What to do about Missing Values in Time Series Cross-Section Data[1]

James Honaker[2] Gary King[3]

July 14, 2006

[1] Our thanks to Matthew Blackwell, Jeff Lewis, Don Rubin, and Ken Scheve 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. The current version of this paper, and software to implement the methods introduced in it, are available at http://gking.harvard.edu/amelia.

[2] University of California, Los Angeles (Department of Political Science, Bunche Hall, Los Angeles, CA 90095-1472; tercer@ucla.edu)

[3] 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).

Abstract

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 these fields 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. 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 build nonignorable missingness models by enabling analysts to incorporate knowledge from area studies experts via priors on individual missing cell values, rather than on difficult-to-interpret model parameters. Third, since 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 made it possible for us to implement our methods in new open source software which, unlike all existing multiple imputation packages, virtually never crashes.

1 Introduction

Multiple imputation is a well accepted and increasingly common approach to missing data problems in many fields. The idea is to use a model to extract relevant information from a data set to impute multiple (around five) values for each missing cell. We use these to fill in multiple "completed" data sets, in which the observed values are the same in all, and the imputations vary depending on the estimated uncertainty in predicting each 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 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 handled by analysis software. As a result, after careful imputation, analysts can ignore the missingness problem (Rubin, 1987; King, Honaker, Joseph and Scheve, 2001).

Existing multiple imputation models and algorithms work well for a limited number of variables from sample surveys and other data with similar rectangular, exchangable, single-sample properties. However, they are especially poorly suited to the types of data available in the fields where missing values are most endemic and where data structures differ markedly from independent draws from a given population, such as in comparative politics and international relations. The time-series cross-section (TSCS) data sets in these fields, especially those from developing countries, are notoriously plagued by incompleteness, and do not come close to fitting the assumptions of existing imputation models. When existing models are applied, they often give absurd results; imputations in an otherwise smooth time series fall far from previous and subsequent observations, or values for countries or time periods that are highly implausible on the basis of genuine local knowledge. Experiments where selected observed values are deleted and then imputed with standard methods often dramatically fail.

Despite often giving these methods a try, most scholars in these fields eschew multiple imputation and instead use 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. Unfortunately, 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 multiple imputation-based approaches (Little and Rubin, 2002).[1]

Applied researchers analyzing TSCS data must then choose between a statistically rig-

[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 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. 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 and are typically unavailable for problems with missingness scattered throughout an entire data matrix of dependent and explanatory variables. Although these approaches will always have an important role to play in the political scientist's toolkit, we focus here on more widely applicable, general purpose algorithms.

orous 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. In this paper, we develop a multiple imputation approach to model missingness that incorporates the insights of applied researchers in a statistically appropriate way. The result should increase the information we are able to extract from data in comparative, IR, other areas with TSCS data, and single cross-sections with larger numbers of variables. Section 2 summarizes the current standard approach to multiple imputation. Section 3 discusses our new algorithm that handles much larger numbers of variables and observations. We then show how to use our new algorithm to recognize the special properties of TSCS data due to time trends and spatial patterns in Section 4 and to incorporate expert knowledge about certain missing values in Section 5. Section 6 concludes.

2 The Standard Multiple Imputation Model

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 post-treatment, endogenously determined, or measures of the same quantity as others can all be helpful to include as long as they have some predictive content. We partition D into its observed and missing elements respectively: $D = \{D^{\text{obs}}, D^{\text{mis}}\}$. We also define a missingness indicator matrix M, each element of which 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 $p(M|D) = p(M|D^{\text{obs}})$. MAR is related to the assumptions of ignorability, nonconfounding, or the absence of omitted variable bias in other areas. MAR assumptions can be wrong, but it would by definition be impossible to know on the basis of the data alone, therefore all general purpose imputation models asssume it. (We alter this assumption and develop a *nonignorable* models by adding priors in Section 5.)

An approach that has become standard for the widest range of uses is based on the assumption that D is multivariate normal. Although this is an approximation, and indeed an approximation not normally 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 keep to 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, using an analysis model that is neither normal nor linear generates no biases (Meng, 1994).

Thus, we assume that $D \sim N(\mu, \Sigma)$, with mean vector μ and variance matrix Σ. The likelihood function for complete data is then:

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

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 over D^{mis}, is

$$L(\mu, \Sigma | D^{\text{obs}}) \propto \prod_{i=1}^{n} N(D_i^{\text{obs}} | \mu_i^{\text{obs}}, \Sigma_i^{\text{obs}}) \tag{2}$$

Thus, each observation i contributes information to differing portions of the parameters, making optimization complex to program.

A common practice is to add "empirical priors" to this likelihood. Empirical priors contain no real external knowledge about the parameter values, and are instead introduced to improve numerical stability and reduce variance by leaving the mean and variance of each variable (μ and $\text{diag}(\Sigma)$) unaffected and shrinking the covariances (the off-diagonal elements of Σ) toward zero (see Schafer, 1997).

An implication of this model is that missing values are imputed from a linear regression. For example, let $\tilde{D}_{ij}$ denote a simulated missing value from the model for observation i and variable j, and let $D_{i,-j}^{\text{obs}}$ 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

$$\tilde{D}_{ij} = D_{i,-j}^{\text{obs}} \tilde{\beta} + \tilde{\epsilon}_i. \tag{3}$$

The systematic component of $\tilde{D}_{ij}$ is thus a linear function of all variables for unit i that are observed, $D_{i,-j}^{\text{obs}}$. The randomness in $\tilde{D}_{ij}$ 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 $\beta = \tilde{\beta}$, but there would still be uncertainty generated by the world.

