The Journal of Open Source Software

cvCovEst: Cross-validated covariance matrix estimator selection and evaluation in R

Philippe Boileau[1, 2], **Nima S. Hejazi**[1, 2], **Brian Collica**[3], **Mark J. van der Laan**[2, 3, 4], **and Sandrine Dudoit**[2, 3, 4]

1 Graduate Group in Biostatistics, University of California, Berkeley **2** Center for Computational Biology, University of California, Berkeley **3** Department of Statistics, University of California, Berkeley **4** Division of Biostatistics, School of Public Health, University of California, Berkeley

DOI: 10.21105/joss.03273

Software

- Review
- Repository
- Archive

Editor: Frederick Boehm

Reviewers:

- @Marie-PerrotDockes
- @yunanwu123

Submitted: 29 April 2021
Published: 26 July 2021

Summary

Covariance matrices play fundamental roles in myriad statistical procedures. When the observations in a dataset far outnumber the features, asymptotic theory and empirical evidence have demonstrated the sample covariance matrix to be the optimal estimator of this parameter. This assertion does not hold when the number of observations is commensurate with or smaller than the number of features. Consequently, statisticians have derived many novel covariance matrix estimators for the high-dimensional regime, often relying on additional assumptions about the parameter's structural characteristics (e.g., sparsity). While these estimators have greatly improved the ability to estimate covariance matrices in high-dimensional settings, objectively selecting the best estimator from among the many possible candidates remains a largely unaddressed challenge. The `cvCovEst` package addresses this methodological gap through its implementation of a cross-validated framework for covariance matrix estimator selection. This data-adaptive procedure's selections are asymptotically optimal under minimal assumptions – in fact, they are equivalent to the selections that would be made if given full knowledge of the true data-generating processes (i.e., an oracle selector) (van der Laan & Dudoit, 2003).

Statement of Need

When the number of observations in a dataset far exceeds the number of features, the estimator of choice for the covariance matrix is the sample covariance matrix. It is efficient under minimal regularity assumptions on the data-generating distribution. In high-dimensional regimes, however, its performance is unsatisfactory: the sample covariance matrix is highly variable, and produces estimates with diverging condition numbers and over-dispersed eigenvalues (Johnstone, 2001). Analyses employing this demonstrably poor estimator may be negatively impacted.

As high-dimensional data have become widespread, researchers have derived many novel covariance matrix estimators to remediate the sample covariance matrix's shortcomings. These estimators come in many flavors, though most are constructed by regularizing the sample covariance matrix. Comprehensive reviews are provided by Fan et al. (2016) and Pourahmadi (2013), and these estimators are implemented across a diversity of R packages: `CovTools` (Lee & You, 2019), `CVTuningCov` (Wang, 2014), and `nlshrink` (Ramprasad, 2016) to name but a few.

This variety brings with it many challenges. Identifying an "optimal" estimator from among a collection of candidates can prove a daunting task, one whose objectivity is often compromised

by the data analyst's decisions. Though data-driven approaches for selecting an optimal estimator from among estimators belonging to certain (limited) classes have been derived, the question of selecting from a diverse collection of candidate procedures remains unaddressed.

`cvCovEst` Framework

The solution provided by `cvCovEst` is a general, cross-validation-based, estimator-agnostic framework for covariance matrix estimator selection. The asymptotic optimality of selections are guaranteed under a few non-restrictive assumptions by extending the seminal work of van der Laan & Dudoit (2003), Dudoit & van der Laan (2005), and van der Vaart et al. (2006) on data-adaptive estimator selection to high-dimensional covariance matrix estimation (Boileau et al., 2021). Here, optimality is defined as choosing an estimator with an equivalent risk difference to that which would have been selected were the underlying data-generating distribution *completely known*.

The `cvCovEst` software package implements this framework for the `R` language and environment for statistical computing (R Core Team, 2021). Included is a collection of covariance matrix estimators spanning the work of many researchers (Table 1). They may be employed independently of the cross-validation procedure. `cvCovEst` also provides a variety of plotting and summary functions. These diagnostic tools allow users to gauge the algorithm's performance, diagnose issues that might arise during estimation procedures, and build intuition about the many estimators' behaviors. Additionally, users have options to increase the cross-validation method's computational efficiency via parallel computation. Parallelization relies on the suite of `future` packages (Bengtsson, 2020) by way of the `origami` package (Coyle & Hejazi, 2018).

