Web Appendix to "A Fast, Easy, and Efficient Estimator for Multiparty Electoral Data"[*]

James Honaker
Department of Political Science
University of California, Los Angeles, Los Angeles, CA 90095-1472
e-mail: tercer@ucla.edu

Jonathan N. Katz
Division of Humanities and Social Science
California Institute of Technology, Pasadena, CA 91125
e-mail: jkatz@caltech.edu

Gary King
Department of Government
Harvard University, Cambridge, MA 02138
e-mail: king@harvard.edu

We begin by describing the existing EMis algorithm for multivariate normal data (Section 1) and briefly summarize some useful properties of the multivariate t density (Section 1.1). We then summarize changes in the EMis algorithm we made to accommodate t-distributed data (Section 2). We follow up on points made in the paper on the effects of numerical integration on mean squared error in the original FIML model of Katz and King (KK). Next we incorporate the constraints for partially contested or uncontested seats (Section3.1) as laid out in the original KK model. We also summarize the use of the t regression analysis model (Section 3.2) and discuss how long the algorithm takes (Section 4).

1. THE EMIS ALGORITHM

The EM algorithm (Expectation Maximization) (Orchard and Woodbury 1972; Dempster et. al 1977; McLachlan and Krishnan 1996) is an increasing popular approach to finding maximum likelihood estimates of systems that are intractable or highly complicated analytically. EM is an iterative deterministic algorithm which under given regularity conditions increases the likelihood of its parameter estimates monotonically on every iteration. The EM algorithm has seen use in political science by Lewis (1998), Bailey (1998), and Jackman (2000).

The EMis (Expectation Maximization with importance resampling) multiple imputation algorithm is an alternative to data augmentation (Schafer 1998) and builds on the EM algorithm by importance resampling (Gelfand and Smith 1990; Gelman et al. 1995; King et al. 2001; Rubin 1987a 192-194 1987b; Tanner 1996; Wei and Tanner 1990). The EMis algorithm is as follows: (1) calculate the maximum posterior of the data using the EM algorithm; (2) estimate the variance of this point estimate in the space of the sufficient statistics; (3) construct an approximating distribution of the posterior likelihood of the sufficient statistics; (4) importance resample m sets of sufficient statistics from this approximating distribution using the actual posterior likelihood; and (5) impute the missing values, D_{mis}, using each of the above samples to create m completed datasets.

Given the maximum posterior estimate of the parameters $\hat{\theta} = (\vec{\hat{\mu}}, \hat{\Sigma})$ one computes its variance $V(\hat{\theta})$ after reparameterizing to unbounded scales using the log for the standard deviations and the Fisher's z for the correlations. For small dimensions the variance can be computed with the negative of the inverse of the Hessian; for moderate dimensions the outer product of the gradient;

2

and for large dimensions the variance of simulations from some appropriate Markov chain run in the vicinity of the maximum can be used. Since calculation of the variance is not effected by dependency in these draws, and thus the typical autocorrelation checking do not need to be decided by user monitoring, this is more easily automated than typical MCMC methods. Although each of these three methods is less accurate than the previous one, our analyses convince us that they serve quite well to create an approximating distributing for θ. From this approximating distribution one then uses an acceptance-rejection algorithm by keeping draws of $\tilde{\theta}$ with probability proportional to the "importance ratio" — the ratio of the actual posterior to the asymptotic normal (or multivariate t) approximation, both evaluated at $\tilde{\theta}$ — and discarding the rest. Without priors, the importance ratio is $L\left(\tilde{\theta} \mid D_{obs}\right) / N\left(\tilde{\theta} \mid \hat{\theta}, V(\hat{\theta})\right)$.

In importance resampling, one often wants the approximating distribution (also known as the covering distribution) to have thicker tails then the true distribution to increase confidence that the distribution is being properly approximated everywhere. This can be done by multiplying the variance computed above by some common factor (generally 1.2–1.5 is used as a rule of thumb), or covering a normal with a t distribution with low degrees of freedom. A useful diagnostic can be extracted from the fact that the larger the ratio of the number of draws from the approximating distribution to the number of acceptances needed (here, m, the number of imputed datasets) the better the approximation.

