The Information Basis of Matching with Propensity Score

Esfandiar Maasoumi* and Ozkan Eren
Southern Methodist University

Abstract

This paper examines the foundations for comparing individuals and treatment subjects in experimental and other program evaluation contexts. We raise the question of multiattribute "characterization" of individuals both theoretically and statistically. The paper examines the information basis of characterizing individuals and offers alternatives motivated by welfare and decision theories. The proposed method helps place propensity scores and other "matching" proposals in context and reveals their advantages and disadvantages. We do not find the implied theoretical assumptions underlying propensity scores to be attractive or robust. Our proposal does not "solve" the matching problem, but provides bounds on inferences and makes clear the arbitrariness of specific solutions.

Key Words: treatment effect; information theory; multivariate scaling; propensity scores; utility functions; fundamentalism; Kullback-Leibler; entropy; aggregation.

1. Introduction

The literature on "treatment" effects and program/policy evaluation deals with a classic problem of unobserved outcomes by "matching" through observed covariates. In order to isolate/identify the effect of treatment alone, the "counterfactual" outcomes must be obtained. This requires control for all other factors that may significantly contribute to the *apparent* differences in outcomes. A mechanism is needed for the way these covariates influence outcomes or choices, such as the decision to participate or probability of treatment. We do not deal with the choice of covariates and its division into observed and unobservables, or exogenous and endogenous variables. A recent insightful survey on this is Heckman and Navaro-Lozano (2004). But we deal with how these covariates are employed in order to "characterize" subjects, through estimation and/or matching, and by any transformations thereof, to determine "similarity" between members of samples or population groups. Our "representation" perspective and techniques further expose the meaning and limitations of Propensity Scores (PS).

We briefly consider aspects of the outcome/treatment *distributions* to be examined, the means, quantiles, or entire distributions. The existing literature is predominantly focused on "averages" or mean treatment effects, conditional on various information sets. Notable exceptions attempt to model for conditional quantiles in an attempt to avoid the "veil of ignorance" and reveal more of the reality that *outcomes are distributed* and "averages" can often mislead. See Imbens (2004) for a survey. Our alternative approach suggests natural ways of dealing with the whole distribution of *both the characterizing covariates and the outcomes*. We only sketch and exemplify full distribution methods such as entropy divergences, equality and/or dominance between whole distributions of *attributes and outcome*, respectively.

The intuition behind our main proposals in this paper is as follows:

Let x_i be a set of m attributes or characteristics for "individual" $i = 1, 2, ...n$. Subsets of x_i determine choices, selection, and/or outcomes of policy or treatment programs. We propose to choose a function $S(x_i)$ to represent or summarize each individual. We believe the choice of functionals $S(x_i)$, perhaps as utility or welfare evaluation functions, must be dealt with explicitly. Powerful implications of welfare axioms of "fundamentality", "impartiality", and "anonymity" may be invoked which par-

*Corresponding author: Esfandiar Maasoumi, Department of Economics, SMU, Dallas, TX 75275-0496; E-mail: maasoumi@mail.smu.edu.

tially justify the same "common representations" for everyone's preferences, in functional forms and for "representative agent" formalisms; see Kolm (1977) and Maasoumi (1986a). Simply put, if we enter into a preference function all the attributes that would matter (i.e., as m increases indefinitely), the need for ex post and often arbitrary heterogeneity in functional representations and other approximations would be diminished. There are philosophical and empirical/data availability issues bearing upon this "ideal", however, leading to practical admission of heterogeneity as well as functional approximations. In what follows, we focus on the cases in which the assumption of "selection on observables" is plausible.

2. Optimal Multiattribute Functions

Let X_{ij} denote a measure of attribute $j = 1, 2, ..., m$, associated with individual (unit, household, country) $i = 1, 2, ..., n$. Define the covariate matrix $X = (X_{ij})$, X_i its *ith* row, X_j its *jth* column, and consider any scalar function of the matrix X. Examples of such scalar functions are inequality measures or Social Welfare Functions (SWFs), or propensity scores. It has proven difficult to develop "consensus" axioms which may characterize an ideal scalar measure of X, such as aggregators or inequality measures

Our approach is based on measures of closeness and affinity which may identify either attributes that are similar in some sense, *and/or* determine a "mean-value", or aggregate, *which most closely represents the constituent attributes*. The difficulty of selecting a measure of "similarity" has had some resolution in "information theory" which seems to suggest members of the Generalized Entropy (GE) family as ideal criteria of "closeness" or "divergence". Axiomatic characterization of GE is, however, beyond the scope of the present paper. The interested reader may refer to Maasoumi (1993).