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. From each analysis, some 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 combining $1/m$ simulations of the quantities of interest from each of the m analyses and summarizing them as is now common practice with single models (e.g., King, Tomz and Wittenberg, 2000).

3 Computational Difficulties and Bootstrapping Solutions

The key steps in imputing a missing value from the model in Section 2 are (1) drawing $\tilde{\mu}$ and $\tilde{\Sigma}$ from their posterior densities, (2) using these to obtain the simulations of $\tilde{\beta}$ and $\tilde{\epsilon}_i$, (3) substituting the results into the right side of Equation 3, and (4) determinstically computing the imputation. The only computational difficulty in this process is Step (1). One reason this is hard is that the $p(p + 3)/2$ elements increase rapidly with the number of variables p. So for example a problem with only 40 variables has 860 parameters, and an 860 × 860 variance matrix of these parameters contains 370,230 elements.

Only two statistically appropriate algorithms are widely used to implement Step (1). The first proposed was the imputation-posterior (IP) approach, which is a Markov-chain, Monte Carlo-based method that takes both expertise to use and considerable computational time. An expectation maximization importance sampling (EMis) algorithm is faster than IP, requires less expertise, and gives virtually the same answers. See King, Honaker, Joseph and Scheve (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 (Rubin and Schenker, 1986; Rubin, 1994; Efron, 1994; Shao and Sitter, 1996; Lahlrl, 2003), but to our knowledge the technique has not been used to develop and implement a general purpose multiple imputation algorithm.

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), Markov chains need not be burnt in and checked for convergence (as in IP). Computer time and the potential for introducing coding bugs is considerably reduced. Even before adapting this approach to TSCS data, the algorithm is capable of imputing data sets with many more variables and observations. Although imputing much more than about 40 variables is difficult or impossible with current implementations of IP and EMis, we experimented with imputing real data sets with up to 150 variables and 36,000 observations and have not yet been able to find the upper limits which this new algorithm can handle. We believe it will be able to accommodate the vast majority of applied problems in the social sciences.

Specifically, our algorithm draws m (≈ 5) samples of size n with replacement from the data D. In each sample, we run the highly reliable and fast EM algorithm to produce point estimates of μ and Σ (see Appendix A 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 will have the right properties to be used in place of draws from the posterior. The two will also likely be very close empirically in large samples (Efron, 1994). In addition, bootstrapping has advantages since it 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, our approach should more faithfully represent the underlying sampling density in smaller samples than the standard approaches. Finally, like sandwich-based standard errors, bootstrap-based variance estimates are also consistent even if the mean function is misspecified.[2]

Our algorithm has the property that computer scientists call "embarrassingly parallel,"

[2] 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 priors.

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 of bootstrapping, 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.

which means that it is easy to segment the computation into separate, parallel processes with no dependence among them until the end. In our case, every EM chain is run from a separate bootstrap of the original data. No two imputed datasets require a previously computed estimate, so each chain can be run independently on an individual processor. EMis is not practically parallelizable until the importance sampling stage, after the EM chain has run. This is likewise true in the IP algorithm. Because burn-in periods in very large data sets can be enormous, and monitoring these is burdensome on the user (and listwise deletion gives poor starting values), standard IP practice is to first run an EM chain, and use the optima found as the starting point for m different MCMC chains. Since, in a parallel environment, our m chains would be completed in the time it takes to run one EM chain, our algorithm would literally finish before IP begins, and about the point that EMis would be able to begin to utilize the parallel environment. Implemented with parallel processing, then, our algorithm would be considerably faster than existing alternatives.

We now replicate the "MAR-1" Monte Carlo experiment in King, Honaker, Joseph and Scheve (2001, p.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 in mean and also has much larger variance.

4 Trends in Time, Shifts in Space

The standard imputation model assumes that the missing values are linear functions of other variables' observed values, that observations are independent conditional on the remaining observed values, and all the observations are exchangable. These assumptions have 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. 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, will not often deviate far from a smooth trend plotted through the data. The smooth trend need not be linear, and so the imputation technique of linear interpo-

