The Journal of Open Source Software

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

Lauren D. Liao [1¶] **and Samuel D. Pimentel** [2]

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

DOI: 10.21105/joss.06093

Software

- Review
- Repository
- Archive

Editor: Andrew Stewart

Reviewers:

- @nhejazi
- @jackmwolf
- @JerryChiaRuiChang

Submitted: 11 August 2023
Published: 01 November 2024

Summary

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

Statement of need

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

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

To improve observational study design, we propose the joint variable importance plot (jointVIP) (Liao et al., 2024), implemented in the `jointVIP` package. The jointVIP represents both treatment and outcome relationships for each variable in a single image: each variable's SMD

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

Development

The `jointVIP` package was created in the R programming language (R Core Team, 2020). The package creates a new S3 class called "jointvip" and uses S3 generic to dispatch `print()`, `summary()`, and `plot()`. Plotting the jointVIP object outputs a plot of the `ggplot2` class. An interactive R Shiny application, available online at https://ldliao.shinyapps.io/jointVIP/, showcases the package.

Usage

The `jointVIP` package is available from the Comprehensive R Archive Network CRAN and GitHub.

```
# installation using CRAN:
# install.packages("jointVIP")

# installation using GitHub
# remotes::install_github('ldliao/jointVIP')

library(jointVIP)
```

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

We demonstrate the utility of this package to investigate the effect of a job training program on earnings (Dehejia & Wahba, 1999; Huntington-Klein & Barrett, 2021; LaLonde, 1986). The treatment is whether the individual is selected for the job training program. The outcome is earnings in 1978. Covariates are age, education, race/ethnicity, and previous earnings in 1974 and 1975. After preprocessing both dataset and log-transforming the earnings,, we use the `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",
                "hisp", "marr", "nodegree",
                "log_re74", "log_re75")

new_jointVIP = create_jointVIP(treatment = treatment,
                               outcome = outcome,
                               covariates = covariates,
```

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

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., 2024)). The y-axis describes outcome correlations in the pilot sample. The `summary()` function outputs the maximum absolute bias and the number of variables required for adjustment above the absolute bias tolerance, *bias_tol*. The *bias_tol* parameter can be used in the `print()` function to see which variables are above the desired tolerance. Additional tuning parameters can be specified in these functions, for details and examples, see the additional options vignette.

```
plot(new_jointVIP)
```

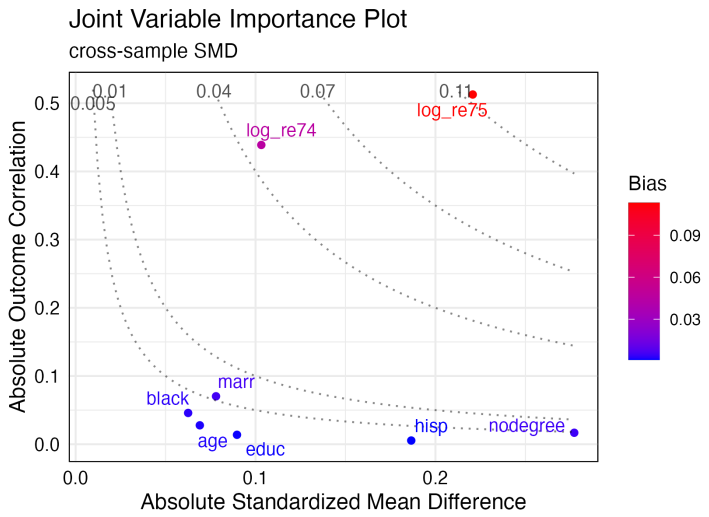

Figure 1: Joint variable importance plot example.

