Using conditional distributions for missing-data imputation*

Andrew Gelman† and T. E. Raghunathan‡

June 21, 2001

1 Introduction

The authors discuss conditionally-specified models in probability theory and for modeling joint distributions in various applications. This theoretical structure is useful, considering that conditional models are becoming standard in many spatial applications, following Besag (1974). (Rather than attempting an exhaustive or ever representative list, we shall just refer to Besag and Higdon (1999) as a recent example with discussion.) In addition, there has been occasional discussion in the literature as to the relative merits of conditionally or jointly-specified models (for example, Besag, 1974, Haslett, 1985, and Ripley, 1988).

Here, however, we would like to address a different topic: the use of conditional distributions, not to model an underlying joint distribution, but for the purpose of imputing missing data. At first this might seem like an unimportant distinction—after all, imputation requires modeling (if only implicitly). However, when the fraction of missing data is not large, imputations can be reasonable even if they are not based on the correct complete-data model (see Meng, 1994, and Rubin, 1996). Thus, it makes sense to consider modeling for imputation separately from modeling of underlying phenomena.

We shall refer to the example of the New York City Social Indicators Survey (Garfinkel and Meyers, 1997), where we had to impute missing responses for family income conditional on demographics and information such as whether or not anyone in the family received government welfare benefits. Conversely, if the "welfare benefits" indicator is missing, then family income is clearly a useful predictor. The whole situation was actually more complicated because the survey asked about several different sources of income, and these questions had different patterns of nonresponse.

*Discussion of "Conditionally specified distributions" by Arnold et al. To appear in *Statistical Science*. We thank the U.S. National Science Foundation for support through grant SES-9987748 and SES-0084368.

†Department of Statistics, Columbia University, New York

‡Department of Biostatistics and Institute for Social Research, University of Michigan, Ann Arbor

2 Inconsistent conditional distributions

As discussed by the authors, a multivariate normal distribution has conditionals that are normal linear regressions, in which case the conditional distributions are automatically compatible. However, when any of these conditions is relaxed—that is, if data are bounded or discrete (and thus cannot be modeled as normal), or regression relationships are nonlinear or have interactions—then, in general, reasonable-seeming conditional models will not be compatible with any single joint distribution.

Nonetheless, imputations can be performed using conditional models: that is, one can start with guesses of all the missing data, then impute $x_1|x_2, x_3, \ldots, x_k$, impute $x_2|x_1, x_3, \ldots, x_k$, and so forth, looping indefinitely through all the variables. If the imputations are stochastic, this is just the notorious "inconsistent Gibbs" algorithm, for which the simulation draws never converge to a single joint distribution; rather, the distribution depends upon the order of the updating and on when the updating is stopped.

With the inconsistent Gibbs sampler, one is always afraid of reasonable-seeming conditional distributions that produce a diverging random walk—for example, if $x_1|x_2 \sim \mathrm{N}(x_2, 1)$, and $x_2|x_1 \sim \mathrm{N}(x_1, 1)$, then the distribution of the simulations simply diffuses out to infinity. However, in practice, with the distributions estimated from data (and using constraints or proper prior distributions when dimensions are high and data sparse), this should not happen.

A big advantage of conditional (rather than joint) modeling is that it splits a k-dimensional problem into k one-dimensional problems, each of which can be attacked flexibly. Thus, conditional imputation using k separate regression models is a popular approach, and it has recently been formalized by Raghunathan et al. (2001) and implemented in SAS-compatible software (Raghunathan et al., 1997). This particular program allows continuous variables to be modeled using normal distributions, binary variables with logistic regression, with other options for ordered and unordered discrete variables and for continuous variables with constraints. The corresponding joint posterior distribution may not exist, of course, which means that the Bayesian inference used to get uncertainties for the imputations is only uncertain. (It could, however, possibly be formalized as a Bayesian counterpart to the pseudolikelihood (Besag, 1975), in which the likelihood function is replaced by the product of conditional densities.)

Performing imputation is awkward without a joint model, and it also results in difficulties in inference for the imputation model itself (for example, how do you correctly adjust for truncation in a bounded-variable model when there is no joint distribution over which to integrate). However, the separate regressions often make more sense than joint models which either assume normality and hope for the best (Gelman et al., 1998) or mix normality with completely unstructured discrete distributions (Schafer, 1997) or mix normality (with random effects) and log-linear structures for

discrete distributions (Raghunathan and Grizzle, 1995) or generalize with the t distribution (Liu, 1995). From a practical perspective, all these approaches provide useful tools, and some of the time it will make sense to go with the inconsistent, but flexible, conditional models such as described by Raghunathan et al. (2001).

One may argue that having a joint distribution in the imputation is less important than incorporating information from other variables and unique features of the data set (such as zero/nonzero features in income components, bounds, skip patterns, nonlinearity, interactions, and so forth). Conditional modeling allows enormous flexibility in dealing with practical problems. We have never been able to apply the joint models to a real data set without making drastic simplifications.

But, if one is modeling some aspect of the nature, then the joint distribution has to be the end point. Specifying just the conditionals without a coherent joint distribution will not be acceptable. Having said that, many of our applied collaborators, are just as happy with conditionals such as p(Hypertension | Body Mass Index) or p(Body Mass Index | Hypertension, Socioeconomic Status), rather than p(Hypertension, Body Mass Index, Socioeconomic Status).