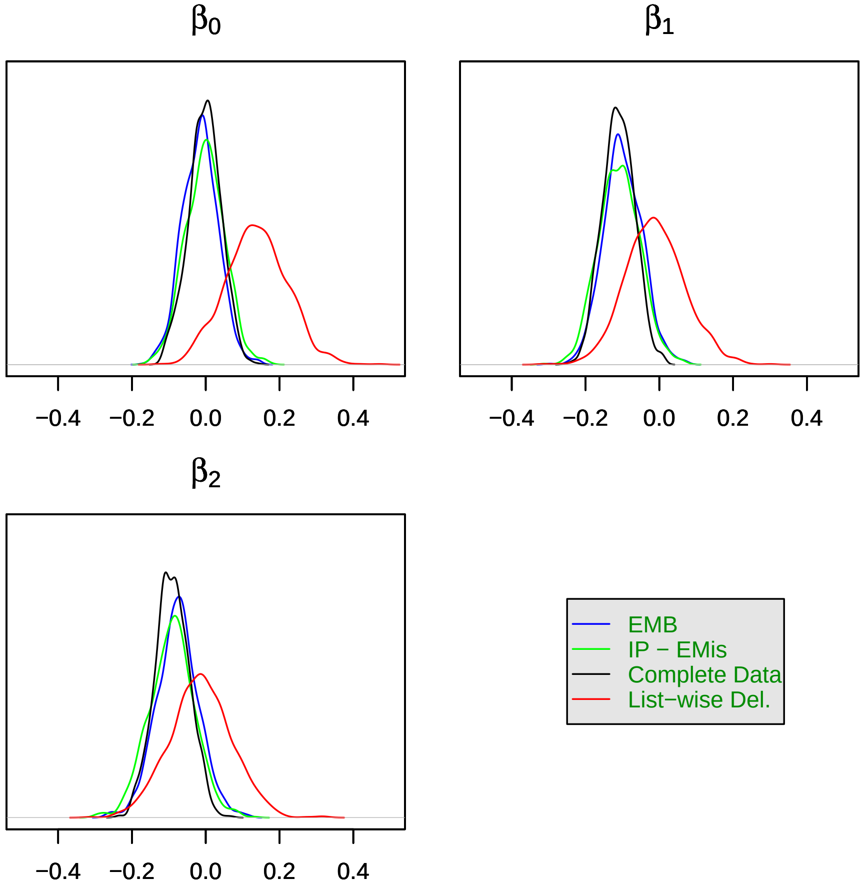

Figure 1: *Histograms representing posterior densities from Monte Carlo simulated data ($n = 500$ and about 78% of the units fully observed), via three algorithms and the complete (normally unobserved) data. IP and EMis, and our algorithm (EMB) are very close in all three graphs, whereas listwise deletion is notably biased with higher variance.*

lation, 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, or other approaches, the choice among which is roughly the same as for an analysis model. If many basis functions are needed, one approach would be to create LOESS or spline 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.

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 as they should, 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 "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.[3] 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 bootstrapping 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 × 6 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. For each imputation model, we construct confidence intervals and plot these in Figure 3. The green confidence intervals represent the distribution of imputed values from an imputation model without time, and the blue confidence intervals include time up to the third order.

The green confidence intervals, created via the standard approach that does not include information about smoothness over time, are so large that the original trends in GDP, from Figure 2, are hard to see at this scaling of the vertical axis. The imputation model that includes polynomials of time has confidence intervals about a quarter the size (25.6 percent on average) of those from the model without time trends. In every country, our imputation approach which allows for smooth trends over time within each cross-section also picks up the gross patterns in the data far better than the standard approach. The blue confidence intervals from our approach are 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

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

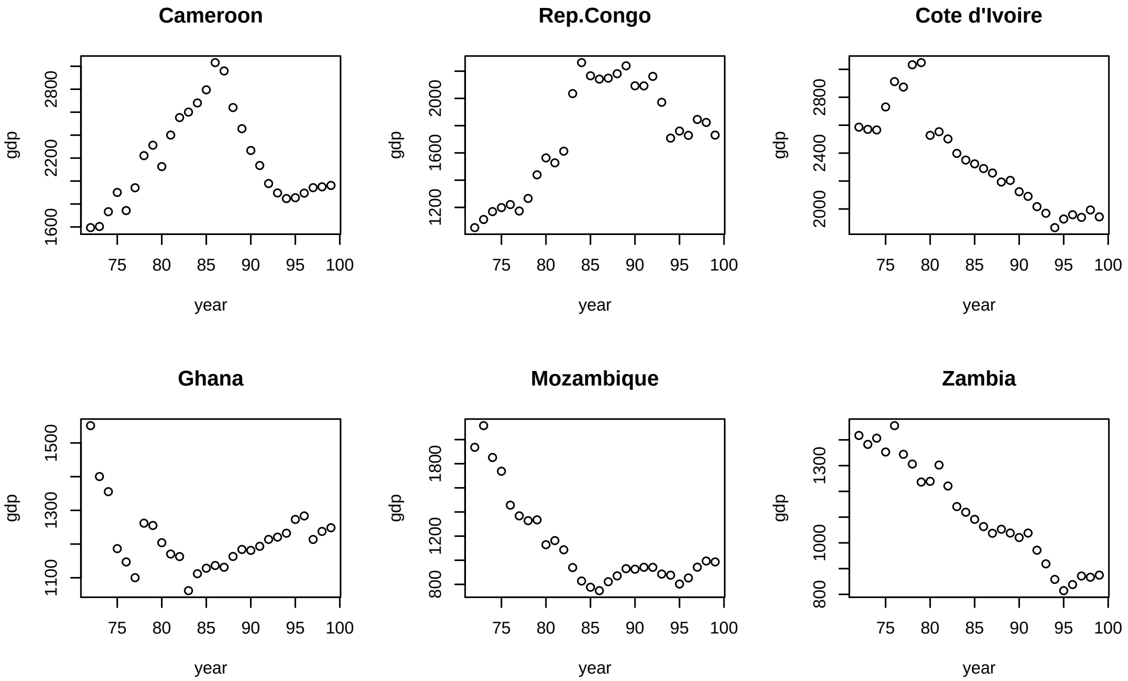

Figure 2: *Time Series of GDP in six African nations with diverse trends and levels.*

smoothing to create the basis functions. We compare the LOESS and polynomial imputation models in Figure 4, with confidence intervals for the two in blue and red, respectively. 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 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.

