Abstract Title Page

Title:
Propensity Score Matching Within Prognostic Strata

Authors and Affiliations:
Ben Kelcey
Wayne State University
ben.kelcey@gmail.com

Background / Context:

A central issue in nonexperimental studies is identifying comparable individuals to remove selection bias. One common way to address this selection bias is through propensity score (PS) matching. PS methods use a model of the treatment assignment to reduce the dimensionality of the covariate space and identify comparable individuals. PSs represent the conditional probability of treatment

$$e(\mathbf{X}) = P(Z = 1 \mid \mathbf{X}) \quad (1)$$

where **X** are the observed covariates and Z the treatment indicator. If adjustment on measured covariates is sufficient for unbiased estimation of the treatment effect, then so is adjustment on PSs (Rosenbaum & Rubin, 1983). As a result, PSs act as unidimensional balancing scores in which all the information relevant to balancing treatment assignment on $\boldsymbol{X}$ is extracted in $e(\boldsymbol{X})$. Accordingly, cases with similar PS values but different treatment assignments can serve as estimates of the counterfactual.

In parallel to the PS, recent literature has developed the prognosis score (PG) to construct models of the potential outcomes (Hansen, 2008). Whereas PSs summarize covariates' association with the treatment groups, PGs summarize covariates' associations with the potential outcomes. PGs collapse a high dimensional covariate space into a unidimensional measure summarizing covariates' associations with the potential outcomes. PGs are constructed by modeling the outcome-covariate relations in the control group

$$\Psi(X) = p(Y^{(0)} \mid X) \quad (2)$$

where $Y^{(0)}$ are the potential outcomes for the control group and $\boldsymbol{X}$ are the measured covariates. Like PSs, if adjustment on measured covariates is sufficient for unbiased estimation of the treatment effect, then so is adjustment on PGs (Hansen, 2008). PGs thus provide a unidimensional balancing score in which all the information relevant to balancing potential on $\boldsymbol{X}$ is extracted in $\Psi(\boldsymbol{X})$. PGs also induce a balancing property similar to that of PSs. Conditioning on PGs brings about prognostic balance among the covariate distributions across potential outcomes

$$Y^{(0)} \perp X \mid \Psi(X) \quad (3)$$

The advantage of PGs is that they help to constrain variation in the outcome due to sources other than the treatment thereby favorably reducing both bias and variance of treatment effect estimators. Because PGs are estimated as a function of the covariates using only the control observations, PGs postpone commitment to a functional form for both potential outcomes without risk of inducing bias associated with use of outcomes in the design stage (Iacus et al., 2009). The separation of control and treatment groups is, in part, what differentiates PGs and helps them outperform other estimators (Hansen, 2006).

Purpose / Objective / Research Question / Focus of Study:

In observational studies where the treatment selection mechanism is not completely known, PGs can be seen as an alternative or, more likely, a complement to PSs (Hansen, 2008). In this study, we explored the utility of combining adjustment on PGs and PSs by successively matching on these scores to examine the extent to which full matching on PSs within strata defined by PGs outperforms alternative PS and PS & PG matching schemes.

Significance / Novelty of study:

Prior literature has largely considered complementary adjustment for PSs and PGs in the form of Mahalanobis metric matching. Matching treated and control cases along combinations defined by Mahalanobis distance attempts to equally weight the scores. With Mahalanobis

distances, net distances between cases are formed by the equal adding up of the variance normalized squared distances of the scores. When estimated scores reflect similar levels of precision and importance, Mahalanobis distances are an optimal way of measuring similarities among treated and control cases.

In contrast, when the estimated PGs and PSs differ in their precision and meaningfulness, the squaring of distances potentially allows one score to obscure information provided by the other. Rather than define discrepancies between potential matches by combining PGs and PSs into a single index using Mahalanobis distance, we considered first stratifying on PGs and subsequently matching on PSs within prognostic strata. In this way, our approach tries to imitate a block randomized study whereby subjects are stratified along the PGs and PSs. However, by selecting finer matching on the PSs, we privilege the bias reduction properties of the PS. By selecting coarser subclassification on the PGs, we used them as a secondary tool to further reduce error variance. In this way, our approach suggests that adjustment on PSs is fundamentally more important and sound than adjustment along PGs.

