Some thoughts on "The Principles Underlying Evaluation Estimators with an Application to Matching" by James J. Heckman

Michael LECHNER[†‡]

1 Introduction

In recent years the rapidly expanding field of "programme evaluation" made substantial advances in providing and understanding (micro-) econometric methods suitable to uncover ceteris paribus effects with a minimum set of assumptions. The actual use of these methods in applied studies is booming particularly in labour economics. It is probably a fair guess that the knowledge obtained in this subfield already influences many other branches of econometrics as discovering causal relation is and was one of the key aims of almost all empirical econometric work (see the excellent review by HECKMAN [2000]). Probably, these powerful econometric methods will "conquer" other fields rather soon, if they have not yet done so.

One reason for their success is the wealth of new data that is now available and is bound to become ever larger and more informative. In our days, the influence of applied econometrics in day-to-day policy advice may be larger than ever, in particular in Europe. The reason why this branch of applied microeconometrics seems to have much larger growth rates in Europe than in the USA (this is of course a subjective observation by the author not validated by objective data) is not only because it started later and from a lower level in Europe than in the USA, but also most likely related to the considerable efforts by European governments and researcher to make the big archives of administrative data combinable and accessible for researchers, and thus usable for policy advice. Thus, the methodological issues raised in Jim Heckman's paper are instrumental in improving actual policy advice and thus economic policy.

The Nobel laureate James Heckman, jointly with many co-authors, was one of the key contributors to this field of econometrics over the last thirty years. In his recent contributions, he develops a unified framework that allows the direct comparison of the key non-structural-model-based identification approaches for causal effects, namely matching, instrumental variables, and control functions. Since it is an important insight of this literature that all methods intended to identify causal relationships require untestable assumptions, understanding the different implications and behavioural restriction that are implied by these different identification strategies is of key importance in obtaining reliable estimates of the causal effects.

The current paper by Jim Heckman reviews these methodological results and discusses their implications for the economic and behavioural content of the so-called matching or selection-on-observables identification strategy. Indeed, matching has become a very popular estimation method in recent years as data improved and software became easily available (see for example the excellent survey by Imbens [2004]). As usual when a fad develops, there have been good matching papers and there have been bad matching papers. I conjecture that the latter are the reason why the current paper by Jim Heckman is very critical about matching and concludes section one with "I next turn to a detailed discussion of the method of matching which is currently 'the method of choice' for evaluating social programs in Europe. When its strong assumptions and severe limitations are understood, perhaps its popularity will wane."

I fully agree with the point that the economic content of the underlying assumptions of any identification strategy needs to be fully understood to be able to derive reliable policy conclusions. However, I will argue somewhat more optimistically that

generally there is nothing like a good or bad identification strategy, but that the cred-
ibility of any such strategy is very context specific. Seen in that way, my guess is that
Jim Heckman's paper will lead to *better* matching papers. Whether it will lead to
fewer matching papers is however not obvious. It is quite conceivable that the quality
of the data will further increase thus making matching more, and not less, attractive.
Indeed, there are two ways to think about identification strategies in general: One the
one hand, there is the consideration, which method to use with the particular data
at hand. On the other hand, one can view these strategies that define the conditions
necessary to identify a particular parameter of interest as a recipe to define the data
required to make them hold. Then, the researcher should try to get the required data,
from government sources or from additional or existing surveys, such that these con-
ditions are plausible. It is certainly helpful for this process of obtaining new data (and
research funding), if the government (or firms, or…) expects in return to obtain infor-
mation that is useful for designing its own policies. Over the last decade or more, the
latter approach clearly played an important role in Europe. With this mindset, I will
now discuss some of the issues raised in the paper by Jim Heckman.

2 The distinction between *X*, *Z*, and *U^D*

The thrust of Jim Heckman arguments relate to the distinction of observable vari-
ables *X* (observable variables determining outcomes) and *Z* (observable variables
determining selection) and the relevant unobservables denoted by U^D (unobservable
variables determining selection). The weaknesses of matching are exposed by show-
ing that it has strong and undesirable implications concerning U^D. Of course, on the
other hand, one of its major advantages is that it is entirely flexible with respect to *X*
and *Z*. In fact, knowing or observing *X* that are not in *Z*, as well as the relation of *X* or
Z to the potential outcomes, does not play any role. This is different for IV or control
function approaches, where the exclusion restrictions on some elements of *Z* play a
key role.[1] This flexibility is one of the key advantages of the method of matching.