5 Incorporating Expert Knowledge

The standard imputation model assumes MAR but, as it turns out, less restrictive assumptions are easier to introduce in TSCS data sets than in the survey data for which most imputation models were developed. In the usual mass survey data, but not TSCS data, rows (respondents) are plausibly exchangable and anonymous, in that the label for any row (which as far as the analyst knows is just a row number rather than a proper noun) can be switched with any other without loss of information. In contrast, no matter how many

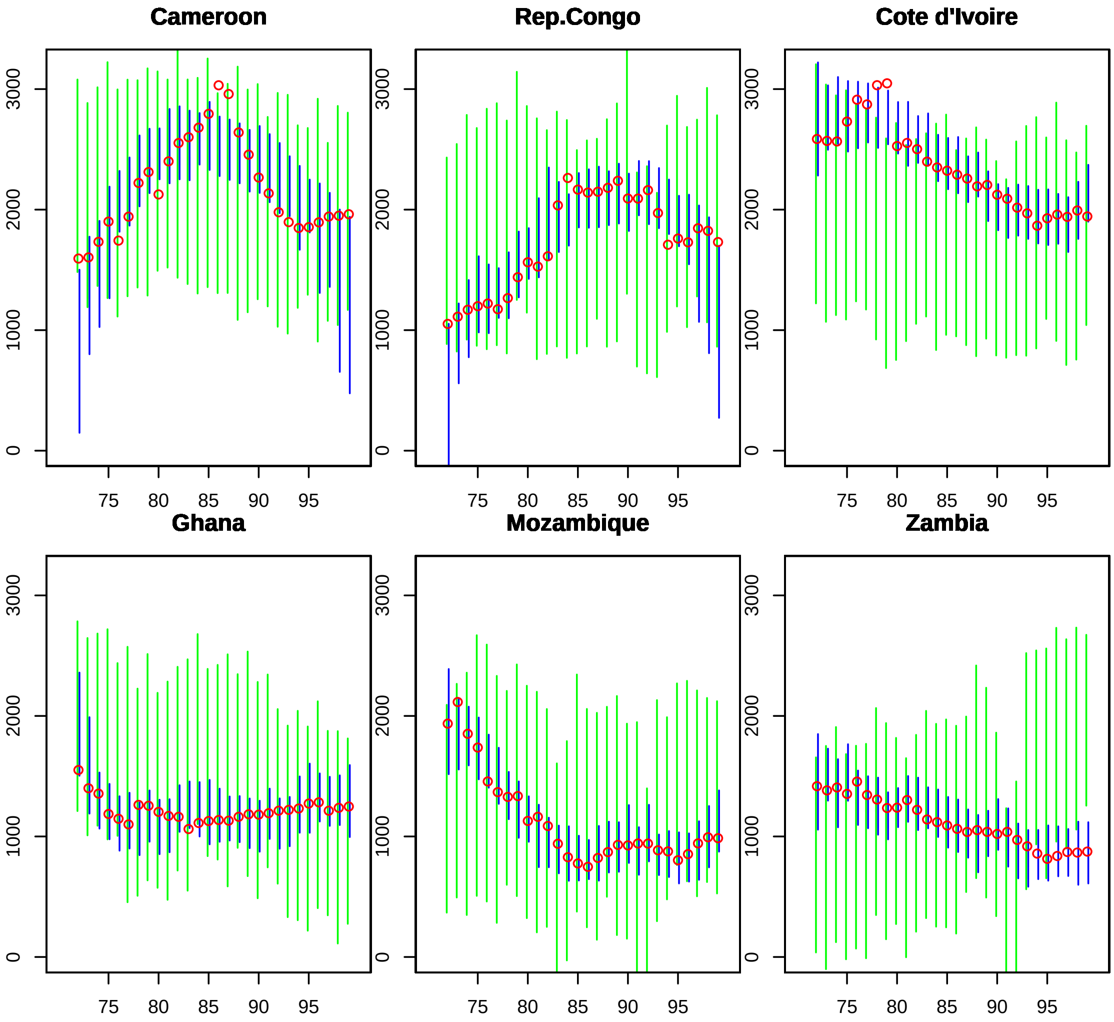

Figure 3: *The vertical lines represent 90 percent confidence intervals of imputed values (with the same true values plotted as red circles as in Figure 2 but on a different vertical scale), from a separate model run for each country-year treating that observation of GDP as missing. The narrow blue confidence intervals come from an imputation model that includes polynomials of time. The green lines, which are on average four times larger, use the standard approach which excludes time from the imputation model.*

