Finding Common Support Using Largest Connected Components

Michael Higgins and Sharif Mahmood
Department of Statistics, Kansas State University

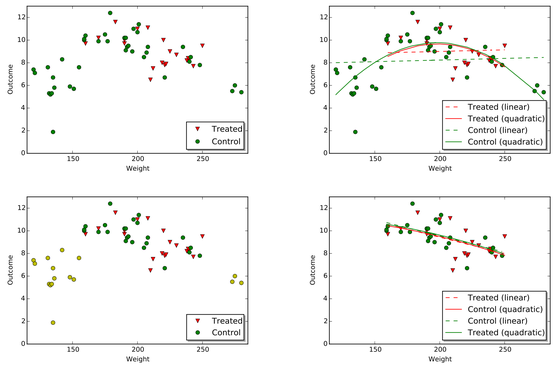
Summary

Treatment effects are only reliably estimated for a subpopulation under which a common support assumption holds—one in which treatment and control covariate spaces overlap. Given a dissimilarity measure between units, we use techniques in graph theory to find common support. We construct an adjacency graph where edges are drawn between similar treated and control units. We then determine regions of common support by finding the largest connected components (LCC) of this graph. We show that LCC efficiently constructs (possibly non-convex) regions that preserve clustering in the data while ensuring interpretability of the region through the distance metric.

Importance of Common Support

Identifying common support has the following advantages [Ho et. al., 2007, King and Zeng, 2006]:

- Analysis is more robust to choice of model
- Better identification of the study population
- Reduction of covariate imbalances between treated and control groups



Current methods for finding common support may struggle in high-dimensional covariate spaces, require considerable computational resources, yield hard-to-interpret regions of support, and/or may require the formation of convex regions of support [Fogarty et. al., 2015, King and Zeng, 2006] [Zubizarreta et. al., 2014]. The method we propose ensures the efficient construction of interpretable and flexible regions of common support.

Common Support Setup

We follow the framework proposed by Fogarty et. al. (2015). Suppose our study contains n_1 treated units and n_0 control units, where $n = n_1 + n_0$. For each unit i, let $\boldsymbol{x}_i = (x_{i1}, \ldots, x_{iK})$ denote a vector of its covariates. A dissimilarity measure $D(\boldsymbol{x}_i, \boldsymbol{x}_{i'})$ is computed between every two units. Intuitively the dissimilarity is small when covariates between i and i' are similar.

Examples of the dissimilarity measure $D(\boldsymbol{x}_i, \boldsymbol{x}_{i'})$ include differences between propensity scores, Mahalanobis distances, Euclidean distances, etc. We advocate the use of a distance measure of the form:

$$D^{\infty}(\boldsymbol{x}_i, \boldsymbol{x}_{i'}) = \max_j \frac{|x_{ij} - x_{i'j}|}{c_j}$$

where c_j is a researcher-selected parameter for how much imbalance on covariate j is tolerable for a match.

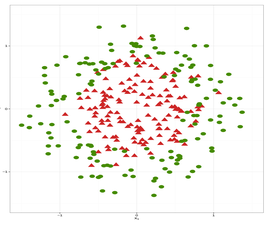
The researcher sets a threshold ω for the amount of acceptable imbalance between two units in the study. That is, treated unit i has an acceptable match if there is a control unit i' such that $D(\boldsymbol{x}_i, \boldsymbol{x}_{i'}) \leq \omega$ and vice versa. For our recommended dissimilarity measure, acceptable matches have $D^{\infty}(\boldsymbol{x}_i, \boldsymbol{x}_{i'}) \leq 1$.

Additionally, the imbalance threshold may be chosen to satisfy certain properties—for example, to minimize mean differences on covariates between treatment and control groups. For a given dissimilarity measure, there are at most $n(n-1)/2$ different sets of acceptable matches corresponding to different values of ω. Hence, a brute force search through these sets will add an $O(n^2)$ complexity to our algorithm.

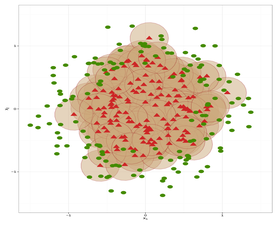
The Algorithm

Our algorithm to find the common support as follows:

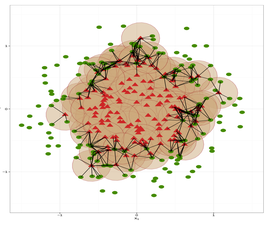
1) View the treated (▲) and control (●) units as vertices in a graph.



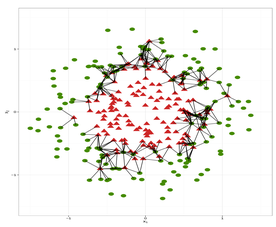
2) Find the acceptable matches (within ω) for the treated units.



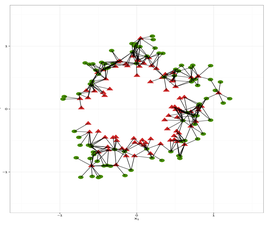
3) Draw edges between treated and control units that are acceptable matches.



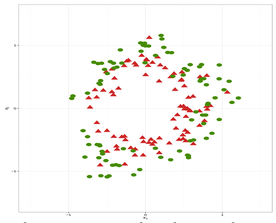
4) Connected components (CC) are subgraphs where, between any two units in a CC, there is a path of edges joining the two units. Find the CC with the greatest number of treated units. This is the LCC.