1.1. *Some Useful Properties of the t Distribution*

We offer here a brief summary of properties of the multivariate t distribution that we find useful. If Y_i is distributed:

$$Y_i \overset{ind}{\sim} \mathrm{N}(\mu, \Psi/u_i) \tag{1}$$

$$u \overset{iid}{\sim} \chi_\nu^2/\nu \tag{2}$$

where $\nu > 0$, then Y is distributed as

$$Y \overset{iid}{\sim} \mathrm{t}\left(\mu, \Psi, \nu\right). \tag{3}$$

The complete-data likelihood, for known weights, is then separable.

$$L(\mu, \Psi, \nu|Y, u) = L_N(\mu, \Psi|Y, u) + L_G(\nu|u),\tag{4}$$

where

$$L_G(\nu|u) = -n \ \ln \ \left(\Gamma\left(\frac{\nu}{2}\right)\right) + \frac{n\nu}{2} \ \ln \ \left(\frac{\nu}{2}\right) + \frac{\nu}{2}\sum_{i=1}^{n} \left(\ln(u_i) - u_i\right).\tag{5}$$

2. AN EMIS ALGORITHM FOR T DISTRIBUTED DATA

We followed the framework of the EMis algorithm to impute the effective vote in constituencies where not all the parties ran, and to deal with missingness we had in the covariates. The EM algorithm itself is often implemented under the assumption that the data are distributed normally, but this distributional assumption can be changed. The EM algorithm retains its simplicity if the E and particularly the M steps are non-iterative themselves and do not involve hard to maximize likelihoods. This can be done easily with the t distribution by the use of the decomposition in equation 1. We take the vector of weights u to be an additional variable (completely unobserved) to be imputed in the dataset, and the degrees of freedom ν as an additional element to θ. The t distributed EM algorithm then resembles the normally distributed EM algorithm and can be driven with the same shortcuts, such as the sweep operator (Schafer 1998), except that the sums and sums of squares and cross-products computed for the M-step need to be appropriately weighted by u.

We began with an EM algorithm for t distributed data but found convergence to be extremely slow. Similar to Lange et.al. (1989), we found results were actually faster by running separate EM algorithms each conditional on some value of ν over a grid of ν values. To speed up convergence we instead used the ECME algorithm (Liu 1994, Liu and Rubin 1994) to find the MLE of θ, a description of which follows.

The E-step of ECME is the same as the E-step in EM. The elements of Y_{mis} are filled in with their expected values from current estimates of μ and Ψ as in the EM algorithm[1]. The vector of weights u^{t+1} is similarly created from the expectation:

$$E(u_i^{t+1}) = \frac{p_i + \nu^{(t)}}{\delta_{i,obs}^{(t+1)} + \nu^{(t)}}\tag{6}$$

4

where p_i is the number of variables and δ known as the Mahalanobis distance is given by:

$$\delta_{i,obs}^{(t+1)} = (Y_{i,obs} - \mu_{i,obs})' \Psi_{i,obs}^{-1} (Y_{i,obs} - \mu_{i,obs}). \tag{7}$$

Thus observations which can be considered as outliers have large Mahalanobis distances and are down-weighted.

After the E-step are two conditional maximization steps (CM). First we maximize the Q-function (the constrained expected log-likelihood) over $\theta_1 = (\mu, \Psi)$ given ν. Then maximize the L-function (the constrained actual log-likelihood) over $\theta_2 = \nu$ given $\theta_1 = (\mu, \Psi)$. To do this a one dimensional search is implemented over equation 5. This function is globally concave and has an analytical derivative making it simple to maximize with a search such as Newton's method.

Under mild conditions the ECME algorithm has the convergent properties of GEM algorithms although it is not itself a special case (Liu and Rubin, 1994). It would be GEM if we maximized the Q-function over $\theta_2 = \nu$ given $\theta_1 = (\mu, \Psi)$ instead, but in most problems this approach leads to much more rapid convergence over ν (Liu 1994). Indeed, with the t distribution, since the likelihood is separable by equation 4, if we maximized the Q-function rather than the L-function we would have again exactly the EM algorithm.

The maximum of the posterior provided by the ECME algorithm substitutes for the value that would be provided by EM in the EMis algorithm. For the importance resampling in the applications that follow we used a covering distribution resembling the "witch's hat" distribution with a t distributed peak trailing to a constant valued "brim" on the joint μ and Σ parameters and a χ^2 distribution on $(\nu - 2)$ with mean $(\hat{\nu} - 2)$. By monitoring the distribution of the importance ratio, and studying simulated data, we were confident that the true distribution had been properly covered. This was also confirmed strongly, albeit indirectly, in the analyses presented in the paper.

