

## CHALLENGES IN ECONOMETRICS

GUIDO W. IMBENS - HARVARD UNIVERSITY, SEPT, 2010

### 1. INTRODUCTION

To frame what is in my view of the main challenges facing researchers in econometrics, let me set the stage by describing the current state of research. Much of the traditional research in econometrics can be divided into two branches, the first comprising cross-section and panel data econometrics and the second time series analysis. In the cross-section branch of econometrics researchers have data on a large number of units, often individuals, or groups of individuals, firms, or markets. For each unit there is information on a relatively small number of variables, sometimes measured at a single point in time, sometimes with repeated measures as in panel data. The units are viewed as exchangeable, or independent in the sense that there is no interaction between the units: what happens to one unit does not affect other units. In time series analysis the typical setting is one with observations on a small number of variables, at many points in time, with relatively unrestricted dependencies between the different variables. For models designed for data configurations of these two types we have learned much in the last few decades. In fully parametric models, as well as in the more flexible semi and non parametric models we have gained an impressive understanding of the appropriate ways of analyzing such data, and the properties of many estimators and methods for inference.

In my view the biggest challenges faced by economists in terms of analyzing economic data concern fundamentally different configurations of the data, with complex, largely unknown, dependence patterns and a relatively large numbers variables per unit. In such cases the current methods to do approximate inference based on large sample results, which are specifically designed to exploit laws of large numbers and central limit theorems, are likely to be inadequate. Moreover, trying to fit these more

complex data configurations into the old methods would be unlikely to lead to much progress. In some cases econometricians and statisticians have made some progress on such alternative data configurations, but for the most these are unexplored areas for research.

## 2. DATA CONFIGURATIONS

More and more data are becoming available to researchers that do not fit the standard mold. We may have information on units located in physical or economic spaces that exhibit strong, but complex and partly unknown, dependencies in economic behavior. These dependencies are likely to weaken as the distances between units increase, but the appropriate notion of distance is likely to be partly unknown. Correlations may be stronger in some parts of the population than in others. One branch of econometrics that has studied such questions is spatial econometrics, but this is still a relatively undeveloped part of the econometrics profession, relative to the number of questions. Much of the work in spatial econometrics relies heavily on methods imported from the statistics literature where the focus was on different questions. For example, in the statistics literature the focus was often on predicting outcomes in particular locations given outcomes in nearby locations, e.g., presence of natural resources at one location given measurements on measures of resources or proxy variables at nearby locations, with often strong prior beliefs about the appropriate distance measures. In economics it may be of more interest to understand how the spatial correlations generate effects of policies implemented in one location on outcomes in another nearby location. Such effects may operate with unknown lags, necessitating the combination of time series methods and spatial analysis. There may be little prior knowledge about the relative importance of different distance measures.

Related to spatial statistics but with a different set of challenges, the dependencies between economic behavior may arise from what is sometimes called peer effects, or social interactions. Here distances between units are often modeled as discrete, typically binary: individuals either influence each other in a constant way, or not at

all. In an important paper Manski (1993) studied identification questions in a special case where a population was divided into peer groups. Important is the fact that the peer groups in Manski's analysis partition the population. Behavior of units in different peer groups is not correlated. Within groups correlations may arise from correlated backgrounds, from a shared environment, or from feedback in behavior. Often individuals within a peer group are viewed as exchangeable: all individuals influence each other to the same degree. Many questions arise when the groups within which the dependencies are present are partly the result of choices made by individuals. Observed correlations may simply affect choices of individuals to team up with similarly minded individuals, rather than effects on peers' behavior. Controlling for shared background is also a difficult challenge. While there have been numerous empirical studies documenting correlations in outcomes for individuals in the same class, both in the short and in the long run, there is still a great deal of uncertainty whether these arise from teacher effects or interactions between students. Ultimately a key question is whether these social interactions can be exploited by policy makers to improve the distribution of outcomes in society, through, for example, tracking in educational settings. An interesting paper in this respect is Carrol, Sacerdote, and West (2010) who attempt to improve average test scores by optimally assigning incoming recruits at the Air Force Academy to different squadrons. If successful, the mechanism would be through the induced interactions associated with the squadron assignments.