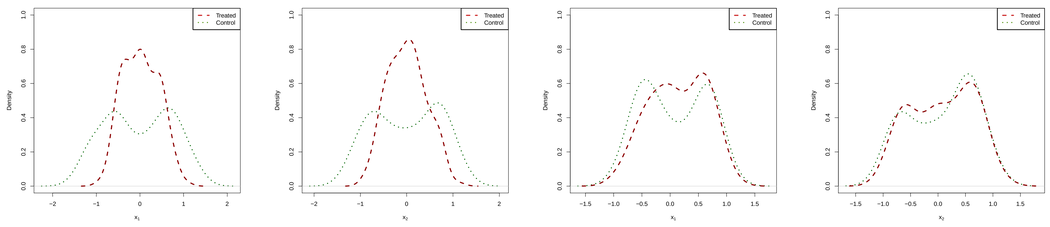
5) Discard all the units that are not in LCC.



6) These units form an interpretable study population.



In this case, our common support algorithm greatly reduces differences in the covariate densities between treated and control groups.



(a) (b) (c) (d)

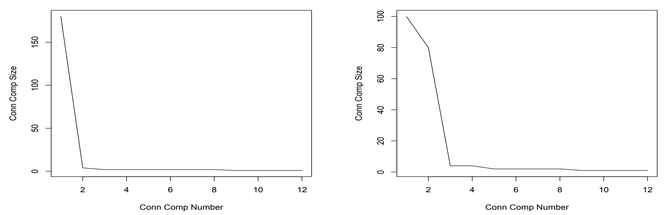
Figure 1: Graphs (1a) and (2b) give the densities of the covariates under the original sample and (1c) and (1d) give the densities under common support.

Clustered Data

In the case of clustered data, multiple large connected components may be preferable.

Plots of the sizes of the largest connected components—similar to skree plots—can be used to denote whether to include multiple connected components.

Differences in sizes may help suggest the number of connected components to include in the common support.



(a) (b)

Figure 2: Left graph: CC size decreases dramatically after the first connected component, suggesting one CC in the common support. Right graph: Drastic difference in size after the second connected component, suggesting two CCs.

Advantages

- Simulations suggest that this method for finding common support successfully reduces model dependency, reduces imbalances between treatment and control groups, and improves estimates of treatment effects.
- This method yields an interpretable region. The LCC is the largest cluster of data that have comparable matches. Interpretability can be aided through the dissimilarity measure, for example, D^{∞}.
- Creating the graphs in Step 3 and finding connected components in Step 4 can be done very efficiently [Cormen et. al. , 2001, Higgins et al., 2015], leading to a low runtime of the algorithm.
- The algorithm allows for formation of common support regions that are non-convex. In our example, our algorithm successfully forms an annulus, a shape that is not possible under many common support methods.

Lalonde Data

- *The National Supported Work (NSW) Program* was a U.S. federally and privately funded experiment that aimed to provide work experience for the people who need a job [LaLonde 1986, Dehejia and Wahba, 1999].
- The benchmark estimate for the regression treatment effect \$1672. The goal of this study is to use the same treated units and observational data for controls to obtain this benchmark.

	Lalonde Data		Under Common Support	
Variable	Treatment	Control	Treatment	Control
Age	25.816	34.851	25.816	32.320
Years of schooling	10.346	12.117	10.346	11.180
Proportion of blacks	0.843	0.251	0.843	0.745
Proportion of Hispanics	0.059	0.033	0.059	0.025
Proportion married	0.189	0.866	0.189	0.642
Real earnings in 1974	2095.574	19428.746	2095.574	14675.844
Real earnings in 1975	1532.056	19063.338	1532.056	13897.591

Table 1: The table shows the average of pretreatment variables of PSID samples which includes 185 treated and 2490 control units. Under common support we select 185 treated and 752 control units.

	Lalonde Data				Under Common Support			
Method	Estimate	Std. Error	t value		Estimate	Std. Error	t value	
MLR	860	907.57	0.95		1,668	893.79	1.87*	

* : p-value < .05

Table 2: Training effect estimate under multiple linear regression (MLR).

References

Cormen, T., Leiserson, C, Rivest, R., and Stein, C. (2001). *Introduction to algorithms*, MIT press Cambridge.

Dehejia, R. H., and Wahba, S. (1999).
Causal Effects in Nonexperimental Studies: Reevaluating the Evaluation of Training Programs. *Journal of the Amer. Stat. Asso.*, 448, 1053-062.

Fogarty, Colin B, Mikkelsen, M., Gaieski, D., and Small, D. (2015).
Discrete optimization for interpretable study populations and randomization inference in an observational study of severe sepsis mortality *Journal of the Amer. Stat. Asso.*, Accepted.

Higgins, M.J. and Savje, F. and Sekhon, J. (2015).
Improving massive experiments with threshold blocking. *PNAS*, Forthcoming.

Ho, D., Imai, K., King, G. and Stuart, E. (2007).
Matching as Nonparametric Preprocessing for Reducing Model Dependence in Parametric Causal Inference. *Political Analysis*, 15, 199-236.

King, G. and Zeng, L., (2006)
The Dangers of Extreme Counterfactuals *Political Analysis*, 14(2), 131-159.

LaLonde, R. (1986)
Evaluating the econometric evaluations of training programs with experimental data *American Economic Review*, 604-620.

Zubizarreta, J., Paredes, R., and Rosenbaum, P., (2014)
Matching for balance, pairing for heterogeneity in an observational study of the effectiveness of for-profit and not-for-profit high schools in Chile *The Annals of Applied Statistics*, 8(1), 204-231.