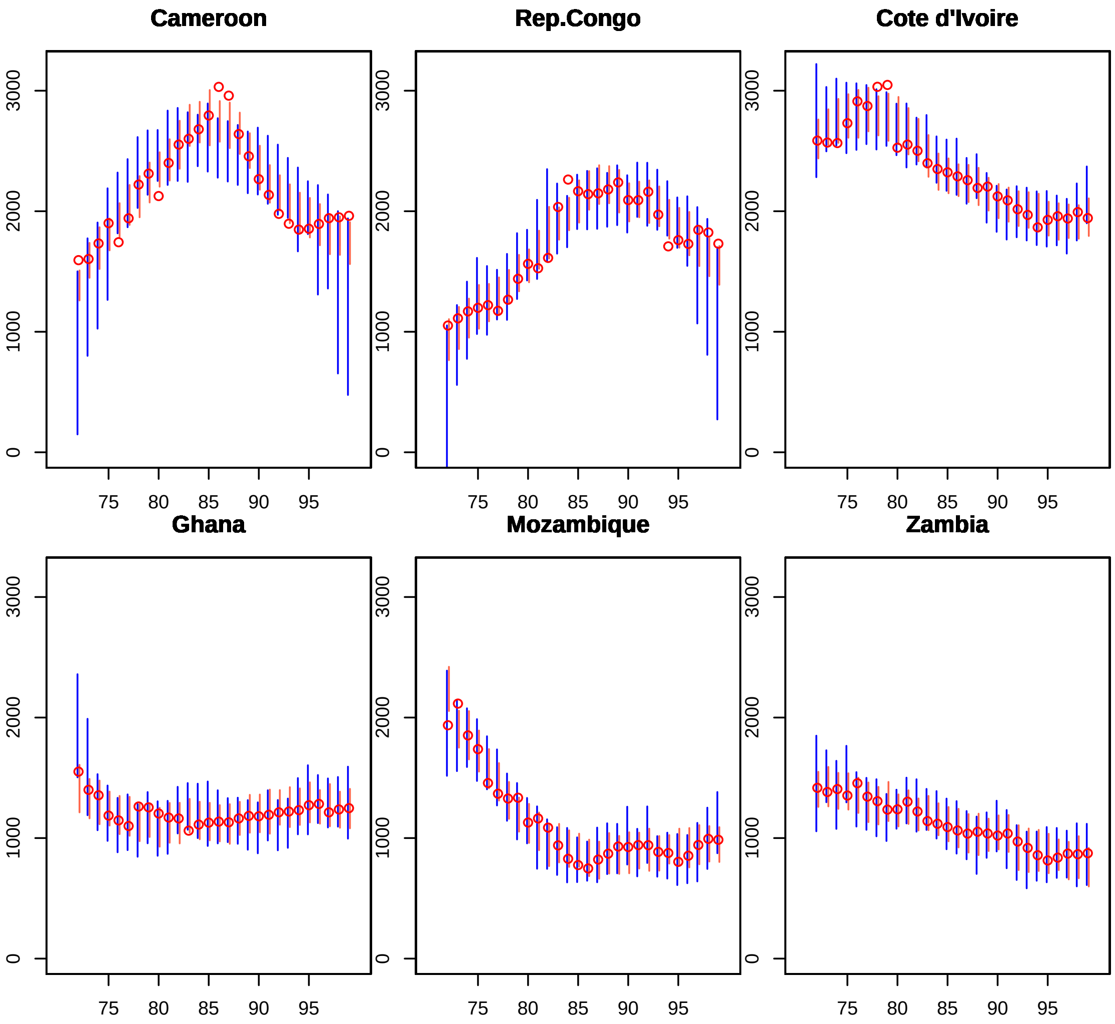

Figure 4: *The blue confidence intervals and data are the same as in Figure 3, from an imputation model with polynomials of time, whereas the shorter red confidence intervals are from a model that uses LOESS smoothing to form the basis functions.*

variables from the standard battery are included, switching "U.S. 2006" with "Barbados 1974" will do considerable violence to the information content in the data. 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 know much about them ex ante. Even when translated into regression coefficients, 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 post-treatment variables.

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 from auxiliary experiences: 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. Therefore mortality, which is surely higher due to the direct and indirect consequences of the conflict, is unobserved (Murray et al., 2002). Some of this information may be predictable from other variables in the data set, but if the MAR assumption is violated, this extra information can dramatically improve the quality of the imputations and ultimate analysis on which it is based. 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.

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 (2007, forthcoming, Chapter 5), if we restrict the transformation to the subspace spanned by the variables taking the role of covariates during an imputation — $D^{\text{obs}}_{i,-j}$ in the linear specification in Equation 3 — 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 $E(\tilde{D}_{ij}) = D^{\text{obs}}_{i,-j}\tilde{\beta}$ can be inverted to yield a prior on $\tilde{\beta} = (D^{\text{obs}\prime}_{i,-j}D^{\text{obs}}_{i,-j})^{-1}D^{\text{obs}\prime}_{i,-j}E(\tilde{D}_{ij})$, 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.[4]

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, 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 (Tsutakawa, 1992*a*; Clogg et al., 1991; Bedrick, Christensen and Johnson, 1996). Empirical priors, discussed in Section 2, 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

[4] In addition to the formal approach introduced for hierarchical models in Girosi and King (2007, forthcoming), 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 (West, Harrison and Migon, 1985; Tsutakawa and Lin, 1986; Tsutakawa, 1992*b*), wavelet analysis (George and Nanopoulos, 2001), and logistic (Clogg et al., 1991) and other generalized linear models (Bedrick, Christensen and Johnson, 1996; Greenland and Christensen, 2001; Greenland, 2001).

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.) This modification to 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. Appendix A gives the technical details.

We now illustrate our approach with a simulation from a model analyzed mathematically in Appendix A. This model is a bivariate normal with a prior on the expected value of observation D_{12}, given a true mean for variable 2 of zero.[5] Here, we add intuition by simulating one set of data from this model, setting the prior to $p(D_{12}) = N(5, \lambda)$, 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 5, we plot the posterior density of imputed values for priors of different strengths. As 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 5 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.)

[5] The parameters of the simulation are $\mu = (0, 0)$, $\Sigma = (1\ 0.4, 0.4\ 1)$.

