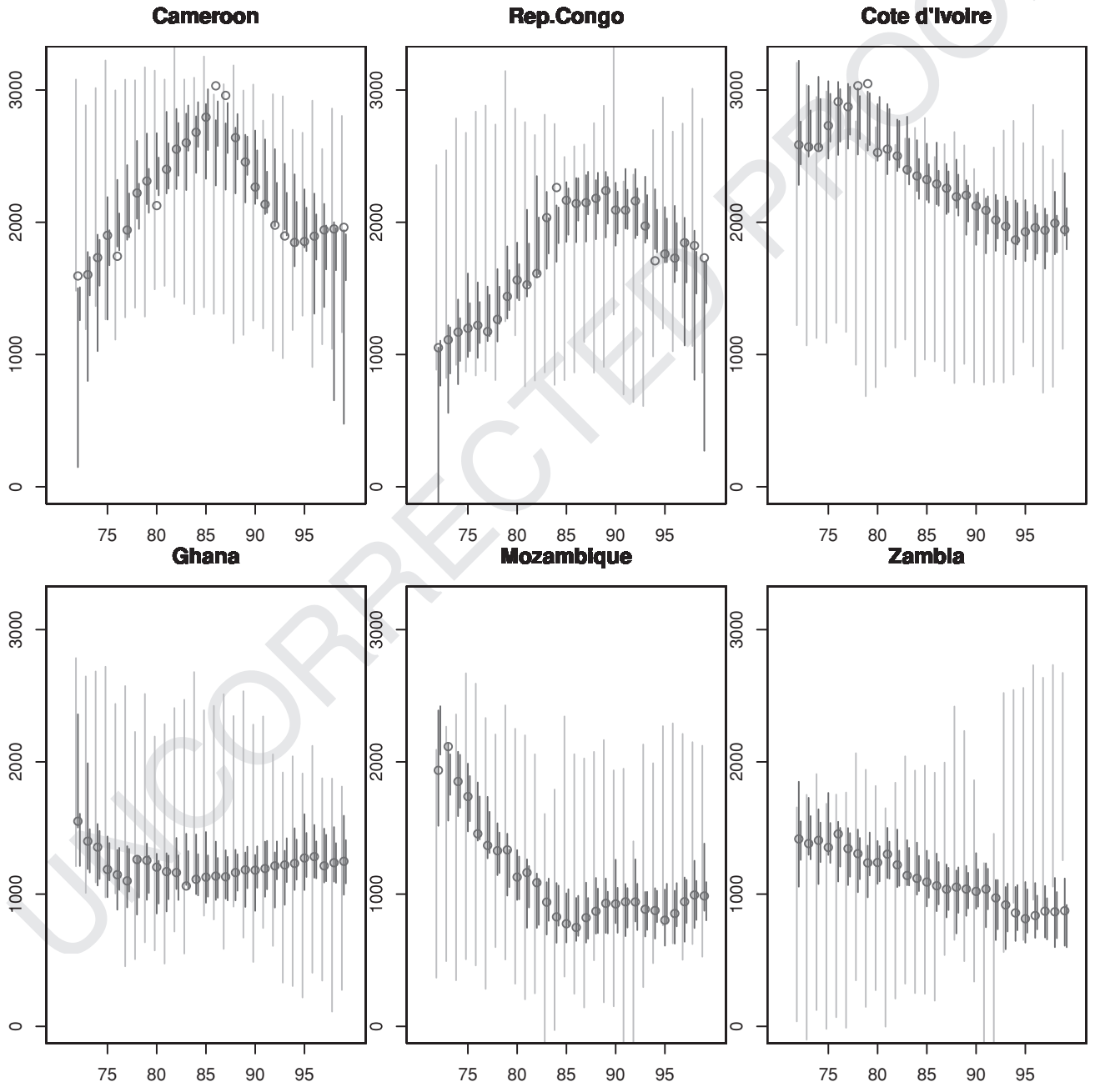
We illustrate our strategy with the data from the Africa Research Program (Bates et al. 2006). The raw data appear in Figure 2, which shows the fully observed levels of GDP in six African countries between 1972 and 1999.[6]

[6]GDP is measured as real per capita purchasing power parity using a chain international price index.

In Cameroon we can see that GDP in any year is close to the previous year, and a trend over time is discernible, whereas in the Republic of Congo the data seems much more scattered. While Cameroon's trend has an interesting narrative with a rise, a fall and then a flat period, Zambia has a much more straightforward, seemingly linear decline. Ghana experiences such a decline, followed by a period of steady growth. Cote d'Ivoire has a break in the middle of the series, possibly attributable to a crisis in the cocoa market. In addition to these values of GDP, we constructed a data set with several of the standard battery of cross-national comparative indicators, including investment, government consumption and trade openness (all three measured as a percentage of GDP), the Freedom House measure of civil freedoms, and the log of total population.

We used our EMB algorithm for all that follows. We ran 120 standard imputation models with this data set, sequentially removing one year's data from each cross-section (20 years $\times$ six countries), trying to impute the now missing value and using the known true value as validation. We then ran another 120 imputations by also including time up to a third order polynomial. For each imputation model, we construct confidence intervals and plot these in Figure 3. The green confidence intervals

FIGURE 3 The Vertical Lines Represent Three 90 percent Confidence Intervals of Imputed Values (with the Same True Values Plotted as Red Circles as in Figure 2 But on a Different Vertical Scale), from a Separate Model Run for Each Country-Year Treating That Observation of GDP as Missing



The green confidence intervals are based on the most common specification which excludes time from the imputation model. The narrower blue confidence intervals come from an imputation model that includes polynomials of time, and the smallest red confidence intervals includes LOESS smoothing to form the basis functions.

represent the distribution of imputed values from an imputation model without variables representing time. Because they were created via the standard approach that does not include information about smoothness over time, they are so large that the original trends in GDP, from Figure 2, are hard to see at this scaling of the vertical axis. The large uncertainty expressed in these intervals is accurate and so inferences based on these data will not

UNCORRECTED PROOF

mislead, but they have such low power that most interesting patterns will be missed.

We then reran the same 120 imputations, this time adding polynomials of time; the results are represented by blue lines in Figure 3 and are about a quarter the size (25.6 percent on average) of those green lines from the model without time trends. In every country, this imputation approach within each cross-section now picks up the gross patterns in the data far better than the standard approach. The blue confidence intervals are not only much smaller, but they also still capture all but a small fraction of the imputations across the 120 tests represented in this figure.

Finally, we also ran a third set of 120 imputation models, this time using LOESS smoothing to create the basis functions. These appear as red lines in Figure 3. LOESS-based smoothing provides a clear advantage over polynomial smoothing: Almost as many points are captured by the 90% confidence intervals as for the polynomials, but the LOESS-based intervals are narrower in almost all cases, especially when the polynomial-based intervals are largest.

The imputations from our preferred model do not fully capture a few patterns in the data, such as the cocoa crisis in Cote d'Ivoire and the drastic economic turnaround in Cameroon. The methods would also be less powerful when applied to data with long stretches of missingness, such as might occur with variables merged from different collections observed over periods that do not completely overlap. In the example presented here, the confidence intervals capture most of the points around, or recover shortly before and after, even extreme outliers like these. We could improve the model further by including additional or more flexible basis functions, or by including expert local knowledge, a subject to which we now turn.

Incorporating Expert Knowledge

In the usual collection of mass survey (type) data, respondents' identities and locations have been removed and so the only information analysts have about an observation is that coded in the numerical variables. In contrast, a great deal is known about the units in TSCS data beyond the quantified variables included in the data set (such as "Iran in 1980" or "the United States in 2008"). This difference between survey and TSCS data thus suggests a new source of valuable information and an opportunity to improve imputations well beyond the standard model. We do this in this section via new types of Bayesian priors.

Prior information is usually elicited for Bayesian analysis as distributions over parameters in the model, which assumes knowledge of the relationships between variables or their marginal distributions. In an imputation model, however, most of the elements of μ and Σ have little *direct* meaning, and researchers are unlikely to have prior believes about their specific values.[7] However, researchers and area studies experts often have information about particular missing values in their data sets that is much more specific and, in the context of imputation models, far more valuable.

Consider three examples. First, a researcher may understand that GDP must have been in a low range: perhaps they visited the country at that time, spoke to migrants from the country, read newspapers from that era, or synthesized the scholarly consensus that the economy was in bad shape at that time. In all these cases, researchers have information about individual missing observations rather than hypothetical parameters. For a second example, in most countries vital registration systems do not operate during wartime, and mortality due to war, which is surely higher due to the direct and indirect consequences of the conflict, is unobserved (Murray et al. 2002). And a final example would be where we do not have much raw information about the level of a variable in a country, but we believe that it is similar to the observed data in a neighboring country. We show how to add information in terms of priors for all these situations.

