Modelling and Forecasting of Realized Covariance Matrices

Inauguraldissertation

zur Erlangung des akademischen Grades
Doctor Rerum Politicarum

an der
Fakultät für Wirtschafts- und Sozialwissenschaften
der Ruprechts-Karls-Universität Heidelberg

vorgelegt von

Michael Stollenwerk

Heidelberg
eingereicht im April 2023

Acknowledgments

I owe a special debt of gratitude to my supervisor, Christian Conrad, for his valuable feedback on my research, funding for data and conference participation, and general availability for advice. Furthermore, I am very thankful to Anne Opschoor for the stimulating and insightful discussions during his research visit to Heidelberg and for being a reviewer of this thesis. Feedback from Nestor Parolya helped to improve the quality of my thesis substantially; it's a pity that our research projects did not work out. I want to thank my coauthor Bastian Gribisch for our collaboration and for his excellent supervision of my master's thesis, which brought me on the path to pursuing a Ph.D. Finally, I am particularly grateful to Michael Rockinger, who hosted my research visit to the Université de Lausanne, where many of the ideas for this thesis were born, and who facilitated our return to Lausanne, which made a significant contribution to the successful completion of this work.

I appreciate the helpful comments by Kris Boudt, Timo Dimitriadis, Jacques Faraut, Roxana Halbleib, Peter Reinhard Hansen, Stephan Laurent, André Lucas, and Christian Müchner, and the participants of the Internal Seminar at the Alfred-Weber-Institute for Economics, the 15th International Conference on Computational and Financial Econometrics, the HeiKaMetrics seminar series, the Quantitiative Finance and Financial Econometrics Conferences (2018, 2019, and 2022), the Society for Financial Econometrics Summer Schools (2018, 2022), and the Research Seminar at the University of Freiburg.

I thank my friends and fellow Ph.D. students, Nabil Bouamara, Zain Chaudhry, Tillmann Eymess, Paula von Haaren, Robbert Schaap, Christian Scherf, Alexander Glas, Onno Kleen, and Christopher Zuber, for their companionship and support. Furthermore, I would like to thank Monika Lülf, Barbara Neef, Freya Schadt, Christoph Schineller, and Oliver Zeh for the bureaucratic and technical support, Melanie Arntz and Zeno Enders for providing employment during the final year, and Thomas Eife, Fabian Krüger, Marina Da Silva Rapp, Manuel Schick, Julius Schölkopf, David Vespermann and all people at the Alfred-Weber Institute for Economics for an enriching time. Finally, I would like to express my heartfelt gratitude to my wife, Rosi and my children Nala-Elikia and Kiano Kimya, for their unwavering support and love. You are a constant source of light, motivation, and inspiration. Thank you for always believing in me!

This work was partially performed on the computational resource *BwForCluster* funded by the Ministry of Science, Research and Arts and the Universities of the State of Baden-Wuerttemberg, Germany, whom I also thank.

Contents

List of Tables

List of Figures

Abbreviations

ARCH	autoregressive conditional heteroskedasticity
BEKK	Baba, Engle, Kroner, Kraft
BIC	Bayes information criterion
CAW	conditional autoregressive Wishart
DCC	dynamic conditional correlation
d.o.f.	degree of freedom
DPC-GARCH	dynamic principal component GARCH
FIM	Fisher information matrix
GARCH	generalized autoregressive conditional heteroskedasticity
GAS	generalized autoregressive score
GMVP	global minimum variance portfolio
HEAVY	high-frequency based volatility
i.i.d.	independent and identically distributed
MCS	model confidence set
MGARCH	multivariate GARCH
ML	maximum likelihood
MSV	multivariate stochastic volatility
NYSE	New York Stock Exchange
NASDAQ	National Association of Securities Dealers Automated Quotations
OGARCH	orthogonal GARCH
p.d.f.	probability density function
QML	quasi maximum likelihood
RC	realized covariance matrix
RCOV	realized covariance
Re-DCC	realized dynamic conditional correlation
RV	realized variance
SD	spectral decomposition
w.r.t.	with respect to

Mathematical Notation

p	cross-sectional dimension, number of assets $i = 1, \ldots, p$
$\mathbf{R}$	$p \times p$ realized covariance matrix; i.e. symmetric positive definite random matrix a realization thereof
$\boldsymbol{\Sigma}$, $\boldsymbol{\Omega}$	$p \times p$ real symmetric positive definite parameter matrices of probability distributions; distributions can be characterized either by $\boldsymbol{\Omega}$ or by $\boldsymbol{\Sigma}$; $\boldsymbol{\Sigma}$ is the expected value matrix
$\mathbf{n}$, $\boldsymbol{\nu}$	$p \times 1$ degree of freedom parameter vectors of probability distributions
n, ν	degree of freedom parameters of probability distributions
$\otimes$	Kronecker product
$\odot$	Hadamart product (element-wise multiplication)
$\mathbf{X}^\top$	transpose of $\mathbf{X}$
$\mathbf{X}^{-1}$	inverse of non-singular $\mathbf{X}$
$\mathbf{X}^{-\top}$	$= (\mathbf{X}^{-1})^\top = (\mathbf{X}^\top)^{-1}$; $\mathbf{X}$ non-singular
$\mathbf{X}^+$	$= (\mathbf{X}^\top\mathbf{X})^{-1}\mathbf{X}^\top$; Moore-Penrose inverse of $\mathbf{X}$
$\mathrm{vec}(\mathbf{X})$	stacks elements of $\mathbf{X}$ column-by-column
$\mathrm{vech}(\mathbf{X})$	stacks elements on and below the main diagonal of symmetric or lower-triangular $\mathbf{X}$
$\mathrm{ivech}(\mathbf{x})$	creates a symmetric matrix from vector $\mathbf{x}$ of suitable size
$\mathrm{vecd}(\mathbf{X})$	stacks diagonal elements of rectangular $\mathbf{X}$; $(\mathrm{vecd}(\mathbf{X}))_i = (\mathbf{X})_{ii}$
$\mathrm{dg}(\mathbf{x})$	creates diagonal matrix from vector $\mathbf{x}$; $(\mathrm{dg}(\mathbf{x}))_{ii} = (\mathbf{x})_i$; $(\mathrm{dg}(\mathbf{x}))_{ij} = 0$, $i \neq j$
$\mathrm{tr}(\mathbf{X})$	$= \sum_{i=1}^{p} (\mathbf{X})_{ii}$; trace; sums elements on main diagonal of $(p \times p)\mathbf{X}$
$\overleftarrow{\mathbf{x}}$	$= (x_p, x_{p-1}, \ldots, x_1)^\top$; $\mathbf{x}$ is $p \times 1$; reverse-order vector
$\overleftarrow{\mathbf{X}}$	$= \mathbf{P}_e\mathbf{X}\mathbf{P}_e$; $\mathbf{P}_e$ see below; $\mathbf{X}$ is square; reverse-order matrix
$\mathrm{vec}^2(\mathbf{X})$	$= \mathrm{vec}(\mathbf{X})\mathrm{vec}(\mathbf{X})^\top$; vec-square
$\mathrm{vech}^2(\mathbf{X})$	$= \mathrm{vech}(\mathbf{X})\mathrm{vech}(\mathbf{X})^\top$; $\mathbf{X}$ symmetric or lower-triangular; vech-square
$\mathbf{X}^{\otimes 2}$	$= (\mathbf{X} \otimes \mathbf{X})$; Kronecker-square
$\mathbf{X}^{-\otimes 2}$	$= (\mathbf{X}^{-1} \otimes \mathbf{X}^{-1}) = (\mathbf{X} \otimes \mathbf{X})^{-1}$

continued on next page

$\Gamma_p(\mathbf{x}), \Gamma_p(x)$	(multivariate) gamma function; Definition 2.2.2
$\lvert\mathbf{X}\rvert_{\mathbf{y}}$	generalized power function; $\mathbf{X}$ is $p \times p$; $\mathbf{y}$ is $p \times 1$; Definition 2.2.1
$\blacktriangledown(\mathbf{X})$	sets all elements of above the main diagonal equal to zero; $\mathbf{X}$ is square; Definition 3.6.1
$\bar{\blacktriangledown}(\mathbf{X})$	$= \blacktriangledown\left(\mathbf{X} - \frac{1}{2}\mathbf{I} \odot \mathbf{X}\right)$; sets all elements above the main diagonal equal to zero and halves all elements on the main diagonal; $\mathbf{X}$ is square; Definition 3.6.2
$\tilde{\blacktriangledown}(\mathbf{X})$	$= \bar{\blacktriangledown}(\mathbf{X}) + \bar{\blacktriangledown}(\mathbf{X})^\top$; sets all elements above the main diagonal equal to the elements below the diagonal; $\mathbf{X}$ is square; Definition 3.6.3
$\mathbf{I}$	identity matrix
$\mathbf{T}, \mathbf{D}$	LDL decomposition of $\mathbf{\Sigma} = \mathbf{T}\mathbf{D}\mathbf{T}^\top$; $\mathbf{T}$ is a lower triangular matrix with ones on the main diagonal, $\mathbf{D}$ is a diagonal matrix
$\mathbf{C}_X$	lower Cholesky factor of symmetric positive definite $\mathbf{X} = \mathbf{C}_X\mathbf{C}_X^\top$
$\mathbf{C}$	without index, $\mathbf{C}$ denotes the lower Cholesky factor of $\mathbf{\Sigma} = \mathbf{C}\mathbf{C}^\top$
$\mathbf{G}$	duplication matrix defined by $\text{vec}(\mathbf{X}) = \mathbf{G}\text{vech}(\mathbf{X})$
$\mathbf{F}$	canonical elimination matrix, which consists only of ones and zeros; $\text{vech}(\mathbf{X}) = \mathbf{F}\text{vec}(\mathbf{X})$
$\mathbf{P}_e$	exchange matrix; square matrix with ones on the diagonal from bottom-left to top-right and zeros elsewhere.

Chapter 1

Introduction

What does modeling and forecasting of *realized covariance matrices* (RCs) mean, and why is it a relevant research topic? This introduction aims to answer these questions. Broadly speaking, RCs contain information about the variability of price changes of multiple financial assets on a given trading day. Since price change variability in the financial system is connected to financial market risk, it is apparent that accurate forecasts of RCs are useful for risk prediction. And in order to obtain accurate forecasts, we have to devise models that fit the data generating process as closely as possible. Thus, modelling and forecasting of realized covariance matrices is a relevant research topic. The following section explains in more detail what RCs are intuitively, what concretely they can be used for, and what they measure theoretically. Then, in the final section of this introduction, a short outline of the thesis is given, highlighting its contributions to the literature.

1.1 Realized Covariance Matrices

Imagine owning a set of financial assets and calculating an asset return on each trading day for each of these. A *realized covariance matrix* is a symmetric positive definite matrix composed of the *realized variances* (RVs) of and the *realized covariances* (RCOVs) between each pair of those asset returns for a given trading day. The RVs are located on the main diagonal entries of an RC; the RCOVs on the off-diagonal elements. For example, the RCOV between the return of asset one and two is located in the first row, second column, and second row, first column entry of the RC. RCs are computed from high-frequency price data, that is, price data over the trading day with a frequency of, say, at least one observation every 15 minutes, as recorded on, for example, stock exchanges like the New York Stock Exchange (NYSE) or the National Association of Securities Dealers Automated Quotations (NASDAQ). The word *realized* reflects that RCs are *accurate* ex-post measurements of the unobservable "true" covariance matrix of the underlying daily financial asset returns.

RCs can be interpreted as making the (co)variances of the underlying financial assets "effectively observable", and as such, it is advocated that they can be modelled directly (see e.g. Andersen et al., 2001, McAleer and Medeiros, 2008, Andersen et al., 2006, Chiriac and Voev, 2011, Golosnoy, Gribisch, and Liesenfeld, 2012, Opschoor et al., 2018). It is important to note that our objects of interest are the RCs, not their constituent RVs and RCOVs. In fact, we can not model the RVs and RCOVs separately and then combine their fitted values and forecasts because this does not guarantee the positive definiteness of the resulting matrix. This is because RCs are restricted to be positive definite and thus have to be modeled directly under this restriction.

1.1.1 Why are Realized Covariance Matrices Important?

As mentioned above, *variance*, also known as *volatility*, is often taken as a synonym for the risk of a financial asset. If we are interested in the risk, or volatility, of an entire portfolio of assets, it is essential to take not only the idiosyncratic variances into account but also their *covariances*, also known as *covolatilites*, which quantify the joint variability of two underlying financial asset returns. If, for example, two assets in the portfolio have a negative covariance, this means that if one has a negative return, the other is likely to have a positive one, leading to so-called diversification effects.

Due to RCs' close connection to financial risk, there are numerous applications in asset management. For example, given a prediction of the next day's RC, one could calculate the portfolio weights, which minimize the overall variance of an entire portfolio; the resulting allocation is called the *global minimum variance portfolio*. Another application is the calculation of the portfolio *value at risk*. That is the portfolio loss, which with a certain probability (usually 95%) will not be exceeded. Furthermore, one can calculate risk-adjusted portfolio returns with the *Sharpe ratio*, which, in this context, is simply the (expected) portfolio return divided by the portfolio's variance. It scales the expected return by its risk, reflecting that riskier investments naturally exhibit higher (expected) returns since investors want to be compensated for the risk they are taking on. Even if one disagrees with its interpretation as risk, it is evident that, holding all else equal, investors prefer low over high variance in a financial asset.

Besides risk, variance can also directly impact the expected return itself as documented by the "low volatility anomaly", which states that low-variance stocks are associated with abnormally high returns (Baker, Bradley, and Wurgler, 2011). This is called an anomaly because stocks with higher variance should offer higher expected returns. After all, investors want to be compensated for the risk they take on.

Finally, there are small specialized applications in quantitative finance where accurate volatility modelling is critical. For example, in asset pricing, since, as we stated before, investors require compensation for the risk that they take on. Notably, option arbitrage can be done by comparing the volatility implied by the current option price to the forecasted volatility of the underlying financial asset.

We have outlined several applications of models for RCs and their forecasts, which can broadly be summarized under the umbrella of asset management. Since asset management touches any citizen that saves in one form or another for their retirement or has purchased any insurance coverage, RCs are indirectly relevant to society as a whole.

Theory 1.1.2

Consider a p-dimensional vector of log-prices $\mathbf{p}(\tau)$, where $\tau \in \mathbb{R}_+$ represents continuous time. Assume that $\mathbf{p}(\tau)$ is a semimartingale,

$$d\mathbf{p}(\tau) = \boldsymbol{\mu}(\tau)d\tau + \mathbf{A}(\tau)d\mathbf{W}(\tau). \tag{1.1}$$

with instantaneous drift $\boldsymbol{\mu}(\tau)$, standard p-variate Wiener process $\mathbf{W}(\tau)$ and $p \times p$ spot covariance matrix $\boldsymbol{\Theta}(\tau) = \mathbf{A}(\tau)\mathbf{A}(\tau)^\top$. Without loss of generality, restricting the trading day to the unit interval, we obtain the "true" *integrated covariance matrix* at day t as

$$\mathbf{IC}_t = \int_{t-1}^{t} \boldsymbol{\Theta}(\tau)\ d\tau.$$

Now, a *realized covariance matrix* (RC), which we denote by $\mathbf{R}_t$, is defined as a non-parametric ex-post estimate of $\mathbf{IC}_t$ exploiting high-frequency asset return information.[1]

The simplest RC is the *standard realized covariance matrix*, which is based on $m+1$ uniformly spaced intraday log-prices, with the j'th intraday return vector on day t ($t = 1, \ldots, T$) given by

$$\mathbf{r}_{j,t} = \mathbf{p}\left(t - 1 + \frac{j}{m}\right) - \mathbf{p}\left(t - 1 + \frac{j-1}{m}\right), \quad j = 1, \ldots, m.$$

It is defined as the sum of outer products of intraday return vectors,

$$\mathbf{R}_t = \sum_{j=1}^{m} \mathbf{r}_{j,t}\mathbf{r}_{j,t}^\top.$$

1. Some authors call a non-parametric ex-post estimate of $\mathbf{IC}_t$ *realized measure* and reserve the name *realized covariance matrix* for what we call *standard realized covariance matrix* below. We prefer our nomenclature, which other authors also choose, since *realized covariance matrix* conveys that the object is a matrix, which *realized measure* does not.

In the absence of market microstructure noise (bid-ask bounce, price discreteness, etc.) and discontinuous price jumps (equation (1.1) explicitly excludes discontinuous price jumps), it can be shown that the standard realized covariance matrix is a consistent estimator of $\mathbf{IC}_t$ as $m \to \infty$ (see Barndorff-Nielsen and Shephard, 2004). In practice, however, we should not let m become very large, i.e. sample every couple of seconds or so, because then the standard realized covariance matrix is contaminated by market microstructure noise and non-synchronous trading, which introduce severe biases (see Zhang, 2011, who also derive an optimal sampling frequency). Pragmatically, it is advocated to sample simply at frequencies of at least one minute and up to 15 minutes (see e.g. Bandi, Russell, and Yang, 2008 and Liu, Patton, and Sheppard, 2015). However, when sampling at these lower frequencies, for example, every five minutes, log-returns are based only on the two price observations at the beginning and end of the five-minute interval, and all price observations in between are discarded. A simple way to use more data and, consequently, make the standard realized covariance matrix estimator more efficient is to span multiple sampling grids over the trading hours (e.g. five five-minute grids starting at 09:00, 09:01, 09:02, 09:03, and 09:04, respectively), calculate the standard realized covariance matrix for each grid, and then taking the average. This is known as the *subsampled realized covariance matrix*. We use it in our empirical sections of Chapters 2 and 3.

There are numerous more sophisticated RCs than the subsampled or standard realized covariance matrix that are robust to market microstructure noise, thus avoiding having to sample at lower frequencies and allowing to use of tick data. Some RCs are also robust to jumps. That is, if we assume an additional jump component in the semimartingale (1.1), these estimators are still able to estimate $\mathbf{\Sigma}_t$ consistently. Popular examples of sophisticated RCs are the *Hayashi-Yoshida realized covariance* (Hayashi and Yoshida, 2005) that uses all ticks of every asset, but does not guarantee positive definiteness, the *realized outlyingness weighted covariation* (Boudt, Croux, and Laurent, 2011) that discards those intraday return vectors for which a statistical test rejects the null of no jumps. In Chapter 4, we use the *multivariate realized kernel* of Barndorff-Nielsen et al. (2011) (see also Barndorff-Nielsen et al., 2009). It is robust to market microstructure noise and jumps and is defined as

$$\mathbf{R}_t = \sum_{h=-m}^{m} k\left(\frac{h}{H}\right) \mathbf{\Gamma}_{h,t}, \tag{1.2}$$

where for $h \geq 0$,

$$\mathbf{\Gamma}_{h,t} = \mathbf{r}_{j,t}\mathbf{r}_{j-h,t}^{\top}$$

is the h-th realized autocovariance matrix with $\mathbf{\Gamma}_{-h} = \mathbf{\Gamma}_h^\top$, and $k(\cdot)$ is a twice continuously differentiable weight function with $k(0) = 1$ and $k'(0) = 0$. The intraday return vectors $\mathbf{r}_{j,t}$ are constructed using refresh time sampling[2] and consequently are not on an equally-spaced grid.[3] Each mentioned, and the numerous non-mentioned RCs have its advantages and drawbacks. For an overview of the large body of literature on RCs, we refer to Jin, Liu, and Yang, 2021 and the references therein.

1.2 Outline of the Thesis

In this thesis, we use *observation-driven models* for time-series of daily RCs. That is, we assume a matrix-variate probability distribution for the daily RCs, whose parameters are updated based on the RC realizations from previous days.

In particular, the next chapter (Chapter 2) looks at different matrix-variate probability distributions for the RCs and their theoretical and empirical properties. Chapter 3 proposes a flexible observation-driven model to update all distribution-specific time-varying parameters, not just the expected value matrix as is done in the literature so far. Chapter 4 introduces an observation-driven updating mechanism that is applicable to high-dimensional time-series of RCs. Each of these three chapters is a self-contained paper. Chapters 2 and 3 are solo-authored working papers. The last chapter is a version of an article peer-reviewed and published in *Quantitative Finance*. It was co-authored with Bastian Gribisch (University of Cologne).

Probability Distributions for Realized Covariance Matrices

In Chapter 2 we compare theoretically and empirically all probability distributions hitherto in the literature applied to time-series of RCs, and we propose a new distribution family of our own. The direct comparison of all the different distributions has not been the focus of a study so far.

We argue that the choice of the probability distribution is crucial. This is because the behavioral implications that can be derived from a forecasted probability distribution of an RC can be different across distributions. A distribution that better fits the observed characteristics of RCs is likely to provide more useful

2. Refresh time sampling refers to the following. Given an initial intraday price vector, a new one is created, using previous tick interpolation, as soon as all corresponding assets have at least one new price observation from an executed trade.

3. Also, the first and last intraday return vectors of a given day, $\mathbf{r}_{0,t}$ and $\mathbf{r}_{m,t}$, need to be constructed from averaged prices over a small time window for the asymptotic theory to work.

behavioral recommendations. We establish that the probability distributions differ in various characteristics, and that some mirror the empirical characteristics of RCs more closely than others. Examples of important empirical characteristics are that realized variances (i.e. the diagonal elements of the RCs) have high variance, are fat-tailed, and are right-skewed (see Opschoor and Lucas, 2022). Furthermore, we show that realized covariances of S&P 500 stocks are more often positive than negative, and that the determinant of the RCs (as a measure of their sizes) has a highly fat-tailed empirical distribution. Moreover, we demonstrate that if the realized variance of one asset is large, the one of another asset is likely to be large too. We call this last property *tail-homogeneity*. One could be interested in many other empirical characteristics, but the point we want to make becomes clear. RCs are complex objects; some probability distributions are better than others in mirroring these complexities, and behavioral recommendations from better-fitting distributions are more useful. Furthermore, from a statistical viewpoint, it is valuable in itself to investigate how well different probability distributions can mirror the RC data.

The specific contributions of this chapter are as follows: We reveal theoretical similarities and differences among the distributions by collecting the relevant theory from the literature and presenting them in a common notation. This is useful for explaining the differences in empirical fit and forecasting performance. As mentioned above, another contribution is, that we derive a new distribution family (the t-Riesz distribution family). We show that this distribution family offers a particularly good fit to the RC data and can be grounded in a realistic low-level assumption on the data-generating process. Furthermore, we visualize how all distributions are related to each other, and we theoretically show the relationships many of which have not been previously derived. The paper contains further minor contributions, for example, a possible formalization of the concept of *fat-tailedness* for RCs.

In the empirical part of this paper, we perform fit and forecasting comparisons of the different distributions with or without the assumption of time-varying parameters. The empirical results show, that without time-varying parameters, the t-Riesz distribution family are the only distributions that can accurately mirror key characteristics of the data. With time-varying parameters, we see that the F-Riesz distribution, recently introduced to financial econometrics by Blasques et al. (2021), becomes a close competitor of the Inverse t-Riesz in the in-sample analysis. One notable difference between the two distributions is that the Inverse t-Riesz assumes tail-homogeneity while the F-Riesz implies *tail-heterogeneity*. We show that the assumption of tail-homogeneity fits very well, especially in more homogeneous, industry-specific datasets. Out-of-sample forecasting comparisons favor the Inverse t-Riesz distribution.

Generalized Autoregressive Score Models for Realized Covariance Matrices

In Chapter 3, we look at a specific class of observation-driven models called *generalized autoregressive score* (GAS) models. This class is very general because it contains many "traditional" observation-driven models. In it, the scaled score of the assumed probability distribution drives the time-variation in the parameters. The score is a non-linear function of the RCs and contains information about the entire shape of the probability density function. Thus the parameter updating dynamics are richer than in "traditional" observation-driven models. Furthermore, GAS models offer a direct way to make any distributional parameter time-varying.

Probability distributions for RCs can be characterized by their symmetric positive definite expected value matrix and a set of d.o.f. parameters. A central contribution of our paper is to endow the distribution-specific *degree of freedom* (d.o.f.) parameters with GAS dynamics to make them time-varying. We show that when estimating time-series models for RCs with static d.o.f. parameters over different time frames, the estimated d.o.f. parameters exhibit systematic and interpretable time-variation. Thus assuming they are time-varying is sensible. Another main contribution is to derive the scores and Fisher information matrices of all hitherto used probability distributions. So far, only a small subset of those have been derived. We further contribute to the literature by deriving a general across-distributions formula for the scaled scores with respect to (w.r.t.) the expected value matrix that reduces the computational burden so that the models can be applied to dimensions of up to 50 assets. We compare our GAS models across distributions. This is sensible since different probability distributions imply different parameter updating processes because their scaled scores differ and it is not clear which one works best.

Empirically, we show that the time-variation in the d.o.f. parameters significantly improves fit and forecast performance, and a statistical test confirms that models without time-varying d.o.f. parameters are too restrictive. Furthermore, the fitted values of the d.o.f. parameters make economic sense as they imply more fat-tailed distributions during economic crisis periods.

Our GAS model constitutes very flexible parameter-updating dynamics that are difficult to generalize further while keeping computational feasibility. Thus the choice of distributional assumption is all the more important. The comparison across distributions reveals, as in the previous chapter, that the Inverse t-Riesz distribution offers the best in- and out-of-sample fit, closely followed by the F-Riesz distribution. Overall the ranking across distributions is very similar to the one in the preceding chapter, where simple time-series dynamics were assumed.

Dynamic Principal Component CAW Models for High-Dimensional Realized Covariance Matrices

The final chapter of this thesis proposes a model that can be applied to high-dimensional time-series of RCs. Since portfolios often contain over 50 assets, high-dimensional models are of considerable practical relevance. At these large dimensions, the GAS models of the previous chapter become prohibitively slow to estimate, even with the computationally efficient formula we derived. Models with simple parameter updating mechanisms, like the scalar-BEKK of Chapter 2, can still be estimated in high dimensions. However, as explained above, their simple dynamics are already too restrictive for moderate dimensional settings (say, 10 to 50 assets). They are even more ill-suited to mirror the complex dynamics of more than 50 asset return (co)variances. Thus, we need models designed explicitly for high-dimensional time-series of RCs.

One strand of the literature proposes to employ so-called sparsity assumptions. That is, the high-dimensional RCs are mapped into lower-dimensional objects, which are assumed to follow rich dynamic models (see e.g. Wang and Zou (2010), Tao et al. (2011), Shen, Yao, and Li (2020), Sheppard and Xu (2019), Asai and McAleer (2015), Jin, Maheu, and Yang (2019)). Another option explored in the literature is to design relatively complex models whose parameters can be estimated iteratively. For example, the well-known Re-DCC model of Bauwens, Storti, and Violante (2012) falls into that category. In this chapter, we follow this latter strand of the literature.

In particular, we propose the dynamic principal component conditional autoregressive Wishart (DPC-CAW) model. In the model the expected value matrix of the Wishart distribution is decomposed into an eigenvector matrix and the corresponding eigenvalues; the latter are sometimes called "principal components". Then we assume separate but interdependent updating processes for the eigenvectors and eigenvalues. The RCs follow a Wishart distribution conditional on the updated expected value matrix. Our model can be seen as an adaptation of the multivariate DPC-GARCH model for daily return vectors proposed by Aielli and Caporin (2015). The decomposition into eigenvectors and eigenvalues is reminiscent of the (Re-)DCC model, where the RCs are decomposed into correlation matrices and (realized) variances, both endowed with distinct but interdependent dynamics.

The contributions of our paper are as follows: We propose the DPC-CAW model and suggest a three-step estimation procedure for its parameters, allowing us to estimate the model in high-dimensional settings. We perform simulation experiments to confirm the reasonable precision of our three-step estimation procedure. Furthermore, we apply the model to a 100-dimensional dataset and show in a fore-

casting experiment that it outperforms its competitors, particularly in forecasting the correlation structure and the global minimum variance portfolio weights.

Chapter 2

Probability Distributions for Realized Covariance Matrices

Introduction 2.1

The covariance matrix of financial asset returns is an essential object in the financial econometrics literature because it entails direct implications for efficient portfolio allocation and can be used in risk management and derivative pricing. A *realized covariance matrix* (RC) is an accurate and consistent ex-post estimate of the integrated covariance matrix over a day of financial asset returns. It is constructed from high-frequency data and can be interpreted as making the daily covariance matrix of the underlying financial asset returns "effectively observable". As such, it is advocated to model time-series of RCs directly (see e.g. Andersen et al., 2001, Andersen et al., 2006, McAleer and Medeiros, 2008, Chiriac and Voev, 2011). This direct modelling is in contrast to traditional multivariate volatility models (pioneered by Engle and Kroner, 1995 and Engle, 2002), which treat the covariance matrix as latent, and are based on low-frequency daily return data. In recent years, the increasing availability of high-frequency data has led to an enormous growth of models designed for RCs. These are mostly *observation-driven models*, in which RCs are treated as random matrices with time-varying conditional distributions. While many different models that feature different probability distributions have been proposed, the explicit comparison of the theoretical and empirical properties of these probability distributions has not been the focus of a study to the best of our knowledge.

This paper aims to fill this gap in the literature. The comparison is important since choosing a probability distribution that accurately reflects the characteristics of RC data is crucial because, depending on the assumed distribution, different courses of action might result for practitioners. For instance, consider an investor who wants to invest in the predicted global minimum variance portfolio. If she is already invested in the underlying assets with some given weights, the covariance matrix forecast almost surely implies different optimal weights than her current

allocation. But should she re-weight her portfolio, or is her current allocation "within reasonable distance" to the optimal weights? To answer this question, she must know the probability distribution of the forecasted covariance matrix, i.e. of the forecasted RC. There are many other practical situations like this where the distribution of the forecasted covariance matrix is of vital importance.

Adopting a well-fitting distribution is also in itself an important aim. After all, if empirically, the realized variances (RVs), i.e. the diagonal elements of the RCs, have high variance, are fat-tailed, and are right-skewed (Opschoor and Lucas, 2022), the realized covariances (RCOVs) are more often positive than negative, and the RV of one asset tends to be large if the one of another asset is large (as is true in our data), it is logical to choose a probability distribution that can reflect these properties theoretically. For one-dimensional models for returns, one might argue that assuming a simple conditional distribution with time-varying volatility yields unconditional distributions that are fat-tailed and thus fit the data well (Brownlees, Engle, and Kelly, 2012). However, Bai and Chen (2008) show that in the traditional *multivariate* GARCH models for daily return vectors, the multivariate normal distribution is rejected, whereas the t-distribution is not. Certainly, distributions that cannot mirror the complex stylized facts of *matrix-variate* RC data cannot do so simply by adding time-varying parameters to the models. Consequently, making an accurate distributional assumption on RCs is imperative.

In this paragraph we outline the specific contributions of the paper. One of the contributions is to collate theoretical knowledge about the different hitherto used probability distributions from the respective papers into one place and common notation.[1] This uncovers theoretical similarities and differences among the distributions, which is useful in explaining their disparity in empirical fit and forecasting performance. Another contribution is that we suggest and derive a new distribution family, the *t-Riesz distribution family*, which is based on the t-Wishart distribution by Sutradhar and Ali (1989). We show that it has desirable theoretical and empirical properties, is closely related to the previously suggested distributions, and can be grounded in a distributional assumption on the intraday return vectors from which an RC is constructed. Furthermore, we demonstrate that the t-Riesz distribution family offers a particularly good fit for the RC data. We contribute further by showing that all distributions, including the t-Riesz distribution family, can be regarded as belonging to a common overarching family. In this context we rigorously show how exactly all distributions are related to each other.

1. We exclude the non-central Wishart distribution, which gives only slight improvements compared to the Wishart in terms of fit and forecasting ability and is not applicable to dimensions higher than five due to computational difficulties involving the matrix-variate hypergeometric function.

Many of these relations have not been previously derived. In the empirical part of this paper, we perform fit and forecasting comparisons of the different distributions in different datasets of time-series of RCs and explain how the theoretical differences translate into differences in fit and forecasting performance. Finally, other minor contributions of this paper are the formalization of *fat-tailedness* for RCs, the discovery and derivation of another distribution (the Inverse F-Riesz), showing that empirically it suffices to consider one of the two versions of Riesz distributions[2], and giving more straightforward proofs of some previously known results.

We now present all probability distributions hitherto applied to time-series of RCs, citing not their original sources, but the ones that introduced them to financial econometrics. Probability distributions for RCs must generate symmetric positive definite random matrices. The first one proposed for RCs was the Wishart (Gourieroux, Jasiak, and Sufana, 2009, and Golosnoy, Gribisch, and Liesenfeld, 2012). It can be derived as the sum of outer products of independent and identically distributed (i.i.d.) normally distributed random vectors, and arises naturally in many areas of statistics. Notably, Golosnoy, Gribisch, and Liesenfeld (2012) propose the conditional autoregressive Wishart (CAW) model, which assumes that the conditional expected value matrix of the Wishart distribution follows BEKK dynamics (cf. Engle and Kroner, 1995 for BEKK dynamics). Gorgi et al. (2019) also use the Wishart matrix but assume *generalized autoregressive score* (GAS) dynamics for the expected value matrix. The Wishart distribution has been well-studied (see e.g. Gupta and Nagar, 2000) and is attractive because of its relative simplicity. However, it has the disadvantage of being a *thin-tailed* distribution, which is in contrast to the empirically *fat-tailed* RCs. Yu, Li, and Ng (2017) generalize the CAW model by assuming a *Non-Central* Wishart distribution, where the non-centrality matrix depends on the lagged values of the RCs. However, their model is not applicable to dimensions of, say, more than three assets because the likelihood computation of the Non-Central Wishart distribution involves the numerical approximation of the matrix-variate hypergeometric function (see Koev and Edelman, 2006), which is prohibitively slow. For this reason, we exclude it from our comparisons. Asai and So (2013) and Jin and Maheu (2016) use the Inverse Wishart distribution in various model frameworks. Jin and Maheu (2016) show that the Inverse Wishart distribution is better at modeling the conditional density of realized covariance matrices than the Wishart distribution. As we explain in the paper, the Inverse Wishart can be considered fatter tailed than the Wishart distribution. The first large advancement in distributional fit was accomplished

2. There is a distribution family called Riesz distributions, where each member has two versions.

by Opschoor et al. (2018) who propose the matrix-F distribution[3] and show that it implies fatter tails for the RVs than previously used distributions. They employ the F distribution in a GAS model and demonstrate that it significantly outperforms previously proposed (Inverse) Wishart distribution-based models. Zhou et al. (2019) confirm these results. Recently, Blasques et al. (2021) introduced the Riesz, Inverse Riesz, and F-Riesz distributions into financial econometrics. These Riesz-type distributions generalize, respectively, the Wishart, Inverse Wishart and F distributions by featuring *degree of freedom* (d.o.f.) parameter vectors instead of scalars, thus adding flexibility. Blasques et al. (2021) propose an efficient algorithm to estimate the Riesz-type distributions. Furthermore, they show that there is again a large increase in fit and forecasting ability from the F to the F-Riesz distribution and attribute this to the fact that the F-Riesz distribution features heterogeneous tails for the realized variances. Gribisch and Hartkopf (2022) show that the Riesz distribution for the standard realized covariance matrix can be derived by assuming a normal distribution on the intraday return innovations and sorting the assets according to their liquidity, where the asset with the most zero intraday returns is interpreted to be least liquid.

The rest of this chapter is structured as follows. The next section derives all probability distributions in a unified framework based on their stochastic representations, analyzes the distribution's stochastic properties, discusses fat-tailedness, shows how the distributions relate to each other and derives the probability density functions. Section 2.3 shows how the newly derived t-Riesz distribution family can be alternatively derived from a reasonable low-level assumption on the intraday return vectors. Section 2.4 introduces time-variation in the expected value parameter matrix of the distributions, 2.5 discusses estimation and performs the in-sample fit comparison and out-of-sample forecasting performance analysis of the different distributions, and 2.6 concludes.

2.2 Probability Distributions

Let $\mathbf{R}$ denote the $p \times p$ symmetric positive definite realized covariance matrix of p asset returns on a given day and assume that it follows a probability distribution $\mathcal{D}$ with support on symmetric positive semidefinite matrices. Later on, we will add a time index t indicating the day $t = 1, \ldots, T$ and write $\mathbf{R}_t$; for now, we opt for better readability.

All hitherto in the literature considered probability distributions for $\mathbf{R}$ are characterized by a $p \times p$ positive definite *parameter matrix* $\mathbf{\Omega}$ and a distribution-specific

3. From now on we call the matrix-F simply the F distribution.

set of *degree of freedom* (d.o.f.) parameters $\boldsymbol{\theta}_{\mathcal{D}}$, such that we write

$$\mathbf{R} \sim \mathcal{D}(\boldsymbol{\Omega}, \boldsymbol{\theta}_{\mathcal{D}}). \tag{2.1}$$

Their characterization in terms of $\boldsymbol{\Omega}$ and $\boldsymbol{\theta}_{\mathcal{D}}$ will become apparent by examining the distributions' stochastic representations, which we will do next. The stochastic representations are central to this paper since based on them we introduce the various distributions, analyze their stochastic properties, derive the distributions' expected values and probability density functions (p.d.f.s) as well as their relationships to one another.

2.2.1 Stochastic Representations

All hitherto in the literature considered probability distributions for $\mathbf{R}$ and the ones newly proposed in this paper can be generated from the $p \times p$ random triangular matrices

$$\underline{\mathbf{B}} = \begin{bmatrix} \sqrt{\chi^2_{n_1-1+1}} & & & 0 \\ & \sqrt{\chi^2_{n_2-2+1}} & & \\ & & \ddots & \\ \mathcal{N}(0,1) & & & \sqrt{\chi^2_{n_p-p+1}} \end{bmatrix}$$

and/or

$$\bar{\mathbf{B}} = \begin{bmatrix} \sqrt{\chi^2_{\nu_1-p+1}} & & & \mathcal{N}(0,1) \\ & \sqrt{\chi^2_{\nu_2-p+2}} & & \\ & & \ddots & \\ 0 & & & \sqrt{\chi^2_{\nu_p-p+p}}, \end{bmatrix},$$

that is

$$(\underline{\mathbf{B}})_{ij} \sim \begin{cases} \mathcal{N}(0,1) & \text{for } i < j, \\ \chi_{n_i-i+1} & \text{for } i = j \end{cases} \tag{2.2}$$

and/or

$$(\bar{\mathbf{B}})_{ij} \sim \begin{cases} \chi_{\nu_i-p+i} & \text{for } i = j, \\ \mathcal{N}(0,1) & \text{for } i > j, \end{cases} \tag{2.3}$$

Distribution	$\boldsymbol{\mathcal{K}}_{\mathcal{D}}$	Distribution	$\boldsymbol{\mathcal{K}}_{\mathcal{D}}$
Wishart ($\mathcal{W}$)	$\underline{\mathcal{B}}\underline{\mathcal{B}}^\top$	Riesz ($\mathcal{R}$)	$\underline{\mathbf{B}}\underline{\mathbf{B}}^\top$
Inv.Wishart ($i\mathcal{W}$)	$\bar{\mathcal{B}}^{-\top}\bar{\mathcal{B}}^{-1}$	Inv.Riesz ($i\mathcal{R}$)	$\bar{\mathbf{B}}^{-\top}\bar{\mathbf{B}}^{-1}$
t-Wishart ($t\mathcal{W}$)	$(\bar{b})^{-2}\underline{\mathcal{B}}\underline{\mathcal{B}}^\top$	t-Riesz ($t\mathcal{R}$)	$(\bar{b})^{-2}\underline{\mathbf{B}}\underline{\mathbf{B}}^\top$
Inv.t-Wishart ($it\mathcal{W}$)	$(\underline{b})^2\bar{\mathcal{B}}^{-\top}\bar{\mathcal{B}}^{-1}$	Inv.t-Riesz ($it\mathcal{R}$)	$(\underline{b})^2\bar{\mathbf{B}}^{-\top}\bar{\mathbf{B}}^{-1}$
F	$\bar{\mathcal{B}}^{-\top}\underline{\mathcal{B}}\underline{\mathcal{B}}^\top\bar{\mathcal{B}}^{-1}$	F-Riesz ($F\mathcal{R}$)	$\bar{\mathbf{B}}^{-\top}\underline{\mathbf{B}}\underline{\mathbf{B}}^\top\bar{\mathbf{B}}^{-1}$
F	$\underline{\mathcal{B}}\bar{\mathcal{B}}^{-\top}\bar{\mathcal{B}}^{-1}\underline{\mathcal{B}}^\top$	Inv.F-Riesz ($iF\mathcal{R}$)	$\underline{\mathbf{B}}\bar{\mathbf{B}}^{-\top}\bar{\mathbf{B}}^{-1}\underline{\mathbf{B}}^\top$

Table 2.1: Stochastic representation kernels $\boldsymbol{\mathcal{K}}_{\mathcal{D}}$ of all distributions for RCs. The complete stochastic representations are given by $\mathbf{C}_\Omega\boldsymbol{\mathcal{K}}_{\mathcal{D}}\mathbf{C}_\Omega^\top$, where $\mathbf{C}_\Omega$ is the lower Cholesky factor of the $p \times p$ symmetric positive definite parameter matrix $\boldsymbol{\Omega} = \mathbf{C}_\Omega\mathbf{C}_\Omega^\top$. $\underline{b}$ and $\bar{b}$ are χ_n and χ_ν distributed random variables, thus can be interpreted as one-dimensional $\underline{\mathbf{B}}$ and $\bar{\mathbf{B}}$, respectively. $\underline{\mathcal{B}}$ ($\bar{\mathcal{B}}$) is the special cases of $\underline{\mathbf{B}}$ ($\bar{\mathbf{B}}$) where for all i, $n_i = n$ ($\nu_i = \nu$).

where all random variables inside the matrices are independent of each other.[4] We refer to these random matrices as *Bartlett matrices* in reference to the well-known Bartlett decomposition of Wishart random matrices. The parameters n_i and ν_i are the aforementioned d.o.f. parameters, which we collect in the $p \times 1$ vectors $\mathbf{n} = (n_1, \ldots, n_p)^\top$ and $\boldsymbol{\nu} = (\nu_1, \ldots, \nu_p)^\top$. The special cases of the Bartlett matrices, where for all i, $n_i = n$ and $\nu_i = \nu$, we denote as $\underline{\mathcal{B}}$ and $\bar{\mathcal{B}}$, respectively. That is $\underline{\mathcal{B}}$ and $\bar{\mathcal{B}}$ are governed by d.o.f. parameters n and ν, and $\underline{\mathbf{B}}$ and $\bar{\mathbf{B}}$ are governed by d.o.f. parameter vectors $\mathbf{n}$ and $\boldsymbol{\nu}$. Note that for $p = 1$, the Bartlett matrices reduce to the random variables χ_n and χ_ν, respectively. For the matrix distributions to exist, we must restrict $n_i > i - 1$ and $\nu_i > p - i$ since otherwise the χ distributions on the main diagonals would not exist.[5]

Let $\mathcal{D} \in (\mathcal{W}, i\mathcal{W}, t\mathcal{W}, it\mathcal{W}, F, \mathcal{R}, i\mathcal{R}, t\mathcal{R}, it\mathcal{R}, F\mathcal{R}, iF\mathcal{R})$ denote the different probability distributions (see Table 2.1 for their full names). In Section 2.2.2 below we will introduce the distributions one by one and analyze their stochastic properties, but for now we continue to present the general stochastic representations. The ones in green are the novel distributions derived in this paper. Assuming that $\mathbf{R}$ follows one of the distributions $\mathcal{D}$, its stochastic representation can be written as

$$\mathbf{R} = \mathbf{C}_\Omega\boldsymbol{\mathcal{K}}_{\mathcal{D}}\mathbf{C}_\Omega^\top, \tag{2.4}$$

4. The χ_n distribution is given in e.g. Walck (2007), Section 8.14.

5. Note that this does not imply the existence of $\mathbb{E}[\mathbf{R}]$. For example the Inverse Wishart distribution is based on $\left(\bar{\mathcal{B}}\bar{\mathcal{B}}^\top\right)^{-1}$ and its mean only exists if in fact $\nu > p + 1$, whereas the distribution exists for $\nu > p - 1$.

Distribution	$\boldsymbol{\theta}_{\mathcal{D}}$	Distribution	$\boldsymbol{\theta}_{\mathcal{D}}$
Wishart	n	Riesz	$\mathbf{n}$
Inv.Wishart	ν	Inv.Riesz	$\boldsymbol{\nu}$
t-Wishart	$(n, \nu)^\top$	t-Riesz	$(\mathbf{n}^\top, \nu)^\top$
Inv.t-Wishart	$(n, \nu)^\top$	Inv.t-Riesz	$(n, \boldsymbol{\nu}^\top)^\top$
F	$(n, \nu)^\top$	F-Riesz	$(\mathbf{n}^\top, \boldsymbol{\nu}^\top)^\top$
F	$(n, \nu)^\top$	Inv.F-Riesz	$(\mathbf{n}^\top, \boldsymbol{\nu}^\top)^\top$

Table 2.2: Degree of freedom parameters of distributions for RCs.

where $\boldsymbol{\mathcal{K}}_{\mathcal{D}}$ is a distribution-specific function of one or both of the Bartlett matrices or their special cases, and $\mathbf{C}_{\Omega}$ denotes the lower Cholesky factor of symmetric positive definite parameter matrix $\boldsymbol{\Omega}$. We call $\boldsymbol{\mathcal{K}}_{\mathcal{D}}$ *stochastic representation kernel*. The distribution parameters are thus given by the parameter matrix $\boldsymbol{\Omega}$ and the d.o.f. parameter(s) in $\boldsymbol{\mathcal{K}}_{\mathcal{D}}$ (one or two of the set $(n, \nu, \mathbf{n}, \boldsymbol{\nu})$). We collect the distribution-specific d.o.f. parameters in the vector $\boldsymbol{\theta}_{\mathcal{D}}$. The exact composition of $\boldsymbol{\theta}_{\mathcal{D}}$ for the different distributions is given in Table 2.2.[6]

Characterization in Terms of Expected Value Matrix

We now standardize the distributions; that is, we characterize them in terms of their $p \times p$ symmetric positive definite *expected value matrix*

$$\boldsymbol{\Sigma} = \mathbf{C}\mathbf{C}^\top := \mathbb{E}\left[\mathbf{R}\right], \tag{2.5}$$

instead of in terms of $\boldsymbol{\Omega}$, where $\mathbf{C}$ is the lower Cholesky factor of $\boldsymbol{\Sigma}$. This standardization allows for a simple two-step estimation strategy, where the $\mathcal{O}(p^2)$ expected value matrix $\boldsymbol{\Sigma}$ is estimated by its obvious method of moments estimator in the first step and the d.o.f. parameters in a second step. Furthermore, since $\boldsymbol{\Sigma}$ has the same interpretation across distributions (unlike $\boldsymbol{\Omega}$), standardization makes comparisons of the distributions easier. Finally, the nesting relationships between the different distributions (see Figure 2.6 on p. 30) are only valid for the standardized, i.e. $\boldsymbol{\Sigma}$-parameterized, distributions.

We denote the expected value of the stochastic representation kernel by

$$\mathbf{M}_{\mathcal{D}} := \mathbb{E}\left[\boldsymbol{\mathcal{K}}_{\mathcal{D}}\right]. \tag{2.6}$$

6. It is easy to see from Table 2.1, which d.o.f. parameter(s) characterize each distribution.

Distribution	$\mathbf{M}_{\mathcal{D}}$	Distribution	$\mathbf{M}_{\mathcal{D}}$
Wishart	$\mathbf{I}n$	Riesz	$\mathrm{dg}(\mathbf{n})$
Inv.Wishart	$\mathbf{I}\frac{1}{\nu-p-1}$	Inv.Riesz	$\mathrm{dg}(\mathring{\boldsymbol{\nu}})$
t-Wishart	$\mathbf{I}\frac{n}{\nu-2}$	t-Riesz	$\mathrm{dg}(\mathbf{n})\frac{1}{\nu-2}$
Inv.t-Wishart	$\mathbf{I}\frac{n}{\nu-p-1}$	Inv.t-Riesz	$\mathrm{dg}(\mathring{\boldsymbol{\nu}})n$
F	$\mathbf{I}\frac{n}{\nu-p-1}$	F-Riesz	$\mathrm{dg}(\mathring{\mathbf{n}})$
F	$\mathbf{I}\frac{n}{\nu-p-1}$	Inv.F-Riesz	$\mathrm{dg}(\mathring{\mathring{\mathbf{n}}})$

Table 2.3: Expected values of stochastic representation kernels, $\mathbf{M}_{\mathcal{D}} = \mathbb{E}[\boldsymbol{\mathcal{K}}_{\mathcal{D}}]$. For the definitions of $\mathring{\boldsymbol{\nu}}$, $\mathring{\mathbf{n}}$ and $\mathring{\mathring{\mathbf{n}}}$, see Theorem 2.2.1.