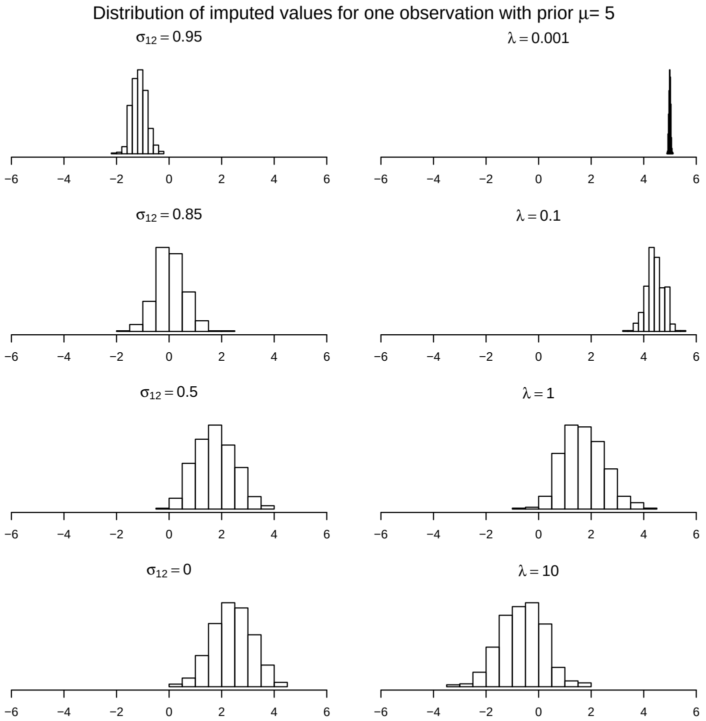

Figure 5: *Posterior densities of the expected value of one imputation generated from a model with a mean of zero and a prior mean of five. The left column holds constant the strength of the prior (summarized by the smallness of its variance, λ at 1) and changes the predictive strength of the data (summarized 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).*

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 6, 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 Figure 6 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[6], 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 6 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

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

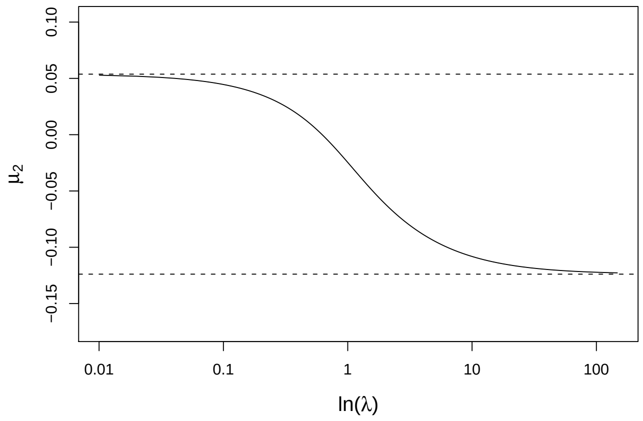

Figure 6: *Values of one model parameter μ_2, the mean of variable 2, with prior $p(x_{12}) = N(5, \lambda)$, across different strengths of the prior, $\ln \lambda$ (that is on the log scale). The parameter is approaching the theoretical limits (represented by dashed lines), where the upper bound is 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.*

priors on observations is putting on the model parameters. The key is to ensure that the covariates span a rich enough space to accomodate 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.

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 Appendix A), but the intuition in this simple case still applies.

6 Concluding Remarks

The new bootstrapping-based EM algorithm developed here makes it possible to include features in the imputation model that would have been difficult or impossible with existing approaches, resulting especially in more accurate imputations in the more time-series cross-sectional data sets. These techniques enable us to impose smoothness over time series variables, shifts over space, interactions between the two, and 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 scenerio 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. We therefore also make available, as a companion to this paper, an easy-to-use software package that implements all the methods discussed here. The nature of the algorithms and models developed here makes this software far more reliable than any existing imputation package. We hope this 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 colleages in American politics have had for years will not be available here. Future researchers may wish to take on the valuable task of using systematic methods of prior ellicitation (Gill and Walker, 2005; Kadane, 1980), and the methods introduced here, to impute some of the available public use data sets in these fields.

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, Hamilton, 1994; Beck and Katz, 1995).

A A Generalized Version of Data Augmentation Priors within EM

A.1 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, $m_i \neq \mathbf{1}\ \forall i$, but so that the ith subscripts represents the ith 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 $\mathbf{X}_{-0}$.

A.2 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 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 is why dropping all incomplete observations via listwise deletion is so pragmatically attactive, even though the resulting estimates are inefficient and often biased. Instead of rectangularizing the dataset by dropping *known* data, the EM algorithm rectangularizes the dataset 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, $T = \mathbf{X}'\mathbf{X}$ summarizes the sufficient statistics. Since the first column of $\mathbf{X}$ is a constant,

$$T = \begin{pmatrix} n & \mathbf{1X}_{-0} \\ \mathbf{X}_{-0}\mathbf{1} & \mathbf{X}'_{-0}\mathbf{X}_{-0} \end{pmatrix} = \sum_i \begin{pmatrix} n & x_{i1} & \dots & x_{ik} \\ x_{i1} & x_{i1}^2 & \dots & x_{i1}x_{ik} \\ \vdots & & \ddots & \\ x_{ik} & \dots & & x_{ik}^2 \end{pmatrix} \tag{4}$$

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 T 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 T on only the first row and column, results in

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

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 T, 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 T 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:

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

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:

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

The second is only slightly more complicated as:

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

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 T swept on the observed variables in some observation, i. Given these expectations, we can create a new rectangularized dataset, $\widehat{\mathbf{X}}$, in which we replace all missing values with their individual expectations given the observed data. Sequentially, every observation of this dataset can be constructed as:

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

The missing values within any observation have a variance-covariance matrix which can be extracted as a submatrix of θ as $\Sigma^{t+1}_{i|x_i^{\text{obs}}} = (M_i' M_i)\theta\{1{-}M_i\}^t$. 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 $\mathrm{E}[x_i' x_i] = \hat{x}_i^{t+1\prime}\hat{x}_i^{t+1} + \Sigma^{t+1}_{i|x_i^{\text{obs}}}$.