The assumptions about the unobservable factor U^D (that influences selection)
imply, for example, that "… that conditional on *X* average returns equal marginal
returns" (section 3.1). In terms of the conditional on *X* treatment effects this implies
considerable homogeneity, thus $\Delta^{MTE}(x) = \Delta^{ATE}(x) = \Delta^{TT}(x) = \Delta^{LATE}(x)$, so that
the marginal treatment effect *MTE(x, u^D)* must be flat in u^D (while the shape of
MTE(x, u^D) with respect to *x* is not restricted). At first sight, this restriction appears
to be very unrealistic for many applications and contrary to economic intuition, as
surely those individuals *expecting* the highest return will select into the treatment
D = 1 and this expected return could be influenced by U^D. But, the key question
in any particular application is for which variables the symbols *X* and U^D stand
for. Quite simply, if *MTE(x, u^D)* is indeed non-flat in some components of u^D and
if these components can be observed (or made observable by suitable data collec-
tion), then this is not a restriction at all. In other words, before stating which factors

---

1. To simplify the exposition of this note, I will always assume that *X* is a subset of *Z*.

these variables actually stand for, it is not clear at all whether such an assumption is highly restrictive (*not plausible*), or not restrictive (*plausible*) at all, or something in-between. If we sample the "right" data, then we may happily live with such an assumption. Similarly, when we manage to sample an instrument, we may happily live with the assumption that this instrument has no direct effect on the potential outcomes.[2] Again, all identification strategies can be seen as recipes on how to obtain the right data necessary to identify the parameter of interest.

Next, I discuss some of the disadvantages of the method of matching mentioned by Jim Heckman in his paper. This comparison is however made locally, *i.e.* considering the main competitors, which is the control function / instrumental variables approach (I will subsume, for example, regression discontinuity and difference-in-difference approaches under the heading of IV methods). Identification of causal effects using structural models has further interesting pros and cons that are however beyond the scope of this short note. To get a more balanced picture, I will mention a few advantages of matching as well. However, as they are not given much emphasis in the paper by Jim Heckman, I will be brief.

3 Some weaknesses and strengths of the method of matching

Weaknesses

Problems that plague all nonparametric estimation schemes

Clearly, there is a price to pay for the big advantages of nonparametric identification and estimation strategies, which are their flexibility and robustness with respect to unnecessary statistical assumptions. We briefly look at three components that may make these methods appear less attractive.

Larger samples necessary

The first and obvious price to pay for econometric models to be flexible and robust (and thus less biased) is a loss in precision. The only way out is to obtain larger data sets, which, given the continuous decline of the price of computer storage and computing power, is becoming easier at least for register data. A bit of an exception are data coming from interview based surveys. They get cheaper as well by using new technologies, but the decline is much slower.

---

2. From these considerations it is clear that I disagree with the statement made in section 3.3 that "Matching assumes access to a richer set of conditioning variables than is assumed in the methods of control functions", as the method of control function requires to *observe* an instrument. Although the method of matching requires the *existence* of an instrument as well (this is implicit in the common support assumption, M-2), there is no need to observe it. It should also be pointed that many IV approach in practice require control variables to make the exclusion restriction of the instrument plausible. It is a priori not clear that their dimension is smaller or larger than the dimension of control variables needed for matching approaches.

Common support assumption limits the reach of the parameters that can be estimated

With nonparametric identification and estimation strategies we can only identify and estimate effects when there are potential comparison observations. In matching, there is the particular issue, what to do if we have available control variables (*Z*) that perfectly predict treatment status. Of course, this does not necessarily mean that matching should not be applied, but that two different cases should be considered: (i) All of those control variables directly influence the outcomes: In this case an average effect is usually not identifiable (because none of them can be used as an instrument; instruments may not even exist, although the assumptions underlying a regression discontinuity approach may sometimes be plausible). (ii) Some of those variables are in fact not influencing the potential outcomes directly, i.e. they are instruments: In this case, if the group of remaining variables fulfils the conditional independence assumption, then the usual average effects are identified despite our "ability" to predict perfectly. We just won't use that ability. In addition, we can use the instruments to form IV type estimators to learn other parameters of interest, and if at least one of the instruments is continuous (and has large support) the matching estimates can be validated by corresponding IV or control function estimates.

There is also the view that being forced to make the empirical statements only in regions of the space of characteristics in which there are actual data is one of the big advantages of nonparametric identification strategies, as it avoids the interpolation coming from functional form assumptions employed by parametric methods.[3] Heckman's paper is very clear on this.

The distribution of the effects is not identified

The matching assumptions identify the marginal distributions of the potential outcomes, but not their joint distribution, and thus not the distribution of the differences of the individual causal effects. The same is true for the nonparametric IV and control variable approaches. This is the price to pay for avoiding the imposition of assumptions on the joint distribution of the potential outcomes, which appear to be particularly hard to justify by economic reasoning. Of course, if (!) we think that in the particular application we should have confidence in restricting heterogeneity by imposing, for example, homogenous effects or a factor structure together with some other distributional assumptions, then the distribution of effects is identifiable as well, as shown in a couple of papers by Jim Heckman (referenced in his paper). However, those assumptions are not credible for many applications.

