Regularizing stock return covariance matrices via multiple testing of correlations*

Richard Luger†
Université Laval, Canada

May 14, 2022

Abstract: This paper presents a large-scale inference approach for the regularization of stock return covariance matrices in well-diversified portfolio settings. The framework allows for the presence of heavy tails and multivariate GARCH-type effects of unknown form among the stock returns. The approach proceeds by simultaneously testing all pairwise correlations and then sets to zero the elements that are not statistically significant. This adaptive thresholding of the sample correlation matrix is obtained by embedding a distribution-free Monte Carlo resampling technique into multiple testing procedures for control of the familywise error rate. A subsequent shrinkage step ensures that the final covariance matrix estimate is positive definite and well conditioned, while preserving the achieved sparsity. When compared to alternative estimators, the new regularization method is found to perform remarkably well both in simulation studies and in an actual portfolio optimization application with the 100 holdings of the Invesco S&P 500 Low Volatility ETF.

JEL classification: C12; C15; C58; G11

Keywords: Covariance estimation; Large-scale inference; Multiple testing; Adaptive thresholding; Shrinkage; Portfolio selection

*This work draws on research supported by the Social Sciences and Humanities Research Council of Canada.

†Correspondence to: Department of Finance, Insurance and Real Estate, Laval University, Quebec City, Quebec G1V 0A6, Canada.
E-mail address: richard.luger@fsa.ulaval.ca

1 Introduction

Estimation of covariance matrices is a ubiquitous problem in multivariate statistical analysis, which has applications in many fields including finance, economics, meteorology, climate research, spectroscopy, signal processing, pattern recognition, and genomics. In finance, covariance matrix estimates capture the dependencies between asset returns – a critical input for portfolio optimization and risk management. Many other techniques also rely on covariance matrix estimates, such as regression analysis, discriminant analysis, principal component analysis, and canonical correlation analysis.

Traditional estimation of the sample covariance matrix is known to perform poorly when the number of variables, N, is large compared to the number of observations, T. As the concentration ratio N/T grows, there are simply too many parameters relative to the available data points and the eigenstructure of the sample covariance matrix gets distorted in the sense that the sample eigenvalues are more spread out than the population ones; see Johnstone (2001). The most egregious case occurs as $N/T > 1$, which causes the sample covariance matrix to become singular (non-invertible). By continuity, this matrix becomes ill-conditioned (i.e., its inverse incurs large estimation errors) as N gets closer to T.

In such situations, it is desirable to find alternative estimates that are more accurate and better conditioned than the sample covariance matrix. Regularization methods for large covariance matrices can be divided into two broad categories: (i) methods that aim to improve efficiency and obtain well-conditioned matrices, and (ii) methods that introduce sparsity (off-diagonal zeros) by imposing special structures on the covariance matrix or its inverse (the precision matrix). The first group includes linear shrinkage (Ledoit and Wolf, 2003, 2004), non-linear shrinkage (Ledoit and Wolf, 2012), condition-number regularization (Won et al., 2013), and split-sample regularization (Abadir et al., 2014). Methods that impose special structure include banding or tapering (Bickel and Levina, 2008b; Wu and Pourahmadi, 2009) and thresholding (Bickel and Levina, 2008a; Cai and Liu, 2011; El Karoui, 2008; Rothman et al., 2009), which involves setting to zero the off-diagonal entries of the covariance matrix that are in absolute value below a certain data-dependent threshold.

Bailey, Pesaran, and Smith (2019), hereafter BPS, develop an alternative thresholding

approach using a multiple hypothesis testing procedure to assess the statistical significance of the elements of the sample correlation matrix; see also El Karoui (2008, p. 2748) who suggests a similar approach. The idea is to test all pairwise correlations *simultaneously*, and then to set to zero the elements that are not statistically significant. As with other thresholding methods, this multiple testing approach preserves the symmetry of the covariance matrix but it does not ensure its positive definiteness. BPS resolve this issue with an additional linear shrinkage step, whereby the correlation matrix estimator is shrunk towards the identity matrix to ensure positive definiteness. It must be emphasized that the BPS approach for reducing the number of spurious correlations is also of interest in the classical *low* N, *large* T setting.

The simultaneous testing of all pairwise correlations gives rise to a multiple comparisons problem. Indeed if the multiplicity of inferences is not taken into account, then the probability that some of the true null hypotheses (of zero correlation) are rejected by chance alone may be unduly large; see Hochberg and Tamhane (1987) and Hsu (1996) for textbook treatments of multiple comparisons. A usual objective in multiple testing is to control the familywise error rate (FWER), which is defined as the probability of rejecting at least one true null hypothesis. BPS use ideas from the multiple testing literature, but from the get-go they state in their introduction (p. 508) that they "will not be particularly concerned with controlling the overall size of the joint $N(N-1)/2$ tests" of zero pairwise correlations. The simulation evidence presented in this paper reveals that the empirical FWER with BPS thresholding can be severely inflated, resulting in far too many spurious rejections of the null hypothesis of zero correlation. This over-rejection problem is exacerbated by the presence of heavy tails, which obviously defeats the purpose of achieving sparsity.

In this paper, the BPS multiple testing regularization approach is extended so that: (i) it is applicable to financial stock returns, and (ii) it achieves control of the FWER. The theory in BPS assumes that the variables are independent over time, which is not tenable with stock returns. Indeed the independence assumption rules out the possibility of time-varying conditional variances and covariances – a well-known feature of financial returns (Cont, 2001). In turn, the presence of such effects gives rise to heavy tails and potential outliers in the distribution of returns. The methods in this paper are developed in a general

framework that allows for the presence of heavy tails and multivariate GARCH-type effects of unknown form. These robust methods achieve control of the FWER, meaning that more spurious correlations are detected and greater sparsity is induced. Simulation studies reveal that the power of the new test procedures is as good as that of the (FWER-adjusted) BPS tests.

Of course, the payoff from multiple testing regularization increases with the true degree of sparsity in the population covariance matrix. In a portfolio context, it is interesting to note that this regularization approach provides a measure of diversification when the asset returns tend to be positively correlated. Indeed the induced sparsity is expected to be proportional to the level of portfolio diversification, since in this case a well-diversified portfolio is precisely one in which the constituent assets demonstrate little or no correlation.

The rest of this paper is organized as follows. Section 2 presents the BPS multiple testing regularization approach. Section 3 establishes the financial context and Section 4 develops the multiple testing procedures. Section 5 presents the results of simulation studies that compare the performance of the new regularization method to BPS and other covariance matrix estimators. Section 6 further illustrates the large-scale inference approach with an application to portfolio optimization using the 100 holdings of the Invesco S&P 500 Low Volatility ETF. Section 7 offers some concluding remarks.

2 Multiple testing regularization

Consider a sample covariance matrix $\hat{\mathbf{\Sigma}} = [\hat{\sigma}_{ij}]_{N\times N}$ based a data sample of size T, and let $\hat{\mathbf{\Gamma}} = [\hat{\rho}_{ij}]_{N\times N}$ denote the corresponding correlation matrix with typical element $\hat{\rho}_{ij} = \hat{\sigma}_{ij}/\sqrt{\hat{\sigma}_{ii}\hat{\sigma}_{jj}}$. As usual the sample covariance and correlation matrices are related via $\hat{\mathbf{\Gamma}} = \hat{\mathbf{D}}^{-1/2}\hat{\mathbf{\Sigma}}\hat{\mathbf{D}}^{-1/2}$, where $\hat{\mathbf{D}} = \text{diag}(\hat{\sigma}^2_{1,1}, ..., \hat{\sigma}^2_{N,N})$. The BPS regularization strategy aims to improve $\hat{\mathbf{\Gamma}}$ by testing the family of $K = N(N-1)/2$ individual hypotheses $H_{i,j}$ in the setting

$$H_{i,j} : \sigma_{ij} = 0 \text{ versus } H'_{i,j} : \sigma_{ij} \neq 0, \tag{1}$$

for $i = 1, ..., N-1$ and $j = i+1, ..., N$, while controlling the FWER; i.e., the probability of at least one Type I error. The elements that are found to be statistically insignificant are then set to zero. Instead of covariances, BPS prefer to base inference on the sample correlations since they are all on the same scale. This leads to multiple testing procedures that are balanced in the sense that all constituent tests have about the same power. Note that the entries of the sample correlation matrix are intrinsically dependent even if the original observations are independent.

There are two types of FWER control. To introduce these, define an index k taking values in the set $\mathcal{K} = \{1, ..., K\}$ as $i = 1, ..., N-1$ and $j = i+1, ..., N$ so that $H_1 = H_{1,2}$, $H_2 = H_{1,3}, ...$ $H_K = H_{N-1,N}$. Furthermore, let $\mathcal{K}_0 = \{k : H_k \text{ is true}\}$ denote the index set of true hypotheses. Given the nominal significance level $\alpha \in (0, 1)$, the FWER is said to be controlled in the *weak* sense when $\Pr\left(\text{Reject at least one } H_k \,|\, \bigcap_{k\in\mathcal{K}} H_k\right) \leq \alpha$, where the conditioning is on the *complete* null hypothesis that $\mathcal{K}_0 = \mathcal{K}$. *Strong* controls is achieved when $\Pr\left(\text{Reject at least one } H_k,\, k \in \mathcal{K}_0 \,|\, \bigcap_{k\in\mathcal{K}_0} H_k\right) \leq \alpha$, regardless of the *partial* null hypothesis (i.e., the particular subset of hypotheses that happens to be true). See Hochberg and Tamhane (1987) and Hsu (1996) for a more detailed discussion of error rate control.

As a simple illustration of the multiple testing problem, consider $T = 100$ random draws from the multivariate normal distribution $N(\mathbf{0}, \mathbf{I}_N)$, where $\mathbf{I}_N$ is the $N \times N$ identity matrix. The multiplicity effect can be appreciated through the FWER and the per-family error rate (PFER), though not really a rate, defined as the expected number of Type I errors (Ge et al., 2003). Algorithm 1 (presented in Section 4.1) yields exact tests of the K null hypotheses in (1), without multiplicity adjustments. Using that algorithm with $\alpha = 5\%$ for increasing values of N, the empirical PFER and FWER (in percentages) are found to be:[1]

N	2	5	10	15
K	1	10	45	105
PFER	4.9	50.4	230.7	528.5
FWER	4.9	41.7	90.5	99.7

[1]Specifically, these results are obtained using Algorithm 1 with $\alpha = 0.05$, $B = 100$, and repeating the simulation experiment 1000 times for each considered value of N.

These results show that it is misleading to use the conventional 5% cutoff for p-values to find statistically significant correlations, since the variables are mutually independent by construction. Indeed, if the 5% cutoff is used, it becomes certain that far too many correlations will be spuriously declared significant as N grows beyond 2. The takeaway message is that it is inappropriate to use p-values that are not adjusted for the multiplicity effect.

The BPS thresholding estimator, denoted here by $\hat{\mathbf{\Gamma}}_{\text{BPS}}$, has entries computed as

$$\hat{\rho}_{\text{BPS},ij} = \hat{\rho}_{ij} \mathbb{1}\left\{|\hat{\rho}_{ij}| > T^{-1/2} c_\alpha(N)\right\}, \tag{2}$$

wherein $\mathbb{1}\{\cdot\}$ denotes the indicator function. The critical value $c_\alpha(N)$ appearing in (2) is given by

$$c_\alpha(N) = \Phi^{-1}\left(1 - \frac{\alpha}{2f(N)}\right), \tag{3}$$

where $\Phi^{-1}(\cdot)$ is the quantile function of a standard normal variate and $f(N)$ is a general function of N chosen to ensure FWER $\leq \alpha$ in the strong sense. Observe that the term $T^{-1/2}c_\alpha(N)$ in (2) is a 'universal' threshold value in that it applies to each off-diagonal element of the correlation matrix $\hat{\mathbf{\Gamma}}$.

Among other asymptotic properties, BPS show that $\hat{\mathbf{\Gamma}}_{\text{BPS}}$ converges to the true $\mathbf{\Gamma}$ as the sample size grows. As one expects, the payoff in terms of noise reduction with this approach increases with the actual number of zeros in $\mathbf{\Gamma}$.

2.1 Positive definiteness

As with other thresholding methods, the matrix $\hat{\mathbf{\Gamma}}_{\text{BPS}}$ obtained via (2) is not necessarily positive definite. BPS solve this problem by shrinking $\hat{\mathbf{\Gamma}}_{\text{BPS}}$ towards $\mathbf{I}_N$, the $N \times N$ identity matrix. Their shrinkage upon multiple testing correlation matrix estimator is given by

$$\hat{\mathbf{\Gamma}}_{\text{BPS}}(\xi) = \xi \mathbf{I}_N + (1 - \xi)\hat{\mathbf{\Gamma}}_{\text{BPS}}, \tag{4}$$

with the shrinkage parameter $\xi \in (\xi_0, 1]$, where ξ_0 is the minimum value of ξ that produces a non-singular $\hat{\mathbf{\Gamma}}_{\text{BPS}}(\xi_0)$ matrix. Note that by shrinking towards the identity matrix, the resulting correlation matrix preserves the zeros achieved in $\hat{\mathbf{\Gamma}}_{\text{BPS}}$ and its diagonal elements

do not deviate from unity. The computation of (4) is made operational by replacing ξ by ξ^*, which is found numerically as

$$\xi^* = \arg \min_{\xi_0+\epsilon \leq \xi \leq 1} \left\| \hat{\mathbf{\Gamma}}_0^{-1} - \hat{\mathbf{\Gamma}}_{\text{BPS}}^{-1}(\xi) \right\|_F^2,$$

where $\|\mathbf{A}\|_F$ denotes the Frobenius of $\mathbf{A}$, ϵ is a small positive constant, and $\hat{\mathbf{\Gamma}}_0$ is a reference matrix.[2] This reference matrix is also found by shrinkage as

$$\hat{\mathbf{\Gamma}}_0 = \hat{\theta}^* \mathbf{I}_N + (1 - \hat{\theta}^*)\hat{\mathbf{\Gamma}},$$

where

$$\hat{\theta}^* = 1 - \frac{\sum\sum_{i \neq j} \hat{\rho}_{ij} \left[\hat{\rho}_{ij} - \frac{\hat{\rho}_{ij}(1-\hat{\rho}_{ij}^2)}{2T} \right]}{\frac{1}{T}\sum\sum_{i \neq j}(1 - \hat{\rho}_{ij}^2)^2 + \sum\sum_{i \neq j} \left[\hat{\rho}_{ij} - \frac{\hat{\rho}_{ij}(1-\hat{\rho}_{ij}^2)}{2T} \right]^2},$$

with the proviso that if $\hat{\theta}^* < 0$ then $\hat{\theta}^*$ is set to 0, and if $\hat{\theta}^* > 1$ then it is set to 1.[3] The resulting covariance matrix is given by

$$\hat{\mathbf{\Sigma}}_{\text{BPS}}(\xi^*) = \hat{\mathbf{D}}^{1/2} \hat{\mathbf{\Gamma}}_{\text{BPS}}(\xi^*) \hat{\mathbf{D}}^{1/2},$$

where $\hat{\mathbf{\Gamma}}_{\text{BPS}}(\xi^*)$ corresponds to (4) evaluated with ξ^*.

3 Financial context

Consider a diversified mix of N financial assets with time-t returns $\mathbf{r}_t = (r_{1,t}, ..., r_{N,t})'$ and let $\mathcal{I}_t = (\mathbf{r}_t', \mathbf{r}_{t-1}', ...)'$. The returns are decomposed as

$$\begin{aligned} \mathbf{r}_t &= \boldsymbol{\mu}_t + \boldsymbol{\varepsilon}_t, \\ \boldsymbol{\varepsilon}_t &= \mathbf{\Sigma}_t^{1/2} \mathbf{z}_t, \end{aligned} \tag{5}$$

[2]Recall that for a matrix $\mathbf{A}$ its Frobenius norm is given by $\|\mathbf{A}\|_F = \sqrt{\text{Tr}(\mathbf{A}'\mathbf{A})}$, where $\text{Tr}(\cdot)$ returns the matrix trace.

