Concluding Remarks and Extensions

Patrick Lam

September 9, 2013

In the first three chapters, I have argued the case for a new approach to causal inference through the direct estimation of individual causal effects, laid out a framework to do so, presented a model to estimate the ICEs through a Bayesian approach with matching, showed that such a model can recover ICEs and other causal estimands through simulation, and demonstrated how the model works in two different applications. I now address some pertinent issues relating to the approach and also suggest some extensions and further applications of the model for future research.

1 Issues Relating to Matching

One of the crucial aspects of the model is the matching process, which chooses the donor observations that ultimately informs the imputation of the missing potential outcomes. The simulations I show suggest that predictive mean matching seems to generally work well across the simulated datasets. However, in any specific application, any number of other matching methods may perform even better. There is likely no single specification that dominates across all datasets. In the causal inference literature, the choice of matching specifications is often ad-hoc and ultimately can cause problems. Researchers can choose from the method to use, the number of matches, the weight of matching variables, etc. The practical solution is often to use a variety of specifications and to compare them on some balance metric to choose the best method. However, this process also has its pitfalls. The choice of balance metric is usually another specification itself. Also, the idea of choosing the best specification may be problematic since it assumes that the specification is the correct one and all others provide no additional information. Short of exact matching, it is unlikely that one specification is the correct one to use. In the case of estimating ICEs, the problems are magnified because it can be the case that one specification works well for certain individuals whereas another specification works well for other individuals. I have also not presented any methods for checking balance in estimating ICEs.

The framework I have presented allows for averaging of specifications by allowing the researcher to choose a different specification for each draw of the algorithm and even possibly for each individual within a draw. I believe this is a more appropriate way to do matching since it does not put all weight on a single specification and leverages the power of model averaging. However, as of now, the pool of specifications and the relative likelihood of choosing any one specification is completely up to the researcher. The prior for the matching specifications completely determines which specifications get used. Ideally, one would be able to calibrate the probabilities of matching specifications through information from the data. For example, for any individual, if some balance metric could be derived such that the specifications are used in proportion to how well they perform on the balance metrics, then the matching specifications would actually be "informed" by the data. This would likely improve the accuracy of the imputations. However, the process for developing a method to incorporate balance metrics within the current ICE algorithm is very computationally intensive and left for future research.

A second way to test the various matching specifications is to see how well they predict observed outcomes. Instead of using the matching to form donor pools to impute Y^{mis}, one can use the same process to predict the observed Y instead and see how well each of the specifications perform. This becomes analogous to a machine learning problem. In fact, the matching method that performs the best does not even have to be a causal inference matching algorithm at all. One can imagine using a myriad of the existing machine learning algorithms to estimate θ^{mis} using Y as the training and test outcomes.

2 Prediction and Population Extrapolation

Testing the performance of the matching algorithms brings up another point about predicting causal effects for individuals outside of the data. As I have alluded to before, ICEs are fundamentally in-sample estimands since there is data only about individuals in our dataset. However, the goal of statistics and causal inference is almost always to predict and generalize to out-of-sample individuals and datasets. Suppose the researcher is presented with the covariate vector for an out-of-sample individual and is asked to predict the individual causal effect of some treatment for this individual. How can the researcher adapt the ICE framework to make a prediction?

The simplest solution for the researcher is to think about the problem as needing to impute two missing potential outcomes, one for the hypothetical treatment and one for the hypothetical control. This involves matching to both in-sample control and treated units, creating two separate donor pools, modeling two separate means, and then drawing two separate Y^{mis}. The resulting ICE would be a predicted ICE for the out-of-sample individual based on the two imputed potential outcomes.