Researchers in many situations are thus perfectly willing to put priors on the expected values of particular missing cell values, even if they have no idea what the priors should be on the parameters of the model. Yet, for Bayesian analysis to work, all priors must ultimately be put on the parameters to be estimated, and so if we have priors on the expected value of missing observations, they must somehow be translated into a prior over the parameters, in our case on μ and Σ. Since according to the model each missing observation is generated by these $p(p+3)/2$ parameters, we need to make a few-to-many transformation, which at first sounds impossible. However, following Girosi and King (2008, chap. 5), if we restrict the transformation to the linear subspace spanned by the variables taking the role of covariates during an imputation, a prior on the expected value of one or more observations is easily transformed into a prior over μ and Σ. In particular, a prior on the expected value $E(\tilde{D}_{ij}) \equiv D^{\text{obs}}_{i,-j}\tilde{\beta}$ (where we

[7] Even when translated into regression coefficients for one variable as a linear function of the others, researchers are highly unlikely to know much about the predictive "effect" of what will be a dependent variable in the analysis model on some explanatory variable that is causally prior to it, or the effect of a treatment controlling for posttreatment variables.

UNCORRECTED PROOF

use a tilde to denote a simulated value) can be inverted to yield a prior on $\tilde{\beta} = (D_{i,-j}^{\text{obs}\prime} D_{i,-j}^{\text{obs}})^{-1} D_{i,-j}^{\text{obs}\prime} E(\tilde{D}_{ij})$, with a constant Jacobian. The parameter β can then be used to reconstruct μ and Σ deterministically. Hence, when researchers can express their knowledge at the level of the observation, we can translate it into what is needed for Bayesian modeling.[8]

We now offer a new way of implementing a prior on the expected value of an outcome variable. Our approach can be thought of as a generalized version of data augmentation priors (which date back at least to Theil and Goldberger 1961), specialized to work within an EM algorithm. We explain each of these concepts in turn. Data augmentation priors (DAPs) are appropriate when the prior on the parameters has the same functional form as the likelihood. They are attractive because they can be implemented easily by adding specially constructed pseudo-observations to the data set, with weights for the pseudo-observations translated from the variance of the prior hyperparameter, and then running the same algorithm as if there were no priors (Bedrick, Christensen, and Johnson 1996; Clogg et al. 1991; Tsutakawa 1992a). Empirical priors (as in Schafer 1997, 155) can be implemented as DAPs.

Unfortunately, implementing priors at the observation-level solely via current DAP technology would not work well for imputation problems.

The first issue is that we will sometimes need different priors for different missing cells in the same unit (say if GDP and fertility are both missing for a country-year). To allow this within the DAP framework would be tedious at best because it would require adding multiple pseudo-observations for each real observation with more than one missing value with a prior, and then adding the appropriate complex combination of weights to reflect the possibly different variances of each prior. A second more serious issue is that the DAPs have been implemented in order to estimate model parameters, in which we have no direct interest. In contrast, our goal is to create imputations, which are predictions conditional on actual observed data.

The EM algorithm iterates between an E-step (which fills in the missing data, conditional on the current model parameter estimates) and an M-step (which estimates the model parameters, conditional on the current imputations) until convergence. Our strategy for incorporating the insights of DAPs into the EM algorithm is to include the prior in the E-step and for it to affect the M-step only indirectly through its affect on the imputations in the E-step. This follows basic Bayesian analysis where the imputation turns out to be a weighted average of the model-based imputation and the prior mean, where the weights are functions of the relative strength of the data and prior: when the model predicts very well, the imputation will downweight the prior, and vice versa. (In contrast, priors are normally put on model parameters and added to EM during the M-step.[9]) This modified EM enables us to put priors on observations in the course of the EM algorithm, rather than via multiple pseudo-observations with complex weights, and enables us to impute the missing values conditional on the real observations rather than only estimate model parameters. The appendix fully describes our derivation of prior distributions for observation-level information.

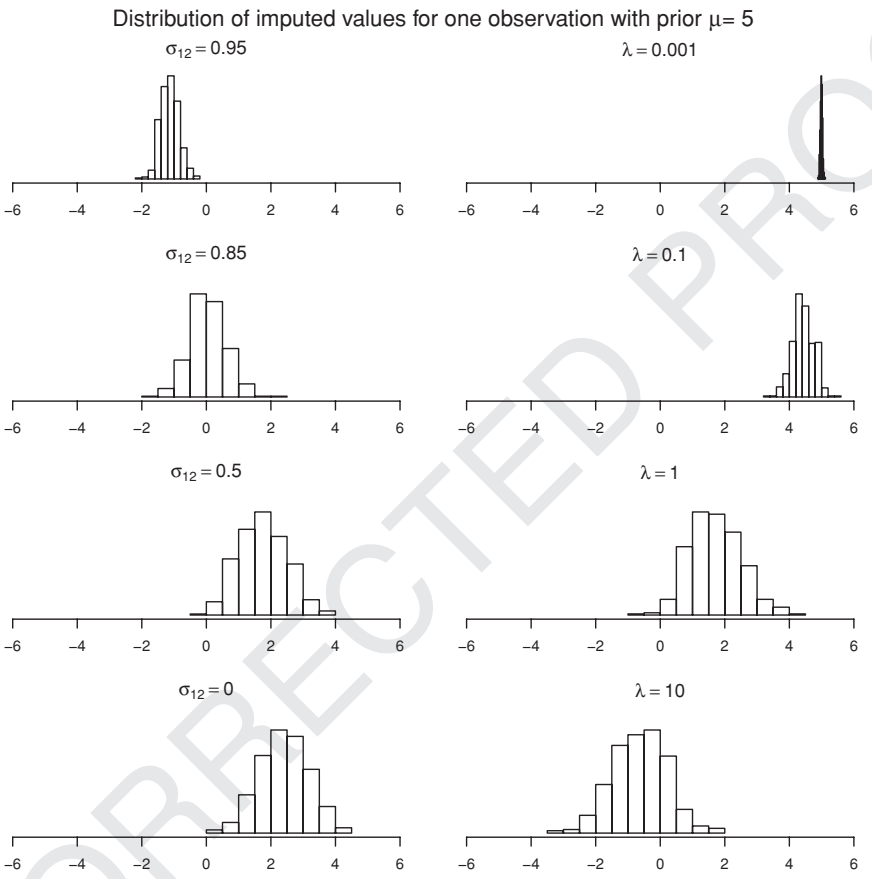
We now illustrate our approach with a simulation from a model analyzed mathematically in the appendix. This model is a bivariate normal (with parameters $\mu = (0, 0)$ and $\Sigma = \{1\ 0.4, 0.4\ 1\}$) and with a prior on the expected value of the one missing observation. Here, we add intuition by simulating one set of data from this model, setting the prior on the observation to $N(5, \lambda)$, and examining the results for multiple runs with different values of λ. (The mean and variance of this prior distribution would normally be set on the basis of existing knowledge, such as from country experts, or from averages of observed values in neighboring countries if we know that adjacent countries are similar.) The prior mean of five is set for illustrative purposes far from the true value of zero. We drew one data set with $n = 30$ and computed the observed mean to be -0.13. In the set of histograms on the right of Figure 4, we plot the posterior density of imputed values for priors of different strengths. As λ

[8] In addition to the formal approach introduced for hierarchical models in Girosi and King (2008), putting priors on observations and then finding the implied prior on coefficients has appeared in work on prior elicitation (see Gill and Walker 2005; Ibrahim and Chen 1997; Kadane 1980; Laud and Ibrahim 1995; Weiss, Wang, and Ibrahim 1997), predictive inference (Tsutakawa 1992; Tsutakawa and Lin 1986; West, Harrison, and Migon 1985), wavelet analysis (Jefferys et al. 2001), and logistic (Clogg et al. 1991) and other generalized linear models (Bedrick, Christensen, and Johnson 1996; Greenland 2001; Greenland and Christensen 2001).

[9] Although the first applications of the EM algorithm were for missing data problems (Dempster, Laird, and Rubin 1977; Orchard and Woodbury 1972), its use and usefulness has expanded to many maximum-likelihood applications (McLachlan and Krishan 2008), and as the conventional M-step is a likelihood maximization EM is considered a maximum-likelihood technique. However, as a technique for missing data, use of prior distributions in the M-step, both informative and simply for numerical stability, are common (as in Schafer 1997) and prior distributions are Bayesian. Missing data models, and multiple imputation in particular, regularly straddle different theories of inference, as discussed by Little (2008).