Table 1: Covariance matrix estimators implemented as of version 1.0.0.

Estimator	Implementation	Description
Sample covariance matrix	`sampleCovEst()`	The sample covariance matrix.
Hard thresholding (Bickel & Levina, 2008b)	`thresholdingEst()`	Applies a hard thresholding operator to the entries of the sample covariance matrix.
SCAD thresholding (Fan & Li, 2001; Rothman et al., 2009)	`scadEst()`	Applies the SCAD thresholding operator to the entries of the sample covariance matrix.
Adaptive LASSO (Rothman et al., 2009)	`adaptiveLassoEst()`	Applies the adaptive LASSO thresholding operator to the entries of the sample covariance matrix.
Banding (Bickel & Levina, 2008a)	`bandingEst()`	Replaces the sample covariance matrix's off-diagonal bands by zeros.
Tapering (Cai et al., 2010)	`taperingEst()`	Tapers the sample covariance matrix's off-diagonal bands, eventually replacing them by zeros.
Optimal Linear Shrinkage (Ledoit & Wolf, 2004)	`linearShrinkLWEst()`	Asymptotically optimal shrinkage of the sample covariance matrix towards the identity.
Linear Shrinkage (Ledoit & Wolf, 2004)	`linearShrinkEst()`	Shrinkage of the sample covariance matrix towards the identity, but the shrinkage is controlled by a hyperparameter.

Estimator	Implementation	Description
Dense Linear Shrinkage (Schäfer & Strimmer, 2005)	`denseLinearShrinkEst()`	Asymptotically optimal shrinkage of the sample covariance matrix towards a dense matrix whose diagonal elements are the mean of the sample covariance matrix's diagonal and whose off-diagonal elements are the mean of the sample covariance matrix's off-diagonal elements.
Nonlinear Shrinkage (Ledoit & Wolf, 2020)	`nlShrinkLWEst()`	Analytical estimator for the nonlinear shrinkage of the sample covariance matrix.
POET (Fan et al., 2013)	`poetEst()`	An estimator based on latent variable estimation and thresholding.
Robust POET (Fan et al., 2018)	`robustPoetEst()`	A robust (and more computationally taxing) take on the POET estimator.

Examples

We briefly showcase `cvCovEst`'s functionality through a toy example and an application to single-cell transcriptomic data.

Toy Dataset Example

Multivariate normal data are simulated using a covariance matrix with a Toeplitz structure and then fed to the `cvCovEst` function. A summary of the cross-validated estimation procedure is provided via the `plot` method.

```
library(MASS)
library(cvCovEst)
set.seed(1584)

# function to generate a toeplitz matrix
toep_sim <- function(p, rho, alpha) {

    times <- seq_len(p)
    H <- abs(outer(times, times, "-")) + diag(p)
    H <- H^-(1 + alpha) * rho
    covmat <- H + diag(p) * (1 - rho)

    sign_mat <- sapply(
      times,
      function(i) {
        sapply(
          times,
          function(j) {
            (-1)^(abs(i - j))
          }
```

