




1 jointVIP: Prioritizing variables in observational study 2 design with joint variable importance plot in R

3 **Lauren D. Liao** ¹ ¶ and **Samuel D. Pimentel** ²

4 **1** Division of Biostatistics, University of California, Berkeley, USA **2** Department of Statistics, University
5 of California, Berkeley, USA ¶ Corresponding author

DOI: [10.xxxxxx/draft](https://doi.org/10.xxxxxx/draft)

Software

- [Review](#) 
- [Repository](#) 
- [Archive](#) 

Editor: [Andrew Stewart](#)  

Reviewers:

- [@nhejazi](#)
- [@jackmwolf](#)

Submitted: 11 August 2023

Published: unpublished

License

Authors of papers retain copyright
and release the work under a
Creative Commons Attribution 4.0
International License ([CC BY 4.0](#)).

6 Summary

7 Credible causal effect estimation requires treated subjects and controls to be otherwise similar.
8 In observational settings, such as analysis of electronic health records, this is not guaranteed.
9 Investigators must balance background variables so they are similar in treated and control groups.
10 Common approaches include matching (grouping individuals into small homogeneous sets)
11 or weighting (upweighting or downweighting individuals) to create similar profiles. However,
12 creating identical distributions may be impossible if many variables are measured, and not
13 all variables are of equal importance to the outcome. The joint variable importance plot
14 (jointVIP) package to guides decisions about which variables to prioritize for adjustment by
15 quantifying and visualizing each variable's relationship to both treatment and outcome.

16 Statement of need

17 Consider an observational study to measure the effect of a binary treatment variable
18 (treated/control) on an outcome, in which additional covariates (background variables) are
19 measured. A covariate may be associated with outcomes, and it may also differ in distribution
20 between treated and controls; if the covariate is associated with treatment and outcome, the
21 covariate in question is a confounder. Ignored confounders introduce bias into treatment effect
22 estimates. For instance, when testing a blood pressure drug, if older patients both take the
23 drug more and have worse initial blood pressure, a simple difference in mean blood pressure
24 between treated and control subjects will understate the drug's benefits. Confounding can be
25 addressed by matching, under which blood pressure is compared only within pairs of patients
26 with similar ages, or by weighting, in which older control subjects receive larger weights than
27 younger control subjects when averaging blood pressure. When many potential confounders
28 are measured, however, neither matching nor weighting can perfectly adjust for all differences,
29 and researchers must select which variables to focus on balancing.

30 Current practice for selecting variables for adjustment focuses primarily on understanding the
31 treatment relationship, via tools such as balance tables and the Love plot ([Ahmed et al., 2006](#);
32 [Greifer & Stuart, 2021](#); [Ben B. Hansen & Bowers, 2008](#); [Rosenbaum & Rubin, 1985](#); [Stuart et al., 2011](#)).
33 A key metric is the standardized mean difference (SMD), or the difference in treated
34 and control means over a covariate measure in standard deviations. Researchers commonly try
35 to adjust so that all SMD values are moderately small, or focus on adjustments for variables
36 with the largest initial SMD. However, these approaches neglect important information about
37 the relationship of each covariate with the outcome variable, which substantially influences the
38 degree of bias incurred by ignoring it.

39 To improve observational study design, we propose the joint variable importance plot (jointVIP)
40 ([Liao et al., 2023](#)), implemented in the jointVIP package. The jointVIP represents both
41 treatment and outcome relationships for each variable in a single image: each variable's SMD

42 is plotted against an outcome correlation measure (computed in a pilot control sample to avoid
43 bias from multiple use of outcome data). Bias curves based on unadjusted, simple one-variable
44 omitted variable bias models are plotted to improve variable comparison. The jointVIP provides
45 valuable insight into variable importance and can be used to specify key parameters in existing
46 matching and weighting methods.

47 Development

48 The jointVIP package was created in the R programming language (R Core Team, 2020).
49 The package uses the S3 object system and leverages system generic functions, `print()`,
50 `summary()`, and `plot()`. Plotting the jointVIP object outputs a plot of the ggplot2 class.
51 An interactive R Shiny application, available online at <https://ldliao.shinyapps.io/jointVIP/>,
52 showcases the package.

53 Usage

54 The jointVIP package is available from the Comprehensive R Archive Network [CRAN](#) and
55 [GitHub](#).

```
# installation using CRAN:  
# install.packages("jointVIP")  
  
# installation using GitHub  
# remotes::install_github('ldliao/jointVIP')
```

```
library(jointVIP)
```

56 To create an object of the jointVIP class, the user needs to supply two datasets and specify
57 the treatment, outcome, and background variable names. Two processed datasets, “pilot” and
58 “analysis” samples, are in the form of data.frames. The analysis sample contains both treated
59 and control groups. The pilot sample contains only control individuals, and they are excluded
60 from the subsequent analysis stage. The treatment variable must be binary: 0 specified for the
61 control group and 1 specified for the treated group. Background variables are measured before
62 both treatment and outcome. The outcome of interest can be either binary or continuous.

