UNSUPERVISED PROPENSITY SCORING: NN & IV PLOTS

Bob Obenchain, US Medical Outcomes Research
Lilly Technology Center South, Indianapolis, IN 46285-5024

Key Words: unsupervised learning, propensity scores, non-overlapping patient clusters, local control, nearest neighbors, instrumental variables, density estimation of local average treatment effects, covariate imbalance, sensitivity analyses.

We illustrate ways to visualize treatment effects using graphical displays of information from within and across "clusters" of patients who have been relatively well matched on their baseline covariate characteristics. We start by motivating use of methods for unsupervised learning to bypass not only parametric estimation of unknown, true propensity scores but also the need to check that conditional covariate independence has been achieved. We then show how nearest neighbor (NN) methods that emphasize within-cluster outcome differences due to treatment systematically differ from the instrumental variable (IV) approach that models within-cluster average outcome regardless of treatment as a function of treatment imbalance across clusters. Detecting differential patient response to treatment then involves fitting mixture-density models to the observed distribution of local average treatment effect (LATE) differences in outcome(s). A major advantage of the proposed graphical, clustering approaches is that they encourage use of up-front sensitivity analyses, where the analyst varies the number of clusters and explores both alternative clustering algorithms and alternative metrics for defining dissimilarity between subjects.

1.0 Introduction

The tutorial by D'Agostino(1998) provides a good introduction to the three most commonly used approaches to Propensity Score (PS) adjustment for treatment selection bias. The sub-classification or "binning" approach is firmly based upon the early work of Cochran(1968). Personally, I first explored nearest neighbors matching as a substitute for a more formal statistical model in a Bell System measurement plan, Obenchain(1979). But it was Rosenbaum and Rubin(1983, 1984) who first discussed the covariate conditional independence property of patient matching on propensity score; see equations [2] and [2′] and their discussions below.

Instrumental Variable (IV) methods for bias adjustment using patient matching and/or clustering have recently been emphasized in the econometric, medical and statistical literature; see Imbens and Angrist(1994), McClellan, McNeil and Newhouse (1994) and Angrist, Imbens and Rubin (1996), respectively. Again, the most basic concepts here appear to be quite "old." For example, the most simple special case of IV estimation using clusters is identical to the "grouping estimator" of Wald(1940); see equations [4] and [5] and their discussion below.

Treatment effects will be visualized here in much more general ways than just as a single degree-of-freedom main-effect within an ANCOVA model. In fact, we address the fundamental questions: Given data on two groups of patients, including their pre-treatment characteristics and their post-treatment outcomes, how might one go about making truly data-driven treatment comparisons? Does it make any real difference if patients/doctors selected the treatment thought to be best? In other words, treatment selection bias may be present in the data, but we do assume that any deliberate treatment decisions were based upon observed, pre-treatment characteristics of the patients (evidence based medicine.)

We then go on to describe our general strategy and tactics for identifying and quantifying treatment effects using non-overlapping clusters to focus attention upon differences in treatment outcomes from Nearest Neighbor (NN) patients who are relatively well matched on whichever pre-treatment characteristics are thought to be most relevant to the treatment outcome(s) of interest. Our approach calls not only for visual display of the across-cluster distribution of within-cluster (NN) local average treatment effect (LATE) differences but also for up-front sensitivity analyses about how the form of this distribution tends to vary with choice of [a] patient dissimilarity metric, [b] unsupervised learning algorithm, and [c] number of clusters.

In prospective design of experiments, one objective of random assignment of treatment to experimental units is to increase the likelihood that the resulting treatment groups will be relatively well balanced on pre-treatment characteristics of the experimental units …both their known characteristics and, perhaps even more importantly, their unknown characteristics. When pre-treatment characteristics (such as patient frailty and disease severity) impact likely treatment outcomes, analyses which ignore observed imbalance lead to biased (unfair) treatment comparisons.

When patient groups receiving different treatments are observed to systematically differ on known baseline characteristics, due either to nonrandom treatment assignment or to relatively poor luck in randomization, a variety of methods are typically used to reduce or

avoid bias in the resulting treatment effect estimates. These methods include covariate adjustment using regression models, propensity scoring adjustments based upon discrete choice models for treatment selection, and instrumental variable adjustment via simultaneous equations models. Here, we show how these relatively well known methods actually suggest meaningful ways to define and summarize NN/LATE difference distributions.

Our symbolic notation for variables (available or missing) for patients will be: y = observed outcome variable(s), t = observed treatment assignment (binary, 0 or 1; usually non-random), x = observed pre-treatment characteristics [covariates, instrumental variables.]