[3]The quantity $\hat{\theta}^*$ is an estimate of the optimal value of the shrinkage parameter that minimizes $E\|\hat{\mathbf{\Gamma}}_0 - \mathbf{\Gamma}\|_F$, assuming that the first two moments of the distribution of $\hat{\mathbf{\Gamma}}$ exist; see BPS for additional details.

where $\boldsymbol{\mu}_t = E(\mathbf{r}_t \,|\, \mathbf{x}_t, \mathcal{I}_{t-1}; \boldsymbol{\theta})$ is a specified model with parameters $\boldsymbol{\theta}$ for the conditional expectation of $\mathbf{r}_t$, given some explanatory variables $\mathbf{x}_t$ and past returns $\mathcal{I}_{t-1}$. The error $\boldsymbol{\varepsilon}_t = (\varepsilon_{1,t}, ..., \varepsilon_{N,t})'$ in (5) consists of an innovation vector $\mathbf{z}_t$ satisfying Assumption 1 below, and an unspecified $N \times N$ "square root" matrix $\boldsymbol{\Sigma}_t^{1/2}$ such that $\boldsymbol{\Sigma}_t^{1/2}\boldsymbol{\Sigma}_t^{1/2} = E(\boldsymbol{\varepsilon}_t\boldsymbol{\varepsilon}_t' \,|\, \mathbf{x}_t, \mathcal{I}_{t-1})$. This framework is compatible with several popular models of time-varying covariances, such as multivariate GARCH models (Silvennoinen and Teräsvirta, 2009) and multivariate stochastic volatility models (Chib et al., 2009).

Assumption 1. *The innovations* $\{\mathbf{z}_t\}_t$ *are independently (but not necessarily identically) distributed according to spherically symmetric distributions, with moments* $E(\mathbf{z}_t \,|\, \mathbf{x}_t, \mathcal{I}_{t-1}) = \mathbf{0}$ *and* $E(\mathbf{z}_t\mathbf{z}_t' \,|\, \mathbf{x}_t, \mathcal{I}_{t-1}) = \mathbf{I}_N$ *for each* t.

This assumption means that $\mathbf{z}_t$ admits the stochastic representation $\mathbf{z}_t \overset{d}{=} \mathbf{H}\mathbf{z}_t$, where the symbol $\overset{d}{=}$ stands for an equality in distribution and $\mathbf{H}$ is any $N \times N$ orthogonal matrix such that $\mathbf{H}'\mathbf{H} = \mathbf{H}\mathbf{H}' = \mathbf{I}_N$. This class includes the multivariate standard normal, Student-t, and logistic distributions, among many others; see Fang et al. (1990) for an in-depth treatment of spherically symmetric distributions.

When $\mathbf{z}_t$ has a well-defined density, then Assumption 1 is equivalent to assuming that the conditional distribution of $\mathbf{r}_t$ is elliptically symmetric, meaning that its density has the form $|\boldsymbol{\Sigma}_t^{-1/2}|g((\mathbf{r}_t - \boldsymbol{\mu}_t)'\boldsymbol{\Sigma}_t^{-1}(\mathbf{r}_t - \boldsymbol{\mu}_t))$ for some non-negative scalar function $g(\cdot)$. Elliptically symmetric distributions play a very important role in mean-variance analysis because they guarantee full compatibility with expected utility maximization regardless of investor preferences (Berk, 1997; Chamberlain, 1983; Owen and Rabinovitch, 1983).

In the context of (5), the complete null hypothesis is formally stated as

$$H_0 : \boldsymbol{\Sigma}_t^{1/2} = \mathbf{D}_t^{1/2}, \tag{6}$$

for each t, where $\mathbf{D}_t^{1/2}$ is a diagonal matrix (i.e., with zeros outside the main diagonal). Observe that conditional heteroskedasticity is permitted under H_0, i.e., the diagonal elements of $\mathbf{D}_t^{1/2}$ may be time-varying. It is easy to see that when Assumption 1 holds and H_0 is true,

the error vector $\boldsymbol{\varepsilon}_t = (\varepsilon_{1,t}, ..., \varepsilon_{N,t})'$ becomes sign-symmetric (Serfling, 2006) in the sense that

$$\boldsymbol{\varepsilon}_t \stackrel{d}{=} \mathbb{S}\boldsymbol{\varepsilon}_t,$$

for all $N \times N$ diagonal matrices $\mathbb{S}$ with ± 1 on the diagonal.

Assumption 2. *The unconditional covariance matrix* $\boldsymbol{\Sigma} = E(\boldsymbol{\varepsilon}_t \boldsymbol{\varepsilon}_t')$ *exists.*

With this assumption the sign-symmetry condition $(\varepsilon_{it}, \varepsilon_{jt}) \stackrel{d}{=} (\pm\varepsilon_{it}, \pm\varepsilon_{jt})$ implies $\sigma_{ij} = 0$, for $i \neq j$, where σ_{ij} is the (i,j)th element of $\boldsymbol{\Sigma}$ (Randles and Wolfe, 1979, Lemma 1.3.28).

4 Multiple testing procedures

Assume momentarily that the value of $\boldsymbol{\theta}$ in (5) is known so that $\boldsymbol{\mu}_t$ is also known. For instance with daily returns it is often reasonable to assume that $\boldsymbol{\mu}_t = \mathbf{0}$. The case of unknown location parameters will be dealt with in Section 4.5.

Given the values of $\boldsymbol{\mu}_t$, centred returns can then be defined as $\mathbf{y}_t = \mathbf{r}_t - \boldsymbol{\mu}_t = (y_{1,t}, ..., y_{N,t})'$, for $t = 1, ..., T$, and these have the same properties as $\boldsymbol{\varepsilon}_t$. The time series of centred returns are collected into $\mathbf{Y} = [\mathbf{y}_1, ..., \mathbf{y}_T]'$. Following BPS, inference is based on the pairwise correlations $\hat{\rho}_{ij} = \hat{\sigma}_{ij}/\sqrt{\hat{\sigma}_{ii}\hat{\sigma}_{jj}}$ that constitute the matrix $\hat{\boldsymbol{\Gamma}} = [\hat{\rho}_{ij}]_{N\times N}$. This matrix can be obtained from the familiar relationship $\hat{\boldsymbol{\Gamma}} = \hat{\mathbf{D}}^{-1/2}\hat{\boldsymbol{\Sigma}}\hat{\mathbf{D}}^{-1/2}$ with $\hat{\mathbf{D}} = \text{diag}(\hat{\sigma}^2_{1,1}, ..., \hat{\sigma}^2_{N,N})$, where the variances and covariances about the origin are computed as $\hat{\sigma}_{ij} = T^{-1}\sum_{t=1}^{T} y_{i,t}y_{j,t}$, for $i, j = 1, ..., N$.

Let $\tilde{\mathbb{S}}_t = \text{diag}(\tilde{s}_{1,t}, ..., \tilde{s}_{N,t})$, for $t = 1, ..., T$, where $\tilde{s}_{i,t}$ are independent Rademacher random draws such that $\Pr(\tilde{s}_{i,t} = 1) = \Pr(\tilde{s}_{i,t} = -1) = 1/2$, for each i, t. An artificial sample $\tilde{\mathbf{Y}} = [\tilde{\mathbf{y}}_1, ..., \tilde{\mathbf{y}}_T]'$ with $\tilde{\mathbf{y}}_t = (\tilde{y}_{1,t}, ..., \tilde{y}_{N,t})'$ is then defined as

$$\tilde{\mathbf{Y}} = \left[\tilde{\mathbb{S}}_1\mathbf{y}_1, ..., \tilde{\mathbb{S}}_T\mathbf{y}_T\right]'. \tag{7}$$

If Assumption 1 holds and H_0 is true, then $\mathbf{Y} \stackrel{d}{=} \tilde{\mathbf{Y}}$, for each of the 2^{NT} possible matrix realizations of $\tilde{\mathbf{Y}}$, given $|\mathbf{Y}|$. Here $|\mathbf{Y}|$ is the matrix of entrywise absolute values of $\mathbf{Y}$. For a

given artificial sample $\tilde{\mathbf{Y}}$, let $\tilde{\mathbf{\Gamma}} = [\tilde{\rho}_{ij}]_{N\times N}$ denote the associated correlation matrix comprising the pairwise correlations about the origin $\tilde{\rho}_{ij} = \tilde{\sigma}_{ij}/\sqrt{\tilde{\sigma}_{ii}\tilde{\sigma}_{jj}}$, where $\tilde{\sigma}_{ij} = T^{-1}\sum_{t=1}^{T}\tilde{y}_{i,t}\tilde{y}_{j,t}$.

Proposition 1. *Suppose that (5) holds along with Assumption 1, and consider* $\tilde{\mathbf{Y}}$ *generated according to (7). If* H_0 *in (6) is true, then* $\Pr(\hat{\mathbf{\Gamma}} = \tilde{\mathbf{\Gamma}} \,|\, |\mathbf{Y}|) = 1/2^{NT}$. *Furthermore under Assumption 2,* $E(\tilde{y}_{i,t}\tilde{y}_{j,t} \,|\, |\mathbf{Y}|) = 0$, *for* $i \neq j$.

From Theorem 1.3.7 in Randles and Wolfe (1979) it is known that if $\mathbf{Y} \overset{d}{=} \tilde{\mathbf{Y}}$ and $\mathcal{F}(\cdot)$ is a measurable function (possibly vector-valued) defined on the common support of $\mathbf{Y}$ and $\tilde{\mathbf{Y}}$, then $\mathcal{F}(\mathbf{Y}) \overset{d}{=} \mathcal{F}(\tilde{\mathbf{Y}})$. The fact that the 2^{NT} possible values (not necessarily distinct) of $\tilde{\mathbf{\Gamma}}$ are equally likely values for $\hat{\mathbf{\Gamma}}$ follows by taking $\mathcal{F}(\cdot)$ to be the covariance function. Observe that (7) conditions on the absolute values of $y_{i,t}$, since only their signs are randomized. The zero covariance property in Proposition 1 follows from the independence of the Rademacher draws used to generate $\tilde{\mathbf{Y}}$.

Proposition 1 shows that $\hat{\mathbf{\Gamma}}$ is conditionally *pivotal* under H_0, meaning that its sign-randomization distribution does not depend on any nuisance parameters. In principle, critical values could be found from the conditional distribution of $\hat{\mathbf{\Gamma}}$ derived from the 2^{NT} equally likely values represented by $\tilde{\mathbf{\Gamma}}$. Determination of this distribution from a complete enumeration of all possible realizations of $\hat{\mathbf{\Gamma}}$ is obviously impractical. To circumvent this problem and still obtain exact p-values, the Monte Carlo test technique (Barnard, 1963; Birnbaum, 1974; Dwass, 1957) is used in the algorithms presented next.

4.1 Unadjusted p-values

It is useful to first describe how to obtain the Monte Carlo p-values without multiplicity adjustments, even if they are not used subsequently for multiple testing regularization. Note that sampling according to (7) yields a discrete distribution of $\tilde{\mathbf{\Gamma}}$ values, which means that ties among the resampled values can occur, at least theoretically. Following Dufour (2006), these are dealt with by working with lexicographic (tie-breaking) ranks. Algorithm 1 details the steps to obtain the unadjusted Monte Carlo p-values.

Algorithm 1 (Unadjusted Monte Carlo p-values).

1. Choose B so that αB is an integer, where $\alpha \in (0,1)$ is the desired significance level.

2. For $b = 1, ..., B-1$, repeat the following steps:

   (a) Generate an artificial data sample $\tilde{\mathbf{Y}}_b$ according to (7).

   (b) Compute the associated matrix of correlations about the origin $\tilde{\mathbf{\Gamma}}_b = [\tilde{\rho}_{ij,b}]_{N\times N}$.

3. Upon completion, create the pairs $(\tilde{\rho}_{ij,1,1}), ..., (\tilde{\rho}_{ij,B-1}, u_{B-1}), (\hat{\rho}_{ij}, u_B)$, where $u_b \sim \mathcal{U}(0,1)$, $b = 1, ..., B$. Next, compute the lexicographic rank of $|\hat{\rho}_{ij}|$ among the $|\tilde{\rho}_{ij}|$'s as

$$R_B^{\mathrm{U}}(|\hat{\rho}_{ij}|) = 1 + \sum_{b=1}^{B-1} \mathbb{1}\left\{|\hat{\rho}_{ij}| > |\tilde{\rho}_{ij,b}|\right\} + \sum_{b=1}^{B-1} \mathbb{1}\left\{|\hat{\rho}_{ij}| = |\tilde{\rho}_{ij,b}|\right\} \mathbb{1}\left\{u_B > u_b\right\}.$$

4. The unadjusted Monte Carlo p-values are then given by

$$\tilde{p}_{\mathrm{U}}(|\hat{\rho}_{ij}|) = \frac{B - R_B^{\mathrm{U}}(|\hat{\rho}_{ij}|) + 1}{B},$$

   for $i = 1, ..., N-1$, and $j = i+1, ..., N$.

Proposition 2. *Suppose that (5) holds along with Assumptions 1 and 2. Then the Monte Carlo p-values computed according to Algorithm 1 are such that* $\Pr(\tilde{p}_U(|\hat{\rho}_{ij}|) \leq \alpha \,|\, H_0) = \alpha$, *for* $1 \leq i < j \leq N$.

This result follows from Proposition 2.4 in Dufour (2006) on the validity of Monte Carlo tests for general statistics; see also Lemma 1 in Romano and Wolf (2005). The key is the recognition that the pairs $(\tilde{\rho}_{ij,1}, u_1), ..., (\tilde{\rho}_{ij,B-1}, u_{B-1}), (\hat{\rho}_{ij}, u_B)$ are *exchangeable* under H_0. This implies that $\Pr(R_B^{\mathrm{U}}(|\hat{\rho}_{ij}|) = b) = 1/B$, for $b = 1, ..., B$, when H_0 holds true. In words, this simply says that the lexicographic rank of exchangeable random variables are uniformly distributed over the integers $1, ..., B$. The Monte Carlo p-values have the usual interpretation: $\tilde{p}_{\mathrm{U}}(|\hat{\rho}_{ij}|)$ is the proportion of $|\tilde{\rho}_{ij}|$ values as extreme or more extreme than the observed $|\hat{\rho}_{ij}|$ value in its resampling distribution. See Dufour and Khalaf (2001) and Kiviet (2011, Ch. 6) for a general overview of the Monte Carlo test technique and further references.

4.2 Single-step adjusted p-values

Westfall and Young (1993) propose several resampling-based methods to adjust p-values so as to account for multiplicity. Adjusted p-values are defined as the smallest significance level for which one still rejects an individual hypothesis $H_{i,j}$, given a particular multiple test procedure. Adapting their *single-step max-t adjusted p-values* to the present context yields the definition

$$p_{\text{SS}}(|\hat{\rho}_{ij}|) = \Pr\left(\max_{1 \leq i < j \leq N} |\tilde{\rho}_{ij}| \geq |\hat{\rho}_{ij}| \,\middle|\, H_0 \right), \tag{8}$$

where H_0 is the complete null hypothesis in (6) that all the covariances are zero; see Westfall and Young (1993, Section 2.3.2). In words, this says that the single-step (SS) adjusted p-value is the probability that the maximum absolute correlation in the artificial data is greater than the observed absolute correlation in the actual data. Let $\hat{m} = \max_{1 \leq i < j \leq N} |\hat{\rho}_{ij}|$ and note that $\hat{m} \overset{d}{=} \tilde{m}$, where $\tilde{m} = \max_{1 \leq i < j \leq N} |\tilde{\rho}_{ij}|$ is computed from an artificial sample $\tilde{\mathbf{Y}}$. Algorithm 2 exploits this result in order to compute the Monte Carlo version of (8).

Algorithm 2 (Single-step adjusted Monte Carlo p-values).

1. Choose B so that αB is an integer, where $\alpha \in (0, 1)$ is the desired FWER.

2. For $b = 1, ..., B - 1$, repeat the following steps:

   (a) Generate an artificial data sample $\tilde{\mathbf{Y}}_b$ according to (7).

   (b) Compute the associated matrix of correlations about the origin $\tilde{\mathbf{\Gamma}}_b = [\tilde{\rho}_{ij,b}]_{N \times N}$.

   (c) Find $\tilde{m}_b = \max_{1 \leq i < j \leq N} |\tilde{\rho}_{ij,b}|$

3. Next, create the pairs $(\tilde{m}_1, u_1), ..., (\tilde{m}_{B-1}, u_{B-1}), (\hat{\rho}_{ij}, u_B)$, where $u_b \sim \mathcal{U}(0, 1)$, $b = 1, ..., B$, and compute the lexicographic rank of $|\hat{\rho}_{ij}|$ among the $\tilde{m}_b$'s as

$$R_B^{\text{SS}}(|\hat{\rho}_{ij}|) = 1 + \sum_{b=1}^{B-1} \mathbb{1}\left\{ |\hat{\rho}_{ij}| > \tilde{m}_b \right\} + \sum_{b=1}^{B-1} \mathbb{1}\left\{ |\hat{\rho}_{ij}| = \tilde{m}_b \right\} \mathbb{1}\left\{ u_B > u_b \right\}.$$

4. The SS adjusted Monte Carlo p-values are given by

$$\tilde{p}_{\text{SS}}(|\hat{\rho}_{ij}|) = \frac{B - R_B^{\text{SS}}(|\hat{\rho}_{ij}|) + 1}{B},$$

for $i = 1, ..., N-1$, and $j = i+1, ..., N$.