Then the stochastic representation of the standardized distributions is

$$\mathbf{R} = \mathbf{C}\mathbf{M}_{\mathcal{D}}^{-\frac{1}{2}}\boldsymbol{\mathcal{K}}_{\mathcal{D}}\mathbf{M}_{\mathcal{D}}^{-\frac{1}{2}}\mathbf{C}^{\top}, \tag{2.7}$$

since

$$\mathbb{E}[\mathbf{C}\mathbf{M}_{\mathcal{D}}^{-\frac{1}{2}}\boldsymbol{\mathcal{K}}_{\mathcal{D}}\mathbf{M}_{\mathcal{D}}^{-\frac{1}{2}}\mathbf{C}^{\top}] = \mathbf{C}\mathbf{M}_{\mathcal{D}}^{-\frac{1}{2}}\mathbb{E}\left[\boldsymbol{\mathcal{K}}_{\mathcal{D}}\right]\mathbf{M}_{\mathcal{D}}^{-\frac{1}{2}}\mathbf{C}^{\top} = \boldsymbol{\Sigma}. \tag{2.8}$$

This implies that

$$\mathbf{C}_{\Omega} = \mathbf{C}\mathbf{M}_{\mathcal{D}}^{-\frac{1}{2}} \Leftrightarrow \mathbf{C} = \mathbf{C}_{\Omega}\mathbf{M}_{\mathcal{D}}^{\frac{1}{2}} \text{ and } \boldsymbol{\Omega} = \mathbf{C}\mathbf{M}_{\mathcal{D}}^{-1}\mathbf{C}^{\top} \Leftrightarrow \boldsymbol{\Sigma} = \mathbf{C}_{\Omega}\mathbf{M}_{\mathcal{D}}\mathbf{C}_{\Omega}. \tag{2.9}$$

The expectations $\mathbf{M}_{\mathcal{D}}$ are straightforward to derive by applying Theorem 2.2.1 (p. 19). They are listed in Table 2.3. Notice that they are all diagonal matrices, so if all its diagonal elements are non-negative, $\mathbf{M}_{\mathcal{D}}^{-1/2}$ is uniquely defined. If an element is negative while the conditions for the existence of the distributions are fulfilled, then the expected value $\boldsymbol{\Sigma}$ does not exist. In this paper, we assume that the expected value always exists, and we can thus equivalently characterize the distribution with $\boldsymbol{\Sigma}$ and write

$$\mathbf{R} \sim \mathcal{D}(\boldsymbol{\Sigma}, \boldsymbol{\theta}_{\mathcal{D}}). \tag{2.10}$$

Theorem 2.2.1. *Let* $\underset{\sim}{\mathbf{B}}$ *and* $\bar{\mathbf{B}}$ *be defined as in equations (2.2) and (2.3). Then*

$$\begin{aligned}
\mathbb{E}[\underset{\sim}{\mathbf{B}}\underset{\sim}{\mathbf{B}}^\top] &= \mathrm{dg}(\mathbf{n}), \textit{ (Díaz-García, 2013)} \\
\mathbb{E}[\left(\bar{\mathbf{B}}\bar{\mathbf{B}}^\top\right)^{-1}] &= \mathrm{dg}(\mathring{\boldsymbol{\nu}}), \textit{ (Louati and Masmoudi, 2015)} \\
\mathbb{E}[\bar{\mathbf{B}}^{-\top}\underset{\sim}{\mathbf{B}}\underset{\sim}{\mathbf{B}}^\top\bar{\mathbf{B}}^{-1}] &= \mathrm{dg}(\mathring{\mathbf{n}}), \textit{ (Blasques et al., 2021) and} \\
\mathbb{E}[\underset{\sim}{\mathbf{B}}\left(\bar{\mathbf{B}}\bar{\mathbf{B}}^\top\right)^{-1}\underset{\sim}{\mathbf{B}}^\top] &= \mathrm{dg}(\mathring{\mathring{\mathbf{n}}}),
\end{aligned}$$

with the $p \times 1$ *vectors*[a]

$$\begin{aligned}
\mathbf{n} &= (n_1, n_2, \ldots, n_p)^\top, \\
\mathring{\boldsymbol{\nu}} &= (\mathring{\nu}_1, \mathring{\nu}_2, \ldots, \mathring{\nu}_p)^\top, \\
\mathring{\nu}_i &= \begin{cases} \frac{1}{\nu_i - p - 1}, & \textit{for } i = 1 \\ \frac{1}{\nu_i - p + i - 2}\left(1 + \sum_{j=1}^{i-1} \mathring{\nu}_j\right) & \textit{for } i = 2, \ldots, p, \end{cases} \qquad (2.11) \\
\mathring{\mathbf{n}} &= (\mathring{n}_1, \mathring{n}_2, \ldots, \mathring{n}_p)^\top, \\
\mathring{n}_i &= \begin{cases} \frac{n_1}{\nu_1 - p - 1}, & \textit{for } i = 1 \\ \frac{1}{\nu_i - p + i - 2}\left(n_i + \sum_{j=1}^{i-1} \mathring{n}_j\right) & \textit{for } i = 2, \ldots, p \end{cases} \textit{ and} \qquad (2.12) \\
\mathring{\mathring{\mathbf{n}}} &= (\mathring{\mathring{n}}_1, \mathring{\mathring{n}}_2, \ldots, \mathring{\mathring{n}}_p)^\top \\
\mathring{\mathring{n}}_i &= \begin{cases} n_1 \, \mathring{\nu}_1, & \textit{for } i = 1 \\ \sum_{j=1}^{i-1} \mathring{\nu}_j + (n_i - i + 1) \, \mathring{\nu}_i, & \textit{for } i = 2, \ldots, p. \end{cases} \qquad (2.13)
\end{aligned}$$

Proof in Appendix.

a. If $\forall i$, $n_i = n$, $\nu_i = \nu$ then $\mathring{\nu}_i = (\nu - p - 1)^{-1}$ and $\mathring{n}_i = \mathring{\mathring{n}}_i = n(\nu - p - 1)^{-1}$. For $p = 1$ we obtain the expectations of a χ^2_n, an inverse χ^2_ν, and the χ^2_n/χ^2_ν ratio distribution.

The Individual Distributions 2.2.2

To understand the various distributions' stochastic properties, we now analyze and compare their stochastic representation kernels (Table 2.1). There are many properties amongst which we could compare the distributions. We focus on the marginal distribution of the diagonal elements (RVs) since the RVs are of particular interest. Furthermore we focus on the dependence amongst all elements in the stochastic representation kernels. We start with the Wishart distribution ($\mathcal{W}$),

which we treat as our baseline case,

$$\underline{\mathcal{B}}\underline{\mathcal{B}}^\top.$$

The diagonal elements of the stochastic representation kernel of the Wishart follow independent χ^2_n-distributions, which is easy to see by noticing that

$$(\underline{\mathcal{B}}\underline{\mathcal{B}}^\top)_{ii} = \sum_{k=1}^{p}(\underline{\mathcal{B}})_{ik}(\underline{\mathcal{B}})_{ki} = \sum_{k=1}^{p}(\underline{\mathcal{B}})^2_{ik} = \sum_{k=1}^{i}(\underline{\mathcal{B}})^2_{ik}, \tag{2.14}$$

which is a sum of a χ^2_{n-i+1} and $i-1$ independent $(N(0,1))^2$ random variables, which yields a χ^2_n random variable. The Riesz distribution ($\mathcal{R}$),

$$\underline{\mathbf{B}}\underline{\mathbf{B}}^\top,$$

adds flexibility by allowing for different d.o.f. parameters of the χ-distributions on the main diagonal of the Bartlett matrix, which leads to the diagonal elements being distributed as $(\underline{\mathbf{B}}\underline{\mathbf{B}}^\top)_{ii} \overset{iid}{\sim} \chi^2_{n_i}$. The t-Wishart ($t\mathcal{W}$) has stochastic representation kernel

$$(\bar{b})^{-2}\underline{\mathcal{B}}\underline{\mathcal{B}}^\top,$$

where the term $(\bar{b})^{-2}$ corresponds to an inverse χ^2_ν random variable. It is immediately obvious, that the scalar multiplication of every element in $\underline{\mathcal{B}}\underline{\mathcal{B}}^\top$ by $(\bar{b})^{-2}$ creates much more dependence among the individual elements as for the Wishart distribution. In particular, the diagonal elements of the stochastic representation kernel are now *dependent* random variables following a ratio distribution $\sim \chi^2_n/\chi^2_\nu$, where the χ^2_ν random variable is common across all diagonal elements and the χ^2_n random variable is specific to the index i. This implies that if one diagonal element has a large realization, the others are likely to have a large realization as well. Again, the Riesz version of the distribution, that is the t-Riesz ($t\mathcal{R}$) with stochastic representation kernel

$$(\bar{b})^{-2}\underline{\mathbf{B}}\underline{\mathbf{B}}^\top,$$

adds flexibility by allowing for the $\chi^2_{n_i}$ random variables to have different d.o.f. parameters. The F distribution has stochastic representation kernel

$$\bar{\mathcal{B}}^{-\top}\underline{\mathcal{B}}\underline{\mathcal{B}}^\top\bar{\mathcal{B}}^{-1}.$$

Note that it is related to the t-Wishart in the sense that if the $p \times p$ Bartlett matrix $\bar{\mathcal{B}}$ was of dimension 1×1, i.e. a scalar, the stochastic representation kernel

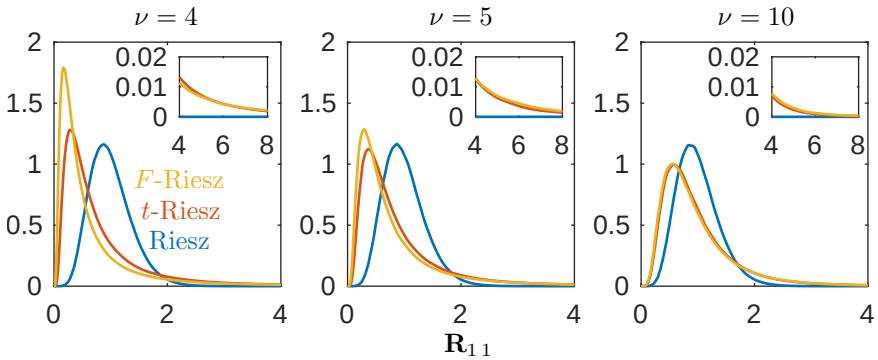

Figure 2.1: Marginal p.d.f.s of the first RV ($\mathbf{R}_{1\,1}$) implied by the Riesz, t-Riesz, and F-Riesz distributions for a two-dimensional RC with the parameter setting $\mathbb{E}[\mathbf{R}] = \mathbf{I}_2$, $\mathbf{n} = (15, 30)^\top$, and with three different settings for the d.o.f. parameters, $\nu = 4, 5, 10$ (t-Riesz) and $\boldsymbol{\nu} = (4, 4)^\top, (5, 5)^\top, (10, 10)^\top$ (F-Riesz).

of the t-Wishart would arise. However, it is important to understand, that the F distribution does *not* nest the t-Wishart and can thus *not* be thought of as a more flexible version of it. Among the notable differences between the two distributions is that the t-Wishart features much higher dependence among the elements in the stochastic representation kernel. This is easy to see as for the t-Wishart every element in $\underset{\sim}{\mathcal{B}}\underset{\sim}{\mathcal{B}}^\top$ is scaled by $(\bar{b})^{-2}$, thus all elements are influenced by one random variable. For the F on the other hand $\underset{\sim}{\mathcal{B}}\underset{\sim}{\mathcal{B}}^\top$ is scaled by $\bar{\mathcal{B}}$ which itself consists of $p(p+1)/2$ independent random variables. The Riesz version of the F distribution, called the F-Riesz ($F\mathcal{R}$) adds flexibility by allowing the $\chi^2_{n_i}$ and $\chi^2_{\nu_i}$ random variables to have different d.o.f. parameters, thus has stochastic representation

$$\bar{\mathbf{B}}^{-\top}\underset{\sim}{\mathbf{B}}\underset{\sim}{\mathbf{B}}^\top\bar{\mathbf{B}}^{-1}.$$

To illustrate the differences between the Riesz, t-Riesz, and F-Riesz we created Figures 2.1 and 2.2. Figure 2.1 plots the marginal p.d.f.s of the first RV ($\mathbf{R}_{1\,1}$) implied by each distribution for the parameter setting $\mathbb{E}[\mathbf{R}] = \mathbf{I}_2$, $\mathbf{n} = (15, 30)^\top$, with three different settings for the d.o.f. parameters, $\nu = 5, 6, 10$ (t-Riesz) and, correspondingly $\boldsymbol{\nu} = (5, 5)^\top, (6, 6)^\top, (10, 10)^\top$ (F-Riesz). We see that the F-Riesz and t-Riesz distributions feature similar marginal distributions for the RV, especially for larger ν. This is not surprising, as the two distributions converge to the same Riesz distribution as ν goes to infinity (see Section 2.2.5). The F-Riesz and t-Riesz distributions have fatter tails than the Riesz, which become smaller with increasing ν. Furthermore, both distributions and have more probability mass on small RVs than the Riesz. In Figure 2.2 we take the setting with $\nu = 5$ (t-Riesz) and $\boldsymbol{\nu} = (5, 5)^\top$ (F-Riesz) from above and plot the marginal joint distribution

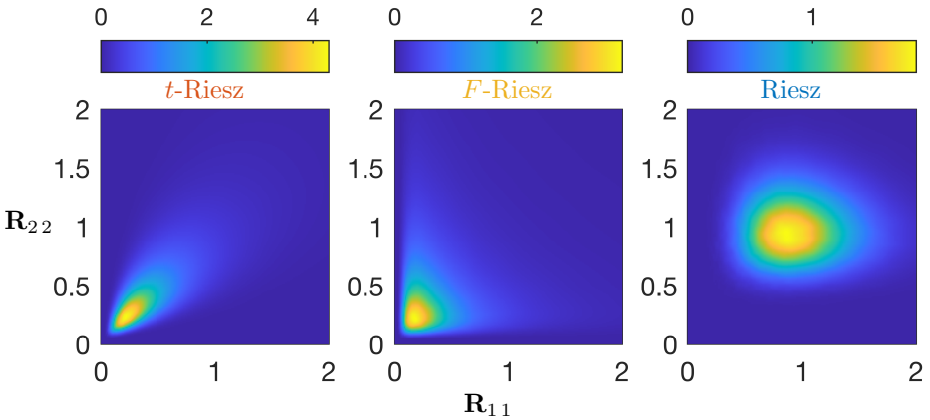

Figure 2.2: Heatmap of marginal joint p.d.f. of the first RV and second RV ($\mathbf{R}_{1\,1}$, $\mathbf{R}_{2\,2}$) implied by the Riesz, t-Riesz, and F-Riesz distributions for a two-dimensional RC with the parameter setting $\mathbb{E}[\mathbf{R}] = \mathbf{I}_2$, $\mathbf{n} = (15, 30)^\top$, $\nu = 5$ (t-Riesz), $\boldsymbol{\nu} = (5, 5)^\top$ (F-Riesz).

of the two RVs ($\mathbf{R}_{1\,1}$, $\mathbf{R}_{2\,2}$). The plots look qualitatively the same for different settings of ν. We first note the Riesz distribution looks drastically different from the others in that most of its probability mass lies around the value of one for the two RVs but the probability mass is less peaked and spread out more (elliptically around the coordinate (1,1)) than for the other two distributions. Comparing subplot one and two, the stark differences between the t-Riesz and F-Riesz distribution now become visible. The t-Riesz distributions' probability mass lies in an elliptical shape around the diagonal from bottom right to top left with a high peak in probability mass on values between 0 and 0.5 for both RVs. This shape implies that the two RVs are more likely to have similar values than drastically different ones for any size of RV realization. In contrast, the F-Riesz probability mass fans out in a triangular fashion from the bottom left corner with more probability mass along the coordinate axes than around the bottom-left to top-right diagonal. It also peaks on values between 0 and 0.5, but the peak is less pronounced than for the t-Riesz and for these small values the probability mass is much more spread out. Furthermore, we see for the F-Riesz, that if one RV is small the probability of the other being small is not nearly as high as for the t-Riesz and there lies substantial probability mass on large realizations for the other RV. Conversely, if one RV has a tail realization the other is more likely not to have one, a property Blasques et al. (2021) call *tail-heterogeneity*. Consequently we call the t-Riesz distribution *tail-homogeneous*. The correlation between the two RVs is 0 for the Riesz, 0.24

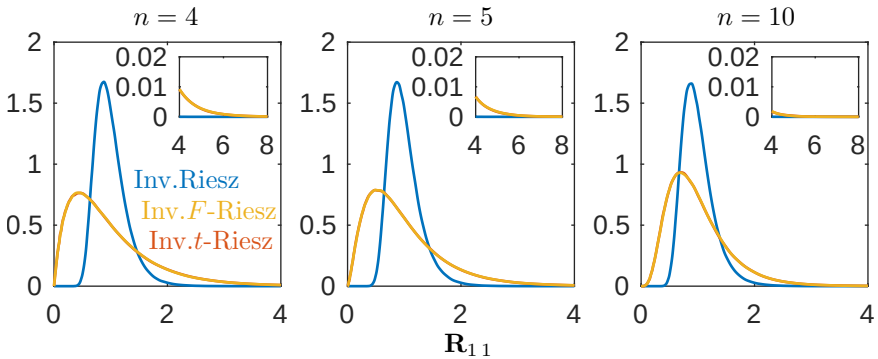

Figure 2.3: Marginal p.d.f.s of the first RV ($\mathbf{R}_{1\,1}$) implied by the Inverse Riesz, Inverse t-Riesz, and Inverse F-Riesz distributions for a two-dimensional RC with the parameter setting $\mathbb{E}[\mathbf{R}] = \mathbf{I}_2$, $\boldsymbol{\nu} = (30, 15)^\top$, and with three different settings for the d.o.f. parameters, $n = 4, 5, 10$ (Inverse t-Riesz) and $\mathbf{n} = (4, 4)^\top, (5, 5)^\top, (10, 10)^\top$ (Inverse F-Riesz).

for the F-Riesz distribution and 0.87 for the t-Riesz. All these observations are in line with the intuition we gained above from analyzing the stochastic representation kernels. The above analysis makes it plausible that the t-Riesz distribution family would perform better during market-wide crises where all assets experience high volatility, whereas the F-Riesz distribution could offer benefits when only a particular asset or subsections of the market experience distress.

To complete the analysis of all distribution, we still have to consider the inverse versions of the distributions mentioned so far. Unfortunately, it is difficult to gain an intuition of the stochastic properties of inverse distributions since due to the inversion of the Bartlett matrices, the marginal distributions of the elements in $\boldsymbol{\mathcal{K}}_{\mathcal{D}}$ and their dependencies are not easily derived. However, we can visualize the same marginal distributions as for non-inverted distributions. Figure 2.3 plots the marginal p.d.f.s of the first RV ($\mathbf{R}_{1\,1}$) implied by the Inverse Riesz, Inverse t-Riesz, and Inverse F-Riesz distributions for the parameter setting $\mathbb{E}[\mathbf{R}] = \mathbf{I}_2$, $\boldsymbol{\nu} = (30, 15)^\top$, with three different settings for the d.o.f. parameters, $n = 5, 6, 10$ (Inverse t-Riesz) and, correspondingly $\mathbf{n} = (5, 5)^\top, (6, 6)^\top, (10, 10)^\top$ (Inverse F-Riesz). We see that the marginal distribution of the RV is almost identical for the Inverse F-Riesz and Inverse t-Riesz distributions. The Inverse F-Riesz and Inverse t-Riesz distributions have is more skewed and has fatter tails than the Inverse Riesz, which become smaller with increasing n. It is also noteworthy, that the Inverse Riesz, compared to the Riesz in Figure 2.1, has larger tails. Even though the Inverse F-Riesz and Inverse t-Riesz have almost identical marginal distributions for the RV, their joint marginal distributions of the first and second

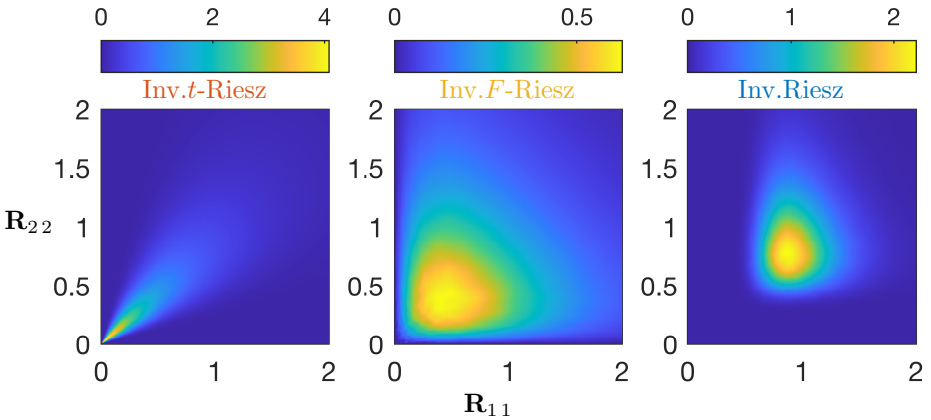

Figure 2.4: Heatmap of marginal joint p.d.f. of the first RV and second RV ($\mathbf{R}_{1\,1}$, $\mathbf{R}_{2\,2}$) implied by the Inverse Riesz, Inverse t-Riesz, and Inverse F-Riesz distributions for a two-dimensional RC with the parameter setting $\mathbb{E}[\mathbf{R}] = \mathbf{I}_2$, $\boldsymbol{\nu} = (30, 15)^\top$, $n = 5$ (Inverse t-Riesz), $\mathbf{n} = (5, 5)^\top$ (Inverse F-Riesz).

RV ($\mathbf{R}_{1\,1}$ and $\mathbf{R}_{2\,2}$) differ extremely, as we can see in Figure 2.4. In particular, the Inverse t-Riesz's probability mass is very concentrated around the lower-left to upper-right diagonal with a very high peak on values below 0.3 for both RVs. As for the t-Riesz's this implies that $\mathbf{R}_{1\,1}$ and $\mathbf{R}_{2\,2}$ are highly correlated. The Inverse F-Riesz's probability mass is spread out in a triangular fashion with a much lower peak than the Inverse t-Riesz. The correlation between $\mathbf{R}_{1\,1}$ and $\mathbf{R}_{2\,2}$ is 0.70, 0.017, and 0.023 for the Inverse t-Riesz, Inverse F-Riesz, and Inverse Riesz, respectively. It is noteworthy, that the Inverse F-Riesz, compared to the F-Riesz in Figure 2.2, is more spread out and features much lower correlation between the RVs.

2.2.3 Fat-Tailedness

In the literature, fat-tailedness of RCs has been measured by or considered synonymous to fat-tailedness of their diagonal elements, i.e. the RVs (c.f. Opschoor et al., 2018 and Blasques et al., 2021). We extend this interpretation by taking the log-determinant of the RCs as the quantity to determine their fat-tailedness because the determinant of a matrix can be geometrically interpreted as its volume. Naturally, we call those matrix-variate distributions fat-tailed that imply fat-tailed marginal distributions for the log-determinant of their random matrices. Conveniently, for random matrices from any distribution for RCs, the log-determinant

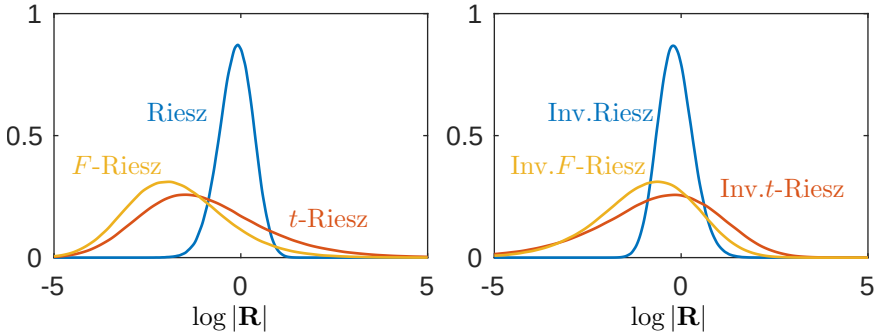

Figure 2.5: Marginal p.d.f. of the log-determinant of the RC for the (Inverse) Riesz, (Inverse) t-Riesz, and (Inverse) F-Riesz distributions for a two-dimensional RC with the parameter setting $\mathbb{E}[\mathbf{R}] = \mathbf{I}_2$, and for the non-inverted distributions, $\mathbf{n} = (15, 30)^\top$, $\nu = 5$ (t-Riesz), $\boldsymbol{\nu} = (5, 5)^\top$ (F-Riesz), and for the inverse distributions, $\boldsymbol{\nu} = (30, 15)^\top$, $n = 5$ (Inverse t-Riesz), $\mathbf{n} = (5, 5)^\top$ (Inverse F-Riesz).

equals the sum of log-χ random variables. This can be seen from the stochastic representations (Table 2.1). For example, for the F-Riesz distribution, the stochastic representation of the log-determinant, omitting the subscript t, is equal to

$$\log|\mathbf{R}| = \log|\mathbf{C}_\Omega \bar{\mathbf{B}}^{-\top} \underline{\mathbf{B}} \underline{\mathbf{B}}^\top \bar{\mathbf{B}}^{-1} \mathbf{C}_\Omega^\top| \tag{2.15}$$

$$= 2\sum_{i=1}^{p} \log(\mathbf{C}_\Omega)_{ii} + 2\sum_{i=1}^{p} \log(\chi_{n_i - i + 1}) - 2\sum_{i=1}^{p} \log(\chi_{\nu_i - p + i}), \tag{2.16}$$

and for the t-Riesz distribution it is

$$\log|\mathbf{R}| = \log|\mathbf{C}_\Omega (\bar{b})^{-2} \underline{\mathbf{B}} \underline{\mathbf{B}}^\top \mathbf{C}_\Omega^\top| \tag{2.17}$$

$$= 2\sum_{i=1}^{p} \log(\mathbf{C}_\Omega)_{ii} + 2\sum_{i=1}^{p} \log(\chi_{n_i - i + 1}) - 2\,p \log(\chi_\nu). \tag{2.18}$$

Thus for all distributions for RCs the marginal distribution of the log-determinant can be theoretically derived along with its tail properties. A distribution $\mathcal{D}_1$ is then considered to be more fat-tailed than another, $\mathcal{D}_2$, if there exists some real number x, such that for all real $y > x$, the p.d.f.s are such that $p_{\log|\mathbf{R}|,\mathcal{D}_1}(y) > p_{\log|\mathbf{R}|,\mathcal{D}_2}(y)$. For the case of 1×1 random matrices, i.e. random variables, this definition corresponds to the classical definition of fat-tailedness for random variables as the determinant of a random variable is simply the random variable itself. We leave for future research the further theoretical investigation of fat-tailedness for RCs.

In Figure 2.5 we plot the marginal p.d.f.s of the log-determinant implied by the Riesz-type distributions. We see, that the (Inverse) F-Riesz and (Inverse) t-Riesz distributions can indeed be considered fat-tailed, as they feature fatter right tails than the (Inverse) Riesz distribution. It is noteworthy that the t-Riesz distribution features a fatter right tail than the F-Riesz.

2.2.4 Asset Ordering

For all Riesz-type distributions, a different ordering of the assets in the RCs yields a different version of the respective probability distribution. That is, if we assume $\mathbf{R} \sim \mathcal{D}(\boldsymbol{\Omega}, \boldsymbol{\theta}_{\mathcal{D}})$, then any other ordering the assets, $\mathbf{PRP}^\top$, where $\mathbf{P}$ denotes an arbitrary permutation matrix[7], has stochastic representation $\mathbf{PC}_\Omega \boldsymbol{\mathcal{K}}_{\mathcal{D}}(\mathbf{C}_\Omega)^\top \mathbf{P}^\top$.[8] This new stochastic representation yields *for Riesz-type distributions* a different and generally unknown probability distribution, which we denote by $\mathcal{D}_{\mathbf{P}}(\boldsymbol{\Omega}, \boldsymbol{\theta}_{\mathcal{D}})$. The next section covers a special permutation matrix for which we do know the probability distributions.

In practice, we are given a randomly ordered RC $\mathbf{PRP}^\top$ for which the model $\mathcal{D}(\boldsymbol{\Omega}, \boldsymbol{\theta}_{\mathcal{D}})$ is only correctly specified if, by chance, the random ordering corresponds to the true one ($\mathbf{P} = \mathbf{I}$). To recover the true ordering, we can treat the ordering as a parameter to optimize over. Blasques et al. (2021) have proposed an efficient algorithm to optimize the likelihood for many different orderings[9] and then choosing the ordering with the highest estimated likelihood value. In a simulation experiment, they find that their algorithms' estimated likelihood values come close to the likelihood value of the true data-generating process and that the ordering of the assets gets close to the true ordering.

Riesz Distribution Versions

For any Riesz-type distribution there are two versions; we call them version-I and version-II (c.f. Blasques et al., 2021, who call them type-I and type-II, respectively). In this subsection we contribute to the literature by showing that the two versions are closely related since assuming $\mathbf{R}$ follows a version-I Riesz distribution is equivalent to assuming that $\mathbf{R}$ with the asset order reversed, denoted by $\overleftarrow{\mathbf{R}}$, follows the corresponding version-II distribution with "reversed" parameters. See Table 2.4 for the exact relationships.

7. A permutation matrix is a square matrix that has exactly one entry of one in each row and each column and zeros elsewhere.

8. Recall the stochastic representation of $\mathbf{R}$ is $\mathbf{R} = \mathbf{C}_\Omega \boldsymbol{\mathcal{K}}_{\mathcal{D}}(\mathbf{C}_\Omega)^\top$.

9. As the number of possible orderings explodes with increasing p, trying *all* possible orderings is infeasible.

Our versions		Alternative versions
$\mathbf{R} \sim \mathcal{R}^I(\mathbf{\Omega}, \mathbf{n})$	$\Leftrightarrow$	$\overleftarrow{\mathbf{R}} \sim \mathcal{R}^{II}(\overleftarrow{\mathbf{\Omega}}, \overleftarrow{\mathbf{n}})$
$\mathbf{R} \sim i\mathcal{R}^{II}(\mathbf{\Omega}, \boldsymbol{\nu})$	$\Leftrightarrow$	$\overleftarrow{\mathbf{R}} \sim i\mathcal{R}^{I}(\overleftarrow{\mathbf{\Omega}}, \overleftarrow{\boldsymbol{\nu}})$
$\mathbf{R} \sim t\mathcal{R}^I(\mathbf{\Omega}, \mathbf{n}, \nu)$	$\Leftrightarrow$	$\overleftarrow{\mathbf{R}} \sim t\mathcal{R}^{II}(\overleftarrow{\mathbf{\Omega}}, \nu, \overleftarrow{\mathbf{n}})$
$\mathbf{R} \sim it\mathcal{R}^{II}(\mathbf{\Omega}, n, \boldsymbol{\nu})$	$\Leftrightarrow$	$\overleftarrow{\mathbf{R}} \sim it\mathcal{R}^{I}(\overleftarrow{\mathbf{\Omega}}, \overleftarrow{\boldsymbol{\nu}}, n)$
$\mathbf{R} \sim F\mathcal{R}^I(\mathbf{\Omega}, \mathbf{n}, \boldsymbol{\nu})$	$\Leftrightarrow$	$\overleftarrow{\mathbf{R}} \sim F\mathcal{R}^{II}(\overleftarrow{\mathbf{\Omega}}, \overleftarrow{\boldsymbol{\nu}}, \overleftarrow{\mathbf{n}})$
$\mathbf{R} \sim iF\mathcal{R}^{II}(\mathbf{\Omega}, \mathbf{n}, \boldsymbol{\nu})$	$\Leftrightarrow$	$\overleftarrow{\mathbf{R}} \sim iF\mathcal{R}^{I}(\overleftarrow{\mathbf{\Omega}}, \overleftarrow{\boldsymbol{\nu}}, \overleftarrow{\mathbf{n}})$

Table 2.4: Equivalence between the Riesz distribution versions. In this paper we choose version-I for the non-inverted distribution and version-II for the inverse ones. Outside of this subsection we do not refer to the different versions and consequently drop the version superscripts for better readability.

To derive the equivalence we have to introduce some concepts. First note that the permutation matrix which achieves a reversal of the asset order in $\mathbf{R}$ is the is the exchange matrix, i.e. a matrix with ones on the diagonal from the upper right-hand corner to the lower left-hand corner and zeros elsewhere,

$$\mathbf{P}_e = \begin{bmatrix} 0 & & 1 \\ & \iddots & \\ 1 & & 0 \end{bmatrix}.$$

We can visualize $\overleftarrow{\mathbf{R}} := \mathbf{P}_e \mathbf{R} \mathbf{P}_e^\top$ as the original matrix rotated by 180 degrees. Furthermore, note that $\mathbf{P}_e \bar{\mathbf{B}} \mathbf{P}_e$ is equal to $\underline{\mathbf{B}}$ but with degrees of freedom $\overleftarrow{\boldsymbol{\nu}} = (\nu_p, \nu_{p-1}, \ldots, \nu_1)$ instead of $\mathbf{n}$. We denote it by $\underline{\mathbf{B}}_{\overleftarrow{\boldsymbol{\nu}}}$. Similarly, $\mathbf{P}_e \underline{\mathbf{B}} \mathbf{P}_e$ is equal to $\bar{\mathbf{B}}$ but with $\overleftarrow{\mathbf{n}}$ instead of $\boldsymbol{\nu}$, denoted by $\bar{\mathbf{B}}_{\overleftarrow{\mathbf{n}}}$. Next, note that $\mathbf{P}_e = \mathbf{P}_e^\top$, and $\mathbf{P}_e \mathbf{P}_e = \mathbf{I}$, such that $\mathbf{P}_e = \mathbf{P}_e^{-1}$. Finally, note that we can write the reverse order $\mathbf{\Omega}$ as

$$\overleftarrow{\mathbf{\Omega}} = \mathbf{P}_e \mathbf{\Omega} \mathbf{P}_e = \mathbf{P}_e \mathbf{C}_\Omega \mathbf{P}_e \mathbf{P}_e \mathbf{C}_\Omega^\top \mathbf{P}_e = \mathbf{U}_{\overleftarrow{\mathbf{\Omega}}} \mathbf{U}_{\overleftarrow{\mathbf{\Omega}}}^\top,$$

where $\mathbf{P}_e \mathbf{C}_\Omega \mathbf{P}_e = \mathbf{U}_{\overleftarrow{\mathbf{\Omega}}}$ is the upper Cholesky factor of $\overleftarrow{\mathbf{\Omega}}$, since the decomposition of a symmetric positive definite matrix into an upper triangular matrix post-multiplied by its transpose, i.e. the Cholesky decomposition, is unique.

We now show the equivalence between the two versions using the example of the F-Riesz distribution. Assume that $\mathbf{R}$ follows our version-I F-Riesz distribution,

$\mathbf{R} \sim F\mathcal{R}(\mathbf{\Omega}, \mathbf{n}, \boldsymbol{\nu})$. Then

$$\begin{aligned}
\mathbf{P}_e\mathbf{R}\mathbf{P}_e &= \mathbf{P}_e\mathbf{C}_\Omega\bar{\mathbf{B}}^{-\top}\underline{\mathbf{B}}\underline{\mathbf{B}}^\top\bar{\mathbf{B}}^{-1}\mathbf{C}_\Omega^\top\mathbf{P}_e \\
&= \mathbf{P}_e\mathbf{C}_\Omega\mathbf{P}_e\mathbf{P}_e\bar{\mathbf{B}}^{-\top}\mathbf{P}_e\mathbf{P}_e\underline{\mathbf{B}}\underline{\mathbf{B}}\mathbf{P}_e\mathbf{P}_e\underline{\mathbf{B}}^\top\mathbf{P}_e\mathbf{P}_e\bar{\mathbf{B}}^{-1}\mathbf{P}_e\mathbf{P}_e\mathbf{C}_\Omega^\top\mathbf{P}_e \\
&= \mathbf{P}_e\mathbf{C}_\Omega\mathbf{P}_e(\mathbf{P}_e\bar{\mathbf{B}}\mathbf{P}_e)^{-\top}\mathbf{P}_e\underline{\mathbf{B}}\underline{\mathbf{B}}\mathbf{P}_e\mathbf{P}_e\underline{\mathbf{B}}^\top\mathbf{P}_e(\mathbf{P}_e\bar{\mathbf{B}}\mathbf{P}_e)^{-1}(\mathbf{P}_e\mathbf{C}_\Omega\mathbf{P}_e)^\top \\
&= \mathbf{U}_{\overleftarrow{\mathbf{\Omega}}}\underline{\mathbf{B}}_{\overleftarrow{\boldsymbol{\nu}}}^{-\top}\bar{\mathbf{B}}_{\overleftarrow{\mathbf{n}}}\bar{\mathbf{B}}_{\overleftarrow{\mathbf{n}}}^\top\underline{\mathbf{B}}_{\overleftarrow{\boldsymbol{\nu}}}^{-1}\mathbf{U}_{\overleftarrow{\mathbf{\Omega}}}^\top,
\end{aligned}$$

which is the stochastic representation of the version-II F-Riesz distribution with parameters $\overleftarrow{\mathbf{\Omega}}$, $\overleftarrow{\boldsymbol{\nu}}$, and $\overleftarrow{\mathbf{n}}$, $F\mathcal{R}^{II}(\overleftarrow{\mathbf{\Omega}}, \overleftarrow{\boldsymbol{\nu}}, \overleftarrow{\mathbf{n}})$ (Theorem 8 in Blasques et al., 2021). We can derive the distributions of $\overleftarrow{\mathbf{R}}$ for all other Riesz-type distributions in similar fashion. Regarding the stochastic representation Kernel $\boldsymbol{\mathcal{K}}_{\mathcal{D}}$ we come to the conclusion, that the alternative versions not used in this paper (Table 2.4) are equal to the ones in our stochastic representation (Table 2.1) where $\underline{\mathbf{B}}$ and $\bar{\mathbf{B}}$ (and their special cases) are interchanged.

Empirically it does not matter which distribution version we assume since, as we mentioned above, we take the asset order as a parameter to optimize over. That is, assuming one version and estimating it via maximum likelihood should yield the "reversed" estimates for the asset order and the parameter values as compared to assuming the other version.

Inverse Distribution Versions

As you can see in Table 2.4, in this paper we choose the version-II distributions for the inverted ones and the version-I distributions for the non-inverted ones. Why we choose to switch versions will become clear below. First, note that assuming that $\mathbf{R}$ follows a version-I (version-II) inverse distribution with parameter matrix $\mathbf{\Omega}^{-1}$ is equivalent to assuming that $\mathbf{R}^{-1}$ follows the corresponding version-I (version-II) non-inverted distribution with parameter matrix $\mathbf{\Omega}$. This is how the inverse distributions are defined (e.g. Blasques et al., 2021).

As an example, lets take a version-I Riesz distribution, $\mathbf{R} \sim \mathcal{R}(\mathbf{\Omega}^{-1}, \mathbf{n})$, and use the stochastic representation to derive its inverse, the version-I Inverse Riesz distribution, as defined above.

$$\begin{aligned}
\mathbf{R} &= \mathbf{C}_{\mathbf{\Omega}^{-1}}\underline{\mathbf{B}}\underline{\mathbf{B}}^\top\mathbf{C}_{\mathbf{\Omega}^{-1}}^\top \\
\Leftrightarrow \mathbf{R}^{-1} &= \mathbf{C}_{\mathbf{\Omega}^{-1}}^{-\top}\underline{\mathbf{B}}^{-\top}\underline{\mathbf{B}}^{-1}\mathbf{C}_{\mathbf{\Omega}^{-1}}^{-1}.
\end{aligned}$$

So the stochastic representation of the version-I Inverse Riesz distribution is $\mathbf{C}_{\mathbf{\Omega}^{-1}}^{-\top}\underline{\mathbf{B}}^{-\top}\underline{\mathbf{B}}^{-1}\mathbf{C}_{\mathbf{\Omega}^{-1}}^{-1}$. The important thing to note here is, that this stochastic representation features $\mathbf{C}_{\mathbf{\Omega}^{-1}}^{-\top} = \mathbf{U}_{\mathbf{\Omega}}$, which is the *upper* Cholesky factor of $\mathbf{\Omega}$.

In fact, we can derive in similar fashion, that the stochastic representations of all inverse version-I distributions feature the upper Cholesky factor of $\mathbf{\Omega}$, while all inverse version-II distributions use the lower Cholesky factor. Thus, in order to have *one* general stochastic representation formula based on the lower Cholesky factor that is valid for all distributions in our paper, i.e. equation (2.4) ($\mathbf{R} = \mathbf{C}_\Omega \boldsymbol{\mathcal{K}}_\mathcal{D} \mathbf{C}_\Omega^\top$), and since empirically it does not matter which distribution version we choose (see above), we use the version-II for the inverse and version-I for the non-inverted ones.

Distribution Relationships 2.2.5

Figure 2.6 shows how the distributions are related to each other. Every Wishart-type distribution is a special case of its Riesz-type counterpart and is obtained by setting the entries in each d.o.f. parameter vector all equal to each other. This is easily seen from the stochastic representations in Table 2.1, where the Wishart-type distributions have the same stochastic representation kernel as their Riesz-type counterparts but with $\underline{\mathcal{B}}$ and $\bar{\mathcal{B}}$ instead of $\underline{\mathbf{B}}$ and $\bar{\mathbf{B}}$. The proofs for the dashed arrows are also immediately evident from the stochastic representations since in the case of $\bar{\mathbf{B}}$ ($\underline{\mathbf{B}}$) being one-dimensional, the stochastic representation of the (Inverse) F-Riesz distribution reduces to the one of the (Inverse) t-Riesz distribution. We derive the remaining relationships in the following Theorem 2.2.2.

Theorem 2.2.2.

$$\mathbf{C}\mathbf{M}_{F\mathcal{R}}^{-\frac{1}{2}} \boldsymbol{\mathcal{K}}_{F\mathcal{R}} \mathbf{M}_{F\mathcal{R}}^{-\frac{1}{2}} \mathbf{C}^\top \xrightarrow[\forall i:\nu_i \to \infty]{d} \mathbf{C}\mathbf{M}_{\mathcal{R}}^{-\frac{1}{2}} \boldsymbol{\mathcal{K}}_{\mathcal{R}} \mathbf{M}_{\mathcal{R}}^{-\frac{1}{2}} \mathbf{C}^\top$$

$$\mathbf{C}\mathbf{M}_{iF\mathcal{R}}^{-\frac{1}{2}} \boldsymbol{\mathcal{K}}_{iF\mathcal{R}} \mathbf{M}_{iF\mathcal{R}}^{-\frac{1}{2}} \mathbf{C}^\top \xrightarrow[\forall i:n_i \to \infty]{d} \mathbf{C}\mathbf{M}_{i\mathcal{R}}^{-\frac{1}{2}} \boldsymbol{\mathcal{K}}_{i\mathcal{R}} \mathbf{M}_{i\mathcal{R}}^{-\frac{1}{2}} \mathbf{C}^\top$$

$$\mathbf{C}\mathbf{M}_{t\mathcal{R}}^{-\frac{1}{2}} \boldsymbol{\mathcal{K}}_{t\mathcal{R}} \mathbf{M}_{t\mathcal{R}}^{-\frac{1}{2}} \mathbf{C}^\top \xrightarrow[\nu \to \infty]{d} \mathbf{C}\mathbf{M}_{\mathcal{R}}^{-\frac{1}{2}} \boldsymbol{\mathcal{K}}_{\mathcal{R}} \mathbf{M}_{\mathcal{R}}^{-\frac{1}{2}} \mathbf{C}^\top$$

$$\mathbf{C}\mathbf{M}_{it\mathcal{R}}^{-\frac{1}{2}} \boldsymbol{\mathcal{K}}_{it\mathcal{R}} \mathbf{M}_{it\mathcal{R}}^{-\frac{1}{2}} \mathbf{C}^\top \xrightarrow[n \to \infty]{d} \mathbf{C}\mathbf{M}_{i\mathcal{R}}^{-\frac{1}{2}} \boldsymbol{\mathcal{K}}_{i\mathcal{R}} \mathbf{M}_{i\mathcal{R}}^{-\frac{1}{2}} \mathbf{C}^\top.$$

Proof in Appendix.

As they are just special cases, Theorem 2.2.2 also holds for the corresponding Wishart-type distributions ($F \to \mathcal{W}$ or $i\mathcal{W}$, $t\mathcal{W} \to \mathcal{W}$, $it\mathcal{W} \to i\mathcal{W}$). Finally, note that an Inverse F distribution is again an F distribution with the degrees of freedom parameters switched and the expected value matrix inverted, as is easy to see from their stochastic representations. However, an Inverse F-Riesz distribution is not again an F-Riesz distribution (see Section 2.7.2 in the appendix). We derive the novel Inverse F-Riesz distribution in this paper.

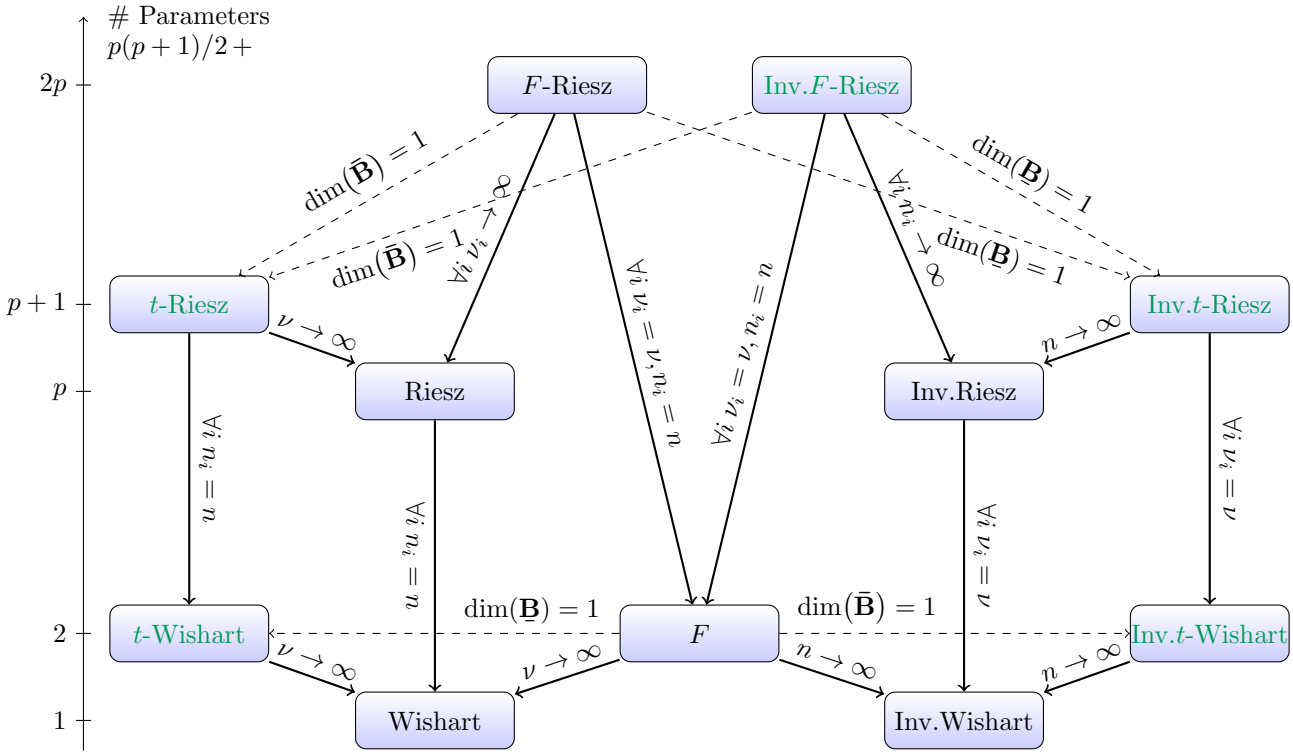

Figure 2.6: Relationships between the standardized probability distributions. Stochastic representations given by equation (2.7) in conjunction with Tables 2.1 and 2.3. Next to every arrow, we indicate how the distributions are related. Dashed arrows indicate a more general relationship; not a nesting. The vertical axis shows the number of parameters of the respective distributions. All distributions have $p(p+1)/2$ distinct parameters in the symmetric positive definite parameter matrix $\mathbf{\Sigma}$ (or $\mathbf{\Omega}$) plus the number of d.o.f. parameters.

We first need to introduce some special functions used in the probability density functions.

Definition 2.2.1 (Generalized Power Function)**.** *Let* $\mathbf{X}$ *be a real* $p \times p$ *matrix,* $\mathbf{n} = (n_1, \ldots, n_p)^\top$ *be a real vector of length* p *and let* $\mathbf{X}_{[i]}$ *denote the square submatrix created by taking the first* i *rows and columns of* $\mathbf{X}$*. Then the generalized power function (a.k.a. highest weight vector), denoted by* $|\mathbf{X}|_\mathbf{n}$ *is defined as*

$$|\mathbf{X}|_\mathbf{n} = |\mathbf{X}_{[1]}|^{n_1 - n_2} |\mathbf{X}_{[2]}|^{n_2 - n_3} \ldots |\mathbf{X}_{[p-1]}|^{n_{p-1} - n_p} |\mathbf{X}|^{n_p}.$$

The generalized power function is defined in e.g. Faraut and Korányi (1994). The determinant-with-subscript notation is taken from Blasques et al. (2021). It makes immediately visible the close relation of the generalized power function to the determinant since for $n_1 = n_2 = \ldots = n_p = n$, $|\mathbf{X}|_\mathbf{n} = |\mathbf{X}|^n$. The next lemma shows that in the case of positive definite $\mathbf{X}$, the generalized power function can be written as a function of the diagonal elements of the lower Cholesky decomposition of $\mathbf{X}$. Blasques et al. (2021) name this special case *power weighted determinant*.

Lemma 2.2.1 (Power Weighted Determinant)**.** *Let* $\mathbf{\Sigma}$ *be positive definite and* $\mathbf{\Sigma} = \mathbf{T}\mathbf{D}\mathbf{T}^\top$ *be the unique decomposition into a lower triangular square matrix with ones on the main diagonal,* $\mathbf{T}$ *and diagonal matrix with positive entries on the diagonal* $\mathbf{D}$*. Then we can rewrite*

$$|\mathbf{\Sigma}|_\mathbf{n} = \prod_{i=1}^{p} \mathbf{D}_{ii}^{n_i} = \prod_{i=1}^{p} \mathbf{C}_{ii}^{2n_i}.$$

Proof in Appendix.

The next lemma lists algebraic equalities for the power weighted determinant. It corresponds to Lemma 3 in Blasques et al. (2021).[10]

10. Part (v) below, where ${}_U|\mathbf{R}|_\mathbf{n}$ denotes their upper power weighted determinant, is only included to see the equivalence between our and their p.d.f. representations.

Lemma 2.2.2 (Lemma 3, Blasques et al., 2021)**.** *Given a scalar n, a vector $\mathbf{n}$ of length p, a vector of ones $\mathbf{1}$ of length p, and a positive definite matrix $\mathbf{R}$, the following identities hold.*

(i) If $\mathbf{n} = n \cdot \mathbf{1}$, then $|\mathbf{R}|_{n\cdot\mathbf{1}} = |\mathbf{R}|^n$

(ii) Let $\mathbf{n}_1, \mathbf{n}_2$ be two vectors of length p, then we have $|\mathbf{R}|_{\mathbf{n}_1} \cdot |\mathbf{R}|_{\mathbf{n}_2} = |\mathbf{R}|_{\mathbf{n}_1+\mathbf{n}_2}$.

(iii) $(|\mathbf{R}|_{\mathbf{n}})^{-1} = |\mathbf{R}|_{-\mathbf{n}}$.

(iv) If $\mathbf{\Sigma} = \mathbf{C}\mathbf{C}^\top$, where $\mathbf{\Sigma}$ is positive definite with lower Cholesky factor $\mathbf{C}$, then $|\mathbf{R}|_{\mathbf{n}} \cdot |\mathbf{\Sigma}|_{-\mathbf{n}} = |\mathbf{C}^{-1}\mathbf{R}\mathbf{C}^{-\top}|_{\mathbf{n}}$. As a special case we have $|\mathbf{C}\mathrm{dg}(\mathbf{n})\mathbf{C}^\top|_{\boldsymbol{\nu}} = \prod_{i=1}^{p} n_i^{\nu_i} |\mathbf{\Sigma}|_{\boldsymbol{\nu}}$.

(v) ${}_U|\mathbf{R}|_{\mathbf{n}} = |\mathbf{R}^{-1}|_{-\mathbf{n}}$.

Next, we define the (multivariate) gamma function as in e.g. equations (5.2.1), (35.3.4) and (35.3.5) of the NIST Digital Library of Mathematical Functions.

Definition 2.2.2 (Multivariate Gamma Function)**.** *Let $\mathbf{n}$ be a real vector of length p. Then the vector-valued multivariate gamma function is defined as*

$$\Gamma_p(\mathbf{n}) = \pi^{p(p-1)/4} \prod_{i=1}^{p} \Gamma\left(n_i - \frac{i-1}{2}\right),$$

with $2n_i > i - 1$, $i = 1, \ldots p$.

Let n be a scalar. Then the scalar-valued multivariate gamma function is defined as

$$\Gamma_p(n) = \pi^{p(p-1)/4} \prod_{i=1}^{p} \Gamma\left(n - \frac{i-1}{2}\right),$$

with $2n > p - 1$.