Following Maasoumi (1986a), let S_i denote the aggregate or mean function for the *ith* unit. It makes little difference to our approach whether S_i is interpreted as an individual's utility evaluations or the "observer's" or policy maker's assessments for individual i. Let us define the following Generalized Multivariate GE measure of closeness or diversity between the m densities of the chosen m attributes:

$$D_\beta(S, X; \alpha) = \sum_{j=1}^{m} \alpha_j \{ \sum_{i=1}^{n} S_i [(S_i/X_{ij})^\beta - 1] / \beta(\beta+1) \} \tag{1}$$

where $\alpha_j s$ are the weights attached to each attribute. Minimizing D_β with respect to S_i such that $\sum\ S_i = 1$, produces the following "optimal" aggregation functions:

$$S_i \propto (\sum_{j}^{m} \alpha_j X_{ij}^{-\beta})^{-1/\beta}, \beta \neq 0, -1 \tag{2}$$

$$S_i \propto \Pi_j X_{ij}^{\alpha_j}\ , \beta = 0 \tag{3}$$

$$S_i \propto \sum_j \alpha_j X_{ij}, \beta = -1 \tag{4}$$

These are, respectively, the hyperbolic, the generalized geometric, and the weighted means of the attributes, see Maasoumi (1986a). Noting the "constant elasticity of substitution", $\sigma = 1/(1+\beta)$, these functional solutions include many of the well known utility functions in economics, as well as some arbitrarily proposed aggregates in empirical applications. For instance, the weighted arithmetic mean subsumes a popular "composite welfare indicator" based on the principal components of X, when $\alpha_j s$ are the elements of the first eigen vector of the $X'X$ matrix; see Ram (1982) and Maasoumi (1989a).

The "divergence measure" $D_\beta(.)$ forces a choice of an aggregate vector S= (S_1, S_2,.....,S_n) with a distribution that is closest to the distributions of its constituent variables. This is especially desirable when the goal of our analysis is the assessment of **distributed outcomes**, such as treatment effects. But it is desirable generally since, empirically, we have no other "information" than the distribution of variables. Information theory establishes that any other S would be extra distortive of the objective information in the data matrix X in the sense of $D_\beta(.)$. Elsewhere it has been argued that such a distributional criterion is desirable for justifying choices of utility, production, and cost functionals since such choices should not distort the actual market allocation signals that are in the observed data. The distribution of the data reflects the outcome of all decisions of all agents in the economy; see Maasoumi (1986b). The divergence criterion

here is α_j-weighted sum/average of pairwise GE divergences between the "distributions" S and X_j, the jth attribute/column in X.[1]

2.1 A Closer Look

The reason we should explicitly deal with this problem of *representation* is to avoid *implicit* aggregation rules and representations which may not be tenable when exposed. For instance, in the propensity scores method a linear function of x would imply infinite substitutability between individual characteristics and other covariates! Any scalar functional of x is a summary measure of it, and possesses an induced "distribution". This summary distribution must not be incompatible with the information available on its constituent variables. While there are no consensus criteria for "closeness" or similarity in sciences, there are some very good ones that are supported by well thought out axioms and properties. The axiom system that identifies GE is constructive and is to be appreciated as an important breakthrough in organizing learning and knowledge; see Maasoumi (1993). The "fundamental welfare axioms" of symmetry, continuity, Principle of Transfers, and decomposability identify GE as the desirable scale invariant family of **relative** "inequality" measures.