Since the SS adjusted p-values in Algorithm 2 are computed under the complete null hypothesis H_0 (meaning that all $H_{i,j}$ are true), it should not be surprising that the FWER is controlled. Indeed if the null hypothesis $H_{i,j}$ is rejected when $\tilde{p}_{\text{SS}}(|\hat{\rho}_{ij}|) \leq \alpha$, then

$$\begin{aligned}
\Pr\left(\text{Reject at least one } H_{i,j} \,|\, H_0\right) &= \Pr\left(\text{At least one } \tilde{p}_{\text{SS}}(|\hat{\rho}_{ij}|) \leq \alpha \,|\, H_0\right) \\
&= \Pr\left(\min_{1 \leq i < j \leq N} \tilde{p}_{\text{SS}}(|\hat{\rho}_{ij}|) \leq \alpha \,|\, H_0\right) \\
&\leq \Pr\left(\tilde{p}_{\text{SS}}(|\hat{\rho}_{ij}|) \leq \alpha \,|\, H_0\right),
\end{aligned}$$

where the last line follows from the fact that $\min_{1 \leq i < j \leq N} \tilde{p}_{\text{SS}}(|\hat{\rho}_{ij}|) \leq \tilde{p}_{\text{SS}}(|\hat{\rho}_{ij}|)$. Applying Algorithm 1 with the statistic $\max_{1 \leq i < j \leq N} |\hat{\rho}_{ij}|$ yields the unadjusted p-value $\tilde{p}_{\text{U}}\left(\max_{1 \leq i < j \leq N} |\hat{\rho}_{ij}|\right)$. If the same artificial samples $\tilde{\mathbf{Y}}_b$, $b = 1, ..., B-1$, and uniform draws u_b, $b = 1, ..., B$, are used in Algorithms 1 and 2, then

$$\tilde{p}_{\text{SS}}\left(|\hat{\rho}_{ij}|\right) \leq \tilde{p}_{\text{U}}\left(\max_{1 \leq i < j \leq N} |\hat{\rho}_{ij}|\right),$$

by virtue of the fact that $|\hat{\rho}_{ij}| \leq \max_{1 \leq i < j \leq N} |\hat{\rho}_{ij}|$. And since Algorithm 1 guarantees $\Pr\left(\tilde{p}_{\text{U}}\left(\max_{1 \leq i < j \leq N} |\hat{\rho}_{ij}|\right) \leq \alpha\right) = \alpha$ under H_0, it follows that Algorithm 2 controls the FWER in the weak sense; i.e., $\Pr\left(\min_{1 \leq i < j \leq N} \tilde{p}_{\text{SS}}(|\hat{\rho}_{ij}|) \leq \alpha \,|\, H_0\right) \leq \alpha$.

The proof in Westfall and Young (1993, p. 53) that their SS adjusted p-values control the FWER in the strong sense relies heavily on the assumption of subset pivotality. That is, they assume that the joint distribution of unadjusted p-values under any partial null hypothesis is identical to that under the complete null hypothesis. As noted by Westfall and Young (1993, p. 42, Example 2.2) and Romano and Wolf (2005, Example 7), this assumption fails in the context of testing pairwise correlations. However, subset pivotality is not a necessary condition for strong control; see, for example, Romano and Wolf (2005), Westfall and Troendle (2008), and Goeman and Solari (2010).

To prove that the SS adjusted p-values computed according to Algorithm 2 achieve strong control of the FWER, recall from Section 2 that H_k corresponds to $H_{i,j}$ and $\mathcal{K}_0$ refers to the set of true null hypotheses. Accordingly, let $\tilde{p}_{\text{U}}(|\hat{\rho}_k|) = \tilde{p}_{\text{U}}(|\hat{\rho}_{ij}|)$ and $\tilde{p}_{\text{SS}}(|\hat{\rho}_k|) = \tilde{p}_{\text{SS}}(|\hat{\rho}_{ij}|)$,

where $\hat{\rho}_1 = \hat{\rho}_{1,2}$, $\hat{\rho}_2 = \hat{\rho}_{1,3}, \ldots, \hat{\rho}_K = \hat{\rho}_{N-1,N}$. From the logical implication $\{H_0 \text{ is true}\} \Rightarrow \{\bigcap_{k\in\mathcal{K}_0} H_k \text{ is true}\}$, for every possible choice $\mathcal{K}_0$, it follows that

$$\Pr\left(\min_{k\in\mathcal{K}_0} \tilde{p}_{\mathrm{SS}}(|\hat{\rho}_k|) > \alpha \,|\, \bigcap_{k\in\mathcal{K}_0} H_k\right) \geq \Pr\left(\min_{k\in\mathcal{K}_0} \tilde{p}_{\mathrm{SS}}(|\hat{\rho}_k|) > \alpha \,|\, H_0\right). \tag{9}$$

Note also that $\min_{k\in\mathcal{K}} \tilde{p}_{\mathrm{SS}}(|\hat{\rho}_k|) \leq \min_{k\in\mathcal{K}_0} \tilde{p}_{\mathrm{SS}}(|\hat{\rho}_k|)$, since $\mathcal{K}_0 \subseteq \mathcal{K}$, and hence

$$\Pr\left(\min_{k\in\mathcal{K}_0} \tilde{p}_{\mathrm{SS}}(|\hat{\rho}_k|) > \alpha \,|\, H_0\right) \geq \Pr\left(\min_{k\in\mathcal{K}} \tilde{p}_{\mathrm{SS}}(|\hat{\rho}_k|) > \alpha \,|\, H_0\right). \tag{10}$$

Combining (9) and (10) yields

$$\Pr\left(\min_{k\in\mathcal{K}_0} \tilde{p}_{\mathrm{SS}}(|\hat{\rho}_k|) > \alpha \,|\, \bigcap_{k\in\mathcal{K}_0} H_k\right) \geq \Pr\left(\min_{k\in\mathcal{K}} \tilde{p}_{\mathrm{SS}}(|\hat{\rho}_k|) > \alpha \,|\, H_0\right)$$

or equivalently

$$\Pr\left(\min_{k\in\mathcal{K}_0} \tilde{p}_{\mathrm{SS}}(|\hat{\rho}_k|) \leq \alpha \,|\, \bigcap_{k\in\mathcal{K}_0} H_k\right) \leq \Pr\left(\min_{k\in\mathcal{K}} \tilde{p}_{\mathrm{SS}}(|\hat{\rho}_k|) \leq \alpha \,|\, H_0\right).$$

The left-hand side equals $\Pr\big(\text{Reject at least one } H_k,\, k \in \mathcal{K}_0 \,|\, \bigcap_{k\in\mathcal{K}_0} H_k\big)$, while the right-hand side is $\leq \alpha$ by virtue of weak control. Therefore, the probability of a false rejection occurring under $\bigcap_{k\in\mathcal{K}_0} H_k$ is bounded above by the probability of a false rejection under H_0. This also shows that the SS procedure becomes more conservative as $\mathcal{K}_0$ shrinks.

4.3 Step-down adjusted p-values

A disconcerting feature of (8) is that all p-values are adjusted according to the distribution of the maximum absolute correlation. Potentially less conservative p-values may be obtained from step-down adjustments that result in uniformly smaller p-values, while retaining the same protection against Type I errors. With the absolute correlations $\hat{\rho}_1, \ldots, \hat{\rho}_K$, let the ordered test statistics have index values $\pi_1, \ldots, \pi_K$ so that $|\hat{\rho}_{\pi_1}| \geq |\hat{\rho}_{\pi_2}| \geq \ldots \geq |\hat{\rho}_{\pi_K}|$. In the present context, the Westfall and Young (1993) *step-down* (SD) *max-t adjusted p-values* can

be defined as

$$
\begin{aligned}
p_{\mathrm{SD}}(|\hat{\rho}_{\pi_1}|) &= \Pr\left(\max_{k=1,\ldots,K} |\tilde{\rho}_{\pi_k}| \geq |\hat{\rho}_{\pi_1}| \,|\, H_0\right), \\
p_{\mathrm{SD}}(|\hat{\rho}_{\pi_2}|) &= \max\left[p_{\mathrm{SD}}(|\hat{\rho}_{\pi_1}|),\, \Pr\left(\max_{k=2,\ldots,K} |\tilde{\rho}_{\pi_k}| \geq |\hat{\rho}_{\pi_2}| \,|\, H_0\right)\right], \\
&\vdots \\
p_{\mathrm{SD}}(|\hat{\rho}_{\pi_\ell}|) &= \max\left[p_{\mathrm{SD}}(|\hat{\rho}_{\pi_{\ell-1}}|),\, \Pr\left(\max_{k=\ell,\ldots,K} |\tilde{\rho}_{\pi_k}| \geq |\hat{\rho}_{\pi_\ell}| \,|\, H_0\right)\right], \\
&\vdots \\
p_{\mathrm{SD}}(|\hat{\rho}_{\pi_K}|) &= \max\left[p_{\mathrm{SD}}(|\hat{\rho}_{\pi_{K-1}}|),\, \Pr\left(|\tilde{\rho}_{\pi_K}| \geq |\hat{\rho}_{\pi_K}| \,|\, H_0\right)\right],
\end{aligned}
\tag{11}
$$

with the sequence of index values $\pi_1, ..., \pi_K$ held *fixed*.[4] Instead of adjusting all p-values according to the distribution of the maximum absolute correlation, this approach only adjusts the p-value of $|\hat{\rho}_{\pi_1}|$ using this distribution. The remaining p-values are then adjusted according to smaller and smaller sets of p-values. This approach can yield power improvements since the SD p-values are uniformly smaller than their SS counterparts. Based on Westfall and Young (1993, Algorithm 4.1, pp. 116–117) and Ge et al. (2003, Box 2), Algorithm 3 shows how to compute the Monte Carlo version of the SD p-values in (11).

Algorithm 3 (Step-down adjusted Monte Carlo p-values).

1. With the actual data, get the index values $\pi_1, ..., \pi_K$ that define the ordering $|\hat{\rho}_{\pi_1}| \geq |\hat{\rho}_{\pi_2}| \geq ... \geq |\hat{\rho}_{\pi_K}|$.

2. Choose B so that αB is an integer, where $\alpha \in (0, 1)$ is the desired FWER.

3. For $b = 1, ..., B - 1$, repeat the following steps:

   (a) Generate an artificial data sample $\tilde{\mathbf{Y}}_b$, according to (7).

   (b) Compute the associated matrix of correlations about the origin $\tilde{\mathbf{\Gamma}}_b = [\tilde{\rho}_{ij,k}]_{N\times N}$.

[4] That is, the adjustments are made by "stepping down" from the largest test statistic to the smallest. Romano and Wolf (2005) discuss the *idealized* step-down method; see also Romano and Wolf (2016). That method is not feasible with the developed resampling scheme since it is not possible to generate artificial data that obey the null hypothesis for each possible intersection of true null hypotheses.

(c) Find the simulated successive maxima as

$$\tilde{m}_{K,b} = |\tilde{\rho}_{\pi_K,b}|,$$
$$\tilde{m}_{k,b} = \max\left(\tilde{m}_{k+1,b}, |\tilde{\rho}_{\pi_k,b}|\right), \text{ for } k = K-1, ..., 1.$$

4. Create the pairs $(\tilde{m}_1, u_1), ..., (\tilde{m}_{B-1}, u_{B-1}), (\hat{\rho}_{\pi_k}, u_B)$, where as before $u_b \sim \mathcal{U}(0,1)$, $b = 1, ..., B$, and compute the lexicographic rank of $|\hat{\rho}_{\pi_k}|$ among the $\tilde{m}_{k,b}$'s as

$$R_B^{\text{SD}}(|\hat{\rho}_{\pi_k}|) = 1 + \sum_{b=1}^{B-1} \mathbb{1}\left\{|\hat{\rho}_{\pi_k}| > \tilde{m}_{k,b}\right\} + \sum_{b=1}^{B-1} \mathbb{1}\left\{|\hat{\rho}_{\pi_k}| = \tilde{m}_{k,b}\right\} \mathbb{1}\left\{u_B > u_b\right\}.$$

5. The SD adjusted Monte Carlo p-values are given by

$$\tilde{p}_{\text{SD}}(|\hat{\rho}_{\pi_k}|) = \frac{B - R_B^{\text{SD}}(|\hat{\rho}_{\pi_k}|) + 1}{B}, \quad \text{for } k = 1, ..., K,$$

with monotonicity of the p-values enforced by setting[5]

$$\tilde{p}_{\text{SD}}(|\hat{\rho}_{\pi_1}|) \leftarrow \tilde{p}_{\text{SD}}(|\hat{\rho}_{\pi_1}|),$$
$$\tilde{p}_{\text{SD}}(|\hat{\rho}_{\pi_k}|) \leftarrow \max\left(\tilde{p}_{\text{SD}}(|\hat{\rho}_{\pi_{k-1}}|), \tilde{p}_{\text{SD}}(|\hat{\rho}_{\pi_k}|)\right), \quad \text{for } k = 2, ..., K.$$

6. In terms of the original matrix indices, the p-values are recovered as $\tilde{p}_{\text{SD}}(|\hat{\rho}_{ij}|) = \tilde{p}_{\text{SD}}(|\hat{\rho}_{\pi_k}|)$ by reversing the mapping $(i,j) \mapsto k$.

To see that Algorithm 3 achieves weak control of the FWER, note the equivalence

$$\{\text{Reject at least one } H_{i,j}\} \iff \{\tilde{p}_{\text{SD}}(|\hat{\rho}_{\pi_1}|) \leq \alpha\}$$

and the inequalities

$$\tilde{p}_{\text{SD}}(|\hat{\rho}_{\pi_1}|) \leq \tilde{p}_{\text{SD}}(|\hat{\rho}_{\pi_k}|) \leq \tilde{p}_{\text{U}}(\max_{k=1,...,K} |\hat{\rho}_{\pi_k}|),$$

conditional on the same underlying $\tilde{\mathbf{Y}}_b$, $b = 1, ..., B-1$, and u_b, $b = 1, ..., B$, being used in Algorithms 1 and 3. It follows that $\Pr\left(\text{Reject at least one } H_{i,j} \,|\, H_0\right) \leq$

[5]This step ensures that the adjusted p-values have the same step-down monotonicity as the original statistics; i.e., smaller p-values are associated with larger values of the $|\hat{\rho}_k|$ statistics.

$\Pr\left(\tilde{p}_{\mathrm{U}}(\max_{k=1,\ldots,K}|\hat{\rho}_{\pi_k}|) \leq \alpha \,|\, H_0\right) = \alpha$, where the last equality holds by virtue of Algorithm 1.

For the proof of strong control, let k^* be the smallest index among the monotonicity enforced p-values in Step 5 of Algorithm 3 where a true hypothesis is rejected. By construction, $\tilde{p}_{\mathrm{SD}}(|\hat{\rho}_{\pi_1}|) \leq \tilde{p}_{\mathrm{SD}}(|\hat{\rho}_{\pi_{k^*}}|)$ and this implies

$$\Pr\left(\tilde{p}_{\mathrm{SD}}(|\hat{\rho}_{\pi_{k^*}}|) > \alpha \,|\, \bigcap_{k\in\mathcal{K}_0} H_k\right) \geq \Pr\left(\tilde{p}_{\mathrm{SD}}(|\hat{\rho}_{\pi_1}|) > \alpha \,|\, \bigcap_{k\in\mathcal{K}_0} H_k\right). \tag{12}$$

As seen before $\{H_0 \text{ is true}\} \Rightarrow \{\bigcap_{k\in\mathcal{K}_0} H_k \text{ is true}\}$, for every possible choice $\mathcal{K}_0$ of true null hypotheses, which here implies

$$\Pr\left(\tilde{p}_{\mathrm{SD}}(|\hat{\rho}_{\pi_1}|) > \alpha \,|\, \bigcap_{k\in\mathcal{K}_0} H_k\right) \geq \Pr\left(\tilde{p}_{\mathrm{SD}}(|\hat{\rho}_{\pi_1}|) > \alpha \,|\, H_0\right). \tag{13}$$

From (12) and (13), one obtains

$$\Pr\left(\tilde{p}_{\mathrm{SD}}(|\hat{\rho}_{\pi_{k^*}}|) > \alpha \,|\, \bigcap_{k\in\mathcal{K}_0} H_k\right) \geq \Pr\left(\tilde{p}_{\mathrm{SD}}(|\hat{\rho}_{\pi_1}|) > \alpha \,|\, H_0\right),$$

or equivalently

$$\Pr\left(\tilde{p}_{\mathrm{SD}}(|\hat{\rho}_{\pi_{k^*}}|) \leq \alpha \,|\, \bigcap_{k\in\mathcal{K}_0} H_k\right) \leq \Pr\left(\tilde{p}_{\mathrm{SD}}(|\hat{\rho}_{\pi_1}|) \leq \alpha \,|\, H_0\right),$$

where the right-hand side is $\leq \alpha$ by weak control of the SD adjusted Monte Carlo p-values. Once again, the probability of a false rejection occurring under $\bigcap_{k\in\mathcal{K}_0} H_k$ (the left-hand side) is bounded above by the probability of a false rejection under H_0.