Another important issue related to prediction is also the issue of generalizing the results to some larger population. In most empirical work, the goal is to use the data to make inferences about some population. Usually, these population inferences rely on some assumptions that may or may not be explicit. For the purposes of the ICE framework, generalizing to a population would theoretically imply knowing covariate information for every individual in the population and then predicting each of their ICEs. However, since the ICE framework allows us to aggregate to calculate average effects in the sample, one can also use these estimates to generalize to the population given certain assumptions. The major assumption that is needed to generalize aggregated ICEs is a random sampling assumption. The sample that one estimates the ICEs on must be a representative sample of the population that one wants to generalize to. One can use population weights or other corrections in the data to meet these assumptions. Given the correct sampling assumption, one can say that the individuals in the data are similar to individuals in the population. And while one cannot say anything about ICEs simply based on this fact, one can say that the aggregated ICE average effects are good estimates of population average effects.

The main way in which estimates of population estimands differ from estimates of sample estimands is in the uncertainty estimates. The variance of population estimates is usually higher to account for the sampling uncertainty. In the ICE framework, one way to simulate this uncertainty to get more accurate population uncertainty estimates is through bootstrapping. There are two possible ways to do the bootstrapping. In the first, one can bootstrap the data first, run the ICE algorithm on the bootstrapped dataset, and then calculate the aggregated effects and repeat. The second way is to calculate the ICEs in the full dataset first, then bootstrap the ICEs themselves and aggregate and repeat. The second way uses all of the information in the dataset to do the matching and imputations while the first way only uses observations in the bootstrapped datasets for each bootstrap iteration. Future research should consider which of the two ways is a better choice for getting uncertainty estimates of population estimands.

3 Non-parametric Imputation

The Bayesian model for the imputation of the missing potential outcomes requires the researcher to specify a distribution to draw from. Additionally, modeling the mean and variance of the donor pools in the matching step requires at least two observations in the donor pool. If either of these requirements are not met, the researcher can still impute via a non-parametric approach. Instead of modeling the mean of the donor pool and then drawing from a specified distribution, the researcher can simply impute by drawing one of the observed outcomes in the donor pool as the imputation. This is analogous to multiple "hot-deck" imputation (Cranmer and Gill, 2013). The assumption is that the empirical distribution of the donor pool is the discrete distribution that is used in the posterior predictive step. This also allows for 1-to-1 matching where the imputation is simply the outcome of the donor observation. A non-parametric approach may be more desirable if the researcher does not want to make any distributional assumptions. However, the tradeoff is that the researcher assumes the the outcome values of the donor pools are sufficient to characterize the distributions of the missing potential outcomes. If there are not enough distinct values for the donor pool outcomes (in the continuous case), then the posterior of the ICEs become very discrete.

4 Convergence

As with any MCMC simulation, convergence to the stationary distribution is necessary and must be checked. The algorithm I propose really only contains dependence among parameters at the matching step (the imputation is only dependent on D, which is dependent on the matching parameters), so non-convergence may be less of an issue than typical MCMC simulations with high dependence among parameters. Nevertheless, convergence should be checked. Unfortunately, the number of parameters in the model is greater than or equal to the number of observations in the data, so checking convergence on each one is tedious at best. However, it is also necessary to check each parameter as non-convergence on even one parameter may be problematic for all the results (Gill, 2008). Further research should be done on ways to test convergence on a large number of parameters. I defer to the vast literature on convergence diagnostics for this. However, one suggestion is that researchers can check convergence on the aggregations of the ICEs. For example, if checking convergence on all N ICEs proves to be too tedious, one can check convergence on the aggregated draws of the ATE or the ATT. If the ATE draws do not seem to converge, then this indicates that one or more of the ICEs have not converged. Unfortunately, the inverse is not true. Convergence on the ATE does not necessarily imply convergence on all the ICEs.

5 Extensions

5.1 Incorporating ICEs into (almost) any possible (causal) model

One benefit of the idea of estimating ICEs is that researchers can incorporate ICEs and potential outcome imputations into nearly any type of causal model that one can run. The potential outcomes framework is a powerful framework that clearly specifies the research design and the problem at hand. The ICE framework simply builds off the potential outcomes framework. Then any regression model, no matter how sophisticated, is really a means to estimate parameters in the potential outcomes framework. Often, including ICEs in a more sophisticated regression model simply boils down to choosing the right relevant set of donor observations.