Matching specific weaknesses

The conditional independence assumption is a very strong assumption

The key weakness of the method of matching is that the required conditional independence assumption is very strong. Imposing an assumption as strong as this is however a feature of every method attempting to uncover causal effects. We have

---

3. The fact that these functional form assumptions are testable does not help, because their validity can only be tested for the regions where the data is located. The assumption that they are valid outside this region is of course untestable.

to rely on untestable and strong assumptions to identify causal effects; otherwise, we have to give up that goal that is so central to econometrics. The strong assumptions underlying matching may be more or less plausible in a particular application than the strong assumptions underlying IV and control function estimation, or estimation of causal effects based on structural models. At the end, we need the context to find those assumptions that are most appropriate. Furthermore, even for a given context we must accept that different researchers having different theoretical models in mind, or different knowledge about the institutional background, come to different conclusions about which identification strategy is most appropriate. Of course, we will trust only those results based on assumption we find most convincing in the specific context.

It is however an important implication of the matching assumption not shared by the others that individuals are not allowed to act on potential outcomes, as potential outcomes are unobservable by definition (i.e. $P(D = 1 \mid X, Y^1, Y^0) = P(D = 1 \mid X)$), and unobservables influencing outcomes are not allowed to influence selection. However, this assumption may seem more restrictive than it is, because agents usually do not know their (future) potential outcomes but must forecast them as well as the econometrician (who tries to forecast their means, which is probably less demanding than forecasting their individual values as has to be done by the agent). The discussion of the different information sets outlined in Jim Heckman's paper is a very helpful way of thinking about the problem (and points to the large value for collecting variables characterizing such individual information sets). It means that from those control variables (X) that the agent uses to predict the potential outcomes, the econometrician needs only those that indeed influence the mean potential outcomes (they may be the same as the agent uses or a subset of them). If the econometrician is successful in doing so, then this assumption is credible. If not, then matching may be biased.[4] Again, this line of thought gives hints about the data that need to be collected, and thinking about information sets explicitly is very helpful in this respect as well.

It is worth pointing out that IV approaches require very similar randomization assumptions. However, they do not directly concern the conditional randomness of D and the potential outcomes but they concern the relation of the instruments to the potential outcomes.

How to select the conditioning variables?

From the considerations so far, as well as from Jim Heckman's paper, it is already clear that the success of matching depends on "finding" the "right" variables to include in the propensity score. From the discussions above, it is clear that data driven methods that depend on improving the fit of a binary choice model explaining participation are deficient, because they would lead to the inclusion of *instruments* that are good predictors of participation. From the point of identification it is undesirable to take out variation (and so probably reduce the support) that could be exploited to identify and estimate the causal effects. Any data driven algorithm for selecting control variables would need to take into account jointly their effect on the outcomes as well as on participation. Unfortunately, for non- or semi-para-

---

4. IMBENS [2004] considers the interesting case when the agent decides on participation by optimizing on outcome variables the econometrician is not interested in. He shows that such a context may help identification.

metric estimators, I am not aware of a joint test that has been developed so far for that specific task. Even worse, one of the advantages of many matching methods is that a regression of Y on X is not strictly necessary. Therefore, introducing such a regression for the purpose of variable selection (i.e. the de-selection of instruments) is not very attractive.

Given the deficiencies of data driven methods for control variable selection, it remains to emphasise the key role of a priori considerations about the variables that *should be* included as they are the cornerstone of any "variable selection algorithm".[5]

One worrying feature remains: Jim Heckman shows in his examples that for a misspecified matching model (i.e. not all necessary components of X are observed), it may be true that "reducing" the misspecification by adding further variables explaining outcomes and treatment selection while some variables are still missing may in fact increase the bias.[6] Unfortunately, it is most likely that such examples may be constructed for (almost?) all econometric estimators.[7] Although, this is not comforting at all, it is not a particular feature akin to the matching principle.

Strengths

At the end of short note let me very briefly remind the reader that the method of matching has some important advantages compared to other methods as well.

One of its major advantages is that it is entirely flexible with respect to X and Z. In fact, knowing or observing the components of Z not in X (or the components of X not in Z) does not play any role, contrary to IV or control function approaches where the exclusion restrictions on some elements of Z play a key role.

Second, in the absence of a continuous instrument (which are very rare in actual applications), parameters identified by matching (ATE, TT) *may* be more interesting as those identified by instrumental variables (without continuous instruments): Not every LATE is an interesting LATE, because the so-called complier population for which this parameter is a causal effect may or may not be interesting for the policy under consideration.