The (usually unknown) true propensity score for a patient is defined to be the conditional probability that the patient (or his/her doctor) will "select" treatment number one given the patient's vector of pre-treatment $\boldsymbol{x}$-characteristics:

$$\text{PS: } p = p(\boldsymbol{x}) = \Pr(t = 1 \mid \boldsymbol{x}) = E(t \mid \boldsymbol{x}). \qquad [1]$$

The fundamental conditional independence theorem of propensity scoring, Rosenbaum and Rubin (1983, 1984), then states that

$$\Pr(\boldsymbol{x}, t \mid p) = \Pr(\boldsymbol{x} \mid p)\,\Pr(t \mid p) \qquad [2]$$

In words, conditional upon any given numerical value of true propensity score, the distribution of baseline patient characteristics is statistically independent of treatment selection. Mathematically, this theorem states that the joint distribution of $\boldsymbol{x}$ and $\boldsymbol{t}$ given the true PS must **factor** as in [2].

This is a relatively simple but truly profound result in statistics and probability that requires only very weak assumptions. In fact, equation [2] appears to have at least four possible interpretations!

[a] Propensity scores are known constants only in randomized studies. Thus [2] can be viewed as the basis for randomized clinical trials in which entire treatment groups are expected to be directly comparable simply because true PS are identical for all patients.

[b] Equation [2] is commonly called the "balancing" theorem because it describes the (expected) behavior of baseline covariate **x**-distributions when propensity scores are either known (and the randomization was relatively "lucky" or balanced) or can be estimated well from the available data (an "unlucky" randomization or treatment selection bias present.)

[c] Equation [2] can also be viewed as establishing the need for blocking or Local Control (LC) analyses whenever propensity scores (local treatment administration fractions) vary widely across **x**-space.

[d] Justifiable variation in propensity scores can be taken as the very definition of and motivation for evidence based medicine, which requires evidence of differential patient response to treatment.

Additionally, Rosenbaum and Rubin (1983) point out that the true PS is the "most coarse" possible balancing score, while the $\boldsymbol{x}$-vector itself is the most highly detailed balancing score:

$$\Pr(\boldsymbol{x}, t) \equiv \Pr(\boldsymbol{x})\,\Pr(t \mid \boldsymbol{x}). \qquad [2']$$

This is extremely important because patient $\boldsymbol{x}$-vectors are usually observable (known) quantities while (coarse) PSs are usually unknown and must be estimated from the available data, say, via a discrete choice (logit or probit) model. When the observed conditional distributions of baseline patient covariates and treatment choices "fail to factor" as in [2], this is quite rightfully interpreted as evidence that one's fitted PS estimates are not even approximately correct. In fact, we will argue below (Section 1.3) that there are distinct advantages to clustering of patients on their entire $\boldsymbol{x}$-vectors rather than matching them "closely" on questionable, numerical PS guesstimates.

1.1 The Non-randomized Abciximab Study

Kereiakes et al. (2000) describe an 18 month study that collected two primary outcome measurements (total cardiac related cost and treatment effectiveness = expected life years preserved due to survival for at least 6 months) for 996 Percutaneous Coronary Intervention (PCI) patients. Researchers used careful telephone follow-up to augment hospital billing and cath-lab records for all patients who had received a PCI (or Percutaneous Transluminal Coronary Angioplasty) at Lindner within 1997.

1.2 Fundamental "Clustering" Concepts

As illustrated in Kereiakes et. al (2000), careful application of propensity scoring methods of adjustment for treatment selection bias requires great attention to detail, including at least the following "Three Initial Steps":

1. Parametric modeling of the treatment assignment mechanism, perhaps using a fitted linear functional of observed patient $\boldsymbol{x}$-factors, to produce an estimated propensity score for each patient;
2. Grouping of patients into, say, 5 adjacent bins (quintiles) using observed PS order statistics; and
3. Testing for "balance" of $\boldsymbol{x}$-factor distributions within each of these (relatively large) bins.

Here we propose a highly graphical, computationally feasible way to bypass these three initial steps and yet end up with an even better (more robust) view of the effects adjusted for treatment selection bias and imbalance.

See Kaufman and Rousseeuw(1990) for descriptions of newer methods for forming non-overlapping clusters

of patients. These methods have recently been implemented in R and are incorporated into the PS functions of Obenchain(2004.) Clustering methods and mixture-density estimation, Fraley and Raftery(2002), are known in standard "data mining" terminology as unsupervised learning algorithms, Barlow(1989.) In sharp contrast, regression models and discriminant analyses are supervised methods in the sense that an observed variable, ***y*** or ***t***, is to be predicted from ***x***. These observed variables can efficiently "guide" selection of (relatively smooth) functions of the ***x***-variables to make the needed predictions. Unsupervised methods need to identify patient and/or treatment outcome "closeness" relationships that may be impossible to visualize in only 2 or 3 dimensions!