3 Choices in setting up the imputation models

We conclude with a discussion of an awkward (or perhaps promising) issue: structural features of the conditional models can affect the distributions of the imputations in ways that are not always obvious. To return to the example introduced at the end of Section 1 of this discussion, suppose we are imputing a continuous income variable y_1, and a binary indicator y_2 for welfare benefits, conditional on a set X of fully-observed covariates.

We can consider two natural approaches. Perhaps simplest is a direct model where, for example $p(y_1|y_2, X)$ is a normal distribution (perhaps a regression model on y_2, X, and the interactions of y_2 and X) and $p(y_2|y_1, X)$ is a logistic regression on y_1, X, and the interactions of y_1 and X. (For simplicity, we ignore the issues of nonnegativity and possible zero values of y_1.)

A more elaborate, and perhaps more appealing model uses hidden variables: let z_2 be a latent continuous variable, defined so that

$$y_2 = \begin{cases} 1 & \text{if } z_2 \geq 0 \\ 0 & \text{if } z_2 < 0. \end{cases} \tag{1}$$

We can then model $p(y_1, z_2|X)$ as a joint normal distribution (that is, a multivariate regression). Compared to the direct model, this latent-variable approach has the advantage of a consistent joint distribution. And, once inference for (y_1, z_2) has been obtained, we can directly infer about y_2 using (1). In addition, this model has the conceptual appeal that z_2 can be interpreted as some sort of continuous "proclivity" for welfare, that is only activated if it exceeds a certain threshold. In fact,

the relation between z_2 and y_2 can be made stochastic if such a model would appear more realistic.

So the latent-variable model is better (except for possible computational difficulties), right? Not necessarily. A perhaps-disagreeable byproduct of the latent model is that, because of the joint normality, the distributions of income among the welfare and non-welfare groups—that is, the distributions $p(y_1|y_2 = 1, X)$ and $p(y_1|y_2 = 0, X)$—must necessarily overlap. In contrast, the direct model allows there to be overlap or non-overlap, depending on the data. Thus, although the latent-variable model seems to be a generalization, it is not.

4 Conclusions

Where does this leave us in practice? Must we just choose a model and hope for the best? Fortunately, we are not completely without tools—in particular, we can use a procedure to impute missing data and then check the fit of the model to the completed dataset (Gelman et al., 1998, 2001). Serious problems (such as overlapping distributions for imputed data amidst nonoverlapping distributions of observed data) should show up. With checking, we should be able to notice major flaws in an imputation model. But we do not have a good sense of how general the models have to be in order to work well, and it is not clear when incompatibility of conditional distributions presents a practical problem.

As with so much of statistics, the study of conditional distributions is an area where theory has not caught up with practice.

References

Besag, J. E. (1974). Spatial interaction and the statistical analysis of lattice systems (with discussion). *Journal of the Royal Statistical Society B* **36**, 192–236.

Besag, J. E. (1975). Statistical analysis of non-lattice data. *The Statistician* **24**, 179–196.

Besag, J. E., and Higdon, D. (1999). Bayesian analysis of agricultural field experiments (with discussion). *Journal of the Royal Statistical Society B* **61**.

Garfinkel, I., and Meyers, M. K. (1999). A tale of many cities: the New York City Social Indicators Survey. School of Social Work, Columbia University.

Gelman, A., King, G., and Liu, C. (1998). Not asked and not answered: multiple imputation for multiple surveys (with discussion and rejoinder). *Journal of the American Statistical Association* **93**, 846–874.

Gelman, A., Van Mechelen, I., Verbecke, G., Heitjan, D. F., and Meulders, M. (2001). Bayesian model checking for missing and latent data problems using posterior predictive simulations.

Technical report, Department of Statistics, Columbia University.

Haslett, J. (1985). Maximum likelihood discriminant analysis on the plane using a Markovian model of spatial content. *Pattern Recognition* **18**, 287–296.

Liu, C. (1995). Monotone data augmentation using the multivariate t distribution. *Journal of Multivariate Analysis* **53**, 139–158.

Meng, X. L. (1994). Multiple-imputation inferences with uncongenial sources of input (with discussion). *Statistical Science* **9**, 538–573.

Raghunathan, T. E.,and Grizzle, J. E. (1995). A split questionnaire survey design. *Journal of American Statistical Association* **90**, 55–63.

Raghunathan, T. E., Lepkowski, J. E., Solenberger, P. W., and Van Hoewyk, J. H. (2001). A multivariate technique for multiply imputing missing values using a sequence of regression models. *Survey Methodology*, to appear.

Raghunathan, T. E., Solenberger, P. W., and Van Hoewyk, J. H. (1997). IVEware: imputation and variance estimation software. `http://www.isr.umich.edu/src/smp/ive`

Ripley, B. D. (1988). *Statistical Inference for Spatial Processes.* Cambridge University Press.

Rubin, D. B. (1996). Multiple imputation after 18+ years (with discussion). *Journal of the American Statistical Association* **91**, 473–520.

Schafer, J. L. (1997). *Analysis of Incomplete Multivariate Data.* London: Chapman and Hall.