The analyses get even more complicated when the peer groups do not simply partition the population. Some individuals may be connected to many others through self-chosen friendship links, and the effect of two different peers on the same individual may be different. Economic theorists have analyzed such network settings in considerable depth (e.g., Jackson, 2008), and empirical work has demonstrated the presence of correlations in behavior associated with such networks, but our understanding of the statistics and econometrics of these models is still in its infancy. For example,

the literature almost exclusively deals with exogenously formed networks, with links between individuals either present or absent rather than of varying intensity, and with little attention to the dynamics of and feedback in the network formation processes. None of these are plausible assumptions, and there is little knowledge about the sensitivity of empirical results to violations of these assumptions. Theorists have focused on the difficulty of defining useful equilibrium concepts in the context of network formation. When taking account of the changing environment the dynamics of the equilibrium may lead to even more problems. Questions of interest for economists include the effect of encouraging interactions by facilitating opportunities to form links, and the effects of interventions in some individuals on outcomes for those connected to them.

There are a number of specific challenges in analyzing such data sets. They arise from common features of such data. They often contain information on a large number of units, as well as detailed information per unit. Especially with some of these data sets drawn from internet communities, one may have information on a very large number of individuals, followed over a period of time during which they were subject to many stimuli from outside and during which many interactions with other individuals took place. With possible dependence in behavior for many individuals in such networks, the basis for conventional large sample results is unclear for even for simple statistics such as sample averages. A general question in this area concerns the presence of data sets with a many variables relative to the number of units, sometimes more even variables than units. For example, we may have for a moderate number of individuals extremely detailed information about their behavior, including all web sites visited, all social interactions experienced, or all purchases made during visits to a supermarket, or, in the biostatistics literature, we may have detailed genetic information on a small number of individuals. Using such data to infer patterns in behavior that can inform policy questions is fundamentally different from that of inferring parameters of parsimonious models in large samples. Simply following the

standard approach of approximating the distribution of estimators by joint normal distributions is unlikely to be a generally satisfactory approach in such settings with many parameters. A specific example of this is the study of regression models with more potential explanatory variables than individuals. Some methods have been developed for the covariate selection problem in the statistics literature (e.g., Lasso and related methods), but these methods have not found many applications yet in economics.

There are also huge computational challenges in this literature. Most of the sophisticated modeling has been done in the context of very small data sets. Even in such settings the number of possible links and networks can quickly be very large. In practice even the number of units in the networks can be very large, leading to even greater computational problems.

Research related to these questions has been conducted in multiple disciplines and is a fertile area for interdisciplinary research. Sociologists have a long tradition of studying communities and social interactions, and have contributed many substantive questions to this area. They have also collected interesting data sets, as well as some statistical methodology. Statisticians have developed methodology for spatial data, although little specifically for network data (see the Holland and Leinhardt (1981) paper and the subsequent literature). Computer scientists have focused on properties of networks emerging from various network formation processes. Specifically they have looked at models that generate few large connected networks rather than many disconnected groups. None of these disciplines have focused much on the type of questions economists tend to be interested in, but many have made progress on related issues.

#### REFERENCES

CARRELL, S., B. SACERDOTE, AND J. WEST (2010), “Beware of Economists Bearing Reduced Forms? An Experiment in How Not To Improve Student Outcomes” Unpublished Working Paper.

HOLLAND, P., AND S. LEINHARDT, (1981), "An Exponential Family of Probability Distributions for Directed Graphs," *Journal of the American Statistical Association*, 76(373): 33-50.

JACKSON, M, (2008) *Social and Economic Networks*, Princeton University Press.

MANSKI, C., (1993), "Identification of Endogenous Social Effects: The Reflection Problem," *Review of Economic Studies*, 60, 531-542.