UNCORRECTED PROOF

FIGURE 4 Posterior Densities of the Expected Value of One Imputation Generated from a Model with a Mean of Zero and a Prior Mean of Five



The left column holds constant the strength of the prior (summarized by the smallness of its variance, λ at 1) and changes the predictive strength of the data (summarized by the covariance between the two variables, σ_{12}). The right column holds constant the predictive strength of the data (at $\sigma_{12} = 0.5$) and changes the strength of the prior (λ).

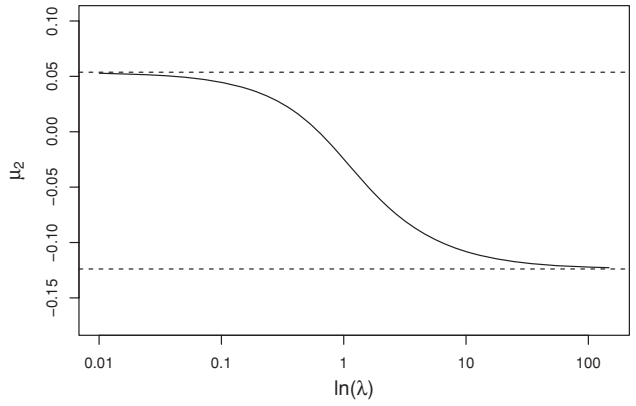
shrinks (shown for the histograms closer to the top of the figure), the imputations collapse to a spike at our value of 5, even though the model and its MAR assumption fit to the observed data without a prior would not support this. As λ becomes larger, and thus our prior becomes weaker given data of the same strength, the observed data increasingly overrides the prior, and we see the distribution of imputations centering close to the observed data value near zero. As importantly, the spread across imputed values, which reflects the uncertainty in the imputation as summarized by the model, increases.

The histograms on the right of Figure 4 keep the predictive strength of the data the same and increase the confidence of the prior. The histograms on the left of the same figure do the opposite: they hold constant the strength of the prior (i.e., λ) and increase the predictive strength of the data (by increasing the covariance between the two variables, σ_{12}). The result is that as the data predicts better (for the histograms higher in the figure on the left), the imputations increasingly reflect the model-based estimates reflecting the raw data (which has a mean value of 1.5) and ignore the prior values. (The histograms in the third position of each column have the same values of λ and σ_{12} and so are the same.)

We also illustrate here the smaller and indirect effect on the model parameters of this prior over one cell in the data matrix with Figure 5 , which plots a model parameter vertically by the log of the strength of the prior horizontally. In particular, with no prior specified, model parameter μ_2 has a value of −0.13, which we represent in

UNCORRECTED PROOF

FIGURE 5 Values of One Model Parameter μ_2, The Mean of Variable 2, with Prior $p(x_{12}) = N(5, \lambda)$, Across Different Strengths of the Prior, $\ln \lambda$ (That is on the Log Scale)



The parameter is approaching the theoretical limits (represented by dashed lines), where the upper bound is the parameter generated when the missing value is simply filled in with the expectation, and the lower bound is the parameter when the model is estimated without priors. The overall movement of this model parameter on the basis of the prior on one observation is small.

Figure 5 with the lower horizontal dashed line. If instead of a prior, we simply filled in our missing cell D_{12} at our prior value of 5, then this parameter rises to 0.05,[10] which we represent in the figure with the vertical dashed line at the top. For any possible prior or value of σ^2, then, these two values act as the limits on how much our prior can change the final estimate. The plotted curve shows how the expected value changes with λ. As $\ln \lambda \to 0$, the expected value converges to what would have resulted had we simply filled in the missing value. Similarly, as $\ln \lambda$ grows large (here about 100) then the prior has no contribution to the final estimate from the EM chain. For a sufficiently weak prior the parameter approaches the lower dashed line at -0.13 which would have resulted had no priors been used on the data set.

Figure 5 shows that the effect on a model parameter of a prior on one observation is relatively small, as it should be. Nevertheless, researchers are advised to use observation-level priors in conjunction with a judicious choice of covariates, since ultimately putting priors on observations is also putting priors on the model parameters. The key is to ensure that the covariates span a rich enough space to accommodate the added prior information, so that the data are fit better rather than the prior values merely creating outliers and biasing the model parameters with respect to the remaining imputations.

We use the same technology for putting priors on individual missing cell values to borrow strength from information in the data of neighboring or similar countries via user-specified proximity matrices. In most applications with priors, users will have information over many of the missing values in the data, rather than just one. In such cases, the computations are somewhat more involved (for details see the appendix), but the intuition in this simple case still applies.

Illustrations

In practice, any analysis using a new method on a given data set only demonstrates what can happen in those data, not in any others. We know from the GDP data analyses in Figure 2 that the effects of our methods can be massive in terms of efficiency and bias. In this section, we go further and replicate two published studies that seek to explain terrorist incidents and economic growth, respectively. We also reanalyze the same data after multiply imputing their missing data with our methods and find some major effects, with some important variables changing sign, uncertainty estimates changing, and some original findings strengthened.[11]

Explaining Terrorism

As an example of our imputation method we replicated Burgoon's (2006) study of the effect of a nation's welfare and economic policies on the number of terrorism incidents caused by citizens of that country. Burgoon estimates six similar model specifications—three different measures of a key variable of interest, with and without lagged levels of the dependent variable and time fixed effects. The number of observations after listwise deletion varies from 1,193 to 1,779. In the model with the fewest observations, 98 countries are present for an average of 12.2 years each.

[10] As shown in the appendix, this is roughly $(n_{obs}\mu_{obs} + \mu_0)/(n_{obs} + 1) = (28 * -0.13 + 5)/29$.

[11] We also replicated Moene and Wallerstein's (2001) analysis of inequality and welfare spending, Fearon's (2005) reassessment of Collier and Hoeffler's (2004) work on natural resources and civil wars, Fearon and Laitin's (2003) work on ethnicity and civil war, and Marinov's (2005) work on economic sanctions. In each of these analyses, imputation of the incomplete data strengthened the original findings, in some cases substantially. Additionally, we are limited to analyzing the effects of our methods on published work, but many research projects have undoubtedly been abandoned altogether because missing data proved too large an obstacle to overcome, or researchers were rightly concerned about the biases and inefficiencies created by listwise deletion; perhaps our methods will bring such works to completion in the future.

UNCORRECTED PROOF

We imputed this data set, bringing the number of rows of data to 2,268, which spans 108 countries for 21 years each. Most of the missing values were scattered over time among various economic indicators. On average, incomplete observations were only missing in 2.3 of the 10 key variables in the analysis (not including all the region and time fixed effects which were of course fully observed). Thus, across the roughly 1,000 incomplete observations, more than three quarters of the variables were present, but none of the observed information are used in the listwise deletion models.

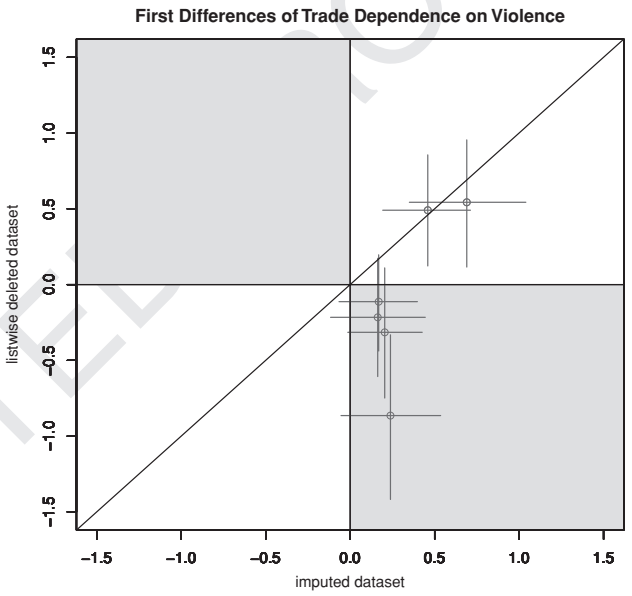
One independent variable Burgoon examines is *trade openness*, the sum of imports and exports as a fraction of gross domestic product. This is an important variable in the literature on growth and development and often used as a proxy in studies in globalization. Burgoon summarizes arguments in the literature that predict opposing causal mechanisms. Succinctly, if trade leads to growth and development, this may reduce domestic tensions and violence, while if trade leads to inequality this may increase violence and the number of terrorist incidents.