Consider the fixed effects regression model often used in economics and political science.

$$Y_{ik} = \alpha_k + X_{ik}\beta_k + \epsilon_i$$

where k denotes a certain cluster (for example, countries). This fixed effects model estimates different intercept and (possibly) slope terms for each cluster. Another way to conceptualize the goal of fixed effects models is simply to match observations only within clusters (Imai and Kim, 2013). Within the ICE framework, this simply boils down to limiting the potential donor pool within each iteration to observations in the same cluster and then aggregating by cluster to get the cluster-specific intercepts and slopes. In more complicated multilevel models, one can simply impute the missing potential outcomes and then aggregate either on a first or second level variable to get specific causal effects.

Now consider complicated regression models that attempt to model time components. Often, such models boil down to including a lagged dependent variable or other terms on the right-hand side of the regression equation. In the matching framework, this simply means adding a variable to the matching specification. To be more precise, the researcher can set the matching algorithm to exact match on certain variables, which is again simply an adjustment on the potential donor pool. More complicated time-dependent models may include certain parametric specifications, such as a spline to account for time in binary dependent variable models (Beck, Katz and Tucker, 1998). Researchers can include such specifications either during or after matching to adjust the imputations. One way would be to run a regression within the donor pool using only the spline variables to estimate θ^{mis}.

The general idea is that including matching and reframing causal inference at the level of ICEs is compatible with almost any existing method. Furthermore, I argue that it has the added benefit of forcing researchers to seriously consider the causal quantities they are estimating by being explicit about modeling individuals. Adding a spline may be simple to implement in any statistical package, but forcing the researcher to understand that the spline simply models how other observations in different time periods contribute to the missing potential outcome of a certain observation of interest is valuable in promoting the understanding of the role of regression models in causal inference.

5.2 ICEs and Causal Inference Assumptions

Another benefit of working with ICEs and a possible avenue for extending the framework is through the testing and relaxing of typical causal inference assumptions. As I alluded to in the examples using a two-stage model model with instrumental variables, the typical exclusion restriction can be tested using the ICE framework by estimating the non-complier average treatment effect (NCATE). If the assumption of the exclusion restriction were correct, then the NCATE should be zero. In the job training example, using the ICE framework to estimate NCATE, I found that there may be some reason to doubt the exclusion restriction that is typically assumed.

Consider also the conventional SUTVA assumption that is required in almost all causal inference studies. The SUTVA assumption has two parts:

1. Treatment assignment on one observation does not affect the potential outcomes of another observation.
2. No varying treatment intensity.

The second part of the assumption may be violated, for example, if individuals assigned to a drug can take either a regular strength or extra strength version. In the corruption monitoring example, the participation treatment actually violated the second part of the assumption since villages received either invitations only or invitations and comment forms. For some parts of the analyses, I assumed that the two were the same. However, since the ICE framework results in a Bayesian posterior, I can actually make probability statements about how true the assumption actually is. Let village i be assigned control (neither invites nor invites and comments). Suppose I then impute the potential outcome for i being assigned invites (by matching to villages that received invites only) and then I also impute the potential outcome for i being assigned invites and comments (by matching to villages that received both invites and comments). Then I would have two posteriors for the two potential outcomes of receiving the two different versions of the participation treatment. I can then compare the two posteriors and calculate the probability that $\tau_i^{(inv)} = \tau_i^{(inv\&com)}$. This would be an estimate of the probability that the second part of SUTVA holds for village i. I can do the same calculation for all i and have a sense of how likely SUTVA is violated. More research must be done into how much to trust the results of such an analysis, but it at leasts suggests potential for testing the sensitivity of certain studies to certain assumptions.