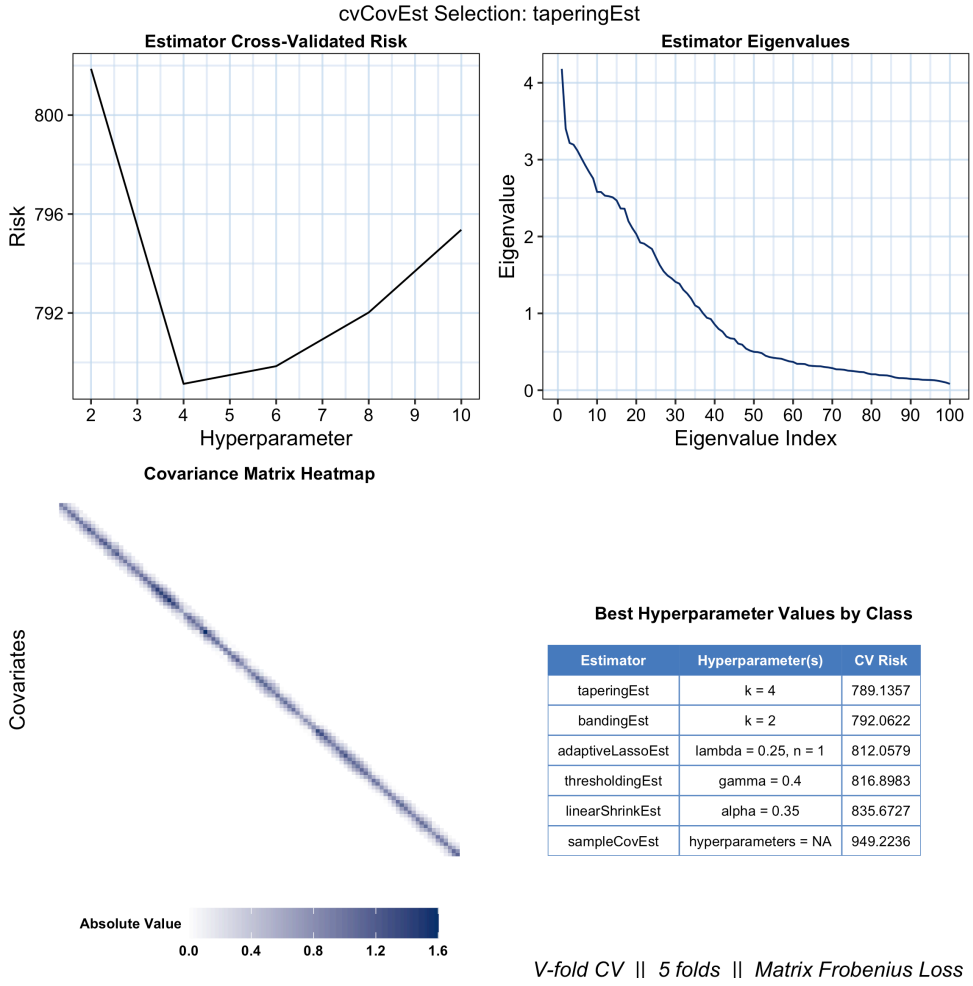
```
        )
      }
    )
    return(covmat * sign_mat)
}

# generate a 100 x 100 covariance matrix
sim_covmat <- toep_sim(p = 100, rho = 0.6, alpha = 0.3)

# sample 75 observations from multivariate normal mean = 0, var = sim_covmat
sim_dat <-  MASS::mvrnorm(n = 75, mu = rep(0, 100), Sigma = sim_covmat)

# run CV-selector
cv_cov_est_sim <- cvCovEst(
  dat = sim_dat,
  estimators = c(
    linearShrinkEst, thresholdingEst, bandingEst, adaptiveLassoEst,
    sampleCovEst, taperingEst
  ),
  estimator_params = list(
    linearShrinkEst = list(alpha = seq(0.25, 0.75, 0.05)),
    thresholdingEst = list(gamma = seq(0.25, 0.75, 0.05)),
    bandingEst = list(k = seq(2L, 10L, 2L)),
    adaptiveLassoEst = list(lambda = c(0.1, 0.25, 0.5, 0.75, 1), n = seq(1, 5)),
    taperingEst = list(k = seq(2L, 10L, 2L))
  ),
  cv_scheme = "v_fold",
  v_folds = 5
)

# plot a summary of the results
plot(cv_cov_est_sim, data_orig = sim_dat)
```



Estimator	Hyperparameter(s)	CV Risk
taperingEst	k = 4	789.1357
bandingEst	k = 2	792.0622
adaptiveLassoEst	lambda = 0.25, n = 1	812.0579
thresholdingEst	gamma = 0.4	816.8983
linearShrinkEst	alpha = 0.35	835.6727
sampleCovEst	hyperparameters = NA	949.2236

Figure 1: A summary of the `cvCovEst` procedure's results. In the top left corner, the selected estimator's risk is plotted against its considered hyperparameters. In the top right, the eigenvalues of the selected estimator's estimate are displayed. The bottom left plot presents the estimated covariance matrix. Entries are colored based on their absolute values. Finally, the table in the bottom right summarizes the performance of the best estimators from each class.

Single Cell Transcriptomic Data

Single-cell transcriptome sequencing (scRNA-seq) measures the gene expression profiles of individual cells within a given population, permitting the identification of rare cell types and the study of developmental trajectories. The datasets resulting from these experiments are typically high-dimensional: expression data for hundreds or thousands of cells are collected for tens of thousands of genes. A critical step in most analytic workflows is therefore that of dimension reduction. In addition to facilitating visualization, this reduction is thought to have a denoising effect. That is, the effects of uninteresting biological variation are typically mitigated in these lower-dimensional embeddings.