If theory cannot predict the effect of trade openness on terrorist incidents, the listwise deleted data is no more instructive. Across the six models, under slightly different model specifications and different complete observations, the effect of trade openness varies considerably in sign and magnitude. In two models the sign is positive, predicting more violence and openness increases. In four models the sign is negative. Both one of the positive and one of the negative models are significant at the 90% confident level.

We present first differences from the six models in Figure 6 . Each circle represents the expected change in the number of terrorist incidents between a country with trade openness one standard deviation below and one above the mean level of openness (holding all other variables at their mean). The vertical lines represent the confidence intervals for these first differences in the six listwise deleted models. The horizontal lines represent the confidence intervals from the six models using multiply imputed data.

If the estimates from listwise deletion and those after imputation agreed with each other, all these plus signs would line up on the $y = x$ (45 degree) line. As they move away from this line, the parameters in these models increasing disagree. The pluses that fall in either of the two shaded quadrants represent parameters whose signs change when the data set is imputed, and here we see four of the six parameters change sign, which of course means that the information discarded by listwise deletion and retained by our imputation model was substantively meaningful. As expected, the confidence interval for the imputed data, which does not discard observed cell values in the data set, are smaller (on average around 14%) than for listwise deletion. Whereas in the listwise deleted data the effect of trade can be positive or negative depending on the model specification, *all* the parameters across all the models in the imputed data predict a positive relationship with two significant at the 90% confidence level. The null test for the parameters from the imputed model can be seen graphically as the horizontal lines do not intersect the horizontal axis at zero. Although not certain, the evidence under listwise deletion indicates no particular pattern whereas under EMB imputation clearly suggests a positive relationship.

FIGURE 6 Each Plus Sign Represents the 90 Percent Confidence Interval for the Change in the Number of Terrorist Incidents When Trade Openness Changes from One Standard Deviation Below to One Deviation Above the Mean



If the parameter estimates from the listwise deleted and imputed data sets agree, then all the stars should fall directly on the 45 degree line. In the listwise deleted data sets, the sign of this parameter varies across models. However, four parameters (in the grey lower-right quadrant) change sign when the data is imputed, making the expected direction of the effect of trade coherent across alternate model specifications.

Explaining Economic Growth

For our second example we reestimate key results from Baum and Lake (2003), who are interested in the effect of democracy on economic growth, both directly (as in

UNCORRECTED PROOF

TABLE 1 Replication of Baum and Lake, Using Listwise Deletion and Multiple Imputation

	Listwise Deletion	Multiple Imputation
Life Expectancy		
Rich Democracies	−.072	.233
	(.179)	(.037)
Poor Democracies	−.082	.120
	(.040)	(.099)
N	1789	5627
Secondary Education		
Rich Democracies	.948	.948
	(.002)	(.019)
Poor Democracies	.373	.393
	(.094)	(.081)
N	1966	5627

The table shows the effect of being a democracy on life expectancy and on the percentage enrolled in secondary education (with p-values in parentheses).

Barro 1997) and indirectly through its intermediate effects on female life expectancy and female secondary education. We reproduce their recursive regression system of linear specifications, using our imputation model, and simple listwise deletion as a point of comparison.[12]

As shown in Table 1, under listwise deletion democracy conflictingly appears to *decrease* life expectancy even though it increases rates of education. These coefficients show the effect of moving one quarter of the range of the Polity democracy scale on female life expectancy and on the percentage enrolled in secondary education. With multiple imputation, the effect of democracy is consistently positive across both variables and types for rich and poor democracies. The effect of democracy on life expectancy has changed direction in the imputed data. Moreover, in the imputed data both rich and poor democracies have a statistically significant relationship to these intermediate variables. Thus the premise of intermediate effects of democracy in growth models through human capital receives increased support, as all types of democracies have a significant relationship to these measures of human capital, and democracy always positively increases human capital.[13]

In each of the examples at least one variable is intermittently measured over time and central to the analysis. We now demonstrate the intuitive fit of the imputation model by showing the distribution of imputed values in several example countries. To do this, we plot the data for three key variables for four selected countries in each row of Figure 7. Observed data appear as black dots, and five imputations are plotted as blue circles. Although five or 10 imputed data sets are commonly enough for estimating model parameters with multiple imputation, we generated 100 imputed data sets so as to obtain a fuller understanding of the range imputations for every missing value, and from these created ninety percent confidence intervals. At each missing observation in the series, these confidence intervals are plotted as vertical lines in grey. The first row shows welfare spending (total social security, health, and education spending) as a percent of GDP, from Burgoon's study. The second row shows female life expectancy from the first model we present from Baum and Lake. The last row shows the percent of female secondary enrollment, from our second model from this study. The confidence intervals and the distribution of imputations line up well with the trends over time. With the life expectancy variable, which has the strongest trends over time, the imputations fall within a narrow range of observed data. Welfare has the least clear trend over time and, appropriately, the largest relative distribution of imputed values.

Concluding Remarks

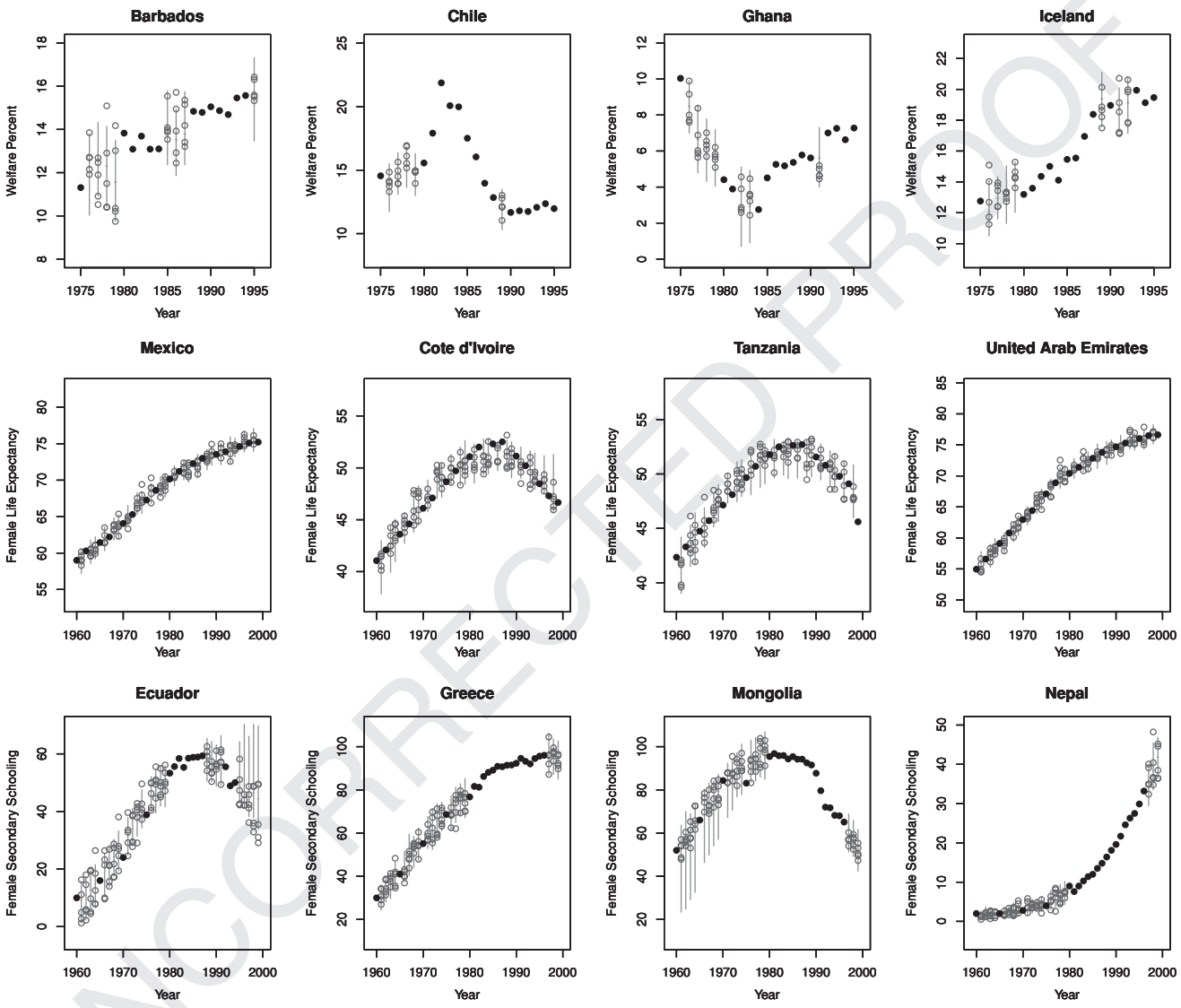
The new EMB algorithm developed here makes it possible to include features in the imputation model that would have been difficult or impossible with existing approaches resulting in more accurate imputations, increased efficiency, and reduced bias. These techniques enable us to impose smoothness over time-series variables, shifts over space, interactions between the two, and

[12]Baum and Lake use a system of overlapping moving averages of the observed data to deal with their missingness problem. Like many seemingly reasonable ad hoc procedures, they can be useful in the hands of expert data analysts but are hard to validate and will still give incorrect standard errors. In the present case, their results are intermediate between our model and listwise deletion with mixed significance and some negative effects of democracy.