Obviously if $n_1 = n_2 = \ldots = n_p = n$, then

$$\Gamma_p(\mathbf{n}) = \Gamma_p(n).$$

Now that we have introduced all special functions used in the probability density functions (p.d.f.s), we list the p.d.f.s for all considered probability distributions in Table 2.5. We derive the ones of the (Inverse) t-Riesz distribution and the Inverse

Distribution	Probability Density Function $p_{\mathcal{D}}(\mathbf{R}\|\mathbf{\Omega}, \boldsymbol{\theta}_{\mathcal{D}})$
Wishart	$\frac{1}{2^{np/2}\Gamma_p(n/2)} \|\mathbf{\Omega}\|^{-\frac{n}{2}} \|\mathbf{R}\|^{\frac{n-p-1}{2}} \operatorname{etr}(-\frac{1}{2}\mathbf{\Omega}^{-1}\mathbf{R})$
Riesz	$\frac{1}{2^{p\bar{\mathbf{n}}/2}\Gamma_p(\mathbf{n}/2)} \|\mathbf{\Omega}\|_{-\frac{\mathbf{n}}{2}} \|\mathbf{R}\|_{\frac{\mathbf{n}-p-1}{2}} \operatorname{etr}(-\frac{1}{2}\mathbf{\Omega}^{-1}\mathbf{R})$
Inv.Wishart	$\frac{1}{2^{\nu p/2}\Gamma_p(\nu/2)} \|\mathbf{\Omega}\|^{\frac{\nu}{2}} \|\mathbf{R}\|^{-\frac{\nu+p+1}{2}} \operatorname{etr}(-\frac{1}{2}\mathbf{\Omega}\mathbf{R}^{-1})$
Inv.Riesz	$\frac{1}{2^{p\overleftarrow{\boldsymbol{\nu}}/2}\Gamma_p(\overleftarrow{\boldsymbol{\nu}}/2)} \|\mathbf{\Omega}\|_{\frac{\boldsymbol{\nu}}{2}} \|\mathbf{R}\|_{-\frac{\boldsymbol{\nu}+p+1}{2}} \operatorname{etr}(-\frac{1}{2}\mathbf{\Omega}\mathbf{R}^{-1})$
t-Wishart	$\frac{\Gamma((\nu+pn)/2)}{\Gamma_p(n/2)\Gamma(\nu/2)} \|\mathbf{\Omega}\|^{-\frac{n}{2}} \|\mathbf{R}\|^{\frac{n-p-1}{2}} (1+\operatorname{tr}(\mathbf{\Omega}^{-1}\mathbf{R}))^{-\frac{\nu+pn}{2}}$
t-Riesz	$\frac{\Gamma((\nu+p\bar{\mathbf{n}})/2)}{\Gamma_p(\mathbf{n}/2)\Gamma(\nu/2)} \|\mathbf{\Omega}\|_{-\frac{\mathbf{n}}{2}} \|\mathbf{R}\|_{\frac{\mathbf{n}-p-1}{2}} (1+\operatorname{tr}(\mathbf{\Omega}^{-1}\mathbf{R}))^{-\frac{\nu+p\bar{\mathbf{n}}}{2}}$
Inv.t-Wishart	$\frac{\Gamma((n+p\nu)/2)}{\Gamma(n/2)\Gamma_p(\nu/2)} \|\mathbf{\Omega}\|^{\frac{\nu}{2}} \|\mathbf{R}\|^{-\frac{\nu+p+1}{2}} (1+\operatorname{tr}(\mathbf{\Omega}\mathbf{R}^{-1}))^{-\frac{n+p\nu}{2}}$
Inv.t-Riesz	$\frac{\Gamma((n+p\bar{\boldsymbol{\nu}})/2)}{\Gamma(n/2)\Gamma_p(\overleftarrow{\boldsymbol{\nu}}/2)} \|\mathbf{\Omega}\|_{\frac{\boldsymbol{\nu}}{2}} \|\mathbf{R}\|_{-\frac{\boldsymbol{\nu}+p+1}{2}} (1+\operatorname{tr}(\mathbf{\Omega}\mathbf{R}^{-1}))^{-\frac{n+p\bar{\boldsymbol{\nu}}}{2}}$
F	$\frac{\Gamma_p((n+\nu)/2)}{\Gamma_p(\nu/2)\Gamma_p(n/2)} \|\mathbf{\Omega}\|^{-\frac{n}{2}} \|\mathbf{R}\|^{\frac{n-p-1}{2}} \|\mathbf{I}+\mathbf{C}_\Omega^{-1}\mathbf{R}\mathbf{C}_\Omega^{-\top}\|^{-\frac{n+\nu}{2}}$
F-Riesz	$\frac{\Gamma_p((\overleftarrow{\mathbf{n}}+\overleftarrow{\boldsymbol{\nu}})/2)}{\Gamma_p(\mathbf{n}/2)\Gamma_p(\overleftarrow{\boldsymbol{\nu}}/2)} \|\mathbf{\Omega}\|_{-\frac{\mathbf{n}}{2}} \|\mathbf{R}\|_{\frac{\mathbf{n}-p-1}{2}} \|\mathbf{I}+\mathbf{C}_\Omega^{-1}\mathbf{R}\mathbf{C}_\Omega^{-\top}\|_{-\frac{\mathbf{n}+\boldsymbol{\nu}}{2}}$
Inv.F-Riesz	$\frac{\Gamma_p((\mathbf{n}+\boldsymbol{\nu})/2)}{\Gamma_p(\mathbf{n}/2)\Gamma_p(\overleftarrow{\boldsymbol{\nu}}/2)} \|\mathbf{\Omega}\|_{\frac{\boldsymbol{\nu}}{2}} \|\mathbf{R}\|_{-\frac{\boldsymbol{\nu}+p+1}{2}} \|(\mathbf{I}+\mathbf{C}_\Omega^{\top}\mathbf{R}^{-1}\mathbf{C}_\Omega)^{-1}\|_{\frac{\mathbf{n}+\boldsymbol{\nu}}{2}}$

Table 2.5: Probability density functions of distributions for RCs. Derivations from stochastic representations are in Section 2.7.2 (appendix). $\bar{\mathbf{n}} = \frac{1}{p}\sum_{i=1}^{p} n_i$.

F-Riesz distribution in Theorem 2.2.3, the ones of the (Inverse) Riesz and F-Riesz distributions are given in Theorems 4, 7 and 8 of Blasques et al. (2021). See Section 2.7.2 in the appendix for more details.

Theorem 2.2.3 (Probability Density Functions)**.** *The probability density functions of* $\mathbf{C}_\Omega\mathcal{K}_{\mathcal{D}}\mathbf{C}_\Omega$ *for* $\mathcal{D} \in (t\mathcal{R}, it\mathcal{R}, iF\mathcal{R})$ *obtain as*

$$p_{t\mathcal{R}}(\mathbf{R}|\mathbf{\Omega}, \mathbf{n}, \nu) = \frac{\Gamma\left((\nu + p\bar{\mathbf{n}})/2\right)}{\Gamma(\nu/2)\Gamma_p(\mathbf{n}/2)} |\mathbf{\Omega}|_{-\frac{\mathbf{n}}{2}} |\mathbf{R}|_{\frac{\mathbf{n}-p-1}{2}} \left(1 + \operatorname{tr}\left(\mathbf{\Omega}^{-1}\mathbf{R}\right)\right)^{-\frac{\nu+p\bar{\mathbf{n}}}{2}},$$

$$p_{it\mathcal{R}}(\mathbf{R}|\mathbf{\Omega}, n, \boldsymbol{\nu}) = \frac{\Gamma\left((n + p\bar{\boldsymbol{\nu}})/2\right)}{\Gamma\left(n/2\right)\Gamma_p\left(\overleftarrow{\boldsymbol{\nu}}/2\right)} |\mathbf{\Omega}|_{\frac{\boldsymbol{\nu}}{2}} |\mathbf{R}|_{-\frac{\boldsymbol{\nu}+p+1}{2}} \left(1 + \operatorname{tr}\left(\mathbf{\Omega}\mathbf{R}^{-1}\right)\right)^{-\frac{n+p\bar{\boldsymbol{\nu}}}{2}},$$

$$p_{iF\mathcal{R}}(\mathbf{R}|\mathbf{\Omega}, \mathbf{n}, \boldsymbol{\nu}) = \frac{\Gamma_p((\boldsymbol{\nu} + \mathbf{n})/2)}{\Gamma_p(\overleftarrow{\boldsymbol{\nu}}/2)\Gamma_p\left(\mathbf{n}/2\right)} |\mathbf{\Omega}|_{\frac{\boldsymbol{\nu}}{2}} |\mathbf{R}|_{-\frac{\boldsymbol{\nu}+p+1}{2}} \left|\left(\mathbf{I} + \mathbf{C}_\Omega^{\top}\mathbf{R}^{-1}\mathbf{C}_\Omega\right)^{-1}\right|_{\frac{\boldsymbol{\nu}+\mathbf{n}}{2}}.$$

Proof in Appendix.

The p.d.f.s of the Wishart-type distributions follow by simply setting all entries in the d.o.f. vectors equal to each other. The p.d.f.s of the standardized probability distributions, which we indicate by writing $p_{\mathcal{D}}(\boldsymbol{\Sigma}, \boldsymbol{\theta}_{\mathcal{D}})$ instead of $p_{\mathcal{D}}(\boldsymbol{\Omega}, \boldsymbol{\theta}_{\mathcal{D}})$, are given in the appendix in Table 2.11. They are easily derived by replacing $\boldsymbol{\Omega}$ with $\mathbf{C}\mathbf{M}_{\mathcal{D}}^{-1}\mathbf{C}^{\top}$.[11]

2.3 The t-Riesz Distribution Family Based on Intraday Return Vectors

This section shows how the newly derived t-Riesz distribution family can be alternatively derived from a reasonable low-level assumption on the intraday return vectors. In particular, we show that it arises naturally as the distribution of the standard realized covariance matrix,

$$\mathbf{R} = \sum_{j=1}^{m} \mathbf{r}_j \mathbf{r}_j^{\top} \tag{2.19}$$

under a set of assumptions on the underlying intraday return vectors $\mathbf{r}_j$. Let us denote the $mp \times 1$ vector of all intraday returns as $\tilde{\mathbf{r}} = (\mathbf{r}_1^{\top}, \mathbf{r}_2^{\top}, \ldots, \mathbf{r}_m^{\top})^{\top}$. Under the assumption that it follows a multivariate elliptically contoured distribution with zero mean vector and block diagonal dispersion matrix $\tilde{\boldsymbol{\Omega}} = (\mathbf{I} \otimes \boldsymbol{\Omega})$, where $\boldsymbol{\Omega}$ is the same $p \times p$ symmetric positive definite parameter matrix as above, the distribution of the realized covariance matrix $\mathbf{R}$ is given by Theorem 2.3.1.

Theorem 2.3.1. *Let $\tilde{\mathbf{r}}$ follow a multivariate elliptically contoured distribution (as defined in Gupta, Varga, and Bodnar, 2013, Definition 2.1) with zero mean vector, dispersion matrix $\tilde{\boldsymbol{\Omega}} = (\mathbf{I} \otimes \boldsymbol{\Omega})$ and p.d.f.*

$$f(\tilde{\mathbf{r}}) = |\boldsymbol{\Omega}|^{-m/2} h(\tilde{\mathbf{r}}^{\top} \tilde{\boldsymbol{\Omega}} \tilde{\mathbf{r}}). \tag{2.20}$$

Then the p.d.f. of $\mathbf{R} = \sum_{j=1}^{m} \mathbf{r}_j \mathbf{r}_j'$ obtains as

$$\frac{\pi^{\frac{pm}{2}}}{\Gamma_p\left(\frac{m}{2}\right)} |\mathbf{R}|^{\frac{m-p-1}{2}} f(tr(\mathbf{R}\boldsymbol{\Omega}^{-1})). \tag{2.21}$$

Proof in Appendix.

For the special case, where $\tilde{\mathbf{r}}$ is assumed to follow a multivariate normal distribution, it is well known that the Wishart distribution $\mathbf{R} \sim \mathcal{W}(\boldsymbol{\Omega}, m)$ is obtained.

11. See Section 2.7.2 in the appendix for more details.

The Wishart is still defined if the d.o.f. parameter is a real number n, rather than natural number m. However, it is common knowledge that the normality assumption is strongly rejected for financial return data. Furthermore, the multivariate normality assumption with block-diagonal covariance matrix $\tilde{\mathbf{\Omega}}$ implicitly assumes that the individual intraday return vectors $\mathbf{r}_j$ are independent of each other, which is a very strong assumption. A more realistic multivariate elliptically contoured distribution is the multivariate t-distribution, as it accommodates the fat tails observed in financial return data. Furthermore, although due to the block-diagonal structure of $\tilde{\mathbf{\Omega}}$ there is no correlation between the $\mathbf{r}_j$, the multivariate t distribution does imply dependence between them. For this special case, i.e. $\tilde{\mathbf{r}} \sim mvt(\mathbf{0}, \tilde{\mathbf{\Omega}}, \nu)$, we obtain the t-Wishart distribution, $\mathbf{R} \sim t\mathcal{W}(\mathbf{\Omega}, m, \nu)$.[12] The more reasonable assumption of fat tails and inter-j dependence is mirrored in the superior performance of the t-Wishart distribution over the Wishart as documented in Section 2.5.[13]

Next, we are going to derive the t-Riesz distribution in the same way as the t-Wishart, but with the notion of asset liquidity included, as proposed by Gribisch and Hartkopf (2022).[14] We measure liquidity for a given asset by the number of intraday intervals where at least one trade occurred, and thus a new price observation was recorded. Only those intraday returns in $\tilde{\mathbf{r}}$ for which there was a new price observation are assumed to jointly follow a multivariate t-distribution, while the others are replaced by zeros.[15] Since there was no new price observation, previous-tick interpolation would make these returns zero anyways. Excluding returns that are zero due to missing price observations is appropriate because the well-documented excessive frequency of zero returns (c.f. Sucarrat and Grønneberg, 2020) undermines the assumption of a continuous distribution, such as the t-distribution, that is commonly used to model returns. Theorem 2.3.2 now derives the t-Riesz distribution.

12. Another version of the t-Wishart was first introduced by Sutradhar and Ali (1989).

13. We also tried the the multivariate hyperbolic distribution (leading to the hyperbolic-Wishart), which is a generalization of the multivariate t distribution and the multivariate Laplace distribution (leading to the Laplace-Wishart). In both a low-dimensional and a high-dimensional estimation there were no substantial likelihood gains over the t-Wishart. The hyperbolic-Wishart numerical maximum likelihood parameter estimates converged to almost exactly its special case, the t-Wishart distribution.

14. Recall that it is a generalization of the t-Wishart distribution with d.o.f. parameter vector $\mathbf{n}$ instead of scalar n, where the two distributions are equal if $\mathbf{n} = (n, \ldots, n)^\top$.

15. This treatment of missing new price observations as deterministic zeros is in line with Gribisch and Hartkopf (2022) and Hassairi, Ktari, and Zine (2022).

Theorem 2.3.2. *Denote by $\dot{\mathbf{r}}$ and $\dot{\mathbf{\Omega}}$ the sub-vector of $\tilde{\mathbf{r}}$ and sub-matrix of $\tilde{\mathbf{\Omega}}$ that are obtained by striking out the rows (and columns) with missing updated prices observations. Assume that for at least one asset, there is a new price observation on all intraday intervals m, and that for the j'th most liquid asset, there is a new price observation on a subset of the intervals of where the $(j+1)$'th most liquid asset has an observation. Furthermore, assume that the assets are sorted from least to most liquid in all intraday return vectors $\mathbf{r}_j$.*

If $\dot{\mathbf{r}}$ follows a multivariate t-distribution $\dot{\mathbf{r}} \sim mvt(\mathbf{0}, \dot{\mathbf{\Omega}}, \nu)$, then the realized covariance matrix $\mathbf{R} = \sum_{j=1}^{m} \mathbf{r}_j \mathbf{r}_j^\top$ follows a t-Riesz distribution with scale matrix $\mathbf{\Omega}$, d.o.f. parameter vector $\mathbf{m} = (m_1, m_2, \ldots, m_p)$, where m_i denotes the number of return observations of asset i and d.o.f. parameter ν, denoted by

$$\mathbf{R} \sim t\mathcal{R}(\mathbf{\Omega}, \mathbf{m}, \nu).$$

Proof in Appendix.

The t-Riesz distribution is still defined if the d.o.f. parameter vector $\mathbf{m}$ is a vector of reals $\mathbf{n} = (n_1, \ldots, n_p)$, rather than of natural numbers m_i. Note that for the t-Wishart distribution, which is obtained under the assumption of no missing new price observations for any of the assets, the liquidity sorted ordering of the assets is irrelevant since it is invariant to the ordering of the assets. Theorem 2.3.2 is a generalization (and slight reformulation) of a finding in Gribisch and Hartkopf (2022) (see also Hassairi, Ktari, and Zine, 2022, and Veleva, 2009). They show that the standard Riesz distribution can be generated by assuming a normal distribution on the intraday returns with heterogeneous liquidity. As for the t-Wishart compared to the Wishart, the superior performance of the t-Riesz compared to the Riesz distribution, shown in the empirical section below, mirrors the more appropriate assumption of a multivariate t compared to the normal distribution on the intraday return vectors.

2.4 Time-Varying Mean

Now, let us add subscripts for the days in our sample, $t = 1, \ldots, T$, to represent the time-series of RCs as $\{\mathbf{R}_t\}_{t=1}^{T}$. In the literature, it is standard to assume time-variation in the mean, i.e., in the $\mathbf{\Sigma}$ matrix of the underlying distribution, while leaving the d.o.f. parameters fixed over time (see e.g. Golosnoy, Gribisch, and Liesenfeld, 2012, Opschoor et al., 2018, Blasques et al., 2021). That is

$$\mathbf{R}_t | \mathcal{F}_{t-1} \sim \mathcal{D}(\mathbf{\Sigma}_t, \boldsymbol{\theta}_{\mathcal{D}}), \tag{2.22}$$

where $\mathcal{F}_{t-1} = \{\mathbf{R}_{t-1}, \mathbf{R}_{t-2}, \dots\}$. A standard updating mechanism for $\mathbf{\Sigma}_t$ is a scalar-BEKK[16] recursion given by

$$\mathbf{\Sigma}_t = (1 - a - b)\mathbf{\Xi} + a\mathbf{R}_{t-1} + b\mathbf{\Sigma}_{t-1}, \tag{2.23}$$

where the intercept parameter matrix $\mathbf{\Xi}$ is symmetric positive definite of dimension $p \times p$ and a and b are scalar parameters, sometimes called ARCH and GARCH parameter, respectively. A necessary condition for stationarity is $a, b > 0 \wedge (a+b) < 1$ under which we have that the unconditional mean

$$\mathbb{E}[\mathbf{R}_t] = \mathbf{\Xi}. \tag{2.24}$$

In this paper, we assume stationarity. Note that if we restrict $a = b = 0$ in equation (2.23), we end up with a static distribution that is, $\mathbf{\Sigma} = \mathbf{\Xi}$, and

$$\mathbf{R}_t \overset{iid}{\sim} \mathcal{D}(\mathbf{\Sigma}, \boldsymbol{\theta}_{\mathcal{D}}).$$

Note that the parameters $\mathbf{\Xi}$, a, and b are specific to the chosen distribution. However, as mentioned above, we omit the subscript $\mathcal{D}$ for readability.

Empirical Analysis 2.5

Estimation 2.5.1

The total number of parameters is dominated by the order $\mathcal{O}(p^2)$ matrix $\mathbf{\Xi}$, which has $p(p+1)/2$ unique elements. This makes one-step numerical maximum likelihood estimation for, say, $p > 5$ very computationally expensive and for, say, $p > 10$ infeasible. To alleviate this so-called *curse of dimensionality*, we estimate $\mathbf{\Xi}$ (or $\mathbf{\Sigma}$ in the static model) with its obvious (see equation (2.24)) method-of-moments estimator

$$\widehat{\mathbf{\Xi}} = \frac{1}{T}\sum_{t=1}^{T} \mathbf{R}_t,$$

and estimate the remaining parameters ($\boldsymbol{\theta}_{\mathcal{D}}$, a, and b) conditional on $\widehat{\mathbf{\Xi}}$ via standard numerical maximum likelihood estimation. This multi-step estimation procedure, sometimes called *targeting*, reduces the size of the numerical optimization problem to the order of at most $\mathcal{O}(p)$, and for the Wishart-type distributions even

16. Named after Yoshi Baba, Robert Engle, Dennis Kraft, and Ken Kroner who wrote an earlier version of the paper Engle and Kroner (1995) in which the BEKK recursion is proposed.

to at most four[17] parameters. The targeting two-step estimation procedure is common in the literature (see e.g. Noureldin, Shephard, and Sheppard, 2012, Opschoor et al., 2018). Its consistency is intuitive and has been shown in the traditional multivariate GARCH framework by Francq, Horváth, and Zakoïan (2014). We expect consistency to carry over to the realized multivariate GARCH framework. In the empirical section of this paper, we always use this two-step estimation procedure, regardless of the cross-sectional dimension p.

A final complication is that, as mentioned above, the ordering of the assets matters for the Riesz-type distributions. We follow the algorithm proposed in Blasques et al. (2021) to optimize over the asset order (see Section 2.2.4).[18]

2.5.2 Data

Our original data are one-minute close prices from all trading days from 01 January 1998 to 05 February 2021 for every stock that was a constituent of the S&P 500 index during the sample period. A *close price* is defined as the latest observed trade price of the respective one-minute interval. So our data can be described as previous tick interpolation on a fixed one-minute grid. We acquired the data from Quantquote[19], who combine, clean and process data directly obtained from various exchanges, where the biggest are NYSE, NASDAQ and AMEX[20].

The aim is to produce the longest possible time-series of accurately estimated daily integrated covariance estimators. We exclude dates before 01 January 2002, because the NYSE fully implemented decimal pricing in 2001[21] and there are numerous other trading irregularities during 2001[22]. This leaves 4808 trading days. We then exclude stocks that have not been traded on one of the remaining days in the sample, which leaves 465 of 983 stocks. We only keep observations from official trading hours to be consistent across trading days. We then choose the 100 stocks with the most one-minute close price observations. Of those, the one with the least observations has, on average, 385.18 one-minute close price observations per trading day. Since the typical trading day has 390 minutes, on average less than five close prices are missing per day. Excluding illiquid stocks is common practice in creating time-series of RCs (see e.g. Lunde, Shephard, and

17. These are n, and/or ν, a and b.

18. The seed for the random generation of permutations to try initially is kept the same for all Riesz-type distributions.

19. The Caltech Quantitative Finance Group recommends the company, see http://quant.caltech.edu/historical-stock-data.html.

20. AMEX was bought by NYSE in 2008, and handled only 10% of trades at its height

21. On 29 January 2001 to be precise.

22. For example the days surrounding the terrorist attacks on 11 September 2001 and "computer systems connectivity problems" on 08 June 2001.

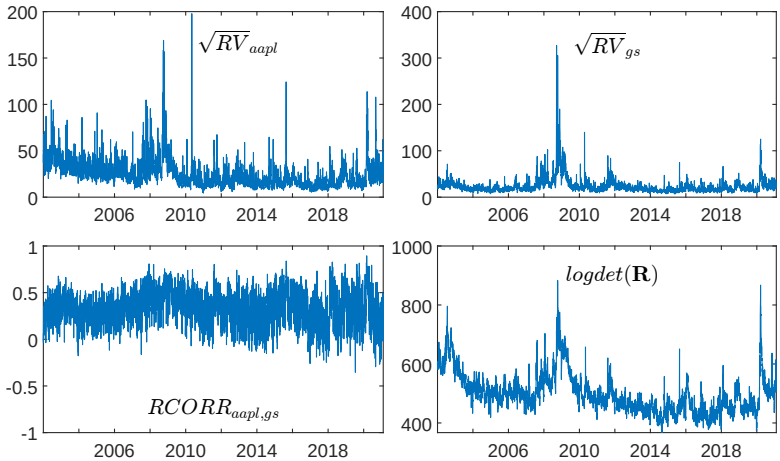

Figure 2.7: Top row: Annualized realized volatilities of Apple (aapl) and Goldman Sachs (gs), i.e. the square root of two of the elements on the main diagonal of the 100 asset $\mathbf{R}_t$ for the complete sample (01 January 2002 to 05 February 2021). Bottom row: Realized correlation between Apple and Goldman Sachs $RCORR_{aapl,gs} = \mathbf{R}_{aapl,gs}/(\sqrt{RV_{aapl}}\sqrt{RV_{gs}})$ and the natural logarithm of the determinant of the 100 assets $\mathbf{R}_t$ over time.

Sheppard, 2016). While this procedure biases the sample towards stocks that were very liquid over the entire sample period[23], it does ensure that the integrated covariance estimates are accurate for those stocks included.

We follow Opschoor et al. (2018) and Blasques et al. (2021) and use five-minute returns to construct the 100-dimensional RCs. In particular, we average for each trading day the five distinct RCs, obtained from constructing five-minute log-return vectors for each of the five distinct five-minute grids over the trading day.[24] This estimator is known as the subsampling realized covariance matrix and has been introduced by Zhang, Mykland, and Aït-Sahalia (2005) (see also Sheppard, 2012). It has the advantage of being more efficient than the simple realized covariance matrix since it uses all our data, not just the data of one of the grids. Furthermore, it produces positive definite matrices, even for high cross-sectional dimensions and low sampling frequencies.[25]

23. Relatively young firms (e.g. Facebook or Tesla) are excluded.

24. The five time-grids start (on a typical trading day) at 09:00, 09:01, 09:02, 09:03 and 09:04, respectively.

25. On a typical trading day we have 390/5=70 intraday return vectors on a five-minute grid. This allows for a maximum 70 assets to generate positive definite RCs. With subsampling, however, RCs are based on 385 five-minute return vectors, so for up to 385 assets the resulting RCs are positive definite.

For a view of the data, see Figure 2.7, which shows the annualized realized volatility for Apple (appl) and Goldman Sachs (gs), as well as their realized correlation and the log-determinant of the 100-asset RCs. We see that the spikes in volatility are of similar magnitude for Apple and Goldman Sachs in the recent COVID-19-induced market turmoil, while the global financial crises of 2008/2009 caused volatility to spike much higher for Goldman Sachs than for Apple. The dot-com crisis (early 2000s), on the other hand, causes more volatility for Apple. We see that correlations are mainly positive and more stable around crisis periods. Finally, we see that the log-determinant of $\mathbf{R}_t$, as a measure of the size of the RCs, does indeed spike in the aforementioned market turmoil periods (dot-com, COVID-19, global financial crisis).

From the 100-dimensional dataset described, we randomly choose a 5, 10, and 25-dimensional principal submatrix, as well as the three principal submatrices corresponding to companies with SIC codes of the (1) Financial, Insurance and Real Estate, (2) Mining and (3) High-End Manufacturing division, for a total of six datasets. The division-specific datasets are chosen to investigate whether the tail-homogeneous distributions better fit more homogeneous data.

2.5.3 In-Sample

Static Distributions

As a first empirical exercise, we fit the different static distributions (i.e. $\mathbf{\Sigma}_t = \mathbf{\Xi}$) to the data. We use the two-step estimation method described above for all distributions and datasets. While it is clear, that a time-varying mean drastically improves the fit of all distributions, it is still imperative to investigate the static distributions first to understand how well their distributional properties match the data. In fact, after adding a time-varying mean we would not be able to dissect anymore, if the differences in fit come from the time-varying mean path, or the distributional properties. For maximum likelihood estimation there would be a different estimated time-varying mean path for each distribution. Even if we imposed the same time-varying mean path for all distributions we would not be sure if the one chosen favors one distribution over another. Furthermore, if one wants to model a time-series of RCs there are several choices to be made. Which distribution? Which time-series structure? Which distribution parameters should be time-varying? We cannot find a good reason to start with one of these decisions before another. In this paper, however, the focus lies on the comparison between distributions; so it is natural to start our empirical investigation with a comparison of the static distributions.

Table 2.12 in the appendix shows the estimated d.o.f. parameters. We focus now on Table 2.6, which shows the estimated log-likelihood values. The distribution

Assets:	Random	Mining	Random	Finance	Random	Manuf.
#Assets:	5	6	10	15	25	25
Wishart	-67794	-84692	-168435	-527223	-537407	-381884
Riesz	-55534	-70180	-108027	-345020	-180751	-124508
Inv.Wishart	-50721	-53737	-111648	-255628	-162663	-78492
Inv.Riesz	-45490	-46495	-77633	-180138	36136	45223
t-Wishart	-24687	-30461	-38384	-25481	86847	233668
t-Riesz	-17543	-23263	-1941	49699	295409	355966
Inv.t-Wishart	-24761	-27528	-40690	-11924	191946	387973
Inv.t-Riesz	-20829	-19584	-8057	57397	376512	445333
F	-48758	-51941	-97299	-221813	-55002	14633
F-Riesz	-26402	-29242	-13809	-30051	307936	333553
Inv.F-Riesz	-29622	-29930	-22537	-70305	276602	308721

Table 2.6: Log-likelihood values for the estimated static distributions on various datasets. The background shades are to be read column-wise, with the lowest log-likelihood value shaded black and the highest shaded white, with linear gray-scaling in between. Largest values in red.

rankings according to log-likelihood values are robust across the different datasets. We see a clustered pattern, where the (Inverse) Wishart, (Inverse) Riesz and F distribution could be considered as a group of similar fit and the (Inverse) t-Wishart, (Inverse) t-Riesz and (Inverse) F-Riesz as another group. The t-Riesz and Inverse t-Riesz distribution stand out from this second group since, for all datasets, one of the two obtains the best log-likelihood value. The industry-specific datasets (Mining, Finance, and High-End Manufacturing) show a clear pattern favoring the t-Riesz distribution family. The F-Riesz and Inverse F-Riesz distributions also fit quite well overall, especially in the randomly drawn datasets. The Riesz and its special case, the Wishart, are the worst-fitting static distributions.

Next, we investigate how the different distributions for RCs match specific marginal characteristics of the data. First, in Figure 2.8, we plot the histograms of the log-determinant and the first diagonal element (RV of Amgen) of the ten-dimensional dataset, as well as the corresponding marginal p.d.f.s for the fitted probability distributions for RCs. The choices of dataset and company are arbitrarily made, but the plots for other datasets and companies look similar. The first obvious observation is that the t-Wishart and t-Riesz distributions are clearly mirror the empirical distribution of both the log-determinant and the RVs best. Secondly, only the t-Riesz distribution family (last row of the figure) exhibits a reasonable marginal distribution of the log-determinant; all other distributions have too much probability mass around their center and too little in the tails.

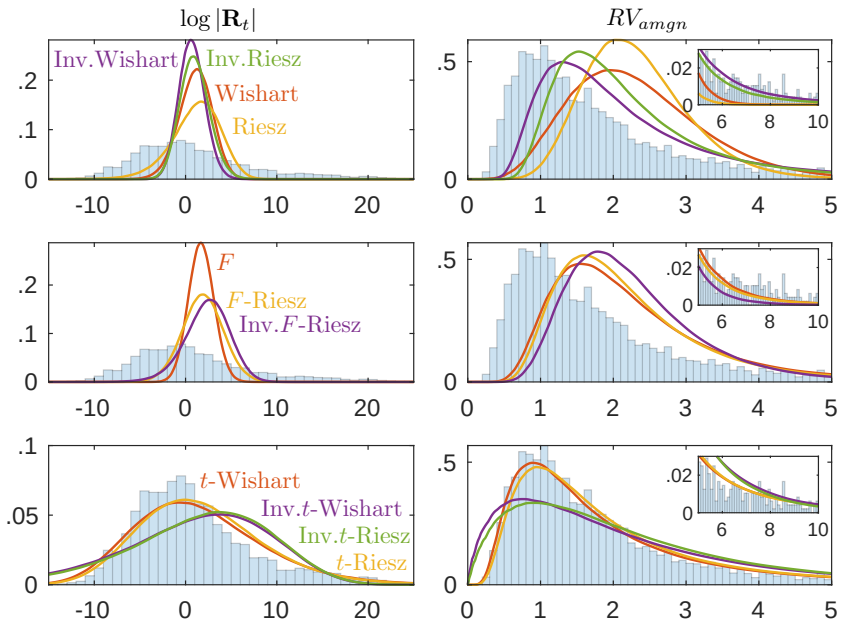

Figure 2.8: Histograms, normed to reflect p.d.f.s, of $\log|\mathbf{R}_t|$, and of the realized variance of Amgen, $RV_{amgn} = (\mathbf{R}_t)_{1\,1}$ of the ten-dimensional dataset, as well as the respective marginal p.d.f.s implied by the fitted static matrix distributions. The p.d.f.s are kernel density estimates on 1.000.000 simulated realizations.

The reason why the F and F-Riesz distributions (second row, first column of the figure) do not match the (tail)-distribution of the log-determinant of the t-Wishart and t-Riesz is that their ν and $\boldsymbol{\nu}$ d.o.f. parameter estimates are much higher in comparison (see Table 2.12 in the appendix). The d.o.f. parameters influence on the distribution of the log-determinant can be seen in the equations (2.15) and (2.17). In fact, the estimated ν parameter is fairly constant and close to 5 across all cross-sectional dimensions, whereas it increases for the F distribution from 11 for the five-dimensional dataset to 46 for the 25-dimensional dataset. A similar pattern is visible for the corresponding Riesz-type distributions.

For the RVs the t-Riesz distribution family also clearly fits best. All other distributions assume too little probability mass on RVs below one and too much on values between one and four. In particular, the match of the t-Riesz and t-Wishart distribution marginal p.d.f.s for the empirical RVs are extremely good, also in the tails. Regarding the tail behavior, we can see that in our definition of fat-tailedness, only the t-Riesz distribution family can be considered fat-tailed, as

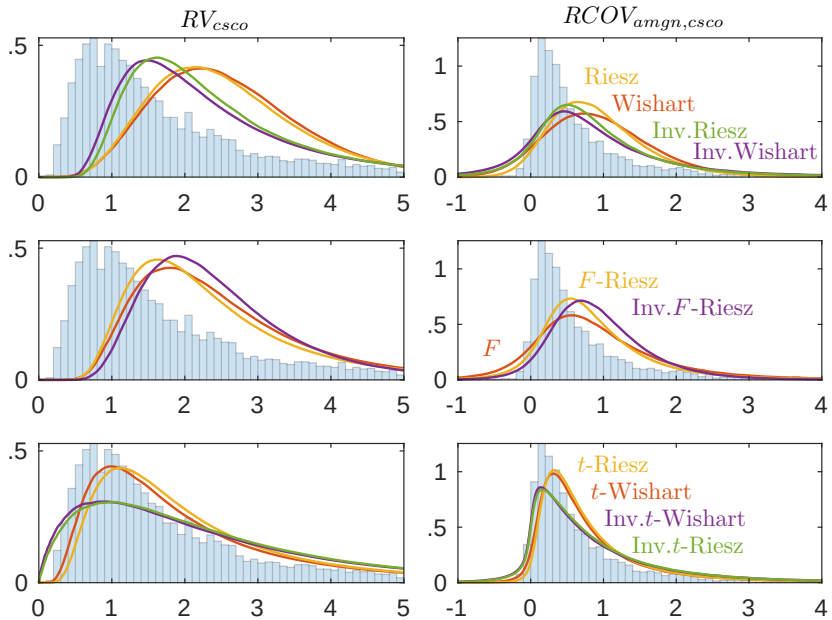

Figure 2.9: Histograms, normed to reflect p.d.f.s, of the realized variance of Cisco, $RV_{csco} = (\mathbf{R}_t)_{2\,2}$, and of the realized covariance between Cisco and Amgen, $RV_{amgen} = (\mathbf{R}_t)_{2\,1}$, in the ten-dimensional dataset, as well as the respective marginal p.d.f.s implied by the fitted static matrix distributions. The p.d.f.s are kernel density estimates on 1.000.000 simulated realizations.

the other distributions do not have sufficient probability mass on large (> 10) log-determinants. In fact, especially the t-Riesz and t-Wishart match the empirical distribution in the tails very well. The F and (Inverse) F-Riesz distributions do imply more probability mass on larger log-determinants than the (Inverse) Wishart and (Inverse) Riesz distribution, and thus are comparatively more fat-tailed, but do not come close to the fat-tailedness of the data or the t-Riesz distribution family. Considering the fat-tailedness of the RVs, the small subplots on the right column in Figure 2.8 show that all but the Wishart and Riesz distribution reasonably well match their empirical tail behavior. Again we note that the t-Riesz distribution family implies the most probability mass on large RVs with the t-Riesz and t-Wishart matching the empirical distribution in the tails best.

Figure 2.9, column one, depicts the RVs of Cisco and confirms our previous observations on RVs, while in column two, we plot the histogram of the empirical realized covariances (RCOVs) between Cisco and Amgen ($(\mathbf{R}_t)_{2\,1}$). The t-Riesz

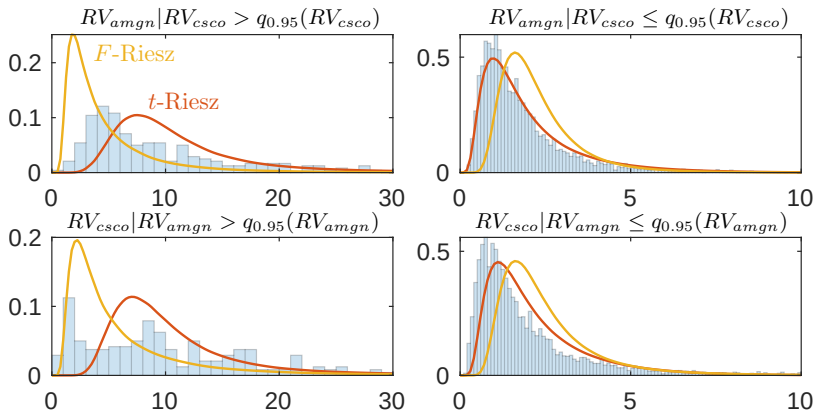

Figure 2.10: Histograms of real data, normed to reflect p.d.f.s, and the respective p.d.f.s of the realized variance of companies Amgen (first row, $RV_{amgn} = (\mathbf{R}_t)_{1\,1}$) and Cisco, (second row, $RV_{csco} = (\mathbf{R}_t)_{2\,2}$), conditional on the respective RV of the other company having a tail-realization (left column) or not (right column). The data comes from the random ten-dimensional dataset described in 2.5.2. The respective conditional p.d.f.s come from the fitted t-Riesz and F-Riesz distribution and are kernel density estimates on 1.000.000 simulated realizations.

family also clearly fits best for the covariances, while the other distributions allocate too little probability mass RCOVs between 0 and 0.5 and too much probability mass on negative RCOVs and on RCOVs between 0.5 and 2.

Next, we turn to tail-heterogeneity versus tail-homogeneity. For this, we plot in Figure 2.10 histograms of the RV of Amgen (Cisco) given that the RV of Cisco (Amgen) has a tail-realization and given that it does not. We plot on top of each histogram the conditional p.d.f.s implied by the t-Riesz and F-Riesz distribution. A tail-realization is defined as an RV of a given asset that exceeds its empirical 95% quantile. As a first observation, comparing the histograms in the left column to those in the right, we see that the empirical distribution of an RV has a much higher probability of large realizations if another RV has a tail-realization (i.e. tail-homogeneity). The next striking observation is that the t-Riesz distribution can mirror this property of the data much better than the F-Riesz. While the F-Riesz has more probability mass on large RVs if another RV has a tail-realization, compared to if the other RV does not, this right-shift in probability mass is not as large as for the t-Riesz. Thus, as we explained intuitively by looking at the respective stochastic representation of the t-Riesz and F-Riesz in Section 2.3, it does make sense to call the t-Riesz tail-homogeneous and the F-Riesz tail-

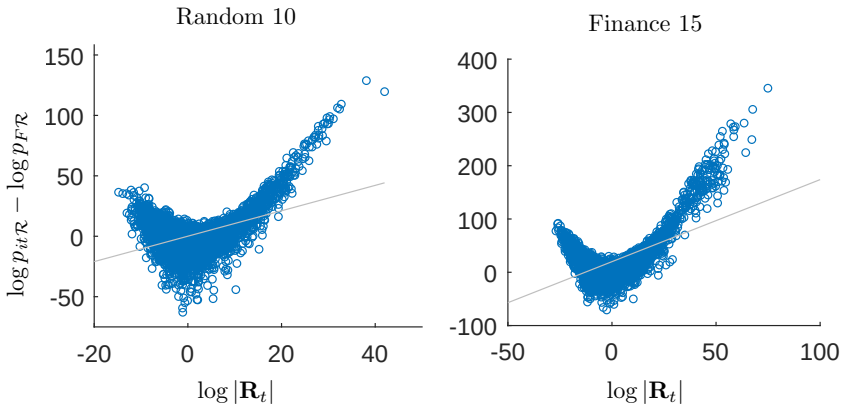

Figure 2.11: Difference in estimated log-likelihood contributions between Inverse t-Riesz and F-Riesz distributions, $\log p_{it\mathcal{R}}(\mathbf{R}_t|\widehat{\boldsymbol{\Sigma}}, \hat{\theta}) - \log p_{F\mathcal{R}}(\mathbf{R}_t|\widehat{\boldsymbol{\Sigma}}, \hat{\theta})$. Least-squares line is in gray.

heterogeneous and our data favours tail-homogeneity.

Finally, we examine the differences in log-likelihood values between the tail-homogeneous Inverse t-Riesz and the tail-heterogeneous F-Riesz distribution more closely.[26] Figure 2.11 shows the difference in log-likelihood contributions between the two distributions depending on the log-determinant of the RCs for the random ten-dimensional and the Finance 15-dimensional dataset. We see that the Inverse t-Riesz distribution gains its advantage in static fit mainly from the RCs with larger log-determinants. This is in line with our expectation that tail-heterogeneity is disadvantageous for crisis periods. The Inverse t-Riesz also fits better for very small RCs, which can be rationalized by the fact that in times of a very calm market, financial assets behave very homogeneously as well.

Time-Varying Mean

Now we fit the dynamic time-varying mean models for the different distributions to the datasets. The estimated d.o.f. parameters can be found in Table 2.13 in the appendix. Table 2.7 contains the estimated ARCH parameters ($\hat{a}$). They are all highly significant. The estimated persistence parameters ($\hat{a} + \hat{b}$) are very close across distributions and datasets and range from 0.976 (Wishart, dataset

26. We take the *Inverse* t-Riesz distribution in this comparison because, in the dynamic setting, it will turn out to be the distribution of the t-Riesz distribution family with the best log-likelihood values and will be in close competition with the F-Riesz.

	$\hat{a}$					
Assets:	Random	Mining	Random	Finance	Random	Manuf.
#Assets:	5	6	10	15	25	25
Wishart	0.364	0.306	0.284	0.299	0.189	0.188
Riesz	0.339	0.286	0.259	0.275	0.161	0.160
Inv.Wishart	0.238	0.211	0.183	0.177	0.115	0.098
Inv.Riesz	0.242	0.206	0.181	0.168	0.108	0.094
t-Wishart	0.196	0.150	0.127	0.090	0.080	0.066
t-Riesz	0.186	0.132	0.117	0.072	0.070	0.052
Inv.t-Wishart	0.153	0.127	0.101	0.074	0.065	0.053
Inv.t-Riesz	0.154	0.122	0.097	0.067	0.059	0.050
F	0.257	0.231	0.198	0.192	0.126	0.110
F-Riesz	0.200	0.166	0.145	0.136	0.091	0.078
Inv.F-Riesz	0.215	0.179	0.156	0.147	0.095	0.084

Table 2.7: Estimated ARCH $\hat{a}$ parameters of the models in equations (2.22) and (2.23) for the different datasets and all distributions for RCs. All estimated $\hat{a}$ are highly significant, with the median (smallest) t-stat equalling 552 (123). The estimated persistence $(\hat{a} + \hat{b})$ is very similar across datasets and distributions and ranges from 0.976 (Wishart, dataset "Random 5") to 0.999 (Inverse t-Wishart, dataset "Finance 15"). All estimated $\hat{b}$ are highly significant, with the median (smallest) t-stat equalling 3060 (399).

"Random 5") to 0.999 (Inverse t-Wishart, dataset "Finance 15") with all GARCH parameters $(\hat{b})$ being highly significant. We see several clear patterns.

First, the estimated ARCH parameters become smaller with increasing cross-sectional dimension p for all distributions. This pattern has been documented by Pakel et al. (2021) to be estimation bias. It is larger, the larger the dimension p and is caused by the method-of-moments estimator $\widehat{\mathbf{\Xi}}$. They show that composite likelihood estimation can mitigate the bias. Unfortunately, composite likelihood estimation is not straightforward to apply on Riesz-type distributions due to their d.o.f. parameter vector(s). Another solution might be to use a shrinkage estimator as in Engle, Ledoit, and Wolf (2019). In this paper, however, we focus on differences between assumed probability distributions for RCs. We do not expect the relative ranking results to change if we use one of the above-mentioned methods to estimate the intercept matrix $\mathbf{\Xi}$.

The second pattern we observe is, that the estimated a are smallest for the t-Riesz distribution family across all dimensions, followed by the F-Riesz distributions and largest for the Riesz distributions and the F distribution. That is, the t-Riesz distribution family reacts least to the previous realizations $\mathbf{R}_{t-1}$ to update

Assets:	Random	Mining	Random	Finance	Random	Manuf.
#Assets:	5	6	10	15	25	25
Wishart	-15599	-11314	-1397	55874	333549	330537
Riesz	-11866	-6289	13107	88881	423237	423510
Inv.Wishart	-7668	1693	32727	132311	541567	547719
Inv.Riesz	-5813	4154	39879	148338	574711	583946
t-Wishart	-5298	5910	33588	149496	469265	467948
t-Riesz	-2650	8572	44108	178339	539295	541439
Inv.t-Wishart	-1422	11277	55196	195532	639630	655575
Inv.t-Riesz	440	13286	62134	213268	666312	684209
F	-6908	2468	35627	142413	571495	572881
F-Riesz	721	12621	62436	190194	669628	674391
Inv.F-Riesz	-122	11619	59471	185342	662615	665143

Table 2.8: Log-likelihood values for the estimated dynamic distributions and different datasets. The background shades are to be read column-wise, with the lowest log-likelihood value shaded black and the highest shaded white, with linear gray-scaling in between. Largest values in red.

the mean $\mathbf{\Sigma}_t$, which we interpret as an indication of the excellent (unconditional) fit of these distributions. In contrast, the Wishart distribution reacts most to the previous $\mathbf{R}_{t-1}$, indicating a worse fit of the distributional assumption. We note here that in terms of fit and forecasting performance, a large mean-shifting reaction to previous RCs (as for the Wishart model) is actually beneficial in crisis periods, where RCs suddenly spike in size and stay large for a short time. The good overall distributional fit of other distributions causes them to react more slowly to those volatility bursts.

Table 2.8 contains the log-likelihood values for the estimated distributions with time-varying mean. As for the static distributions, the ranking across distributions is relatively stable over the cross-sectional dimension p. However, now in the dynamic setting, the ranking across distributions is less clustered. The Inverse t-Riesz and the F-Riesz distributions emerge as the clear winners exhibiting the largest likelihood values. They are very close, with the former winning the three industry-specific datasets and the latter winning the three random datasets. This is again in line with our economic intuition that tail-homogeneity is advantageous for RCs of homogeneous assets. One hypothesis why the Inverse t-Riesz now clearly beats the t-Riesz compared to the static case, can be constructed by recalling Figures 2.2 and 2.4. There it is visible, that the probability mass of the Inverse t-Riesz is more concentrated than for the t-Riesz with similar parameters. So, if our dynamic mean-shifting model approximates the conditional mean reasonably well,

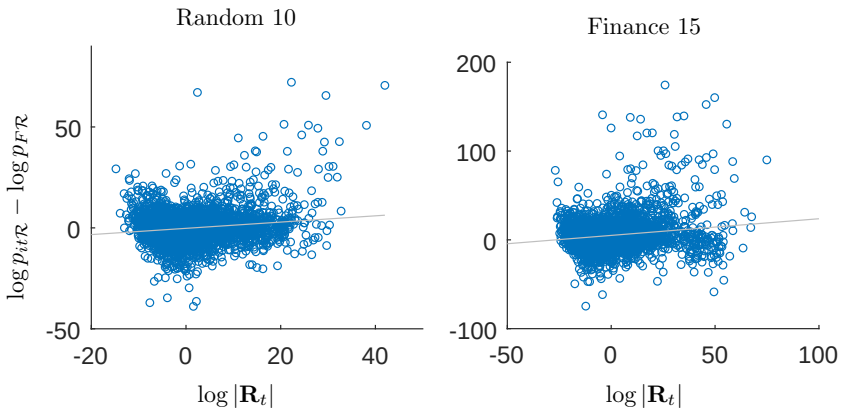

Figure 2.12: Difference in estimated log-likelihood contributions between Inverse t-Riesz and F-Riesz distributions with time-varying expected value matrix, $\log p_{it\mathcal{R}}(\mathbf{R}_t|\widehat{\mathbf{\Sigma}}_t, \hat{\theta}) - \log p_{F\mathcal{R}}(\mathbf{R}_t|\widehat{\mathbf{\Sigma}}_t, \hat{\theta})$. Least-squares line in gray.

the higher dispersion of probability mass of the t-Riesz might be disadvantageous.

The Riesz and its special case, the Wishart distribution, are unambiguously the worst-fitting distributions. In general, inverse distributions fit better than non-inverted ones. This is not surprising since fitting the inverse distributions to $\{\mathbf{R}_1, \mathbf{R}_2, \ldots, \mathbf{R}_T\}$ is equivalent to fitting the corresponding non-inverted ones to $\{\mathbf{R}_1^{-1}, \mathbf{R}_2^{-1}, \ldots, \mathbf{R}_T^{-1}\}$. The inverted RCs, also known as precision or concentration matrices, exhibit much thinner tails; hence the good fit of inverse distributions. Through this reasoning, we can call all inverse distributions fat-tailed distributions. Furthermore, by construction, every Wishart-type distribution has a lower estimated likelihood value than its Riesz-type counterpart. However, it is noteworthy that the difference in likelihood values is particularly large between the (Inverse) F-Riesz distribution and the F distribution.

Finally, Figure 2.12 compares the log-likelihood contributions of the Inverse t-Riesz and F-Riesz distributions similar to Figure 2.11 but for the time-varying mean specification. As in the static case, larger RCs are associated with higher log-likelihood contributions for the tail-homogeneous Inverse t-Riesz distributions, in line with our economic intuition that more volatile trading days exhibit more dependence among financial assets, favoring tail-homogeneity. As in the static case the smallest RCs are associated with higher log-likelihood contributions for the Inverse t-Riesz distribution, indicating that when markets are very calm, tail-homogeneity might be favored as well.

Assets:	Random	Mining	Random	Finance	Random	Manuf.
#Assets:	5	6	10	15	25	25
Wishart	319	616	541	1565	1870	900
Riesz	282	559	413	1061	1066	103
Inv.Wishart	224	425	174	-32	-101	-999
Inv.Riesz	214	404	126	-226	-423	-1271
t-Wishart	153	327	15	-685	-274	-972
t-Riesz	139	307	-44	-869	-662	-1432
Inv.t-Wishart	135	304	-94	-912	-1137	-1978
Inv.t-Riesz	128	292	-128	-1026	-1302	-2162
F	210	414	136	-230	-396	-1172
F-Riesz	147	319	-89	-715	-1173	-1959
Inv.F-Riesz	155	332	-59	-651	-1112	-1876

Table 2.9: Average of log-score loss over one-month forecasting period (22 trading days), $-\sum_{j=1}^{22} p_{\mathcal{D}}(\mathbf{R}_{t+j}|\widehat{\boldsymbol{\Sigma}}_{j+1}, \hat{\boldsymbol{\theta}}_{\mathcal{D},j+1})$, for the entire forecasting window; each model is re-estimated every ten trading days; 90% model confidence sets in red.

2.5.4 Out-of-Sample Forecasting Performance

We re-estimate the models every ten trading days on a rolling window of 1250 observations (roughly five years of data). The forecasting window starts at observation 1251 (18 December 2006) and ends on 05 February 2021.