In the context of our discussion above, PS first takes a convenient function (typically linear, but often augmented with polynomial terms), $g(x)$ say, and then transforms this into the unit interval by inversion through some cumulative distribution function (CDF), F *say*; Normal or logistic CDFs predominate in determining "treatment probability". While we may view this last transformation as a convenient standardization, we are less comfortable with its meaning. Firstly, neither $g(x)$ nor $F(g(x))$ has any clear justification as a representation of individuals/households/policy makers in the sense made clear earlier. Second, the parameter values, or indeed even semi-parametric estimates of these functionals, are determined with reference to optimal estimation rules which, again, have no clear connection to optimal representation or aggregation. In addition, the criticism in Heckman and Navaro-Lozano (2004), about the choice of covariates in x being based on estimation and model fit measures, applies. In addition, estimating conditional means or quantiles, further impinges on the plausibility of variable choices and coefficient values/factor loadings. Another discomforting aspect of the PS type solution is that the conditioning set is almost always made of the same fundamental and frequently observed/cited x variables, whatever the experiment or treatment. In the extreme, it is therefore possible that $g(x)$ will be called upon to determine selection into a drug treatment program, as well as a welfare program! Finally, matching based on propensity scores is a decision on similarity of individuals based on *estimated means* of whole distributions, not the true means. Furthermore, many different distributions of individuals, and of heterogeneous groups, may have similar or even the same means. The "control function" method favored by Heckman and Navaro-Lozano (2004) is a second moment-based approach to partially deal with this shortcoming of the PS method.

[1] In unpublished work, the authors have considered hyperbolic means of these same pairwise divergences. This generalization is capable of producing more flexible functional forms for S_i.

3. An Application

We use the data from Dehejia and Wahba (1999), which is based on Lalonde's (1986) seminal study on the comparison between experimental and non-experimental methods for the evaluation of causal effects. The data combine the treated units from a randomized evaluation of the National Supported Work (NSW) demonstration with non-experimental comparison units drawn from PSID data. We restrict our analysis to the so-called NSW-PSID-1 subsample consisting of the male NSW treatment units, and the largest of the three PSID subsample.[2]

The outcome of interest is real earnings in 1978; the treatment is participation in the NSW treatment group. Control covariates are age, education, real earnings in 1974 and 1975, and binary variables for black, hispanic, marital status and a degree. The treatment group contains 185 observations, the control group 2490 observations, for a total of 2675. PS specification includes the quadratic terms for age, education, real earnings in 1974 and 1975, an interaction term between black and unemployment as well as the linear covariates described

[2] The same data set was also utilized in Becker and Ichino (2000) and Smith and Todd (2005).

above; S_i, on the other hand, includes only the linear covariates.[3]

In Table 1, matching is based on the aggregator functions, S_i, as well as on the (Normal) CDF transform of the S_i; $\Phi(S_i)$. The latter is done for direct comparability with the PS results, and to draw out the impact of different substitution and other parametric values.[4] Matching based on S_i utilize the equal weights of α_j (1/8), which is the number of linear covariates. We use the resulting coefficients of the linear probit model for α_j when we match based on $\Phi(S_i)$. The standard errors were obtained by simple bootstrap and are given in parentheses. It is clear that inferences vary dramatically depending on which functional forms, weight parameters or βs are used. The case with $\beta = -1$ is the case of "infinite substitutability" between covariates in the linear in parameters models. When less substitution is allowed, we have dramatically different results, including the reversal of the sign for the ATT! This further reinforces the criticisms offered recently in Heckman and Navaro-Lozano (2004), and Smith and Todd (2005).[5]

Table 1: ATT Estimations with Different Matching Techniques

	ATT (Standard Error)	Kullback-Leibler Divergence	
		Treated/Nontreated	Nontreated/Treated
Matching with Propensity Score	1654.566 (1042.161)	0.0015	0.0014
Matching with S(i), β=-1	2263.860 (1785.033)	3.5109e-008	3.5106e-008
Matching with S(i), β=-1/2	606.618 (1082.455)	2.2390e-005	2.2456e-005
Matching with S(i), β=-2/3	-557.414 (1663.907)	2.3058e-007	2.3079e-007
Matching with Φ(S(i)), β=-1	1545.518 (928.496)	0.0054	0.0051
Matching with Φ(S(i)), β=-1/2	-6929.172 (1709.479)	1.9903e-005	1.9836e-005
Matching with Φ(S(i)), β=-2/3	-1398.497 (1156.810)	4.2654e-004	4.2509e-004