[13]The number of observations more than doubles after imputation compared to listwise deletion, although of course the amount of information included is somewhat less than this because the additional rows in the data matrix are in fact partially observed. We used a first-order autoregressive model to deal with the time series properties of the data in these analyses; if we had used a lagged dependent variable there would have been only 303 and 1578 observations, respectively, in these models after listwise deletion, because more cases would be lost. The mean per capita GDP in these 303 observations where female life expectancy was collected for two sequential years was $14,900, while in the other observations the mean observed GDP was only $4,800.

UNCORRECTED PROOF

FIGURE 7 Fit of the Imputation Model



Black disks are the observed data. Blue open circles are five imputations for each missing value, and grey vertical bars represent 95% confidence intervals for these imputations. Countries in the second row have missing data for approximately every other year

observation-level priors for as many missing cells as a researcher has information about. The new algorithm even enables researchers to more reliably impute single cross-sections such as survey data with many more variables and observations than has previously been possible.

Multiple imputation was originally intended to be used for "shared (i.e., public use) data bases, collected and imputed by one entity with substantial resources but analyzed by a variety of users typically armed with only standard complete-data software" (Rubin 1994). This scenario has proved valuable for imputing a small number of public use data sets. However, it was not until software was made available directly to researchers, so they could impute their own data, that the technique began to be widely used (King et al. 2001). We hope our software, and the developments outlined here, will make it possible for scholars in comparative and international relations and other fields with similar TSCS data to extract considerably more information from their data and generate more reliable inferences. The benefits their colleagues in American politics have had for years will now be available here. Future researchers may also wish to take on the valuable task of using systematic methods of prior elicitation (Gill and Walker 2005; Kadane 1980), and the methods introduced here, to impute some of the available public use data sets in these fields.

UNCORRECTED PROOF

What will happen in the next data set to which our method is applied depends on the characteristics of those data. The method is likely to have its largest effect in data that deviates the most from the standard sample survey analyzed a few variables at a time. The leading example of such data include TSCS data sets collected over country-years or country-dyads and presently most common in comparative politics and international relations. Of course, the methods we introduce also work for more than six times as many variables as previous imputation approaches and so should also help with data analyses where standard surveys are common, such as in American politics and political behavior.

Finally, we note that users of data sets imputed with our methods should understand that, although our model has features to deal with TSCS data, analyzing the resulting multiply imputed data set still requires the same attention that one would give to TSCS problems as if the data had been fully observed (see, for example, Beck and Katz 1995; Hsiao 2003).

Appendix A. Generalized Version of Data Augmentation Priors within EM

Notation

As in the body of the paper, elements of the missingness matrix, M, are 1 when missing and 0 when observed. For notational and computational convenience, let $\mathbf{X} \equiv D$ (where D is defined in the text as a partially observed latent data matrix), where x_i is the ith row (unit), and x_{ij} the jth element (variable) in this row. Then, create a rectangularized version of D^{obs}, called $\mathbf{X}^{\text{obs}}$ by replacing missing elements with zeros: $\mathbf{X}^{\text{obs}} = \mathbf{X} * (1 - \mathbf{M})$, where the asterisk denotes an element-wise product. As is common in multivariate regression notation, assume the first column of $\mathbf{X}$ is a constant. Since this can never be missing, no row is completely unobserved (that is $m_i \neq \mathbf{1}\ \forall i$), but so that the jth subscript represents the jth variable, subscript these constant elements of the first column of $\mathbf{X}$ as x_{i0}. Denote the data set without this zero-th constant column as $\mathbf{X}_{-0}$.

The Likelihood Framework

We assume that $D \sim N(\mu, \Sigma)$, with mean μ and variance Σ. The likelihood for complete data is:

$$L(\mu, \Sigma \mid D) \propto \prod_{i=1}^{n} N(D_i \mid \mu, \Sigma) = \prod_{i=1}^{n} N(x_i \mid \mu, \Sigma). \quad (1)$$

where D_i refers to row i ($i = 1, \ldots, n$) of D. We also denote D_i^{obs} as the observed elements of row i of D, and μ_i^{obs} and Σ_i^{obs} as the corresponding subvector and submatrix of μ and Σ, respectively. Then, because the marginal densities are normal, the observed data likelihood, which we obtain by integrating (1) over D^{mis}, is

$$\begin{aligned} L(\mu, \Sigma \mid D^{\text{obs}}) &\propto \prod_{i=1}^{n} N\left(D_i^{\text{obs}} \,\middle|\, \mu_i^{\text{obs}}, \Sigma_i^{\text{obs}}\right) \\ &= \prod_{i=1}^{n} N\left(x_i^{\text{obs}} \,\middle|\, (1 - M_i) * \mu_i,\right. \\ &\qquad \left. \times (1 - M_i)'(1 - M_i) * \Sigma + M_i' M_i * H\right) \end{aligned} \quad (2)$$

where $H = \mathbf{I}(2\pi)^{-1}$, for identity matrix $\mathbf{I}$, is a placeholding matrix that numerically removes the dimensions in M_i from the calculation of the normal density since $N(\mathbf{0}|\mathbf{0}, H) = 1$. What is key here is that each observation i contributes information to differing portions of the parameters, making optimization complex to program. Each pattern of missingness contributes in a unique way to the likelihood.

An implication of this model is that missing values are imputed from a linear regression. For example, let $\tilde{x}_{ij}$ denote a simulated missing value from the model for observation i and variable j, and let $x_{i,-j}^{\text{obs}}$ denote the vector of values of all observed variables in row i, except variable j (the missing value we are imputing). The true coefficient β (from a regression of D_j on the variables with observed values in row j) can be calculated deterministically from μ and Σ since they contain all available information in the data under this model. Then, to impute, we use

$$\tilde{x}_{ij} = x_{i,-j}^{\text{obs}} \tilde{\beta} + \tilde{\epsilon}_i. \quad (3)$$

The systematic component of $\tilde{x}_{ij}$ is thus a linear function of all other variables for unit i that are observed, $x_{i,-j}^{\text{obs}}$. The randomness in $\tilde{x}_{ij}$ is generated by both estimation uncertainty due to not knowing β (i.e., μ and Σ) exactly, and fundamental uncertainty $\tilde{\epsilon}_i$ (i.e., since Σ is not a matrix of zeros). If we had an infinite sample, we would find that $\tilde{\beta} = \beta$, but there would still be uncertainty in $\tilde{x}_{ij}$ generated by the world. In the terminology of King, Tomz, and Wittenberg (2000) these imputations are *predicted values*, drawn from the distribution of x_{ij}, rather than *expected values*, or best guesses, or simulations of $\hat{x}_{ij}$ that average away the distribution of $\tilde{\epsilon}_i$.

EM Algorithms for Incomplete Data

The EM algorithm is a commonly used technique for finding maximum-likelihood estimates when the likelihood function cannot be straightforwardly constructed but a likelihood "simplified" by the addition of unknown

UNCORRECTED PROOF

parameters is easily maximized (Dempster, Laird, and Rubin 1977). In models for missing data, the likelihood conditional on the *observed* (but incomplete) data in (2) cannot be easily constructed as it would require a separate term for each of the up to 2^k patterns of missingness. However, the likelihood of a rectangularized data set (that is, for which all cells are treated as observed) like that in (1) is easy to construct and maximize, especially under the assumption of multivariate normality. The simplicity of rectangularized data is why dropping all incomplete observations via listwise deletion is so pragmatically attractive, even though the resulting estimates are inefficient and often biased. Instead of rectangularizing the data set by dropping *known* data, the EM algorithm rectangularizes the data set by filling in *estimates* of the missing elements, generated from the observed data. In the E-step, missing information are filled-in (using a generalized version of (3)) with their conditional expectations, given the current estimate of the sufficient statistics (which are estimates of μ and Σ) and the observed data. In the M-step, a new estimate of the sufficient statistics is computed from the current version of the completed data.