For out-of-sample comparisons between different probability distributions, a natural loss function is the log-score, also known as the log posterior predictive likelihood, since it indicates how much probability mass the predictive distribution assigns to the observed outcome (compare e.g. Hautsch and Voigt, 2019 and Blasques et al., 2021). The log-score can also be justified as the consistent choice of loss function for maximum likelihood estimation in the following sense. It evaluates the out-of-sample data with the same loss function used to estimate the models in-sample. This is in line with Hansen and Dumitrescu (2022), who show that coherency between the estimation criterion and the actual objective is essential.

Since we are interested in overall distribution fit, it is important to not only look at the $t+1$ forecasting performance of the different distributions. To this end, Table 2.9 contains the log-score losses over a one-month forecasting period (22 trading days) for the entire forecasting window. The 90% *model confidence sets* (MCS, see Hansen, Lunde, and Nason, 2011) are shaded in gray.[27]

27. For calculation of the MCS, we choose 5000 stationary bootstrap replications with block length set equal to the maximum number of consecutive significant partial autocorrelations of the losses. We use the MFE toolbox by Kevin Sheppard for MCS calculation.

We see that the Inverse t-Riesz distribution emerges as the clear winner for log-score losses over a one-month forecasting period as it is the only member of the MCS for all datasets. Across datasets, the entire t-Riesz distribution family fits very well out-of-sample, and slightly better than the F-Riesz, except for the 25-dimensional datasets.

If we take a volatile period (2007 - 2011) and a calm period (2012 - 2019) forecasting window (see Tables 2.14 and 2.15 in the appendix), the Inverse t-Riesz distribution remains the sole member of the MCS except for the random 25-dimensional dataset in the calm period (even here it has the lowest loss value), where also the F-Riesz distribution is in the MCS.

For completeness, in Table 2.10 we also report the one-day-ahead log-score loss results for the entire sample. Here the Inverse t-Riesz distribution is the sole member of the MCS for the industry-specific Mining and Finance datasets, while for the other datasets, the F-Riesz distribution is also in the MCS. Still the Inverse t-Riesz distribution has lower losses for all datasets than the F-Riesz.

The above observations confirm our intuition that tail-homogeneity is a reasonable assumption. Clearly, the worst fitting distributions out-of-sample are the Riesz and its special case the Wishart distributions.

Assets:	Random	Mining	Random	Finance	Random	Manuf.
#Assets:	5	6	10	15	25	25
Wishart	7.22	15.97	3.51	-18.61	-0.79	-29.75
Riesz	6.53	15.08	0.60	-25.22	-19.40	-49.70
Inv.Wishart	5.35	12.94	-4.84	-35.97	-49.62	-83.52
Inv.Riesz	5.04	12.52	-6.14	-39.49	-56.69	-91.34
t-Wishart	4.85	12.23	-4.79	-40.61	-32.63	-61.30
t-Riesz	4.38	11.67	-6.95	-46.42	-46.68	-76.80
Inv.t-Wishart	4.01	11.02	-9.59	-50.55	-70.32	-104.08
Inv.t-Riesz	3.70	10.67	-10.86	-53.90	-75.18	-110.57
F	5.18	12.85	-5.44	-38.59	-56.51	-87.99
F-Riesz	3.75	10.85	-10.81	-48.56	-76.43	-109.33
Inv.F-Riesz	3.93	11.09	-10.16	-47.36	-74.87	-107.25

Table 2.10: Average of log-score loss, $-p_{\mathcal{D}}(\mathbf{R}_{t+1}|\widehat{\boldsymbol{\Sigma}}_{t+1}, \hat{\boldsymbol{\theta}}_{\mathcal{D},t+1})$, for the entire forecasting window, where each model is re-estimated every 10 trading days. 90% model confidence sets in red.

Conclusion 2.6

In conclusion, this paper provides a comprehensive comparison of probability distributions used to model realized covariance matrices (RCs) in financial applications. We reveal theoretical similarities and differences among the distributions, which are useful in explaining their disparity in empirical fit and forecasting performance. We formalize the concept of fat-tailedness for RCs and introduce the notion of tail-homogeneity as opposed to the tail-heterogeneity implied by the F-Riesz distribution. Furthermore, we derive the novel t-Riesz distribution family, which features tail-homogeneity. We show that the novel t-Riesz distribution family can be rooted in a realistic low-level assumption on the intraday return vectors from which the realized covariance matrices are constructed. The empirical part of the paper performs fit and forecasting comparisons of the different distributions in different datasets of time-series of RCs and explains how the theoretical differences translate into differences in fit and forecasting performance. It emerges that when assuming a static distribution for the RCs, the t-Riesz distribution fits best in terms of log-likelihood values and matching the fat-tails, the marginal distributions of the RVs and covariances, and the tail-homogeneity that financial data exhibits. In the dynamic setting where the mean of the distributions is assumed to be time-varying, the distribution rankings are less pronounced. Here, the Inverse t-Riesz and F-Riesz distribution fit best in-sample and out-of-sample with an advantage in favor of the former. We show that, especially in times of high market volatility and for assets of the same industry sector, tail-homogeneity is a more fitting assumption to the RC time-series data. Overall, the paper provides important insights for practitioners and researchers who want to model RCs of financial asset returns.

2.7 Appendix

2.7.1 Proofs

Proof of Theorem 2.2.1

Proof.

Proof of $\mathbb{E}[\underline{\mathbf{B}}\underline{\mathbf{B}}^\top]$: This result has been proven in Díaz-García (2013). However, our proof is more straightforward as it directly uses the stochastic representations in terms of the Bartlett matrices. We have

$$(\underline{\mathbf{B}}\underline{\mathbf{B}}^\top)_{ij} = \sum_{k=1}^{p} \underline{\mathbf{B}}_{ik}(\underline{\mathbf{B}}^\top)_{kj} = \sum_{k=1}^{p} \underline{\mathbf{B}}_{ik}\underline{\mathbf{B}}_{jk}. \tag{2.25}$$

For the off-diagonal elements, i.e. $i \neq j$, we have

$$\mathbb{E}[(\underline{\mathbf{B}}\underline{\mathbf{B}}^\top)_{ij}] = \sum_{k=1}^{p} \mathbb{E}[\underline{\mathbf{B}}_{ik}\underline{\mathbf{B}}_{jk}] = \sum_{k=1}^{p} \mathbb{E}[\underline{\mathbf{B}}_{ik}]\mathbb{E}[\underline{\mathbf{B}}_{jk}] = 0, \tag{2.26}$$

where we have used independence of the elements in $\underline{\mathbf{B}}$ and the fact that at least one of the elements in each summand above is a mean zero normal random variable. For the diagonal elements, i.e. $i = j$, we have

$$(\underline{\mathbf{B}}\underline{\mathbf{B}}^\top)_{ii} = \sum_{k=1}^{p} \underline{\mathbf{B}}_{ik}^2 = \sum_{k=1}^{i} \underline{\mathbf{B}}_{ik}^2, \tag{2.27}$$

which is the sum of a $\chi^2_{n_i-i+1}$ and $(i-1)$ independent $\mathcal{N}(0,1)^2$ random variables, which implies that

$$\sum_{k=1}^{i} \underline{\mathbf{B}}_{ik}^2 \sim \chi^2_{n_i} \tag{2.28}$$

with expectation n_i. Thus

$$\mathbb{E}[(\underline{\mathbf{B}}\underline{\mathbf{B}}^\top)_{ii}] = n_i. \tag{2.29}$$

Proof of $\mathbb{E}[(\bar{\mathbf{B}}\bar{\mathbf{B}}^\top)^{-1}]$: See Louati and Masmoudi (2015).
Proof of $\mathbb{E}[\bar{\mathbf{B}}^{-\top}\underline{\mathbf{B}}\underline{\mathbf{B}}^\top\bar{\mathbf{B}}^{-1}]$: See Theorem 10 in Blasques et al. (2021).
Proof of $\mathbb{E}[\underline{\mathbf{B}}(\bar{\mathbf{B}}\bar{\mathbf{B}}^\top)^{-1}\underline{\mathbf{B}}^\top]$: Due to independence, we have

$$\mathbb{E}[\underline{\mathbf{B}}\bar{\mathbf{B}}^{-\top}\bar{\mathbf{B}}^{-1}\underline{\mathbf{B}}^\top] = \mathbb{E}[\underline{\mathbf{B}}\mathrm{dg}(\mathring{\boldsymbol{\nu}})\underline{\mathbf{B}}^\top],$$

where $\mathring{\boldsymbol{\nu}}$ is given in (2.11). Denote

$$\mathbf{T} = \underline{\mathbf{B}}(\mathring{\boldsymbol{\nu}})^{1/2},$$

with elements $\mathbf{T}_{ij} = \underline{\mathbf{B}}_{ij}\sqrt{\mathring{\nu}_j}$. The (i,j)'th element of $\mathbf{R} = \mathbf{T}\mathbf{T}^\top$ is

$$\mathbf{R}_{ij} = \sum_{k=1}^{p} \mathbf{T}_{ik}(\mathbf{T}^\top)_{kj} = \sum_{k=1}^{p} \mathbf{T}_{ik}\mathbf{T}_{jk} = \sum_{k=1}^{p} \mathring{\nu}_k \, \underline{\mathbf{B}}_{ik}\underline{\mathbf{B}}_{jk},$$

which for $i \neq j$ we have

$$\mathbb{E}[\mathbf{R}_{ij}] = \sum_{k=1}^{p} \mathring{\nu}_k \, \mathbb{E}[\underline{\mathbf{B}}_{ik}\underline{\mathbf{B}}_{jk}] = \sum_{k=1}^{p} \mathring{\nu}_k \, \mathbb{E}[\underline{\mathbf{B}}_{ik}]\mathbb{E}[\underline{\mathbf{B}}_{jk}] = 0,$$

because of the independence of the elements in $\underline{\mathbf{B}}$ and the fact that at least one of the elements in each summand is mean zero. Furthermore, for $i = j$ we have

$$\mathbb{E}[\mathbf{R}_{ii}] = \sum_{k=1}^{p} \mathring{\nu}_k \, \mathbb{E}[\underline{\mathbf{B}}_{ik}^2] = \sum_{k=1}^{i} \mathring{\nu}_k \, \mathbb{E}[\underline{\mathbf{B}}_{ik}^2],$$

with

$$\mathbb{E}[\underline{\mathbf{B}}_{ik}^2] = \begin{cases} 1, & \text{for } i \neq k \\ n_k - k + 1 & \text{for } i = k. \end{cases}$$

Thus the elements of $\mathbb{E}_{iF\mathcal{R}}[\mathbf{R}] = \mathrm{dg}(\mathring{\mathring{\mathbf{n}}})$ are given by

$$\begin{aligned} \mathbb{E}[\mathbf{R}_{1\,1}] &= (n_1 - 1 + 1)\, \mathring{\nu}_1, \\ \mathbb{E}[\mathbf{R}_{2\,2}] &= \mathring{\nu}_1 + (n_2 - 2 + 1)\, \mathring{\nu}_2, \\ \mathbb{E}[\mathbf{R}_{3\,3}] &= \mathring{\nu}_1 + \mathring{\nu}_2 + (n_3 - 3 + 1)\, \mathring{\nu}_3, \\ &\vdots \end{aligned}$$

or

$$\mathring{\nu}_i = \sum_{j=1}^{i-1} \mathring{\nu}_j + (n_i - i + 1)\, \mathring{\nu}_i$$

or more precisely

$$\mathring{\nu}_i = \begin{cases} n_1 \, \mathring{\nu}_1, & \text{for } i = 1 \\ \sum_{j=1}^{i-1} \mathring{\nu}_j + (n_i + i - 1)\, \mathring{\nu}_i, & \text{for } i > 1, \end{cases} \tag{2.30}$$

which for $n_i = n$ and $\nu_i = \nu$ for all i equals $\frac{n}{\nu - p - 1}$. □

Proof of Lemma 2.2.1

Proof. The equivalence between the two different representations is proofed in Maaß (1971), pp. 69-70. This proof is closely based on it. Using the LDL and Cholesky decomposition of $\mathbf{\Sigma}$,

$$\mathbf{\Sigma} = \mathbf{T}\mathbf{D}\mathbf{T}^\top = \mathbf{C}\mathbf{C}^\top.$$

Then

$$\mathbf{\Sigma}_{[j]} = \mathbf{C}_{[j]}\mathbf{C}_{[j]}^\top = \mathbf{T}_{[j]}\mathbf{D}_{[j]}\mathbf{T}_{[j]}^\top,$$

where $\mathbf{X}_{[i]}$ denotes the square submatrix created by taking the first i rows and columns of $\mathbf{X}$. So

$$|\mathbf{\Sigma}_{[j]}| = \prod_{i=1}^{j} \mathbf{D}_{ii}$$

and thus

$$|\mathbf{\Sigma}_{[1]}| = \mathbf{D}_{1\,1} \text{ and for } j > 1 \text{ we have } |\mathbf{\Sigma}_{[j]}|/|\mathbf{\Sigma}_{[j-1]}| = \mathbf{D}_{jj}.$$

Finally

$$\prod_{i=1}^{p} \mathbf{D}_{ii}^{s_i} = |\mathbf{\Sigma}_{[1]}|^{s_1} \prod_{i=2}^{p} (|\mathbf{\Sigma}_{[i]}|/|\mathbf{\Sigma}_{[i-1]}|)^{s_i} = |\mathbf{\Sigma}_{[1]}|^{s_1-s_2}|\mathbf{\Sigma}_{[2]}|^{s_2-s_3}\ldots|\mathbf{\Sigma}_{[p]}|^{s_p}.$$

□

Proof of Theorem 2.2.2

Proof. We will make use of properties of probability limits of products of (inverse) random matrices and of Slutzky's Theorems for random matrices (see e.g. Theorems 5.6, 5.9 and 5.10 in Mittelhammer, 2013).

The non-zero off-diagonal elements of the lower triangular matrix $\mathrm{dg}(\mathbf{n})^{-\frac{1}{2}}\underset{\sim}{\mathbf{B}}$, $i < j$ are given by

$$(\mathrm{dg}(\mathbf{n})^{-\frac{1}{2}}\underset{\sim}{\mathbf{B}})_{ij} = \frac{(\underset{\sim}{\mathbf{B}})_{ij}}{\sqrt{n_i}} \xrightarrow{p} 0, \text{ as } n_i \to \infty,$$

since $(\underset{\sim}{\mathbf{B}})_{ij} \sim \mathcal{N}(0,1)$ for $i < j$.

Furthermore, note that for the *squared* diagonal ($i = j$) elements we have for $n_i \to \infty$,

$$\mathbb{E}\left[\frac{((\underline{\mathbf{B}})_{ii})^2}{n_i}\right] = \frac{n_i - i + 1}{n_i} \to 1 \quad \text{and}$$
$$\mathbf{Var}\left(\frac{((\underline{\mathbf{B}})_{ii})^2}{n_i}\right) = 2\frac{n_i - i + 1}{n_i^2} \to 0,$$

since $((\underline{\mathbf{B}})_{ii})^2$ is $\chi^2_{n_i - i + 1}$ distributed, and thus as $n_i \to \infty$

$$\frac{((\underline{\mathbf{B}})_{ii})^2}{n_i} \xrightarrow{p} 1 \Leftrightarrow \frac{(\underline{\mathbf{B}})_{ii}}{\sqrt{n_i}} \xrightarrow{p} 1,$$

where the equivalence follows from the Continuous Mapping Theorem. Finally, we can conclude that as $n_i \to \infty$ for all i,

$$\operatorname*{plim}_{\mathbf{n}\to\infty}(\mathrm{dg}(\mathbf{n})^{-\frac{1}{2}}\underline{\mathbf{B}}) = \mathbf{I}, \tag{2.31}$$

where $\mathbf{n} \to \infty$ means that all elements in $\mathbf{n}$ converge to infinity and the plim operator on a matrix is to be understood element-wise. By similar arguments we get that as $\nu_i \to \infty$ for all i,

$$\operatorname*{plim}_{\boldsymbol{\nu}\to\infty}(\bar{\mathbf{B}}\mathrm{dg}(\boldsymbol{\nu})^{-\frac{1}{2}}) = \mathbf{I}$$

and consequently

$$\operatorname*{plim}_{\boldsymbol{\nu}\to\infty}((\bar{\mathbf{B}}\mathrm{dg}(\boldsymbol{\nu})^{-\frac{1}{2}})^{-1}) = \mathbf{I}. \tag{2.32}$$

Now, we have

$$\mathrm{dg}(\mathring{\mathbf{n}})^{-\frac{1}{2}}\bar{\mathbf{B}}^{-\top}\underline{\mathbf{B}} = \underbrace{\mathrm{dg}(\mathring{\mathbf{n}})^{-\frac{1}{2}}\mathrm{dg}(\boldsymbol{\nu})^{-\frac{1}{2}}}_{\xrightarrow[\boldsymbol{\nu}\to\infty]{} \mathrm{dg}(\mathbf{n})^{-\frac{1}{2}}}\underbrace{\mathrm{dg}(\boldsymbol{\nu})^{\frac{1}{2}}\,\bar{\mathbf{B}}^{-\top}}_{\xrightarrow[\boldsymbol{\nu}\to\infty]{p} \mathbf{I}}\;\underline{\mathbf{B}} \xrightarrow[\boldsymbol{\nu}\to\infty]{d} \mathrm{dg}(\mathbf{n})^{-\frac{1}{2}}\underline{\mathbf{B}}$$

and

$$\mathrm{dg}(\mathring{\bar{\mathbf{n}}})^{-\frac{1}{2}}\underline{\mathbf{B}}\bar{\mathbf{B}}^{-\top} = \underbrace{\mathrm{dg}(\mathring{\bar{\mathbf{n}}})^{-\frac{1}{2}}\mathrm{dg}(\mathbf{n})^{\frac{1}{2}}}_{\xrightarrow[\mathbf{n}\to\infty]{} \mathrm{dg}(\mathring{\boldsymbol{\nu}})^{\frac{1}{2}}}\;\underbrace{\mathrm{dg}(\mathbf{n})^{-\frac{1}{2}}\underline{\mathbf{B}}}_{\xrightarrow[\mathbf{n}\to\infty]{p} \mathbf{I}}\;\bar{\mathbf{B}}^{-\top} \xrightarrow[\mathbf{n}\to\infty]{d} \mathrm{dg}(\mathring{\boldsymbol{\nu}})^{\frac{1}{2}}\bar{\mathbf{B}}^{-\top}$$

Finally,

$$\mathrm{dg}(\mathring{\mathbf{n}})^{-\frac{1}{2}}\bar{\mathbf{B}}^{-\top}\underline{\mathbf{B}}\underline{\mathbf{B}}^{\top}\bar{\mathbf{B}}^{-1}\mathrm{dg}(\mathring{\mathbf{n}})^{-\frac{1}{2}} \xrightarrow[\boldsymbol{\nu}\to\infty]{d} \mathrm{dg}(\mathbf{n})^{-\frac{1}{2}}\underline{\mathbf{B}}\underline{\mathbf{B}}^{\top}\mathrm{dg}(\mathbf{n})^{-\frac{1}{2}} \quad \text{and}$$
$$\mathrm{dg}(\mathring{\bar{\mathbf{n}}})^{-\frac{1}{2}}\underline{\mathbf{B}}\bar{\mathbf{B}}^{-\top}\bar{\mathbf{B}}^{-1}\underline{\mathbf{B}}^{\top}\mathrm{dg}(\mathring{\bar{\mathbf{n}}})^{-\frac{1}{2}} \xrightarrow[\mathbf{n}\to\infty]{d} \mathrm{dg}(\mathring{\boldsymbol{\nu}})^{\frac{1}{2}}\bar{\mathbf{B}}^{-\top}\bar{\mathbf{B}}^{-1}\mathrm{dg}(\mathring{\boldsymbol{\nu}})^{\frac{1}{2}},$$

which are the stochastic representations of the Riesz and Inverse Riesz, respectively.

The proofs for

$$(\nu-2)(\chi^2_\nu)^{-1}\mathrm{dg}(\mathbf{n})^{-\frac{1}{2}}\underline{\mathbf{B}}\underline{\mathbf{B}}^{\top}\mathrm{dg}(\mathbf{n})^{-\frac{1}{2}} \xrightarrow[\nu\to\infty]{d} \mathrm{dg}(\mathbf{n})^{-\frac{1}{2}}\underline{\mathbf{B}}\underline{\mathbf{B}}^{\top}\mathrm{dg}(\mathbf{n})^{-\frac{1}{2}} \quad \text{and}$$
$$\frac{\chi^2_n}{n}\mathrm{dg}(\mathring{\boldsymbol{\nu}})^{\frac{1}{2}}\bar{\mathbf{B}}^{-\top}\bar{\mathbf{B}}^{-1}\mathrm{dg}(\mathring{\boldsymbol{\nu}})^{\frac{1}{2}} \xrightarrow[n\to\infty]{d} \mathrm{dg}(\mathring{\boldsymbol{\nu}})^{\frac{1}{2}}\bar{\mathbf{B}}^{-\top}\bar{\mathbf{B}}^{-1}\mathrm{dg}(\mathring{\boldsymbol{\nu}})^{\frac{1}{2}}$$

are very easy, noticing that $(\nu-2)(\chi^2_\nu)^{-1}$ and χ^2_n/n converge in probability to 1. □

Proof of Theorem 2.2.3

Proof. All proofs start from the stochastic representations given in Table 2.1. The two integrals in the following lemma are important for the derivation of the p.d.f.s of the Riesz-type distributions.

Lemma 2.7.1. *(Faraut and Korányi, 1994) For* $\mathbf{n}$ *with* $n_i > i-1$ *we have,*

$$\int_{\mathbf{A}>\mathbf{0}} |\mathbf{A}|_{\frac{\mathbf{n}-p-1}{2}} \mathrm{etr}\left(-\frac{1}{2}\mathbf{B}\mathbf{A}\right) d\mathbf{A} = 2^{p\bar{\mathbf{n}}/2}\Gamma_p\left(\frac{\mathbf{n}}{2}\right)|\mathbf{B}^{-1}|_{\frac{\mathbf{n}}{2}} \tag{2.33}$$

and for $n_i < i-p$ *we have,*

$$\int_{\mathbf{A}>\mathbf{0}} |\mathbf{A}^{-1}|_{\frac{\mathbf{n}+p+1}{2}} \mathrm{etr}\left(-\frac{1}{2}\mathbf{B}\mathbf{A}\right) d\mathbf{A} = \frac{1}{2^{p\bar{\mathbf{n}}/2}}\Gamma_p\left(-\frac{\overleftarrow{\mathbf{n}}}{2}\right)|\mathbf{B}|_{\frac{\mathbf{n}}{2}}\,. \tag{2.34}$$

Proof. The proofs can be found in Faraut and Korányi (1994) chapter VII.[28]. Throughout, according to their table on p. 97, for the cone of symmetric positive definite matrices, we have the dimension $n = p(p+1)/2$, the rank $r = p$ and $d = 1$.[29] Throughout their book, they use the Euclidean measure on a Euclidean

28. Further references are Díaz-García (2014), Maaß (1971) p. 76, Gupta and Nagar (2000), Theorem 1.4.7, which is based on Olkin (1959), which in turn is based on the generalized Ingham formula in Bellman (1956).

29. For the notation see their Example 2 on p. 8 and p. 9.

space, which translated into our notation is $dx = \prod_{i=1}^{p} a_{ii} 2^{p(p-1)/4} \prod_{i<j} a_{ij} = 2^{p(p-1)/4} d\mathbf{A}$ and leads to a slightly different multivariate gamma function.[30] In particular, from their Theorem VII.1.1.

$$\Gamma_\Omega(\mathbf{n}) = 2^{p(p-1)/4} \Gamma_p(\mathbf{n}), \tag{2.35}$$

with $\Gamma_p(\mathbf{n})$ as in Definition 2.2.2. Their Proposition VII.1.2., with $x = \mathbf{A}$, $y = \frac{1}{2}\mathbf{B}$ and $\mathbf{s} = \frac{\mathbf{n}}{2}$ translates to

$$\begin{aligned}\int_{\mathbf{A}>\mathbf{0}} |\mathbf{A}|_{\frac{\mathbf{n}-p-1}{2}} \operatorname{etr}\left(-\frac{1}{2}\mathbf{B}\mathbf{A}\right) 2^{p(p-1)/4} d\mathbf{A} &= 2^{p(p-1)/4} \Gamma_p\left(\frac{\mathbf{n}}{2}\right) |2\mathbf{B}^{-1}|_{\frac{\mathbf{n}}{2}} \\ &= 2^{p(p-1)/4} \Gamma_p\left(\frac{\mathbf{n}}{2}\right) 2^{p\bar{\mathbf{n}}/2} |\mathbf{B}^{-1}|_{\frac{\mathbf{n}}{2}}.\end{aligned}$$

Their last equation on page 129, together with Proposition VII.1.5 (ii) and $x = \mathbf{A}$, $y = \frac{1}{2}\mathbf{B}$ and $\mathbf{s} = \frac{\mathbf{n}}{2}$ translates to

$$\begin{aligned}\int_{\mathbf{A}>\mathbf{0}} |\mathbf{A}^{-1}|_{\frac{\mathbf{n}+p+1}{2}} \operatorname{etr}\left(-\frac{1}{2}\mathbf{B}\mathbf{A}\right) 2^{p(p-1)/4} d\mathbf{A} &= 2^{p(p-1)/4} \Gamma_p\left(-\frac{\overleftarrow{\mathbf{n}}}{2}\right) \left|\frac{1}{2}\mathbf{B}\right|_{\frac{\mathbf{n}}{2}} \\ &= 2^{p(p-1)/4} \Gamma_p\left(-\frac{\overleftarrow{\mathbf{n}}}{2}\right) \frac{1}{2^{p\bar{\mathbf{n}}/2}} |\mathbf{B}|_{\frac{\mathbf{n}}{2}}.\end{aligned}$$

□

<u>t-Riesz distribution:</u> The stochastic representation is $\mathbf{R} = \mathbf{C}_\Omega(\bar{b})^{-2} \underline{\mathbf{B}}\underline{\mathbf{B}}^\top \mathbf{C}_\Omega^\top$, which can be written as $\mathbf{R} = w^{-1}\mathbf{A}$, with $\mathbf{A} \sim \mathcal{R}(\mathbf{\Omega}, \mathbf{n})$ independent of $w \sim \chi^2_\nu$. The joint p.d.f. of w and $\mathbf{A}$ is given by

$$\frac{1}{\Gamma(\nu/2)\, 2^{\nu/2}} w^{\frac{\nu}{2}-1} \exp\left(-\frac{w}{2}\right) \frac{|\mathbf{A}|_{\frac{\mathbf{n}-p-1}{2}} \exp\left(-\frac{1}{2}\operatorname{tr}\left(\mathbf{\Omega}^{-1}\mathbf{A}\right)\right)}{|\mathbf{\Omega}|_{\frac{\mathbf{n}}{2}} \Gamma_p(\mathbf{n}/2)\, 2^{p\bar{\mathbf{n}}/2}}.$$

Transforming $\mathbf{R} = w^{-1}\mathbf{A}$, with Jacobian $J(w, \mathbf{A} \to w, \mathbf{R}) = w^{p(p+1)/2}$ (see e.g. Gupta and Nagar, 2000, equation 1.3.5.), we get the joint density of w and $\mathbf{R}$ as

$$\begin{aligned}&\frac{1}{\Gamma(\nu/2)\, 2^{\nu/2}} w^{\frac{\nu}{2}-1} \exp\left(-\frac{w}{2}\right) \frac{|w\mathbf{R}|_{\frac{\mathbf{n}-p-1}{2}} \exp\left(-\frac{w}{2}\operatorname{tr}\left(\mathbf{\Omega}^{-1}\mathbf{R}\right)\right)}{|\mathbf{\Omega}|_{\frac{\mathbf{n}}{2}} \Gamma_p(\mathbf{n}/2)\, 2^{p\bar{\mathbf{n}}/2}} w^{\frac{p(p+1)}{2}} \\ =&\frac{|\mathbf{\Omega}|_{-\frac{\mathbf{n}}{2}} |\mathbf{R}|_{\frac{\mathbf{n}-p-1}{2}}}{\Gamma(\nu/2)\, \Gamma_p(\mathbf{n}/2)\, 2^{(\nu+p\bar{\mathbf{n}})/2}} w^{\frac{\nu+p\bar{\mathbf{n}}}{2}-1} \exp\left(-\frac{w}{2}\left(1 + \operatorname{tr}\left(\mathbf{\Omega}^{-1}\mathbf{R}\right)\right)\right),\end{aligned}$$

30. I thank Jacques Faraut for pointing this out to me.

where

$$|w\mathbf{R}|_{\frac{\mathbf{n}-p-1}{2}} = |\mathbf{R}|_{\frac{\mathbf{n}-p-1}{2}} \prod_{i=1}^{p} w^{\frac{n_i-p-1}{2}} = |\mathbf{R}|_{\frac{\mathbf{n}-p-1}{2}} w^{p\frac{\bar{\mathbf{n}}-(p+1)}{2}}.$$

Now integrating out w we get the p.d.f. of $\mathbf{R}$ as

$$\begin{aligned} p_{t\mathcal{R}}(\mathbf{R}|\mathbf{\Omega},\mathbf{n},\nu) &= \frac{|\mathbf{\Omega}|_{-\frac{\mathbf{n}}{2}}|\mathbf{R}|_{\frac{\mathbf{n}-p-1}{2}}}{\Gamma(\nu/2)\,\Gamma_p(\mathbf{n}/2)\,2^{(\nu+p\bar{\mathbf{n}})/2}} \\ &\quad \times \int_0^\infty w^{\frac{\nu+p\bar{\mathbf{n}}}{2}-1} \exp\left(-\frac{w}{2}\left(1+\mathrm{tr}\left(\mathbf{\Omega}^{-1}\mathbf{R}\right)\right)\right) dw \\ &= \frac{|\mathbf{\Omega}|_{-\frac{\mathbf{n}}{2}}|\mathbf{R}|_{\frac{\mathbf{n}-p-1}{2}}}{\Gamma(\nu/2)\,\Gamma_p(\mathbf{n}/2)\,2^{(\nu+p\bar{\mathbf{n}})/2}} \Gamma\left((\nu+p\bar{\mathbf{n}})/2\right)\left[\frac{1}{2}\left(1+\mathrm{tr}\left(\mathbf{\Omega}^{-1}\mathbf{R}\right)\right)\right]^{-(\nu+p\bar{\mathbf{n}})/2} \\ &= \frac{\Gamma\left((\nu+p\bar{\mathbf{n}})/2\right)}{\Gamma(\nu/2)\Gamma_p(\mathbf{n}/2)}|\mathbf{\Omega}|_{-\frac{\mathbf{n}}{2}}|\mathbf{R}|_{\frac{\mathbf{n}-p-1}{2}}\left(1+\mathrm{tr}\left(\mathbf{\Omega}^{-1}\mathbf{R}\right)\right)^{-\frac{\nu+p\bar{\mathbf{n}}}{2}}, \end{aligned}$$

where we used equation (5.9.1) of the NIST Digital Library of Mathematical Functions.

Inverse t-Riesz distribution: We have $\mathbf{R} = \mathbf{C}_\Omega(\underline{b})^2\left(\bar{\mathbf{B}}\bar{\mathbf{B}}^\top\right)^{-1}\mathbf{C}_\Omega^\top$, which can be written as $\mathbf{R} = w\mathbf{A}$, with $\mathbf{A} \sim i\mathcal{R}(\mathbf{\Omega},\boldsymbol{\nu})$ independent of $w \sim \chi^2_n$. The joint p.d.f. of w and $\mathbf{A}$ is given by

$$\frac{1}{\Gamma(n/2)\,2^{n/2}} w^{\frac{n}{2}-1}\exp\left(-\frac{w}{2}\right)\frac{|\mathbf{\Omega}|_{\frac{\boldsymbol{\nu}}{2}}|\mathbf{A}|_{-\frac{\boldsymbol{\nu}+p+1}{2}}}{\Gamma_p\left(\overleftarrow{\boldsymbol{\nu}}/2\right)2^{p\bar{\boldsymbol{\nu}}/2}}\exp\left(-\frac{1}{2}\mathrm{tr}\left(\mathbf{\Omega}\mathbf{A}^{-1}\right)\right).$$

Transforming $\mathbf{R} = w\mathbf{A}$, with Jacobian $J(w,\mathbf{A}\to w,\mathbf{R}) = w^{-p(p+1)/2}$ (see e.g. Gupta and Nagar, 2000, equation 1.3.5.), we get the joint density of w and $\mathbf{R}$ as

$$\begin{aligned} &\frac{w^{n/2-1}}{\Gamma(n/2)\,2^{n/2}}\exp\left(-\frac{w}{2}\right)\frac{|\mathbf{\Omega}|_{\frac{\boldsymbol{\nu}}{2}}|w^{-1}\mathbf{R}|_{-\frac{\boldsymbol{\nu}+p+1}{2}}}{\Gamma_p\left(\overleftarrow{\boldsymbol{\nu}}/2\right)2^{p\bar{\boldsymbol{\nu}}/2}}\exp\left(-\frac{1}{2}\mathrm{tr}\left(w\mathbf{\Omega}\mathbf{R}^{-1}\right)\right)w^{-\frac{p(p+1)}{2}} \\ =&\frac{|\mathbf{\Omega}|_{\frac{\boldsymbol{\nu}}{2}}|\mathbf{R}|_{-\frac{\boldsymbol{\nu}+p+1}{2}}}{\Gamma(n/2)\,\Gamma_p\left(\overleftarrow{\boldsymbol{\nu}}/2\right)2^{(n+p\bar{\boldsymbol{\nu}})/2}}w^{\frac{n+p\bar{\boldsymbol{\nu}}}{2}-1}\exp\left(-\frac{w}{2}\left(1+\mathrm{tr}\left(\mathbf{\Omega}\mathbf{R}^{-1}\right)\right)\right), \end{aligned}$$

where

$$|w\mathbf{R}|_{-\frac{\boldsymbol{\nu}+p+1}{2}} = |\mathbf{R}|_{-\frac{\boldsymbol{\nu}+p+1}{2}}\prod_{i=1}^{p} w^{\frac{\nu_i+p+1}{2}} = |\mathbf{R}|_{-\frac{\boldsymbol{\nu}+p+1}{2}} w^{p\frac{\bar{\boldsymbol{\nu}}+(p+1)}{2}}.$$

Now integrating out w we get the p.d.f. of $\mathbf{R}$ as

$$
\begin{aligned}
p_{it\mathcal{R}}(\mathbf{R}|\mathbf{\Omega}, n, \boldsymbol{\nu}) &= \frac{|\mathbf{\Omega}|^{\frac{\nu}{2}} |\mathbf{R}|^{-\frac{\nu+p+1}{2}}}{\Gamma(n/2)\,\Gamma_p\left(\overleftarrow{\boldsymbol{\nu}}/2\right) 2^{(n+p\bar{\nu})/2}} \\
&\quad \times \int_0^\infty w^{\frac{n+p\bar{\nu}}{2}-1} \exp\left(-\frac{w}{2}\left(1+\operatorname{tr}\left(\mathbf{\Omega}\mathbf{R}^{-1}\right)\right)\right) dw \\
&= \frac{|\mathbf{\Omega}|^{\frac{\nu}{2}} |\mathbf{R}|^{-\frac{\nu+p+1}{2}}}{\Gamma(n/2)\,\Gamma_p\left(\overleftarrow{\boldsymbol{\nu}}/2\right) 2^{(n+p\bar{\nu})/2}} \Gamma\left((n+p\bar{\boldsymbol{\nu}})/2\right) \left(\frac{1}{2}\left(1+\operatorname{tr}\left(\mathbf{\Omega}\mathbf{R}^{-1}\right)\right)\right)^{-\frac{n+p\bar{\nu}}{2}} \\
&= \frac{\Gamma\left((n+p\bar{\boldsymbol{\nu}})/2\right)}{\Gamma(n/2)\,\Gamma_p\left(\overleftarrow{\boldsymbol{\nu}}/2\right)} |\mathbf{\Omega}|^{\frac{\nu}{2}} |\mathbf{R}|^{-\frac{\nu+p+1}{2}} \left(\left(1+\operatorname{tr}\left(\mathbf{\Omega}\mathbf{R}^{-1}\right)\right)\right)^{-\frac{n+p\bar{\nu}}{2}},
\end{aligned}
$$

where we used equation (5.9.1) of the NIST Digital Library of Mathematical Functions.

Inverse F-Riesz distribution: The stochastic representation an F-Riesz distribution of type II with scale matrix $\mathbf{\Omega}^{-1}$, and d.o.f. parameter vectors $\boldsymbol{\nu}$ and $\mathbf{n}$ is $\mathbf{U}_{\mathbf{\Omega}^{-1}} \underline{\mathbf{B}}^{-1} \bar{\mathbf{B}} \bar{\mathbf{B}}^\top \underline{\mathbf{B}}^{-1} \mathbf{U}_{\mathbf{\Omega}^{-1}}^\top$, where $\mathbf{U}_{\mathbf{\Omega}^{-1}}$ is the upper Cholesky factor of $\mathbf{\Omega}^{-1}$.[31] Thus the stochastic representation of the Inverse F-Riesz distribution of type II is given by

$$
\mathbf{R} = \mathbf{C}_\Omega \underline{\mathbf{B}} \bar{\mathbf{B}}^{-\top} \bar{\mathbf{B}}^{-1} \underline{\mathbf{B}}^\top \mathbf{C}_\Omega^\top, \tag{2.36}
$$

which translate to $\mathbf{R} \sim i\mathcal{R}^{II}(\mathbf{Y}, \boldsymbol{\nu})$, $\mathbf{Y} \sim \mathcal{R}^I(\mathbf{\Omega}, \mathbf{n})$.[32] For the p.d.f. we can

31. See Blasques et al. (2021).
32. Recall that $\mathbf{U}_{\mathbf{\Omega}^{-1}}^{-\top} = \mathbf{C}_\Omega$.

consequently use

$$
\begin{aligned}
p_{iF\mathcal{R}}(\mathbf{R}|\mathbf{\Omega},\mathbf{n},\boldsymbol{\nu}) &= \int_{\mathbf{Y}>\mathbf{0}} p_{i\mathcal{R}^{ii}}(\mathbf{R}|\mathbf{Y},\boldsymbol{\nu}) p_{\mathcal{R}^{i}}(\mathbf{Y}|\mathbf{\Omega},\mathbf{n}) d\mathbf{Y} \\
&= \int_{\mathbf{Y}>\mathbf{0}} \left(|\mathbf{R}|_{-\frac{\boldsymbol{\nu}+p+1}{2}} |\mathbf{Y}|_{\frac{\boldsymbol{\nu}}{2}} \mathrm{etr}\left(-\frac{1}{2}\mathbf{Y}\mathbf{R}^{-1}\right) \frac{1}{\Gamma_p(\overleftarrow{\boldsymbol{\nu}}/2) 2^{p\boldsymbol{\nu}/2}} \right. \\
&\qquad \left. \times |\mathbf{Y}|_{\frac{\mathbf{n}-p-1}{2}} |\mathbf{\Omega}|_{-\frac{\mathbf{n}}{2}} \mathrm{etr}\left(-\frac{1}{2}\mathbf{\Omega}^{-1}\mathbf{Y}\right) \frac{1}{\Gamma_p(\mathbf{n}/2)\, 2^{p\bar{\mathbf{n}}/2}} \right) d\mathbf{Y} \\
&= \frac{1}{\Gamma_p(\overleftarrow{\boldsymbol{\nu}}/2)\Gamma_p(\mathbf{n}/2)\, 2^{p(\bar{\boldsymbol{\nu}}+\bar{\mathbf{n}})/2}} |\mathbf{R}|_{-\frac{\boldsymbol{\nu}+p+1}{2}} |\mathbf{\Omega}|_{-\frac{\mathbf{n}}{2}} \\
&\quad \times \int_{\mathbf{Y}>\mathbf{0}} |\mathbf{Y}|_{\frac{\mathbf{n}+\boldsymbol{\nu}-p-1}{2}} \mathrm{etr}\left(-\frac{1}{2}\mathbf{Y}(\mathbf{\Omega}^{-1}+\mathbf{R}^{-1})\right) d\mathbf{Y} \\
&= \frac{2^{p(\bar{\boldsymbol{\nu}}+\bar{\mathbf{n}})/2}\Gamma_p((\boldsymbol{\nu}+\mathbf{n})/2)}{\Gamma_p(\overleftarrow{\boldsymbol{\nu}}/2)\Gamma_p(\mathbf{n}/2)\, 2^{p(\bar{\boldsymbol{\nu}}+\bar{\mathbf{n}})/2}} |\mathbf{R}|_{-\frac{\boldsymbol{\nu}+p+1}{2}} |\mathbf{\Omega}|_{-\frac{\mathbf{n}}{2}} \left| \left(\mathbf{\Omega}^{-1}+\mathbf{R}^{-1}\right)^{-1} \right|_{\frac{\boldsymbol{\nu}+\mathbf{n}}{2}} \\
&= \frac{\Gamma_p((\boldsymbol{\nu}+\mathbf{n})/2)}{\Gamma_p(\overleftarrow{\boldsymbol{\nu}}/2)\Gamma_p(\mathbf{n}/2)} |\mathbf{R}|_{-\frac{\boldsymbol{\nu}+p+1}{2}} |\mathbf{\Omega}|_{-\frac{\mathbf{n}}{2}} |(\mathbf{\Omega}^{-1}+\mathbf{R}^{-1})^{-1}|_{\frac{\boldsymbol{\nu}+\mathbf{n}}{2}}, \qquad (2.37)
\end{aligned}
$$

where we used Theorem 2.7.1. Now rewrite using Lemma 2.2.2. □

Proof of Theorem 2.3.2

Proof. It is well-known that $\dot{\mathbf{r}}$ has stochastic representation $\dot{\mathbf{r}} = \sqrt{y}\mathbf{C}_{\dot{\mathbf{\Omega}}}\dot{\mathbf{z}}$ with $\dot{z}_i \overset{iid}{\sim} \mathcal{N}(0,1)$ and $y \sim \Gamma(\nu/2, 2/\nu)$. Equivalently, we can fix all entries with missing observations in $\tilde{\mathbf{r}}$ equal to zero and write $\tilde{\mathbf{r}} = \sqrt{y}\mathbf{C}_{\tilde{\mathbf{\Omega}}}\tilde{\mathbf{z}}$, with

$$
\tilde{z}_i = \begin{cases} 0 \text{ if there is a missing observation,} \\ \overset{iid}{\sim} \mathcal{N}(0,1) \text{ else.} \end{cases}
$$

The RC can be written as

$$
\mathbf{R} = \sum_{j=1}^{n} \mathbf{r}_j \mathbf{r}_j^\top = y\mathbf{C}_\Omega \mathbf{Z}\mathbf{Z}^\top \mathbf{C}_\Omega^\top,
$$

where the $p \times n$ matrix $\mathbf{Z} = (\mathbf{z}_1, \mathbf{z}_2, \ldots \mathbf{z}_n)$ with $\mathbf{z}_j = \mathbf{C}_\Omega^{-1}\mathbf{r}_j$. Hassairi, Ktari, and Zine (2022) show that if the assets in $\mathbf{Z}$ are sorted according to their liquidity with the least liquid asset in the first row, then $\mathbf{Z}\mathbf{Z}^\top$ follows a Riesz distribution with parameter matrix $\mathbf{\Omega} = \mathbf{I}$ and d.o.f. parameter vector $\mathbf{n}$, which implies that $\mathbf{C}_\Omega \mathbf{Z}\mathbf{Z}^\top \mathbf{C}_\Omega^\top \sim \mathcal{R}(\mathbf{\Omega}, \mathbf{n})$ follows a Riesz distribution with parameter matrix $\mathbf{\Omega}$ and d.o.f. parameter vector $\mathbf{n}$. Then according to Theorem 2.2.3 $\mathbf{R} \sim t\mathcal{R}(\mathbf{\Omega}, \mathbf{n}, \nu)$. □

Proof of Theorem 2.3.1

Proof. This Theorem is closely based on Gupta, Varga, and Bodnar (2013). If $\tilde{\mathbf{r}}$ follows an elliptically contoured distribution, $\tilde{\mathbf{r}} \sim E_{np}(\mathbf{0}, \mathbf{I}_n \otimes \mathbf{\Omega}, \psi)$, then according to their Theorem 2.1 $\mathbf{X}^\top \sim E_{n,p}(\mathbf{0}, \mathbf{I}_n \otimes \mathbf{\Omega}, \psi)$ and then according to their Theorem 2.3 $\mathbf{X} \sim E_{p,n}(\mathbf{0}, \mathbf{\Omega} \otimes \mathbf{I}_n, \psi)$. Then our Theorem follows from their Corollary 5.1 and by noticing that

$$\begin{aligned} \operatorname{tr}\left(\mathbf{X}^\top \mathbf{\Omega}^{-1} \mathbf{X}\right) &\overset{(3.14)}{=} \operatorname{vec}(\mathbf{X})^\top \operatorname{vec}(\mathbf{\Omega}^{-1}\mathbf{X}) \overset{(3.12)}{=} \operatorname{vec}(\mathbf{X})^\top (\mathbf{I} \otimes \mathbf{\Omega}^{-1}) \operatorname{vec}(\mathbf{X}) \\ &= \tilde{\mathbf{r}}^\top (\mathbf{I} \otimes \mathbf{\Omega}^{-1}) \tilde{\mathbf{r}} \text{ and} \\ \mathbf{R} &= \mathbf{X}\mathbf{X}^\top = \sum_{j=1}^{n} \mathbf{r}_j \mathbf{r}_j^\top . \end{aligned}$$

□

Probability Density Functions 2.7.2

Our first aim is to derive the p.d.f.s in Table 2.5 from the stochastic representations of the respective distribution given by equation (2.4) in conjunction with Table 2.1.

For the derivation of the Riesz, Inverse Riesz and F-Riesz p.d.f.s we refer to Blasques et al. (2021), where to translate their notation to ours we use $\bar{\Gamma}_U(\mathbf{n}) = \Gamma_p(\overleftarrow{\mathbf{n}})$ (Lemma 2.7.2), ${}_U|\mathbf{X}|_{\mathbf{n}} = |\mathbf{X}^{-1}|_{-\mathbf{n}}$ (Lemma 2.2.2 (iv)) and $\mathbf{\Sigma}_{Blasques} = \mathbf{\Omega}$.

Lemma 2.7.2. *Let the upper generalized multivariate gamma function, $\bar{\Gamma}_U(\cdot)$, be defined as in Blasques et al. (2021) and denote a vector with its elements in reverse order by a superscript left arrow, e.g. $\overleftarrow{\mathbf{n}} = (n_p, n_{p-1}, \ldots, n_1)^\top$, then*

$$\Gamma_p\left(\overleftarrow{\mathbf{n}}\right) = \bar{\Gamma}_U(\mathbf{n}) .$$

Proof. We have $(\overleftarrow{\mathbf{n}})_i = n_{p-i+1}$, such that

$$\begin{aligned} \Gamma_p\left(\overleftarrow{\mathbf{n}}\right) &= \pi^{p(p-1)/2} \prod_{i=1}^{p} \Gamma\left(n_{p-i+1} - \frac{i-1}{2}\right) \\ &= \pi^{p(p-1)/2} \prod_{i=1}^{p} \Gamma\left(n_i - \frac{p-1}{2}\right) = \bar{\Gamma}_U(\mathbf{n}) . \end{aligned}$$

□

In particular, the p.d.f. of the Riesz distribution ($\mathbf{C}_\Omega \underline{\mathbf{B}} \underline{\mathbf{B}}^\top \mathbf{C}_\Omega^\top$) has been derived in Theorems 4 (i) of Blasques et al. (2021).

Our stochastic representation of the Inverse Riesz distribution ($\mathbf{C}_\Omega \bar{\mathbf{B}}^{-\top} \bar{\mathbf{B}}^{-1} \mathbf{C}_\Omega^\top$) is the same as the one of the Inverse Riesz type *II* in Blasques et al. (2021) ($\mathbf{U}^{-\top} \bar{\mathbf{B}}^{-\top} \bar{\mathbf{B}}^{-1} \mathbf{U}^{-1}$) (see their Theorem 4 (ii) and Definition 6 (ii)), because $\mathbf{U}$ is the upper Cholesky factor of $\mathbf{\Omega}^{-1}$ and consequently $\mathbf{U}^{-\top} = \mathbf{C}_\Omega$ is the lower Cholesky factor of $\mathbf{\Omega}$. The corresponding p.d.f. is given in their Theorem 7 (ii).

For the F-Riesz distribution ($\mathbf{C}_\Omega \bar{\mathbf{B}}^{-\top} \underline{\mathbf{B}} \underline{\mathbf{B}}^\top \bar{\mathbf{B}}^{-1} \mathbf{C}_\Omega^\top$) use their Theorem 8 (i) and notice that their $\mathbf{Y} = \mathbf{C}_\Omega \bar{\mathbf{B}}^{-\top} \bar{\mathbf{B}}^{-1} \mathbf{C}_\Omega^\top$ and thus, according to their Theorem 4 (i) their $\mathbf{X}|\mathbf{Y} = \mathbf{C}_\Omega \bar{\mathbf{B}}^{-\top} \underline{\mathbf{B}} \underline{\mathbf{B}}^\top \bar{\mathbf{B}}^{-1} \mathbf{C}_\Omega^\top$. So the stochastic representations are identical, and the corresponding p.d.f. is given in their Theorem 8 (i) and can be rewritten using Lemma 2.2.2 (iv) as

$$\begin{aligned} p_{F\mathcal{R}} &= \frac{\Gamma_p((\overleftarrow{\mathbf{n}} + \overleftarrow{\boldsymbol{\nu}})/2)}{\Gamma_p(\mathbf{n}/2)\Gamma_p(\overleftarrow{\boldsymbol{\nu}}/2)} |\mathbf{\Omega}|^{\frac{\nu}{2}} |\mathbf{R}|^{\frac{\mathbf{n}-p-1}{2}} |\mathbf{\Omega} + \mathbf{R}|^{-\frac{\mathbf{n}+\boldsymbol{\nu}}{2}} \\ &= \frac{\Gamma_p((\overleftarrow{\mathbf{n}} + \overleftarrow{\boldsymbol{\nu}})/2)}{\Gamma_p(\mathbf{n}/2)\Gamma_p(\overleftarrow{\boldsymbol{\nu}}/2)} |\mathbf{\Omega}|^{-\frac{\mathbf{n}}{2}} |\mathbf{R}|^{\frac{\mathbf{n}-p-1}{2}} |\mathbf{I} + \mathbf{C}_\Omega^{-1} \mathbf{R} \mathbf{C}_\Omega^{-\top}|^{-\frac{\mathbf{n}+\boldsymbol{\nu}}{2}}. \end{aligned} \tag{2.38}$$

The p.d.f.s of the t-Riesz, Inverse t-Riesz, and Inverse F-Riesz distributions are derived in Theorem 2.2.3. The one of the Inverse F-Riesz can be rewritten using Lemma 2.2.2 (iv) as

$$\begin{aligned} p_{iF\mathcal{R}} &= \frac{\Gamma_p((\boldsymbol{\nu} + \mathbf{n})/2)}{\Gamma_p(\overleftarrow{\boldsymbol{\nu}}/2)\Gamma_p(\mathbf{n}/2)} |\mathbf{\Omega}|^{-\frac{\mathbf{n}}{2}} |\mathbf{R}|^{-\frac{\nu+p+1}{2}} |(\mathbf{\Omega}^{-1} + \mathbf{R}^{-1})^{-1}|^{\frac{\nu+\mathbf{n}}{2}} \\ &= \frac{\Gamma_p((\boldsymbol{\nu} + \mathbf{n})/2)}{\Gamma_p(\overleftarrow{\boldsymbol{\nu}}/2)\Gamma_p(\mathbf{n}/2)} |\mathbf{\Omega}|^{-\frac{\mathbf{n}}{2}} |\mathbf{R}|^{\frac{\mathbf{n}-p-1}{2}} |(\mathbf{I} + \mathbf{C}_R^\top \mathbf{\Omega}^{-1} \mathbf{C}_R)^{-1}|^{\frac{\nu+\mathbf{n}}{2}} \\ &= \frac{\Gamma_p((\boldsymbol{\nu} + \mathbf{n})/2)}{\Gamma_p(\overleftarrow{\boldsymbol{\nu}}/2)\Gamma_p(\mathbf{n}/2)} |\mathbf{\Omega}|^{\frac{\nu}{2}} |\mathbf{R}|^{-\frac{\nu+p+1}{2}} |(\mathbf{I} + \mathbf{C}_\Omega^\top \mathbf{R}^{-1} \mathbf{C}_\Omega)^{-1}|^{\frac{\nu+\mathbf{n}}{2}}. \end{aligned}$$

Now that we have derived all Riesz-type p.d.f.s, it is easy to get the Wishart-type p.d.f.s since they are just special cases where all elements in the d.o.f. parameter vectors are equal to each other and using that for $\mathbf{n} = (n, n, \ldots, n)$, $\Gamma_p(\mathbf{n}) = \Gamma_p(n)$ and $|\mathbf{X}|_{\mathbf{n}}| = |\mathbf{X}|^n$ (see Definitions 2.2.1 and 2.2.2).