In other words, efficient clustering of patients in ***x***-space and outcomes in LATE difference space is a truly difficult (NP hard) problem, and "greedy" computing algorithms almost surely need to be avoided. Our proposals for clustering approachs to adjustment for treatment selection bias thus definitely calls for "standing on the shoulders of computational giants."

1.3 Technical Problems in Estimation of Propensity "Balancing" Scores

Consider the following explanation of why grouping of subjects on their estimated propensity scores does not automatically assure ***x***-factor balance. One's prediction formula is frequently of the form

$$PS = Pr(t = 1 | \boldsymbol{x}) = \text{function}(\boldsymbol{x}'\boldsymbol{\beta}) \text{ for a specific, estimated } \boldsymbol{\beta} \text{ vector,}$$

at least when using a logit or probit model. In the above equation, $\boldsymbol{x}'\boldsymbol{\beta}$ is the fitted **linear functional**, and the elements of the β vector obviously need to be estimated to predict the "outcome" …which here is actually just a treatment assignment (0 or 1) indicator. Again, the real problem for the analyst is that he/she does not know whether "interaction" terms (products of two or more individual ***x*** variables) or "curvature" terms (powers of individual ***x*** variables) will be needed. If the final ***x***-vector (including cross terms and power terms) contains, say, 9 components, then $\boldsymbol{x}'\boldsymbol{\beta}$ = constant actually defines an unbounded 8-dimensional hyperplane (linear subspace) embedded within 9-dimensional ***x***-space.

The important thing to note here is that two different subjects with the exact same PS estimates may still have very different ***x***-vectors, $\boldsymbol{x}_1$ and $\boldsymbol{x}_2$. All we really know for sure here is simply that $\boldsymbol{x}_1'\boldsymbol{\beta} = \boldsymbol{x}_2'\boldsymbol{\beta}$ for one specific vector of $\boldsymbol{\beta}$ estimates.

"Cluster-binning" of patients (especially when clusters are relatively numerous and small) assures that any two subjects within the same cluster will be fairly well matched on all components of their entire ***x***-vectors. Here, we denote this by $\boldsymbol{x}_1 \approx \boldsymbol{x}_2$ …where "≈" denotes "approximately equal." Note that $\boldsymbol{x}_1'\boldsymbol{\beta} \approx \boldsymbol{x}_2'\boldsymbol{\beta}$ would then follow for a variety of different $\boldsymbol{\beta}$ vectors. In other words, the numerical values of $\boldsymbol{x}'\boldsymbol{\beta}$ will be assured to be approximately equal for all patients within any relatively compact cluster simply because their corresponding ***x***-vectors are then nearly equal!

Furthermore, $\boldsymbol{x}_1 \approx \boldsymbol{x}_2$ within any single cluster is the very definition of ***x***-factor "balance" when the two subjects (numbered 1 and 2 here) being compared actually received different treatments.

In fact, let us now consider a heuristic restatement of the fundamental "balancing" theorem in which we condition upon cluster-bin membership rather than upon propensity score. Suppose that the current clusters are C = 1, 2, …, K and that we are interested in the joint distribution of patient ***x***-characteristics and *t*-selections within a cluster:

$$\begin{aligned} Pr(\boldsymbol{x}, t | C) &\equiv Pr(\boldsymbol{x} | C) \, Pr(t | \boldsymbol{x}, C) \\ &\approx Pr(\boldsymbol{x} | C) \, Pr(t | C) \end{aligned} \qquad [2'']$$

The first line in equation [2″] again follows from the very definition of conditional probability. When a cluster is relatively compact, the Pr(***x*** | C) distribution should tend to be both unimodal and fairly tight about the cluster ***x***-centroid. Furthermore, if treatment selection [t = 0 or 1] does not depend upon the highly limited ***x***-variation allowed within a cluster, then Pr(*t* | ***x***, C) ≈ Pr(*t* | C) = expected fraction of patients with t = 1 within cluster C.

Conditioning upon membership in the same ***x***-cluster, as in [2″], is conceptually somewhere between the two possible extremes of "most coarse" and "most detailed" balancing score whenever clusters are relatively numerous and thus are both small and compact.

2. Hierarchical Clustering of Subjects in *x*-Space

The objective of this sort of analysis is to partition subjects with observed ***x***-factors into disjoint subsets (*clusters*) such that:

- Subjects within a cluster are as similar as possible on their ***x***-factors, and
- Subjects in different clusters are as dissimilar as possible on their ***x***-factors.

2.1 Choice of Patient Dissimilarity Metric

The underlying concept needed for partitioning of a set of objects into subgroups or clusters is a metric for measuring dissimilarity between pairs of patients. A variety of distance and similarity measures, such as the Dice coefficient, the Jaccard coefficient and the cosine coefficient, are apparently widely used. The algorithms

illustrated here in the abciximab case study use Mahalanobis distance, Rubin (1980):

$$d_{ij}^2 = (x_i - x_j)' \left[\hat{\Sigma} \right]^{-1} (x_i - x_j) . \quad [3]$$

Suppose that some patient *x*-characteristics are qualitative factors with either only relatively few levels or else unordered levels. The analyst may then wish to require that all patients within the same cluster match exactly on this particular ***x***. Alternatively, an ***x***-factor with *k* levels can be recoded as *k*–1 "dummy" (binary) variables in equation [3].