Sufficient Statistics. Because the data are jointly normal, $Q = \mathbf{X}'\mathbf{X}$ summarizes the sufficient statistics. Since the first column of $\mathbf{X}$ is a constant,

$$Q = \begin{pmatrix} n & \mathbf{1}\mathbf{X}_{-0} \\ \mathbf{X}_{-0}\mathbf{1} & \mathbf{X}'_{-0}\mathbf{X}_{-0} \end{pmatrix} = \sum_i \begin{pmatrix} n & x_{i1} & \dots & x_{ik} \\ x_{i1} & x_{i1}^2 & \dots & x_{i1}x_{ik} \\ \vdots & & \ddots & \\ x_{ik} & \dots & & x_{ik}^2 \end{pmatrix} \quad (4)$$

We now transform this matrix by means of the *sweep operator* into parameters of the conditional mean and unconditional covariance between the variables. Let s be a binary vector indicating which columns and rows to sweep and denote $\theta\{s\}$ as the matrix resulting from Q swept on all rows and columns for which $s_i = 1$ but not swept on rows and columns where $s_i = 0$. For example, sweeping Q on only the first row and column, results in

$$\theta\{s = (1\,0\,\dots\,0)\} = \begin{pmatrix} -1 & \mu \\ \mu' & \Sigma \end{pmatrix}, \quad (5)$$

where μ is a vector of the means of the variables, and Σ the variance-covariance matrix. This is the most common way of expressing the sufficient statistics, since $\mathbf{X}_{-0} \sim N(\mu, \Sigma)$ and all these terms are found in this version of θ. However, transformations exist to move between different parameterizations of θ and Q, as all contain the same information.

The E-step. In the E-step we compute the expectation of all quantities needed to make estimation of the sufficient statistics simple. The matrix Q requires $x_{ij}x_{ik}$ $\forall i, j, k$. Only when neither are missing can this be calculated straightforwardly from the observed data. Treating observed data as known, one of three cases holds:

$$\mathrm{E}[x_{ij}x_{ik}] = \begin{cases} x_{ij}x_{ik}, & \text{if } m_{ij}, & m_{ik} = 0 \\ \mathrm{E}[x_{ij}]x_{ik}, & \text{if } m_{ij} = 1, & m_{ik} = 0 \\ \mathrm{E}[x_{ij}x_{ik}], & \text{if } m_{ij}, & m_{ik} = 1 \end{cases} \quad (6)$$

Thus we need to calculate both $E[x_{ij} : m_{ij} = 1]$, the expectations of all missing values, and $E[x_{ij}x_{ik} : m_{ij}, m_{ik} = 1]$ the expected product of all pairs of elements missing in the same observation. The first of these can be computed simply as:

$$E[x_{ij}] = x_i^{\text{obs}}\theta\{1 - M_i\}_j^t \quad (7)$$

where the superscript t, here and below, denotes the iteration round of the EM algorithm in which that statistic was generated.

The second is only slightly more complicated as:

$$E[x_{ij}x_{ik}] = E[x_{ij}]E[x_{ik}] + \theta\{1 - M_i\}_{jk}^t \quad (8)$$

where the latter term is the estimated covariance of j and k, conditional on the observed variables in observation i.

Both (7) and (8) are functions simply of the observed data, and the matrix Q swept on the observed variables in some observation, i. Given these expectations, we can create a new rectangularized data set, $\widehat{\mathbf{X}}$, in which we replace all missing values with their individual expectations given the observed data. Sequentially, every observation of this data set can be constructed as:

$$\hat{x}_i^{t+1} = x_i^{\text{obs}} + M_i * \left(x_i^{\text{obs}}\theta\{1 - M_i\}^t\right) \quad (9)$$

The missing values within any observation have a variance-covariance matrix which can be extracted as a submatrix of θ as $\Sigma_{i|x_i^{\text{obs}}}^{t+1} = M_i' M_i * \theta\{1 - M_i\}^t$. By construction with M this will be zero for all σ_{ij} unless i and j are both missing in this observation. The expectation of the contribution of one observation, i, to T is thus $\mathrm{E}[x_i' x_i] = \hat{x}_i^{t+1'}\hat{x}_i^{t+1} + \Sigma_{i|x_i^{\text{obs}}}^{t+1}$.

The M-step. Given the construction of the expectations above, it is now simple to create an updated expectation

UNCORRECTED PROOF

of the sufficient statistics, T, by:

$$Q^{t+1} = \sum_i \left(\hat{x}_i^{t+1\prime} \hat{x}_i^{t+1} + \Sigma_{i\,|\,x_i^{\text{obs}}}^{t+1} \right) = \widehat{\mathbf{X}}^{t+1\prime} \widehat{\mathbf{X}}^{t+1} + \sum_i \left(\Sigma_{i\,|\,x_i^{\text{obs}}}^{t+1} \right). \quad (10)$$

Convergence to the Observed Data Sufficient Statistics. Throughout the iterations, the values of the observed data are constant, and generated from the sufficient statistics of the true data-generating process we would like to estimate. In each iteration, the unobserved values are filled in with the current estimate of these sufficient statistics. One way to conceptualize EM is that the sufficient statistics generated at the end of any iteration, θ^t, are a weighted sum of the "true" sufficient statistics contained within the observed data, θ^{MLE}, and the erroneous sufficient statistics, θ^{t-1} that generated the expected values. The previous parameters in θ^{t-1} used to generate these expectations may have been far from the true values, but in the next round these parameters will only be given partial weight in the construction of θ^t together with the true relationships in the observed data. Thus each sequential value of θ by necessity must be closer to the truth, since it is a weighting of the truth with the previous estimate. Like Zeno's paradox, where runners are constantly moving a set fraction of the remaining distance to the finishing line, we never quite get to the end point, but we are confident we are always moving closer. If we iterate the sequence long enough, we can get arbitrarily close to the truth, and usually we decide to end the process when the change between successive values of θ seems tolerably small that we believe we are within a sufficient neighborhood of the optimum. Convergence is guaranteed to at least a local maximum of the likelihood space under simple regularity conditions (McLachlan and Krishan 2008). When the possibility of multiple modes in the likelihood space exists, a variety of starting points, θ^0, can be used to monitor for local maxima, as is common in maximum-likelihood techniques. However, modes caused by underidentification or symmetries in the likelihood, while leading to alternate sets of sufficient statistics, often lead to the same model fit and the same distribution of predicted values for the missing data, and so are commonly less problematic than when multiple modes occur in analysis models. We provide diagnostics in our software to identify local modes in the likelihood surface as well as identify which variables in the model are contributing to these modes.

Incorporating a Single Prior

Existing EM algorithms incorporate prior information in the M-step, because this is the step where the parameters are updated, and prior information has always been assumed to inform the posterior of the parameters. Instead, we have information that informs the distribution of particular missing cells in the data set and so we modify the E-step to incorporate our priors. If the priors are over elements, it should be intuitive that it will be advantageous to apply this information over the construction of expected elements, rather than the maximization of the parameters. It is possible to map information over elements to restrictions on parameters, as demonstrated in Girosi and King (2008), but in the EM algorithm for missing data we have to construct expectations explicitly anyway for the objects for which we have information, so it is opportune to bind our information to this estimate.

Let individuals have a prior for the realized value of any individual observation, $x_{ij} : m_{ij} = 1$, as $p(x_{12}) = N(\mu_0, \lambda)$. Given this prior, we need to update $E[x_{ij}]$, and $E[x_{ij}x_{ik} : m_{ik} = 1]$ in the E-step. Conditional only on X^{obs} and the current sufficient statistics, Q, these are given by (7) and (8). Incorporating the prior, the expectation becomes:

$$E\left[x_{ij} \,|\, \mu_0, \lambda, T^t, x_i^{\text{obs}}\right] = \frac{\mu_0 \lambda^{-1} + \hat{x}_{ij}\sigma_{jj}^{-1}}{\lambda^{-1} + \sigma_{22|1}^{-1}} \quad (11)$$

where $\hat{x}_{ij} = x_i^{\text{obs}}\theta\{1 - M_i\}_j^t$ and $\sigma_{jj} = \theta\{1 - M_i\}_{jj}^t$, as previously detailed. For (8) in addition to these new expectations, we need to understand how the covariances and variance change. The variance is given by: $\text{Var}(x_{ij}, x_{ij}) = [\lambda^{-1} + (\theta\{1 - M_i\}_{jk}^t)^{-1}]^{-1}$, and calculation of the covariances are left for the more general explanation of multivariate priors in the fifth next section of this appendix.