Remark that since $\tilde{p}_{\mathrm{SD}}(|\hat{\rho}_{ij}|) \leq \tilde{p}_{\mathrm{SS}}(|\hat{\rho}_{ij}|)$, given the same underlying randomization (i.e., the same artificial samples $\tilde{\mathbf{Y}}_b$, $b = 1, ..., B-1$ and uniform draws u_b, $b = 1, ..., B$) in Algorithms 2 and 3, the SD adjustments will tend to be less conservative than their SS counterparts as the set $\mathcal{K}_0$ shrinks.

4.4 Correlation estimators

The next step in the construction of the proposed multiple testing regularized estimators is to set to zero the statistically insignificant entries of the sample correlation matrix $\hat{\mathbf{\Gamma}} = [\hat{\rho}_{ij}]_{N\times N}$, defined previously. Let $\tilde{p}_\bullet(|\hat{\rho}_{ij}|)$ represent either $\tilde{p}_{\text{SS}}(|\hat{\rho}_{ij}|)$ or $\tilde{p}_{\text{SD}}(|\hat{\rho}_{ij}|)$ computed according to Algorithms 2 and 3, respectively. The corresponding correlation estimator $\hat{\mathbf{\Gamma}}_\bullet = [\hat{\rho}_{\bullet,ij}]_{N\times N}$ has entries given by

$$\hat{\rho}_{\bullet,ij} = \hat{\rho}_{ij}\mathbb{1}\big\{\tilde{p}_\bullet(|\hat{\rho}_{ij}|) \leq \alpha\big\}. \tag{14}$$

These adjustments to the sample correlation matrix are made for $i = 1, ..., N-1$ and $j = i+1, ..., N$; the diagonal elements of $\hat{\mathbf{\Gamma}}_\bullet$ are obviously set as $\hat{\rho}_{i,i} = 1$; and symmetry is imposed by setting $\hat{\rho}_{j,i} = \hat{\rho}_{ij}$. In contrast to the 'universal' threshold value used in (3) for each of the pairwise correlations, note the *adaptive* nature of (14) that makes entrywise threshold adjustments to the correlation matrix.

Proceeding to the shrinkage step in (4) of the BPS approach with $\hat{\mathbf{\Gamma}}_\bullet$ instead of $\hat{\mathbf{\Gamma}}_{\text{BPS}}$ yields the positive definite correlation matrix estimator $\hat{\mathbf{\Gamma}}_\bullet(\xi^*) = \xi^*\mathbf{I}_N + (1-\xi^*)\hat{\mathbf{\Gamma}}_\bullet$, where $\xi^* = \arg\min_{\xi_0+\epsilon\leq\xi\leq 1}\left\|\hat{\mathbf{\Gamma}}_0^{-1} - \hat{\mathbf{\Gamma}}_\bullet^{-1}(\xi)\right\|_F^2$, and the associated covariance matrix estimator $\hat{\mathbf{\Sigma}}_\bullet(\xi^*) = \hat{\mathbf{D}}^{1/2}\hat{\mathbf{\Gamma}}_\bullet(\xi^*)\hat{\mathbf{D}}^{1/2}$.

4.5 Unknown location parameters

When the parameters comprising $\boldsymbol{\theta}$ in (5) are unknown, it will be assumed that they can be estimated consistently at least under H_0. Denote by $\hat{\boldsymbol{\theta}}$ the consistent estimator of $\boldsymbol{\theta}$, and let $\hat{\boldsymbol{\mu}}_t = E(\mathbf{r}_t \,|\, \mathbf{x}_t, \mathcal{I}_{t-1}; \hat{\boldsymbol{\theta}})$. The Monte Carlo procedures proceed as before except that $\hat{\mathbf{y}}_t = \mathbf{r}_t - \hat{\boldsymbol{\mu}}_t$ replaces $\mathbf{y}_t$. Let the correlations estimated on the basis of $\hat{\mathbf{Y}} = [\hat{\mathbf{y}}_1, ..., \hat{\mathbf{y}}_T]'$ be written as $\hat{\rho}_{ij}(\hat{\boldsymbol{\theta}})$.

Proposition 3. *Suppose that (5) holds along with Assumptions 1–2, and that* $\hat{\boldsymbol{\theta}} \xrightarrow{p} \boldsymbol{\theta}$ *under* H_0. *Then the Monte Carlo p-values computed according to Algorithm 1 are asymptotically valid in the sense that* $\Pr(\tilde{p}(|\hat{\rho}_{ij}(\hat{\boldsymbol{\theta}})|) \leq \alpha \,|\, H_0) = \alpha + o_p(1)$, *for* $1 \leq i < j \leq N$, *as* T *increases.*

This proposition follows from Theorem 3 in Toulis and Bean (2021) and its intuition is

obvious: the Monte Carlo p-values are a function of only $\hat{\mathbf{Y}}$, and the consistency of $\hat{\boldsymbol{\theta}}$ implies that $\hat{\mathbf{Y}} \to \mathbf{Y}$ as $T \to \infty$. An immediate corollary is that the SS and SD adjusted p-values computed according to Algorithms 2 and 3 are also asymptotically valid.

5 Simulation studies

This section studies the performance of the proposed Monte Carlo regularized covariance estimators, with the BPS approach serving as the natural benchmark for comparisons. For this purpose, the data-generating process for daily returns $\mathbf{r}_t = (r_{1,t}, ..., r_{N,t})'$ is specified as a CCC model (Bollerslev, 1990) of the form

$$\mathbf{r}_t = \boldsymbol{\mu} + \boldsymbol{\Sigma}_t^{1/2} \mathbf{z}_t,$$
$$\boldsymbol{\Sigma}_t = \mathbf{D}_t^{1/2} \boldsymbol{\Gamma} \mathbf{D}_t^{1/2},$$

where $\mathbf{D}_t = \text{diag}(\sigma_{1,t}^2, ..., \sigma_{N,t}^2)$ is an $N \times N$ diagonal matrix comprising the time-t conditional variances, and $\boldsymbol{\Gamma}$ is a constant conditional correlation (CCC) matrix. The vector $\boldsymbol{\mu}$ is set to zero, but this is not assumed known and the multiple testing procedures are applied with $\mathbf{y}_t = \mathbf{r}_t - \bar{\mathbf{r}}$, where $\bar{\mathbf{r}}$ is the vector of sample means. The conditional variances comprising $\mathbf{D}_t$ evolve according to the standard GARCH model $\sigma_{i,t}^2 = \theta_0 + \theta_1 r_{i,t-1}^2 + \theta_2 \sigma_{i,t-1}^2$ with common parameters across assets set as $\theta_0 = 0.01$, $\theta_1 = 0.1$, and $\theta_2 = 0.85$. These are typical values found with daily returns data. The innovation terms $z_{i,t}$ are i.i.d. according either to a standard normal distribution; or a heavier-tailed t-distribution with 12 or 6 degrees of freedom.

The correlation structure is defined as follows. Given a value $0 \leq \delta \leq 1$, the vector $\mathbf{c} = (c_1, ..., c_N)'$ is filled in with $N_c = \lfloor \delta N \rfloor$ non-zero elements drawn from a uniform distribution $\mathcal{U}(-1, 1)$, and the remaining $N - N_c$ elements are set to zero. The positions of the zero and non-zero elements within $\mathbf{c}$ are random. Following BPS, this vector is then used to obtain a well-defined correlation matrix as

$$\boldsymbol{\Gamma} = \mathbf{I}_N + \mathbf{c}\mathbf{c}' - \text{diag}(\mathbf{c}\mathbf{c}').$$

The complete null hypothesis in (6) is thus represented by $\delta = 0$ and increasing the value of δ towards 1 results in a smaller set $\mathcal{K}_0$ of true hypotheses. With $\mathbf{\Gamma}$ in hand, the unconditional covariance matrix is then found as $\mathbf{\Sigma} = \mathbf{D}^{1/2}\mathbf{\Gamma}\mathbf{D}^{1/2}$, where $\mathbf{D} = \text{diag}(\sigma_1^2, ..., \sigma_N^2)$ with $\sigma_i^2 = \theta_0/(1-\theta_1-\theta_2)$, for $i = 1, ..., N$.

Two choices are used to complete the definition of the universal critical value in (3). First $f(N) = N(N-1)/2$, which corresponds to the Bonferroni rule; and second $f(N) = N^2$, as suggested by BPS. In the tables presented next the results based on these choices are labelled BPS_1 and BPS_2, respectively. The number of Monte Carlo replications for the SS and SD methods is set with $B = 100$ and all the tests are conducted with $\alpha = 0.05$.

5.1 Numerical results

Table 1 reports the empirical probability of rejecting at least one individual hypothesis $H_{i,j}$ when $N = 30$, 100, and 200 assets, and with samples of size $T = 60$, 120, and 240. The empirical FWER in Panels A and B is based on counting occurrences of at least one true $H_{i,j}$ being rejected, whereas the empirical measure of disjunctive power (Bretz et al., 2010, Section 2.1.1) in Panel C is based on counting occurrences of at least one false $H_{i,j}$ being rejected. Under the complete null hypothesis ($\delta = 0$) the SS and SD methods have identical rejection rates, indicated on the lines labelled SS/SD in Panel A. It is seen that the SS (SD) method does a good job at keeping the FWER close to the desired 5% value in that case. These methods are seen in Panel B to also maintain control of the FWER in the strong sense. Comparing Panels A and B shows that the SS and SD methods become more conservative when the number of true null hypotheses diminishes (δ increasing), as expected from the developed theory. The BPS method also achieves good control of the FWER under normality. However when the error terms are non-normal, the BPS approach tends to spuriously over-reject.

Table 1 reveals that the BPS over-rejection problem worsens as: (i) the degree of tail heaviness increases (from t_{12} to t_6 errors), (ii) N increases, and (iii) T increases. The most egregious instance occurs with $T = 240$ where the FWER of BPS_1 and BPS_2 attains 99% when $N = 200$ under t_6 errors. This makes clear that finite-sample FWER control with the BPS thresholding estimator based on $c_\alpha(N)$ in (3) is heavily dependent on normality, even

as the sample size T increases.

Therefore in order to ensure a fair comparison, all the power results for the BPS approach are based on FWER-adjusted critical values. Panel C of Table 1 shows the power of the three multiple testing procedures when $\delta = 0.1$. Note that BPS_1 and BPS_2 have identical size-adjusted power, reported on the lines labelled BPS. The results in Table 1 show that the disjunctive power of SS (SD) is quite close to that of BPS. As expected power increases with T and N, and decreases as tail heaviness increases.

The performance of the multiple testing methods is further examined by computing the true positive rate, defined as

$$\text{TPR}(\hat{\mathbf{\Sigma}}, \mathbf{\Sigma}) = \frac{\sum_{i=1}^{N-1} \sum_{j=i+1}^{N} \mathbb{1}\{\hat{\sigma}_{ij} \neq 0 \text{ and } \sigma_{ij} \neq 0\}}{\sum_{i=1}^{N-1} \sum_{j=i+1}^{N} \mathbb{1}\{\sigma_{ij} \neq 0\}}.$$

Table 2 reports the average TPR in percentage when $\delta = 0.1$, 0.5, and 0.9. Specifically, the lines labelled BPS report $\text{TPR}(\hat{\mathbf{\Sigma}}_{\text{BPS}}(\xi^*), \mathbf{\Sigma})$, while lines labelled SS and SD report $\text{TPR}(\hat{\mathbf{\Sigma}}_{\text{SS}}(\xi^*), \mathbf{\Sigma})$ and $\text{TPR}(\hat{\mathbf{\Sigma}}_{\text{SD}}(\xi^*), \mathbf{\Sigma})$, respectively. Obviously, the value of ξ^* is recomputed in each case. As expected Table 2 reveals that, all else equal, the TPR performance generally improves as: (i) tail heaviness decreases towards normality; (ii) δ increases away from zero; (iii) T increases; and (iv) N decreases. The TPR with the SS method is lower than with the BPS and SD methods. Observe that the TPRs achieved with SD and BPS are very close indeed, and that SD tends to deliver a relatively better TPR as tail heaviness increases.

Table 3 reports the average Frobenius norm of the matrix losses $\Delta(\hat{\mathbf{\Sigma}}, \mathbf{\Sigma}) = \hat{\mathbf{\Sigma}} - \mathbf{\Sigma}$ for the covariance matrix estimators obtained from the (FWER-adjusted) BPS method, and from the SS and SD methods. It is immediately clear that these three methods result in similar losses. To further the comparisons, Table 4 reports the Frobenius norm losses for the sample covariance matrix and two shrinkage covariance matrix estimators. The latter are based on the linear shrinkage (LS) method of Ledoit and Wolf (2004) that shrinks the sample covariance matrix towards the identity matrix, and the second one uses the non-linear shrinkage (NLS) estimator proposed by Ledoit and Wolf (2015).[6]

[6]Specifically, the Ledoit-Wolf shrinkage covariance matrix estimates are computed with the

Comparing Tables 3 and 4 reveals that the NLS method results in the smallest losses when the innovations are Gaussian. Under the heavier-tailed t_{12} and t_6 distributions, however, the LS method tends to do better at least when $T = 60$ (Panel A) and when $\delta = 0.1$ (i.e., when there are few non-zero correlations). As δ increases to 0.5 and 0.9, the multiple testing regularized covariance matrix estimators (BPS, SS, SD) tend to yield the smallest norm losses. This becomes particularly apparent for larger values of the (N, T) pair in Panels B and C of Tables 2 and 3. Of course, the FWER-adjusted BPS method is not feasible in practice. It is merely used here as a benchmark for the SS and SD methods.

6 Application to portfolio optimization

The proposed multiple testing regularization method is further illustrated in this section with an application to portfolio optimization using the $N = 100$ holdings of the Invesco S&P 500 Low Volatility ETF (fund ticker: SPLV) as of May 28, 2021. Table A1 in the Appendix lists these 100 holdings of SPLV. This ETF invests at least 90% of its total assets in the securities that comprise the S&P 500 Low Volatility Index, which consists of the 100 securities from the S&P 500 with the lowest realized volatility over the past 12 months. Volatility is defined as the standard deviation of the security's daily price returns, in local currency, over the prior one year of trading days. The index constituents are weighted relative to the inverse of their corresponding volatility, with the least volatile stocks receiving the highest weights.

Price data on the SPLV securities were obtained from Yahoo Finance and returns were computed as $r_{i,t} = (p_{i,t} - p_{i,t-1})/p_{i,t-1}$, for $i = 1, ..., N$, where $p_{i,t}$ is the adjusted (for splits and dividend and/or capital gain distributions) closing price for asset i on day t. The resulting time series of $T = 2095$ daily returns cover the period from February 4, 2013 to May 28, 2021. Table A1 gives the mean and standard deviation of the return series for each asset over the full sample period. Although the sample correlation matrix is too big to be shown, it is sufficient to note that none of its entries is negative.

The S&P 500 Low Volatility Index does not consider correlation among stocks, which means that SPLV is a basket of low-volatility securities, not a global minimum variance

`linshrink_cov` and `nlshrink_cov` commands available with the R 'nlshrink' package.

(GMV) portfolio. Consider an investor whose objective is indeed to achieve a GMV portfolio with the 100 SPLV holdings. The GMV portfolio selection problem is well suited for the evaluation of covariance matrix estimators since it does not depend on expected returns. Any other point on the efficient frontier would put some emphasis on forecasts of expected returns (Chan et al., 1999).