```
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
```

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

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

Klopfer, 2006; Stuart et al., 2011) to select pairs for analysis, a post-adjustment dataset, *post_analysis_df*, can be used to create a post adjustment object of class `post_jointVIP`. The `create_post_jointVIP()` function can be used to visualize and summarize the post adjustment results, as seen in Figure 2. The functions: `summary()`, `print()`, and `plot()` all can take in 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)
```

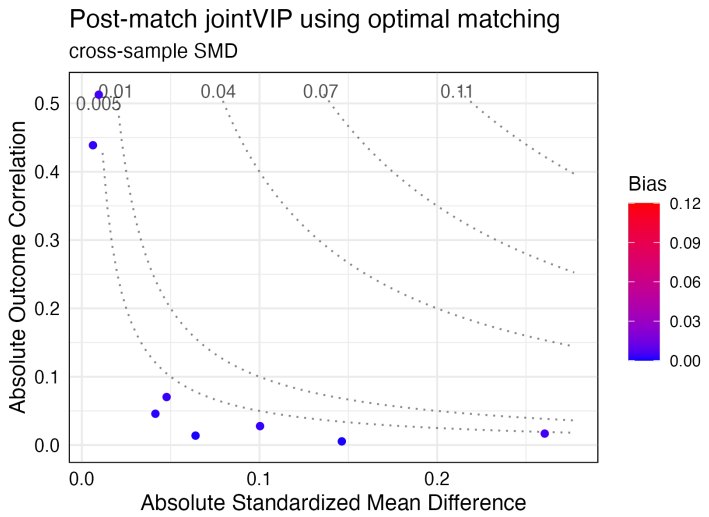

Figure 2: Post match example showing balanced sample based on new mean differences.

```
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
```

Discussion

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

Acknowledgements

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

References

Ahmed, A., Husain, A., Love, T. E., Gambassi, G., Dell'Italia, L. J., Francis, G. S., Gheorghiade, M., Allman, R. M., Meleth, S., & Bourge, R. C. (2006). Heart failure, chronic diuretic use, and increase in mortality and hospitalization: An observational study using propensity score methods. *European Heart Journal*, *27*(12), 1431–1439. https://doi.org/10.1093/eurheartj/ehi890

Dehejia, R. H., & Wahba, S. (1999). Causal effects in nonexperimental studies: Reevaluating the evaluation of training programs. *Journal of the American Statistical Association*, *94*(448), 1053–1062. https://doi.org/10.1080/01621459.1999.10473858

Greifer, N., & Stuart, E. A. (2021). Choosing the estimand when matching or weighting in observational studies. *arXiv Preprint arXiv:2106.10577*.

Hansen, Ben B., & Bowers, J. (2008). Covariate balance in simple, stratified and clustered comparative studies. *Statistical Science*, 219–236. https://doi.org/10.1214/08-sts254

Hansen, Ben B., & Klopfer, S. O. (2006). Optimal full matching and related designs via network flows. *Journal of Computational and Graphical Statistics*, *15*(3), 609–627. https://doi.org/10.1198/106186006x137047

Huntington-Klein, N., & Barrett, M. (2021). *Causaldata: Example data sets for causal inference textbooks*. https://doi.org/10.32614/cran.package.causaldata

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

Liao, L. D., Zhu, Y., Ngo, A. L., Chehab, R. F., & Pimentel, S. D. (2024). Prioritizing variables for observational study design using the joint variable importance plot. *The American Statistician*, 1–9. https://doi.org/10.1080/00031305.2024.2303419

R Core Team. (2020). *R: A language and environment for statistical computing*. R Foundation for Statistical Computing. https://www.R-project.org/

Rosenbaum, P. R., & Rubin, D. B. (1985). Constructing a control group using multivariate matched sampling methods that incorporate the propensity score. *The American Statistician*, *39*(1), 33–38. https://doi.org/10.1017/cbo9780511810725.019

Stuart, E. A., King, G., Imai, K., & Ho, D. (2011). MatchIt: Nonparametric preprocessing for parametric causal inference. *Journal of Statistical Software*. https://doi.org/10.18637/jss.v042.i08