3. MEAN SQUARED ERROR IN KK WITH INCREASINGLY PRECISE NUMERICAL INTEGRATION

To explore the effects of imprecision of numerical integration, we replace the methods used by KK which we used in figure 1 in the paper with midpoint quadrature. For low numbers of terms of

quadrature the FIML method does much more poorly then our method, while for very many terms of quadrature it seems to exceed the performance of our method. Figure 1 compares the mean squared error of the FIML method with seven, forty and five thousand terms of quadrature as the respective descending solid lines. Comparing these to the MSE reported by our method in the earlier figure (here shown again as the dotted line) we see the sandwiching of the MSE by the two extremes. Thus, in this three party example, devoting enough computational resources to the FIML method will allow it to outperform our approximating method.

[Figure 1 about here.]

3.1. *Imposing Uncontestedness Constraints*

Sometimes imputations from missing data models are not appropriate to the user's analysis or fail known bounds or identities of the data. For some special cases, such as with ordinal and nominal variables, it is possible to directly transform the normal model posterior to address the constraint of discreteness in a logically consistent way. We have analogous, although somewhat more complicated, constraints to implement.

The KK model imposes the constraint that "the noncontesting parties would have received fewer votes than the parties which did nominate candidates" and thus the effective vote in some district of any party which did not contest in that district must be lower than the effective vote of all other parties which did run candidates in that district. We impose this constraint in the imputation model by rejection sampling/resampling from the t model. For a given imputed dataset, $j \in 1, \ldots, M$, with sufficient statistics θ_j, each observation is checked as to whether it meets the model constraint[2]. In each round, any observation, y_i, which fails the constraint is redrawn from $P(\theta_j | y_i, u_i)$. The number of failing observations, which is a useful diagnostic and reported by our software, is necessarily non-increasing in each round. This is iterated until all observations pass. This approach can be tailored, with different check functions, to a broad range of analyst constraints that might fit in any particular application. As discussed in the paper, the distribution from which the final imputations are drawn will then be the truncation (to the limits of the constraints) of the unbounded distribution that maximizes the likelihood of the observed data, whereas the FIML model derives the truncated

6

distribution that maximizes the likelihood of the observed data.

3.2. *The t Regression Analysis Model*

Part of the appeal of the multiple imputation framework is that it separates the model of missingness from the model of analysis. Once the data sets are imputed, the user can apply whatever model he or she would have used if the dataset had arrived fully observed. In the present application, the researcher can analyze the data as if all parties contested elections in every district, which would normally require the application of a multivariate t-regression model.

As an easier alternative, a reasonable approximation would be to use t-regressions conditional on the weights output from the imputation stage. This means that the user only needs to run one (noniterative) weighted least squares analysis, for each imputation, and to average the results as in multiple imputation. Thus, Ameila will provide (say) five sets of imputed effective vote data, along with any covariates provided (with their missing data, if any, also imputed) and a weight. The user will then run a set of weighted least squares regressions, using any statistical package. The dependent variable is the log-ratios of the effective votes, the weight is as provided by Amelia, and the results are averaged.

4. HOW LONG DOES IT TAKE?

To run the series of ten elections under the original KK FIML approach took 35 minutes. To multiply impute the effective votes (and the missing values in the covariates) for all ten elections, the first stage of our alternative algorithm, took 22 minutes, after which the analysts model must be run on each imputed dataset. Using our augmented weighted least squares approach for all ten takes only a few seconds, so for all practical purposes the time for analysis is essentially the time taken by the imputation model. In practice, researchers are thus asked to invest 22 minutes in imputation time, and can then run as many analysis models as they like, each nearly as quickly as any other regression analysis. In cases with more parties, KK is infeasible but our approach scales up well, approximately as does Amelia.

For a model that cannot be rewritten as Weighted Least Squares, the researcher must balance

the additional computational time required to run the original model on each of the M imputed datasets[3] versus the analyst's time in writing and programming a more complicated model. In addition, if there are missing values in the covariates, the imputation approach has the further benefit of increasing efficiency and potentially correcting bias.

