corrselect: Flexible and exact variable subset selection based on correlation and association matrices

2025-09-09

Summary

`corrselect` (Colling 2025) is an R package that selects variable subsets whose pairwise correlations or associations do not exceed a user-defined threshold. Instead of returning a single heuristic solution, it enumerates all maximal admissible subsets, supporting transparent and reproducible preprocessing when redundancy reduces interpretability or stability.

The package handles both numeric and mixed-type data. For correlation-based workflows, measures such as Pearson, Spearman, Kendall, and biweight midcorrelation (Langfelder and Horvath 2008) take values in $[-1, 1]$. For association-based workflows, measures are normalized to $[0, 1]$ for consistent thresholding, including distance correlation (Székely et al. 2007; Székely and Rizzo 2009), the maximal information coefficient (Reshef et al. 2011), ANOVA η^2, and Cramér's V.

Statement of Need

Collinearity among predictors is widespread and can degrade model inference and prediction (Dormann et al. 2013). Popular utilities such as `caret::findCorrelation()` apply greedy, order-dependent filtering and return a single solution. Embedded and wrapper methods like the elastic net (Zou and Hastie 2005) or recursive feature elimination (Witten et al. 2009) can be powerful but couple selection to a specific model and add opacity.

`corrselect` formulates a global admissible set problem. Given variables $X_1, \ldots, X_p$ and pairwise measures r_{ij}, the goal is to find all maximal subsets S such that

$$|r_{ij}| \leq t \quad \text{for all } i \neq j \in S,$$

with a user threshold $t \in (0, 1)$. The software supports mixed variable types, optional forced inclusion of key predictors, and exhaustive coverage of all maximal solutions.

Functionality

Two user functions cover common workflows:

- `corrSelect()` takes a precomputed correlation matrix with entries in $[-1, 1]$ and selects admissible subsets at threshold t.
- `assocSelect()` computes pairwise associations for a mixed-type data frame, maps them to $[0, 1]$, and selects admissible subsets at threshold t.

Both return a `CorrCombo` object with the list of maximal subsets and summary statistics, along with `print`, `summary`, and `as.data.frame` methods.

Internally, the package implements two exact algorithms:

- Efficient Local Search (ELS): recursive branch-and-bound that grows admissible subsets and prunes early, effective when users supply `forced_in` seeds.
- Bron–Kerbosch: classical maximal clique enumeration on the complement of the thresholded association graph (Bron and Kerbosch 1973), guaranteeing exhaustive coverage and performing well when the graph is sparse.

Both methods ensure non-redundant and exhaustive enumeration of admissible subsets.

Related Work

Heuristic correlation filters are simple and common but are order dependent and yield a single result. `corrselect` extends this space by providing exhaustive enumeration, support for mixed data, and user control via `forced_in`. Compared with embedded or wrapper selection, it is model agnostic and interpretable. The graph-theoretic basis links admissible subsets to maximal cliques and independent sets, with ELS offering a complementary search strategy.

Other approaches to feature selection include embedded methods such as the elastic net (Zou and Hastie 2005), recursive feature elimination (Witten et al. 2009), or permutation-based approaches such as Boruta. These methods can be

powerful but are tied to specific modeling frameworks, non-deterministic, and less interpretable in the presence of multicollinearity. By contrast, `corrselect` is fast, deterministic, and model agnostic, linking statistical association to well-studied optimization problems.

Applications

The approach supports bioclimatic predictor filtering, feature screening in high-dimensional settings, and exploratory mapping of alternative, equally valid predictor sets. The inclusion of biweight midcorrelation (Langfelder and Horvath 2008), distance correlation (Székely et al. 2007; Székely and Rizzo 2009), and the maximal information coefficient (Reshef et al. 2011) further extends its applicability to genomics, network analysis, and large heterogeneous datasets.

References

Bron, Coen, and Joep Kerbosch. 1973. "Algorithm 457: Finding All Cliques of an Undirected Graph." *Communications of the ACM* 16 (9): 575–77. https://doi.org/10.1145/362342.362367.

Colling, Gilles. 2025. *Corrselect: Correlation-Based Variable Subset Selection*. https://doi.org/10.32614/CRAN.package.corrselect.

Dormann, Carsten F., Jane Elith, Sven Bacher, and et al. 2013. "Collinearity: A Review of Methods to Deal with It and a Simulation Study Evaluating Their Performance." *Ecography* 36 (1): 27–46. https://doi.org/10.1111/j.1600-0587.2012.07348.x.

Langfelder, Peter, and Steve Horvath. 2008. "WGCNA: An r Package for Weighted Correlation Network Analysis." *BMC Bioinformatics* 9 (1): 559. https://doi.org/10.1186/1471-2105-9-559.

Reshef, David N., Yakir A. Reshef, Hilary K. Finucane, and et al. 2011. "Detecting Novel Associations in Large Data Sets." *Science* 334 (6062): 1518–24. https://doi.org/10.1126/science.1205438.

Székely, Gábor J., and Maria L. Rizzo. 2009. "Brownian Distance Covariance." *The Annals of Applied Statistics* 3 (4): 1236–65. https://doi.org/10.1214/09-AOAS312.

Székely, Gábor J., Maria L. Rizzo, and N. K. Bakirov. 2007. "Measuring and Testing Dependence by Correlation of Distances." *The Annals of Statistics* 35 (6): 2769–94. https://doi.org/10.1214/009053607000000505.

Witten, Daniela M., Robert Tibshirani, and Trevor Hastie. 2009. "A Penalized Matrix Decomposition, with Applications to Sparse Principal Components and Canonical Correlation Analysis." *Biostatistics* 10 (3): 515–34. https://doi.org/10.1093/biostatistics/kxp008.

Zou, Hui, and Trevor Hastie. 2005. "Regularization and Variable Selection via the Elastic Net." *Journal of the Royal Statistical Society: Series B* 67 (2): 301–20. https://doi.org/10.1111/j.1467-9868.2005.00503.x.