A popular method for the dimensionality reduction of scRNA-seq is uniform manifold approximation and projection (UMAP) (McInnes et al., 2018), capable of capturing non-linear relationships between features, applied to the dataset's leading principal components. Since these principal components (PCs) are derived from the sample covariance matrix, however, they are likely to be poor estimates of the true PCs when the number of genes exceeds the number of cells. Instead, the `cvCovEst` estimate could be used to compute the initial

dimensionality reduction.

Indeed, we find that the two-dimensional UMAP embedding resulting from the `cvCovEst`-based approach improves upon that of the standard PCA-based approach when applied to a dataset of 285 mouse visual cortex's cells' 1,000 most variable genes (Tasic et al., 2016). Fewer rare cells are misclustered, engendering a 47% improvement in average silhouette width. For further discussion, see Boileau et al. (2021).

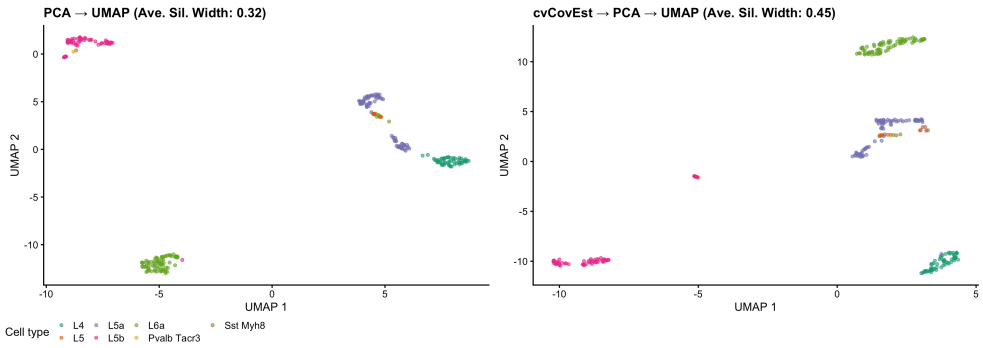


Figure 2: A comparison of UMAP embeddings using the 20 leading PCs from traditional PCA and from `cvCovEst`-based PCA as initializations.

Availability

A stable release of the `cvCovEst` package is freely-available via the Comprehensive R Archive Network. Its development version can be found on GitHub. Documentation and examples are contained in each version's manual pages, vignette, and `pkgdown` (Wickham & Hesselberth, 2020) website at https://philboileau.github.io/cvCovEst.

Acknowledgments

Philippe Boileau's contribution to this work was supported by the Fonds de recherche du Québec - Nature et technologies (B1X) and by the National Institute of Environmental Health Sciences [P42ES004705] Superfund Research Program at UC Berkeley.

We thank Jamarcus Liu for his contributions to the software package.

References

Bengtsson, H. (2020). *A unifying framework for parallel and distributed processing in R using futures*. https://arxiv.org/abs/2008.00553

Bickel, P. J., & Levina, E. (2008a). Regularized estimation of large covariance matrices. *Annals of Statistics*, *36*(1), 199–227. https://doi.org/10.1214/009053607000000758

Bickel, P. J., & Levina, E. (2008b). Covariance regularization by thresholding. *Annals of Statistics*, *36*(6), 2577–2604. https://doi.org/10.1214/08-AOS600

Boileau, P., Hejazi, N. S., van der Laan, M. J., & Dudoit, S. (2021). *Cross-validated loss-based covariance matrix estimator selection in high dimensions*. http://arxiv.org/abs/2102.09715

Cai, T. T., Zhang, C.-H., & Zhou, H. H. (2010). Optimal rates of convergence for covariance matrix estimation. *Annals of Statistics*, *38*(4), 2118–2144. https://doi.org/10.1214/09-AOS752

Coyle, J. R., & Hejazi, N. S. (2018). Origami: A generalized framework for cross-validation in r. *Journal of Open Source Software*, *3*(21), 512. https://doi.org/10.21105/joss.00512

Dudoit, S., & van der Laan, M. J. (2005). Asymptotics of cross-validated risk estimation in estimator selection and performance assessment. *Statistical Methodology*, *2*(2), 131–154. https://doi.org/10.1016/j.stamet.2005.02.003