Notice that the stochastic representations of the F-Riesz and Inverse F-Riesz are, if $\forall i$, $n_i = n$ and $\nu_i = \nu$, (i.e. in case of an F distribution) $\mathbf{C}_\Omega \bar{\mathcal{B}}^{-\top} \underline{\mathcal{B}} \underline{\mathcal{B}}^\top \bar{\mathcal{B}}^{-1} \mathbf{C}_\Omega^\top$

and $\mathbf{C}_\Omega \underset{\sim}{\mathcal{B}} \bar{\mathcal{B}}^{-\top} \bar{\mathcal{B}}^{-1} \underset{\sim}{\mathcal{B}}^\top \mathbf{C}_\Omega^\top$, respectively, and their p.d.f.s are identical,

$$\begin{aligned}
&p_{F\mathcal{R}}(\mathbf{R}|\mathbf{\Omega}, (n,\dots,n), (\nu,\dots,\nu)) \\
&\quad= \frac{\Gamma_p((n+\nu)/2)}{\Gamma_p(n/2)\Gamma_p(\nu/2)} |\mathbf{\Omega}|^{-\frac{n}{2}} |\mathbf{R}|^{\frac{n-p-1}{2}} |\mathbf{I} + \mathbf{C}_\Omega^{-1}\mathbf{R}\mathbf{C}_\Omega^{-\top}|^{-\frac{n+\nu}{2}} \\
&\quad= \frac{\Gamma_p((n+\nu)/2)}{\Gamma_p(n/2)\Gamma_p(\nu/2)} |\mathbf{\Omega}|^{\frac{\nu}{2}} |\mathbf{R}|^{\frac{n-p-1}{2}} |\mathbf{\Omega} + \mathbf{R}|^{-\frac{n+\nu}{2}} \\
&\quad= \frac{\Gamma_p((n+\nu)/2)}{\Gamma_p(n/2)\Gamma_p(\nu/2)} |\mathbf{\Omega}|^{\frac{\nu}{2}} |\mathbf{R}|^{-\frac{\nu+p+1}{2}} |\mathbf{\Omega}\mathbf{R}^{-1} + \mathbf{I}|^{-\frac{n+\nu}{2}} \\
&\quad= \frac{\Gamma_p((n+\nu)/2)}{\Gamma_p(n/2)\Gamma_p(\nu/2)} |\mathbf{\Omega}|^{\frac{\nu}{2}} |\mathbf{R}|^{-\frac{\nu+p+1}{2}} |(\mathbf{I} + \mathbf{C}_\Omega^\top \mathbf{R}^{-1}\mathbf{C}_\Omega)^{-1}|^{\frac{\nu+n}{2}} \\
&\quad= p_{iF\mathcal{R}}(\mathbf{R}|\mathbf{\Omega}, (n,\dots,n), (\nu,\dots,\nu)) \\
&\quad= p_F(\mathbf{R}|\mathbf{\Omega}, n, \nu).
\end{aligned}$$

This proves that both its alternative stochastic representations given in Table 2.1 yield the F distribution. Note that $\mathbf{R} \sim F\mathcal{R}(\mathbf{\Omega}^{-1}, \mathbf{n}, \boldsymbol{\nu}) \nRightarrow \mathbf{R}^{-1} \sim F\mathcal{R}(\mathbf{\Omega}, \boldsymbol{\nu}, \mathbf{n})$ for either type.[33] This is in contrast to the F distribution. Also, note that the standardized F-Riesz distribution cannot be obtained by mixing a standardized Riesz with a standardized Inverse Riesz but only by mixing the non-standardized versions and then standardizing the resulting distribution, as done above. This is also in contrast to the F distribution and can be seen since

$$\begin{aligned}
&\mathbf{C}\mathrm{dg}(\mathring{\mathbf{n}})^{-\frac{1}{2}} \bar{\mathbf{B}}^{-\top} \underset{\sim}{\mathbf{B}} \underset{\sim}{\mathbf{B}}^\top \bar{\mathbf{B}}^{-1} \mathrm{dg}(\mathring{\mathbf{n}})^{-\frac{1}{2}} \mathbf{C}^\top \\
&\qquad\qquad \neq \mathbf{C}\mathrm{dg}(\mathring{\boldsymbol{\nu}})^{-\frac{1}{2}} \bar{\mathbf{B}}^{-\top} \mathrm{dg}(\mathbf{n})^{-\frac{1}{2}} \underset{\sim}{\mathbf{B}} \underset{\sim}{\mathbf{B}}^\top \mathrm{dg}(\mathbf{n})^{-\frac{1}{2}} \bar{\mathbf{B}}^{-1} \mathrm{dg}(\mathring{\boldsymbol{\nu}})^{-\frac{1}{2}} \mathbf{C}^\top.
\end{aligned}$$

Next, in Table 2.11 we list the p.d.f.s of the standardized distributions $p_{\mathcal{D}}(\mathbf{R}|\mathbf{\Sigma}, \boldsymbol{\theta}_{\mathcal{D}})$. They can be derived by replacing in the non-standardized p.d.f.s $\mathbf{\Omega} = \mathbf{C}\mathbf{M}_{\mathcal{D}}^{-1}\mathbf{C}$, where $\mathbf{C}$ is the lower Cholesky factor of $\mathbf{\Sigma} = \mathbf{C}\mathbf{C}^\top$ and using Lemma 2.2.2 (iv). They are interesting because (1) in the GAS models we use $\frac{\partial p_{\mathcal{D}}(\mathbf{R}|\mathbf{\Sigma}, \boldsymbol{\theta}_{\mathcal{D}})}{\partial \mathrm{vech}(\mathbf{\Sigma})^\top}$, not $\frac{\partial p_{\mathcal{D}}(\mathbf{R}|\mathbf{\Omega}, \boldsymbol{\theta}_{\mathcal{D}})}{\partial \mathrm{vech}(\mathbf{\Omega})^\top}$, (2) the presence of $|\mathbf{R}|^{-(p+1)/2}$ in all standardized p.d.f.s makes clear that the likelihood value and likelihood-based information criteria can be manipulated, given an estimated expected value $\widehat{\mathbf{\Sigma}}$, by choosing a different scaling of the RCs, and (3) programming their explicit functional form gives computational efficiency gains as opposed to simply inputting $\mathbf{\Omega} = \mathbf{C}\mathbf{M}_{\mathcal{D}}^{-1}\mathbf{C}$ in the non-standardized p.d.f. programs.

33. See the derivation of the Inverse F-Riesz type II. The derivation of the Inverse F-Riesz type I is very similar.

Distribution	Probability Density Function $p_{\mathcal{D}}(\mathbf{R}\|\mathbf{\Sigma},\boldsymbol{\theta}_{\mathcal{D}})$
Wishart	$\frac{n^{np/2}}{2^{np/2}}\frac{1}{\Gamma_p(n/2)}\ \|\mathbf{R}\|^{-\frac{p+1}{2}}\ \|\mathbf{Z}\|^{\frac{n}{2}}\ \operatorname{etr}(-\frac{1}{2}n\mathbf{Z})$
Riesz	$\frac{\prod_{i=1}^p n_i^{n_i/2}}{2^{p\bar{\mathbf{n}}/2}}\frac{1}{\Gamma_p(\mathbf{n}/2)}\ \|\mathbf{R}\|^{-\frac{p+1}{2}}\ \|\mathbf{Z}\|_{\frac{\mathbf{n}}{2}}\ \operatorname{etr}(-\frac{1}{2}\operatorname{dg}(\mathbf{n})\mathbf{Z})$
Inv.Wishart	$\frac{(\nu-p-1)^{\nu p/2}}{2^{\nu p/2}}\frac{1}{\Gamma_p(\nu/2)}\ \|\mathbf{R}\|^{-\frac{p+1}{2}}\ \|\mathbf{Z}\|^{-\frac{\nu}{2}}\ \operatorname{etr}(-\frac{1}{2}(\nu-p-1)\mathbf{Z}^{-1})$
Inv.Riesz	$\frac{\prod_{i=1}^p \mathring{\nu}_i^{-\nu_i/2}}{2^{p\bar{\boldsymbol{\nu}}/2}}\frac{1}{\Gamma_p(\overleftarrow{\boldsymbol{\nu}}/2)}\ \|\mathbf{R}\|^{-\frac{p+1}{2}}\ \|\mathbf{Z}\|_{-\frac{\boldsymbol{\nu}}{2}}\ \operatorname{etr}(-\frac{1}{2}\operatorname{dg}(\mathring{\boldsymbol{\nu}})^{-1}\mathbf{Z}^{-1})$
t-Wishart	$(\frac{n}{\nu-2})^{pn/2}\frac{\Gamma((\nu+pn)/2)}{\Gamma_p(n/2)\Gamma(\nu/2)}\ \|\mathbf{R}\|^{-\frac{p+1}{2}}\ \|\mathbf{Z}\|^{\frac{n}{2}}\ (1+\frac{n}{\nu-2}\operatorname{tr}(\mathbf{Z}))^{-\frac{\nu+pn}{2}}$
t-Riesz	$\frac{\prod_{i=1}^p n_i^{n_i/2}}{(\nu-2)^{p\bar{\mathbf{n}}/2}}\frac{\Gamma((\nu+p\bar{\mathbf{n}})/2)}{\Gamma_p(\mathbf{n}/2)\Gamma(\nu/2)}\ \|\mathbf{R}\|^{-\frac{p+1}{2}}\ \|\mathbf{Z}\|_{\frac{\mathbf{n}}{2}}\ (1+\frac{1}{\nu-2}\operatorname{tr}(\operatorname{dg}(\mathbf{n})\mathbf{Z}))^{-\frac{\nu+p\bar{\mathbf{n}}}{2}}$
Inv.t-Wishart	$(\frac{\nu-p-1}{n})^{\frac{\nu p}{2}}\frac{\Gamma((n+p\nu)/2)}{\Gamma(n/2)\Gamma_p(\nu/2)}\ \|\mathbf{R}\|^{-\frac{p+1}{2}}\ \|\mathbf{Z}\|^{-\frac{\nu}{2}}\ (1+\frac{\nu-p-1}{n}\operatorname{tr}(\mathbf{Z}^{-1}))^{-\frac{n+p\nu}{2}}$
Inv.t-Riesz	$\frac{\prod_{i=1}^p \mathring{\nu}_i^{-\nu_i/2}}{n^{p\bar{\boldsymbol{\nu}}/2}}\frac{\Gamma((n+p\bar{\boldsymbol{\nu}})/2)}{\Gamma(n/2)\Gamma_p(\overleftarrow{\boldsymbol{\nu}}/2)}\ \|\mathbf{R}\|^{-\frac{p+1}{2}}\ \|\mathbf{Z}\|_{-\frac{\boldsymbol{\nu}}{2}}\ (1+\frac{1}{n}\operatorname{tr}(\operatorname{dg}(\mathring{\boldsymbol{\nu}})^{-1}\mathbf{Z}^{-1}))^{-\frac{n+p\bar{\boldsymbol{\nu}}}{2}}$
F	$(\frac{n}{\nu-p-1})^{\frac{np}{2}}\frac{\Gamma_p((n+\nu)/2)}{\Gamma_p(n/2)\Gamma_p(\nu/2)}\ \|\mathbf{R}\|^{-\frac{p+1}{2}}\ \|\mathbf{Z}\|^{\frac{n}{2}}\ \|\mathbf{I}+\frac{n}{\nu-p-1}\mathbf{Z}\|^{-\frac{\nu+n}{2}}$
F-Riesz	$\prod_{i=1}^p \mathring{n}_i^{\frac{n_i}{2}}\frac{\Gamma_p((\overleftarrow{\mathbf{n}}+\overleftarrow{\boldsymbol{\nu}})/2)}{\Gamma_p(\mathbf{n}/2)\Gamma_p(\overleftarrow{\boldsymbol{\nu}}/2)}\ \|\mathbf{R}\|^{-\frac{p+1}{2}}\ \|\mathbf{Z}\|_{\frac{\mathbf{n}}{2}}\ \|\mathbf{I}+\operatorname{dg}(\mathring{\mathbf{n}})^{\frac{1}{2}}\mathbf{Z}\operatorname{dg}(\mathring{\mathbf{n}})^{\frac{1}{2}}\|_{-\frac{\mathbf{n}+\boldsymbol{\nu}}{2}}$
Inv.F-Riesz	$\prod_{i=1}^p \mathring{\mathring{n}}_i^{-\frac{\nu_i}{2}}\frac{\Gamma_p((\mathbf{n}+\boldsymbol{\nu})/2)}{\Gamma_p(\mathbf{n}/2)\Gamma_p(\overleftarrow{\boldsymbol{\nu}}/2)}\ \|\mathbf{R}\|^{-\frac{p+1}{2}}\ \|\mathbf{Z}\|_{-\frac{\boldsymbol{\nu}}{2}}\ \|(\mathbf{I}+\operatorname{dg}(\mathring{\mathbf{n}})^{-\frac{1}{2}}\mathbf{Z}^{-1}\operatorname{dg}(\mathring{\mathbf{n}})^{-\frac{1}{2}})^{-1}\|_{\frac{\mathbf{n}+\boldsymbol{\nu}}{2}}$

Table 2.11: Standardized probability density functions. We define $\mathbf{Z}=\mathbf{C}^{-1}\mathbf{R}\mathbf{C}^{-\top}$, where $\mathbf{C}$ is the lower Cholesky factor of $\mathbf{\Sigma}$. For the definition of $\mathring{\boldsymbol{\nu}}$, $\mathring{\mathbf{n}}$ and $\mathring{\mathring{\mathbf{n}}}$ Theorem 2.2.1. To derive these representations from the ones in Table 2.5 use Lemma 2.2.2 (iv).

Empirical Results

Assets:		Rnd	Mngn	Rnd	Fin	Rnd	Manf
#Assets:		5	6	10	15	25	25
Wishart	n	7.0	8.3	12.8	15.0	27.7	29.4
Riesz	$\mathbf{n}_{min}$	1.4	1.9	1.2	0.8	1.1	1.5
	$\bar{\mathbf{n}}$	6.9	7.9	12.8	12.8	25.0	25.9
	$\mathbf{n}_{max}$	12.4	16.0	24.4	27.8	49.5	52.5
t-Wishart	n	17.1	17.5	22.7	28.4	38.2	42.3
t-Riesz	$\mathbf{n}_{min}$	3.9	4.4	2.2	2.2	2.1	4.4
	$\bar{\mathbf{n}}$	17.5	16.8	23.1	27.5	35.5	37.4
	$\mathbf{n}_{max}$	30.8	31.0	43.3	51.3	63.9	68.2
Inv.t-Wishart	n	3.1	4.1	3.6	1.7	3.8	3.5
Inv.t-Riesz	n	3.3	4.1	3.8	1.8	3.9	3.7
F	n	30.1	45.8	40.6	53.1	78.7	85.0
F-Riesz	$\mathbf{n}_{min}$	3.1	4.6	2.7	1.7	5.2	2.3
	$\bar{\mathbf{n}}$	43.3	41.9	57.5	63.6	74.5	75.0
	$\mathbf{n}_{max}$	110.4	108.8	168.0	191.2	180.5	190.4
Inv.F-Riesz	$\mathbf{n}_{min}$	1.7	2.8	1.5	1.0	3.2	2.6
	$\bar{\mathbf{n}}$	1297.5	1098.2	1658.0	981.2	697.4	700.0
	$\mathbf{n}_{max}$	6429.7	6344.4	8213.8	5285.6	7480.9	7440.0
Inv.Wishart	ν	9.2	11.2	16.2	19.3	34.6	36.3
Inv.Riesz	$\boldsymbol{\nu}_{min}$	4.0	4.2	4.2	4.1	4.2	4.3
	$\bar{\boldsymbol{\nu}}$	8.5	11.0	15.7	19.3	31.7	32.4
	$\boldsymbol{\nu}_{max}$	11.9	16.8	24.1	31.0	48.2	46.6
t-Wishart	ν	4.1	4.8	4.9	3.0	5.2	5.3
t-Riesz	$\boldsymbol{\nu}_{min}$	4.3	4.5	5.4	3.1	5.7	5.5
Inv.t-Wishart	ν	17.3	18.4	22.6	29.4	41.2	47.7
Inv.t-Riesz	$\boldsymbol{\nu}_{min}$	8.0	5.3	3.9	6.2	4.4	6.7
	$\bar{\boldsymbol{\nu}}$	18.7	19.6	24.3	32.8	42.3	45.5
	$\boldsymbol{\nu}_{max}$	27.9	30.9	36.4	50.5	70.3	67.9
F	ν	10.9	12.9	20.7	21.5	43.1	45.9
F-Riesz	$\boldsymbol{\nu}_{min}$	4.0	4.2	5.0	3.7	3.6	3.8
	$\bar{\boldsymbol{\nu}}$	7.9	9.8	14.1	15.1	31.8	31.8
	$\boldsymbol{\nu}_{max}$	10.7	13.7	18.6	25.2	51.4	53.8
Inv.F-Riesz	$\boldsymbol{\nu}_{min}$	4.1	3.6	4.6	4.2	3.6	3.9
	$\bar{\boldsymbol{\nu}}$	22.4	25.2	34.6	33.2	47.5	50.4
	$\boldsymbol{\nu}_{max}$	49.4	62.3	83.9	77.4	103.7	111.3

Table 2.12: Estimated degree of freedom parameters of static distributions for the different datasets.

Assets:		Rnd	Mngn	Rnd	Fin	Rnd	Manf
#Assets:		5	6	10	15	25	25
Wishart	n	17.6	20.9	26.3	33.2	45.2	45.1
Riesz	$\mathbf{n}_{min}$	6.3	7.6	6.0	5.4	5.0	5.0
	$\bar{\mathbf{n}}$	16.6	19.6	24.0	29.7	39.5	39.5
	$\mathbf{n}_{max}$	26.4	33.2	41.0	51.1	66.9	66.9
t-Wishart	n	24.8	30.4	33.4	42.8	50.7	50.7
t-Riesz	$\mathbf{n}_{min}$	8.3	10.3	7.9	7.1	6.3	6.2
	$\bar{\mathbf{n}}$	23.1	27.9	30.1	38.5	44.3	44.3
	$\mathbf{n}_{max}$	35.2	41.1	49.0	64.5	71.7	73.2
Inv.t-Wishart	n	24.2	24.3	23.9	16.5	24.2	21.8
Inv.t-Riesz	n	23.1	23.9	22.1	14.4	22.5	20.9
F	n	81.6	109.4	122.1	134.2	171.3	179.4
F-Riesz	$\mathbf{n}_{min}$	17.2	18.5	14.5	12.8	16.2	12.9
	$\bar{\mathbf{n}}$	71.1	84.7	93.9	100.5	117.3	124.4
	$\mathbf{n}_{max}$	142.4	169.3	203.2	210.3	225.8	264.6
Inv.F-Riesz	$\mathbf{n}_{min}$	10.2	12.4	8.5	7.8	6.8	6.7
	$\bar{\mathbf{n}}$	3078.3	2237.5	1005.8	1604.8	1965.0	1187.7
	$\mathbf{n}_{max}$	15248.2	13161.6	9356.7	12236.0	21896.1	10330.9
Inv.Wishart	ν	20.9	25.4	31.9	39.9	54.1	54.4
Inv.Riesz	$\boldsymbol{\nu}_{min}$	9.9	11.1	10.1	8.7	9.2	9.7
	$\bar{\boldsymbol{\nu}}$	19.6	23.8	29.0	35.8	48.1	48.9
	$\boldsymbol{\nu}_{max}$	25.2	32.6	38.7	46.3	62.7	64.7
t-Wishart	ν	19.8	20.3	20.6	15.3	23.7	20.4
t-Riesz	ν	19.8	18.8	20.6	13.7	23.7	19.9
Inv.t-Wishart	ν	26.8	33.0	37.9	48.3	58.7	59.6
Inv.t-Riesz	$\boldsymbol{\nu}_{min}$	11.0	13.5	11.2	8.7	10.5	11.7
	$\bar{\boldsymbol{\nu}}$	25.5	31.3	35.5	45.7	54.1	55.2
	$\boldsymbol{\nu}_{max}$	34.1	42.7	46.9	61.6	70.6	74.0
F	$\boldsymbol{\nu}_{min}$	27.3	32.3	41.1	53.4	70.8	69.7
F-Riesz	$\boldsymbol{\nu}_{min}$	9.3	10.4	9.5	10.6	10.7	10.5
	$\bar{\boldsymbol{\nu}}$	19.2	22.6	27.4	35.7	50.0	49.7
	$\boldsymbol{\nu}_{max}$	26.0	28.6	34.7	53.3	73.2	72.7
Inv.F-Riesz	$\boldsymbol{\nu}_{min}$	9.0	10.1	9.5	10.1	9.5	9.5
	$\bar{\boldsymbol{\nu}}$	33.5	39.3	45.1	51.3	63.4	63.5
	$\boldsymbol{\nu}_{max}$	56.8	82.1	79.0	83.7	102.3	102.1

Table 2.13: Estimated degree of freedom parameters of dynamic mean shifting distributions for the different datasets.

Assets:	Random	Mining	Random	Finance	Random	Manuf.
#Assets:	5	6	10	15	25	25
Wishart	485	600	1121	5104	4829	3769
Riesz	449	552	984	4305	4066	2986
Inv.Wishart	420	497	775	2526	3052	2201
Inv.Riesz	407	475	730	2269	2630	1891
t-Wishart	344	364	548	1697	2423	1723
t-Riesz	330	347	510	1491	2104	1344
Inv.t-Wishart	335	346	468	1538	1809	1018
Inv.t-Riesz	325	335	445	1357	1610	848
F	410	484	738	2341	2656	1979
F-Riesz	346	389	494	1765	1833	1164
Inv.F-Riesz	355	400	533	1847	1885	1237

Table 2.14: Average of log-score loss over a one-month forecasting period (22 trading days), $-\sum_{j=1}^{22} p_{\mathcal{D}}(\mathbf{R}_{t+j}|\widehat{\boldsymbol{\Sigma}}_{j+1}, \hat{\boldsymbol{\theta}}_{\mathcal{D},j+1})$, for the forecasting window from 1 January 2007 to 31 December 2011, where each model is re-estimated every ten trading days. 90% model confidence sets in red.

Assets:	Random	Mining	Random	Finance	Random	Manuf.
#Assets:	5	6	10	15	25	25
Wishart	117	357	-87	-1467	-1502	-1940
Riesz	89	324	-197	-1695	-2160	-2611
Inv.Wishart	76	290	-295	-1786	-2823	-3534
Inv.Riesz	67	270	-349	-1948	-3075	-3792
t-Wishart	15	227	-393	-2302	-2548	-3037
t-Riesz	0	203	-468	-2483	-2994	-3561
Inv.t-Wishart	-14	192	-528	-2586	-3584	-4297
Inv.t-Riesz	-21	180	-571	-2667	-3729	-4490
F	57	280	-340	-2026	-3058	-3655
F-Riesz	-5	189	-554	-2455	-3764	-4428
Inv.F-Riesz	-1	198	-537	-2414	-3711	-4343

Table 2.15: Average of log-score loss over a one month forecasting period (22 trading days), $-\sum_{j=1}^{22} p_{\mathcal{D}}(\mathbf{R}_{t+j}|\widehat{\boldsymbol{\Sigma}}_{j+1}, \hat{\boldsymbol{\theta}}_{\mathcal{D},j+1})$, for the forecasting window from 1 January 2012 to 31 December 2019, where each model is re-estimated every ten trading days. 90% model confidence sets in red.

Riesz Covariance

For the derivation of the Fisher information matrix of the t-Riesz distribution in Chapter 3 we need the covariance matrix of $\text{vech}(\mathbf{R})$, if $\mathbf{R} \sim \mathcal{R}(\mathbf{\Sigma}, \mathbf{n})$.

Lemma 2.7.3. *Let* $\mathbf{R}$ *follow a Riesz distribution,* $\mathbf{R} \sim \mathcal{R}(\mathbf{\Sigma}, \mathbf{n})$ *and let* $\boldsymbol{\mathcal{K}}_{\mathcal{R}}$ *be its stochastic representation kernel as defined in Table 2.1. Then*

$$\begin{aligned}
\mathbb{E}[\text{vech}^2(\boldsymbol{\mathcal{K}}_{\mathcal{R}})] &= \text{vech}^2(\text{dg}(\mathbf{n})) + 2\mathbf{G}^+(\text{dg}(\mathbf{n}) \otimes \mathbf{I})(\mathbf{G}^+)^\top, \\
\mathbf{Cov}\left(\text{vech}\left(\boldsymbol{\mathcal{K}}_{\mathcal{R}}\right)\right) &= 2\mathbf{G}^+\left(\text{dg}(\mathbf{n}) \otimes \mathbf{I}\right)\left(\mathbf{G}^+\right)^\top, \\
\mathbb{E}[\text{vech}^2(\mathbf{R})] &= \text{vech}^2(\mathbf{\Sigma}) + 2\mathbf{G}^+(\mathbf{\Sigma} \otimes \mathbf{C}\text{dg}(\mathbf{n})^{-1}\mathbf{C}^\top)(\mathbf{G}^+)^\top, \text{ and} \\
\mathbf{Cov}\left(\text{vech}(\mathbf{R})\right) &= 2\mathbf{G}^+\left(\mathbf{\Sigma} \otimes \mathbf{C}\text{dg}(\mathbf{n})^{-1}\mathbf{C}^\top\right)\left(\mathbf{G}^+\right)^\top.
\end{aligned}$$

Proof. First, see §2.1.2 of Kollo and von Rosen (2005) for the characteristic function of a patterned (in our case symmetric) matrix-variate distribution. Note there are two approaches here. Either we ignore symmetry and get the characteristic function of $\text{vec}(\mathbf{R})$, or we take it into account by getting the characteristic function of e.g. $\text{vech}\left(\mathbf{R}\right)$. In consistency with the rest of this paper, we take symmetry into account. Díaz-García (2013) and Gribisch and Hartkopf (2022) don't. Gupta and Nagar (2000) and Kollo and von Rosen (2005) do. The characteristic function of $\text{vech}\left(\boldsymbol{\mathcal{K}}_{\mathcal{R}}\right)$ where, as defined in Table 2.1, $\boldsymbol{\mathcal{K}}_{\mathcal{R}} = \underline{\mathbf{B}}\underline{\mathbf{B}}^\top$ is given by

$$\begin{aligned}
\phi(\mathbf{Z}) &= \mathbb{E}\left[e^{i\,\text{vech}(\mathbf{Z})^\top \text{vech}(\boldsymbol{\mathcal{K}}_{\mathcal{R}})}\right] \\
&= \mathbb{E}\left[\text{etr}\left(i\frac{1}{2}\left(\mathbf{Z}+\mathbf{Z}\right)\boldsymbol{\mathcal{K}}_{\mathcal{R}}\right)\right] \text{ p. 244 Kollo and von Rosen (2005)} \\
&= \frac{1}{2^{p\bar{\mathbf{n}}/2}\Gamma_p\left(\mathbf{n}/2\right)}\int_{\boldsymbol{\mathcal{K}}_{\mathcal{R}}>\mathbf{0}} |\boldsymbol{\mathcal{K}}_{\mathcal{R}}|_{\frac{\mathbf{n}-p-1}{2}}\,\text{etr}\left(i\frac{1}{2}\left(\mathbf{Z}+\mathbf{Z}\right)\boldsymbol{\mathcal{K}}_{\mathcal{R}}\right)\text{etr}\left(-\frac{1}{2}\mathbf{I}\boldsymbol{\mathcal{K}}_{\mathcal{R}}\right) \\
&= \frac{1}{2^{p\bar{\mathbf{n}}/2}\Gamma_p\left(\mathbf{n}/2\right)}\int_{\boldsymbol{\mathcal{K}}_{\mathcal{R}}>\mathbf{0}} |\boldsymbol{\mathcal{K}}_{\mathcal{R}}|_{\frac{\mathbf{n}-p-1}{2}}\,\text{etr}\left(-\frac{1}{2}\left(\mathbf{I}-i\left(\mathbf{Z}+\mathbf{Z}\right)\right)\boldsymbol{\mathcal{K}}_{\mathcal{R}}\right) \\
&= \frac{1}{2^{p\bar{\mathbf{n}}/2}\Gamma_p\left(\mathbf{n}/2\right)}\Gamma_p\left(\mathbf{n}/2\right)\left|\frac{1}{2}\left(\mathbf{I}-i\left(\mathbf{Z}+\mathbf{Z}\right)\right)^{-1}\right|_{\frac{\mathbf{n}}{2}} \\
&= \frac{1}{2^{p\bar{\mathbf{n}}/2}}2^{p\bar{\mathbf{n}}/2}\left|\left(\mathbf{I}-i\left(\mathbf{Z}+\mathbf{Z}\right)\right)^{-1}\right|_{\frac{\mathbf{n}}{2}} \\
&= \left|\left(\mathbf{I}-i\left(\mathbf{Z}+\mathbf{Z}\right)\right)^{-1}\right|_{\frac{\mathbf{n}}{2}},
\end{aligned}$$

where $\mathbf{Z}$ is a diagonal matrix with elements $\text{dg}(\mathbf{Z})$ and where we used Lemma (2.7.1). See also Díaz-García (2013), Lemma 1.

Denote $\boldsymbol{\Xi} = \mathbf{I} - i\,(\mathbf{Z} + \mathbf{Z})$, then

$$\frac{\partial \text{vech}(\boldsymbol{\Xi})}{\partial \text{vech}(\mathbf{Z})^\top} = -i2(\mathbf{G}^\top\mathbf{G})^{-1} = -i2(\mathbf{G}^\top\mathbf{G})^{-\top}$$

Then

$$\begin{aligned}
&\frac{\partial^2\phi(\mathbf{Z})}{\partial \text{vech}(\mathbf{Z})\partial \text{vech}(\mathbf{Z})^\top} = i\frac{\partial}{\partial \text{vech}(\mathbf{Z})}|\boldsymbol{\Xi}^{-1}|_{\frac{\mathfrak{n}}{2}}\text{vech}(\dot{\mathbf{C}}_\xi\,\text{dg}(\mathbf{n})\dot{\mathbf{C}}_\xi{}^\top)^\top \\
&= i\frac{\partial|\boldsymbol{\Xi}^{-1}|_{\frac{\mathfrak{n}}{2}}}{\partial \text{vech}(\mathbf{Z})}\text{vech}(\dot{\mathbf{C}}_\xi\,\text{dg}(\mathbf{n})\dot{\mathbf{C}}_\xi{}^\top)^\top + i|\boldsymbol{\Xi}^{-1}|_{\frac{\mathfrak{n}}{2}}\frac{\partial \text{vec}(\dot{\mathbf{C}}_\xi\,\text{dg}(\mathbf{n})\dot{\mathbf{C}}_\xi{}^\top)}{\partial \text{vech}(\mathbf{Z})}(\mathbf{G}^+)^\top \\
&= i^2|\boldsymbol{\Xi}^{-1}|_{\frac{\mathfrak{n}}{2}}\text{vech}^2(\dot{\mathbf{C}}_\xi\text{dg}(\mathbf{n})\dot{\mathbf{C}}_\xi{}^\top) \\
&\quad - i|\boldsymbol{\Xi}^{-1}|_{\frac{\mathfrak{n}}{2}}\left(\frac{\partial \text{vec}(\dot{\mathbf{C}}_\xi\,\text{dg}(\mathbf{n})\dot{\mathbf{C}}_\xi{}^\top)}{\partial \text{vech}(\boldsymbol{\Xi})^\top}\frac{\partial \text{vech}(\boldsymbol{\Xi})}{\partial \text{vech}(\mathbf{Z})^\top}\right)^\top(\mathbf{G}^+)^\top \\
&= i^2|\boldsymbol{\Xi}^{-1}|_{\frac{\mathfrak{n}}{2}}\text{vech}^2(\dot{\mathbf{C}}_\xi\text{dg}(\mathbf{n})\dot{\mathbf{C}}_\xi{}^\top) + i^2 2|\boldsymbol{\Xi}^{-1}|_{\frac{\mathfrak{n}}{2}} \\
&\quad \times(\mathbf{G}\mathbf{G}^+(\dot{\mathbf{C}}_\xi\,\text{dg}(\mathbf{n})\otimes\mathbf{I})\mathbf{F}^\top(\mathbf{G}^+(\dot{\mathbf{C}}_\xi{}^{-\top}\otimes\boldsymbol{\Xi})\mathbf{F}^\top)^{-1}(\mathbf{G}^\top\mathbf{G})^{-\top})^\top(\mathbf{G}^+)^\top \\
&= i^2|\boldsymbol{\Xi}^{-1}|_{\frac{\mathfrak{n}}{2}}\text{vech}^2(\dot{\mathbf{C}}_\xi\text{dg}(\mathbf{n})\dot{\mathbf{C}}_\xi{}^\top) + i^2 2|\boldsymbol{\Xi}^{-1}|_{\frac{\mathfrak{n}}{2}} \\
&\quad \times(\mathbf{G}^\top\mathbf{G})^{-1}(\mathbf{G}\mathbf{G}^+(\dot{\mathbf{C}}_\xi\,\text{dg}(\mathbf{n})\otimes\mathbf{I})\mathbf{F}^\top(\mathbf{G}^+(\dot{\mathbf{C}}_\xi{}^{-\top}\otimes\boldsymbol{\Xi})\mathbf{F}^\top)^{-1})^\top(\mathbf{G}^+)^\top,
\end{aligned}$$

where $\dot{\mathbf{C}}_\xi$ is the lower Cholesky factor of $\boldsymbol{\Xi}^{-1}$, such that

$$\begin{aligned}
\mathbb{E}[\text{vech}^2(\boldsymbol{\mathcal{K}}_\mathcal{R})] &= \left.\frac{\partial^2\phi(\mathbf{Z})}{i^2\partial \text{vech}(\mathbf{Z})\partial \text{vech}(\mathbf{Z})^\top}\right|_{\mathbf{Z}=\mathbf{0}} \\
&= \text{vech}^2(\text{dg}(\mathbf{n})) + 2(\mathbf{G}^\top\mathbf{G})^{-1}(\mathbf{G}\mathbf{G}^+(\text{dg}(\mathbf{n})\otimes\mathbf{I})\mathbf{F}^\top(\mathbf{G}^+\mathbf{I}^{\otimes 2}\mathbf{F}^\top)^{-1})^\top(\mathbf{G}^+)^\top \\
&= \text{vech}^2(\text{dg}(\mathbf{n})) + 2\mathbf{G}^+((\text{dg}(\mathbf{n})\otimes\mathbf{I})\mathbf{G}\mathbf{G}^+\mathbf{F}^\top(\mathbf{G}^+\mathbf{F}^\top)^{-1})^\top(\mathbf{G}^+)^\top \\
&= \text{vech}^2(\text{dg}(\mathbf{n})) + 2\mathbf{G}^+(\text{dg}(\mathbf{n})\otimes\mathbf{I})(\mathbf{G}^+)^\top,
\end{aligned}$$

and

$$\begin{aligned}
\mathbb{E}[\text{vec}^2(\mathbf{R})] &= \mathbf{C}_\Omega^{\otimes 2}\mathbb{E}[\text{vec}^2(\boldsymbol{\mathcal{K}}_\mathcal{R})](\mathbf{C}_\Omega^{\otimes 2})^\top \\
&= \mathbf{C}_\Omega^{\otimes 2}\mathbf{G}\mathbb{E}[\text{vech}^2(\boldsymbol{\mathcal{K}}_\mathcal{R})]\mathbf{G}^\top(\mathbf{C}_\Omega^{\otimes 2})^\top \\
&= \mathbf{C}_\Omega^{\otimes 2}\text{vec}^2(\text{dg}(\mathbf{n}))(\mathbf{C}_\Omega^{\otimes 2})^\top + 2(\mathbf{C}_\Omega^{\otimes 2})\mathbf{G}\mathbf{G}^+(\text{dg}(\mathbf{n})\otimes\mathbf{I})(\mathbf{C}_\Omega^{\otimes 2}\mathbf{G}\mathbf{G}^+)^\top \\
&= \text{vec}^2(\mathbf{C}_\Omega\text{dg}(\mathbf{n})\mathbf{C}_\Omega^\top) + 2\mathbf{G}\mathbf{G}^+(\mathbf{C}_\Omega\text{dg}(\mathbf{n})\mathbf{C}_\Omega^\top\otimes\boldsymbol{\Omega})\mathbf{G}\mathbf{G}^+, \qquad (2.39)
\end{aligned}$$

and consequently

$$\begin{aligned}\mathbb{E}[\text{vech}^2(\mathbf{R})] &= \mathbf{G}^+\mathbb{E}[\text{vec}^2(\mathbf{R})](\mathbf{G}^+)^\top \\ &= \text{vech}^2(\mathbf{C}_\Omega \text{dg}(\mathbf{n})\mathbf{C}_\Omega^\top) + 2\mathbf{G}^+(\mathbf{C}_\Omega \text{dg}(\mathbf{n})\mathbf{C}_\Omega^\top \otimes \boldsymbol{\Omega})(\mathbf{G}^+)^\top \\ &= \text{vech}^2(\boldsymbol{\Sigma}) + 2\mathbf{G}^+(\boldsymbol{\Sigma} \otimes \mathbf{C}\text{dg}(\mathbf{n})^{-1}\mathbf{C}^\top)(\mathbf{G}^+)^\top.\end{aligned}$$

Thus

$$\mathbf{Cov}(\text{vech}(\boldsymbol{\mathcal{K}}_\mathcal{R})) = 2\mathbf{G}^+(\text{dg}(\mathbf{n}) \otimes \mathbf{I})(\mathbf{G}^+)^\top,$$

and

$$\begin{aligned}\mathbf{Cov}(\text{vech}(\mathbf{R})) &= \mathbf{Cov}(\text{vech}(\mathbf{C}_\Omega \boldsymbol{\mathcal{K}}_\mathcal{R} \mathbf{C}_\Omega^\top)) \\ &= \mathbf{G}^+\mathbf{C}_\Omega^{\otimes 2}\mathbf{G}\mathbf{Cov}((\text{vech}(\boldsymbol{\mathcal{K}}_\mathcal{R})))(\mathbf{G}^+\mathbf{C}_\Omega^{\otimes 2}\mathbf{G})^\top \\ &= 2\mathbf{G}^+\mathbf{C}_\Omega^{\otimes 2}\mathbf{G}\mathbf{G}^+(\text{dg}(\mathbf{n}) \otimes \mathbf{I})(\mathbf{G}^+)^\top(\mathbf{G}^+\mathbf{C}_\Omega^{\otimes 2}\mathbf{G})^\top \\ &= 2\mathbf{G}^+\mathbf{C}_\Omega^{\otimes 2}(\text{dg}(\mathbf{n}) \otimes \mathbf{I})(\mathbf{C}_\Omega^{\otimes 2})^\top(\mathbf{G}^+)^\top \\ &= 2\mathbf{G}^+(\boldsymbol{\Sigma} \otimes \mathbf{C}\text{dg}(\mathbf{n})^{-1}\mathbf{C}^\top)(\mathbf{G}^+)^\top.\end{aligned}$$

□

Chapter 3

Generalized Autoregressive Score Models for Realized Covariance Matrices

Introduction 3.1

The covariance matrix of financial asset returns is a crucial concept in financial econometrics, with direct implications for efficient portfolio allocation, risk management, and derivative pricing. A *realized covariance matrix* (RC) is an ex-post estimate of the daily integrated covariance matrix constructed from high-frequency data. It can be interpreted as providing an "effectively observable" daily covariance matrix of the underlying financial asset returns, and consequently, directly modeling the time-series of RCs has been advocated in the literature (e.g., Andersen et al., 2001, Andersen et al., 2006, McAleer and Medeiros, 2008, Chiriac and Voev, 2011).

One way to model time-series of RCs is the class of *observation-driven models* as defined by Cox (1981). In this class of models, the RC of a given day is assumed to follow a conditional probability distribution with time-varying parameters, where the parameters are updated using the previous RC realization(s). Probability distributions for RCs can be characterized by their $p \times p$ symmetric positive definite *expected value matrix* and a set of *degree of freedom* (d.o.f.) parameters (see Chapter 2). In the literature on observation-driven models for RCs, it is assumed that only the expected value matrix is time-varying, while the d.o.f. parameters are constant (see e.g. Golosnoy, Gribisch, and Liesenfeld, 2012, Noureldin, Shephard, and Sheppard, 2012 and Opschoor et al., 2018).

To make the models applicable to reasonable data dimension the updating process typically has to be restricted to scalar dynamics. That is, next day's expected value matrix is determined by the previous RC(s) multiplied by scalar parameters, which implies that the conditional expected value of any element in the RC on a

given day depends only on the realizations of this element in the previous RCs. This represents a very strong assumption, as there can be no spillovers between the realized (co)variances.

A very general subclass of observation-driven models that avoids this assumption is given by the *generalized autoregressive score* (GAS) models, introduced by Creal, Koopman, and Lucas (2011), Creal, Koopman, and Lucas (2013) and Harvey (2013). In GAS models, information about the entire shape of the assumed conditional probability distribution is incorporated directly into the parameter updating process, by using the distribution-specific *score*, often scaled by the corresponding inverse of the *Fisher information matrix* (FIM), as the forcing variable.[1] Advantages of GAS models over traditional observation-driven models are as follows. First, GAS models feature rich dynamic updating dynamics even if scalar dynamics are assumed. Second, they offer an intuitive way to make *any* parameter of a probability distribution time-varying, not just those that have an obvious observable forcing variable like the expected value matrix of distributions for RCs. Furthermore, GAS models have been shown to possess desirable information-theoretic optimality (see Blasques, Koopman, and Lucas, 2015). Finally, it is noteworthy that GAS models contain some traditional observation-driven models as special cases. One of these special cases is discovered in this paper.[2]

While the general literature on GAS models is very extensive (see Harvey, 2022 and www.gasmodel.com), and for time-series of RCs, the literature on traditional observation-driven is quite large (see references above and, among others, Asai and So, 2013, Jin and Maheu, 2016 and Zhou et al., 2019), the literature on GAS models for time-series of RCs is very scarce. To the best of our knowledge, there are only two papers. The first is by Opschoor et al. (2018), who propose to model the daily return vectors and the RCs jointly. They assume that the daily return vectors follow a multivariate standardized t-distribution with a covariance matrix that is equal to the expected value matrix of the standardized matrix-F distribution, assumed for the RCs. This parameter matrix follows GAS dynamics. The authors show that their model significantly outperforms all previously proposed competitors in terms of in-sample fit and out-of-sample forecasting ability. The second paper on GAS models for RCs is by Gorgi et al. (2019). They propose a model that is similar to the one in Opschoor et al. (2018). However, they assume a multivariate standardized normal (nested by the t) and Wishart (nested by the matrix-F) distribution for the daily returns and RCs, respectively. They generalize the GAS dynamics of Opschoor et al. (2018) by assuming that the covariance matrix of the normal distribution is equal to the *scaled* expected value matrix of

1. We call the variable that drives the updating process the "forcing variable".
2. See Lemma 3.3.1.

the Wishart, and by assuming full GAS rather than scalar-GAS dynamics. While the full GAS dynamics make their model more flexible, it also makes it infeasible even for moderate cross-sectional dimensions.

The scarcity of literature on GAS models for RCs is despite there having been proposed many different probability distributions for RCs (see Chapter 2). Each distribution would entail a different parameter updating forcing variable in a GAS model since the score is distribution-specific. This paper aims to extend the literature by deriving GAS models for RCs for all probability distributions used in the literature. This extension is important because, in GAS models the natural question arises which distributions and their scores work best in practice. We contribute to the literature by deriving a general (across all distributions) representation of the scaled score w.r.t. the expected value matrix that makes the GAS models computationally feasible for medium to large RCs (say, five to 50 assets).

Furthermore, we provide empirical evidence consistent with economic intuition, which suggests that the d.o.f. parameters should also be modeled as time-varying. Consequently, another contribution we make is to assume time-varying d.o.f. parameters by endowing them with GAS dynamics. For realized variances, i.e. one-dimensional RCs, time-varying d.o.f. parameters with GAS dynamics have been proposed by Opschoor and Lucas (2022), who consider the special case of the F distribution. Using various loss functions, they show that their model outperforms competitors without time-varying d.o.f. parameters in an out-of-sample forecasting exercise. For RCs there is, to the best of our knowledge, no study to date considering time-varying d.o.f. parameters.

We contribute by deriving all the necessary theoretical inputs for GAS models with time-varying expected value matrix and time-varying d.o.f. parameters. That is, we derive the scores and the FIMs w.r.t. the expected value matrix and the d.o.f. parameters for all probability distributions for RCs.[3]. To the best of our knowledge, only a very small subset of these quantities (e.g. the score and FIM w.r.t. the expected value matrix of the Wishart and the score w.r.t. the expected value matrix of the matrix-F) have been derived before.

Finally, in the empirical section, likelihood ratio tests and forecast comparisons reveal that time-varying d.o.f. parameters are indeed important for all distributions. Furthermore, we show that in our model as in Chapter 2, which features much simpler parameter updating dynamics, the t-Riesz distribution family and the F-Riesz distribution family are the best distributions in terms of fit and forecasting performance.

The rest of this paper is structured as follows. The next section examines the drawbacks of non-GAS observation-driven models for RCs and explains how the

3. We do not obtain the FIMs w.r.t. the expected value matrix for the Inverse t-Riesz and the (Inverse) F-Riesz distributions.

GAS models alleviate them. Section 3.3 introduces the GAS model framework for time-series of RCs in general and presents, in particular, (the derivations for) our GAS models for all distributions for RCs. Section 3.4 contains the empirical application with in-sample fit and out-of-sample forecasting comparison that investigates among others the relevance of time-varying d.o.f. parameters across all distributions. Section 3.5 concludes.

3.2 Drawbacks of Traditional Observation-Driven Models for RCs

To be more precise on the traditional observation-driven model setup, that is a model without GAS dynamics, remember that all probability distributions $\mathcal{D}$ for RCs can be characterized by a symmetric positive definite expected value matrix $\mathbf{\Sigma}$ and their distribution-specific d.o.f. parameters, which we collect in $\boldsymbol{\theta}_{\mathcal{D}}$. Thus, the distributional assumption can be written as,

$$\mathbf{R}_t|\mathcal{F}_{t-1} \sim \mathcal{D}(\mathbf{\Sigma}_t, \boldsymbol{\theta}_{\mathcal{D}}), \tag{3.1}$$

where $\mathbf{R}_t$ denotes an RC at time t and $\mathcal{F}_{t-1} = \{\mathbf{R}_{t-1}, \mathbf{R}_{t-2} \dots\}$. The scalar dynamics for the conditional expected value matrix in a traditional observation-driven model might be, for example, a scalar-BEKK specification:

$$\mathbf{\Sigma}_{t+1} = (1 - a - b)\mathbf{\Xi} + a\mathbf{R}_t + b\mathbf{\Sigma}_t. \tag{3.2}$$

This scalar-dynamic assumption, which is necessary to make the models applicable to more than, say, five assets, restrictively implies that each element in $\mathbf{\Sigma}_{t+1}$ only depends on the corresponding previous element in $\mathbf{\Sigma}_t$ and the corresponding previous realization in $\mathbf{R}_t$. That is, there is no possibility for (co)volatility spillovers, which is a very strong assumption. In reality there is complex non-linear interdependence between different firms. On the other hand, a scalar-GAS model, where $\mathbf{R}_t$ in equation (3.2) is replaced by the score of the observational density, naturally accommodates volatility spillovers, since $\mathbf{R}_t$ enters non-linearly in the score.

The second drawback of traditional observation-driven models for RCs is the central assumption of constant d.o.f. parameters. To illustrate this we estimated the scalar-BEKK model given by equations (3.1) and (3.2) for all distributions for RCs $\mathcal{D}$ on a rolling window of 1250 observations for the ten-dimensional dataset described in Section 3.4.1. We plot the resulting d.o.f. parameter estimates in Figure 3.1. We see that the estimated d.o.f. parameters vary substantially over time for all distributions, even if we introduce time-variation in the expected value matrix ($\mathbf{\Sigma}_t$) via the scalar-BEKK specification. For an intuition, recall that the

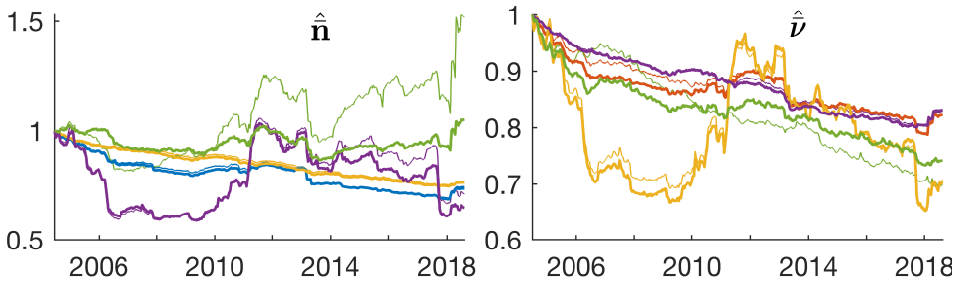

Figure 3.1: (Average) estimated d.o.f. parameters $\mathbf{n}$ and $\boldsymbol{\nu}$ from estimation on a moving window of 1250 observations of scalar-BEKK model for all distributions for the ten-dimensional dataset described in 3.4.1. All time-series are normed to the first estimated (average) degree of freedom. The plotted points correspond to the mid-point of the moving window sample. The color coding is: Riesz, Wishart, Inv.Riesz, Inv.Wishart, t-Riesz, t-Wishart, Inv.t-Riesz, Inv.t-Wishart, F-Riesz, F. The thick line corresponds to the Riesz-type distributions, for which we plot the average of their d.o.f. parameter vectors, the thin line to the corresponding Wishart-named (and the F) ones.

d.o.f. parameter ν of the t-Wishart distribution regulates its fat-tailedness, with lower values implying a more fat-tailed distribution. In Figure 3.1, we indeed see, as we would expect, that $\hat{\nu}$ of the t-Wishart decreases substantially (around 30 %) during the 2008 global financial crisis and in the COVID-19 crisis. We conclude that the assumption of static d.o.f. parameters is restrictive and that there is statistical and economic value in exploring time-variation in the d.o.f. parameters. For the d.o.f. parameters in an observation-driven model, there is no observable realization to use in an updating mechanism. Here, GAS models, with their use of the lagged score as the forcing variable, offer an intuitive way to dynamically update the d.o.f. parameters.