NOTES: (i) Propensity Score estimation include: age, age squared, education, education squared, real earnings in 1974 and its square, real earnings in 1975 and its square, dummies for black, hispanic marital status, nodegree, and an interaction term between black and nonemployment in 1974. (ii) S(i) index include : age, education, real earnings in 1974, real earnings in 1975 and dummies for black, hispanic, marital status and nodegree with weight given by 1/8; the number of linear covariates. (iii) Φ(S(i)) includes the same specification as in S(i). The weights for Φ(S(i)) are the corresponding probability weights from the probit equation. (iv) Standard errors are obtained via 500 bootstrap replications. (v) Sample Size=2675, Treated Units=185 and Nontreated Units=2490. (v) Kullback-Leibler Divergence Measure is obtained via a Gaussian kernel with fixed bandwidth, see text for further details.

[3] Unemployment is a binary variable defined for those who report non-zero earnings in 1974.

[4] We utilize nearest (single) neighbor matching in estimation of ATT. Kernel matching with different kernels/bandwidths produce similar results for ATT as of nearest neighbor matching.

[5] We cannot use the interesting case of $\beta = 0$ when there are binary variables since this will produce a zero value for the generalized geometric function!

In Table 1 we also provide an idea of proximity of the distributions of the propensity scores between the experimental and the matched non-experimental groups. This can be used to evaluate the efficacy of the matching technique employed (here, the nearest neighbor), not as in the traditional way of assessing how close two or more matched pairs are, but how well "similar" the two samples are. This is of some importance given the recent debate regarding the sensitivity and lack of robustness to the choice of subsamples being used by various investigators; see Smith and Todd (2005) and response by Dahejia (2005). KL and other entropic measures give an idea of how far apart the two distributions are.

The KL measure is defined as $I(f,g) = \int f \log \frac{f}{g} d_x$ where f and g are the distributions of the treated and control groups for a variable x, respectively. A symmetrized version of this is the well-known KL measure, but we report both of the asymmetric measures which need to be averaged to get KL in the third column of Table 1.[6] The KL measures are smaller for S_i functions and the comparable $\Phi(S_i)$ than matching with the traditional PS.

Figures 1-2 demonstrate the distribution of the matched pairs by each technique. Specifically, these are frequency distributions of PS for each individual, and the comparable transformation $\Phi(S_i)$. Better "distribution" matches are obtained when $\Phi(S_i)$ is used with finite substitutability.

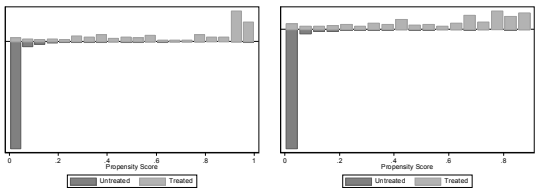

Figure 1: Frequency Distributions of PS and $\Phi(S_i)$ for $\beta = -1$

[6] The KL measures are obtained using a Gaussian kernel and a fixed bandwidth.

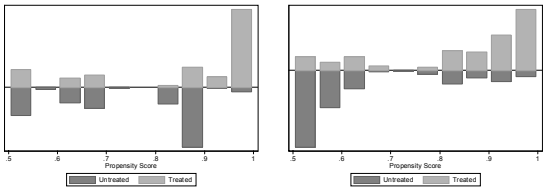

Figure 2: Frequency Distributions of $\Phi(S_i)$ for $\beta = -1/2$ and $\beta = -2/3$

Traditional PS matching seems to correspond to very different distributions of scores for the treated and untreated. The non-experimental subjects in the PSID were not considered for treatment, but should have similar scores if the PS matching theory is valid. It is evident that at lower levels of substitution between characteristics, treatment probability is increased for the untreated sample. Again these figures are computed using probit regression estimates for the α weights in the $\Phi(S_i)$ functions. Naturally, there would be some variation in these results as we change the α values. Basing these values, and the β, on estimation techniques is a decision to vary these graphs (PS values) to obtain statistically optimal values for treatment/selection probability. It is difficult a priori to see very many contexts in which this would be reflective of policy optimality, or in keeping with random assignment.