Fan, J., & Li, R. (2001). Variable selection via nonconcave penalized likelihood and its oracle properties. *Journal of the American Statistical Association*, *96*(456), 1348–1360. https://doi.org/10.1198/016214501753382273

Fan, J., Liao, Y., & Liu, H. (2016). An overview of the estimation of large covariance and precision matrices. *The Econometrics Journal*, *19*(1), C1–C32. https://doi.org/10.1111/ectj.12061

Fan, J., Liao, Y., & Mincheva, M. (2013). Large covariance estimation by thresholding principal orthogonal complements. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, *75*(4), 603–680. https://doi.org/10.2139/ssrn.1977673

Fan, J., Liu, H., & Wang, W. (2018). Large covariance estimation through elliptical factor models. *Annals of Statistics*, *46*(4), 1383–1414. https://doi.org/10.1214/17-AOS1588

Johnstone, I. M. (2001). On the distribution of the largest eigenvalue in principalcomponents analysis. *Annals of Statistics*, *29*(2), 295–327. https://doi.org/10.1214/aos/1009210544

Ledoit, O., & Wolf, M. (2004). A well-conditioned estimator for large-dimensional covariance matrices. *Journal of Multivariate Analysis*, *88*(2), 365–411. https://doi.org/10.1016/S0047-259X(03)00096-4

Ledoit, O., & Wolf, M. (2020). Analytical nonlinear shrinkage of large-dimensional covariance matrices. *Annals of Statistics*, *48*(5), 3043–3065. https://doi.org/10.1214/19-AOS1921

Lee, K., & You, K. (2019). *CovTools: Statistical tools for covariance analysis*. https://CRAN.R-project.org/package=CovTools

McInnes, L., Healy, J., & Melville, J. (2018). *UMAP: Uniform Manifold Approximation and Projection for Dimension Reduction*. http://arxiv.org/abs/1802.03426

Pourahmadi, M. (2013). *High-dimensional covariance estimation*. Wiley. ISBN: 978-1-118-03429-3

R Core Team. (2021). *R: A language and environment for statistical computing*. R Foundation for Statistical Computing. https://www.R-project.org/

Ramprasad, P. (2016). *nlshrink: Non-linear shrinkage estimation of population eigenvalues and covariance matrices*. https://CRAN.R-project.org/package=nlshrink

Rothman, A. J., Levina, E., & Zhu, J. (2009). Generalized thresholding of large covariance matrices. *Journal of the American Statistical Association*, *104*(485), 177–186. https://doi.org/10.1198/jasa.2009.0101

Schäfer, J., & Strimmer, K. (2005). A shrinkage approach to large-scale covariance matrix estimation and implications for functional genomics. *Statistical Applications in Genetics and Molecular Biology*, *4*(1). https://doi.org/10.2202/1544-6115.1175

Tasic, B., Menon, V., Nguyen, T. N., Kim, T. K., Jarsky, T., Yao, Z., Levi, B., Gray, L. T., Sorensen, S. A., Dolbeare, T., Bertagnolli, D., Goldy, J., Shapovalova, N., Parry, S., Lee, C., Smith, K., Bernard, A., Madisen, L., Sunkin, S. M., … Zeng, H. (2016). Adult mouse cortical cell taxonomy revealed by single cell transcriptomics. *Nature Neuroscience*, *19*(2), 335–346. https://doi.org/10.1038/nn.4216

van der Laan, M. J., & Dudoit, S. (2003). *Unified cross-validation methodology for selection among estimators and a general cross-validated adaptive epsilon-net estimator: Finite sample oracle inequalities and examples* (Working Paper No. 130; U.c. Berkeley Division of Biostatistics Working Paper Series). University of California, Berkeley. https://biostats.bepress.com/ucbbiostat/paper130

van der Vaart, A. W., Dudoit, S., & van der Laan, M. J. (2006). Oracle inequalities for multi-fold cross validation. *Statistics and Decisions*, *24*, 351–371. https://doi.org/10.1524/stnd.2006.24.3.351

Wang, B. (2014). *CVTuningCov: Regularized estimators of covariance matrices with CV tuning*. https://CRAN.R-project.org/package=CVTuningCov

Wickham, H., & Hesselberth, J. (2020). *pkgdown: Make static HTML documentation for a package*. https://CRAN.R-project.org/package=pkgdown