The M-step Given the construction of the expectations above, it is now simple to create an updated expectation of the sufficient statistics, T, by:

$$T^{t+1} = \sum_i \left(\hat{x}_i^{t\prime}\hat{x}_i^{t} + \Sigma^{t+1}_{i|x_i^{\text{obs}}}\right) = \widehat{\mathbf{X}}^{t+1\prime}\widehat{\mathbf{X}}^{t+1} + \sum_i \left(\Sigma^{t+1}_{i|x_i^{\text{obs}}}\right). \tag{10}$$

Convergence to the Observed Data Sufficient Statistics Throughout the iterations, the values of the observed data are of course constant, and generated from the sufficient statistics of the true data generating process we would like to estimate. In each iteration, the unobserved values have been filled in with our current guess of these sufficient statistics. One way to conceptualize the EM process is to realize that the sufficient statistics generated at the end of any iteration, θ^t, will be 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 arbitrarily bad and exceptionally far from the true values, but in the next round these parameters that were used 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 chain 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.

A.3 Incorporating a Single Prior

Conventionally, prior information is elicited as distributions over parameters in the model, which assumes knowledge of the relationships between variables or their marginal distributions. In contrast, researchers seem to have information about the realized value of an element missing from the data set and thus we need to be able to add prior information about one observation.

EM algorithms incorporate prior information in the M-step, because this is the step where the parameters are updated, and prior information is always assumed to inform the posterior of the parameters. Instead, we have information that informs the distribution of particular missing information in the dataset. As the elements of the dataset are updated in the E-step, we want to 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 (2007, forthcoming), but in the EM algorithm for missing data we have to explicitly construct expectations 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, T, these are given by (7) and (8). Incorporating the prior, the expectation becomes:

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

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: $\text{Var}(x_{ij}, x_{ij}) = \left[\lambda^{-1} + (\theta\{1-M_i\}_{jk}^t)^{-1}\right]^{-1}$, and calculation of the covariances are left for the more general explanation of multivariate priors in Section A.4.

Example Consider the following simplified example with a latent bivariate dataset 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:

$$M = \begin{pmatrix} 0 & 0 & 1 \\ 0 & 0 & 1 \\ 0 & 0 & 0 \\ \vdots & \vdots & \vdots \\ 0 & 0 & 0 \end{pmatrix} \tag{12}$$

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

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

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

$$\theta\{(1\ 1\ 0)\}^t = \begin{pmatrix} . & . & \mu_2 - \mu_1\sigma_{11}^{-1}\sigma_{12} \\ . & . & \sigma_{11}^{-1}\sigma_{12} \\ \mu_2 - \mu_1\sigma_{11}^{-1}\sigma_{12} & \sigma_{11}^{-1}\sigma_{12} & \sigma_{22} - \sigma_{21}\sigma_{11}^{-1}\sigma_{12} \end{pmatrix} \tag{14}$$

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

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.

A.4 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.[7] Thus,

$$\Lambda^{-1} = \begin{pmatrix} \lambda_{11}^{-1} & 0 & \dots & 0 \\ 0 & \lambda_{22}^{-1} & & \vdots \\ \vdots & & \ddots & \\ 0 & \dots & & \lambda_{kk}^{-1} \end{pmatrix} \tag{16}$$

and furthermore define $\lambda_{jj}^{-1} = 0$ for all observed elements j or unobserved elements for which there is no prior.

The posterior distribution x_i^{mis} can now be defined as:

$$\begin{aligned} \mu_i^* &= (\Lambda_i^{-1} + (\Sigma_{i|x_i^{\text{obs}}}^{t+1})^{-1})^{-1}(\Lambda_i^{-1}\mu_{0_i} + (\Sigma_{i|x_i^{\text{obs}}}^{t+1})^{-1}\hat{x}_i^{\text{mis}^{t+1}}) && (17) \\ \Sigma_i^* &= (\Lambda_i^{-1} + (\Sigma_{i|x_i^{\text{obs}}}^{t+1})^{-1})^{-1} && (18) \end{aligned}$$

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).[8] When the EM algorithm has converged, these terms will be also be used for the final imputations as:

$$(\tilde{x}_i | \mathbf{X}^{\text{obs}}, M, \lambda, \mu_0) \sim N(\mu_i^*, \Sigma^*) \tag{19}$$

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 (17) and (18) and the I-step replaced by a draw from (19).

A.5 Convergence Diagnostics

EM is a highly reliable algorithm but, like most optimization procedures, it is only guaranteed to converge to a local maxima. Usually the local maximum is also the global

[7] 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.

[8] In (17) μ_{ij}^* simplifies to $\hat{x}_{ij}^{t+1}$ for any missing element ij for which there is no prior specified, that is, where $\lambda_{jj}^{-1} = 0$.