GAS Model Setup 3.3

As mentioned above, probability distributions for RCs can be characterized by a symmetric positive definite expected value matrix $\boldsymbol{\Sigma}$ and their distribution-specific d.o.f. parameters (one or two of the set $(n, \nu, \mathbf{n}, \boldsymbol{\nu})$), which we collect in the vector $\boldsymbol{\theta}_{\mathcal{D}}$. For example, for the F-Riesz distribution we have $\boldsymbol{\theta}_{F\mathcal{R}} = (\mathbf{n}^\top, \boldsymbol{\nu}^\top)^\top$ with d.o.f. parameter vectors $\mathbf{n} = (n_1, \ldots, n_p)^\top$ and $\boldsymbol{\nu} = (\nu_1, \ldots, \nu_p)^\top$, the t-Riesz distribution has d.o.f. parameter vector $\mathbf{n}$ and scalar d.o.f. ν, thus $\boldsymbol{\theta}_{t\mathcal{R}} = (\mathbf{n}^\top, \nu)^\top$, and the Wishart distribution has just one d.o.f. parameter, thus $\boldsymbol{\theta}_{\mathcal{W}} = n$.

Assuming that all parameters are time-varying, the conditional distribution can be written as

$$\mathbf{R}_t|\mathcal{F}_{t-1} \sim \mathcal{D}(\boldsymbol{\Sigma}_t, \boldsymbol{\theta}_{\mathcal{D},t}). \tag{3.3}$$

In the GAS model framework, the parameter updating at time $t+1$ is based on the score vector of the respective distribution $\mathcal{D}$ at time t,

$$\nabla_{\mathcal{D},t} = \left(\frac{\partial \log p_{\mathcal{D}}(\boldsymbol{\Sigma}_t, \boldsymbol{\theta}_{\mathcal{D},t}|\mathbf{R})}{\partial \text{vech}(\boldsymbol{\Sigma}_t)^\top}, \frac{\partial \log p_{\mathcal{D}}(\boldsymbol{\Sigma}_t, \boldsymbol{\theta}_{\mathcal{D},t}|\mathbf{R})}{\partial \boldsymbol{\theta}_{\mathcal{D},t}^\top} \right)^\top,$$

where $p_{\mathcal{D}}(\cdot)$ is the probability density function (p.d.f.) of distribution $\mathcal{D}$, and the vech$(\cdot)$-operator takes a lower-triangular or symmetric matrix as input and stacks its elements on and below the main diagonal column-wise into a column vector. We denote the distinct parts of the score by

$$\nabla^{\Sigma}_{\mathcal{D},t} = \left(\frac{\partial \log p_{\mathcal{D}}(\boldsymbol{\Sigma}_t, \boldsymbol{\theta}_{\mathcal{D},t}|\mathbf{R})}{\partial \text{vech}(\boldsymbol{\Sigma}_t)^\top} \right)^\top \text{ and}$$

$$\nabla^{\theta}_{\mathcal{D},t} = \left(\frac{\partial \log p_{\mathcal{D}}(\boldsymbol{\Sigma}_t, \boldsymbol{\theta}_{\mathcal{D},t}|\mathbf{R})}{\partial \boldsymbol{\theta}_{\mathcal{D},t}^\top} \right)^\top.$$

See Section 3.6.1 in the appendix for a treatment of multivariate derivatives.[4]

The time-t score vector gives the parameter updating direction in which the log-likelihood at time t can be improved most. Thus, it is a natural choice for a forcing variable in a dynamic updating equation. However, empirically, it behaves too erratically in order to extract valuable signals, and thus it is usually advocated to scale the score in some way (c.f. Creal, Koopman, and Lucas, 2013). A widely used scaling matrix is the inverse Fisher information matrix (FIM) (Gorgi et al., 2019, Blasques et al., 2022), which embodies information about the curvature of the log-likelihood,

$$\mathcal{I}_{\mathcal{D},t} = \mathbb{E}\begin{bmatrix} \nabla^{\Sigma}_{\mathcal{D},t}(\nabla^{\Sigma}_{\mathcal{D},t})^\top & \nabla^{\Sigma}_{\mathcal{D},t}(\nabla^{\theta}_{\mathcal{D},t})^\top \\ \nabla^{\theta}_{\mathcal{D},t}(\nabla^{\Sigma}_{\mathcal{D},t})^\top & \nabla^{\theta}_{\mathcal{D},t}(\nabla^{\theta}_{\mathcal{D},t})^\top \end{bmatrix}, \tag{3.4}$$

to obtain the scaled score vector

$$\mathbf{s}_{\mathcal{D},t} = \mathcal{I}^{-1}_{\mathcal{D},t} \nabla_{\mathcal{D},t}.$$

4. It is not immediately obvious how to arrange the individual derivatives $\frac{\partial \log p_{\mathcal{D}}(\boldsymbol{\Sigma}_t, \boldsymbol{\theta}_{\mathcal{D},t}|\mathbf{R})}{\partial (\boldsymbol{\Sigma}_t)_{ij}}$. Also, it is important to take the derivative w.r.t. vech$(\boldsymbol{\Sigma})$, not vec$(\boldsymbol{\Sigma})$, to take the symmetry of $\boldsymbol{\Sigma}$ into account. Both points are explained in more detail in Section 3.6.1.

In order to be able to estimate the parameters of the updating equation, one has to choose scalar dynamics, akin to the scalar-BEKK model, because full matrix dynamics suffer from the *curse of dimensionality*. However, since it seems unrealistic that $\mathbf{\Sigma}_t$ and the d.o.f. parameters $(\mathbf{n}_t, n_t, \boldsymbol{\nu}_t, \nu_t)$ follow the same scalar dynamics, we assume separate scalar-GAS updating equations for them. The scaled score w.r.t. $\mathbf{\Sigma}_t$ is defined as

$$\mathbf{S}^{\Sigma}_{\mathcal{D},t} = \text{ivech}((\mathcal{I}^{\Sigma}_{\mathcal{D},t})^{-1}\nabla^{\Sigma}_{\mathcal{D},t}), \tag{3.5}$$

$$\text{with } \mathcal{I}^{\Sigma}_{\mathcal{D},t} = \mathbb{E}[\nabla^{\Sigma}_{\mathcal{D},t}(\nabla^{\Sigma}_{\mathcal{D},t})^{\top}], \tag{3.6}$$

where $\text{ivech}(\cdot)$ is the inverse vech operator that creates a symmetric matrix from a vector of suitable size. For the d.o.f. parameters in $\boldsymbol{\theta}_{\mathcal{D},t}$ we assume separate scalar-GAS updating equations for $\mathbf{n}_t, n_t, \boldsymbol{\nu}_t$ and ν_t, with the scaled scores w.r.t. $\mathbf{n}_t$ and n_t given by

$$\mathbf{s}^{\mathbf{n}}_{\mathcal{D},t} = (\mathcal{I}^{\mathbf{n}}_{\mathcal{D},t})^{-1}\nabla^{\mathbf{n}}_{\mathcal{D},t}$$
$$\text{and } s^{n}_{\mathcal{D},t} = \frac{\nabla^{n}_{\mathcal{D},t}}{\mathcal{I}^{n}_{\mathcal{D},t}},$$

where

$$\nabla^{\mathbf{n}}_{\mathcal{D},t} = \left(\frac{\partial \log p_{\mathcal{D}}(\cdot)}{\partial \mathbf{n}_t^{\top}}\right)^{\top},$$
$$\nabla^{n}_{\mathcal{D},t} = \frac{\partial \log p_{\mathcal{D}}(\cdot)}{\partial n_t},$$

and

$$\mathcal{I}^{\mathbf{n}}_{\mathcal{D},t} = \mathbb{E}[\nabla^{\mathbf{n}}_{\mathcal{D},t}(\nabla^{\mathbf{n}}_{\mathcal{D},t})^{\top}],$$
$$\mathcal{I}^{n}_{\mathcal{D},t} = \mathbb{E}[(\nabla^{n}_{\mathcal{D},t})^2].$$

The scaled scores w.r.t $\boldsymbol{\nu}_t$ and ν_t are defined analogously to those of $\mathbf{n}_t$ and n_t. Which of the d.o.f. parameter updating equations is present in the respective model depends, of course, on the chosen distribution.

Distribution	$2\triangle^{\Omega}_{\mathcal{D}}$
Wishart	$\boldsymbol{\Omega}^{-1}\mathbf{R}\boldsymbol{\Omega}^{-1} - n\boldsymbol{\Omega}^{-1}$
Riesz	$\boldsymbol{\Omega}^{-1}\mathbf{R}\boldsymbol{\Omega}^{-1} - \mathbf{C}_{\Omega}^{-\top}\mathrm{dg}(\mathbf{n})\mathbf{C}_{\Omega}^{-1}$
Inv.Wishart	$\nu\boldsymbol{\Omega}^{-1} - \mathbf{R}^{-1}$
Inv.Riesz	$\mathbf{C}_{\Omega}^{-\top}\mathrm{dg}(\boldsymbol{\nu})\mathbf{C}_{\Omega}^{-1} - \mathbf{R}^{-1}$
t-Wishart	$\frac{\nu+pn}{1+\mathrm{tr}(\boldsymbol{\Omega}^{-1}\mathbf{R})}\boldsymbol{\Omega}^{-1}\mathbf{R}\boldsymbol{\Omega}^{-1} - n\boldsymbol{\Omega}^{-1}$
t-Riesz	$\frac{\nu+p\bar{\mathbf{n}}}{1+\mathrm{tr}(\boldsymbol{\Omega}^{-1}\mathbf{R})}\boldsymbol{\Omega}^{-1}\mathbf{R}\boldsymbol{\Omega}^{-1} - \mathbf{C}_{\Omega}^{-\top}\mathrm{dg}(\mathbf{n})\mathbf{C}_{\Omega}^{-1}$
Inv.t-Wishart	$\nu\boldsymbol{\Omega}^{-1} - \frac{\nu+pn}{1+\mathrm{tr}(\boldsymbol{\Omega}\mathbf{R}^{-1})}\mathbf{R}^{-1}$
Inv.t-Riesz	$\mathbf{C}_{\Omega}^{-\top}\mathrm{dg}(\boldsymbol{\nu})\mathbf{C}_{\Omega}^{-1} - \frac{n+p\bar{\nu}}{1+\mathrm{tr}(\boldsymbol{\Omega}\mathbf{R}^{-1})}\mathbf{R}^{-1}$
F	$\mathbf{C}_{\Omega}^{-\top}\mathrm{dg}(\boldsymbol{\nu})\mathbf{C}_{\Omega}^{-1} - \mathbf{C}_{H}^{-\top}\mathrm{dg}(\boldsymbol{\nu}+\mathbf{n})\mathbf{C}_{H}^{-1}$
F-Riesz	$\mathbf{C}_{\Omega}^{-\top}\mathrm{dg}(\boldsymbol{\nu})\mathbf{C}_{\Omega}^{-1} - \mathbf{C}_{H}^{-\top}\mathrm{dg}(\boldsymbol{\nu}+\mathbf{n})\mathbf{C}_{H}^{-1}$
Inv.F-Riesz	$\mathbf{C}_{\Omega}^{-\top}\mathrm{dg}(\mathbf{n})\mathbf{C}_{\Omega}^{-1} - \boldsymbol{\Omega}^{-1}\mathbf{C}_{J}\mathrm{dg}(\mathbf{n}+\boldsymbol{\nu})\mathbf{C}_{J}^{\top}\boldsymbol{\Omega}^{-1}$

Table 3.1: $\triangle^{\Omega}_{\mathcal{D}}$ is the $p \times p$ score matrix w.r.t. $\boldsymbol{\Omega}$, as defined in equation (3.33) in the appendix, with subscripts t omitted for readability. The scores w.r.t. $\boldsymbol{\Omega}$ are given by $\nabla^{\Omega}_{\mathcal{D}} = \mathbf{G}^{\top}\mathrm{vec}(\triangle^{\Omega}_{\mathcal{D}})$ and the ones w.r.t. $\boldsymbol{\Sigma}$ then easily follow from Lemma 3.6.9. The proofs are straightforward using Lemmas 3.6.6 and 3.6.7. As an example, see Section 3.6.2 in the appendix for the derivation of the F-Riesz score. $\mathbf{C}_H$ and $\mathbf{C}_J$ denote the lower Cholesky factor of $\mathbf{H} = \boldsymbol{\Omega} + \mathbf{R}$ and $\mathbf{J} = (\boldsymbol{\Omega}^{-1} + \mathbf{R}^{-1})^{-1}$, respectively.

3.3.1 Expected Value Matrix Recursion

The scores and FIMs w.r.t. $\boldsymbol{\Sigma}_t$ for the different distributions are given in Tables 3.1 and 3.2. The derivations are given in Section 3.6.2 in the appendix. The formulas for the FIMs contain $p^2 \times p^2$ matrix multiplications and inversions which makes the direct calculation of the scaled score (equation (3.5)) prohibitively slow. Thus, we propose to scale the score w.r.t. $\boldsymbol{\Sigma}_t$ of all Riesz-type distributions with the FIMs of the corresponding Wishart-type distributions. That is, for example, we scale the F-Riesz distribution score ($\nabla^{\Sigma}_{F\mathcal{R},t}$) by the F distribution inverse FIM ($\mathcal{I}^{\Sigma}_{F,t}$) instead of its own ($\mathcal{I}^{\Sigma}_{F\mathcal{R},t}$), by setting $n_{F,t} = \bar{\mathbf{n}}_{F\mathcal{R},t}$ and $\nu_{F,t} = \bar{\boldsymbol{\nu}}_{F\mathcal{R},t}$.

We can be flexible with the scaling of the scores since the most important information is contained in the score itself, which at time t defines the steepest ascend direction of the log-likelihood in which the parameters should be updated in $t+1$. The scaling just serves to stabilize the time-series behavior of the scores. In fact, in a small dimensional example using the t-Riesz distribution we found

Distribution	$\mathcal{I}_{\mathcal{D}}^{\Sigma}$
Wishart	$\frac{n}{2}\mathbf{G}^\top\boldsymbol{\Sigma}^{-\otimes 2}\mathbf{G}$
Inv.Wishart	$\frac{\nu}{2}\mathbf{G}^\top\boldsymbol{\Sigma}^{-\otimes 2}\mathbf{G}$
t-Wishart	$\frac{n}{2}\mathbf{G}^\top\left(\frac{\nu+pn}{\nu+pn+2}\boldsymbol{\Sigma}^{-\otimes 2}-\frac{n}{(\nu+pn+2)}\text{vec}^2\left(\boldsymbol{\Sigma}^{-1}\right)\right)\mathbf{G}$
Inv.t-Wishart	$\frac{\nu}{2}\mathbf{G}^\top\left(\frac{n+p\nu}{n+p\nu+2}\boldsymbol{\Sigma}^{-\otimes 2}-\frac{\nu}{(n+p\nu+2)}\text{vec}^2\left(\boldsymbol{\Sigma}^{-1}\right)\right)\mathbf{G}$
F	$\frac{1}{2}\mathbf{G}^\top\left((\nu+(n+\nu)(c_3+c_4))\,\boldsymbol{\Sigma}^{-\otimes 2}+(n+\nu)c_4\text{vec}^2\left(\boldsymbol{\Sigma}^{-1}\right)\right)\mathbf{G}$

Table 3.2: Fisher information matrices w.r.t $\boldsymbol{\Sigma}$ of Wishart-type distributions. The subscripts t are omitted for readability. For the derivations (and the definitions of c_3 and c_4) see Section 3.6.2 in the appendix. The Fisher information matrices of the (inverse) Riesz and t-Riesz distributions are also derived in the appendix.

that the $\boldsymbol{\Sigma}_t$-dynamics obtained from scaling with the "correct" t-Riesz FIM are very similar to the ones obtained from scaling with the t-Wishart FIM. Opschoor et al. (2018) scale the matrix F distribution with the inverse FIM of the Wishart distribution. Furthermore, Blasques, Francq, and Laurent (2022) show that it is possible to disentangle the distributional assumption from the score dynamics completely, i.e. one could update the parameters with a scaled score from a different distribution than the conditional distributional assumption on the data.

In Theorem 3.3.1 we show that in the formulas for the scaled scores that result from this scaling scheme, only $p \times p$ matrix operations remain, which makes the models empirically feasible again.

Theorem 3.3.1. *Consider the scaled score w.r.t.* $\boldsymbol{\Sigma}_t$*,*

$$\mathbf{S}_{\mathcal{D},t}^{\Sigma} = \text{ivech}((\mathcal{I}_{\mathcal{D},t}^{\Sigma})^{-1}\nabla_{\mathcal{D},t}^{\Sigma}),$$

as defined in equation (3.5). If, for any Riesz-type distribution, we use $\mathcal{I}_{\mathcal{D},t}^{\Sigma}$ *of its Wishart-type counterpart, instead of its own, by setting the degree(s) of freedom equal to the average of the corresponding d.o.f. parameter vector(s). Then, for all distributions for RCs the scaled score w.r.t.* $\boldsymbol{\Sigma}_t$ *can be written as*

$$\dot{\mathbf{S}}_{\mathcal{D},t}^{\Sigma} = \alpha_{\mathcal{D}}\boldsymbol{\Sigma}_t\triangle_{\mathcal{D},t}^{\Sigma}\boldsymbol{\Sigma}_t + \beta_{\mathcal{D}}\text{tr}\left(\boldsymbol{\Sigma}_t\triangle_{\mathcal{D},t}^{\Sigma}\right)\boldsymbol{\Sigma}_t,$$

where $\triangle_{\mathcal{D},t}^{\Sigma}$ *is the* $p \times p$ *score matrix w.r.t.* $\boldsymbol{\Sigma}_t$ *ignoring symmetry (as defined in equation (3.30)), and* $\alpha_{\mathcal{D}}$ *and* $\beta_{\mathcal{D}}$ *are scalars that depend only on the d.o.f. parameters of the respective distribution. Proof in Appendix.*

We omit the FIMs w.r.t. $\boldsymbol{\Sigma}_t$ for the Riesz-type distributions in Table 3.2 since using Theorem 3.3.1 we don't need them and their formulas are long and cumbersome.[5]

Now we present our specification for the expected value matrix equation. We assume

$$\begin{aligned}\boldsymbol{\Sigma}_{t+1} = (1 - b_1 - b_2 - b_3)\boldsymbol{\Xi} + a_1\boldsymbol{\Sigma}_t\triangle^{\Sigma}_{\mathcal{D},t}\boldsymbol{\Sigma}_t + a_2\text{tr}(\boldsymbol{\Sigma}_t\triangle^{\Sigma}_{\mathcal{D},t})\boldsymbol{\Sigma}_t \\ + b_1\boldsymbol{\Sigma}_t + b_2\bar{\boldsymbol{\Sigma}}_{t:t-4} + b_3\bar{\boldsymbol{\Sigma}}_{t:t-21},\end{aligned} \tag{3.7}$$

where $\boldsymbol{\Xi}$ is a $p \times p$ symmetric positive definite parameter matrix and a_1, a_2, b_1, b_2 and b_3 are scalar parameters. Following Opschoor et al. (2018), we assume a HAR-type (see Corsi, 2009) structure, where $\bar{\boldsymbol{\Sigma}}_{t:t-4} = 1/5\sum_{j=0}^{4}\boldsymbol{\Sigma}_{t-j}$ and $\bar{\boldsymbol{\Sigma}}_{t:t-21} = 1/22\sum_{j=0}^{21}\boldsymbol{\Sigma}_{t-j}$ are the average expected value matrix of the previous trading week and month, respectively. In our empirical analysis, the HAR structure improved the fit and forecasting ability substantially over the version where only $\boldsymbol{\Sigma}_t$ goes into the updating process. Note that our scaling nest the one in Opschoor et al. (2018) by setting $a_2 = 0$. It is important to see, that instead of straightforwardly incorporating Theorem 3.3.1 by using $a\dot{\mathbf{S}}^{\Sigma}_{\mathcal{D},t} = a(\alpha_{\mathcal{D}}\boldsymbol{\Sigma}_t\triangle^{\Sigma}_{\mathcal{D},t}\boldsymbol{\Sigma}_t + \beta_{\mathcal{D}}\text{tr}(\boldsymbol{\Sigma}_t\triangle^{\Sigma}_{\mathcal{D},t})\boldsymbol{\Sigma}_t)$ as the forcing variable in our updating equation, we allow for different stand-alone parameters (a_1,a_2) for the two terms that constitute $\dot{\mathbf{S}}^{\Sigma}_{\mathcal{D},t}$. This comes at the small cost of adding one more parameter but has the advantage of adding flexibility and making the updating equation completely independent of the FIMs such that we don't need to calculate the complicated formulas for $\alpha_{\mathcal{D}}$ and $\beta_{\mathcal{D}}$.[6] Note that under this small generalization the mean-zero property of the two scaled score terms is preserved since $\mathbb{E}[\triangle^{\Sigma}_{\mathcal{D},t}|\mathcal{F}_{t-1}] = 0$. Furthermore, it is directly visible that for all distributions, even for scalar parameters there are spillovers of the realized (co)variances of all assets to the expected realized (co)variances of all other assets.

Note that our specification endows the expected value matrix $\boldsymbol{\Sigma}_t$ with GAS dynamics which is different from a model that assumes GAS dynamics for time-varying parameter matrix $\boldsymbol{\Omega}_t$ (recall from Chapter 2, that the probability distributions can be parameterized in terms of expected value matrix $\boldsymbol{\Sigma}_t$ or parameter matrix $\boldsymbol{\Omega}_t$). This is because the score w.r.t. $\boldsymbol{\Omega}$ is not simply a multiple of the score w.r.t. $\boldsymbol{\Sigma}$. In contrast, in a scalar-BEKK specification as in Blasques et al. (2021), it does *not* matter if the distributions are parameterized in terms of $\boldsymbol{\Omega}_t$ or $\boldsymbol{\Sigma}_t$. We choose the model $\boldsymbol{\Sigma}_t$ with a GAS specification instead of $\boldsymbol{\Omega}_t$ because there are several advantages to it. First, unlike $\boldsymbol{\Omega}_t$, the $\boldsymbol{\Sigma}_t$ parameter has the same meaning across distributions and an intuitive interpretation of yielding a mean-shifting process. Second, the distributions when parameterized in terms of $\boldsymbol{\Sigma}_t$ nest each

5. You can find the FIMs for some Riesz-type distributions expressions in Section 3.6.2 in the appendix.

6. These formulas are especially complicated for the F distribution.

other according to the distribution family tree in figure 2.6 of Chapter 2, and this nesting translates over to the GAS models. This makes comparisons of likelihood values via information criteria valid and opens the possibility of likelihood ratio tests. One could also use the optimized parameters of a GAS model with nested distributions as starting points for the estimation of the GAS models with nesting distributions. Third, the targeting estimation of the intercept matrix in the $\boldsymbol{\Sigma}_t$-GAS recursion is feasible and very easy. In fact, if we introduce time-variation in the d.o.f. parameters, like we do in the next section, it is impossible to target the intercept matrix of a $\boldsymbol{\Omega}_t$-GAS specification since the unconditional mean of $\boldsymbol{\Omega}_t$ depends on the d.o.f. parameters.

As an interesting side note, in Lemma 3.3.1, we show that, for the case of the Riesz (and Wishart) distribution, the original scaled score $\mathbf{S}^{\Sigma}_{\mathcal{D},t}$ boils down to the difference between $\mathbf{R}_t$ and $\boldsymbol{\Sigma}_t$ (and thus also does not require $p^2 \times p^2$ matrix multiplications).

Lemma 3.3.1. *For the Riesz (and consequently also the Wishart) distribution, we have that the scaled score w.r.t.* $\boldsymbol{\Sigma}$, $\mathbf{S}^{\Sigma}_{\mathcal{D},t}$, *in equation (3.5) can be rewritten as*

$$\mathbf{S}^{\Sigma}_{\mathcal{R},t} = \mathbf{R}_t - \boldsymbol{\Sigma}_t.$$

Proof in Appendix.

Consequently, a scalar-GAS (1,1) specification would simplify to

$$\begin{aligned}\boldsymbol{\Sigma}_{t+1} &= (1-b)\boldsymbol{\Xi} + a\mathbf{S}^{\Sigma}_{\mathcal{R},t} + b\boldsymbol{\Sigma}_t \\ &= (1-b)\boldsymbol{\Xi} + a\mathbf{R}_t + (b-a)\boldsymbol{\Sigma}_t,\end{aligned}$$

which is equivalent to the well-known scalar-BEKK specification in equation (3.2). For the other Riesz-type distributions, however, such simplifications are not possible.

Degree of Freedom Parameters Recursion 3.3.2

We assume that the time-varying d.o.f. parameters in $\boldsymbol{\theta}_{\mathcal{D},t}$ follow GARCH-type recursions,

$$\mathbf{n}_{t+1} = (1-b^{\mathbf{n}})\boldsymbol{\xi}^{\mathbf{n}} + a^{\mathbf{n}}\mathbf{s}^{\mathbf{n}}_{\mathcal{D},t} + b^{\mathbf{n}}\mathbf{n}_t, \tag{3.8}$$

$$n_{t+1} = (1-b^{n})\xi^{n} + a^{n}s^{n}_{\mathcal{D},t} + b^{n}n_t, \tag{3.9}$$

where $\boldsymbol{\xi}^{\mathbf{n}}$ is a parameter vector of size $p \times 1$, $\xi^n, a^{\mathbf{n}}, b^{\mathbf{n}}, a^n$ and b^n are scalar parameters, and $\mathbf{s}^{\mathbf{n}}_{\mathcal{D},t}$ $(p \times 1)$ and $\mathbf{s}^{n}_{\mathcal{D},t}$ (scalar) are the scaled scores with respect

to $\mathbf{n}_t$ and n_t, respectively.[7] The specifications for $\boldsymbol{\nu}_t$ and ν_t are analogous to those of $\mathbf{n}_t$ and n_t. Of course, which of the recursions ($\mathbf{n}_t, \boldsymbol{\nu}_t, n_t$ and/or ν_t) are present in the respective model depends on the chosen distribution $\mathcal{D}$. We collect all distribution-specific d.o.f.-recursion parameters in the vector $\boldsymbol{\vartheta}_{\mathcal{D}}$. For example, $\boldsymbol{\vartheta}_{\mathcal{W}} = (\xi^n, a^n, b^n)^\top$ and $\boldsymbol{\vartheta}_{F\mathcal{R}} = ((\boldsymbol{\xi}^{\mathbf{n}})^\top, a^{\mathbf{n}}, b^{\mathbf{n}}, (\boldsymbol{\xi}^{\boldsymbol{\nu}})^\top, a^{\boldsymbol{\nu}}, b^{\boldsymbol{\nu}})^\top$. The scores and FIMs w.r.t. $\mathbf{n}_t, n_t, \boldsymbol{\nu}_t$ and ν_t, which are needed to construct the forcing variables (scaled scores) in the recursions, are listed in Tables 3.3 and 3.4, respectively, and are derived in Section 3.6.2 in the appendix. Note that for the special case that the non-intercept parameters $(a^{\mathbf{n}}, a^n, a^{\boldsymbol{\nu}}, a^\nu, b^{\mathbf{n}}, b^n, b^{\boldsymbol{\nu}}, b^\nu)$ are set to equal zero, we recover a restricted GAS model with constant d.o.f. parameters (GAS dynamics only in $\boldsymbol{\Sigma}_t$).

To give some intuition on the interpretation of the d.o.f. parameters, consider omitting the subscript t, the one-dimensional case, i.e. $\mathbf{R}$ being of size 1×1. In that case the stochastic representation of all distributions[8] can be written as

$$\mathbf{R} = \begin{cases} \boldsymbol{\Sigma}\,\Gamma(\frac{n}{2}, \frac{n}{2}) & (\mathcal{W}, \mathcal{R}) \\ \boldsymbol{\Sigma}\,\Gamma^{-1}(\frac{\nu}{2}, \frac{\nu-2}{2}) & (i\mathcal{W}, i\mathcal{R}) \\ \boldsymbol{\Sigma}\,\Gamma(\frac{n}{2}, \frac{n}{2})\Gamma^{-1}(\frac{\nu}{2}, \frac{\nu-2}{2}) & (t\mathcal{W}, t\mathcal{R}, it\mathcal{W}, it\mathcal{R}, F, F\mathcal{R}, iF\mathcal{R}), \end{cases}$$

where $\boldsymbol{\Sigma}$ is the 1×1 expected value and $\Gamma(n/2, n/2)$ and $\Gamma^{-1}(\nu/2, (\nu-2)/2)$ are mean-zero Gamma and Inverse Gamma distributions, respectively. The variance and excess kurtosis of both the $\Gamma(n/2, n/2)$ and the $\Gamma^{-1}(\nu/2, (\nu-2)/2)$ increases, if the d.o.f. parameters n and ν decrease. That is, the distributions get "wider" and more "fat-tailed", which carries over to the distribution of the one-dimensional $\mathbf{R}_t$. This is economically important because if we are interested in predicting $\mathbf{R}_{t+1}$, simple point predictions are of limited use without an accompanying statement about the distributional properties of this forecast, either in the form of confidence bands or by stating variance and tail-measures. Because of the outlined interpretation, we expect the d.o.f. parameters to fall during times of crisis. Thus forecasts of the d.o.f. parameters could also be used to identify upcoming market turmoil.

7. To mirror the $\boldsymbol{\Sigma}_t$ specification, we also tried HAR-type recursions for the d.o.f. parameters. However, in the empirical application, it turned out that the lagged weekly and monthly average d.o.f. parameters were insignificant, and their estimates were often close to zero.

8. Recall equation (2.7): $\mathbf{R} = \mathbf{C}\mathbf{M}_{\mathcal{D}}^{-\frac{1}{2}}\boldsymbol{\mathcal{K}}_{\mathcal{D}}\mathbf{M}_{\mathcal{D}}^{-\frac{1}{2}}\mathbf{C}^\top$.

Distribution	$2\nabla^{\mathbf{n}}$ or $2\nabla^{n}$
Wishart	$-p\log(2) - \sum_{j=1}^{p}\psi(\frac{n-j+1}{2}) + \log\lvert\boldsymbol{\Omega}^{-1}\mathbf{R}\rvert$
Riesz	$-\log(2) - \boldsymbol{\psi}(\frac{\mathbf{n}}{2}) + 2\log\operatorname{vecd}(\mathbf{C}_{\Omega}^{-1}\mathbf{C}_{R})$
Inv.Wishart	-
Inv.Riesz	-
t-Wishart	$p\psi(\frac{\nu+pn}{2}) - \sum_{j=1}^{p}\psi(\frac{n-j+1}{2}) - \log\lvert\boldsymbol{\Omega}^{-1}\mathbf{R}\rvert - p\log(1+\operatorname{tr}(\boldsymbol{\Omega}^{-1}\mathbf{R}))$
t-Riesz	$\psi(\frac{\nu+p\bar{\mathbf{n}}}{2}) - \boldsymbol{\psi}(\frac{\mathbf{n}}{2}) + 2\log\operatorname{vecd}(\mathbf{C}_{\Omega}^{-1}\mathbf{C}_{R}) - \log(1+\operatorname{tr}(\boldsymbol{\Omega}^{-1}\mathbf{R}))$
Inv.t-Wishart	$\psi(\frac{n+p\nu}{2}) - \psi(\frac{n}{2}) - \log(1+\operatorname{tr}(\boldsymbol{\Omega}\mathbf{R}^{-1}))$
Inv.t-Riesz	$\psi(\frac{n+p\bar{\boldsymbol{\nu}}}{2}) - \psi(\frac{n}{2}) - \log(1+\operatorname{tr}(\boldsymbol{\Omega}\mathbf{R}^{-1}))$
F	$\sum_{j=1}^{p}\psi(\frac{\nu+n-j+1}{2}) - \psi(\frac{n-j+1}{2}) + \log\lvert\mathbf{R}\rvert - \log\lvert\mathbf{H}\rvert$
F-Riesz	$\overleftarrow{\boldsymbol{\psi}}(\frac{\overleftarrow{\mathbf{n}}+\overleftarrow{\boldsymbol{\nu}}}{2}) - \boldsymbol{\psi}(\frac{\mathbf{n}}{2}) + 2\log\operatorname{vecd}(\mathbf{C}_{R}) - 2\log\operatorname{vecd}(\mathbf{C}_{H})$
Inv.F-Riesz	$\boldsymbol{\psi}(\frac{\mathbf{n}+\boldsymbol{\nu}}{2}) - \boldsymbol{\psi}(\frac{\mathbf{n}}{2}) - 2\log\operatorname{vecd}(\mathbf{C}_{\Omega}) + 2\log\operatorname{vecd}(\mathbf{C}_{J})$
Distribution	$2\nabla^{\boldsymbol{\nu}}$ or $2\nabla^{\nu}$
Wishart	-
Riesz	-
Inv.Wishart	$-p\log(2) - \sum_{j=1}^{p}\psi(\frac{\nu-j+1}{2}) + \log\lvert\boldsymbol{\Omega}\mathbf{R}^{-1}\rvert$
Inv.Riesz	$-\log(2) - \overleftarrow{\boldsymbol{\psi}}(\frac{\overleftarrow{\boldsymbol{\nu}}}{2}) + 2\log\operatorname{vecd}(\mathbf{C}_{R}^{-1}\mathbf{C}_{\Omega})$
t-Wishart	$\psi(\frac{\nu+pn}{2}) - \psi(\frac{\nu}{2}) - \log(1+\operatorname{tr}(\boldsymbol{\Omega}^{-1}\mathbf{R}))$
t-Riesz	$\psi(\frac{\nu+p\bar{\mathbf{n}}}{2}) - \psi(\frac{\nu}{2}) - \log(1+\operatorname{tr}(\boldsymbol{\Omega}^{-1}\mathbf{R}))$
Inv.t-Wishart	$p\psi(\frac{n+p\nu}{2}) - \sum_{j=1}^{p}\psi(\frac{\nu-j+1}{2}) + \log\lvert\boldsymbol{\Omega}\mathbf{R}^{-1}\rvert - p\log(1+\operatorname{tr}(\boldsymbol{\Omega}\mathbf{R}^{-1}))$
Inv.t-Riesz	$\psi(\frac{n+p\bar{\boldsymbol{\nu}}}{2}) - \overleftarrow{\psi}(\frac{\overleftarrow{\boldsymbol{\nu}}}{2}) + 2\log\operatorname{vecd}(\mathbf{C}_{R}^{-1}\mathbf{C}_{\Omega}) - \log(1+\operatorname{tr}(\boldsymbol{\Omega}\mathbf{R}^{-1}))$
F	$\sum_{j=1}^{p}\psi(\frac{\nu+n-j+1}{2}) - \psi(\frac{\nu-j+1}{2}) + \log\lvert\boldsymbol{\Omega}\rvert - \log\lvert\mathbf{H}\rvert$
F-Riesz	$\overleftarrow{\boldsymbol{\psi}}(\frac{\overleftarrow{\mathbf{n}}+\overleftarrow{\boldsymbol{\nu}}}{2}) - \overleftarrow{\boldsymbol{\psi}}(\frac{\overleftarrow{\boldsymbol{\nu}}}{2}) + 2\log\operatorname{vecd}(\mathbf{C}_{\Omega}) - 2\log\operatorname{vecd}(\mathbf{C}_{H})$
Inv.F-Riesz	$\boldsymbol{\psi}(\frac{\mathbf{n}+\boldsymbol{\nu}}{2}) - \overleftarrow{\boldsymbol{\psi}}(\frac{\overleftarrow{\boldsymbol{\nu}}}{2}) - 2\log\operatorname{vecd}(\mathbf{C}_{R}) + 2\log\operatorname{vecd}(\mathbf{C}_{J})$

Table 3.3: Scores w.r.t. $\mathbf{n}$, n, $\boldsymbol{\nu}$ and ν of all distributions for RCs. $\mathbf{C}_H$ and $\mathbf{C}_J$ denote the lower Cholesky factors of $\mathbf{H} = \boldsymbol{\Omega} + \mathbf{R}$ and $\mathbf{J} = (\boldsymbol{\Omega}^{-1} + \mathbf{R}^{-1})^{-1}$, respectively. The derivations are straightforward by applying Lemmas 3.6.10 and 3.6.12 (appendix) directly on the log of the p.d.f.s given in Table 2.5. Also note Lemma 3.6.11.

Distribution	$4\mathcal{I}^{\mathbf{n}}$ or $4\mathcal{I}^{n}$	$4\mathcal{I}^{\boldsymbol{\nu}}$ or $4\mathcal{I}^{\nu}$
Wishart	$\sum_{j=1}^{p}\psi'(\frac{n-j+1}{2})$	-
Riesz	$\boldsymbol{\psi}'(\frac{\mathbf{n}}{2})$	-
Inv.Wishart	-	$\sum_{j=1}^{p}\psi'(\frac{\nu-j+1}{2})$
Inv.Riesz	-	$\overleftarrow{\boldsymbol{\psi}}'(\frac{\overleftarrow{\boldsymbol{\nu}}}{2})$
t-Wishart	$\sum_{j=1}^{p}\psi'(\frac{n-j+1}{2}) - p^2\psi'(\frac{\nu+pn}{2})$	$\psi'(\frac{\nu}{2}) - \psi'(\frac{\nu+pn}{2})$
t-Riesz	$\boldsymbol{\psi}'(\frac{\mathbf{n}}{2}) - \psi'(\frac{\nu+p\bar{\mathbf{n}}}{2})$	$\psi'(\frac{\nu}{2}) - \psi'(\frac{\nu+p\bar{\mathbf{n}}}{2})$
Inv.t-Wishart	$\psi'(\frac{n}{2}) - \psi'(\frac{n+p\nu}{2})$	$\sum_{j=1}^{p}\psi'(\frac{\nu-j+1}{2}) - p^2\psi'(\frac{n+p\nu}{2})$
Inv.t-Riesz	$\psi'(\frac{n}{2}) - \psi'(\frac{n+p\bar{\boldsymbol{\nu}}}{2})$	$\overleftarrow{\psi}'(\frac{\overleftarrow{\boldsymbol{\nu}}}{2}) - \psi'(\frac{n+p\bar{\boldsymbol{\nu}}}{2})$
F	$\psi'(\frac{n-j+1}{2}) - \sum_{j=1}^{p}\psi'(\frac{\nu+n-j+1}{2})$	$\psi'(\frac{\nu-j+1}{2}) - \sum_{j=1}^{p}\psi'(\frac{\nu+n-j+1}{2})$
F-Riesz	$\boldsymbol{\psi}'(\frac{\mathbf{n}}{2}) - \overleftarrow{\boldsymbol{\psi}}'(\frac{\overleftarrow{\mathbf{n}}+\overleftarrow{\boldsymbol{\nu}}}{2})$	$\overleftarrow{\boldsymbol{\psi}}'(\frac{\overleftarrow{\boldsymbol{\nu}}}{2}) - \overleftarrow{\boldsymbol{\psi}}'(\frac{\overleftarrow{\mathbf{n}}+\overleftarrow{\boldsymbol{\nu}}}{2})$
Inv.F-Riesz	$\boldsymbol{\psi}'(\frac{\mathbf{n}}{2}) - \boldsymbol{\psi}'(\frac{\mathbf{n}+\boldsymbol{\nu}}{2})$	$\overleftarrow{\boldsymbol{\psi}}'(\frac{\overleftarrow{\boldsymbol{\nu}}}{2}) - \boldsymbol{\psi}'(\frac{\mathbf{n}+\boldsymbol{\nu}}{2})$

Table 3.4: Fisher information matrices w.r.t. $\mathbf{n}$, n, $\boldsymbol{\nu}$ and ν directly obtained by applying Lemma 3.6.12 (appendix) on the scores in Table 3.3.

3.3.3 Summary

To summarize, we collect all model parameters in the vector

$$\boldsymbol{\delta}_{\mathcal{D}} = (\text{vech}(\boldsymbol{\Xi})^{\top}, a_1, a_2, b_1, b_2, b_3, \boldsymbol{\vartheta}_{\mathcal{D}}^{\top})^{\top},$$

such that we can write

$$\mathbf{R}_t|\mathcal{F}_{t-1} \sim \mathcal{D}\left(\boldsymbol{\Sigma}_t(\boldsymbol{\delta}_{\mathcal{D}}), \boldsymbol{\theta}_{\mathcal{D},t}(\boldsymbol{\delta}_{\mathcal{D}})\right).$$

Both hyperparameters $\boldsymbol{\Sigma}_t$ and $\boldsymbol{\theta}_{\mathcal{D},t}$ are dependent on all model parameters $\boldsymbol{\delta}_{\mathcal{D}}$ since their scaled scores on both $\boldsymbol{\Sigma}_t$ and $\boldsymbol{\theta}_{\mathcal{D},t}$.

With our assumed specifications for $\boldsymbol{\Sigma}_t$ and $\boldsymbol{\theta}_t$, we believe that we have made the time-variation of the parameter space as rich as feasible while avoiding the curse of dimensionality. Thus, the remaining differences in fit and forecasting ability reflect mainly the appropriateness of the distributional assumption.

Empirical Application 3.4

Data 3.4.1

We take the same datasets as in the previous chapter for the same reasons as stated there (see Section 2.5.2).

Estimation of the GAS Models 3.4.2

Since $\mathbb{E}[\triangle^{\Sigma}_{\mathcal{D},t}|\mathcal{F}_{t-1}] = \mathbf{0}$, it is easy to show (see equation (3.7)) that

$$\mathbb{E}[\mathbf{R}_t] = \boldsymbol{\Xi}. \tag{3.10}$$

We thus choose the well-established two-step estimation method where we "target" $\boldsymbol{\Xi}$ in the first step; that is, we apply the obvious method of moments estimator

$$\widehat{\boldsymbol{\Xi}} = \frac{1}{T}\sum_{t=1}^{T}\mathbf{R}_t, \tag{3.11}$$

and we estimate the rest of the parameters in a second step via standard numerical maximum likelihood estimation, conditional on our estimate for $\boldsymbol{\Xi}$. As in the previous chapter, we follow the algorithm proposed in Blasques et al. (2021) to optimize over the asset ordering for Riesz-type distributions.

Note that the overall number of parameters that must be estimated via numerical maximum likelihood is constant for Wishart-type distributions and is of order $\mathcal{O}(p)$ for Riesz-type distributions. This is because the intercepts of the d.o.f. recursions in the Riesz-type distributions ($\boldsymbol{\xi}^{\mathbf{n}}$ and/or $\boldsymbol{\xi}^{\boldsymbol{\nu}}$) contain p elements, while they are scalars for Wishart-type distributions.

In our empirical analysis, we noticed that the estimation of the dynamic d.o.f. recursion parameters (the a's and b's in equations (3.8) and (3.9)) is very sensitive to the chosen starting values since the likelihood is relatively flat in their directions. A "bottom-up" estimation method that proved to recover sensible estimates relatively reliably is to first estimate the restricted model with constant d.o.f. parameters, that is, restricting all present a and b d.o.f. recursion parameters $(a^{\mathbf{n}}, a^{n}, a^{\boldsymbol{\nu}}, a^{\nu}, b^{\mathbf{n}}, b^{n}, b^{\boldsymbol{\nu}}, b^{\nu})$ to be equal to zero. This provides good estimates for the rest of the parameters (i.e. for $\boldsymbol{\Xi}$, a_1, a_2, b_1, b_2 and b_3 and for the intercepts in the d.o.f. recursions $\boldsymbol{\xi}^{\mathbf{n}}$, $\boldsymbol{\xi}^{\boldsymbol{\nu}}$, ξ^{n} and ξ^{ν}). Then, we set the obtained estimates as starting values for a second restriction-less optimization. As a useful by-product, we receive the estimated restricted GAS models with constant d.o.f. parameters, which we can compare via likelihood ratio test and forecast comparison to the non-restricted versions to determine the importance of time-varying d.o.f. parameters.

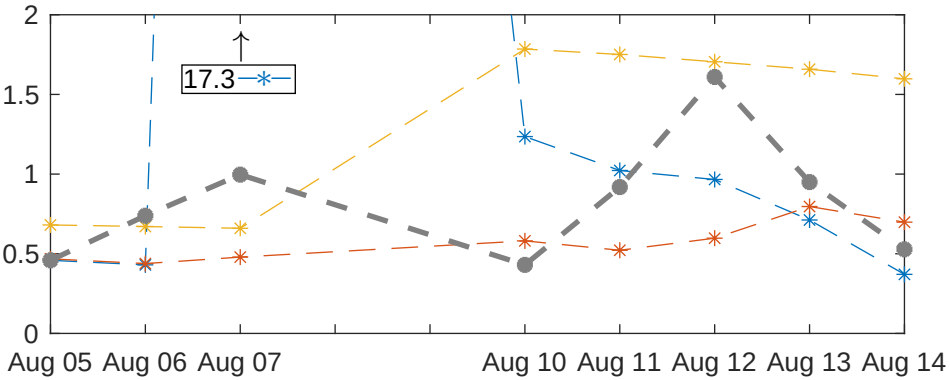

Figure 3.2: Time frame: 05 August 2015 - 14 August 2015. Blue line: Realized variance of American Express (AXP). Grey line: Median realized variance of the 15 stocks in *Financial* dataset. Both the yellow and the red line assume the best-fitting Inverse t-Riesz distribution ($\mathbf{R}_t|\mathcal{F}_{t-1} \sim it\mathcal{R}(\mathbf{\Sigma}_t, \mathbf{n}, \boldsymbol{\nu})$) and depict the estimated conditional expected variance $(\widehat{\mathbf{\Sigma}}_t)_{3\,3}$.) Yellow line: scalar-BEKK model ($\mathbf{\Sigma}_{t+1} = (1-a-b)\mathbf{\Xi} + a\mathbf{R}_t + b\mathbf{\Sigma}_t$). Red line: scalar-GAS model ($\mathbf{\Sigma}_{t+1} = (1-a-b)\mathbf{\Xi} + a_1\mathbf{\Sigma}_t\triangle^{\Sigma}_{it\mathcal{R},t}\mathbf{\Sigma}_t + a_2\mathrm{tr}\left(\mathbf{\Sigma}_t\triangle^{\Sigma}_{it\mathcal{R},t}\right)\mathbf{\Sigma}_t + b\mathbf{\Sigma}_t$).

3.4.3 Illustration of GAS Advantages in $\mathbf{\Sigma}_t$ Updating Equation

In Figure 3.2, we plot an illustration of the advantages of scalar-GAS model over scalar-BEKK dynamics in the updating equation for $\mathbf{\Sigma}_t$. There, we take the 15-dimensional "Financial" dataset, assume the best-fitting distribution (Inverse t-Riesz, see Section 3.4), and consider the estimated model with scalar-BEKK and scalar-GAS dynamics for $\mathbf{\Sigma}_t$. The scalar-GAS model does not feature the HAR-type structure (only $\mathbf{\Sigma}_{t-1}$ goes into updating $\mathbf{\Sigma}_t$) and has constant d.o.f. parameters to make the comparison as fair as possible. The blue line corresponds to the realized variance (RV) of American Express (AXP), and the yellow and red lines correspond to the estimated expected conditional RV $(\widehat{\mathbf{\Sigma}}_t)_{3\,3}$ of the scalar-BEKK and scalar-GAS model, respectively. The gray line depicts the median RV of the 15 financial assets. On 07 August 2015, "reports emerged activist investor Value Act Capital had acquired an undisclosed stake in the credit card issuer with plans to pursue shareholder-friendly changes at the company."[9] This constitutes an idiosyncratic one-day event, which only influences the RV of AXP on the 07 August 2015, and should not affect the financial system. Consequently, the conditional

9. See https://www.nasdaq.com/articles/financial-sector-update-08072015-axphtgcalex-2015-08-07.

expected RV of AXP should not be updated strongly. However, the scalar-BEKK model cannot capture this, as it updates the RV too strongly. This overestimation takes a long time to revert back downwards. In contrast, the GAS model takes into account that the entire RC matrix on 07 August was not a tail event, as only AXP had a tail-realization and updates the conditional expected RV of AXP only slightly upwards. The ability to take into account the *entire* RC in order to update the individual constituent time-series is also visible in the more complex updating dynamics of the scalar-GAS model in the relatively calm period from 10 August to 14 August. The median RV increased relatively much from 11 to 12 August. The GAS model takes this into account and updates upwards its expected RV of AXP from 12 August to 13 August, even while AXP's RV dropped a little. The BEKK model on the other hand just incorporates the decrease in RV of AXP and updates its expected RV for AXP downwards accordingly.

In general, for fat-tailed distributions, GAS models inherently down-weight the impact of extreme realizations of the RCs on the updating process since extreme realizations are less unexpected by fat-tailed distributions and thus yield less extreme score realizations. This type of modeling behavior has also been advocated for in the literature on modeling RCs by e.g. Bollerslev, Patton, and Quaedvlieg (2018), whose *dynamic attenuation model* down-weights the impact of extreme RC realizations by incorporating the fact that they are more inaccurate estimates of integrated covariance.

In-Sample Fit 3.4.4

For the in-sample fit comparison we add two competitor models, namely those of the only other two papers featuring GAS models for RCs. These are the model by Gorgi et al. (2019) based on the Wishart distribution and the one by Opschoor et al. (2018) based on the F distribution. Note that our model does not model the daily return vectors, but both of these competitors do. In order to keep comparability we drop the daily return vector equations from their models and only take their RC specification. Furthermore, note that the main difference to our model is that Gorgi et al. (2019) and Opschoor et al. (2018) take the Wishart distribution FIM to scale the score. Thus, their models are nested by our restricted GAS model. That is, if we impose the restriction that $\alpha_2 = 0$ in equation (3.7), we arrive at their models.[10]

Table 3.5 displays the log-likelihood values of our estimated restricted GAS model, i.e. the model where the d.o.f.s are restricted to be constant, for the different datasets and distributions. The Bayes information criterion (BIC) rankings are

10. Actually the model of Gorgi et al. (2019) does not feature the HAR structure, but we impose it for their model to keep the comparison fair.

Assets:	Random	Mining	Random	Finance	Random	Manuf.
#Assets:	5	6	10	15	25	25
Wishart (GHJK)	-14742	-9977	1784	64253	349331	348292
Wishart	-13984	-8426	5831	83409	384626	382252
Riesz	-10915	-4285	18362	112218	462677	463780
Inv.Wishart	-8228	6367	46776	169822	634566	645663
Inv.Riesz	-6340	8319	52720	182317	658765	669844
t-Wishart	-3290	8612	39554	161069	496899	495073
t-Riesz	-1392	10657	48182	184176	554587	554640
Inv.t-Wishart	1510	15169	63864	216533	690170	702292
Inv.t-Riesz	3068	16519	69474	228078	708591	721055
F (OJLV)	-4231	5197	44130	161781	618250	626190
F	-3076	7132	49361	178930	650237	657571
F-Riesz	2691	15007	68534	213218	716322	726208
Inv.F-Riesz	1877	14259	66107	209555	710940	719874

Table 3.5: Log-likelihood values for the estimated restricted GAS model where d.o.f. parameters are restricted to be constant. As competitor models we have (GHJK), which resembles the model of Gorgi et al. (2019), and (OJLV) resembles the one of Opschoor et al. (2018). The background shades are to be read column-wise, with the lowest log-likelihood value being shaded black and the highest one being shaded white, with a linear scaling in between.

equal to those of the log-likelihood since the BIC penalty term for the number of parameters is dominated by the number of parameters in $\mathbf{\Sigma}$, which is common to all distributions and is of order p^2.

