corrselect: Exhaustive variable subset selection based on correlation and association matrices

2025-09-09

Summary

`corrselect` (Colling 2025) is a model-agnostic R package for selecting 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. This allows users to select subsets before model fitting, avoiding the common problems of highly correlated or associated predictors, which inflate variance estimates, destabilize coefficient estimates, and obscure the relative importance of variables. The package also supports forced inclusion of user-specified predictors (`forced_in`), ensuring that key variables are retained while admissibility constraints govern the remainder.

The package supports both numeric and mixed-type data. Correlation-based workflows include measures such as Pearson, Spearman, Kendall, and biweight midcorrelation (Langfelder and Horvath 2008), which take values in $[-1, 1]$. Association-based workflows use measures 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 common in applied modeling and can degrade 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 reduce transparency.

`corrselect` instead 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

Three core functions implement the main subset selection tasks:

- `corrSelect()` takes a numeric data frame, computes pairwise correlations in $[-1, 1]$, and selects admissible subsets at threshold t.
- `assocSelect()` handles mixed-type data, computes normalized association measures in $[0, 1]$, and selects admissible subsets at threshold t.
- `MatSelect()` identifies all maximal subsets of variables from a symmetric matrix (typically a correlation or association matrix) such that all pairwise absolute values are below a specified threshold.

All return a `CorrCombo` object containing maximal subsets, summary statistics, and standard methods (`print`, `summary`, `as.data.frame`). For example, given a data frame `df` in wide format (variables in columns, observations in rows), `corrSelect(df, t = 0.7)` returns all maximal subsets of numeric variables whose pairwise correlations are below 0.7. The function `assocSelect(df, t = 0.7)` generalizes this to mixed-type variables (numeric, binary, or categorical) using normalized association measures.

To apply the selected subsets to the original data, `corrSubset()` uses a `CorrCombo` object together with the input data frame `df` to return one or more filtered data frames. By default it returns the "best" subset, defined as the largest subset with the smallest average correlation. Other options allow selecting the nth subset, the top k subsets, or all subsets at once, with the option to retain extra columns. This makes it straightforward to continue with modeling or analysis using only the admissible variable sets.

Internally, the package implements two exact algorithms in C++ for efficient exhaustive enumeration:

- **Efficient Local Search (ELS)**: a recursive branch-and-bound algorithm that expands admissible subsets while pruning early, particularly effective when `forced_in` seeds are specified.

- **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 complete enumeration of admissible subsets.

Related Work

Heuristic correlation filters are widely used but are order dependent and return only 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. Its graph-theoretic foundation links admissible subsets to maximal cliques and independent sets, with ELS offering a complementary search strategy.

Other feature selection methods include embedded approaches such as the elastic net (Zou and Hastie 2005), recursive feature elimination (Witten et al. 2009), or permutation-based algorithms 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, formulating subset selection as a well-defined graph optimization problem.

Applications

The approach supports feature screening in high-dimensional modelling and exploratory mapping of alternative, equally valid predictor sets. With support for correlation and association measures such as 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), `corrselect` is applicable across domains including genomics, network analysis, environmental modeling, and machine learning.

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.