Assessment of propensity score and Mahalanobis distance matching in the exposure-response setting

Dan Polhamus, Ph.D., Jonathan L. French, Sc.D.
Metrum Research Group

Abstract

Objectives: Matching methodology in pharmacometrics has been used by the FDA [5] for determining drug effect in the presence of confounding variables. Common matching methods include Mahalanobis distance (MD) and propensity scores (PS) [3]. MD is covariance matrix normalized Euclidean distance and expects continuous data, while PS [1] models conditional treatment probability and allows mixed data. Assessment after matching is crucial and typically includes univariate measures like standardized bias and QQ-plots. Curiously, multivariate comparisons of the covariate distribution are rarely considered. In the case where all data is continuous, we can expect methods that explicitly account for the multivariate distribution of covariates (MD) will retain multivariate balance. However, in data more typical of medical trials can we expect the same? We investigate MD and PS methods to evaluate matching bias and examine new methods for comparing the covariance structure between matched patient samples of mixed data types.

Methods: Using publicly available oncology data (R::survival::colon), we generate hypothetical exposure and confoundedness with exposure as a function of qualitative and quantitative covariates. Matches are found on distances of: A) MD, B) PS matching (0.25 calipers), and C) MD on prognostic covariates (within 0.25 PS calipers). We bootstrap univariate and multivariate summaries of the matched data to quantify similarity.

Results: PS methods result in smaller bias between matched samples, MD methods better preserved the multivariate structure of the data.

Conclusions: Matching with mixed data is simplified by using propensity score methods, but it is important to assess both univariate and multivariate balance after matching. Our results indicate that while PS leads to low bias in matched samples, it does not preserve pairwise correlations as well as MD. We suggest use of MD within propensity score calipers or inclusion of pairwise interaction terms in PS models to preserve the multivariate structure.

Methods

Data: Table 1 describes the R::survival::colon data set. Hypothetical exposure and confoundedness with exposure as a function of the quantitative covariates was generated in order to perform matching between low exposure (Q1) patients and control, a typical comparison used to assess benefit of low exposure patients to higher doses.

Covariate selection: To choose variables for matching, we take a different approach as compared to that recommended in [5]: 1) to avoid any concern of gaming the results, we remain blinded to outcome data for the treated patients when selecting covariates for matching and 2) we aim to match on the complete set of covariates which are possibly predictive of outcome either directly, or indirectly by means of a confounded effect on exposure [2]. We additionally identify a set of highly prognostic covariates in the control both to demonstrate one possible matching procedure, and to assist in assessment of matching quality as related to covariates with high implications for outcome modification [4]. Prognostic covariates *using the control data only* are identified via a simple univariate screen, selecting those that significantly improve fit (via LRT, $\alpha < 0.1$).

Matching: Matches are found on distances of: A) MD, B) PS matching (0.25 calipers), and C) MD on prognostic covariates (within 0.25 PS calipers) using R:::MatchIt. Calipers are generally defined as relative to the standard deviation of the distance metric, hence a 0.25 propensity caliper is $0.25 * \sigma_{\text{PS}}$. The propensity score distance is modeled simply as the logit of the probability of treatment as explained by the observed covariates:

$$PS(T|\beta, X) = \text{logit}P(T = t|\beta, X) = X'\beta$$

Assessing balance: We bootstrap the following univariate and multivariate summaries of the matched data to quantify similarity:

- Standardized bias, typically required to at least be below 0.2:

$$B_p = \frac{x^{\text{Treated}}_{p,\text{Matched}} - x^{\text{Control}}_{p,\text{Matched}}}{\sigma^{\text{Treated}}_{p,\text{Pre-match}}}, \quad B_p^{\text{Max}} = \arg\max_p B, \quad B_p^{\text{Mean}} = \frac{1}{p}\sum_i B_i$$

- Variance/Covariance:

$$\Sigma = \begin{cases} \sigma_p^2 & \text{if } p = p'; \\ \rho_{pp'} & \text{if } p \neq p'. \end{cases}$$