These results make clear the difficulties of assessing treatment effect, and expose the nature of the implicit choices being made which appear disconnected from common program objectives. Data "snooping", such as balancing, and estimation of flexible or semi-parametric functional forms for PS regressions, are attempts at changing the above distribution matches, but without the benefit of guidance by whole distribution measures of "match", and often distracted by statistical cost functions that need to be justified.

4. Treatment Effects: Stochastic Dominance

Following the IT reasoning above, it is possible and desirable to assess the similarity, or any relation between the **distributions** of outcomes for the treated and untreated/counterfactual groups. Entropy measures such as KL can be used both to quantify differences and to test if the differences are significant. Any summary measure of this divergence/distance corresponds to a cardinal evaluation function that puts different weights on different subjects. This is true of the "average" treatment effects. Averages are particularly non-robust to outliers and special outcomes which may have nothing to do with the treatment. But averages, or specific quantiles, have the advantage of easy monetization for standard cost benefit analysis of policy outcomes. It would be useful to examine whether the *distribution* of outcomes is rankable between two scenarios, no matter what weights are attached to different subgroups. If it is desired to assess the impact of treatment on the entire "sample", and to avoid full cardinalizationrios, such as in "average effect" or choice of any quantile, weak uniform ranking presents itself as a useful method. Such rankings are achieved by tests for stochastic dominance. Various orders of SD can provide guidance and eliminate certain summary measures. For instance, if no First Order SD is found, one cannot employ averages, medians, and other single quantiles to summarize "the" treatment effect.

Let X and Y be two variables. Let U_1 denote the class of all utility functions u such that $u' \geq 0$, (increasing). Also, let U_2 denote the class of all utility functions in U_1 for which $u'' \leq 0$ (strict concavity). Assume $F(x)$ and $G(x)$ are continuous and monotonic cumulative distribution functions (CDFs) of X and Y, respectively. X First Order Dominates Y if F is everywhere to the right of G. X Second Order dominates Y if integrated F is everywhere to the right of integrated G. Lower order dominance implies higher order ones. See Linton et al (2005) for statistical tests of the relevant hypotheses. Below we examine the sample values graphically. No significance levels are reported as necessary descriptions cannot be included in this limited space.

4.1 SD rankings for DW data

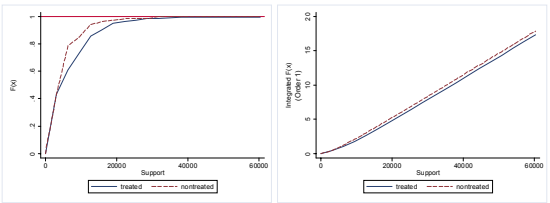

Figure 3: CDFs and Integrated CDFs of Propensity Score (FSD not ruled out, SSD indicated)
Similar results for the CDFs and Integrated CDFs for S_i when $\beta = -1$; FSD not ruled out, SSD likely.

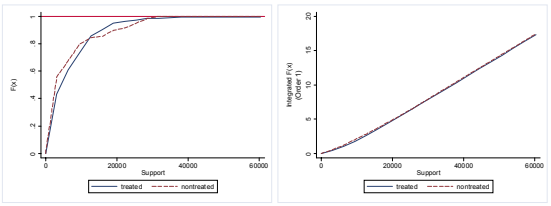

Figure 4: CDFs and Integrated CDFs for S_i for $\beta = -1/2$

This case is instructive. The CDFs cross, indicating different groups are impacted differently and not uniformly. A discussion of averages and quantiles leaves the policy maker in a quandary. But since even SSD is not certain, a simple risk aversion, or inequality aversion, would not allow the policy maker to make uniform statements about program effects. Given the indication of 3rd order SD, it appears that only a policy maker, that is a welfare summary measure, which cares increasingly more about transfers at the lower end of earnings distribution, might declare the program a benefit (irrespective of costs). The case for $\beta = -2/3$ is qualitatively similar to the case of $\beta = -1/2$.