The first thing to note is that our restricted GAS model, i.e. including the a_2 parameter, assuming the Wishart and F distribution do indeed improve the fit of significantly over the models of Gorgi et al. (2019) and Opschoor et al. (2018) ($a_2 = 0$). When comparing the fit across distributions, note that nested distributions[11] must have a lower estimated log-likelihood values than the nesting ones. However, it is not clear if the differences are statistically significant and how non-nested distributions, e.g. the (Inverse) F-Riesz versus the (Inverse) t-Riesz distributions compare to each other. The Inverse t-Riesz and the F-Riesz distributions emerge as the clear winners. The former wins the four lower-dimensional datasets, and the latter wins the two 25-dimensional datasets. Both distributions are very close in terms of fitted log-likelihood values, thus the difference in terms of fit between tail-homogeneity (Inverse t-Riesz) and tail-heterogeneity (F-Riesz) is very slim. The gap favoring the Inverse t-Riesz distribution is larger for the

11. See Figure 2.6.

Assets:	Random	Mining	Random	Finance	Random	Manuf.
#Assets:	5	6	10	15	25	25
Wishart	-13814	-8409	6501	83541	386600	387104
Riesz	-10741	-4230	18813	112455	465362	469443
Inv.Wishart	-6979	6425	46776	172471	635234	646137
Inv.Riesz	-6053	8342	52727	182444	658864	669930
t-Wishart	-3113	8870	40283	162996	500667	501338
t-Riesz	-1241	10760	48554	184606	557429	559547
Inv.t-Wishart	1641	15351	64294	217966	692004	704941
Inv.t-Riesz	3157	16590	69656	228469	709889	722258
F	-3030	7213	49563	180031	651463	658949
F-Riesz	2776	15044	68572	213372	716891	726828
Inv.F-Riesz	2014	14356	66405	210311	712416	721365

Table 3.6: Log-likelihood values for the estimated GAS models with time-varying d.o.f. parameters and different datasets. The background shades are to be read column-wise, with the lowest log-likelihood value being shaded black and the highest one being shaded white, with a linear gray-scaling in between. Largest values in red.

industry-specific Mining and Finance datasets. This aligns with our economic intuition that having homogeneous tails is beneficial in more homogeneous datasets. The F-Riesz distribution has the advantage for both high-dimensional (25 assets) datasets, suggesting that in higher dimensions, the tail-homogeneity imposed by the Inverse t-Riesz distribution could be too restrictive. The Riesz and its special case, the Wishart distribution, are unequivocally the worst-fitting distributions.

Now we return to our unrestricted GAS model, i.e. the one where the d.o.f. parameters are assumed to be time varying as well. Table 3.6 displays the corresponding log-likelihood values. Comparing this table to the one for our restricted GAS model, it becomes clear that the likelihood-gains from assuming time-varying d.o.f. parameters are very small. Furthermore, the ranking across distributions stays the same as for the restricted GAS model. Thus, it is not that some distributions gain more from time-varying d.o.f.s than others. This is surprising since one could think that the worst fitting distributions could gain more from dynamically changing their shape via time-varying d.o.f. parameters than those distributions whose shape fits the RC data already quite well.

Additionally we see that the distribution ranking and relative distances between distributions are also very similar to those of the previous chapter's simple scalar-BEKK parameter updating process (compare Table 2.8). This indicates that adding the most flexible dynamic parameter specification with GAS dynamics in both the $\mathbf{\Sigma}_t$ and d.o.f. parameter recursions does not influence the ranking in fit

across distributions. It is noteworthy that the Inverse t-Wishart distribution with only two d.o.f. parameters[12] provides a reasonably good fit across datasets and gets closer to the best-fitting distributions with increasing cross-sectional dimension. It is thus a natural candidate for empirical applications on vast-dimensional (say 50 or more assets) RCs, where, for the Riesz-type distributions, the number of d.o.f. recursion parameters (of order p) becomes problematic for numerical maximum likelihood estimation.

Although the likelihood gains from time-varying d.o.f.s are rather small we still want to test statistically whether the data supports the GAS models with the d.o.f. parameters constraint to be constant. To this end we perform the simple likelihood-ratio test with the test statistic

$$\lambda_{LR} = -2 \left(\sum_{t=1}^{T} \log p_{\mathcal{D}} \left(\boldsymbol{\Sigma}_t(\widehat{\boldsymbol{\vartheta}}_{\mathcal{D}}^R), \boldsymbol{\theta}_{\mathcal{D},t}(\widehat{\boldsymbol{\vartheta}}_{\mathcal{D}}^R) \right) - \sum_{t=1}^{T} \log p_{\mathcal{D}} \left(\boldsymbol{\Sigma}_t(\widehat{\boldsymbol{\vartheta}}_{\mathcal{D}}), \boldsymbol{\theta}_{\mathcal{D},t}(\widehat{\boldsymbol{\vartheta}}_{\mathcal{D}}) \right) \right),$$

where $\widehat{\boldsymbol{\vartheta}}_{\mathcal{D}}^R$ is the parameter vector with the non-intercept d.o.f. recursion parameters restricted to be equal to zero. As $T \to \infty$,

$$\lambda_{LR} \sim \chi_d^2,$$

where $d = \dim(\boldsymbol{\vartheta}_{\mathcal{D}}) - \dim(\boldsymbol{\vartheta}_{\mathcal{D}}^R)$ is the difference of dimensionality between the restricted and unrestricted parameter vector. Across all distributions and datasets the median (largest) p-value we obtain equals 0 (0.0009) up to computer precision. Thus, even though the likelihood gains are small from assuming time-varying d.o.f. parameters, we can still reject the models where the d.o.f. parameters are constant. This provides strong evidence in favor of time-varying d.o.f. parameters.

Next, we are interested in the time-evolution of the d.o.f. parameters to determine if it makes sense economically. For this, we present the estimated n_t and ν_t of all Wishart-based distributions from our estimated GAS model for the ten-dimensional dataset in Figure 3.3. The time-series are normed to their first values to make them comparable. We see substantial time-variation, supporting the importance of time-varying d.o.f. parameters. Furthermore, we see that the estimated $\hat{\nu}_t$ does indeed drop down substantially, making the distributions "wider" and more "fat-tailed", in economically sensible time-periods, for example, during the 2008 financial crisis, the 2015 European sovereign debt crisis and the 2020 COVID-19 crises.

12. In total, the GAS models with Wishart-type distributions contain only 10 to 15 parameters (depending on whether they feature one or two d.o.f. parameters) that have to be estimated via numerical maximum likelihood.

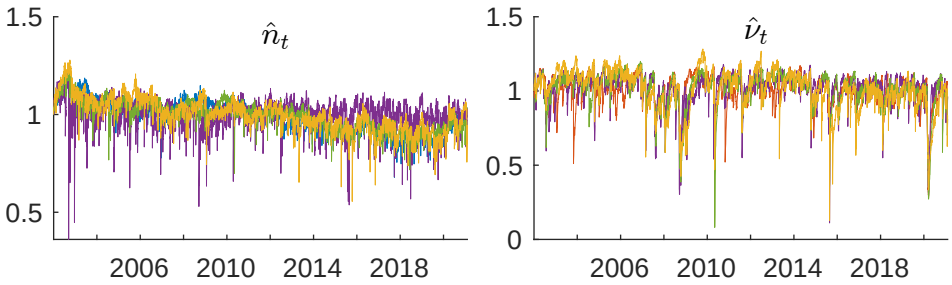

Figure 3.3: Estimated dynamic d.o.f. parameters $\hat{n}_t$ and $\hat{\nu}_t$ from estimation of our GAS model describe in Section 3.3 for all distributions for the ten-dimensional dataset described in 3.4.1. All time-series are normed to the first estimated d.o.f. The color coding is: Wishart, Inv.Wishart, t-Wishart, Inv.t-Wishart, F. The Riesz-type distributions have been left out for clarity of the plot, but their average d.o.f. vectors behave very similarly to their Wishart-based counterpart d.o.f. parameters.

Next, we look at the model parameter estimates. Table 3.7 shows the estimated score parameters $\hat{a}_1$ and $\hat{a}_2$. All are highly significant. It is interesting to note that the estimated parameters are similar across distributions. Furthermore, we see that the higher the cross-sectional dimension p, the smaller the estimated parameters, which, according to Pakel et al. (2021), is estimation bias caused by the targeting estimator for $\boldsymbol{\Xi}$. The estimated GARCH (b_1, b_2 and b_3) and persistence ($b_1 + b_2 + b_3$) parameters are highly significant as well and very close to each other across cross-sectional dimensions and distributions, with the lowest (highest) estimated persistence amounting to 0.983 (0.9994). See Table 3.13 in Section of the appendix. This aligns with the empirical observation that realized (co)variances exhibit persistent time-series behavior.

For the estimated score parameters of the d.o.f. parameter recursions, see Table 3.8. We note that even with our "bottom-up" estimation method described in Section 3.4.2 we can, in a few cases, not come up with sensible estimates. However, in the overwhelming majority of cases, we obtain sensible estimates, that is, the estimated d.o.f. recursions exhibit high persistence, as for the $\boldsymbol{\Sigma}_t$ recursion parameters. For these cases, estimated d.o.f. recursion parameters, that is, the intercepts, the score parameters, and the GARCH parameters are significant at the 1% level. We thus think that insignificant results are only due to the technical difficulty of obtaining reasonable estimates. We tried multiple solvers, solver specifications, and starting points but have yet to develop a robust estimation method. As we can see in Table 3.8, the F, F-Riesz, and Inverse F-Riesz distributions are the most difficult to estimate as the solver does not come up with sensible estimates

Assets:	Random	Mining	Random	Finance	Random	Manuf.
#Assets:	5	6	10	15	25	25
			$\hat{a}_1 \times 100$			
Wishart	1.610	1.193	0.872	0.517	0.340	0.335
Riesz	1.135	0.809	0.588	0.336	0.236	0.218
Inv.Wishart	0.197	0.541	0.332	0.157	0.167	0.166
Inv.Riesz	0.121	0.480	0.299	0.113	0.144	0.155
t-Wishart	0.998	0.767	0.651	0.419	0.341	0.325
t-Riesz	0.813	0.663	0.527	0.292	0.247	0.233
Inv.t-Wishart	0.711	0.575	0.456	0.302	0.220	0.210
Inv.t-Riesz	0.636	0.527	0.398	0.258	0.192	0.185
F	0.890	0.704	0.536	0.332	0.248	0.235
F-Riesz	0.477	0.371	0.300	0.207	0.168	0.156
Inv.F-Riesz	0.548	0.416	0.344	0.221	0.169	0.166
			$\hat{a}_2 \times 100$			
Wishart	0.607	0.517	0.249	0.204	0.085	0.080
Riesz	0.791	0.630	0.346	0.251	0.103	0.097
Inv.Wishart	0.328	0.418	0.212	0.149	0.067	0.071
Inv.Riesz	0.399	0.435	0.237	0.149	0.074	0.076
t-Wishart	4.249	4.198	3.370	3.703	2.146	2.221
t-Riesz	4.084	4.285	3.283	3.950	2.238	2.184
Inv.t-Wishart	3.598	3.474	2.851	3.192	1.600	1.756
Inv.t-Riesz	3.409	3.231	2.768	3.134	1.493	1.686
F	0.623	0.477	0.276	0.179	0.089	0.086
F-Riesz	1.183	0.915	0.490	0.293	0.141	0.138
Inv.F-Riesz	1.121	0.884	0.478	0.280	0.133	0.129

Table 3.7: Estimated scaled score parameters $\hat{a}_1 \times 100$ and $\hat{a}_2 \times 100$. All estimates are highly significant. The median (smallest) t-statistic for $\hat{a}_1$ is 114 (68), and for $\hat{a}_2$ is 150 (42).

for the d.o.f.-recursions most often for these distributions. In our opinion, this is an argument in favor of the (Inverse) t-Riesz or (Inverse) t-Wishart distribution, which are analytically (see the score and FIM derivations in the appendix) and numerically more tractable.

Finally, we estimated a more general GAS model for the Wishart distribution for the five-dimensional "Random" dataset, where we generalized the dynamics in the $\mathbf{\Sigma}_t$-equation to diagonal matrices instead of scalars. The Wishart, providing the worst fit for all distributions, should benefit most from such a generalization.

Assets:	Random	Mining	Random	Finance	Random	Manuf.
#Assets:	5	6	10	15	25	25
			$\widehat{a^{\mathbf{n}}} \times 100$ ($\widehat{a^{n}} \times 100$)			
Wishart	0.827	0.707	1.633	0.598	1.002	1.235
Riesz	0.863	1.096	1.292	0.507	0.691	0.917
Inv.Wishart	-	-	-	-	-	-
Inv.Riesz	-	-	-	-	-	-
t-Wishart	6.735	8.149	6.658	12.561	8.098	9.925
t-Riesz	1.533	1.500	0.955	1.905	0.829	1.002
Inv.t-Wishart	1.591	1.892	1.076	0.987	1.323	1.183
Inv.t-Riesz	1.824	1.904	0.992	1.093	1.295	1.223
F	13.723	5.032	0.570	0.000	2.893	0.003
F-Riesz	7.017	2.612	2.647	0.842	1.535	1.524
Inv.F-Riesz	2.549	5.600	0.273	0.001	3.422	3.325
			$\widehat{a^{\boldsymbol{\nu}}} \times 100$ ($\widehat{a^{\nu}} \times 100$)			
Wishart	-	-	-	-	-	-
Riesz	-	-	-	-	-	-
Inv.Wishart	3.077	3.595	2.794	2.408	1.230	3.486
Inv.Riesz	6.687	59.902	5.449	1.677	2.503	4.108
t-Wishart	1.071	1.078	1.548	1.816	1.274	2.108
t-Riesz	1.347	1.025	1.396	1.654	1.049	2.392
Inv.t-Wishart	12.924	12.572	14.124	20.952	16.173	19.947
Inv.t-Riesz	8.756	7.097	5.356	1.917	0.645	0.456
F	14.264	11.900	13.637	17.600	10.968	10.406
F-Riesz	9.910	7.857	0.934	4.683	1.093	2.213
Inv.F-Riesz	9.924	9.946	9.327	8.631	6.056	6.188

Table 3.8: Estimated score parameter $\widehat{a^{\mathbf{n}}} \times 100$ and $\widehat{a^{\boldsymbol{\nu}}} \times 100$ ($\widehat{a^{n}} \times 100$ and $\widehat{a^{\nu}} \times 100$ for Wishart-type distributions). No background color indicates significance at the 1% level, light gray and mid-light gray indicate significance at the 5% and 10% level, dark gray indicates insignificance.

We wanted to investigate if the significance of our dynamic d.o.f. specification comes from a too restrictive specification in the mean-shifting process $\mathbf{\Sigma}_t$. This does not seem to be the case. In each diagonal parameter matrix, the estimated parameters are very close to each other, and the d.o.f. recursion parameters stayed highly significant.

Assets:	Random	Mining	Random	Finance	Random	Manuf.
#Assets:	5	6	10	15	25	25
			BEKK			
Inv.t-Riesz	3.70	10.67	-10.86	-53.90	-75.18	-110.57
F-Riesz	3.75	10.85	-10.81	-48.56	-76.43	-109.33
			Restricted GAS			
Inv.t-Riesz	3.13	10.05	-12.52	-57.12	-85.48	-118.53
F-Riesz	3.28	10.46	-12.12	-53.44	-86.20	-119.24
			GAS			
Inv.t-Riesz	3.11	10.05	-12.55	-57.19	-85.57	-118.56
F-Riesz	3.27	10.45	-12.13	-53.51	-86.23	-119.26

Table 3.9: Average of log-score loss, $-p_{\mathcal{D}}(\mathbf{R}_{t+1}|\widehat{\boldsymbol{\Sigma}}_{t+1}, \hat{\boldsymbol{\theta}}_{\mathcal{D},t+1})$, for the entire forecasting window, where each model is re-estimated every ten trading days. 90% model confidence sets in red. BEKK refers to the scalar-BEKK model from Chapter 2 (see equation (2.23)), Restricted GAS is our GAS model, where the d.o.f. parameters are constant, GAS is our full GAS model.

Out-of-Sample Forecasting Performance 3.4.5

We re-estimate the models every ten trading days on a rolling window of 1250 observations (roughly five years of data). The forecasting window starts at observation 1251 (18 December 2006) and ends on 05 February 2021.

We take the log-score as the natural loss function since, as Hansen and Dumitrescu (2022) show, coherency between the estimation criterion, that is, the loss function chosen to obtain the estimates, and the actual objective, i.e. the out-of sample quantity we want to forecast, is essential. We only consider the Inverse t-Riesz and the F-Riesz distribution to save computational resources, and since it is evident from the Chapter 2 and the in-sample analysis that these two distributions provide the best out-of-sample forecasting results. Table 3.9 contains the log-score one-day-ahead forecast losses for the entire forecasting window. The 90% model confidence sets (MCS, see Hansen, Lunde, and Nason, 2011) are shaded in gray.[13]

We see that our unrestricted GAS model with dynamic d.o.f. parameters is the clear winner in this comparison as it is the only member of the MCS for all datasets except the Mining dataset, where also the GAS model with d.o.f. parameters restricted to be constant is a member of the MCS. The scalar-BEKK model, with its simple mean shifting dynamics, is rejected out-of-sample by the data. Further-

13. For calculation of the MCS, we choose 5000 stationary bootstrap replications with block length set equal to the maximum number of consecutive significant partial autocorrelations of the losses. We use the MFE toolbox by Kevin Sheppard for MCS calculation.

Assets:	Random	Mining	Random	Finance	Random	Manuf.
#Assets:	5	6	10	15	25	25
			BEKK			
Inv.t-Riesz	1.368	2.046	0.736	1.403	0.687	0.836
F-Riesz	1.369	2.042	0.735	1.375	0.678	0.824
			Restricted GAS			
Inv.t-Riesz	1.368	2.045	0.738	1.441	0.698	0.841
F-Riesz	1.391	2.073	0.758	1.468	0.705	0.859
			GAS			
Inv.t-Riesz	1.368	2.045	0.738	1.440	0.698	0.841
F-Riesz	1.392	2.073	0.757	1.472	0.704	0.859

Table 3.10: Average GMVP loss, $RV_{GMPV,t+1}$, for the entire forecasting window, where each model is re-estimated every ten trading days. 90% model confidence sets in red. BEKK refers to the scalar-BEKK model from Chapter 2 (see equation (2.23)), Restricted GAS is our GAS model, where the d.o.f. parameters are constant, GAS is our full GAS model.

more, it is apparent that for the datasets with smaller cross-sectional dimensions p, the Inverse t-Riesz distribution provides the best forecasts. In contrast, for the 25-dimensional dataset, the F-Riesz distribution is the sole member of the MCS. This aligns with the in-sample fit (see Table 3.6). We conjecture that with increasing cross-sectional dimension, the assumption of tail-homogeneity (Inverse t-Riesz) might become too restrictive since it imposes too much dependence between the different elements in the RCs.

Finally, we consider an economic loss function by computing the realized variance of the forecasted global minimum variance portfolio (GMVP). That is, suppose we want to minimize the portfolio variance given our forecasted RC, $\widehat{\mathbf{\Sigma}}_{t+1}$, by choosing the optimal weights,

$$\min_{\mathbf{w}} \mathbf{w}^\top \widehat{\mathbf{\Sigma}}_{t+1} \mathbf{w}, \qquad \text{given } \mathbf{w}^\top \mathbf{1} = 1,$$

where $\mathbf{1}$ is an p-dimensional vector of ones. The optimal weights that solve this minimization problem are given by

$$\hat{\mathbf{w}} = \frac{\widehat{\mathbf{\Sigma}}_{t+1}\mathbf{1}}{(\mathbf{1}^\top \widehat{\mathbf{\Sigma}}_{t+1}^{-1}\mathbf{1})}.$$

Now for each forecast the GMVP loss is equal to the actually realized variance if

an investor had chosen these weights,

$$RV_{GMPV,t+1} = \hat{\mathbf{w}}^\top \mathbf{R}_{t+1} \hat{\mathbf{w}}.$$

Table 3.10 contains the average GMVP loss for all datasets and models. We see that for the three datasets of dimension 10 or lower, there emerges no clear winner across time-series specifications since the BEKK, Restricted GAS and GAS all are members of the MCS. Comparing distributions, for these three datasets the Inverse t-Riesz distribution performs better since it is in the MCS for all three times versus one time for the F-Riesz. For the higher dimensional dataset the picture is more clear. Here the F-Riesz with BEKK dynamics is the sole member of the MCS. However, due to the nonlinear formula of the GMVP loss function it is difficult to speculate which distributional features might lead to the advantage of the F-Riesz for high-dimensional datasets.

3.5 Conclusion

In this paper, we propose a novel GAS model for time-series of realized covariance matrices. This model features a time-varying expected value matrix and allows for time-varying degree of freedom parameters, thus making all distribution parameters dynamic. We derive the model for all hitherto used probability distributions in order to understand which distributions' score yields the best fit and forecasting performance. In particular we derive all necessary scores and Fisher information matrices for all probability distributions hitherto used for RCs. Our model features an easy-to-compute closed-form formula for the scaled scores, making the model computationally feasible.

Empirically, we find evidence for time-varying degree of freedom parameters. We show that their time-variation is economically interpretable and is in line with the notion that realized covariance matrices become more fat-tailed during crises periods. Our GAS model with time-varying d.o.f. parameters performs best in- and out-of-sample against the competitor with constant d.o.f. parameters using the log-likelihood as the loss-function. Using the economically relevant global minimum variance portfolio loss function we see no advantage for our GAS model. Furthermore, we confirm the finding from the last chapter that the Inverse t-Riesz and F-Riesz distributions exhibit the best in-sample fit for time-series of RCs. No clear winner between these two distributions can be made out, but there seems to be a relation to the cross-sectional dimension. In particular, the F-Riesz is more often preferred in higher-dimensional settings, whereas the t-Riesz outperforms in low- to medium-dimensional settings. Furthermore, we show using an illustrative example that GAS models imply richer dynamic updating properties than traditional models, even for scalar dynamics.

Appendix 3.6

Preliminaries 3.6.1

Some Matrix Relations

For matrices $\mathbf{W}$, $\mathbf{X}$, $\mathbf{Y}$ and $\mathbf{Z}$ with appropriate dimensions we have (Magnus and Neudecker (2019), p.12, p. 35)

$$\operatorname{vec}(\mathbf{XYZ}) = (\mathbf{Z}^\top \otimes \mathbf{X})\operatorname{vec}(\mathbf{Y}), \tag{3.12}$$

$$\operatorname{tr}(\mathbf{XYZ}) = \operatorname{tr}(\mathbf{YZX}) = \operatorname{tr}(\mathbf{ZXY}), \tag{3.13}$$

$$\operatorname{tr}(\mathbf{X}^\top\mathbf{Y}) = \operatorname{vec}(\mathbf{X})^\top\operatorname{vec}(\mathbf{Y}) \text{ and} \tag{3.14}$$

$$\operatorname{tr}(\mathbf{WXYZ}) = \operatorname{vec}(\mathbf{W})^\top\operatorname{vec}(\mathbf{XYZ}) = \operatorname{vec}(\mathbf{W})^\top(\mathbf{X} \otimes \mathbf{Z})\operatorname{vec}(\mathbf{Y}), \tag{3.15}$$

where for the last equality we used (3.14) and (3.12).

Special Functions

Definition 3.6.1. *Let* $\mathbf{X}$ *be a square matrix, then the function* $\blacktriangle(\mathbf{X})$ *returns a lower triangular matrix by setting all elements of* $\mathbf{X}$ *above the main diagonal equal to zero,*

$$\blacktriangle_{ij}(\mathbf{X}) = \begin{cases} \mathbf{X}_{ij} \text{ for } i \geq j \text{ and} \\ 0 \text{ for } i < j. \end{cases}$$

Definition 3.6.2. *Let* $\mathbf{X}$ *be a square matrix, then the function* $\bar{\blacktriangle}(\mathbf{X})$ *returns a lower triangular matrix by setting all elements of* $\mathbf{X}$ *above the main diagonal equal to zero and halving all elements on the main diagonal,*

$$\bar{\blacktriangle}_{ij}(\mathbf{X}) = \begin{cases} \mathbf{X}_{ij} \text{ for } i > j, \\ \frac{1}{2}\mathbf{X}_{ii} \text{ for } i = j \text{ and} \\ 0 \text{ for } i < j. \end{cases}$$

That is, $\bar{\blacktriangle}(\mathbf{X}) = \blacktriangle\left(\mathbf{X} - \frac{1}{2}\mathbf{I} \odot \mathbf{X}\right)$.

Definition 3.6.3. *Let* $\mathbf{X}$ *be a square matrix, then the function* $\tilde{\blacktriangle}(\mathbf{X})$ *returns a symmetric matrix by setting all elements above the main diagonal equal to the elements of* $\mathbf{X}$ *below the diagonal,*

$$\tilde{\blacktriangle}_{ij}(\mathbf{X}) = \begin{cases} \mathbf{X}_{ij} \text{ for } i \geq j, \\ \mathbf{X}_{ji} \text{ for } i < j. \end{cases}$$

That is, $\tilde{\blacktriangle}(\mathbf{X}) = \bar{\blacktriangle}(\mathbf{X}) + \bar{\blacktriangle}(\mathbf{X})^\top$.

Duplication, Elimination, and Commutation Matrices

As a reference, see Lütkepohl (2005), A.12.2. $\mathbf{G}_p$ denotes the *duplication matrix* defined by

$$\text{vec}(\mathbf{X}) = \mathbf{G}_p \text{vech}(\mathbf{X}), \tag{3.16}$$

where $\mathbf{X}$ is an arbitrary *symmetric* $p \times p$ matrix.

For symmetric $\mathbf{X}$, the *duplication matrix* $\mathbf{G}$ is unique. However, the so-called *elimination matrix*, which converts vec($\mathbf{X}$) to vech($\mathbf{X}$), is not unique (since for every lower-diagonal element of $\mathbf{X}$, we can take a fraction c of the corresponding upper- and a fraction $1 - c$ of the lower-diagonal element of $\mathbf{X}$). One possible choice is the Moore-Penrose inverse of $\mathbf{G}_p$,

$$\mathbf{G}_p^+ = (\mathbf{G}_p^\top \mathbf{G}_p)^{-1} \mathbf{G}_p^\top, \tag{3.17}$$

for which obviously

$$\mathbf{G}_p^+ \text{vec}(\mathbf{X}) = \mathbf{G}_p^+ \mathbf{G}_p \text{vech}(\mathbf{X}) = \text{vech}(\mathbf{X}). \tag{3.18}$$

Another possible choice is the canonical elimination matrix $\mathbf{F}_p$, which sets the aforementioned fraction $c = 0$.

For *lower-triangular* $p \times p$ matrix $\mathbf{Y}$ Magnus and Neudecker (1980) note (Lemma 3.3 (i)) that the *unique elimination* and *duplication matrices* are given by

$$\text{vec}(\mathbf{Y}) = \mathbf{F}_p^\top \text{vech}(\mathbf{Y}) \text{ and} \tag{3.19}$$

$$\text{vech}(\mathbf{Y}) = \mathbf{F}_p \text{vec}(\mathbf{Y}). \tag{3.20}$$

$\mathbf{K}_{pq}$ denotes the *commutation matrix* defined by

$$\text{vec}(\mathbf{Z}^\top) = \mathbf{K}_{pq} \text{vec}(\mathbf{Z}), \tag{3.21}$$

for arbitrary $p \times q$ matrix $\mathbf{Z}$. Note that the exact size and structure of $\mathbf{G}_p$, $\mathbf{F}_p$ and $\mathbf{K}_{pq}$ depends on the size of $\mathbf{X}$, but for better readability, we choose to omit the size-indicating subscripts in the rest of this paper.

Magnus and Neudecker (2019) show (Theorem 3.12) that

$$(\mathbf{I} + \mathbf{K}) = 2\mathbf{G}\mathbf{G}^+. \tag{3.22}$$

Furthermore, it holds that

$$\mathbf{G}\mathbf{G}^+ = \mathbf{G}(\mathbf{G}^\top\mathbf{G})^{-1}\mathbf{G}^\top = (\mathbf{G}(\mathbf{G}^\top\mathbf{G})^{-1}\mathbf{G}^\top)^\top = (\mathbf{G}\mathbf{G}^+)^\top \tag{3.23}$$

and

$$(\mathbf{G}^+)^\top\mathbf{G}^\top\text{vec}(\mathbf{X}) = ((\mathbf{G}^\top\mathbf{G})^{-1}\mathbf{G}^\top)^\top\mathbf{G}^\top\mathbf{G}\text{vech}(\mathbf{X}) \tag{3.24}$$

$$= \mathbf{G}\text{vech}(\mathbf{X}) = \text{vec}(\mathbf{X}). \tag{3.25}$$

For square matrix $\mathbf{X}$ we have (see Magnus and Neudecker, 2019, p.57)

$$\mathbf{G}\mathbf{G}^+\mathbf{X}^{\otimes 2} = \mathbf{X}^{\otimes 2}\mathbf{G}\mathbf{G}^+. \tag{3.26}$$

For nonsingular matrix $\mathbf{X}$ it holds that (see Lütkepohl, 2005, p. 664 or Magnus and Neudecker (2019), Theorem 3.13)

$$(\mathbf{G}^\top\mathbf{X}^{\otimes 2}\mathbf{G})^{-1} = \mathbf{G}^+\mathbf{X}^{-\otimes 2}(\mathbf{G}^+)^\top. \tag{3.27}$$

Lemma 3.6.1. *For scalar α we have*

$$(\mathbf{G}^\top(\mathbf{\Sigma}^{\otimes 2} + \alpha\text{vec}^2(\mathbf{\Sigma}))\mathbf{G})^{-1} = \mathbf{G}^+\left(\mathbf{\Sigma}^{-\otimes 2} + \frac{\alpha}{1+\alpha p}\text{vec}^2(\mathbf{\Sigma}^{-1})\right)\mathbf{G}^+.$$

Proof. Using Magnus and Neudecker (1980) Lemma 4.4 (i) and Lemma 4.7 (iv), we have

$$\begin{aligned}
(\mathbf{F}(\mathbf{\Sigma}^{\otimes 2} + \alpha\text{vec}^2(\mathbf{\Sigma})))^{-1} &= (\mathbf{G}^+\mathbf{G}\mathbf{F}(\mathbf{\Sigma}^{\otimes 2} + \alpha\text{vec}^2(\mathbf{\Sigma}))\mathbf{G})^{-1} \\
&= (\mathbf{G}^+\mathbf{G}\mathbf{F}(\mathbf{\Sigma}^{\otimes 2} + \alpha\text{vec}^2(\mathbf{\Sigma}))\mathbf{G})^{-1} \\
&= (\mathbf{G}^+(\mathbf{\Sigma}^{\otimes 2} + \alpha\text{vec}^2(\mathbf{\Sigma}))\mathbf{G})^{-1} \\
&= (\mathbf{G}^\top(\mathbf{\Sigma}^{\otimes 2} + \alpha\text{vec}^2(\mathbf{\Sigma}))\mathbf{G})^{-1}\mathbf{G}^\top\mathbf{G} \qquad (3.28) \\
&= \mathbf{F}\left(\mathbf{\Sigma}^{-\otimes 2} + \frac{\alpha}{1+\alpha p}\text{vec}^2\left(\mathbf{\Sigma}^{-1}\right)\right)\mathbf{G} \\
&= \mathbf{G}^+\left(\mathbf{\Sigma}^{-\otimes 2} + \frac{\alpha}{1+\alpha p}\text{vec}^2\left(\mathbf{\Sigma}^{-1}\right)\right)\mathbf{G}. \qquad (3.29)
\end{aligned}$$

Then

$$(3.28) = (3.29)$$
$$\Leftrightarrow$$
$$(\mathbf{G}^\top(\boldsymbol{\Sigma}^{\otimes 2} + \alpha \text{vec}^2(\boldsymbol{\Sigma}))\mathbf{G})^{-1} = \mathbf{G}^+ \left(\boldsymbol{\Sigma}^{-\otimes 2} + \frac{\alpha}{1+\alpha p}\text{vec}^2(\boldsymbol{\Sigma}^{-1})\right)\mathbf{G}^+.$$

□

Matrix Derivatives

We first want to clarify exactly how we take the derivative of a (matrix) function of a matrix of variables since there is no inherently "right" way to arrange the individual univariate derivatives. Throughout, we follow Magnus (2010), who recommend for an $m \times p$ matrix function $\mathbf{F}$ of an $n \times q$ matrix of variables $\mathbf{X}$ to define the derivative as the $mp \times nq$ matrix

$$\frac{\partial \text{vec}(\mathbf{F}(\mathbf{X}))}{\partial \text{vec}(\mathbf{X})^\top}.$$

In this derivative matrix, for example, in its third row and second column, we have the derivative of the third element of $\text{vec}(\mathbf{F}(\mathbf{X}))$ with respect to the second element in $\text{vec}(\mathbf{X})$. Hence the column vector notation in the nominator ($\text{vec}(\mathbf{F}(\mathbf{X}))$) and the row vector notation in the denominator ($\text{vec}(\mathbf{X})^\top$).

For our scalar-valued log-likelihood function ($m = p = 1$), we have, in accordance with this definition, that its derivative with respect to $\boldsymbol{\Sigma}$ is a row vector,

$$\frac{\partial \log p_{\mathcal{D}}(\boldsymbol{\Sigma}, \boldsymbol{\theta}_{\mathcal{D}}|\mathbf{R})}{\partial \text{vec}(\boldsymbol{\Sigma})^\top} = \text{vec}(\triangle_{\mathcal{D}}^{\Sigma})^\top,$$

with

$$(\triangle_{\mathcal{D}}^{\Sigma})_{ij} = \frac{\partial \log p_{\mathcal{D}}(\boldsymbol{\Sigma}, \boldsymbol{\theta}_{\mathcal{D}}|\mathbf{R})}{\partial (\boldsymbol{\Sigma})_{ij}}, \tag{3.30}$$

where we omit the time subscript t for readability in this and the following section. However, we still have to consider that $\boldsymbol{\Sigma}$ is a symmetric matrix. That is, for $i \neq j$, an infinitesimally small change in $\boldsymbol{\Sigma}_{(i,j)}$, changes $\boldsymbol{\Sigma}_{(j,i)}$ by the same amount. To do so, we can use the chain rule,

$$\frac{\partial \log p_{\mathcal{D}}(\boldsymbol{\Sigma}, \boldsymbol{\theta}_{\mathcal{D}}|\mathbf{R})}{\partial \text{vech}(\boldsymbol{\Sigma})^\top} = \frac{\partial \log p_{\mathcal{D}}(\boldsymbol{\Sigma}, \boldsymbol{\theta}_{\mathcal{D}}|\mathbf{R})}{\partial \text{vec}(\boldsymbol{\Sigma})^\top} \frac{\partial \text{vec}(\boldsymbol{\Sigma})}{\partial \text{vech}(\boldsymbol{\Sigma})^\top}.$$

For symmetric $\mathbf{\Sigma}$ we have according to Lemma 3.6.2 (see p. 101) that

$$\frac{\partial \text{vec}\left(\mathbf{\Sigma}\right)}{\partial \text{vech}\left(\mathbf{\Sigma}\right)^\top} = \mathbf{G},$$

where $\mathbf{G}$ is the duplication matrix. Finally, it is usually the case that the score is defined as a column vector, such that we write

$$\nabla_{\mathcal{D}}^{\Sigma} = \left(\frac{\partial \log p_{\mathcal{D}}(\mathbf{\Sigma}, \boldsymbol{\theta}_{\mathcal{D}} | \mathbf{R})}{\partial \text{vech}(\mathbf{\Sigma})^\top} \right)^\top.$$

In summary, for a given distribution $\mathcal{D}$,

> $\nabla_{\mathcal{D}}^{\Sigma}(\nabla_{\mathcal{D}}^{\Omega})$ is the $p(p+1)/2 \times 1$ score *vector* w.r.t. $\mathbf{\Sigma}$ ($\mathbf{\Omega}$) taking symmetry into account, and
>
> $\triangle_{\mathcal{D}}^{\Sigma}(\triangle_{\mathcal{D}}^{\Omega})$ is the $p \times p$ score *matrix* w.r.t. $\mathbf{\Sigma}$ ($\mathbf{\Omega}$) *ignoring* symmetry (used in e.g. Theorem 3.3.1 and Table 3.1).

Lemma 3.6.2. *(Magnus and Neudecker, 1980, Lemma 3.8). Let $\mathbf{X}$ be a square matrix of variables. Then*

$$\frac{\partial \text{vec}(\mathbf{X})}{\partial \text{vech}(\mathbf{X})^\top} = \begin{cases} \mathbf{F}^\top, & \text{for lower triangular } \mathbf{X}, \\ \mathbf{G}, & \text{for symmetric } \mathbf{X}, \end{cases}$$

Lemma 3.6.3. *(Harville, 1997, p. 371). For non-singular symmetric matrix $\mathbf{X}$,*

$$\frac{\partial \text{vech}(\mathbf{X}^{-1})}{\partial \text{vech}(\mathbf{X})^\top} = -\mathbf{G}^+ \mathbf{X}^{-\otimes 2} \mathbf{G}.$$

Lemma 3.6.4. *(Lütkepohl, 1989, Lemma 1 and Murray, 2016, equation (15).) For lower triangular matrix $\mathbf{C}$ we have*

$$\frac{\partial \text{vech}(\mathbf{C})}{\partial \text{vech}(\mathbf{C}\mathbf{C}^\top)^\top} = \frac{1}{2}(\mathbf{G}^+(\mathbf{C} \otimes \mathbf{I})\mathbf{F}^\top)^{-1} = \mathbf{F}(\mathbf{I} \otimes \mathbf{C})\mathbf{Z}\mathbf{C}^{-\otimes 2}\mathbf{G},$$

where $\mathbf{Z}$ is a diagonal matrix defined such that for any square matrix $\mathbf{A}$, $\mathbf{Z}\text{vec}(\mathbf{A}) = \text{vec}(\bar{\blacksquare}(\mathbf{A}))$ with $\bar{\blacksquare}(\cdot)$ defined in Definition 3.6.2. In fact it can be shown that $\mathbf{Z} = \frac{1}{2}(\mathbf{G}\mathbf{F})^\top(\mathbf{G}\mathbf{F})$.

Lemma 3.6.5. *Let* $\mathbf{X}$ *be a non-singular symmetric matrix,* $\mathbf{C}$ *its lower Cholesky factor, and* $\mathbf{Y}$ *be a diagonal matrix. Then*

$$\begin{aligned}\frac{\partial \text{vec}(\mathbf{C}\mathbf{Y}\mathbf{C}^\top)}{\partial \text{vech}(\mathbf{X})^\top} &= \mathbf{G}\mathbf{G}^+(\mathbf{C}\mathbf{Y}\otimes\mathbf{I})\mathbf{F}^\top(\mathbf{G}^+(\mathbf{C}^{-\top}\otimes\mathbf{X}^{-1})\mathbf{F}^\top)^{-1}\mathbf{G}^+\mathbf{X}^{-\otimes 2}\mathbf{G} \\ &= \mathbf{G}\mathbf{G}^+(\mathbf{C}\mathbf{Y}\otimes\mathbf{I})\mathbf{F}^\top(\mathbf{G}^+(\mathbf{C}\otimes\mathbf{I})\mathbf{F}^\top)^{-1},\end{aligned}$$

$$\begin{aligned}\frac{\partial \text{vec}(\mathbf{C}\mathbf{Y}\mathbf{C}^\top)}{\partial \text{vech}(\mathbf{X}^{-1})^\top} &= -\mathbf{G}\mathbf{G}^+(\mathbf{C}\mathbf{Y}\otimes\mathbf{I})\mathbf{F}^\top(\mathbf{G}^+(\mathbf{C}\otimes\mathbf{I})\mathbf{F}^\top)^{-1}\mathbf{G}^+\mathbf{X}^{\otimes 2}\mathbf{G} \\ &= -\mathbf{G}\mathbf{G}^+(\mathbf{C}\mathbf{Y}\otimes\mathbf{I})\mathbf{F}^\top(\mathbf{G}^+(\mathbf{C}^{-\top}\otimes\mathbf{X}^{-1})\mathbf{F}^\top)^{-1},\end{aligned}$$

$$\begin{aligned}&\frac{\partial \text{vec}(\mathbf{C}^{-\top}\mathbf{Y}\mathbf{C}^{-1})}{\partial \text{vech}(\mathbf{X}^{-1})^\top} \\ &\quad= \mathbf{G}\mathbf{G}^+(\mathbf{C}^{-\top}\otimes\mathbf{C}^{-\top}\mathbf{Y}\mathbf{C}^{-1})\mathbf{F}^\top(\mathbf{G}^+(\mathbf{C}\otimes\mathbf{I})\mathbf{F}^\top)^{-1}\mathbf{G}^+\mathbf{X}^{\otimes 2}\mathbf{G} \\ &\quad= \mathbf{G}\mathbf{G}^+(\mathbf{C}^{-\top}\otimes\mathbf{C}^{-\top}\mathbf{Y}\mathbf{C}^{-1})\mathbf{F}^\top(\mathbf{G}^+(\mathbf{C}^{-\top}\otimes\mathbf{X}^{-1})\mathbf{F}^\top)^{-1},\end{aligned}$$

$$\begin{aligned}&\frac{\partial \text{vec}(\mathbf{C}^{-\top}\mathbf{Y}\mathbf{C}^{-1})}{\partial \text{vech}(\mathbf{X})^\top} \\ &\quad= -\mathbf{G}\mathbf{G}^+(\mathbf{C}^{-\top}\otimes\mathbf{C}^{-\top}\mathbf{Y}\mathbf{C}^{-1})\mathbf{F}^\top(\mathbf{G}^+(\mathbf{C}^{-\top}\otimes\mathbf{X}^{-1})\mathbf{F}^\top)^{-1}\mathbf{G}^+\mathbf{X}^{-\otimes 2}\mathbf{G} \\ &\quad= -\mathbf{G}\mathbf{G}^+(\mathbf{C}^{-\top}\otimes\mathbf{C}^{-\top}\mathbf{Y}\mathbf{C}^{-1})\mathbf{F}^\top(\mathbf{G}^+(\mathbf{C}\otimes\mathbf{I})\mathbf{F}^\top)^{-1}.\end{aligned}$$

Proof. We have

$$\begin{aligned}\text{dvec}(\mathbf{C}\mathbf{Y}\mathbf{C}^\top) &= \text{vec}(\text{d}\mathbf{C}\mathbf{Y}\mathbf{C}^\top) + \text{vec}(\mathbf{C}\mathbf{Y}\text{d}\mathbf{C}^\top) \\ &= (\mathbf{I}+\mathbf{K}_{pp})\text{vec}(\text{d}\mathbf{C}\mathbf{Y}\mathbf{C}^\top) \\ &= (\mathbf{I}+\mathbf{K}_{pp})(\mathbf{C}\mathbf{Y}\otimes\mathbf{I})\text{vec}(\text{d}\mathbf{C}) \\ &= 2\mathbf{G}\mathbf{G}^+(\mathbf{C}\mathbf{Y}\otimes\mathbf{I})\mathbf{F}^\top\text{dvech}(\mathbf{C}),\end{aligned}$$

$$\begin{aligned}\text{dvec}(\mathbf{C}^{-\top}\mathbf{Y}\mathbf{C}^{-1}) &= \text{vec}(\text{d}\mathbf{C}^{-\top}\mathbf{Y}\mathbf{C}^{-1}) + \text{vec}(\mathbf{C}^{-\top}\mathbf{Y}\text{d}\mathbf{C}^{-1}) \\ &= -\mathbf{K}_{pp}\text{vec}(\mathbf{C}^{-\top}\mathbf{Y}\mathbf{C}^{-1}\text{d}\mathbf{C}\mathbf{C}^{-1}) - \text{vec}(\mathbf{C}^{-\top}\mathbf{Y}\mathbf{C}^{-1}\text{d}\mathbf{C}\mathbf{C}^{-1}) \\ &= -(\mathbf{I}+\mathbf{K}_{pp})\text{vec}(\mathbf{C}^{-\top}\mathbf{Y}\mathbf{C}^{-1}\text{d}\mathbf{C}\mathbf{C}^{-1}) \\ &= -(\mathbf{I}+\mathbf{K}_{pp})(\mathbf{C}^{-\top}\otimes\mathbf{C}^{-\top}\mathbf{Y}\mathbf{C}^{-1})\text{vec}(\text{d}\mathbf{C}) \\ &= -2\mathbf{G}\mathbf{G}^+(\mathbf{C}^{-\top}\otimes\mathbf{C}^{-\top}\mathbf{Y}\mathbf{C}^{-1})\mathbf{F}^\top\text{dvech}(\mathbf{C}),\end{aligned}$$

where we used (3.21), (3.12) and (3.22), such that

$$\frac{\partial \mathrm{vec}(\mathbf{C}\mathbf{Y}\mathbf{C}^\top)}{\partial \mathrm{vech}(\mathbf{C})^\top} = 2\mathbf{G}\mathbf{G}^+(\mathbf{C}\mathbf{Y}\otimes\mathbf{I})\mathbf{F}^\top \tag{3.31}$$

and

$$\frac{\partial \mathrm{vec}(\mathbf{C}^{-\top}\mathbf{Y}\mathbf{C}^{-1})}{\partial \mathrm{vech}(\mathbf{C})^\top} = -2\mathbf{G}\mathbf{G}^+(\mathbf{C}^{-\top}\otimes\mathbf{C}^{-\top}\mathbf{Y}\mathbf{C}^{-1})\mathbf{F}^\top. \tag{3.32}$$

Furthermore,

$$\begin{aligned}
\mathrm{dvech}(\mathbf{X}^{-1}) &= \mathbf{G}^+\mathrm{dvec}(\mathbf{C}^{-\top}\mathbf{C}^{-1}) \\
&= -\mathbf{G}^+\mathrm{vec}([\mathbf{C}^{-1}\mathrm{d}\mathbf{C}\mathbf{C}^{-1}]^\top\mathbf{C}^\top) - \mathbf{G}^+\mathrm{vec}(\mathbf{C}^{-\top}\mathbf{C}^{-1}\mathrm{d}\mathbf{C}\mathbf{C}^{-1}) \\
&= -\mathbf{G}^+(\mathbf{I}+\mathbf{K}_{pp})\mathrm{vec}(\mathbf{C}^{-\top}\mathbf{C}^{-1}\mathrm{d}\mathbf{C}\mathbf{C}^{-1}) \\
&= -\mathbf{G}^+\mathbf{G}\mathbf{G}^+\mathrm{vec}(\mathbf{C}^{-\top}\mathbf{C}^{-1}\mathrm{d}\mathbf{C}\mathbf{C}^{-1}) \\
&= -2\mathbf{G}^+(\mathbf{C}^{-\top}\otimes\mathbf{C}^{-\top}\mathbf{C}^{-1})\mathrm{vec}(\mathrm{d}\mathbf{C}) \\
&= -2\mathbf{G}^+(\mathbf{C}^{-\top}\otimes\mathbf{X}^{-1})\mathbf{F}^\top\mathrm{dvech}(\mathbf{C}).
\end{aligned}$$

such that

$$\frac{\partial \mathrm{vech}(\mathbf{C})}{\partial \mathrm{vech}(\mathbf{X}^{-1})^\top} = -\frac{1}{2}(\mathbf{G}^+(\mathbf{C}^{-\top}\otimes\mathbf{X}^{-1})\mathbf{F}^\top)^{-1}.$$

Finally, recall Lemma 3.6.4, then the lemma follows by applying the chain rule,

$$\begin{aligned}
\frac{\partial \mathrm{vec}(\mathbf{C}\mathbf{Y}\mathbf{C}^\top)}{\partial \mathrm{vech}(\mathbf{X})^\top} &= \frac{\partial \mathrm{vec}(\mathbf{C}\mathbf{Y}\mathbf{C}^\top)}{\partial \mathrm{vech}(\mathbf{C})^\top}\frac{\partial \mathrm{vech}(\mathbf{C})}{\partial \mathrm{vech}(\mathbf{X}^{-1})^\top}\frac{\partial \mathrm{vech}(\mathbf{X}^{-1})}{\partial \mathrm{vech}(\mathbf{X})^\top} \\
&= \frac{\partial \mathrm{vec}(\mathbf{C}\mathbf{Y}\mathbf{C}^\top)}{\partial \mathrm{vech}(\mathbf{C})^\top}\frac{\partial \mathrm{vech}(\mathbf{C})}{\partial \mathrm{vech}(\mathbf{X})^\top},
\end{aligned}$$

$$\begin{aligned}
\frac{\partial \mathrm{vec}(\mathbf{C}\mathbf{Y}\mathbf{C}^\top)}{\partial \mathrm{vech}(\mathbf{X}^{-1})^\top} &= \frac{\partial \mathrm{vec}(\mathbf{C}\mathbf{Y}\mathbf{C}^\top)}{\partial \mathrm{vech}(\mathbf{C})^\top}\frac{\partial \mathrm{vech}(\mathbf{C})}{\partial \mathrm{vech}(\mathbf{X})^\top}\frac{\partial \mathrm{vech}(\mathbf{X})}{\partial \mathrm{vech}(\mathbf{X}^{-1})^\top} \\
&= \frac{\partial \mathrm{vec}(\mathbf{C}\mathbf{Y}\mathbf{C}^\top)}{\partial \mathrm{vech}(\mathbf{C})^\top}\frac{\partial \mathrm{vech}(\mathbf{C})}{\partial \mathrm{vech}(\mathbf{X}^{-1})^\top},
\end{aligned}$$

$$\begin{aligned}
\frac{\partial \mathrm{vec}(\mathbf{C}^{-\top}\mathbf{Y}\mathbf{C}^{-1})}{\partial \mathrm{vech}(\mathbf{X}^{-1})^\top} &= \frac{\partial \mathrm{vec}(\mathbf{C}^{-\top}\mathbf{Y}\mathbf{C}^{-1})}{\partial \mathrm{vech}(\mathbf{C})^\top}\frac{\partial \mathrm{vech}(\mathbf{C})}{\partial \mathrm{vech}(\mathbf{X})^\top}\frac{\partial \mathrm{vech}(\mathbf{X})}{\partial \mathrm{vech}(\mathbf{X}^{-1})^\top} \\
&= \frac{\partial \mathrm{vec}(\mathbf{C}^{-\top}\mathbf{Y}\mathbf{C}^{-1})}{\partial \mathrm{vech}(\mathbf{C})^\top}\frac{\partial \mathrm{vech}(\mathbf{C})}{\partial \mathrm{vech}(\mathbf{X}^{-1})^\top},
\end{aligned}$$

$$\begin{aligned}\frac{\partial \operatorname{vec}(\mathbf{C}^{-\top}\mathbf{Y}\mathbf{C}^{-1})}{\partial \operatorname{vech}(\mathbf{X})^\top} &= \frac{\partial \operatorname{vec}(\mathbf{C}^{-\top}\mathbf{Y}\mathbf{C}^{-1})}{\partial \operatorname{vech}(\mathbf{C})^\top}\frac{\partial \operatorname{vech}(\mathbf{C})}{\partial \operatorname{vech}(\mathbf{X}^{-1})^\top}\frac{\partial \operatorname{vech}(\mathbf{X}^{-1})}{\partial \operatorname{vech}(\mathbf{X})^\top} \\ &= \frac{\partial \operatorname{vec}(\mathbf{C}^{-\top}\mathbf{Y}\mathbf{C}^{-1})}{\partial \operatorname{vech}(\mathbf{C})^\top}\frac{\partial \operatorname{vech}(\mathbf{C})}{\partial \operatorname{vech}(\mathbf{X})^\top}.\end{aligned}$$

There are two versions for each derivative. The longer versions immediately make obvious the nesting of the case where $\mathbf{Y} = c\mathbf{I}$. The equality between the two versions can be derived based on the two different chain rule applications or by noticing that

$$\begin{aligned}&(\mathbf{G}^+(\mathbf{C}^{-\top}\otimes\mathbf{X}^{-1})\mathbf{F}^\top)^{-1}\mathbf{G}^+\mathbf{X}^{-\otimes 2}\mathbf{G} \\ &\quad=(\mathbf{G}^\top(\mathbf{C}^{-\top}\otimes\mathbf{X}^{-1})\mathbf{F}^\top)^{-1}\mathbf{G}^\top\mathbf{G}\mathbf{G}^+\mathbf{X}^{-\otimes 2}\mathbf{G} \\ &\quad=(\mathbf{G}^\top(\mathbf{C}^{-\top}\otimes\mathbf{X}^{-1})\mathbf{F}^\top)^{-1}\mathbf{G}^\top\mathbf{X}^{-\otimes 2}\mathbf{G} \\ &\quad=(\mathbf{G}^\top(\mathbf{C}^{-\top}\otimes\mathbf{X}^{-1})\mathbf{F}^\top)^{-1}(\mathbf{G}^+\mathbf{X}^{\otimes 2}(\mathbf{G}^+)^\top)^{-1} \\ &\quad=(\mathbf{G}^+\mathbf{X}^{\otimes 2}(\mathbf{G}^+)^\top\mathbf{G}^\top(\mathbf{C}^{-\top}\otimes\mathbf{X}^{-1})\mathbf{F}^\top)^{-1} \\ &\quad=(\mathbf{G}^+\mathbf{X}^{\otimes 2}(\mathbf{G}\mathbf{G}^+)^\top(\mathbf{C}^{-\top}\otimes\mathbf{X}^{-1})\mathbf{F}^\top)^{-1} \\ &\quad=(\mathbf{G}^+\mathbf{G}\mathbf{G}^+\mathbf{X}^{\otimes 2}(\mathbf{C}^{-\top}\otimes\mathbf{X}^{-1})\mathbf{F}^\top)^{-1} \\ &\quad=(\mathbf{G}^+\mathbf{X}^{\otimes 2}(\mathbf{C}^{-\top}\otimes\mathbf{X}^{-1})\mathbf{F}^\top)^{-1} \\ &\quad=(\mathbf{G}^+(\mathbf{C}\otimes\mathbf{I})\mathbf{F}^\top)^{-1},\end{aligned}$$

and

$$\begin{aligned}&(\mathbf{G}^+(\mathbf{C}\otimes\mathbf{I})\mathbf{F}^\top)^{-1}\mathbf{G}^+\mathbf{X}^{\otimes 2}\mathbf{G} \\ &\quad= (\mathbf{G}^\top(\mathbf{C}\otimes\mathbf{I})\mathbf{F}^\top)^{-1}\mathbf{G}^\top\mathbf{G}\mathbf{G}^+\mathbf{X}^{\otimes 2}\mathbf{G} \\ &\quad= (\mathbf{G}^\top(\mathbf{C}\otimes\mathbf{I})\mathbf{F}^\top)^{-1}\mathbf{G}^\top\mathbf{X}^{\otimes 2}\mathbf{G} \\ &\quad= (\mathbf{G}^\top(\mathbf{C}\otimes\mathbf{I})\mathbf{F}^\top)^{-1}(\mathbf{G}^+\mathbf{X}^{-\otimes 2}(\mathbf{G}^+)^\top)^{-1} \\ &\quad= (\mathbf{G}^+\mathbf{X}^{-\otimes 2}(\mathbf{G}^+)^\top\mathbf{G}^\top(\mathbf{C}\otimes\mathbf{I})\mathbf{F}^\top)^{-1} \\ &\quad= (\mathbf{G}^+\mathbf{X}^{-\otimes 2}(\mathbf{G}\mathbf{G}^+)^\top(\mathbf{C}\otimes\mathbf{I})\mathbf{F}^\top)^{-1} \\ &\quad= (\mathbf{G}^+\mathbf{G}\mathbf{G}^+\mathbf{X}^{-\otimes 2}(\mathbf{C}\otimes\mathbf{I})\mathbf{F}^\top)^{-1} \\ &\quad= (\mathbf{G}^+(\mathbf{C}^{-\top}\otimes\mathbf{X}^{-1})\mathbf{F}^\top)^{-1},\end{aligned}$$

where we have used Theorem 3.13 (c) of Magnus and Neudecker (2019) for the third equality.