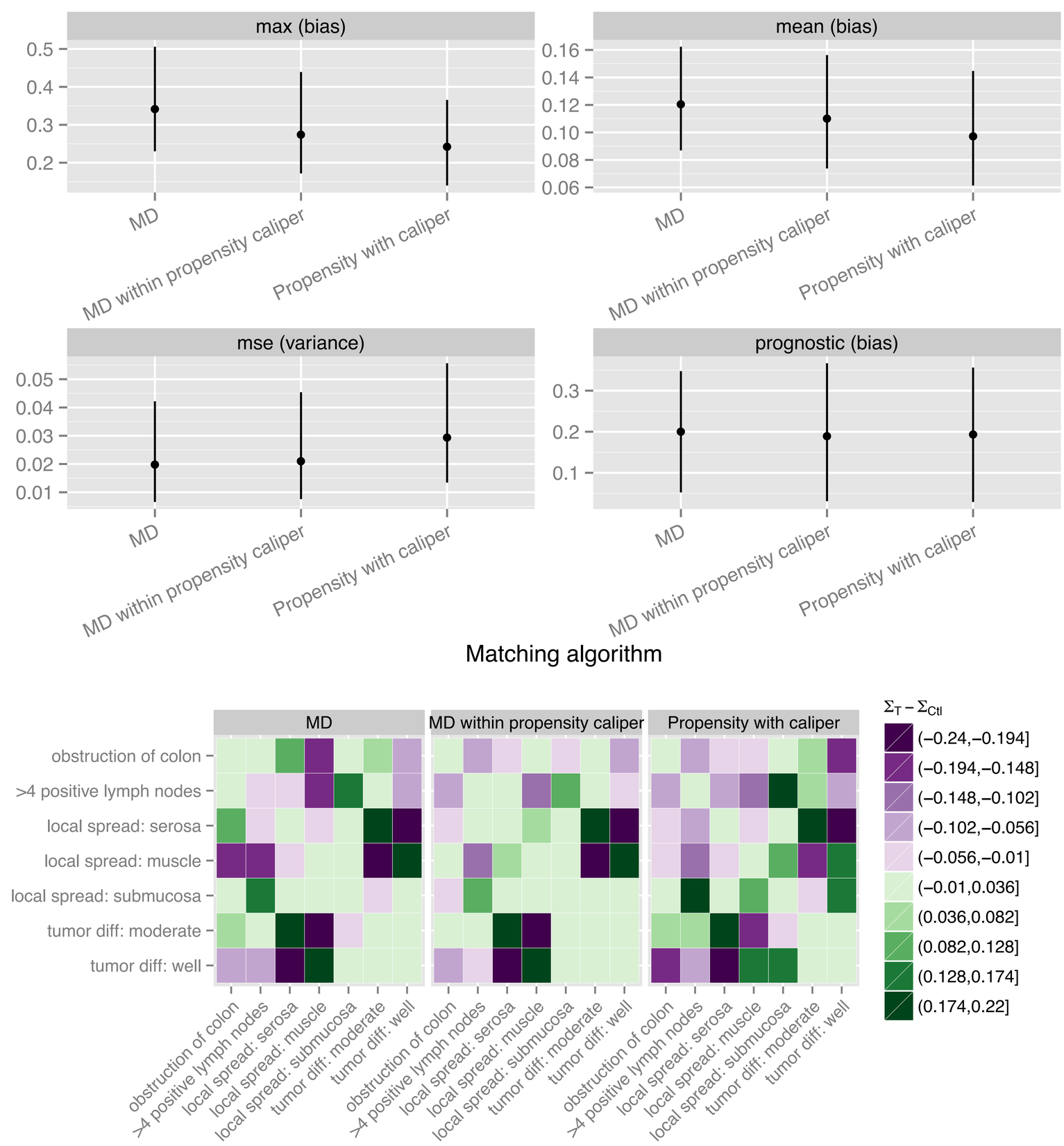
$$MSE^{\text{Cov}} = \frac{1}{p^2}\sum\left(\Sigma^{\text{Treated}}_{\text{Matched}} - \Sigma^{\text{Control}}_{\text{Matched}}\right)^2$$