If certain ***x***-covariates are being used primarily as "instrumental variables," the analyst may wish to give extra **weight** to these variables in defining patient dissimilarity. For example, McClellan, McNeil and Newhouse (1994) used approximate "distance from the hospital of admission" (derived from ZIP codes) as their initial, key variable in clustering 205,021 elderly patients; the only other available ***x***-characteristics were age, sex and race. With such a gigantic number of subjects, the logical strategy is to start by **stratifying** patients into several distinct distance-from-the-hospital "bands." Smaller clusters can be easily formed within these initial strata by, say, matching patients on both sex and race and then grouping them into age ranges.

2.2 Choice of Clustering Algorithm

Agglomerative (bottom-up) clustering methods start with each subject in his/her own cluster and iteratively combine subjects and/or clusters of subjects to form larger and larger clusters. This is the "natural" way to do unsupervised analyses, and the vast majority of clustering algorithms do work this way.

Divisive (top-down) clustering methods start with a single cluster containing all subjects. Some rather new unsupervised algorithms, such as the "diana" method of Kaufman and Rousseeuw (1990), are divisive.

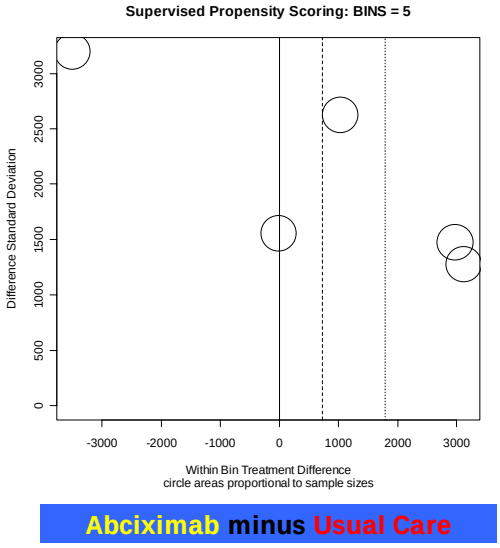
Different choices for ***Number of Clusters*** can then be explored using the clustering "dendrogram," see Figure 3.2.1 below. This graphic depicts the complete "hierarchical" structure derived using a specific patient dissimilarity metric and a specific clustering algorithm. Selecting overall numbers of clusters involves determining heights for a set of horizontal lines that cut across the dendrogram (tree) and produce the desired alternative numbers of clusters.

3.0 Nearest Neighbor (NN) Plots

Let us now consider a "NN snowball plot" to graphically display within-cluster-bin treatment differences and across-bin weighted averages. For example, in Figure 3.0.1, each within-bin average outcome difference (treated minus untreated) provides the horizontal coordinate for a circular plotting symbol, while the corresponding standard deviation provides the vertical coordinate. Finally, the area of the snowball represents the total number of patients (regardless of treatment choice) within that cluster-bin. In addition to a **solid vertical line** corresponding to an outcome difference of zero between treatments, the weighted averages across bins, with weights either (i) proportional to total number of patients or (ii) inversely proportional to the variance of the estimated treatment difference, are also shown using **vertical lines** that are **dashed** and **dotted**, respectively.

While different cluster-bins may be of very different overall sizes (total number of patients regardless of treatment), here they are equal (199 or 200 patients each.) Finally, by specifically displaying within-bin measures of uncertainty (vertical coordinates of snowballs), NN plots illustrate key issues related to homoscedasticity (constant variance) assumptions commonly made in parametric regression and econometric models.

Figure 3.0.1 "NN Snowball Plot" with 5 PS Bins for the LATE distribution of Abciximab on Cost



3.1 "Pure," "Impure" and "Fully Informative" Cluster-Bins

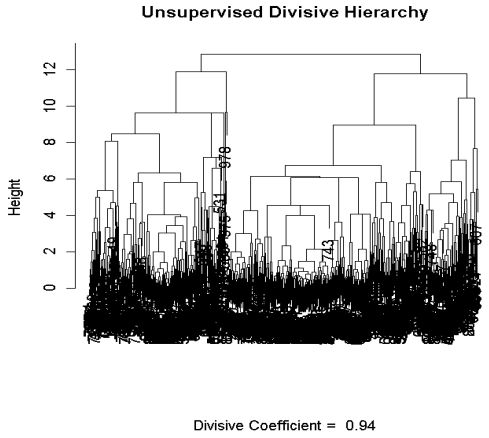
Since subjects are being clustered solely on the basis of their ***x***-characteristics, the outcome variables, *y*, and treatment assignment variable, *t*, may take on essentially any values (consistent with the available data) for the subjects within any single cluster. Here, a cluster will be said to be "pure" if all subjects within that cluster were assigned to the same treatment (either all $t = 0$ or all $t = 1$.) There is no possibility of observing any local outcome (*y*) difference between treatments using only subjects from within a "pure" cluster! In this sense, NN methods end up "discarding" all outcome information that ends up being isolated within pure clusters.