Statistical, Measurement, or Econometric Model:

Unto themselves, PGs have several important limitations that are not shared with PSs. For instance, although both summaries require the nontrivial assumption that all confounding variables have been observed, the asymmetric estimation of PGs necessitates an additional assumption; all effect modifications are captured in the prognostic model. In other words, use of PGs in isolation additionally assumes the sufficiency of the prognostic model for the control potential outcomes is transferable to treated potential outcomes. Moreover, whereas PS balance can be appraised across the entire sample, PG balance can only be appraised in the control sample. However, when paired with PSs, PGs have been shown to reduce both bias and variance of estimated treatment effects (Hansen, 2006).

To integrate PGs and PSs, adjustment has largely come through Mahalanobis distances. Specifically, the two scores are combined into a single index using

$$D_M = \sqrt{(\boldsymbol{X} - \boldsymbol{\mu})^T \boldsymbol{S}^{-1} (\boldsymbol{X} - \boldsymbol{\mu})^T} \quad (4)$$

where $\mathbf{X}$ are the PGs and PSs with means $\boldsymbol{\mu}$ and sample covariance matrix $\boldsymbol{S}$. The comparability of all control and treatment cases is thus defined by a single index, D_M. Matching along this combined index has shown to be effective at both reducing bias and variance (Hansen, 2006). Use of PGs in combination with PSs through Mahalanobis metric matching tends to be particularly favorable in the presence of nonlinear relationships because matching, to some extent, alleviates assumptions of functional form. Accordingly, this type of matching has been shown to outperform both regression models and estimators which base matching only on PSs (Hansen, 2006).

Combining PGs and PSs into a single index, however, may not always be well suited for investigating treatment effects. In particular, matching on unified Mahalanobis distances leaves open the potential for one score to dominate the other. Although Mahalanobis distances are designed to take into account differences in the distributions of the scores through the covariance matrix, the contribution of scores to the distances, and thus matches, can be obscured.

Poor performance of Mahalanobis metric matching has been noted in the presence of rare events (e.g., Rubin & Thomas, 2000). Mahalanobis metric matching tends to place undue emphasis on variables with low prevalence rates because their variance tends to be so small. As a result, when combining PGs and PSs through Mahalanobis distance, it is possible that the distance will favor close matches on the PGs and more coarse matches on PSs. The implications

of this weakness may be particularly acute for studies investigating rare dichotomous outcomes. Conversely, when the control to treatment cases ratio is large, matches may be dominated by PSs.

Similarly, when scores contain outliers or contain excessive noise, the estimated variance can diminish its contributions to Mahalanobis distance. Such sensitivity may be important when outcomes in the control group have a low signal to noise ratio. These difficulties may be further exacerbated because PSs and PGs are constructed using different samples; PSs with the full data and PGs with the control only. As a result, the variance in PGs may often reflect a larger portion of error than do PSs. In part, the poor performance of Mahalanobis metric matching, especially in high covariate space, is what in part led to the development of the PS (e.g., Rubin & Thomas, 2000).

There are also critical differences in the balancing properties of PGs and PSs. The balancing property of PSs can be explicitly appraised and ensured for the full sample. In contrast, prognostic balance can only be examined and ensured for the control sample. And although balance is a necessary but not sufficient condition for bias reduction, the authentication of PG balance using only control cases inherently produces balance and results that are more sensitive to misspecifications. Consequently, the balancing properties of PSs and their implications would seem to be more robust and important than those of PGs.

Finally, our review of the literature suggested that bias reduction should be fundamentally more important than variance reduction in observational studies. Because PGs are primarily used to reduce error variance whereas PSs are primarily used to reduce bias, our review suggests that PSs and PGs should not be equally weighted.

To reduce bias and secondarily to reduce variance while maintaining both propensity and prognostic balance, we propose PS matching within prognostic strata. In other words, to improve the robustness of simultaneous adjustment on PGs and PSs, we propose stratification on PGs followed by within stratum matching on PSs. To this end, matching on PSs within prognostic strata privileges PSs over PGs. In other words, our approach trades coarser matching on PGs for finer matching on PSs.

Usefulness / Applicability of Method:

As a preliminary examination of the utility of this approach, we performed multiple Monte Carlo simulations and highlight one for brevity. This simulation examined the performance of three estimators:

1) Full matching with a 0.1 caliper on PSs only
2) Full matching on PGs and PSs using Mahalanobis distance
3) Full matching on PSs within PG strata (quintiles)

We describe the simplest simulation in which we generated ten covariates (X_i) using independent standard normal distributions with sample sizes of 1000. We designed the treatment, Z, as the realization of a dichotomous variable given covariates which followed

$$logit(P(Z=1)) = \beta_0 + \sum_{m=1}^{10} \beta_m X_{mi} + \beta_{m+1} X_{2i}^2 \quad (5)$$

with β_m set at 0.3. The true outcome model was simulated as a nonlinear function with treatment by covariate interactions so that PGs constructed using only the control group would not be sufficient for the treatment group (different functional forms)

$$Y_i = \beta_0 + \delta Z_i + [\sum_{m=1}^{10} \beta_m X_{mi}] + \beta_{m+1} Z_i X_{1i} + \beta_{m+2} Z_i X_{2i} + \beta_{m+3} X_{2i}^2 + \varepsilon_i \quad (6)$$

where coefficients were analogous to the treatment model and the average treatment effect, δ, was 0.5.