63 We demonstrate the utility of this package to investigate the effect of a job training program
64 on earnings (Dehejia & Wahba, 1999; Huntington-Klein & Barrett, 2021; LaLonde, 1986). The
65 treatment is whether the individual is selected for the job training program. The outcome is
66 earnings in 1978. Covariates are age, education, race/ethnicity, and previous earnings in 1974
67 and 1975. After [preprocessing both dataset and log-transforming the earnings](#), we use the
68 `create_jointVIP()` function to create a jointVIP object stored as `new_jointVIP`.

```
# first define and get pilot_df and analysis_df  
# they should both be data.frame objects  
  
treatment <- "treat"  
outcome <- "log_re78"  
covariates <- c("age", "educ", "black",  
               "hispanic", "marr", "nodegree",  
               "log_re74", "log_re75")  
  
new_jointVIP = create_jointVIP(treatment = treatment,  
                              outcome = outcome,  
                              covariates = covariates,
```

```
pilot_df = pilot_df,
analysis_df = analysis_df)
```

69 The `plot()` function displays a jointVIP (Figure 1). The x-axis describes treatment imbalance in SMD (computed with a denominator based on the pilot sample as in (Liao et al., 2023)).
 70 in SMD (computed with a denominator based on the pilot sample as in (Liao et al., 2023)).
 71 The y-axis describes outcome correlations in the pilot sample. The `summary()` function outputs
 72 the maximum absolute bias and the number of variables required for adjustment above the
 73 absolute bias tolerance, `bias_tol`. The `bias_tol` parameter can be used in the `print()` function
 74 to see which variables are above the desired tolerance. Additional tuning parameters can be
 75 specified in these functions, for details and examples, see [the additional options vignette](#).

```
plot(new_jointVIP)
```

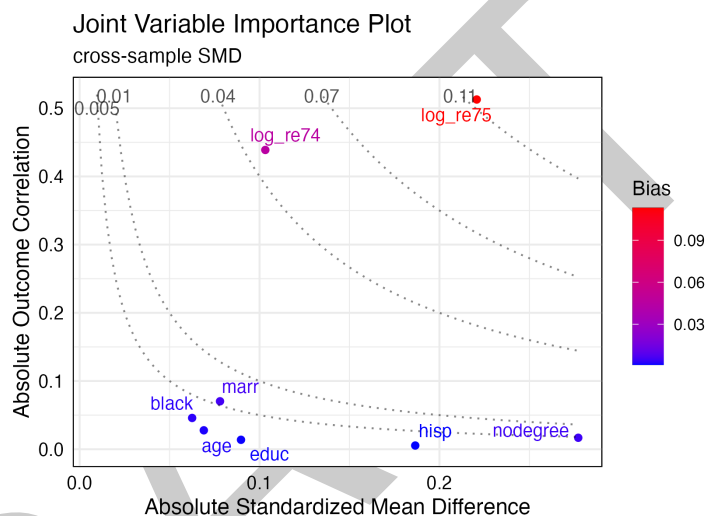


Figure 1: Joint variable importance plot example.

```
summary(new_jointVIP,
smd = "cross-sample",
use_abs = TRUE,
bias_tol = 0.01)
# > Max absolute bias is 0.113
# > 2 variables are above the desired 0.01 absolute bias tolerance
# > 8 variables can be plotted

print(new_jointVIP,
smd = "cross-sample",
use_abs = TRUE,
bias_tol = 0.01)

# > bias
# > log_re75 0.113
# > log_re74 0.045
```

76 To interpret our working example, the most important variables are the previous earning
 77 variables in 1975 and 1974, `log_re75` and `log_re74` variables, respectively. Using the traditional
 78 visualization method, the Love plot, would only identify variables based on the SMD. The same
 79 information can be interpreted from the x-axis of the jointVIP. For example, the Love plot would
 80 indicate variables, `nodegree` and `hisp`, to be more important for adjustment than `log_re74`.
 81 In comparison, those variables, `nodegree` and `hisp`, show low bias using the jointVIP.