To state the problem more formally, let t_{b} refer to the day when the portfolio is initially formed and subsequent rebalancing days. On those days, the investor uses the returns over the past L days to estimate the covariance matrix. In the simplest case, this estimate is given by the rolling-window sample covariance matrix

$$\hat{\boldsymbol{\Sigma}}_{t_{\mathrm{b}}} = \frac{1}{L} \sum_{t=t_{\mathrm{b}}-L+1}^{t_{\mathrm{b}}} (\mathbf{r}_t - \bar{\mathbf{r}}_t)(\mathbf{r}_t - \bar{\mathbf{r}}_t)', \tag{15}$$

where $\bar{\mathbf{r}}_t = L^{-1} \sum_{t=t_{\mathrm{b}}-L+1}^{t_{\mathrm{b}}} \mathbf{r}_t$. The investor then uses the covariance matrix estimate to find the GMV portfolio. This portfolio is found by solving the problem

$$\begin{aligned} &\min_{\boldsymbol{\omega}} \ \boldsymbol{\omega}'\boldsymbol{\Sigma}\boldsymbol{\omega} \\ &\text{subject to } \ \boldsymbol{\iota}'\boldsymbol{\omega} = 1, \end{aligned} \tag{16}$$

with the assignment $\boldsymbol{\Sigma} \leftarrow \hat{\boldsymbol{\Sigma}}_{t_{\mathrm{b}}}$, and where $\boldsymbol{\iota}$ denotes an $N \times 1$ vector of ones. The analytical solution to this problem yields the GMV portfolio weights $\hat{\boldsymbol{\omega}}_{t_{\mathrm{b}}} = (\omega_{1,t_{\mathrm{b}}}, ..., \omega_{N,t_{\mathrm{b}}})' = (\boldsymbol{\iota}'\hat{\boldsymbol{\Sigma}}_{t_{\mathrm{b}}}^{-1}\boldsymbol{\iota})^{-1}\boldsymbol{\iota}'\hat{\boldsymbol{\Sigma}}_{t_{\mathrm{b}}}^{-1}$. The initialization $t_{\mathrm{b}} \leftarrow L$ references the day when the portfolio is first formed. This portfolio is then held for H days and the resulting realized out-of-sample portfolio returns are $\hat{\boldsymbol{\omega}}_{t_{\mathrm{b}}}'\mathbf{r}_\tau$, for $\tau = t_{\mathrm{b}} + 1, ..., t_{\mathrm{b}} + H$. After its initial formation, the portfolio is rebalanced on days $t_{\mathrm{b}} \leftarrow t_{\mathrm{b}} + H$. Rebalancing consists of finding new GMV weights by solving (16) with the updated covariance matrix estimate.

The sample covariance matrix in (15) is a natural benchmark to compare results obtained with the multiple testing regularization approaches, namely BPS_1, BPS_2, SS, and SD, which proceed by testing the significance of the 4950 distinct covariances in the rolling-window scheme.[7] Those approaches are further compared to the portfolio optimization results ob-

[7]The nominal FWER is set to $\alpha = 5\%$ and $B - 1 = 999$ resampling draws are used in the computation of the Monte Carlo p-values.

tained by two direct shrinkage estimators. The first is the linear method of Ledoit and Wolf (2004) that shrinks the sample covariance matrix towards the identity matrix, and the second one uses the non-linear shrinkage estimator proposed by Ledoit and Wolf (2015).[8] Following the rolling-window scheme in (15), all the considered covariance matrix estimators are based on the past L days of (demeaned) returns whenever the portfolio weights are computed. Finally the performance evaluation also includes the equally weighted portfolio, which bypasses (16) altogether and simply sets $\hat{\omega}_{i,t_\mathrm{b}} = 1/N$. This naive portfolio strategy is a standard benchmark for comparisons; see DeMiguel et al. (2009) and Kirby and Ostdiek (2012), among others.

Note that there is no estimation risk associated with the naive diversification strategy, which helps reduce the portfolio turnover (i.e., the fraction of invested wealth that is traded in a given period). On the contrary, active stratgeies that generate high turnover will suffer more in the presence of transaction costs. To see the impact of transaction costs, note that for every dollar invested in the portfolio at time t_b, there are $\hat{\omega}_{i,t_\mathrm{b}} \prod_{\tau=t_\mathrm{b}+1}^{t}(1+r_{i,\tau})$ dollars invested in asset i at time t, for $t > t_\mathrm{b}$ and as long as the portfolio is not rebalanced. Hence, at any time t until the the next rebalancing occurs, the actual weight of asset i in the portfolio is

$$\omega_{i,t}^{*} = \frac{\hat{\omega}_{i,t_\mathrm{b}} \prod_{\tau=t_\mathrm{b}+1}^{t}(1+r_{i,\tau})}{\sum_{i=1}^{N} \hat{\omega}_{i,t_\mathrm{b}} \prod_{\tau=t_\mathrm{b}+1}^{t}(1+r_{i,\tau})}.$$

When rebalancing occurs, the portfolio turnover can be defined as $\mathrm{TO}_t = \sum_{i=1}^{N} |\hat{\omega}_{i,t} - \omega_{i,t}^{*}|$, where $\hat{\omega}_{i,t}$ is the updated weight for asset i at rebalancing time $t = t_\mathrm{b}$. Denoting by κ the transaction cost in proportion to the amount of wealth invested, the proportional cost of rebalancing all the portfolio positions is $\mathrm{TC}_t = \kappa \mathrm{TO}_t$ when $t = t_\mathrm{b}$. Therefore, starting with $t_\mathrm{b} \leftarrow L$ and a normalized inital wealth of $W_{t_\mathrm{b}} = 1$ on that first day, wealth subsequently evolves according to

$$W_{t+1} = \begin{cases} W_t \left(1 + \sum_{i=1}^{N} \hat{\omega}_{i,t} r_{i,t+1}\right) (1 - \mathrm{TC}_t), & \text{when } t = t_\mathrm{b}, \\ W_t \left(1 + \sum_{i=1}^{N} \omega_{i,t}^{*} r_{i,t+1}\right), & \text{when } t \neq t_\mathrm{b}, \end{cases}$$

[8] Specifically, the Ledoit-Wolf shrinkage covariance matrix estimates are computed with the `linshrink_cov` and `nlshrink_cov` commands available with the R 'nlshrink' package.

for $t = L, ..., T-1$, with the updating rule $t_{\text{b}} \leftarrow t_{\text{b}} + H$ to determine the rebalancing days. The out-of-sample portfolio return net of transaction costs is then given by $(W_{t+1} - W_t)/W_t$. To see the impact of transaction costs, results are presented with κ first set to zero and then to 25 basis points (bps).[9]

For each portfolio strategy, the resulting out-of-sample daily returns are used to compute: (i) the mean return (annualized by multiplying by 252); (ii) the standard deviation of returns (annualized by multiplying by $\sqrt{252}$); and (iii) the information ratio (IR) given as the annualized mean divided by the annualized standard deviation. As Engle et al. (2019) argue, the most important performance measure in the context of GMV portfolios is the out-of-sample standard deviation. Indeed, the GMV portfolio is meant to minimize the portfolio variance and thus the standard deviation. Any portfolio pursuing the GMV objective should therefore be evaluated primarily by the standard deviation it achieves out of sample. While a high mean and IR are naturally desirable, they should be considered of secondary importance when judging a covariance matrix estimator. In addition to the active strategies that involve rebalancing the portfolio periodically, the performance measures are also computed for the passive investment strategy in which the investor simply buys and holds shares of the SPLV exchange traded fund.

Tables 5 and 6 present the results with $\kappa = 0$ and $\kappa = 25$ bps, respectively. When $N > L$, the sample covariance matrix is singular. Those cases are indicated with n/a in the tables. Note also that the performance of the passive SPLV strategy does not depend on H, but it does depend on L since the first day of the trading period is set equal to February 4, 2013 + $(L-1)$ trading days. The main findings with respect to the out-of-sample standard deviation are summarized as follows:

1. When $L = 60$ the NLS portfolio achieves the lowest standard deviation for all considered holding periods, H. Even as the estimation window grows to $L = 120$, 240, the NLS portfolio tends to perform well when $H = 60$ and when there are no transaction costs (Table 5).

[9] French (2008) estimates the cost of trading stocks listed on the NYSE, AMEX, and NASDAQ, including total commissions, bid-ask spreads, and other costs investors pay for trading services. He finds that these costs have dropped significantly over time "from 146 basis points in 1980 to a tiny 11 basis points in 2006."

2. As the length of the estimation window L grows to 120 and 240, the SS and SD portfolios are seen to perform better. In particular, when $L = 240$ (third column) the SS portfolio delivers the smallest out-of-sample standard deviation for holding periods $H = 120$, 240, both in the absence of transactions costs (Table 5) and in the presence of proportional transactions costs of $\kappa = 25$ bps (Table 6).

Figures 1 (for $\kappa = 0$) and 2 (for $\kappa = 25$ bps) show the growth of 1 dollar invested according to the various portfolio strategies over the trading period from January 15, 2014 to May 28, 2021. The depicted scenarios in those figures correspond, respectively, to the cases in Tables 4 and 5 where $L = 240$, $H = 120$ trading days. The GMV portfolio using the SD covariance matrix (solid black line in each subfigure) is seen to yield higher growth in comparison to the other portfolio strategies. A comparison of Figures 1 and 2 shows the eroding effects of transactions costs on wealth growth. For instance, in the absence of transaction costs the wealth growth in Figure 1 (c) from the SD strategy clearly outpaces the wealth growth from naive portfolio strategy. The introduction of transactions costs is seen in Figure 2 (c) to shrink the gap between those two wealth growth paths.

The erosion of wealth growth seen in Figure 2 is primarily caused by turnover. Figure 3 shows the amount of portfolio turnover (defined previously as TO_t) generated by the various strategies, given that the estimation window is of length $L = 240$ and the holding period is $H = 120$ trading days, as before. As expected, the naive portfolio is seen in Figure 3 (c) to have near zero turnover After that, the next lowest turnover is obtained from the SS strategy (solid black line in each subfigure). Observe in Figure 3 that the turnover from the SD strategy is the one that remains closest to that of the SS strategy. The next lowest turnover comes from BPS, while all the other portfolio strategies generate greater turnover.

Further insight into these results can be gleaned from Figure 4, which shows the number of statistically significant covariances (out of 4950) detected by the multiple testing procedures each time the portfolio is rebalanced. The solid black line in Figure 4 shows that the SS procedure sets to zero the greatest number of covariances, which explains the low standard deviations seen in Tables 4 and 5 when $L = 240$ and $H = 120$, 240. In contrast, the BPS approach rejects the null hypothesis far more often and thus declares the greatest number of non-zero covariances. Given the over-rejection problem seen in Table 1 with the BPS

approach, it is quite likely that many of these rejections are spurious in nature. Recall that in comparison to the SS adjustments, the SD approach improves power while retaining the same FWER protection. This makes SD a "Goldilocks" solution, whose number of declared non-zero covariances appears in Figure 4 between SS and BPS. Striking this balance results in the greater wealth growth seen in Figures 1 and 2.

Finally note that Figure 4 provides a gauge of the portfolio's overall diversification profile, which improves when there are fewer significant correlations among these assets that generally tend to move in the same direction.

7 Concluding remarks

This paper has developed a Monte Carlo resampling method to regularize stock return covariance matrices. Following BPS, the method begins by testing the significance of the pairwise correlations and then sets to zero the sample correlations whose multiplicity adjusted p-values fall above the specified threshold level. A subsequent shrinkage step ensures that the final covariance matrix estimate is positive definite and well conditioned, while preserving the zeros achieved by thresholding.

The multiple testing procedures developed in this paper have two important advantages compared to the BPS approach. First, they achieve control of the traditional FWER, which is defined as the probability of falsely rejecting one or more true null hypotheses. When the conditional location is known, the FWER is controlled exactly in finite samples under any partial configuration of true and false null hypotheses. This strong control of the FWER holds asymptotically when the conditional location parameters are consistently estimated.

The second advantage is that the proposed resampling approach allows for the presence of heavy tails and multivariate GARCH-type effects, which are prominent features of stock returns. Indeed the Monte Carlo resampling scheme proceeds conditional on the absolute values of the error terms, since only their signs are randomized. The Lehmann and Stein (1949) impossibility theorem shows that such sign-based tests are the *only* ones that yield valid inference in the presence of non-normalities and heteroskedasticity of unknown form; see also Dufour (2003) for more on this point.

In exploratory research, control of the FWER may be too stringent. That is, if it is only necessary to have no more than a certain number or a certain proportion of errors, then the use of the FWER may not be appropriate. In such cases, the FWER can be replaced by the k-FWER, defined in Lehmann and Romano (2005) as the probability of k or more false rejections. Another proposal is to control the false discovery proportion (FDP), defined as the number of false rejections divided by the total number of rejections. (If no hypotheses are rejected, then the FDP is defined to be 0.) The false discovery rate (FDR) proposed by Benjamini and Hochberg (1995) corresponds to the expected value of FDP. Following the approaches in Korn et al. (2004) for the k-FWER and FDP and in Ge et al. (2008) for the FDR, the method developed in this paper can be extended to control these alternatives to the traditional FWER.

Table 1. Empirical probability of rejecting at least one $H_{i,j}$

	Normal			t_{12}			t_6		
	$N = 30$	100	200	$N = 30$	100	200	$N = 30$	100	200
Panel A: FWER ($\delta = 0$)									
$T = 60$									
BPS_1	3.4	1.5	0.5	9.7	9.7	8.9	36.8	48.4	49.7
BPS_2	1.5	0.6	0.1	5.3	5.0	4.4	25.8	34.5	37.2
SS/SD	6.6	6.8	5.4	5.4	6.7	5.9	5.3	6.2	5.8
$T = 120$									
BPS_1	3.0	3.4	3.0	18.5	25.3	34.1	62.4	84.5	92.0
BPS_2	1.1	1.3	0.9	11.1	15.6	23.1	50.3	74.2	84.7
SS/SD	4.6	5.7	5.8	4.7	5.2	5.4	4.5	6.0	5.1
$T = 240$									
BPS_1	5.3	5.8	4.4	25.8	43.8	59.6	78.1	97.4	99.8
BPS_2	2.7	3.4	2.6	15.9	30.1	46.0	68.4	93.9	99.0
SS/SD	6.5	6.3	5.6	5.1	4.4	5.9	5.5	4.6	6.3
Panel B: FWER ($\delta = 0.5$)									
$T = 60$									
BPS_1	1.9	0.9	0.4	7.2	6.3	5.9	31.4	37.6	39.0
BPS_2	0.7	0.6	0.1	3.7	3.7	2.3	22.1	26.3	28.3
SS	0.6	0.9	0.3	2.3	1.1	0.8	2.7	2.5	1.8
SD	3.6	3.4	2.3	4.0	3.7	3.1	4.1	4.2	4.2
$T = 120$									
BPS_1	2.8	1.8	1.4	14.3	19.9	26.1	52.2	74.1	85.0
BPS_2	1.2	1.1	0.6	8.4	13.2	16.8	39.2	64.6	76.2
SS	1.2	0.1	0.0	1.8	0.8	0.8	2.0	2.6	1.8
SD	3.8	3.8	3.0	3.9	3.2	4.2	3.6	5.1	3.7
$T = 240$									
BPS_1	2.7	3.8	3.9	20.2	35.6	44.2	69.0	93.1	98.7
BPS_2	1.4	1.3	1.8	13.1	24.9	31.8	58.3	88.0	96.6
SS	1.2	0.1	0.2	1.4	0.7	0.6	2.0	2.1	1.8
SD	3.6	4.5	3.4	4.4	3.8	3.2	3.5	3.8	4.7
Panel C: Power ($\delta = 0.1$)									
$T = 60$									
BPS	40.7	92.0	99.6	39.0	87.8	98.3	26.9	76.1	92.3
SS/SD	42.6	92.2	99.6	39.8	88.6	98.8	31.6	81.4	95.8
$T = 120$									
BPS	58.0	97.9	100.0	48.4	96.2	99.9	34.6	86.9	98.4
SS/SD	56.8	98.3	100.0	48.6	96.4	99.9	39.6	91.1	99.2
$T = 240$									
BPS	72.2	99.6	100.0	64.1	99.3	100.0	46.7	93.2	99.4
SS/SD	72.4	99.6	100.0	65.8	99.2	100.0	51.2	95.9	99.7

Notes: Panels A and B report the empirical FWER of the multiple testing procedures, while Panel C reports their power, in percentages given a nominal FWER $\alpha = 5\%$. The power results of BPS_1 and BPS_2, reported on the lines labelled BPS, are based on FWER-adjusted critical values.