To be "fully informative" about a within-cluster local treatment difference without assuming homoscedasticity of outcomes, a cluster must contain at least 2 patients on each treatment. After all, this many patients are needed to compute the heteroscedastic standard errors of the two treatment outcome means and, thus, the conventional standard error of the resulting local treatment difference!

3.2 NN plots for the Abciximab Case Study

Let us now consider a variety of "highly visual" alternative NN analyses for our abciximab case study. As suggested in section 2.1, on "unsupervised" learning in analyses of non-randomized studies, the starting point for these analyses is the calculation of a dendrogram for hierarchical clustering of all 996 patients on their observed, baseline *x*-characteristics …ignoring their ultimate "*y*" outcomes (survival for at least six months and accumulated cardiac related cost) as well as their "treatment" assignment (abciximab or usual-care-alone.) The resulting dendrogram is displayed in Figure 3.2.1, below.

Figure 3.2.1 Cluster-Bin Dendrogram for the Abciximab Study



Once this sort of dendrogram has been computed, results for a wide range of alternative numbers of "cluster-bins" can be generated quite quickly and efficiently. For example, Figures 3.2.2 (for 30 cluster-bins) and 3.2.3 (for 90 cluster-bins) display abciximab cost adjustment results which turn out to be quite similar to those from "conventional" propensity (quintile) binning, as displayed in Figures 3.0.1. With 30 cluster–bins, the number of patients per bin ranged from 5 to 302. When 90 cluster-bins (a relatively large number of bins for only 996 subjects) were requested, the number of patients per bin then ranged all of the way from 1 (with 32 bins of this minimum size) to 127 (the single, largest bin.)

Note, in particular, that "total cardiac related cost" differences (abciximab minus usual-care-alone) are displayed in Figure 3.2.3 for only the 34 "fully informative" cluster-bins out of the 90 requested.

Twenty additional cluster-bins out of the original 90 contained just one patient on one (or both) of the two treatments and the remaining 36 were "pure." No within cluster-bin treatment difference could be observed for these 36 cluster-bins …meaning that cost outcomes for 82 of the 996 patients had to be ignored in this particular analysis. A wide range of treatment cost differences (from –$9,000 to +$18,000) as well as a wide range in uncertainty about within cluster-bin cost differences (up to almost ±$15,000) were also observed. But alternative estimates of the overall main-effect of treatment are embodied by the vertical lines in Figure 3.2.3. The dotted vertical line at +$2,921 (±$428) represents the inverse-variance weighted average cost increase (abciximab minus usual-care-alone) while the dashed line at +$1,432 (±$672) is the average cost increase resulting from weighting within-cluster differences by the overall size of each cluster (total number of patients within that cluster.)

Figure 3.2.2: NN Plot for 30 Requested Clusters

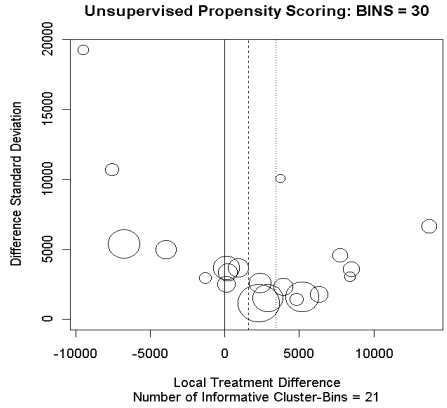
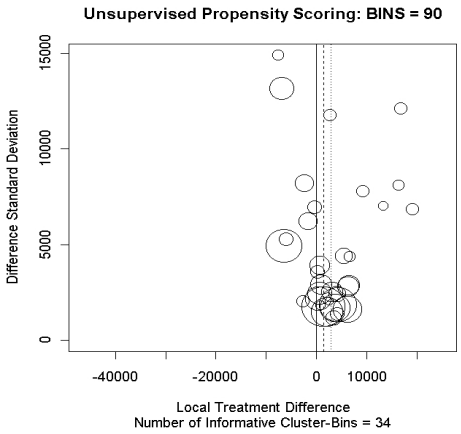


Figure 3.2.3: NN Plot for 90 Requested Clusters



4. Simultaneous Equations Models with Instrumental Variables

While we have seen some rather obvious advantages of using cluster-bins with propensity scoring, having to disregard information from "pure" bins could be a disadvantage. Thus, let us now consider using cluster-bins with econometric instrumental variable (IV) methods to fit models across ALL cluster-bins.