Timing is greatly effected by the number of variables in the imputation model, as in the original EMis algorithm (King, Honaker, Joseph, and Scheve, 2001). The number of variables increases with the number of patterns of party contestation across districts. For a given number of (possibly incomplete) covariates, a dataset with a very large number of parties, but where almost all parties contest all districts may take less time to impute the effective vote than a dataset with a small number of parties each of which contest randomly[4].

Notes

[1] Two families of EM algorithms are possible. In one, the completed data is stored (Beale and Little, 1975) in the other the sufficient statistics (sums, sums of squares, and sums of cross products) are stored (Dempster, Laird, and Rubin 1977). Done properly they are equivalent. (A mixture of the two is also possible, storing sufficient statistics for nearly completed observations, and raw data otherwise (Little and Rubin, 1987).) We opt for the first of these methods in the exposition in this paper and in our code because it seems conceptually simpler and more intuitive, and was a faster implementation in GAUSS as it can be written to draw on GAUSS's strength in large matrix algebra computations and avoid GAUSS's weakness in looping.

[2] To do this, partition the imputed effective vote of party j in partially contested district i by $V_{ij} \in R_i^+$ if j originally ran and $V_i j \in R_i^-$ if the party did not run a candidate. The boolean is then $\max(R_i^-) < \min(R_i^+)$.

[3] In the KK example, although we could run WLS we also ran the original maximum likelihood model. To run through the ten imputed elections took only 13.5 minutes, roughly two-fifths the time of the original model, but this needed to be iterated on each of the M datasets, where we chose $M = 10$ for a total of 157 minutes including the time taken by the imputation model.

[4] For k covariates and n parties, of whom p partially contest some districts, there may be up to $k + \sum_{i=0}^{p} \binom{p}{i}(n-i-1)$ total variables in the imputation model.

REFERENCES

Aitchison, J. 1986. *The Statistical Analysis of Compositional Data.* London: Chapman and Hall.

Bailey, Michael. 1998. "Ideal Point Estimation with a Small Number of Votes: A Random-Effects Approach." Political Analysis 9(3): 192-210.

Beale, E.M.L. and R.J.A. Little 1975. "Missing data in multivariate analysis." *Journal of the Royal Statistical Society Series B*, 37: 129-145.

Dempster, A.P., N.M. Laird, and D. Rubin. 1977. "Maximum Likelihood From Incomplete Data Via the EM Algorithm (with discussion)." *Journal of the Royal Statistical Society*, Ser. B, 39, 1-38.

Gelfand, A.E. and A.F.M. Smith. 1990. "Sampling-based approaches to calculating marginal densities." *Journal of the American Statistical Association.* 85: 398–409.

Gelman, Andrew, John Carlin, Hal Stern, and Donald Rubin. 1995. *Bayesian Data Analysis.* New York: Chapman and Hall.

Gelman, Andrew and Gary King. 1994. "A Unified Method of Evaluating Electoral Systems and Redistricting Plans." *American Journal of Political Science* 38(2)(May): 514-554.

Gibson, John and Anna Cielecka. 1995. "Economic Influences on the Political Support for Market Reform in Post-communist Transitions: Some Evidence from the 1993 Polish Parliamentary Elections." *Europe-Asia Studies* 47(5): 765-785.

Katz, Jonathan and Gary King. 1999. "A Statistical Model for Multiparty Electoral Data." *American Political Science Review* 93(1)(March): 15-32.

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.

Lange, Kenneth L., Roderick J. A. Little, and Jeremy M. G. Taylor. 1989. "Robust Statistical Modeling Using the t Distribution." *Journal of the American Statistical Association* 84, 408: 881-896.

Lewis, Jeffrey B. 1998. "Estimating Voter Preference Distributions from Individual-Level Voting Data." *Political Analysis* 9(3): 275-297.

Little, Roderick J. A. 1988. "Robust Estimation of the Mean and Covariance Matrix from Data with Missing Values." Applied Statistics 37, 1: 23-38.

Little, Roderick J. A., Donald Rubin. 1987. *Statistical Analysis with Missing Data.* New York: Wiley.

Liu, Chuanhai. 1994. *Statistical Analysis Using the Multivariate t Distribution.* Disertation. Harvard University, Cambridge MA.

Liu, Chuanhai and Donald Rubin. 1994. "A Simple Extension to EM and ECM with Faster Monotone Convergence." *Biometrika* 81(4): 633-648.