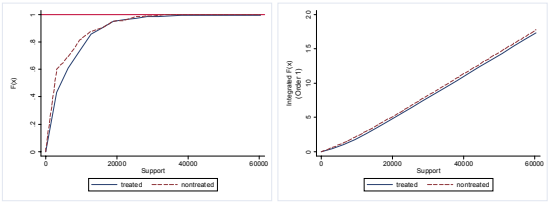

Figure 5: CDFs and Integrated CDFs for $\Phi(S_i)$ for $\beta = -1$

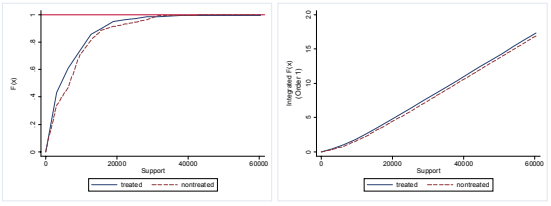

Figure 6: CDFs and Integrated CDFs for $\Phi(S_i)$ for $\beta = -2/3$

The case of $\beta = -1/2$ is qualitatively identical to this case.

References

[1] Becker, S. O. and A. Ichino (2000), "Estimation of Average Treatment Effect Based on Propensity Score", *Stata Journal,* 2, 1-19.

[2] Dahejia, R.H. (2005),"Program evaluation as a Decision Problem", *Journal of Econometrics*, 125, 141-173.

[3] Dahejia, R., and S. Wahba (1999), "Causal Effects in Non-Experimental Studies: Reevaluating the evaluation of training programs", *Journal of American Statistical Association*, 94, 1053-1062.

[4] Deheija, R. and S. Wahba (2002), "Propensity Score-Matching Methods for Nonexperimental Causal Studies," *Review of Economics & Statistics*, 84, 151-161.

[5] Granger, C., E. Maasoumi and J. S. Racine (2004), "A Dependence Metric for Possibly Nonlinear Time Series", *Journal of Time Series Analysis*, vol. 25, 5, pages 649-669.

[6] Heckman, J. and S. Navarro-Lozano (2004), "Using Matching, Instrumental Variables, and Control Functions to Estimate Economic Choice Models," *Review of Economics & Statistics*, 86, 30-57.

[7] Imbens, G.W. (2004), "Nonparametric Estimation of Average Treatment Effects Under Exogeneity: A Review," *Review of Economics & Statistics*, 86, 4-29.

[8] Kolm, S.C. (1977),"Multidimensional egalitarianism," *Quarterly Journal of Economics*, 91**,** 1-13.

[9] Lalonde, R. (1986),"Evaluating the Econometric evaluations of the Training Programs with Experimental Data", *American Economic Review*, 76, 604-620.

[10] Linton, O., E. Maasoumi, and Y. J. Whang (2005),"Consistent Testing for Stochastic Dominance under general Sampling Schemes", *Review of Economic Studies*, forthcoming.

[11] Maasoumi, E.(1986a),"The Measurement and Decomposition of Multi-Dimensional Inequality," *Econometrica*, 54, 991-97.

[12] _____(1986b),"Unknown regression functions and efficient functional forms: An interpretation," *Advances in Econometrics,* 5**,** 301-9.

[13] Maasoumi, E.(1989a),"Composite Indices of Income and Other Developmental Indicators : A General Approach," *Research on Economic Inequality*, Vol. 1, 269-286.

[14] _____. (1993),"A Compendium to Information Theory in Economics and Econometrics," *Econometric Reviews*, Vol 12, 3, 1-49.

[15] Maasoumi, E., and G. Nickelsburg (1988),"Multivariate Measures of Well Being and an Analysis of Inequality in the Michigan Data," *Journal of Business and Economic Statistics*, 6, 3, 327-334.

[16] Ram, R.(1982),"Composite Indices of Physical Quality of Life, Basic needs Fulfillment, and Income : A ·Principal Component· representation," *Journal of Development Economics*, 11, 227-47.

[17] Sen, A. (1970a) Collective choice and social welfare, Holden Day: San Francisco, (reprinted, North-Holland, 1979).

[18] _____(1970b),"Degrees of cardinality and aggregate partial orderings," *Econometrica*, 43, 393-409.

[19] _____(1977) ,"On Weights and Measures: Informational constraints in social welfare analysis," *Econometrica*, 45, 7, 1539-1572.

[20] _____(1980),"The class of additively decomposable inequality measures," *Econometrica*, 48, 613-625.

[21] Smith, J. A. and P.E. Todd (2005),"Does Matching overcome LaLonde's critique of nonexperimental estimators", *Journal of Econometrics,*125, 305-353.