Example. Consider the following simplified example with a latent bivariate data set of n observations drawn from $\mathbf{X}_{1,2} \sim N(\mu_1, \mu_2, \sigma_{11}, \sigma_{12}, \sigma_{22})$ where the first variable is fully observed, and the first two observations of the second variable are missing. Thus the missingness matrix looks like:

$$M = \begin{pmatrix} 0 & 0 & 1 \\ 0 & 0 & 1 \\ 0 & 0 & 0 \\ \vdots & \vdots & \vdots \\ 0 & 0 & 0 \end{pmatrix} \quad (12)$$

recalling that the first column represents the constant in the data set. Assume a solitary prior exists for the missing element of the first observation: $p(x_{12}) = N(\mu_0, \lambda)$. After

UNCORRECTED PROOF

the tth iteration of the EM sequence,

$$\theta\{(1\ 0\ 0)\}^t = \begin{pmatrix} -1 & \mu_1 & \mu_2 \\ \mu_1 & \sigma_{11} & \sigma_{12} \\ \mu_2 & \sigma_{12} & \sigma_{22} \end{pmatrix}. \quad (13)$$

If we sweep T on the observed elements of row one we return

$$\theta\{(1\ 1\ 0)\}^t$$

$$= \begin{pmatrix} . & . & \mu_2 - \mu_1\sigma_{11}^{-1}\sigma_{12} \\ . & . & \sigma_{11}^{-1}\sigma_{12} \\ \mu_2 - \mu_1\sigma_{11}^{-1}\sigma_{12} & \sigma_{11}^{-1}\sigma_{12} & \sigma_{22} - \sigma_{21}\sigma_{11}^{-1}\sigma_{12} \end{pmatrix} \quad (14)$$

$$= \begin{pmatrix} . & . & \beta_0 \\ . & . & \beta_1 \\ \beta_0 & \beta_1 & \sigma_{22|1} \end{pmatrix} \quad (15)$$

where .'s represent portions of the matrix no longer of use to this example, and β_0, β_1 and $\sigma_{22|1}$ are the parameters of the regression of x_2 on x_1, from which we can determine our expectation of the missing data element, x_{12}, conditional only on the current iteration of θ, defined as $p(x_{12}|\theta^t) = N(\mu_{12}, \sigma^2)$, $\mu_{12}^{t+1} = \beta_0 + \beta_1 * x_{11}$, and $\sigma^2 = \sigma_{22|1}$.

Therefore our expected value from this distribution is simply, $E[x_{12}|\theta^t] = \mu_{12}^{t+1}$. Then our posterior is $p(x_{ij}|\theta^t, \mu_0, \lambda) = N(\mu^*, \sigma^{2*})$, where $\sigma^{2*} = (\lambda^{-1} + \sigma_{22|1}^{-1})^{-1}$ and $\mu_{12}^* = (\lambda^{-1}\mu_0 + \sigma_{22|1}^{-1}\mu_{12}^{t+1})\sigma^{2*}$. If θ has not converged, then μ^* becomes our new expectation for x_{12} in the E-step. If θ *has* converged, then $p(x_{ij}|\theta^t, \mu_0, \lambda)$ becomes the distribution from which we draw our imputed value.

Incorporating Multiple Priors

More generally, priors may exist for multiple observations and multiple missing elements within the same observation. Complications arise especially from the latter since the strength of the prior may vary across the different elements within an observation. Conditional only on the current value of θ^t the mean expectation of the missing values in some row can be computed (by the rightmost term of equation 9) as $\hat{x}_i^{\text{mis}^{t+1}} = M_i * (x_i^{\text{obs}}\theta\{1 - M_i\}^t)$, which is a vector with zeros for observed elements, and gives the mean value of the multivariate normal distribution for unobserved values, conditional on the observed values in that observation and the current value of the sufficient statistics.

For observation i, assume a prior of $p(x_i^{\text{mis}}) = N(\mu_{0_i}, \Lambda)$, where μ_{0_i} is a vector of prior means, and where we define Λ to be a diagonal matrix: $\lambda_{ij} = 0$ for all $i \neq j$. Assuming off-diagonal elements of Λ are zero is computationally convenient, and it is appropriate when we do not have prior beliefs about how missing elements within an observation covary.[14] Thus, Λ^{-1} is a diagonal matrix with diagonal element j ($j = 1, \ldots, k$) equal to λ_{jj}^{-1} for missing values with priors, and zero for elements that are missing with no prior or are observed.

The posterior distribution of x_i^{mis} has parameters:

$$\mu_i^* = \left(\Lambda_i^{-1} + \left(\Sigma_{i|x_i^{\text{obs}}}^{t+1}\right)^{-1}\right)^{-1} \times \left(\Lambda_i^{-1}\mu_{0_i} + \left(\Sigma_{i|x_i^{\text{obs}}}^{t+1}\right)^{-1}\hat{x}_i^{\text{mis}^{t+1}}\right) \quad (16)$$

$$\Sigma_i^* = \left(\Lambda_i^{-1} + \left(\Sigma_{i|x_i^{\text{obs}}}^{t+1}\right)^{-1}\right)^{-1}. \quad (17)$$

The vector μ^* becomes our new expectation for the E-step as in the rightmost term in (9) in the construction of $\widehat{\mathbf{X}}^{t+1}$, while Σ_i^* replaces $\Sigma_{i|x_i^{\text{obs}}}^{t+1}$ in (10).[15] When the EM algorithm has converged, these terms will be also be used for the final imputations as:

$$(\tilde{x}_i|\mathbf{X}^{\text{obs}}, M, \lambda, \mu_0) \sim N(\mu_i^*, \Sigma^*) \quad (18)$$

Implicitly, note that this posterior is normally distributed, thus the priors are conjugate normal, which is convenient for the normal EM algorithm. Although we constructed our technique of observation-level priors to easily incorporate such prior information into EM chains and our EMB imputation algorithm, clearly the same observation priors could be incorporated into the IP algorithm. Here, instead of parameter priors updating the P-step, observation priors would modify the I-step through the exact same calculation of (16) and (17) and the I-step replaced by a draw from (18).

References

Barro, Robert J. 1997. *Determinants of Economic Growth.* Cambridge, MA: MIT Press.

Bates, Robert, Karen Feree, James Habyarimana, Macartan Humphreys, and Smita Singh. 2006. "The Africa Research Program." http://africa.gov.harvard.edu.

Baum, Matthew A., and David A. Lake. 2003. "The Political Economy of Growth: Democracy and Human Capital." *American Journal of Political Science* 47 (2): 333–47.

Beck, Nathaniel, and Jonathan Katz. 1995. "What to Do (and Not to Do) with Time-Series-Cross-Section Data." *American Political Science Review* 89: 634–47.

[14]This prior can be used if off-diagonal elements of Λ are nonzero. However, using the diagonal formulation is computationally convenient as it allows us to store the priors for a data set $\mathbf{X}$ of size $n \times k$ in two similarly sized $n \times k$ matrices, one matrix containing every μ_0 and one for the diagonals of each of the n different Λ^{-1}'s.

[15]In (16) μ_{ij}^* simplifies to $\hat{x}_{ij}^{t+1}$ for any missing element ij for which there is no prior specified, that is, where $\lambda_{jj}^{-1} = 0$.

UNCORRECTED PROOF

Bedrick, Edward J., Ronald Christensen, and Wesley Johnson. 1996. "A New Perspective on Priors for Generalized Linear models." *Journal of the American Statistical Association* 91 (436): 1450–60.

Burgoon, Brian. 2006. "On Welfare and Terror." *Journal of Conflict Resolution* 50 (2): 176–203.

Clogg, Clifford C., Donald B. Rubin, Nathaniel Schenker, Bradley Schultz, and Lynn Weidman. 1991. "Multiple Imputation of Industry and Occupation Codes in Census Public-Use Samples Using Bayesian Logistic Regression." *Journal of the American Statistical Association* 86 (413): 68–78.

Collier, Paul, and Anke Hoeffler. 2004. "Greed and Grievance in Civil War." *Oxford Economic Papers* 56 (4): 563–95.