For example, for the first two clusters, the "local" IV estimator corresponds to connecting those clusters with a straight line on the graph where mean within cluster health outcome (regardless of treatment) is plotted against a non-parametric estimate (fraction treated) of within-cluster propensity (PS) to receive treatment:

$$\beta^{IV} = \frac{\mathrm{E}(y \mid C=1) - \mathrm{E}(y \mid C=2)}{\Pr(t=1 \mid C=1) - \Pr(t=1 \mid C=2)} \quad . \qquad [4]$$

In other words, the IV estimate of the causal beta-coefficient is a simple ratio of mean differences. The denominator propensity score estimates in [4] play a key role in econometric IV modeling approaches to adjustment for treatment selection bias.

When many cluster-bin centroids are plotted on the "IV plane," an obvious generalization of the pairwise IV approach would be to fit a line (or smoothing spline) through all of these centroids. The line of best fit to this scatter of cluster centroids can usually be visualized as a weighted average over all pairwise IV estimands resulting from connecting pairs of distinct centroids. Specifically, equation (3.1) of Wu(1986) can be rewritten as

$$\beta^{OLS} = \sum_{i<j}^{n} w_{ij}\beta_{ij}^{IV} \quad \text{for} \quad \beta_{ij}^{IV} = \left(y_i - y_j\right) / \left(x_i - x_j\right), \qquad [5]$$

where the weight given to a pair of cluster-bin centroids to produce the overall OLS estimate is proportional to the square of their separation along the x (propensity score) axis,

$$w_{ij} = \left(x_i - x_j\right)^2 / \sum_{i<j}^{n} \left(x_i - x_j\right)^2 .$$

The above sort of "smoothing through cluster centroids" reasoning appears to be the primary motivation for the clustering approach of McClellan, McNeil and Newhouse (1994); see their Figure 1.

5. Instrumental Variable (IV) Plots

In this section, we define the graphical elements of an "IV plot," discuss the extreme cases where the number of cluster-bins is either very small or very large, and display and interpret a pair of IV plots for the cost outcome in the abciximab case study.

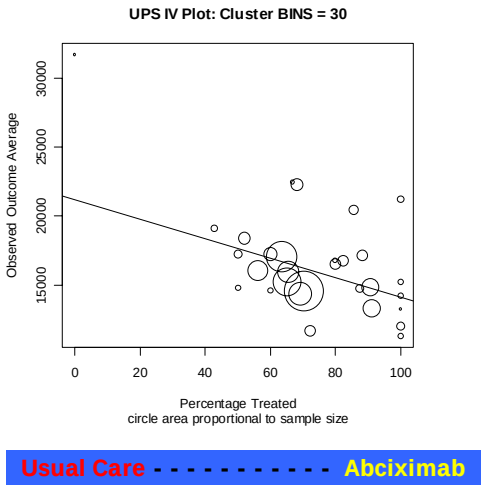
Note that econometric "IV Plots" display information from the "pure" cluster-bins that must be discarded when visualizing local treatment differences using "NN Plots." While attempting to use only relatively few clusters tends to avoid "pure" clusters, using relatively many clusters (and thus forcing some clusters to be pure) frequently appears to be needed to coerce the resulting "IV Plot" into agreeing (at least approximately) with the corresponding "NN Plot!"

Varying the number of clusters used provides visual "sensitivity analyses" for comparing the NN and IV approaches. Researchers need to be able to literally "see" which of these alternative analyses appear to be most realistic, relevant and robust for their data …not just which approach is "generally" recommended.

5.1 IV Plots for the Abciximab Case Study

Let us now discuss the "IV adjustment" analyses displayed in Figures 5.2.1 and 5.2.2 for 30 and 90 cluster-bins, respectively. Here we are looking at within-bin average cost (regardless of treatment assignment) plotted versus within-bin estimates of propensity score expressed as a percentage of patients treated with abciximab within each cluster-bin. Instead of ignoring "pure" cluster-bins in these graphics, pure cluster-bins now contribute observed results at the left-hand and right-hand extremes, 0% and 100% of subjects treated with abciximab, respectively. In fact, outcomes from these extreme cluster-bins may have high leverage on any fitted across-cluster smooth!