Table 2. True positive rates

	Normal			t_{12}			t_6		
	$\delta = 0.1$	0.5	0.9	$\delta = 0.1$	0.5	0.9	$\delta = 0.1$	0.5	0.9
Panel A: $T = 60$									
$N = 30$									
BPS	16.8	17.3	17.5	16.0	15.5	15.5	9.9	9.8	10.1
SS	17.3	15.3	13.1	16.3	13.8	11.7	11.9	10.3	9.2
SD	17.5	17.6	17.4	16.5	15.9	15.6	12.0	11.7	11.8
$N = 100$									
BPS	13.8	13.5	13.7	10.8	10.7	10.7	6.4	6.4	6.3
SS	13.5	10.1	8.3	11.3	8.7	7.2	8.3	6.9	5.8
SD	13.9	13.2	12.7	11.5	11.1	10.5	8.6	8.3	8.0
$N = 200$									
BPS	12.3	12.1	12.2	9.0	8.9	8.8	4.5	4.4	4.4
SS	11.6	8.1	6.6	9.3	6.8	5.5	6.7	5.2	4.3
SD	12.2	11.5	10.7	9.7	9.2	8.6	6.9	6.5	6.3
Panel B: $T = 120$									
$N = 30$									
BPS	29.2	29.6	29.5	22.2	24.3	24.5	14.5	16.1	16.1
SS	28.3	25.7	22.8	22.3	21.7	19.5	17.3	16.9	15.2
SD	28.5	29.1	29.0	22.4	24.3	24.7	17.4	18.8	19.1
$N = 100$									
BPS	23.8	23.9	24.0	18.4	18.6	18.9	9.8	9.8	9.6
SS	23.7	19.2	16.6	18.8	15.9	14.0	13.3	11.7	10.5
SD	24.1	24.0	23.7	19.0	19.2	19.4	13.6	13.6	13.5
$N = 200$									
BPS	22.2	22.0	21.9	16.4	16.2	16.2	7.7	7.6	7.7
SS	21.6	16.3	14.1	16.8	13.4	11.7	11.1	9.4	8.5
SD	22.3	21.8	21.3	17.3	17.0	16.9	11.3	11.1	11.2
Panel C: $T = 240$									
$N = 30$									
BPS	41.5	40.9	40.9	34.4	33.6	33.4	21.3	21.7	21.9
SS	41.5	37.5	34.0	35.2	31.4	28.4	24.3	22.8	21.3
SD	41.6	41.2	41.6	35.3	34.8	35.2	24.5	25.1	26.0
$N = 100$									
BPS	34.3	35.3	35.1	29.0	28.8	28.3	14.5	14.6	14.6
SS	34.2	29.9	26.9	28.8	25.1	22.3	18.2	16.5	15.3
SD	34.7	35.6	35.6	29.4	29.5	29.2	18.5	18.8	19.1
$N = 200$									
BPS	33.2	33.2	33.3	24.9	24.7	24.8	12.0	12.1	12.0
SS	32.2	26.2	23.9	25.4	21.3	19.5	15.1	13.8	12.8
SD	33.2	33.1	33.2	26.0	25.9	26.1	15.4	15.8	15.9

Notes: This table reports the true positive rates for $\mathbf{\Sigma}$, in percentages, achieved by the multiple testing regularized covariance matrix estimators. The results for BPS are based on FWER-adjusted critical values.

Table 3. Frobenius norm losses of multiple testing regularized covariance matrix estimators

	Normal			t_{12}			t_6		
	$\delta = 0.1$	0.5	0.9	$\delta = 0.1$	0.5	0.9	$\delta = 0.1$	0.5	0.9
Panel A: $T = 60$									
$N = 30$									
BPS	0.50	0.85	1.41	1.31	1.50	1.90	5.61	5.73	6.08
SS	0.50	0.87	1.48	1.31	1.51	1.94	5.60	5.72	6.02
SD	0.50	0.85	1.41	1.31	1.50	1.89	5.60	5.74	6.05
$N = 100$									
BPS	1.05	2.95	5.38	2.60	3.81	5.85	10.88	11.05	12.13
SS	1.05	3.03	5.55	2.60	3.86	6.01	10.88	11.02	12.12
SD	1.05	2.95	5.41	2.60	3.80	5.86	10.88	11.01	12.05
$N = 200$									
BPS	1.71	6.23	11.37	3.95	7.21	11.83	16.12	17.22	20.71
SS	1.71	6.35	11.58	3.95	7.31	12.03	16.10	17.18	20.73
SD	1.71	6.25	11.43	3.94	7.21	11.85	16.10	17.12	20.58
Panel B: $T = 120$									
$N = 30$									
BPS	0.41	0.68	1.15	1.53	1.66	1.86	10.65	10.79	10.83
SS	0.41	0.72	1.27	1.52	1.65	1.90	10.64	10.77	10.73
SD	0.41	0.69	1.16	1.52	1.65	1.86	10.64	10.82	10.84
$N = 100$									
BPS	0.85	2.67	5.11	2.84	3.76	5.62	24.72	24.16	23.63
SS	0.85	2.79	5.32	2.84	3.81	5.81	24.72	24.10	23.55
SD	0.85	2.67	5.12	2.84	3.74	5.60	24.72	24.11	23.51
$N = 200$									
BPS	1.45	5.91	11.06	4.28	7.12	11.53	37.20	36.13	38.02
SS	1.45	6.07	11.31	4.28	7.22	11.78	37.15	36.05	37.92
SD	1.45	5.92	11.08	4.28	7.09	11.50	37.14	36.04	37.84
Panel C: $T = 240$									
$N = 30$									
BPS	0.31	0.50	0.89	1.47	1.67	1.77	22.69	21.91	20.13
SS	0.31	0.52	1.03	1.47	1.64	1.77	22.67	21.85	19.90
SD	0.31	0.49	0.87	1.47	1.67	1.77	22.67	21.95	20.33
$N = 100$									
BPS	0.63	2.34	4.77	2.86	3.50	5.21	61.07	58.71	60.64
SS	0.63	2.49	5.05	2.86	3.58	5.46	61.06	58.61	60.51
SD	0.63	2.32	4.75	2.86	3.48	5.16	61.06	58.77	60.56
$N = 200$									
BPS	1.13	5.53	10.72	4.17	6.65	11.08	154.59	156.26	253.76
SS	1.14	5.77	11.06	4.16	6.79	11.37	154.58	156.18	252.66
SD	1.13	5.53	10.72	4.16	6.60	11.01	154.59	156.25	253.63

Notes: The results for BPS are based on FWER-adjusted critical values.

Table 4. Frobenius norm losses of sample and shrinkage covariance matrix estimators

	Normal			t_{12}			t_6		
	$\delta = 0.1$	0.5	0.9	$\delta = 0.1$	0.5	0.9	$\delta = 0.1$	0.5	0.9
Panel A: $T = 60$									
$N = 30$									
Sample	0.91	0.97	1.10	1.93	2.08	2.35	6.92	7.42	8.74
LS	0.30	0.74	1.02	0.99	1.41	1.89	4.69	5.42	6.96
NLS	0.32	0.67	0.95	1.19	1.55	1.99	5.75	6.42	7.97
$N = 100$									
Sample	2.77	3.00	3.45	5.42	6.03	7.22	17.28	19.44	23.95
LS	0.70	2.37	3.25	1.84	3.82	5.72	8.20	11.69	17.63
NLS	0.68	1.92	2.86	2.42	3.97	5.82	12.14	15.31	20.87
$N = 200$									
Sample	5.39	5.84	6.85	10.34	11.78	14.14	31.42	35.66	44.37
LS	1.34	4.72	6.46	2.80	7.36	11.20	11.35	19.19	31.30
NLS	1.20	3.68	5.64	3.99	7.51	11.31	20.18	26.73	37.86
Panel B: $T = 120$									
$N = 30$									
Sample	0.68	0.73	0.84	1.90	2.04	2.31	11.68	12.49	14.22
LS	0.27	0.62	0.81	1.27	1.61	2.01	9.52	10.48	12.38
NLS	0.29	0.55	0.75	1.44	1.71	2.08	10.84	11.76	13.64
$N = 100$									
Sample	1.98	2.20	2.64	4.67	5.48	7.07	30.14	33.64	39.92
LS	0.65	1.96	2.57	2.18	4.01	6.07	20.92	25.44	33.03
NLS	0.60	1.53	2.28	2.67	4.12	6.17	26.03	30.26	37.40
$N = 200$									
Sample	3.84	4.31	5.13	8.72	10.74	13.80	51.21	60.58	78.33
LS	1.27	3.83	5.02	3.26	7.68	11.69	29.62	42.23	63.43
NLS	1.05	2.90	4.39	4.27	7.81	11.90	41.43	53.01	72.97
Panel C: $T = 240$									
$N = 30$									
Sample	0.49	0.53	0.62	1.70	1.90	2.25	23.44	23.34	23.92
LS	0.22	0.49	0.61	1.31	1.64	2.06	21.29	21.31	22.05
NLS	0.23	0.41	0.56	1.42	1.70	2.11	22.85	22.82	23.51
$N = 100$									
Sample	1.41	1.59	1.90	4.01	4.98	6.62	65.23	69.39	82.73
LS	0.59	1.51	1.89	2.39	4.08	5.98	55.70	61.17	74.99
NLS	0.49	1.15	1.68	2.74	4.13	6.07	62.16	66.85	80.78
$N = 200$									
Sample	2.72	3.08	3.73	7.06	9.33	12.81	166.05	185.98	485.87
LS	1.17	2.93	3.72	3.46	7.43	11.49	140.13	162.52	456.15
NLS	0.83	2.17	3.26	4.16	7.49	11.66	158.43	179.88	480.35

Table 5. Annualized portfolio performance measures with no transaction costs

	$L = 60$			$L = 120$			$L = 240$		
	$H = 60$	120	240	$H = 60$	120	240	$H = 60$	120	240
Sample									
Mean	n/a	n/a	n/a	10.93	4.90	16.59	11.68	11.25	13.34
Std dev.	n/a	n/a	n/a	25.72	32.72	29.14	15.88	18.26	19.01
IR	n/a	n/a	n/a	0.42	0.14	0.56	0.73	0.61	0.70
SS									
Mean	16.38	15.79	15.28	17.96	16.11	14.86	16.19	15.67	15.75
Std dev.	14.44	14.74	14.73	14.55	14.66	14.72	14.51	**14.32**	**14.30**
IR	1.13	1.07	1.03	1.23	1.09	1.00	1.11	1.09	1.10
SD									
Mean	15.82	15.27	15.27	16.46	15.98	14.18	17.01	18.44	17.88
Std dev.	14.49	14.80	14.85	14.85	**14.55**	**14.13**	14.74	14.61	14.94
IR	1.09	1.03	1.02	1.10	1.09	1.00	1.15	1.26	1.19
BPS_1									
Mean	18.44	17.72	16.02	14.57	15.01	14.04	15.57	16.25	17.02
Std dev.	14.59	15.03	14.91	15.22	14.96	15.32	14.87	14.90	15.19
IR	1.26	1.17	1.07	0.95	1.00	0.91	1.04	1.09	1.12
BPS_2									
Mean	18.21	17.23	15.85	14.47	14.95	13.97	15.59	16.90	17.04
Std dev.	14.34	14.91	14.80	15.40	15.26	15.61	14.82	14.68	15.05
IR	1.27	1.15	1.07	0.93	0.97	0.89	1.05	1.15	1.13
LS									
Mean	16.80	14.84	14.50	14.97	13.01	14.38	13.78	13.33	12.80
Std dev.	13.89	14.37	14.44	14.78	16.14	15.24	14.61	15.23	16.00
IR	1.20	1.03	1.00	1.01	0.80	0.94	0.94	0.87	0.79
NLS									
Mean	15.04	14.45	14.27	14.75	14.52	13.86	14.03	13.77	13.40
Std dev.	**13.74**	**14.16**	**14.21**	**14.45**	15.39	14.99	**14.41**	15.05	15.75
IR	1.09	1.02	1.00	1.02	0.94	0.92	0.97	0.91	0.85
Naive									
Mean	16.18	16.05	15.99	16.13	16.01	15.97	16.18	16.04	16.01
Std dev.	14.55	14.55	14.56	14.62	14.62	14.67	14.88	14.88	14.87
IR	1.11	1.10	1.09	1.10	1.09	1.08	1.08	1.07	1.07
SPLV	Mean = 11.45 Std dev. = 15.93 IR = 0.71			Mean = 11.63 Std dev. = 16.01 IR = 0.72			Mean = 11.94 Std dev. = 16.32 IR = 0.73		

Notes: This table reports annualized mean, standard deviation, and information ratio (IR) of the out-of-sample returns (in percentage) achieved by various rolling-window portfolio strategies. L is the length of the estimation window, while H is the length of the holding period until the next rebalancing. The trading period is from February 4, 2013 + $(L - 1)$ trading days to May 28, 2021. Bold entries indicate the lowest standard deviation achieved, given L and H. The SPLV results do not depend on H, since they correspond to a passive buy-and-hold strategy.

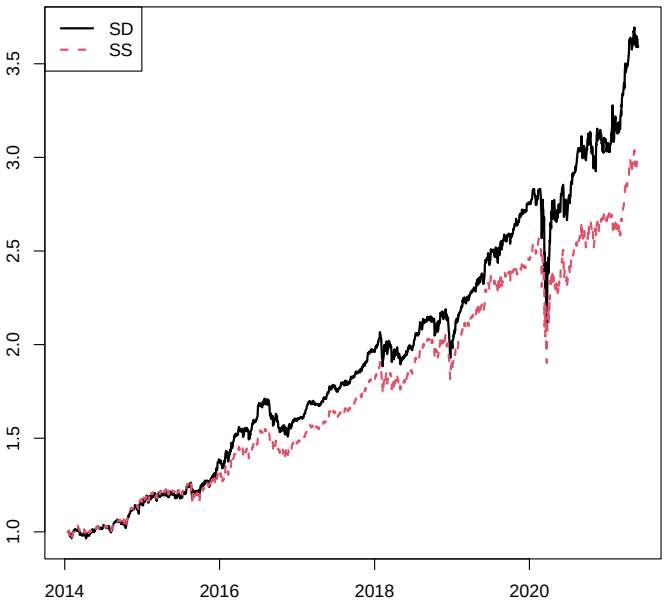
Table 6. Annualized portfolio performance measures with proportional transaction costs of 25 bps

	$L = 60$			$L = 120$			$L = 240$		
	$H = 60$	120	240	$H = 60$	120	240	$H = 60$	120	240
Sample									
Mean	n/a	n/a	n/a	-7.95	-6.64	10.73	6.17	7.31	11.00
Std dev.	n/a	n/a	n/a	27.45	33.56	29.41	15.98	18.29	19.12
IR	n/a	n/a	n/a	-0.28	-0.19	0.36	0.38	0.40	0.57
SS									
Mean	15.04	15.12	14.92	16.34	15.14	14.24	14.78	14.69	15.16
Std dev.	14.47	14.75	14.73	14.58	14.68	14.70	14.52	**14.32**	**14.33**
IR	1.03	1.02	1.01	1.12	1.03	0.96	1.01	1.02	1.05
SD									
Mean	14.09	14.41	14.83	14.20	14.68	13.43	15.11	17.13	17.09
Std dev.	14.53	14.82	14.86	14.91	**14.56**	**14.12**	14.77	14.64	14.98
IR	0.96	0.97	0.99	0.95	1.00	0.95	1.02	1.17	1.14
BPS_1									
Mean	16.04	16.57	15.46	11.80	13.25	13.03	13.12	14.66	16.13
Std dev.	14.65	15.06	14.92	15.28	15.01	15.33	14.89	14.91	15.23
IR	1.09	1.09	1.03	0.77	0.88	0.85	0.88	0.98	1.05
BPS_2									
Mean	16.03	16.15	15.37	11.81	13.24	12.96	13.22	15.36	16.15
Std dev.	14.39	14.94	14.81	15.46	15.31	15.62	14.84	14.70	15.09
IR	1.11	1.08	1.03	0.76	0.86	0.82	0.89	1.04	1.07
LS									
Mean	12.35	12.54	13.26	10.81	10.10	12.81	10.73	11.05	11.30
Std dev.	14.03	14.47	14.48	14.88	16.21	15.26	14.64	15.24	16.07
IR	0.88	0.86	0.91	0.72	0.62	0.83	0.73	0.72	0.70
NLS									
Mean	11.63	12.76	13.39	11.40	12.05	12.54	11.48	11.84	12.12
Std dev.	**13.80**	**14.22**	**14.23**	**14.52**	15.43	14.98	**14.42**	15.04	15.81
IR	0.84	0.89	0.94	0.78	0.78	0.83	0.79	0.78	0.76
Naive									
Mean	16.09	15.98	15.93	16.04	15.94	15.91	16.09	15.97	15.95
Std dev.	14.56	14.55	14.57	14.62	14.62	14.67	14.88	14.88	14.87
IR	1.10	1.09	1.09	1.09	1.08	1.08	1.08	1.07	1.07
SPLV	Mean = 11.45 Std dev. = 15.93 IR = 0.71			Mean = 11.63 Std dev. = 16.01 IR = 0.72			Mean = 11.94 Std dev. = 16.32 IR = 0.73		

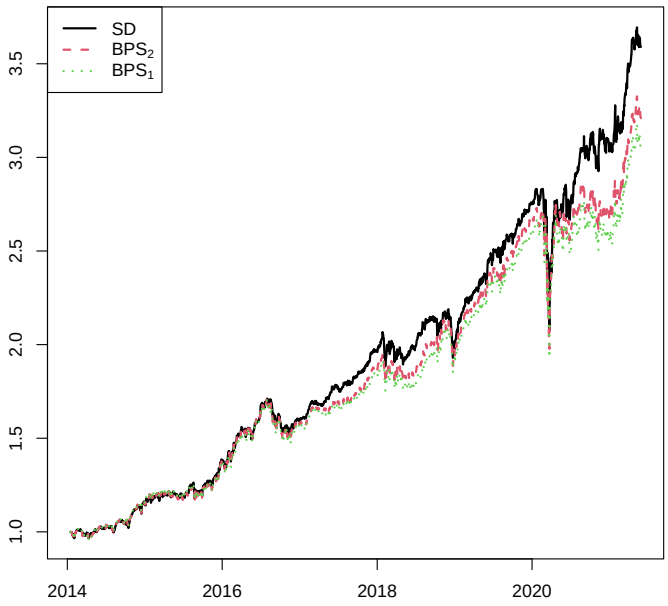
Notes: See notes of Table 5.