Davey, Adam, Michael J. Shanahan, and Joseph L. Schafer. 2001. "Correcting for Selective Nonresponse in the National Longitudinal Survey of Youth Using Multiple Imputation." *The Journal of Human Resources* 36 (3): 500–19.

Dempster, Arthur P., N.M. Laird, and D.B. Rubin. 1977. "Maximum Likelihood Estimation from Incomplete Data via the EM Algorithm." *Journal of the Royal Statistical Association* 39: 1–38.

Efron, Bradley. 1994. "Missing Data, Imputation, and the Bootstrap." *Journal of the American Statistical Association* 89 (426): 463–75.

Fearon, James D. 2005. "Primary Commodity Exports and Civil War." *Journal of Conflict Resolution* 49 (4): 483–507.

Fearon, James D., and David D. Laitin. 2003. "Ethnicity, Insurgency, and Civil War." *American Political Science Review* 97 (1): 75–90.

Gill, Jeff, and Lee Walker. 2005. "Elicited Priors for Bayesian Model Specification in Political Science Research." *Journal of Politics* 67 (3): 841–72.

Girosi, Federico, and Gary King. 2008. *Demographic Forecasting*. Princeton: Princeton University Press. http://gking.harvard.edu/files/smooth/.

Greenland, Sander. 2001. "Putting Background Information About Relative Risks into conjugate Prior Distributions." *Biometrics* 57(September): 663–70.

Greenland, Sander, and Ronald Christensen. 2001. "Data Augmentation Priors for Bayesian and Semi-Bayes Analyses of Conditional-logistic and Proportional-Hazards Regression." *Statistics in Medicine* 20: 2421–28.

Heckman, James. 1990. "Varieties of Selection Bias." *The American Economic Review* 80 (2): 313–18.

Honaker, James, Gary King, and Matthew Blackwell. 2009. "Amelia II: A Program for Missing Data.". http://gking.harvard.edu/amelia.

Horowitz, Joel L. 2001. "The Bootstrap." *Handbook of Econometrics* 5: 3159–3228.

Horton, Nicholas J., and Ken P. Kleinman. 2007. "Much Ado About Nothing: A Comparion of Missing Data Methods and Software to Fit Incomplete Data Regression Models." *The American Statistician* 61 (1): 79–90.

Hsiao, C. 2003. *Analysis of Panel Data*. New York: Cambridge University Press.

Ibrahim, Joseph G., and Ming-Hui Chen. 1997. "Predictive Variable Selection for the Multivariate Linear Model." *Biometrics* 53(June): 465–78.

Iversen, Torben, and David Soskice. 2006. "Electoral Institutions and the Politics of Coalitions: Why Some Democracies Redistribute More Than Others." *American Political Science Review* 100 (2): 165–81.

Kaciroti, Niko A., Trivellore E. Raghunathan, M. Anthony Schork, and Noreen M. Clark. 2008. "A Bayesian Model for Longitudinal Count Data with Nongnorable Dropout." *Journal of the Royal Statistical Society Series C-Applied Statistics* 57 (5): 521–34.

Kadane, Joseph B. 1980. Predictive and Structural Methods for Eliciting Prior Distributions. In *Bayesian Analysis in Econometrics and Statistics*, ed. Arnold Zellner. Amsterdam: North-Holland, 845–54.

King, Gary, Christopher J.L. Murray, Joshua A. Salomon, and Ajay Tandon. 2004. "Enhancing the Validity and Cross-cultural Comparability of Measurement in Survey Research." *American Political Science Review* 98 (1): 191–207. http://gking.harvard.edu/files/abs/vign-abs.shtml.

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): 49–69. http://gking.harvard.edu/files/abs/evil-abs.shtml.

King, Gary, and Langche Zeng. 2006. "The Dangers of Extreme Counterfactuals." *Political Analysis* 14 (2): 131–59. http://gking.harvard.edu/files/abs/counterft-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): 341–55. http://gking.harvard.edu/files/abs/making-abs.shtml.

Lahlrl, P. 2003. "On the Impact of Boostrapping in Survey Sampling and Small Area Estimation." *Statistical Science* 18 (2): 199–210.

Lake, David A., and Matthew A. Baum. 2001. "The Invisible Hand of Democracy: Political Control and the Provision of Public Services." *Comparative Political Studies* 34 (6): 587–621.

Laud, Purushottam W., and Joseph G. Ibrahim. 1995. "Predictive Model Selection." *Journal of the Royal Statistical Society, B* 57 (1): 247–62.

Little, Roderick. 1995. "Modeling the Drop-Out Mechanism in Repeated-Measures Studies." *jasa* 90 (431): 1112–21.

Little, Roderick. 2008. "Calibrated Bayes: A Bayes/Frequentist Roadmap." *American Statistician* 60 (1): 1–11.

Little, Roderick J.A., and Donald B. Rubin. 2002. *Statistical Analysis with Missing Data, 2nd Edition*. New York, New York: John Wiley and Sons.

Marinov, Nikolay. 2005. "Do Economic Sanctions Destabilize Country Leaders?" *American Journal of Political Science* 49 (3): 564–76.

McLachlan, Geoffrey J., and Thriyambakam Krishan. 2008. *The EM Algorithm and Extensions, Second Edition*. New York: Wiley.

Meng, Xiao-Li. 1994. "Multiple-Imputation Inferences with Uncongenial Sources of input." *Statistical Science* 9 (4): 538–73.

Moene, Karl Ove, and Michael Wallerstein. 2001. "Inequality, Social Insurance, and Redistribution." *American Political Science Review* 95 (4): 859–74.

UNCORRECTED PROOF

Molenberghs, Geert, and Geert Verbeke. 2005. *Models for Discrete Longitudinal Data*. New York: Wiley.

Murray, Christopher J.L., Gary King, Alan D. Lopez, Niels Tomijima, and Etienne Krug. 2002. "Armed Conflict as a Public Health Problem." *BMJ (British Medical Journal)* 324 (9): 346–49. http://gking.harvard.edu/files/abs/armedph-abs.shtml.

Orchard, T., and M.A. Woodbury. 1972. A Missing Information Principle: Theory and Applications. In *Proceedings of the 6th Berkeley Symposium on Mathematical Statistics and Probability*, eds. Lucien M. Le Cam, Jerzy Neyman, and Elizabeth L. Scott. Berkeley: University of California Press, 697–715.

Rodrik, Dani. 1998. "Why Do More Open Economies Have Bigger Governments?" *Journal of Political Economy* 106 (5): 997–1032.

Ross, Michael. 2006. "Is Democracy Good for the Poor?" *American Journal of Political Science* 50 (4): 860–74.

Rubin, Donald B. 1987. *Multiple Imputation for Nonresponse in Surveys*. New York: John Wiley.

Rubin, Donald B. 1994. "Missing Data, Imputation, and the Bootstrap: Comment." *Journal of the American Statistical Association* 89 (426): 475–78.

Rubin, Donald, and Nathaniel Schenker. 1986. "Multiple Imputation for Interval Estimation for Simple Random Samples with Ignorable Nonresponse." *Journal of the American Statistical Association* 81 (394): 366–74.

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): 545–71.

Shao, Jun, and Randy R. Sitter. 1996. "Bootstrap for Imputed Survey Data." *Journal of the American Statistical Association* 91 (435): 1278–88.

Spence, Matthew J. 2007. "Do Governments Spend More to Compensate for Openness." Working paper, UCLA.

Theil, H., and AS Goldberger. 1961. "On Pure and Mixed Estimation in Econometrics." *International Economic Review* 2: 65–78.

Timmons, Jeffrey F. 2005. "The Fiscal Contract: States, Taxes, and Public Services." *World Politics* 57 (4): 530–67.

Tsutakawa, Robert K. 1992a. "Moments Under Conjugate Distributions in Bioassay." *Statistics & Probability Letters* 15(October): 229–33.

Tsutakawa, Robert K. 1992b. "Prior Distribution for Item Response Curves." *British Journal of Mathematical and Statistical Psychology* 45: 51–74.

Tsutakawa, Robert K., and Hsin Ying Lin. 1986. "Bayesian Estimation of Item Response Curves." *Psychometrika* 51 (2): 251–67.

Weiss, Robert E., Yan Wang, and Joseph G. Ibrahim. 1997. "Predictive Model Selection for Repeated Measures Random Effects Models Using Bayes Factors." *Biometrics* 53(June): 592–602.

West, Mike, P. Jeff Harrison, and Helio S. Migon. 1985. "Dynamic Generalized Linear Models and Bayesian Forecasting." *Journal of the American Statistical Association* 80 (389): 73–83.

UNCORRECTED PROOF