82 After adjusting for variables, for example, using optimal matching (Ben B. Hansen &

83 [Klopfer, 2006](#); [Stuart et al., 2011](#)) to select pairs for analysis, a post-adjustment dataset,
84 `post_analysis_df`, can be used to create a post adjustment object of class `post_jointVIP`. The
85 `create_post_jointVIP()` function can be used to visualize and summarize the post adjustment
86 results, as seen in Figure 2. The functions: `summary()`, `print()`, and `plot()` all can take in
87 the `post_jointVIP` object and provide comparison between original and post adjusted jointVIPs.

```
post_optmatch_jointVIP <- create_post_jointVIP(new_jointVIP,  
                                              post_analysis_df = optmatch_df)
```

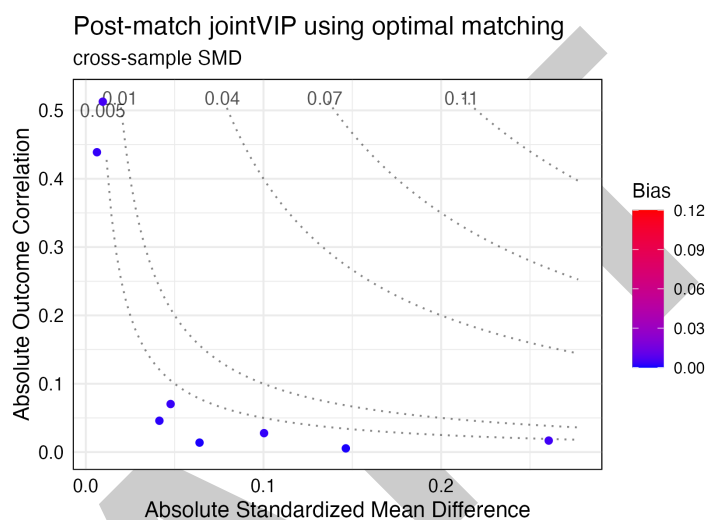


Figure 2: Post match example showing balanced sample based on new mean differences.

```
summary(post_optmatch_jointVIP)
# > Max absolute bias is 0.113
# > 2 variables are above the desired 0.01 absolute bias tolerance
# > 8 variables can be plotted
# >
# > Max absolute post-bias is 0.005
# > Post-measure has 0 variable(s) above the desired 0.005 absolute bias tolerance

print(post_optmatch_jointVIP)
# >      bias post_bias
# > log_re75 0.113    0.005
# > log_re74 0.045    0.003
```

88 Discussion

89 We have developed user-friendly software to prioritize variables for adjustment in observational
90 studies. This package can help identify important variables related to both treatment and
91 outcome. One limitation is that each background variable is individually evaluated for bias.
92 Thus, conditional relationships, interactions, or higher moments of variables need to be carefully
93 considered or preprocessed by the user.

94 Acknowledgements

95 The authors thank Emily Z. Wang for helpful comments. SDP is supported by Hellman Family
96 Fellowship and by the National Science Foundation (grant 2142146). LDL is supported by
97 National Science Foundation Graduate Research Fellowship (grant DGE 2146752).

98 **References**

- 99 Ahmed, A., Husain, A., Love, T. E., Gambassi, G., Dell'Italia, L. J., Francis, G. S., Gheorghiade,
100 M., Allman, R. M., Meleth, S., & Bourge, R. C. (2006). Heart failure, chronic diuretic
101 use, and increase in mortality and hospitalization: An observational study using propensity
102 score methods. *European Heart Journal*, 27(12), 1431–1439.
- 103 Dehejia, R. H., & Wahba, S. (1999). Causal effects in nonexperimental studies: Reevaluating
104 the evaluation of training programs. *Journal of the American Statistical Association*,
105 94(448), 1053–1062.
- 106 Greifer, N., & Stuart, E. A. (2021). Choosing the estimand when matching or weighting in
107 observational studies. *arXiv Preprint arXiv:2106.10577*.
- 108 Hansen, Ben B., & Bowers, J. (2008). Covariate balance in simple, stratified and clustered
109 comparative studies. *Statistical Science*, 219–236.
- 110 Hansen, Ben B., & Klopfer, S. O. (2006). Optimal full matching and related designs via
111 network flows. *Journal of Computational and Graphical Statistics*, 15(3), 609–627.
- 112 Huntington-Klein, N., & Barrett, M. (2021). *Causaldata: Example data sets for causal
113 inference textbooks*. <https://CRAN.R-project.org/package=causaldata>
- 114 LaLonde, R. J. (1986). Evaluating the econometric evaluations of training programs with
115 experimental data. *The American Economic Review*, 604–620.
- 116 Liao, L. D., Zhu, Y., Ngo, A. L., Chehab, R. F., & Pimentel, S. D. (2023). Using joint variable
117 importance plots to prioritize variables in assessing the impact of glyburide on adverse birth
118 outcomes. *arXiv Preprint arXiv:2301.09754*.
- 119 R Core Team. (2020). *R: A language and environment for statistical computing*. R Foundation
120 for Statistical Computing. <https://www.R-project.org/>
- 121 Rosenbaum, P. R., & Rubin, D. B. (1985). Constructing a control group using multivariate
122 matched sampling methods that incorporate the propensity score. *The American Statistician*,
123 39(1), 33–38.
- 124 Stuart, E. A., King, G., Imai, K., & Ho, D. (2011). MatchIt: Nonparametric preprocessing for
125 parametric causal inference. *Journal of Statistical Software*.