




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

Gilles Colling ¹

¹ Department of Botany and Biodiversity Research, University of Vienna, Austria

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

Software

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

Editor: 

Submitted: 10 September 2025

Published: unpublished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](https://creativecommons.org/licenses/by/4.0/)).

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 & 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 & 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 & 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, \dots, 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 n th 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:

- **Eppstein-Löffler-Strash (ELS)**: a near-optimal maximal clique enumeration algorithm for sparse graphs (Eppstein et al., 2010), adapted here for admissible subset selection. It is particularly effective when `forced_in` seeds are specified, since the search can be anchored around these variables, pruning the space of possible subsets more efficiently.
- **Bron-Kerbosch**: the classical maximal clique enumeration algorithm (Bron & Kerbosch, 1973), applied to the complement of the thresholded association graph so that admissible subsets correspond to maximal cliques. In practice it is often faster when enumerating all maximal subsets without seeding, while still guaranteeing exhaustive coverage.

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 & 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 & Horvath, 2008), distance correlation (Székely et al., 2007; Székely & Rizzo, 2009), and the maximal information coefficient (Reshef et

80 al., 2011), corrselect is applicable across domains including genomics, network analysis,
81 environmental modeling, and machine learning.

82 References

- 83 Bron, C., & Kerbosch, J. (1973). Algorithm 457: Finding all cliques of an undirected graph.
84 *Communications of the ACM*, 16(9), 575–577. <https://doi.org/10.1145/362342.362367>
- 85 Colling, G. (2025). *Corrselect: Correlation-based variable subset selection*. <https://doi.org/10.32614/CRAN.package.corrselect>
- 87 Dormann, C. F., Elith, J., Bacher, S., & al., et. (2013). Collinearity: A review of methods to
88 deal with it and a simulation study evaluating their performance. *Ecography*, 36(1), 27–46.
89 <https://doi.org/10.1111/j.1600-0587.2012.07348.x>
- 90 Eppstein, D., Löffler, M., & Strash, D. (2010). Listing all maximal cliques in sparse graphs
91 in near-optimal time. *Algorithms and Computation (ISAAC 2010)*, 6506, 403–414. https://doi.org/10.1007/978-3-642-17517-6_36
- 93 Langfelder, P., & Horvath, S. (2008). WGCNA: An r package for weighted correlation network
94 analysis. *BMC Bioinformatics*, 9(1), 559. <https://doi.org/10.1186/1471-2105-9-559>
- 95 Reshef, D. N., Reshef, Y. A., Finucane, H. K., & al., et. (2011). Detecting novel associations in
96 large data sets. *Science*, 334(6062), 1518–1524. <https://doi.org/10.1126/science.1205438>
- 97 Székely, G. J., & Rizzo, M. L. (2009). Brownian distance covariance. *The Annals of Applied*
98 *Statistics*, 3(4), 1236–1265. <https://doi.org/10.1214/09-AOAS312>
- 99 Székely, G. J., Rizzo, M. L., & Bakirov, N. K. (2007). Measuring and testing dependence by
100 correlation of distances. *The Annals of Statistics*, 35(6), 2769–2794. <https://doi.org/10.1214/009053607000000505>
- 102 Witten, D. M., Tibshirani, R., & Hastie, T. (2009). A penalized matrix decomposition, with
103 applications to sparse principal components and canonical correlation analysis. *Biostatistics*,
104 10(3), 515–534. <https://doi.org/10.1093/biostatistics/kxp008>
- 105 Zou, H., & Hastie, T. (2005). Regularization and variable selection via the elastic net. *Journal*
106 *of the Royal Statistical Society: Series B*, 67(2), 301–320. <https://doi.org/10.1111/j.1467-9868.2005.00503.x>
- 107