Data

Variable	Level	Control % or mean (sd)	Low exposure (Q1) % or mean (sd)	High exposure (Q2-4) % or mean (sd)	Prognostic p-value
Obstruction of colon by tumor	no	79.68	78.67	82.96	0.09
	yes	20.32	21.33	17.04	
Differentiation of tumor	well	11.94	4	11.66	0.09
	moderate	73.87	58.67	76.68	
	poor	14.19	37.33	11.66	
Extent of local spread	submucosa	0.97	1.33	4.04	< 0.01
	muscle	11.61	4	12.56	
	serosa	83.55	88	80.72	
	contiguous structures	3.87	6.67	2.69	
More than 4 positive lymph nodes	no	71.29	36	86.55	< 0.01
	yes	28.71	64	13.45	
Sex	F	42.9	52	54.71	0.13
	M	57.1	48	45.29	
Age (years)		60.1 (11.6)	52 (13)	62.3 (10.9)	0.18
Perforation of colon	no	96.77	98.67	96.86	0.83
	yes	3.23	1.33	3.14	
Adherence to nearby organs	no	84.19	88	86.55	0.33
	yes	15.81	12	13.45	
Time from surgery to registration	short	74.19	76	73.99	0.18
	long	25.81	24	26.01	

Table 1. Covariate distribution across the exposure grouping: For control, quartile 1, and quartiles 2-4 of exposure, the percent or mean and standard deviation of each covariate level is shown. The prognostic p-value is taken from the LRT of the Cox PH model with the covariate against the null model. Significant covariates are used to create the "prognostic" summary covariate for all patients, and is used to assess quality of the match on those covariates likely to be most predictive of outcome.

Results



Figures 1 and 2. Figure 1 shows summaries of balance after matching (90% bootstrap CI): maximum standardized bias over the matched covariates, mean standardized bias over the matched covariates, and the prognostic score standardized bias. Figure 2 uses a heat map to summarize the bootstrapped median cell by cell deviations between the correlation matrices (with variance on the diagonal) of the treatment group and matched control.

Conclusion

With a simple PS model (including only main effects of the balancing covariates), the PS match outperforms MD regarding bias minimization but underperforms when evaluated on preservation of multivariate structure of the treated group. Heat maps prove to be useful for model building: Figure 2 implies that the interaction between local spread and tumor differentiation, and the interaction between positive lymph nodes and local spread should be matched upon in addition to the main effects. Prognostic scores are comparable between methods in this exercise, but provide reassurance when matching on large sets of covariates that the most directly predictive of outcome are being adequately handled.
Due to the general familiarity of logistic regression, the strong standardized bias performance, and the clear interpretation of how categorical variables are adjusted for, we recommend use of propensity matching or MD within PS calipers. Inspection and model building to preserve the multivariate structure is strongly recommended when using PS. Overlooking this may be safe with MD, but PS matching appears to require this extra scrutiny.

References

[1] Paul R. Rosenbaum and Donald B. Rubin. The central role of the propensity score in observational studies for causal effects. *Biometrika*, 70(1):pp. 41–55, 1983.

[2] D B Rubin and N Thomas. Matching using estimated propensity scores: relating theory to practice. *Biometrics*, 52(1):249–64, Mar 1996.

[3] Elizabeth A Stuart. Matching methods for causal inference: A review and a look forward. *Stat Sci*, 25(1):1–21, Feb 2010.

[4] Elizabeth A Stuart, Brian K Lee, and Finbarr P Leacy. Prognostic score-based balance measures can be a useful diagnostic for propensity score methods in comparative effectiveness research. *J Clin Epidemiol*, 66(8 Suppl):S84–S90.e1, Aug 2013.

[5] Jun Yang, Hong Zhao, Christine Garnett, Atiqur Rahman, Jogarao V Gobburu, William Pierce, Genevieve Schechter, Jeffery Summers, Patricia Keegan, Brian Booth, and Yaning Wang. The combination of exposure-response and case-control analyses in regulatory decision making. *J Clin Pharmacol*, 53(2):160–6, Feb 2013.