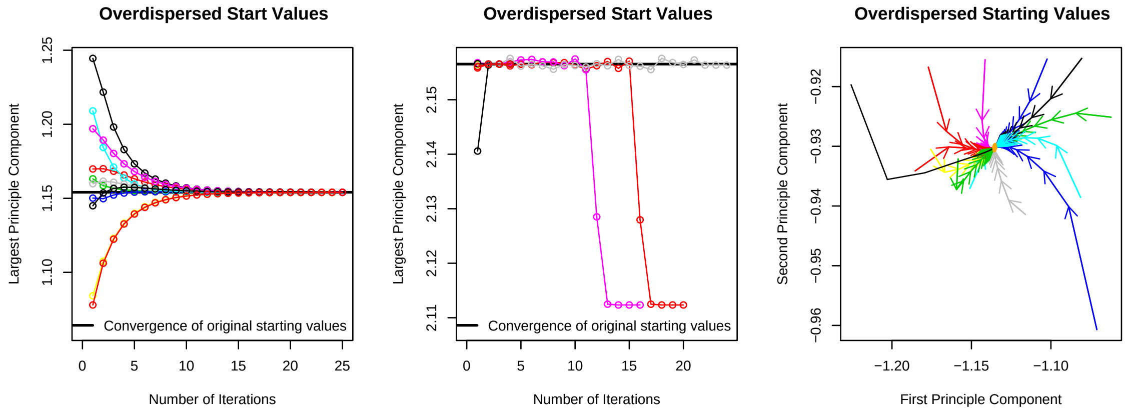

Figure 7: EM Convergence Diagnostics

maximum, but sometimes this does not happen and so it is worth checking. We thus introduce some graphical methods of assessing convergence in EM algorithms. We do this by starting multiple chains at overdispersed starting values. Since the trace of the iterations for each chain is in very high dimensions, and so obviously cannot be displayed directly, we summarize them with the first principal component or two of the

parameters at value of the final converged mode or modes. Figure 7 gives examples of three such plots. The first two graphs plot iteration numbers (horizontally) by the largest principal component of all the parameters at the point or points of convergence (vertically). All the overdispersed starting values in the first graph converge cleanly to a single mode, whereas in the second graph the overdispersed points converge to two separate modes. When problems like that in the middle graph are discovered, it is a simple matter to decide which is right based on which converged to a higher likelihood value. The third graph gives an example of a way to view the first two principal components (displayed on the two axes), with the length of each connected arrow representing the distance traversed between successive iterations. All the chains of iterations in this graph also clearly converge to a single point.

References

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

Beck, Nathaniel and Jonathan Katz. 1995. ""What to Do (and Not to Do) with Time-Series-Cross-Section Data"." *American Political Science Review* 89:634–647.

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):1450–1460.

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, March):68–78.

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, June):463–475.

George, E. and P. Nanopoulos, eds. 2001. *Model Selection for Cepheid Star Oscillations.* Jefferys, Barnes and Rodrigues Univ TX at Austin, Berger and Muller, Duke Univ: ISBA and Eurostat, Official Publications of the European Communities, Luxembourg 253 –252.

Gill, Jeff and Lee Walker. 2005. "Elicited Priors for Bayesian Model Specification in Political Science Research." *Journal of Politics* 67(3, August):841–872.

Girosi, Federico and Gary King. 2007, forthcoming. *Demographic Forecasting.* Princeton: Princeton University Press. http://gking.harvard.edu/files/abs/smooth-abs.shtml.

Greenland, Sander. 2001. "Putting Background Information About Relative Risks into conjugate Prior Distributions." *Biometrics* 57(September):663–670.

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* 20:2421–2428.

Hamilton, James Douglas. 1994. *Time Series Analysis.* Princeton: Princeton University Press.

Ibrahim, Joseph G. and Ming-Hui Chen. 1997. "Predictive Variable Selection for the Multivariate Linear Model." *Biometrics* 53(June):465–478.

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

King, Gary, James Honaker, Anne Joseph and Kenneth Scheve. 2001. "Analyzing Incomplete Political Science Data: An Alternative Algorithm for Multiple Imputation." *American Political Science Review* 95(1, March):49–69. http://gking.harvard.edu/files/abs/evil-abs.shtml.

King, Gary and Langche Zeng. 2006. "When Can History Be Our Guide? The Pitfalls of Counterfactual Inference." *International Studies Quarterly* . forthcoming, copy at http://gking.harvard.edu/files/counterf.pdf.

King, Gary, Michael Tomz and Jason Wittenberg. 2000. "Making the Most of Statistical Analyses: Improving Interpretation and Presentation." *American Journal of Political Science* 44(2, April):341–355. http://gking.harvard.edu/files/abs/making-abs.shtml.

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

Laud, Purushottam W. and Joseph G. Ibrahim. 1995. "Predictive Model Selection." *Journal of the Royal Statistical Society, B* 57(1):247–262.

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

Meng, Xiao-Li. 1994. "Multiple-Imputation Inferences with Uncongenial Sources of input." *Statistical Science* 9(4):538–573.

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(February 9):346–349. http://gking.harvard.edu/files/abs/armedph-abs.shtml.

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, Jun):475–478.

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):366–374.

Schafer, Joseph L. 1997. *Analysis of incomplete multivariate data*. London: Chapman & Hall.

Schafer, Joseph L. and Maren K. Olsen. 1998. "Multiple Imputation for multivariate Missing-Data Problems: A Data Analyst's Perspective." *Multivariate Behavioral Research* 33(4):545–571.

Shao, Jun and Randy R. Sitter. 1996. "Bootstrap for Imputed Survey Data." *Journal of the American Statistical Association* 91(435, September):1278–1288.

Tsutakawa, Robert K. 1992*a*. "Moments Under Conjugate Distributions in Bioassay." *Statistics & Probability Letters* 15(October):229–233.

Tsutakawa, Robert K. 1992*b*. "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, June):251–267.

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, March):73–83.