Figure 1: Normalized wealth growth with no transaction costs

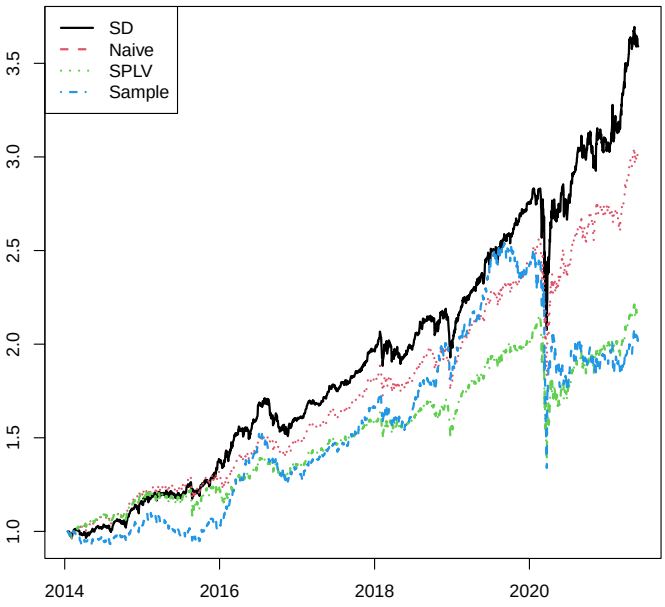
(a) SD and SS



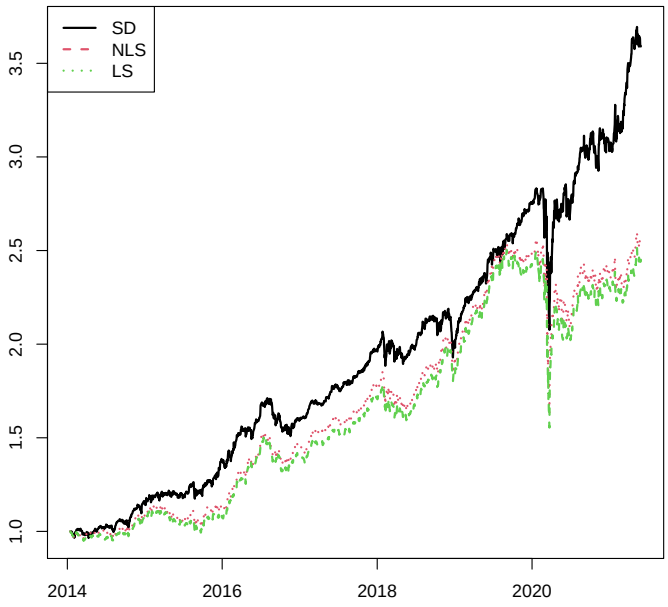
(b) SD, BPS_2, and BPS_1



(c) SD, Naive, SPLV, and Sample



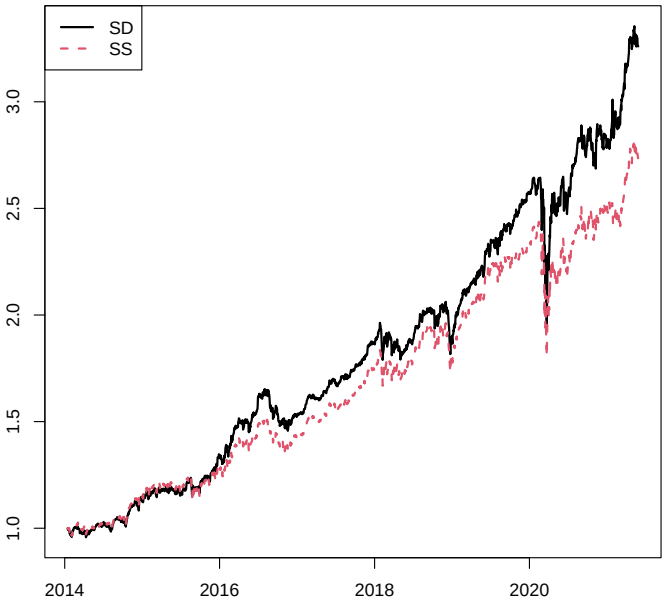
(d) SD, NLS, and LS



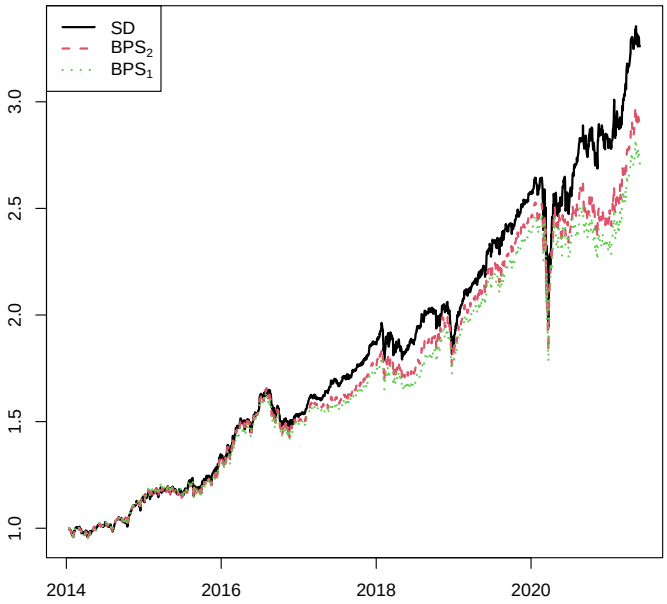
This figure shows the growth of 1 dollar invested according to the various portfolio strategies over the trading period from January 15, 2014 to May 28, 2021. The estimation window is of length $L = 240$ trading days and the portfolios are held for $H = 120$ trading days until the next rebalancing. The SPLV results correspond to a passive buy-and-hold strategy.

Figure 2: Normalized wealth growth with proportional transaction costs of 25 bps

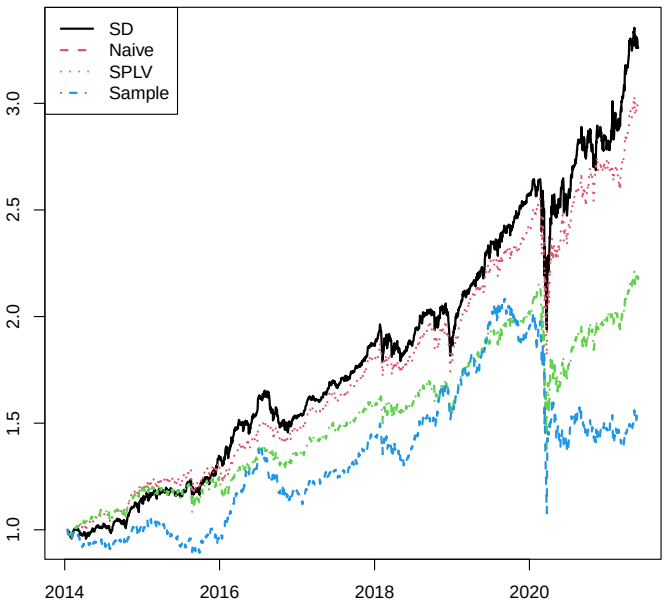
(a) SD and SS



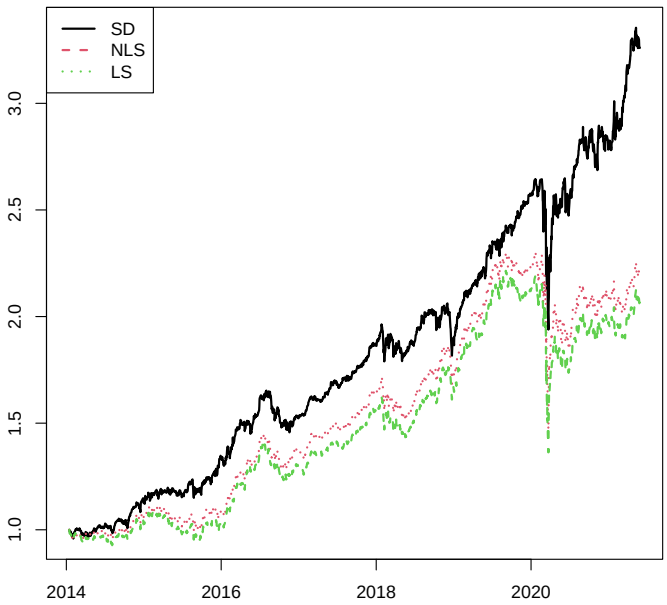
(b) SD, BPS_2, and BPS_1



(c) SD, Naive, SPLV, and Sample



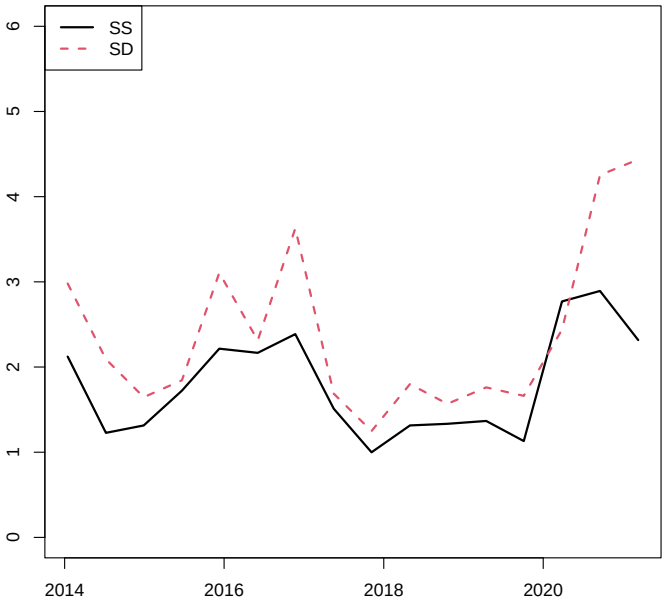
(d) SD, NLS, and LS



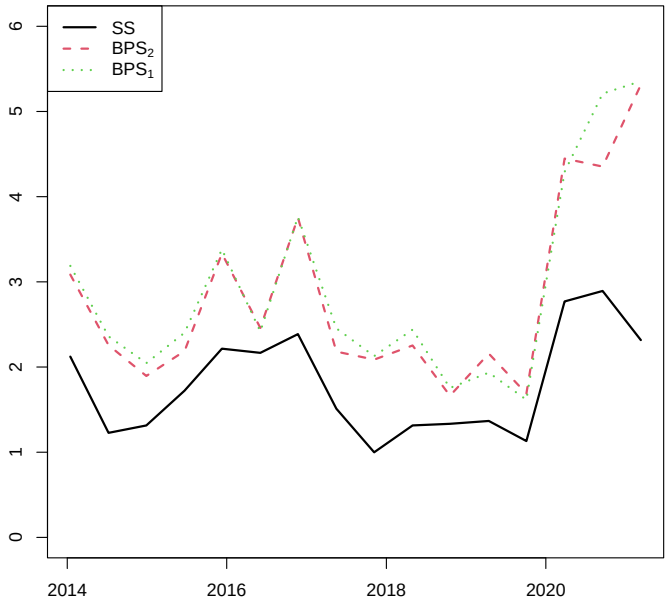
See notes of Figure 1.

Figure 3: Portfolio turnover, TO_t

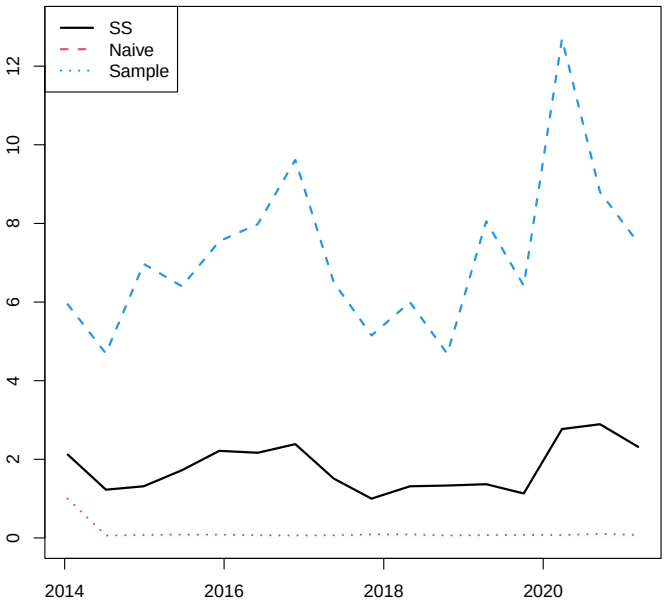
(a) SS and SD



(b) SS, BPS_2, and BPS_1



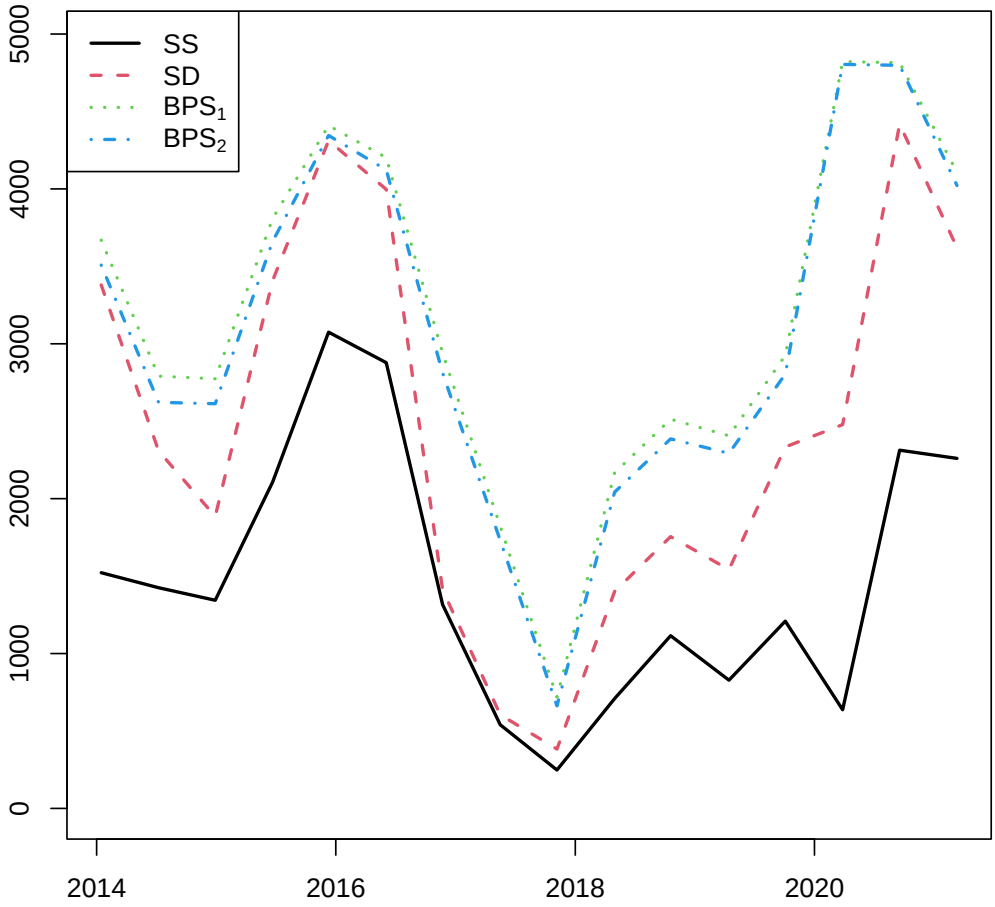
(c) SS, Naive, SPLV, and Sample



(d) SS, NLS, and LS

This figure shows the amount of turnover generated by the various portfolio strategies over the trading period from January 15, 2014 to May 28, 2021. The estimation window is of length $L = 240$ trading days and the portfolios are held for $H = 120$ trading days until the next rebalancing.