Figure 5.2.1: IV Plot for 30 Requested Clusters

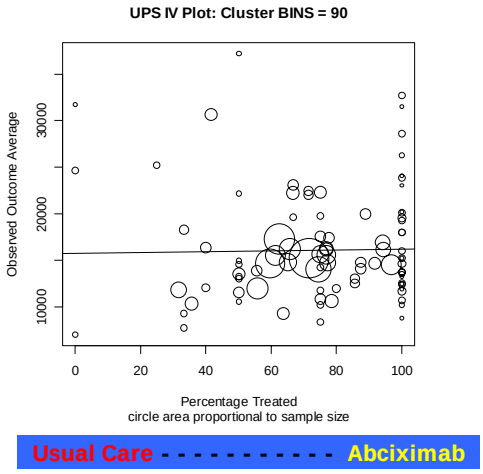


Note that Figure 5.2.1 represents a truly "very different" sort of result from all of the other abciximab cost analyses displayed so far. Specifically, the "IV adjustment" analysis using 30 (or fewer) cluster-bins quite clearly suggests that total cardiac related costs are expected to decrease with increased abciximab use.

On the other hand, Figure 5.2.2 reconfirms instead our earlier cluster-bin analyses (Figures 3.3.2 and 3.3.3) …but only when using a relatively large number of

cluster-bins (90 or more) in an "IV adjustment" analysis. Because cluster-bins tend, on average, to be much smaller (as measured by total numbers of patients) when there are many more of them, these more numerous cluster-bins also tend to contain relatively "more well-matched" patients …and to produce estimates the are more realistic overall.

Figure 5.2.2 IV Plot for 90 Requested Clusters



In my opinion, the problem in Figure 5.2.1 is that the **x**-covariates in the Kereiakes study are NOT instruments; they are clear predictors of mortality and, thus, of cost. Instead of suggesting that abciximab treatment reduces cost, Figure 5.2.1 actually shows simply that early death truncates cost (i.e. relatively frail patients are more likely to be administered abciximab.)

5.2 Extreme Numbers of Clusters: One per Subject versus One Overall

Tables 5.3.1 and 5.3.2 as well as Figure 5.3.3, below, summarize my current opinions and/or best conjectures about the "optimal" number of cluster-bins for display in NN and IV plots when the objective is to adjust for selection bias when comparing treatments.

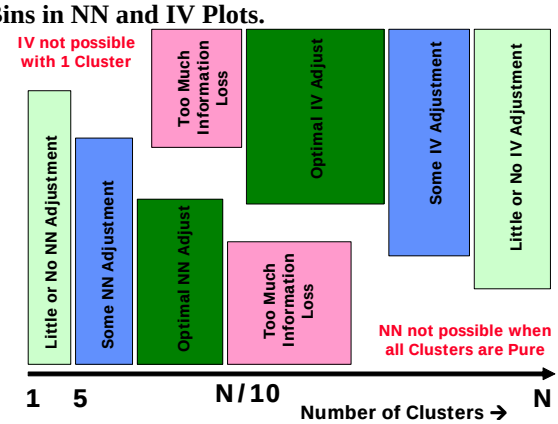
Table 5.3.1. Summary of Plot Content for Extreme Numbers of Cluster-Bins.

	Case [a]: Only One Cluster	Case [b]: Only Pure Clusters
NN plot	No Adjustment for Covariate Imbalance	Plot is "Blank" (Zero Information)
IV plot	Plot void of Treatment Difference Information	No Adjustment for Covariate Imbalance

Table 5.3.2. Conjectures about "Optimal" Numbers of Cluster-Bins.

	Relatively Low Number of Clusters (Too Few?)	Relatively High Number of Clusters (Too Many?)
NN plot	Relatively Little Bias Removal	Uncertainty Inflation due to Pure Clusters
IV plot	Potential Conflict with PS Results	More Consistency with PS Results

Figure 5.3.3. More Conjectures about Loss-of-Information and "Optimal" Numbers of Cluster-Bins in NN and IV Plots.



6. Discussion

Perhaps randomized studies remain the "gold standard" for all types of scientific research. This is unfortunate, especially in health outcomes research settings, where performing "prospective experiments" implies imposing enrollment / participation incentives that result in unrealistic (unnatural) behaviors from both patients and clinicians. There is absolutely nothing wrong with using real-world data to try to realistically answer real-world questions! On the other hand, it is quite obvious that more insights and much better insights could be developed if we had better (much more complete) data and better (interactive, graphical) analysis software. For example, my "R" functions, Obenchain(2004), are clearly simplistic and primitive relative to the full spectrum of clustering concepts envisioned here.

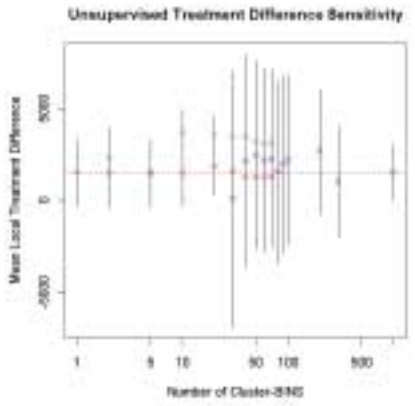
The NN plotting approach discussed here is interesting primarily because it directly addresses the highly relevant subject of the distribution of LATE differences in an "almost" non-parametric way. In fact, the only obvious down-side of this approach is that taking this difference literally doubles the variance of the resulting outcome point-estimates. And the clear

plausibility of using mixture-density estimates to detect differential patient responses is truly exciting.