McLachlan, Geoffrey and Thiriyambakam Krishnan. 1996. *The EM Algorithm and Extensions.* New York: Wiley.

Orchard, T. and Woodbury, M. A. 1972. "A missing information principle: Theory and applications." *Proceedings of the 6th Berkeley Symposium on Mathematical Statistics and Probability.* Berkeley: University of California Press, 697-715.

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

Rubin, Donald. 1987b. "A noniterative sampling/importance resampling alternative to the data augmentation algorithm for creating a few imputations when fractions of missing information are modest: the SIR algorithm. Discussion of Tanner and Wong." *Journal of the American Statistical Association* 82: 543–546.

Schafer, Joseph L. 1997. *Analysis of Incomplete Multivariate Data.* New York: Chapman and Hall.

Tanner, Martin A. 1996. *Tools for Statistical Inference: Methods for the Exploration of Posterior Distributions and Likelihood Functions*, third edition. New York: Springer-Verlag.

Wei, Greg C. G. and Martin A. Tanner. 1990. "A Monte Carlo Implementation of the EM Algorithm and the Poor Man's Data Augmentation Algorithms." *Journal of the American Statistical Association* 85: 699–704.
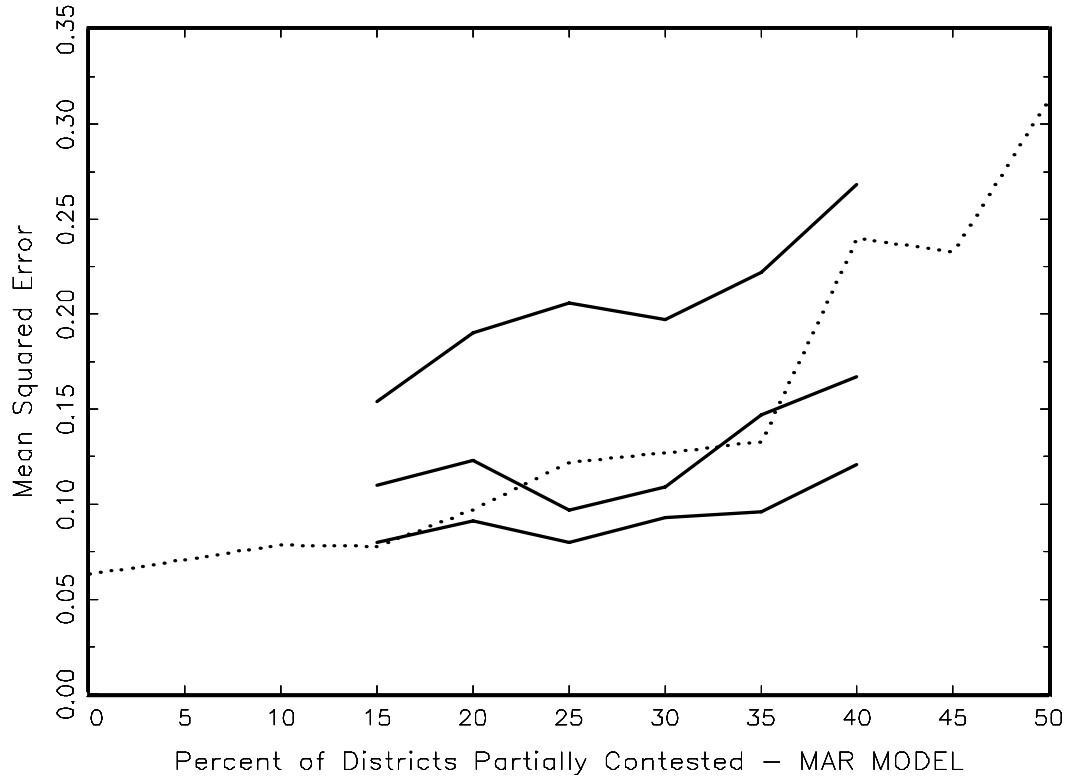
Figure 1: *Mean Square Error Comparisons. MSE is plotted for our method as in the last graph (dotted line) against three implementations of the KK FIML method with increasingly precise numerical integration. Traveling down the graph each solid line has more terms of quadrature (7, 40, 5000 respectively). Thus the FIML method has lower mean squared error, but only at very high computational cost.*