We estimated PSs as

$$logit(P(Z=1)) = \beta_0 + \sum_{m=1}^{10} \beta_m X_{mi} \quad (7)$$

and PGs by estimating and extrapolating the fit using

$$Y_i^{(0)} = \beta_0 + [\sum_{m=1}^{10} \beta_m X_{mi}] + \varepsilon_i \quad (8)$$

Subsequently, using full matching with a 0.1 PS caliper, units were matched on (1) PSs; (2) PGs and PSs through Mahalanobis distances; and (3) PGs and PSs by first stratifying along quintiles of PGs and then matching units on PSs within strata. To estimate the treatment effect for each data set and approach, we combined PS/PG adjustment with linear regression (e.g. Hirano & Imbens, 2002). Using match indicators, we modeled the outcome as:

$$Y_i = \beta_0 + \hat{\delta} Z_i + \sum_{q=1}^{Q} \beta_q M_{qi} + \varepsilon_i \quad (9)$$

where $\hat{\delta}$ is the estimated treatment, M is the matched group, and Q is the number of matched groups minus one. To compare the estimates, we used the results of the Monte Carlo simulations to estimate the bias and mean-squared error (MSE) of each approach. We estimated these quantities using:

$$Bias = \frac{1}{N} \sum_{i=1}^{N} \hat{\delta}_i - \delta \quad (10)$$

and

$$MSE = \frac{1}{N} \sum_{i=1}^{N} (\hat{\delta}_i - \delta)^2 \quad (11)$$

where N represents the number of simulated data sets (1000).

Findings / Results:

The results suggested incorporating PGs into PS matching via PG stratification decreased bias by about 11% when matching on the PS only (1) and by about 7% when matching on PG & PS using Mahalanobis distance (2) (Table 1). Similarly, using the new approach reduced the variance of the estimator by 6% as compared with PS matching only and 10% as compared with PS & PG Mahalanobis matching. Together, we found that the new method reduced mean-squared error by about 17% and 11% as compared with PS matching and PS & PG Mahalanobis matching.

Conclusions:

The results suggested that adjustment using both PSs and PGs provides two avenues by which to block bias and reduce variance. The gains demonstrated in this very limited simulation suggest that strategies which reduce estimator variation while maintaining bias reduction may improve treatment effect estimates in ways parallel to choices in experimental design. Among many limitations of this very small initial investigation, the sequential use of PGs and PSs may have the undesirable effect of reducing the availability of PS matches within strata. In turn, the quality of PS matches, and thus bias reduction, may be negatively impacted. In understanding the utility of incorporating PG adjustment alongside PS adjustment in unequal ways, it will be important to explicate the risks and benefits of these situations.

Appendices

Appendix A. References

Hansen, Ben. 2006. Bias reduction in observational studies via prognosis scores. Technical report #441, University of Michigan Statistics Department.

Hansen, Ben. 2008. "The Prognostic Analogy of the Propensity Score." Biometrika 95(2):481–488.

Hirano, K., & Imbens, G., (2002). Estimation of Causal Effects using Propensity Score Weighting: An Application to Data on Right Heart Catheterization. Health Services & Outcomes Research Methodology, 2, pp. 259–278.

Iacus, S., King, G., & Porro, G. (2009). Causal Inference Without Balance Checking: Coarsened Exact Matching. http://gking.harvard.edu.cem

Rosenbaum. P. R., and Rubin, D. B. (1983). The central role of the propensity score in observational studies for casual effects, *Biometrika,* 70, 41-55.

Rubin, D., & Thomas, N. (2000). Combining propensity score matching with additional adjustments for prognostic covariates.

Appendix B. Tables and Figures

Table 1: Comparison of different matching estimators

	PS Only	PG & PS Using Mahalanobis	PG & PS Using Matching Within Prognostic Strata
Bias x 10^2	8.64	8.24	7.70
Variance x 10^2	1.52	1.43	1.29
Mean-squared Error x 10	2.26	2.11	1.88