Another key assumption that is often made in causal inference with instrumental variables is the monotonicity assumption. Let Z be an instrument for W with outcome Y. The monotonicity assumption (also called the no defiers assumption) states that $W_i(1) \geq W_i(0)$. In simple terms, the assumption is that there are no individuals who would take the treatment when assigned control but not take the treatment when assigned treatment. In many situations, this assumption makes sense. However, for certain studies, this assumption is very important and may not be fully satisfied. Consider the now famous paper on the effect of institutions on economic performance by Acemoglu, Johnson and Robinson (2001). In that paper, the authors use settler mortality as an instrument for extractive institutions. The theory states that in countries where settler mortality was high, settlers built extractive institutions since they did not settle there themselves. In countries with low settler mortality, the settlers actually installed less extractive and "better" institutions for economic growth. The author use this design to conclude that institutions matter in economic growth.

To simplify the analysis, let settler mortality (Z) and non-extractive institutions (W) both be measured with binary variables. The monotonicity assumption states that the relationship between settler mortality and institutions can only go one way for all countries. No country exists that would establish extractive institutions with low settler mortality but non-extractive institutions with high settler mortality. However, this assumption is fundamentally untested and given the myriad of variables that interact with both Z and W, it is conceivable that the monotonicity assumption could be violated. For example, one can make the case that the relationship would depend on which groups of settlers were affected the most by mortality. Suppose the settlers can be partitioned into "royalists" who supported the Crown and "colonialists" who supported more independent institutions. If mortality affected the royalist camp disproportionately, then it could be the case that increasing mortality actually increases the odds of less extractive institutions while absent high mortality, the royalists have enough political power to enact extractive institutions. This is but one possible scenario in which the monotonicity assumption is violated. Using the ICE framework, one can actually relax the monotonicity assumption and estimate the probability that a country is a "defier" country. By allowing for defiers and jointly estimating the compliance group memberships, one can get a better estimate for LATE and also estimate a defier average treatment effect.

By thinking about causal inference at the individual level and estimating ICEs, researchers are given tools to think about the assumptions they make and relax some of the assumptions or test the sensitivity of their results.

6 Final Words

In this dissertation, I have presented an argument for why researchers should shift their focus from estimating average effects to estimating individual effects. I am in no way arguing that existing methods for average effects should no longer be used. I believe that estimating ICEs in conjunction with existing methods can produce great results. The contribution of the dissertation is in opening up new avenues and helping scholars rethink how existing methods fit into the causal inference framework using potential outcomes. The algorithm and the models themselves are very much works in progress. There are other areas that my work touch on, such as the merging of matching and Bayesian methods, the use of model averaging with matching, or the idea that complicated regression models can be integrated with a matching approach. All of these areas deserve much more future research. I simply hope that my dissertation will spur more interest in how to deal with treatment effect heterogeneity and how to reconcile small n research with large n studies as well as provide a unified and straightforward framework for thinking about causal inference.

References

Acemoglu, Daron, Simon Johnson and James A. Robinson. 2001. “The Colonial Origins of Comparative Development: An Empirical Investigation.” *The American Economic Review* 91(5):1369–1401.

Beck, Nathaniel, Jonathan N. Katz and Richard Tucker. 1998. “Taking Time Seriously: Time-Series-Cross-Section Analysis with a Binary Dependent Variable.” *American Journal of Political Science* 42(4):1260–1288.

Cranmer, Skyler J. and Jeff Gill. 2013. “We Have to Be Discrete About This: A Non-Parametric Imputation Technique for Missing Categorical Data.” *British Journal of Political Science* 43(2):425–449.

Gill, Jeff. 2008. “Is Partial-Dimension Convergence a Problem for Inferences from MCMC Algorithms?” *Political Analysis* 16(2):153–178.

Imai, Kosuke and In Song Kim. 2013. “On the Use of Linear Fixed Effects Regression Models for Causal Inference.” Unpublished.