In contrast, the IV plotting approach discussed here requires parametric (or semi-parametric) modeling across clusters. By averaging outcomes across treatments within cluster-bins, at least IV methods thereby avoid doubling the variance in outcome point-estimates. But failing to examine LATE Differences still strikes me as being a very high price to pay.

Figure 6.1 illustrates how point estimates and confidence intervals for the overall main-effect of treatment with abciximab (mean of its LATE difference distribution) on cardiac related cost for PCI patients varies with number of clusters.

Figure 6.1: Cost Sensitivity Analysis Graphic



NN Cluster Size Weighted Difference
NN Inverse Variance Weighted Difference
IV Predicted Total Treatment Difference

Relative to the "unadjusted" observed difference (abciximab-plus-usual-care minus usual-care-alone) of +$1,513 (±$913), almost all of the other analyses (somewhat curiously) suggest that the true difference is either

- smaller but with higher uncertainty, or else
- larger but with possibly lower uncertainty.

But can't one really say much more here? Perhaps we could if we had much better (more interactive) software for clustering patients and/or mixture-density decomposition of outcome LATE distributions. But it is fascinating that our most clearly relevant comparisons (using 50 to 100 clusters for ~1000 patients) are quite different from the two "extremes."

Anyway, the clear "good news" is that much more sensitive, robust and data-driven methods for assessing treatment effects will soon be available. Meanwhile, the "bad news" is that these approaches will rely heavily on unsupervised learning methods that address the most challenging / difficult problems in statistics!

7. References

Angrist JD, Imbens GW, Rubin DB. "Identification of Causal Effects Using Instrumental Variables." ***J Amer Stat Assoc*** 1996; 91: 444-472.

Barlow HB. "Unsupervised Learning." ***Neural Computation*** 1989; 1: 295-311.

Cochran WG. "The Effectiveness of Adjustment by Subclassification in Removing Bias in Observational Studies." ***Biometrics*** 1968; 24: 205-213.

D'Agostino RB Jr. "Tutorial in Biostatistics: Propensity score methods for bias reduction in the comparison of a treatment to a non-randomized control group." ***Stat Med*** 1998; 17: 2265-2281.

Fraley C, Raftery AE. "Model-based clustering, discriminant analysis, and density estimation." ***J Amer Stat Assoc*** 2002; 97: 611-631.

Imbens GW, Angrist JD. "Identification and Estimation of Local Average Treatment Effects." ***Econometrica*** 1994; 62: 467-475.

Kaufman L, Rousseeuw PJ. ***Finding Groups in Data. An Introduction to Cluster Analysis***. New York: John Wiley and Sons. 1990.

Kereiakes DJ, Obenchain RL, Barber BL, et al. "Abciximab provides cost effective survival advantage in high volume interventional practice." ***Am Heart J*** 2000; 140: 603-610.

McClellan M, McNeil BJ, Newhouse JP. "Does More Intensive Treatment of Myocardial Infarction in the Elderly Reduce Mortality?: Analysis Using Instrumental Variables." ***JAMA*** 1994; 272: 859-866.

Obenchain RL. "Nearest Neighbors Analysis for PRRAP, the Probable Report Rate Analysis Plan." ***AT&T Bell Laboratories*** TM-79-1711-4. 1979. Holmdel, NJ.

Obenchain RL. Unsupervised and Supervised Methods for Propensity Score and Instrumental Variable Adjustment for Treatment Selection Bias in "R." http://www.math.iupui.edu/~indyasa/download.htm. March 2004.

Rosenbaum PR. ***Observational Studies, 2nd Edition***. New York: Springer-Verlag. 2002.

Rosenbaum PR, Rubin RB. "The Central Role of the Propensity Score in Observational Studies for Causal Effects." ***Biometrika*** 1983; 70: 41-55.

Rosenbaum PR, Rubin DB. "Reducing Bias in Observational Studies Using Subclassification on a Propensity Score." ***J Amer Stat Assoc*** 1984; 79: 516-524.

Rubin DB. "Bias Reduction using Mahalanobis Metric Matching." ***Biometrics*** 1980; 36: 293-298.

Wald A. "The Fitting of Straight Lines if Both Variables are Subject to Error." ***Ann Math Stat*** 1940; 11: 284-300.

Wu CFJ. Jackknife, bootstrap and other resampling methods in regression analysis (invited paper with discussion). ***Ann Stat*** 1986; 14**:** 1261-1350.