Furthermore, a matching estimator may be much more precise than IV estimators, as in practice credible instruments seems to share the feature that they are also weak instruments, thus requiring very large sample sizes for sufficiently precise estimation. Of course, this does not mean that matching should be used despite being biased in a particular application, but if there is reason to assume that any remaining bias is very small, it may not be worth (in mean squared error sense) to remove it at the cost of a huge loss of precision. Such additional precision may not only be important for the precise estimation of mean population parameters, but it may allow discovering interesting heterogeneity relating to some components of X that one might not be able to discover with a less precise estimator.

These are just two arguments that can be put forward supporting the matching approach. Further arguments supporting matching can be found for example in the

---

5. Of course, when control variables are required to make the exclusion restrictions of the instruments plausible, similar issues appear.
6. Similar results have been obtained by Bhattacharya and Vogt [2007] who look at the case of including instruments in a misspecified matching estimation.
7. For example, it is a priori not clear, whether including another invalid instrument into an (invalid) IV estimation will increase or decrease the bias of this IV.

surveys of Heckman, LaLonde, AND Smith [1999] and Imbens [2004], as well in the huge statistics literature.[8]

4 Conclusion

What is the conclusion from James Heckman's important paper for the choice of a particular identification strategy in a specific empirical study? Clearly, a wrong conclusion would be to stop using matching methods and engage in a permanent search for instruments coming from some more or less natural or unnatural experiments, or to estimate restrictive structural models only, or both. None of these methods is based on assumptions less restrictive than the assumptions required for the method of matching. Those assumptions are different, but a priori not less restrictive.

In fact, Jim Heckman's paper is extremely valuable to understand better what data are required such that matching is a powerful empirical tool. In particular, thinking about the identification problem explicitly in terms of information sets is a new twist in that discussion that appears to be very valuable to understand what data are required to arrive at an identification strategy that appears to be credible. Obtaining these data from whatever sources is a substantial but worthwhile challenge that the applied econometrician should consider as being part of his or her "job description". Maybe, in the past econometricians saw themselves too much as passive consumers of whatever the "data generating process" spits out. In our days, due to the recent success of microeconometric methods in applied policy, the applied researcher has much more leeway in influencing this process, and thus to create data that are suitable for one of the many available econometric identification and estimation strategies to uncover causal effects.

All in all, there are at least four interesting scenarios concerning the choice of identification and estimation strategies:

(i) We have good data in the sense that we are confident to observe the relevant information set jointly determining selection and potential outcomes, and there is enough variation left so that the common support assumption holds: The method of matching is the method of choice.

(ii) We have good data in the sense that we are confident to observe one or more variables influencing selection but not potential outcomes, and if the particular LATE implied by a discrete instrument is interesting or there is a continuous instrument available on a large support, then instrumental variables or control function approaches are the preferred method. The exact choice of method will depend on whether the instrument is continuous or discrete, and on the power / weakness of the instruments.

(iii) The matching assumptions hold and in addition there are instruments available. Now, we can estimate additional causal parameters compared to (i) and (ii). Given large enough samples, the more continuity these instruments have, the more

---

8. The paper collection RUBIN [2006] may provide a good starting point for that literature.

parameters we can estimate (see the various papers by Jim Heckman). For a continuous instrument (together with large support assumptions) we can even test the matching assumptions. Such cases are however very rare in practice.[9]

(iv) The data is not good enough so neither the matching, nor the IV, nor control function assumptions are valid. In this case, either we have to get better data, or we have to "substitute" the missing data by economic modelling assumptions, or we have to give up on obtaining causal effects.

I like to conclude by thanking Jim Heckman for his important and continuous efforts in improving policy relevant econometric studies by requiring us to think hard about the economic content of the various identifying strategies available in the applied econometrician's toolbox.■

References

BHATTACHARYA J. and VOGT W.B. (2007). – "Do Instrumental Variables Belong in Propensity Scores?" NBER Technical Working Paper Nº 343.

FRÖLICH M. and LECHNER M. (2006). – "Exploiting Regional Treatment Intensity for the Evaluation of Labour Market Policies", CEPR Discussion Paper Nº 5728.

HECKMAN J.J. (2000). – "Causal Parameters and Policy Analysis in Economics: A Twentieth Century Retrospective", *Quarterly Journal of Economics* 115, 45-97.

IMBENS G.W. (2004). – "Nonparametric Estimation of Average Treatment Effects Under Exogeneity: A Review", *The Review of Economics and Statistics* 86, 4-29.

RUBIN D. (2006). – *"Matched Sampling for Causal Effects"*, Cambridge, New York: Cambridge University Press.

---

9. FRÖLICH and LECHNER [2006] consider a case in which matching is combined with a binary instrument.