Figure 4: Number of significant covariances



This figure shows the number of statistically significant covariances detected by the multiple testing procedures over the trading period from January 15, 2014 to May 28, 2021. The estimation window is of length $L = 240$ trading days and the portfolios are held for $H = 120$ trading days until the next rebalancing.

Appendix

Table A1. List of SPLV fund holdings

Index	Ticker	Name	Average return (%)	Std dev. of returns (%)
1	CL	Colgate-Palmolive Co	0.036	1.187
2	PG	Procter & Gamble Co	0.045	1.145
3	PEP	PepsiCo Inc	0.052	1.162
4	JNJ	Johnson & Johnson	0.057	1.136
5	VZ	Verizon Communications Inc	0.035	1.125
6	COST	Costco Wholesale Corp	0.077	1.220
7	MDLZ	Mondelez International Inc	0.056	1.385
8	HSY	Hershey Co	0.055	1.354
9	MCD	McDonald's Corp	0.062	1.280
10	MMC	Marsh & McLennan Cos Inc	0.079	1.218
11	WMT	Walmart Inc	0.050	1.268
12	RSG	Republic Services Inc	0.074	1.137
13	HRL	Hormel Foods Corp	0.064	1.342
14	KMB	Kimberly-Clark Corp	0.040	1.250
15	EXPD	Expeditors International of Washington Inc	0.066	1.405
16	K	Kellogg Co	0.026	1.304
17	WM	Waste Management Inc	0.081	1.152
18	MKC	McCormick & Co Inc	0.065	1.335
19	BRK/B	Berkshire Hathaway Inc	0.059	1.215
20	MRK	Merck & Co Inc	0.049	1.321
21	CMS	CMS Energy Corp	0.063	1.281
22	GIS	General Mills Inc	0.040	1.276
23	YUM	Yum! Brands Inc	0.064	1.582
24	A	Agilent Technologies Inc	0.085	1.566
25	ICE	Intercontinental Exchange Inc	0.081	1.435
26	BAX	Baxter International Inc	0.053	1.367
27	KO	Coca-Cola Co	0.037	1.148
28	AEE	Ameren Corp	0.068	1.374
29	XEL	Xcel Energy Inc	0.065	1.274
30	LNT	Alliant Energy Corp	0.064	1.287
31	WEC	WEC Energy Group Inc	0.063	1.356
32	CHD	Church & Dwight Co Inc	0.065	1.267
33	HD	Home Depot Inc	0.093	1.452
34	AWK	American Water Works Co Inc	0.084	1.361
35	AJG	Arthur J Gallagher & Co	0.084	1.284
36	AEP	American Electric Power Co Inc	0.053	1.269
37	NDAQ	Nasdaq Inc	0.102	1.483
38	ARE	Alexandria Real Estate Equities Inc	0.065	1.402
39	BMY	Bristol-Myers Squibb Co	0.051	1.589
40	BR	Broadridge Financial Solutions Inc	0.108	1.398
41	ED	Consolidated Edison Inc	0.038	1.290
42	DUK	Duke Energy Corp	0.043	1.276
43	SO	Southern Co	0.044	1.345

Continued on next page

– continued from previous page

Index	Ticker	Name	Average return (%)	Std dev. of returns (%)
44	D	Dominion Energy Inc	0.041	1.362
45	CERN	Cerner Corp	0.042	1.529
46	CAG	Conagra Brands Inc	0.042	1.616
47	BF/B	Brown-Forman Corp	0.070	1.447
48	DG	Dollar General Corp	0.086	1.611
49	ZTS	Zoetis Inc	0.098	1.578
50	MNST	Monster Beverage Corp	0.103	1.942
51	PM	Philip Morris International Inc	0.033	1.424
52	SHW	Sherwin-Williams Co	0.094	1.563
53	ORCL	Oracle Corp	0.054	1.546
54	GRMN	Garmin Ltd	0.091	1.698
55	PSA	Public Storage	0.051	1.323
56	ORLY	O'Reilly Automotive Inc	0.097	1.684
57	ROP	Roper Technologies Inc	0.076	1.441
58	DRE	Duke Realty Corp	0.078	1.512
59	DHR	Danaher Corp	0.092	1.294
60	ITW	Illinois Tool Works Inc	0.081	1.444
61	VRSN	VeriSign Inc	0.088	1.572
62	ALL	Allstate Corp	0.070	1.369
63	ES	Eversource Energy	0.054	1.384
64	T	AT&T Inc	0.021	1.285
65	ATO	Atmos Energy Corp	0.065	1.342
66	CCI	Crown Castle International Corp	0.069	1.435
67	LMT	Lockheed Martin Corp	0.091	1.363
68	VRSK	Verisk Analytics Inc	0.064	1.398
69	BDX	Becton Dickinson and Co	0.065	1.340
70	V	Visa Inc	0.098	1.547
71	AON	Aon PLC	0.085	1.362
72	AME	AMETEK Inc	0.070	1.523
73	MCO	Moody's Corp	0.106	1.737
74	PAYX	Paychex Inc	0.076	1.449
75	IEX	IDEX Corp	0.086	1.388
76	FAST	Fastenal Co	0.060	1.704
77	MAA	Mid-America Apartment Communities Inc	0.067	1.461
78	SJM	J M Smucker Co	0.039	1.399
79	AZO	AutoZone Inc	0.075	1.559
80	ABBV	AbbVie Inc	0.085	1.775
81	MMM	3M Co	0.053	1.363
82	NOC	Northrop Grumman Corp	0.100	1.448
83	ETR	Entergy Corp	0.051	1.451
84	SPGI	S&P Global Inc	0.107	1.638
85	DTE	DTE Energy Co	0.060	1.403
86	GILD	Gilead Sciences Inc	0.047	1.779
87	GWW	WW Grainger Inc	0.058	1.748
88	UNH	UnitedHealth Group Inc	0.115	1.631

Continued on next page

– continued from previous page

Index	Ticker	Name	Average return (%)	Std dev. of returns (%)
89	CHTR	Charter Communications Inc	0.120	1.780
90	ACN	Accenture PLC	0.082	1.452
91	CMCSA	Comcast Corp	0.070	1.494
92	PNW	Pinnacle West Capital Corp	0.045	1.387
93	MO	Altria Group Inc	0.047	1.377
94	DPZ	Domino's Pizza Inc	0.123	1.754
95	CPB	Campbell Soup Co	0.035	1.496
96	ABT	Abbott Laboratories	0.077	1.458
97	STE	STERIS PLC	0.093	1.482
98	PFE	Pfizer Inc	0.041	1.290
99	PGR	Progressive Corp	0.092	1.345
100	BLL	Ball Corp	0.075	1.500

References

Abadir, K., W. Distaso, and F. Žikeš (2014). Design-free estimation of variance matrices. *Journal of Econometrics 181*, 165–180.

Bailey, N., H. Pesaran, and V. Smith (2019). A multiple testing approach to the regularisation of large sample correlation matrices. *Journal of Econometrics 208*, 507–534.

Barnard, G. (1963). Comment on 'The spectral analysis of point processes' by M.S. Bartlett. *Journal of the Royal Statistical Society (Series B) 25*, 294.

Benjamini, Y. and Y. Hochberg (1995). Controlling the false discovery rate: A practical and powerful approach to multiple testing. *Journal of the Royal Statistical Society. Series B 57*, 289–300.

Berk, J. (1997). Necessary conditions for the CAPM. *Journal of Economic Theory 73*, 245–257.

Bickel, P. and E. Levina (2008a). Covariance regularization by thresholding. *Annals of Statistics 36*, 2577–2604.

Bickel, P. and E. Levina (2008b). Regularized estimation of large covariance matrices. *Annals of Statistics 36*, 199–227.

Birnbaum, Z. (1974). Computers and unconventional test statistics. In F. Proschan and R. Serfling (Eds.), *Reliability and Biometry*, pp. 441–458. SIAM, Philadelphia.

Bollerslev, T. (1990). Modelling the coherence in short-run nominal exchange rates: A multivariate generalized ARCH model. *Review of Economics and Statistics 72*, 498–505.

Bretz, F., T. Hothorn, and P. Westfall (2010). *Multiple Comparisons Using R*. Chapman & Hall/CRC.

Cai, T. and W. Liu (2011). Adaptive thresholding for sparse covariance matrix estimation. *Journal of the American Statistical Association 106*, 673–684.

Chamberlain, G. (1983). A characterization of the distributions that imply mean-variance utility functions. *Journal of Economic Theory 29*, 185–201.

Chan, L., J. Karceski, and J. Lakonishok (1999). On portfolio optimization: Forecasting covariances and choosing the risk model. *Review of Financial Studies 12*, 937–974.

Chib, S., Y. Omori, and M. Asai (2009). Multivariate stochastic volatility. In T. Andersen, R. Davis, J.-P. Kreiss, and T. Mikosch (Eds.), *Handbook of Financial Time Series*, pp. 365–400. Springer-Verlag, Berlin.

Cont, R. (2001). Statistical inference for computable general equilibrium models, with application to a model of the Maroccan economy. *Quantitative Finance 1*, 223–236.

DeMiguel, V., L. Garlappi, and R. Uppal (2009). Optimal versus naive diversification: How inefficient is the 1/N portfolio strategy? *Review of Financial Studies 22*(5), 1915–1953.

Dufour, J.-M. (2003). Identification, weak instruments, and statistical inference in econometrics. *Canadian Journal of Economics 36*(4), 767–808.

Dufour, J.-M. (2006). Monte Carlo tests with nuisance parameters: A general approach to finite-sample inference and nonstandard asymptotics in econometrics. *Journal of Econometrics 133*, 443–477.

Dufour, J.-M. and L. Khalaf (2001). Monte Carlo test methods in econometrics. In B. Baltagi (Ed.), *A Companion to Theoretical Econometrics*, pp. 494–510. Basil Blackwell, Oxford, UK.

Dwass, M. (1957). Modified randomization tests for nonparametric hypotheses. *Annals of Mathematical Statistics 28*, 181–187.

El Karoui, N. (2008). Operator norm consistent estimation of large-dimensional sparse covariance matrices. *Annals of Statistics 36*, 2717–2756.

Engle, R., O. Ledoit, and M. Wolf (2019). Large dynamic covariance matrices. *Journal of Business and Economic Statistics 37*, 363–375.

Fang, K.-T., S. Kotz, and K. Ng (1990). *Symmetric Multivariate and Related Distributions*. Chapman and Hall/CRC, Boca Raton.

French, K. (2008). Presidential address: The cost of active investing. *Journal of Finance 63*(4), 1537–1573.

Ge, Y., S. Dudoit, and T. Speed (2003). Resampling-based multiple testing for microarray data analysis. *Test 12*, 1–77.

Ge, Y., S. Sealfon, and T. Speed (2008). Some step-down procedures controlling the false discovery rate under dependence. *Statistica Sinica 18*(3), 881–904.

Goeman, J. and A. Solari (2010). The sequential rejection principle of familywise error control. *Annals of Statistics 38*, 3782–3810.

Hochberg, Y. and A. Tamhane (1987). *Multiple Comparison Procedures*. John Wiley & Sons, New York.

Hsu, J. (1996). *Multiple Comparisons: Theory and Methods*. Chapman & Hall/CRC, Boca Raton.

Johnstone, I. (2001). On the distribution of the largest eigenvalue in principal components analysis. *Annals of Statistics 29*, 295–327.

Kirby, C. and B. Ostdiek (2012). It's all in the timing: Simple active portfolio strategies that outperform naïve diversification. *Journal of Financial and Quantitative Analysis 47*(2), 437–467.

Kiviet, J. (2011). Monte Carlo simulation for econometricians. *Foundations and Trends in Econometrics 5*, 1–181.

Korn, E., J. Troendle, L. McShane, and R. Simon (2004). Controlling the number of false discoveries: application to high-dimensional genomic data. *Journal of Statistical Planning and Inference 124*(2), 379–398.

Ledoit, O. and M. Wolf (2003). Improved estimation of the covariance matrix of stock returns with an application to portfolio selection. *Journal of Empirical Finance 10*, 603–621.

Ledoit, O. and M. Wolf (2004). A well-conditioned estimator for large-dimensional covariance matrices. *Journal of Multivariate Analysis 88*, 365–411.

Ledoit, O. and M. Wolf (2012). Nonlinear shrinkage estimation of large-dimensional covariance matrices. *Annals of Statistics 40*, 1024–1060.

Ledoit, O. and M. Wolf (2015). Spectrum estimation: A unified framework for covariance matrix estimation and pca in large dimensions. *Journal of Multivariate Analysis 139*, 360–384.

Lehmann, E. and J. Romano (2005). Generalizations of the familywise error rate. *Annals of Statistics 33*(3), 1138–1154.

Lehmann, E. and C. Stein (1949). On the theory of some non-parametric hypotheses. *Annals of Mathematical Statistics 20*(1), 28–45.

Owen, J. and R. Rabinovitch (1983). On the class of elliptical distributions and their applications to the theory of portfolio choice. *Journal of Finance 38*, 745–752.

Randles, R. and D. Wolfe (1979). *Introduction to the Theory of Nonparametric Statistics*. Wiley, New York.

Romano, J. and M. Wolf (2005). Exact and approximate stepdown methods for multiple hypothesis testing. *Journal of the American Statistical Association 100*, 94–108.

Romano, J. and M. Wolf (2016). Efficient computation of adjusted p-values for resampling-based stepdown multiple testing. *Statistics and Probability Letters 113*, 38–40.

Rothman, A., E. Levina, and J. Zhu (2009). Generalized thresholding of large covariance matrices. *Journal of the American Statistical Association 104*, 177–186.

Serfling, R. (2006). Multivariate symmetry and asymmetry. In S. Kotz, C. Read, N. Balakrishnan, B. Vidakovic, and N. Johnson (Eds.), *Encyclopedia of Statistical Sciences*. John Wiley & Sons.

Silvennoinen, A. and T. Teräsvirta (2009). Multivariate GARCH models. In T. Andersen, R. Davis, J.-P. Kreiss, and T. Mikosch (Eds.), *Handbook of Financial Time Series*, pp. 201–229. Springer-Verlag, Berlin.

Toulis, P. and J. Bean (2021). Randomization inference of periodicity in unequally spaced time series with application to exoplanet detection. *arXiv preprint arXiv:2105.14222*.

Westfall, P. and J. Troendle (2008). Multiple testing with minimal assumptions. *Biometrical Journal 50*, 745–755.

Westfall, P. and S. Young (1993). *Resampling-Based Multiple Testing: Examples and Methods for p-Value Adjustment*. Wiley, New York.

Won, J.-H., J. Lim, S.-J. Kim, and B. Rajaratnam (2013). Condition-number-regularized covariance estimation. *Journal of the Royal Statistical Society: Series B 75*, 427–450.

Wu, W. and M. Pourahmadi (2009). Banding sample autocovariance matrices of stationary processes. *Statistica Sinica 19*, 1755–1768.