□

Lemma 3.6.6. *Let* $\mathbf{X}$ *be a non-singular symmetric matrix,* $\mathbf{C}$ *its lower Cholesky factor, and* $\mathbf{Y}$ *and* $\mathbf{Z}$ *be square matrices. Then*

$$\frac{\partial \mathrm{tr}(\mathbf{XY})}{\partial \mathrm{vech}(\mathbf{X})^\top} = \mathrm{vec}(\mathbf{Y})^\top \mathbf{G},$$

$$\frac{\partial \mathrm{tr}(\mathbf{X}^{-1}\mathbf{Y})}{\partial \mathrm{vech}(\mathbf{X})^\top} = -\mathrm{vec}(\mathbf{X}^{-1}\mathbf{Y}\mathbf{X}^{-1})^\top \mathbf{G},$$

$$\frac{\partial \mathrm{tr}(\mathbf{C}^{-\top}\mathbf{Y}\mathbf{C}^{-1}\mathbf{Z})}{\partial \mathrm{vech}(\mathbf{C})^\top} = -2\mathrm{vec}(\mathbf{C}^{-\top}\mathbf{Y}\mathbf{C}^{-1}\mathbf{Z}\mathbf{C}^{-\top})^\top \mathbf{F}^\top,$$

and

$$\frac{\partial^2 \mathrm{tr}(\mathbf{XY})}{\partial \mathrm{vech}(\mathbf{X}) \partial \mathrm{vech}(\mathbf{X})^\top} = \mathbf{0}.$$

Proof. We have

$$\begin{aligned} \mathrm{d}\mathrm{tr}(\mathbf{XY}) &= \mathrm{tr}(\mathbf{Y}\mathrm{d}\mathbf{X}) \\ &= \mathrm{vec}(\mathbf{Y})^\top \mathrm{vec}(\mathrm{d}\mathbf{X}) \\ &= \mathrm{vec}(\mathbf{Y})^\top \mathbf{G}\mathrm{dvech}(\mathbf{X}), \end{aligned}$$

using (3.14) and (3.16),

$$\begin{aligned} \mathrm{d}\mathrm{tr}(\mathbf{X}^{-1}\mathbf{Y}) &= -\mathrm{tr}(\mathbf{X}^{-1}\mathrm{d}\mathbf{X}\mathbf{X}^{-1}\mathbf{Y}) \\ &= -\mathrm{tr}(\mathbf{X}^{-1}\mathbf{Y}\mathbf{X}^{-1}\mathrm{d}\mathbf{X}) \\ &= -\mathrm{vec}(\mathbf{X}^{-1}\mathbf{Y}\mathbf{X}^{-1})^\top \mathrm{vec}(\mathrm{d}\mathbf{X}) \\ &= -\mathrm{vec}(\mathbf{X}^{-1}\mathbf{Y}\mathbf{X}^{-1})^\top \mathbf{G}\mathrm{dvech}(\mathbf{X}), \end{aligned}$$

using (3.13), (3.14) and (3.16),

$$\mathrm{d}^2\mathrm{tr}(\mathbf{XY}) = \mathrm{d}\mathrm{tr}(\mathbf{Y}\mathrm{d}\mathbf{X}) = 0.$$

Furthermore,

$$\begin{aligned}
\mathrm{d}\mathrm{tr}(\mathbf{C}^{-\top}\mathbf{Y}\mathbf{C}^{-1}\mathbf{Z}) &= -\mathrm{tr}((\mathbf{C}^{-1}\mathrm{d}\mathbf{C}\mathbf{C}^{-1})^{-\top}\mathbf{Y}\mathbf{C}^{-1}\mathbf{Z} + \mathbf{C}^{-\top}\mathbf{Y}\mathbf{C}^{-1}\mathrm{d}\mathbf{C}\mathbf{C}^{-1}\mathbf{Z}) \\
&= -\mathrm{tr}(\mathrm{d}\mathbf{C}^{\top}\mathbf{C}^{-\top}\mathbf{Y}\mathbf{C}^{-1}\mathbf{Z}\mathbf{C}^{-\top} + \mathbf{C}^{-1}\mathbf{Z}\mathbf{C}^{-\top}\mathbf{Y}\mathbf{C}^{-1}\mathrm{d}\mathbf{C}) \\
&= -2\mathrm{tr}(\mathbf{C}^{-1}\mathbf{Z}\mathbf{C}^{-\top}\mathbf{Y}\mathbf{C}^{-1}\mathrm{d}\mathbf{C}) \\
&= -2\mathrm{vec}(\mathbf{C}^{-1}\mathbf{Z}\mathbf{C}^{-\top}\mathbf{Y}\mathbf{C}^{-1})^{\top}\mathrm{d}\mathrm{vec}(\mathbf{C}) \\
&= -2\mathrm{vec}(\mathbf{C}^{-1}\mathbf{Z}\mathbf{C}^{-\top}\mathbf{Y}\mathbf{C}^{-1})^{\top}\mathbf{F}^{\top}\mathrm{d}\mathrm{vech}(\mathbf{C}),
\end{aligned}$$

For properties of the differential "d", see Magnus and Neudecker (2019) pp. 163-169 and pp. 434-436. To convert differentials to derivatives, see Tables 9.2 and 10.1 in Magnus and Neudecker (2019).

□

Lemma 3.6.7. *Let* $\mathbf{\Sigma}$ *be a symmetric positive definite matrix and* $\mathbf{C}$ *its lower Cholesky factor. Then*

$$\begin{aligned}
\frac{\partial \log|\mathbf{\Sigma}|_{\mathbf{n}}}{\partial \mathrm{vech}(\mathbf{\Sigma})^{\top}} &= \mathrm{vec}(\mathbf{C}^{-\top}\mathrm{dg}(\mathbf{n})\mathbf{C}^{-1})^{\top}\mathbf{G}, \\
\frac{\partial \log|\mathbf{\Sigma}|_{\mathbf{n}}}{\partial \mathrm{vech}(\mathbf{\Sigma}^{-1})^{\top}} &= -\mathrm{vec}(\mathbf{C}\,\mathrm{dg}(\mathbf{n})\mathbf{C}^{\top})^{\top}\mathbf{G},
\end{aligned}$$

$$\begin{aligned}
&\frac{\partial^2 \log|\mathbf{\Sigma}|_{\mathbf{n}}}{\partial \mathrm{vech}(\mathbf{\Sigma})\partial \mathrm{vech}(\mathbf{\Sigma})^{\top}} \\
&\quad = -\mathbf{G}^{\top}(\mathbf{C}^{-\top} \otimes \mathbf{C}^{-\top}\mathrm{dg}(\mathbf{n})\mathbf{C}^{-1})\mathbf{F}^{\top}(\mathbf{G}^{\top}(\mathbf{C}^{-\top} \otimes \mathbf{\Sigma}^{-1})\mathbf{F}^{\top})^{-1}\mathbf{G}^{\top}\mathbf{\Sigma}^{-\otimes 2}\mathbf{G} \\
&\quad = -\mathbf{G}^{\top}(\mathbf{C}^{-\top} \otimes \mathbf{C}^{-\top}\mathrm{dg}(\mathbf{n})\mathbf{C}^{-1})\mathbf{F}^{\top}(\mathbf{G}^{+}(\mathbf{C} \otimes \mathbf{I})\mathbf{F}^{\top})^{-1},
\end{aligned}$$

$$\begin{aligned}
&\frac{\partial^2 \log|\mathbf{\Sigma}|_{\mathbf{n}}}{\partial \mathrm{vech}(\mathbf{\Sigma}^{-1})\partial \mathrm{vech}(\mathbf{\Sigma}^{-1})^{\top}} \\
&\quad = \mathbf{G}^{\top}(\mathbf{C}\mathrm{dg}(\mathbf{n}) \otimes \mathbf{I})\mathbf{F}^{\top}(\mathbf{G}^{+}(\mathbf{C} \otimes \mathbf{I})\mathbf{F}^{\top})^{-1}\mathbf{G}^{+}\mathbf{\Sigma}^{\otimes 2}\mathbf{G} \\
&\quad = \mathbf{G}^{\top}(\mathbf{C}\mathrm{dg}(\mathbf{n}) \otimes \mathbf{I})\mathbf{F}^{\top}(\mathbf{G}^{+}(\mathbf{C}^{-\top} \otimes \mathbf{\Sigma}^{-1})\mathbf{F}^{\top})^{-1}.
\end{aligned}$$

Proof. Decompose $\mathbf{\Sigma} = \mathbf{T}\mathbf{D}\mathbf{T}^{\top}$, where $\mathbf{T}$ is a lower triangular matrix with diagonal elements being 1 and $\mathbf{D}$ is a diagonal matrix with positive diagonal elements, such

that $\mathbf{C} = \mathbf{T}\mathbf{D}^{\frac{1}{2}}$ is the lower Cholesky factor. Note that

$$\frac{\partial \sum_{i=1}^{p} n_i \log(\mathbf{D}_{ii})}{\partial \mathbf{D}} = \begin{bmatrix} \frac{\mathbf{n}_1}{\mathbf{D}_{1\,1}} & & & \\ & \frac{\mathbf{n}_2}{\mathbf{D}_{2\,2}} & & \\ & & \ddots & \\ & & & \frac{\mathbf{n}_p}{\mathbf{D}_{pp}} \end{bmatrix} = \mathbf{D}^{-\frac{1}{2}} \mathrm{dg}(\mathbf{n}) \mathbf{D}^{-\frac{1}{2}},$$

such that

$$\frac{\partial \sum_{i=1}^{p} n_i \log(\mathbf{D}_{ii})}{\partial \mathrm{vec}(\mathbf{D})^\top} = \mathrm{vec}(\mathbf{D}^{-\frac{1}{2}} \mathrm{dg}(\mathbf{n}) \mathbf{D}^{-\frac{1}{2}})^\top$$

and

$$\mathrm{dvec}(\mathbf{\Sigma}) = \mathrm{dvec}(\mathbf{T}\mathbf{D}\mathbf{T}^\top) = (\mathbf{T}^{\otimes 2}) \mathrm{dvec}(\mathbf{D}),$$

such that

$$\frac{\partial \mathrm{vec}(\mathbf{D})}{\partial \mathrm{vec}(\mathbf{\Sigma})^\top} = (\mathbf{T}^{\otimes 2})^{-1}.$$

Then

$$\begin{aligned} \frac{\partial \log |\mathbf{\Sigma}|_{\mathbf{n}}}{\partial \mathrm{vech}(\mathbf{\Sigma})^\top} &= \frac{\partial \sum_{i=1}^{p} n_i \log(\mathbf{D}_{ii})}{\partial \mathrm{vec}(\mathbf{D})^\top} \frac{\partial \mathrm{vec}(\mathbf{D})}{\partial \mathrm{vec}(\mathbf{\Sigma})^\top} \frac{\partial \mathrm{vec}(\mathbf{\Sigma})}{\partial \mathrm{vech}(\mathbf{\Sigma})^\top} \\ &= \mathrm{vec}(\mathbf{D}^{-\frac{1}{2}} \mathrm{dg}(\mathbf{n}) \mathbf{D}^{-\frac{1}{2}})^\top (\mathbf{T}^{\otimes 2})^{-1} \mathbf{G} \\ &= \mathrm{vec}(\mathbf{D}^{-\frac{1}{2}} \mathrm{dg}(\mathbf{n}) \mathbf{D}^{-\frac{1}{2}})^\top (\mathbf{T}^{-\top} \otimes \mathbf{T}^{-\top})^\top \mathbf{G} \\ &= \mathrm{vec}(\mathbf{T}^{-\top} \mathbf{D}^{-\frac{1}{2}} \mathrm{dg}(\mathbf{n}) \mathbf{D}^{-\frac{1}{2}} \mathbf{T}^{-1})^\top \mathbf{G} \\ &= \mathrm{vec}(\mathbf{C}^{-\top} \mathrm{dg}(\mathbf{n}) \mathbf{C}^{-1})^\top \mathbf{G}, \end{aligned}$$

where we used (3.12). Furthermore

$$\begin{aligned} \frac{\partial \log |\mathbf{\Sigma}|_{\mathbf{n}}}{\partial \mathrm{vech}(\mathbf{\Sigma}^{-1})^\top} &= \frac{\partial \log |\mathbf{\Sigma}|_{\mathbf{n}}}{\partial \mathrm{vech}(\mathbf{\Sigma})^\top} \frac{\partial \mathrm{vech}(\mathbf{\Sigma})}{\partial \mathrm{vech}(\mathbf{\Sigma}^{-1})^\top} \\ &= -\mathrm{vec}(\mathbf{C}^{-\top} \mathrm{dg}(\mathbf{n}) \mathbf{C}^{-1})^\top \mathbf{G} \mathbf{G}^{+} \mathbf{\Sigma}^{\otimes 2} \mathbf{G} \\ &= -\mathrm{vec}(\mathbf{\Sigma} \mathbf{C}^{-\top} \mathrm{dg}(\mathbf{n}) \mathbf{C}^{-1} \mathbf{\Sigma})^\top \mathbf{G} \\ &= -\mathrm{vec}(\mathbf{C} \mathrm{dg}(\mathbf{n}) \mathbf{C}^\top)^\top \mathbf{G}, \end{aligned}$$

where we used Lemma 3.6.3 and equation (3.12). Now applying Lemma 3.6.5 on

$$\frac{\partial \log |\mathbf{\Sigma}|_{\mathbf{n}}}{\partial \mathrm{vech}(\mathbf{\Sigma}) \mathrm{vech}(\mathbf{\Sigma})^\top} = \mathbf{G}^\top \frac{\partial \mathrm{vec}(\mathbf{C}^{-\top} \mathrm{dg}(\mathbf{n}) \mathbf{C}^{-1})}{\partial \mathrm{vech}(\mathbf{\Sigma})^\top} \text{ and}$$

$$\frac{\partial \log |\mathbf{\Sigma}|_{\mathbf{n}}}{\partial \mathrm{vech}(\mathbf{\Sigma}^{-1}) \mathrm{vech}(\mathbf{\Sigma}^{-1})^\top} = -\mathbf{G}^\top \frac{\partial \mathrm{vec}(\mathbf{C} \mathrm{dg}(\mathbf{n}) \mathbf{C}^\top)}{\partial \mathrm{vech}(\mathbf{\Sigma}^{-1})^\top}$$

and using $\mathbf{G}^\top \mathbf{G}\mathbf{G}^+ = \mathbf{G}^\top$ directly gives the desired results.

□

Lemma 3.6.8. *Let* $\mathbf{X}$ *be a square matrix. Then*

$$\mathbf{F}^\top \mathbf{F} \text{vec}(\mathbf{X}) = \text{vec}(\blacksquare(\mathbf{X}))$$

Proof. $\mathbf{F}$ eliminates those elements from $\text{vec}(\mathbf{X})$ which are on $\mathbf{X}$'s upper triangular part. Then $\mathbf{F}^\top$ we know from equation (3.19) is the matrix which for a triangular matrix maps its half-vectorization $\text{vech}(\cdot)$ to its vectorization $\text{vec}(\cdot)$.

□

Finally, this helps us with rewriting the scores in a format that is quick to evaluate since it avoids Kronecker product products or inversions. To see this, consider

Lemma 3.6.9. *Let* $\mathcal{D} \in (\mathcal{W}, i\mathcal{W}, t\mathcal{W}, it\mathcal{W}, F, \mathcal{R}, i\mathcal{R}, t\mathcal{R}, it\mathcal{R}, F\mathcal{R}, iF\mathcal{R})$ *be one of the distributions for RCs. Then*

$$\nabla_{\mathcal{D}}^{\Sigma} = \mathbf{G}^\top \text{vec}(\triangle_{\mathcal{D}}^{\Sigma}),$$

with

$$\triangle_{\mathcal{D}}^{\Sigma} = \mathbf{C}^{-\top} \tilde{\blacksquare}(\mathbf{C}^\top \blacksquare(\triangle_{\mathcal{D}}^{\Omega} \mathbf{C} \mathbf{M}_{\mathcal{D}}^{-1})) \mathbf{C}^{-1},$$

where $\triangle_{\mathcal{D}}^{\Omega}$ *is symmetric with elements*

$$(\triangle_{\mathcal{D}}^{\Omega})_{ij} = \frac{\partial p_{\mathcal{D}}}{\partial (\mathbf{\Omega})_{ij}}, \tag{3.33}$$

$\mathbf{M}_{\mathcal{D}}$ *is the expectation of the stochastic representation kernel for the respective distribution (see Table 2.3),* $\mathbf{C}$ *is the lower Cholesky factor of the expected value matrix* $\mathbf{\Sigma}$ *and* $\blacksquare(\cdot)$ *and* $\tilde{\blacksquare}(\cdot)$ *are defined in 3.6.1 and 3.6.3. For all Wishart-type distributions* $\mathbf{M}_{\mathcal{D}} = m_{\mathcal{D}} \mathbf{I}$*, with scalar* $m_{\mathcal{D}}$*, and the expression for* $\triangle_{\mathcal{D}}^{\Sigma}$ *reduces to*

$$\triangle_{\mathcal{D}}^{\Sigma} = \frac{1}{m_{\mathcal{D}}} \triangle_{\mathcal{D}}^{\Omega}.$$

Proof.

$$
\begin{aligned}
\frac{\partial p_{\mathcal{D}}}{\partial \mathrm{vech}(\mathbf{\Sigma})^\top} &= \frac{\partial p_{\mathcal{D}}}{\partial \mathrm{vech}(\mathbf{\Omega})^\top}\frac{\partial \mathrm{vech}(\mathbf{\Omega})}{\partial \mathrm{vech}(\mathbf{\Sigma})^\top} \\
&= \mathrm{vec}(\triangle_{\mathcal{D}}^{\Omega})^\top \mathbf{G}\frac{\partial \mathrm{vech}(\mathbf{C}\mathbf{M}_{\mathcal{D}}^{-1}\mathbf{C}^\top)}{\partial \mathrm{vech}(\mathbf{C}\mathbf{C}^\top)^\top} \\
&= \mathrm{vec}(\triangle_{\mathcal{D}}^{\Omega})^\top \mathbf{G}\mathbf{G}^+(\mathbf{C}\mathbf{M}_{\mathcal{D}}^{-1}\otimes\mathbf{I})\mathbf{F}^\top(\mathbf{G}^+(\mathbf{C}\otimes\mathbf{I})\mathbf{F}^\top)^{-1} \\
&= \mathrm{vec}(\triangle_{\mathcal{D}}^{\Omega})^\top (\mathbf{C}\mathbf{M}_{\mathcal{D}}^{-1}\otimes\mathbf{I})\mathbf{F}^\top(\mathbf{G}^+(\mathbf{C}\otimes\mathbf{I})\mathbf{F}^\top)^{-1} \\
&= \mathrm{vec}(\triangle_{\mathcal{D}}^{\Omega}\mathbf{C}\mathbf{M}_{\mathcal{D}}^{-1})^\top \mathbf{F}^\top(\mathbf{G}^+(\mathbf{C}\otimes\mathbf{I})\mathbf{F}^\top)^{-1} \\
&= 2\mathrm{vec}(\triangle_{\mathcal{D}}^{\Omega}\mathbf{C}\mathbf{M}_{\mathcal{D}}^{-1})^\top \mathbf{F}^\top\mathbf{F}(\mathbf{I}\otimes\mathbf{C})\mathbf{Z}\mathbf{C}^{-\otimes 2}\mathbf{G} \\
&= 2\mathrm{vec}(\blacktriangleright(\triangle_{\mathcal{D}}^{\Omega}\mathbf{C})\mathbf{M}_{\mathcal{D}}^{-1})^\top (\mathbf{I}\otimes\mathbf{C})\mathbf{Z}\mathbf{C}^{-\otimes 2}\mathbf{G} \\
&= 2\mathrm{vec}(\mathbf{C}^\top\blacktriangleright(\triangle_{\mathcal{D}}^{\Omega}\mathbf{C})\mathbf{M}_{\mathcal{D}}^{-1})^\top \mathbf{Z}\mathbf{C}^{-\otimes 2}\mathbf{G} \\
&= 2\mathrm{vec}(\bar{\blacktriangleright}(\mathbf{C}^\top\blacktriangleright(\triangle_{\mathcal{D}}^{\Omega}\mathbf{C})\mathbf{M}_{\mathcal{D}}^{-1}))^\top \mathbf{C}^{-\otimes 2}\mathbf{G} \\
&= 2\mathrm{vec}(\mathbf{C}^{-\top}\bar{\blacktriangleright}(\mathbf{C}^\top\blacktriangleright(\triangle_{\mathcal{D}}^{\Omega}\mathbf{C})\mathbf{M}_{\mathcal{D}}^{-1})\mathbf{C}^{-1})^\top \mathbf{G} \\
&= 2\mathrm{vec}(\mathbf{C}^{-\top}\bar{\blacktriangleright}(\mathbf{C}^\top\blacktriangleright(\triangle_{\mathcal{D}}^{\Omega}\mathbf{C}\mathbf{M}_{\mathcal{D}}^{-1}))\mathbf{C}^{-1})^\top \mathbf{G},
\end{aligned}
$$

where we used Lemmas 3.6.5, 3.6.8, and 3.6.4, Definition 3.6.2 and equation (3.25). For $\bar{\blacktriangleright}(\cdot)$ see Definition 3.6.2. Note that, even though

$$
\mathrm{vec}(\triangle_{\mathcal{D}}^{\Sigma})^\top\mathbf{G} = \frac{\partial \log p_{\mathcal{D}}}{\partial \mathrm{vech}(\mathbf{\Sigma})^\top} = 2\mathrm{vec}(\mathbf{C}^{-\top}\bar{\blacktriangleright}(\mathbf{C}^\top\blacktriangleright(\triangle_{\mathcal{D}}^{\Omega}\mathbf{C}\mathbf{M}_{\mathcal{D}}^{-1}))\mathbf{C}^{-1})^\top\mathbf{G},
$$

this does **not** imply that

$$
\triangle_{\mathcal{D}}^{\Sigma} = 2\mathbf{C}^{-\top}\bar{\blacktriangleright}(\mathbf{C}^\top\blacktriangleright(\triangle_{\mathcal{D}}^{\Omega}\mathbf{C}\mathbf{M}_{\mathcal{D}}^{-1}))\mathbf{C}^{-1}
$$

since $\triangle_{\mathcal{D}}^{\Sigma}$ is symmetric, but the r.h.s. is not necessarily symmetric. However, since we know that $\triangle_{\mathcal{D}}^{\Sigma}$ is the **unique symmetric** matrix for which $\frac{\partial p_{\mathcal{D}}}{\partial \mathrm{vech}(\mathbf{\Sigma})^\top} = \mathrm{vec}(\triangle_{\mathcal{D}}^{\Sigma})^\top\mathbf{G}$ holds, it must be true that

$$
\begin{aligned}
\triangle_{\mathcal{D}}^{\Sigma} &= \mathbf{C}^{-\top}\bar{\blacktriangleright}(\mathbf{C}^\top\blacktriangleright(\triangle_{\mathcal{D}}^{\Omega}\mathbf{C}\mathbf{M}_{\mathcal{D}}^{-1}))\mathbf{C}^{-1} + (\mathbf{C}^{-\top}\bar{\blacktriangleright}(\mathbf{C}^\top\blacktriangleright(\triangle_{\mathcal{D}}^{\Omega}\mathbf{C}\mathbf{M}_{\mathcal{D}}^{-1}))\mathbf{C}^{-1})^\top \\
&= \mathbf{C}^{-\top}\tilde{\blacktriangleright}(\mathbf{C}^\top\blacktriangleright(\triangle_{\mathcal{D}}^{\Omega}\mathbf{C}\mathbf{M}_{\mathcal{D}}^{-1}))\mathbf{C}^{-1}.
\end{aligned}
$$

Furthermore, note that

$$
\begin{aligned}
\mathbf{C}^{-\top}\bar{\blacktriangleright}(\mathbf{C}^\top\blacktriangleright(\mathbf{C}^{-\top}\mathrm{dg}(\mathbf{n})))\mathbf{C}^{-1} &= \mathbf{C}^{-\top}\bar{\blacktriangleright}(\mathbf{C}^\top\mathrm{dg}(\mathbf{C}^{-\top})\mathrm{dg}(\mathbf{n}))\mathbf{C}^{-1} \\
&= \mathbf{C}^{-\top}\mathrm{dg}(\mathbf{C}^\top)\mathrm{dg}(\mathbf{C}^{-\top})\mathrm{dg}(\mathbf{n})\mathbf{C}^{-1} \\
&= \mathbf{C}^{-\top}\mathrm{dg}(\mathbf{n})\mathbf{C}^{-1}.
\end{aligned}
$$

□

Lemma 3.6.10. *Let* $\mathbf{n}$ *be a real column vector of length* p*, then*

$$\frac{\partial \log |\boldsymbol{\Sigma}|_{\mathbf{n}}}{\partial \mathbf{n}^\top} = 2 \log \operatorname{vecd}(\mathbf{C})^\top.$$

Proof.

$$\frac{\partial \log |\boldsymbol{\Sigma}|_{\mathbf{n}}}{\partial \mathbf{n}^\top} = \frac{\partial \log(\prod_{i=1}^{p} \mathbf{C}_{ii}^{2n_i})}{\partial \mathbf{n}^\top} = \frac{\partial \sum_{i=1}^{p} 2n_i \log(\mathbf{C}_{ii})}{\partial \mathbf{n}^\top} = 2 \log \operatorname{vecd}(\mathbf{C})^\top.$$

□

The following Lemma is just for rewriting of the scores in Table 3.3.

Lemma 3.6.11. *Let* $\mathbf{X}$ *and* $\mathbf{Y}$ *be symmetric* $p \times p$ *positive definite matrices with lower Cholesky factors* $\mathbf{C}_X$ *and* $\mathbf{C}_Y$*, respectively, and let* $\mathbf{n}$ *be a real vector of length* p*. Then*

$$\log \operatorname{vecd}(\mathbf{C}_X) - \log \operatorname{vecd}(\mathbf{C}_Y) = \log \operatorname{vecd}(\mathbf{C}_Y^{-1} \mathbf{C}_X)$$

Proof.

$$\begin{aligned}
\log \operatorname{vecd}(\mathbf{C}_X) - \log \operatorname{vecd}(\mathbf{C}_Y) &= \frac{1}{2} \frac{\partial (\log |\mathbf{X}|_{\mathbf{n}} - \log |\mathbf{Y}|_{\mathbf{n}})}{\partial \mathbf{n}^\top} \\
&= \frac{1}{2} \frac{\partial \log (|\mathbf{X}|_{\mathbf{n}} |\mathbf{Y}|_{-\mathbf{n}})}{\partial \mathbf{n}^\top} \\
&= \frac{1}{2} \frac{\partial \log (|\mathbf{C}_Y^{-1} \mathbf{C}_X \mathbf{C}_X^\top \mathbf{C}_X^{-\top}|_{\mathbf{n}})}{\partial \mathbf{n}^\top} = \log \operatorname{vecd}(\mathbf{C}_Y^{-1} \mathbf{C}_X),
\end{aligned}$$

where we used Lemma 2.2.2. □

Lemma 3.6.12. *Consider the multivariate gamma function as in Definition 2.2.2. We have*

$$\frac{\partial \log \Gamma_p(\mathbf{n})}{\partial \mathbf{n}^\top} = \left[\psi\left(n_1\right), \psi\left(n_2 - \frac{1}{2}\right), \ldots, \psi\left(n_p - \frac{1}{2}(p-1)\right)\right] = \boldsymbol{\psi}(\mathbf{n})^\top,$$
$$\frac{\partial \log \Gamma_p(\overleftarrow{\mathbf{n}})}{\partial \mathbf{n}^\top} = \overleftarrow{\boldsymbol{\psi}}(\overleftarrow{\mathbf{n}})^\top,$$

with $\boldsymbol{\psi}_i(\mathbf{n}) = \psi(n_i - \frac{1}{2}(i-1))$ *and* $\overleftarrow{\boldsymbol{\psi}}_i(\overleftarrow{\mathbf{n}}) = \psi(n_{p-i+1} - \frac{1}{2}(p-i))$*, where* $\psi(\cdot)$ *is defined in equation (5.2.2) of NIST Digital Library of Mathematical Functions. Furthermore*

$$\frac{\partial^2 \log \Gamma_p(\mathbf{n})}{\partial \mathbf{n} \partial \mathbf{n}^\top} = \mathrm{dg}(\boldsymbol{\psi}'(\mathbf{n})), \text{ and}$$
$$\frac{\partial^2 \log \Gamma_p(\mathbf{n})}{\partial \mathbf{n} \partial \mathbf{n}^\top} = \mathrm{dg}(\overleftarrow{\boldsymbol{\psi}}'(\overleftarrow{\mathbf{n}}))$$

with $\boldsymbol{\psi}'(\mathbf{n}) = \left[\psi'\left(n_1\right), \psi'\left(n_2 - \frac{1}{2}\right), \ldots, \psi'\left(n_p - \frac{1}{2}(p-1)\right)\right]^\top$ *and* $\overleftarrow{\boldsymbol{\psi}}'(\overleftarrow{\mathbf{n}})$ *similar to* $\overleftarrow{\boldsymbol{\psi}}_i(\overleftarrow{\mathbf{n}})$ *above. Finally,*

$$\frac{\partial \log \Gamma_p(n)}{\partial n} = \sum_{j=1}^{p} \psi\left(n - \frac{1}{2}(j-1)\right),$$

and

$$\frac{\partial^2 \log \Gamma_p(n)}{\partial n^2} = \sum_{j=1}^{p} \psi'\left(n - \frac{1}{2}(j-1)\right).$$

Proof.

$$\begin{aligned}
\frac{\partial \log \Gamma_p(\mathbf{n})}{\partial \mathbf{n}^\top} &= \frac{\partial \log\left(\pi^{p(p-1)/4} \prod_{j=1}^{p} \Gamma\left(n_j - \frac{1}{2}(j-1)\right)\right)}{\partial \mathbf{n}^\top} \\
&= \frac{\sum_{j=1}^{p} \partial \log\left(\Gamma\left(n_j - \frac{1}{2}(j-1)\right)\right)}{\partial \mathbf{n}^\top} \\
&= \left[\psi\left(n_1\right), \psi\left(n_2 - \frac{1}{2}\right), \ldots, \psi\left(n_p - \frac{1}{2}(p-1)\right)\right] \\
&= \boldsymbol{\psi}(\mathbf{n})^\top,
\end{aligned}$$

with $\boldsymbol{\psi}_i(\mathbf{n}) = \psi\left(n_i - \frac{1}{2}(i-1)\right)$ as defined in equation (5.2.2) of NIST Digital

Library of Mathematical Functions, and

$$\begin{aligned}\frac{\partial^2 \log \Gamma_p(\mathbf{n})}{\partial \mathbf{n} \partial \mathbf{n}^\top} &= \mathrm{dg}\left(\left[\psi'\left(n_1\right), \psi'\left(n_2 - \frac{1}{2}\right), \ldots, \psi'\left(n_p - \frac{1}{2}(p-1)\right)\right]\right) \\ &= \mathrm{dg}\left(\boldsymbol{\psi}'(\mathbf{n})\right)\end{aligned}$$

since for $i \neq j$, $\frac{\partial^2 \log \Gamma_p(\mathbf{n})}{\partial n_i \partial n_j} = 0$, with $\boldsymbol{\psi}_i'(\mathbf{n}) = \psi'\left(n_i - \frac{1}{2}(i-1)\right)$. Obviously, we have

$$\frac{\partial \log \Gamma_p(\overleftarrow{\mathbf{n}})}{\partial \overleftarrow{\mathbf{n}}^\top} = \boldsymbol{\psi}(\overleftarrow{\mathbf{n}})^\top,$$

such that

$$\frac{\partial \log \Gamma_p(\overleftarrow{\mathbf{n}})}{\partial \mathbf{n}^\top} = \overleftarrow{\frac{\partial \log \Gamma_p(\overleftarrow{\mathbf{n}})}{\partial \overleftarrow{\mathbf{n}}^\top}} =: \overleftarrow{\boldsymbol{\psi}}(\overleftarrow{\mathbf{n}})^\top.$$

Similarly for

$$\frac{\partial^2 \log \Gamma_p(\mathbf{n})}{\partial \mathbf{n} \partial \mathbf{n}^\top} = \mathrm{dg}(\overleftarrow{\boldsymbol{\psi}}'(\overleftarrow{\mathbf{n}})).$$

Finally,

$$\begin{aligned}\frac{\partial \log \Gamma_p(n)}{\partial n} &= \frac{\sum_{j=1}^p \partial \log\left(\Gamma\left(n - \frac{1}{2}(j-1)\right)\right)}{\partial n} \\ &= \sum_{j=1}^p \psi\left(n - \frac{1}{2}(j-1)\right),\end{aligned}$$

and

$$\frac{\partial^2 \log \Gamma_p(n)}{\partial n^2} = \sum_{j=1}^p \psi'\left(n - \frac{1}{2}(j-1)\right).$$

□

3.6.2 Scores and Fisher Information Matrices

Scores w.r.t. $\boldsymbol{\Omega}$

We want to derive the scores listed in Table 3.1. The scores w.r.t. $\boldsymbol{\Omega}$ of the Riesz, Inverse Riesz, t-Riesz, and Inverse t-Riesz distributions are easy to derive

by taking the logarithm of the p.d.f.s given in Table 2.5 and straightforwardly applying Lemmas 3.6.7 and 3.6.6. For the F-Riesz and Inverse F-Riesz, we have to start from the p.d.f. representations in equations (2.38) and (2.37), respectively, because their representations in Table 2.5 contain elements of the form $|\mathbf{I}+\mathbf{X}|_{\mathbf{n}}$, which are not necessarily positive definite and for which we can thus not apply Lemma 3.6.7. We show, as an example, the most complicated derivation, which is the one of the Inverse F-Riesz. We denote $\mathbf{J} = (\mathbf{\Omega}^{-1} + \mathbf{R}^{-1})^{-1}$ with lower Cholesky factor $\mathbf{C}_J$ to derive

$$\begin{aligned}
\frac{\partial \log p_{iF\mathcal{R}}}{\partial \text{vech}(\mathbf{\Omega})^\top} &= \frac{\partial \log |\mathbf{\Omega}|_{-\frac{\mathbf{n}}{2}}}{\partial \text{vech}(\mathbf{\Omega})^\top} + \frac{\partial \log |\mathbf{J}|_{\frac{\boldsymbol{\nu}+\mathbf{n}}{2}}}{\partial \text{vech}(\mathbf{J})^\top}\frac{\partial \text{vech}(\mathbf{J})}{\partial \text{vech}(\mathbf{\Omega})^\top} \\
&= -\frac{1}{2}\text{vec}(\mathbf{C}_\Omega^{-\top}\text{dg}(\mathbf{n})\mathbf{C}_\Omega^{-1})^\top\mathbf{G} \\
&\quad + \frac{1}{2}\text{vec}(\mathbf{C}_J^{-\top}\text{dg}(\mathbf{n}+\boldsymbol{\nu})\mathbf{C}_J^{-1})^\top\mathbf{G}\frac{\partial \text{vech}(\mathbf{J})}{\partial \text{vech}(\mathbf{J}^{-1})^\top}\frac{\partial \text{vech}(\mathbf{J}^{-1})}{\partial \text{vech}(\mathbf{\Omega})^\top} \\
&= -\frac{1}{2}\text{vec}(\mathbf{C}_\Omega^{-\top}\text{dg}(\mathbf{n})\mathbf{C}_\Omega^{-1})^\top\mathbf{G} \\
&\quad + \frac{1}{2}\text{vec}(\mathbf{C}_J^{-\top}\text{dg}(\mathbf{n}+\boldsymbol{\nu})\mathbf{C}_J^{-1})^\top\mathbf{G}\mathbf{G}^+(\mathbf{J}^{-1})^{-\otimes 2}\mathbf{G}\mathbf{G}^+\mathbf{\Omega}^{-\otimes 2}\mathbf{G} \\
&= -\frac{1}{2}\text{vec}(\mathbf{C}_\Omega^{-\top}\text{dg}(\mathbf{n})\mathbf{C}_\Omega^{-1})^\top\mathbf{G} \\
&\quad + \frac{1}{2}\text{vec}(\mathbf{C}_J^{-\top}\text{dg}(\mathbf{n}+\boldsymbol{\nu})\mathbf{C}_J^{-1})^\top\mathbf{G}\mathbf{G}^+(\mathbf{J}\mathbf{\Omega}^{-1})^{\otimes 2}\mathbf{G} \\
&= -\frac{1}{2}\text{vec}(\mathbf{C}_\Omega^{-\top}\text{dg}(\mathbf{n})\mathbf{C}_\Omega^{-1} + \mathbf{\Omega}^{-1}\mathbf{J}\mathbf{C}_J^{-\top}\text{dg}(\mathbf{n}+\boldsymbol{\nu})\mathbf{C}_J^{-1}\mathbf{J}\mathbf{\Omega}^{-1})^\top\mathbf{G} \\
&= -\frac{1}{2}\text{vec}(\mathbf{C}_\Omega^{-\top}\text{dg}(\mathbf{n})\mathbf{C}_\Omega^{-1} + \mathbf{\Omega}^{-1}\mathbf{C}_J\text{dg}(\mathbf{n}+\boldsymbol{\nu})\mathbf{C}_J^\top\mathbf{\Omega}^{-1})^\top\mathbf{G},
\end{aligned}$$

where apart from the aforementioned Lemmas we used Lemma 3.6.3 and equation (3.26). The scores w.r.t. $\mathbf{\Omega}$ of the other Riesz-type distributions are, as mentioned above, much simpler, and the ones of the Wishart-type distributions then follow by setting all elements in the degree of freedom parameter vectors of the corresponding Riesz-type score equal to each other. Finally, to construct Table 3.1, recall equation (3.33).

Scores w.r.t. $\mathbf{\Sigma}$

The scores w.r.t. $\mathbf{\Sigma}$ are then straightforward to derive using Lemma 3.6.9 and using the distribution specific maps from $\mathbf{\Omega}$ to $\mathbf{\Sigma}$ (see equation (2.9) and Table 2.3)

Fisher Information Matrices w.r.t. $\boldsymbol{\Omega}$

For the Riesz (and Wishart) and t-Riesz (and t-Wishart) distribution we will use

$$\begin{aligned}\mathcal{I}_{\mathcal{D}}^{\Omega} &= \mathbb{E}\left[\left(\frac{\partial \log p_{\mathcal{D}}}{\partial \operatorname{vech}\left(\boldsymbol{\Omega}^{-1}\right)^{\top}} \frac{\partial \operatorname{vech}\left(\boldsymbol{\Omega}^{-1}\right)}{\partial \operatorname{vech}\left(\boldsymbol{\Omega}\right)^{\top}}\right)^{\top} \frac{\partial \log p_{\mathcal{D}}}{\partial \operatorname{vech}\left(\boldsymbol{\Omega}^{-1}\right)^{\top}} \frac{\partial \operatorname{vech}\left(\boldsymbol{\Omega}^{-1}\right)}{\partial \operatorname{vech}\left(\boldsymbol{\Omega}\right)^{\top}}\right] \\ &= \left(\frac{\partial \operatorname{vech}\left(\boldsymbol{\Omega}^{-1}\right)}{\partial \operatorname{vech}\left(\boldsymbol{\Omega}\right)^{\top}}\right)^{\top} \mathcal{I}_{\mathcal{D}}^{(\Omega^{-1})} \frac{\partial \operatorname{vech}\left(\boldsymbol{\Omega}^{-1}\right)}{\partial \operatorname{vech}\left(\boldsymbol{\Omega}\right)^{\top}}. \end{aligned} \tag{3.34}$$

Riesz

We have

$$\begin{aligned}\mathcal{I}_{\mathcal{R}}^{(\Omega^{-1})} &= -\mathbb{E}\left[\frac{\partial^2 \log p_{\mathcal{R}}}{\partial \operatorname{vech}\left(\boldsymbol{\Omega}^{-1}\right) \partial \operatorname{vech}\left(\boldsymbol{\Omega}^{-1}\right)^{\top}}\right] \\ &= -\mathbb{E}\left[\frac{\partial^2 \log |\boldsymbol{\Omega}|_{-\frac{\mathbf{n}}{2}}}{\partial \operatorname{vech}\left(\boldsymbol{\Omega}^{-1}\right) \partial \operatorname{vech}\left(\boldsymbol{\Omega}^{-1}\right)^{\top}} - \frac{\partial^2 \frac{\operatorname{tr}\left(\boldsymbol{\Omega}^{-1}\mathbf{R}\right)}{2}}{\partial \operatorname{vech}\left(\boldsymbol{\Omega}^{-1}\right) \partial \operatorname{vech}\left(\boldsymbol{\Omega}^{-1}\right)^{\top}}\right] \\ &= \frac{1}{2}\mathbf{G}^{\top}\left(\mathbf{C}_{\Omega} \operatorname{dg}(\mathbf{n}) \otimes \mathbf{I}\right)\mathbf{F}^{\top}\left(\mathbf{G}^{\top}\left(\mathbf{C}_{\Omega} \otimes \mathbf{I}\right)\mathbf{F}^{\top}\right)^{-1}\mathbf{G}^{\top}\boldsymbol{\Omega}^{\otimes 2}\mathbf{G}. \end{aligned}$$

such that using equation (3.34) and Lemma (3.6.3), we have

$$\begin{aligned}\mathcal{I}_{\mathcal{R}}^{\Omega} &= \mathbf{G}^{\top}\boldsymbol{\Omega}^{-\otimes 2}(\mathbf{G}^{+})^{\top}\frac{1}{2}\mathbf{G}^{\top}(\mathbf{C}_{\Omega} \operatorname{dg}(\mathbf{n}) \otimes \mathbf{I})\mathbf{F}^{\top}(\mathbf{G}^{\top}(\mathbf{C}_{\Omega} \otimes \mathbf{I})\mathbf{F}^{\top})^{-1}\mathbf{G}^{\top}\boldsymbol{\Omega}^{\otimes 2}\mathbf{G} \\ &\quad \times \mathbf{G}^{+}\boldsymbol{\Omega}^{-\otimes 2}\mathbf{G} \\ &= \frac{1}{2}\mathbf{G}^{\top}\boldsymbol{\Omega}^{-\otimes 2}\mathbf{G}\mathbf{G}^{+}(\mathbf{C}_{\Omega} \operatorname{dg}(\mathbf{n}) \otimes \mathbf{I})\mathbf{F}^{\top}(\mathbf{G}^{+}(\mathbf{C}_{\Omega} \otimes \mathbf{I})\mathbf{F}^{\top})^{-1} \\ &= \frac{1}{2}\mathbf{G}^{\top}\boldsymbol{\Omega}^{-\otimes 2}(\mathbf{C}_{\Omega} \operatorname{dg}(\mathbf{n}) \otimes \mathbf{I})\mathbf{F}^{\top}(\mathbf{G}^{+}(\mathbf{C}_{\Omega} \otimes \mathbf{I})\mathbf{F}^{\top})^{-1} \\ &= \frac{1}{2}\mathbf{G}^{\top}(\mathbf{C}_{\Omega}^{-\top} \operatorname{dg}(\mathbf{n}) \otimes \boldsymbol{\Omega}^{-1})\mathbf{F}^{\top}(\mathbf{G}^{+}(\mathbf{C}_{\Omega} \otimes \mathbf{I})\mathbf{F}^{\top})^{-1}. \end{aligned} \tag{3.35}$$

Wishart

Starting from equation (3.35) and setting $n_1, \ldots, n_p = n$ we have

$$\begin{aligned}\mathcal{I}_{\mathcal{W}}^{\Omega} &= \frac{n}{2}\mathbf{G}^{\top}\boldsymbol{\Omega}^{-\otimes 2}\mathbf{G}\mathbf{G}^{+}(\mathbf{C}_{\Omega} \otimes \mathbf{I})\mathbf{F}^{\top}(\mathbf{G}^{+}(\mathbf{C}_{\Omega} \otimes \mathbf{I})\mathbf{F}^{\top})^{-1} \\ &= \frac{n}{2}\mathbf{G}^{\top}\boldsymbol{\Omega}^{-\otimes 2}\mathbf{G}. \end{aligned}$$

Inverse Riesz

$$\begin{aligned}\mathcal{I}^{\Omega}_{i\mathcal{R}} &= -\mathbb{E}\left[\frac{\partial^2 \log|\boldsymbol{\Omega}|_{\frac{\nu}{2}}}{\partial \text{vech}\,(\boldsymbol{\Omega})\,\partial \text{vech}\,(\boldsymbol{\Omega})^\top} - \frac{\partial^2 \frac{\text{tr}(\boldsymbol{\Omega}\mathbf{R}^{-1})}{2}}{\partial \text{vech}\,(\boldsymbol{\Omega})\,\partial \text{vech}\,(\boldsymbol{\Omega})^\top}\right] \\ &= \frac{1}{2}\mathbf{G}^\top(\mathbf{C}_\Omega^{-\top}\otimes\mathbf{C}_\Omega^{-\top}\text{dg}(\boldsymbol{\nu})\mathbf{C}_\Omega^{-1})\mathbf{F}^\top(\mathbf{G}^\top(\mathbf{C}_\Omega^{-\top}\otimes\boldsymbol{\Omega}^{-1})\mathbf{F}^\top)^{-1}\mathbf{G}^\top\boldsymbol{\Omega}^{-\otimes 2}\mathbf{G} \quad (3.36) \\ &= \frac{1}{2}\mathbf{G}^\top(\mathbf{C}_\Omega^{-\top}\otimes\mathbf{C}_\Omega^{-\top}\text{dg}(\boldsymbol{\nu})\mathbf{C}_\Omega^{-1})\mathbf{F}^\top(\mathbf{G}^+(\mathbf{C}_\Omega\otimes\mathbf{I})\mathbf{F}^\top)^{-1}\end{aligned}$$

Inverse Wishart

Starting form equation (3.36) and setting $\nu_1,\ldots,\nu_p = \nu$ we have

$$\begin{aligned}\mathcal{I}^{\Omega}_{i\mathcal{W}} &= \frac{\nu}{2}\mathbf{G}^\top(\mathbf{C}_\Omega^{-\top}\otimes\mathbf{C}_\Omega^{-\top}\mathbf{C}_\Omega^{-1})\mathbf{F}^\top(\mathbf{G}^\top(\mathbf{C}_\Omega^{-\top}\otimes\boldsymbol{\Omega}^{-1})\mathbf{F}^\top)^{-1}\mathbf{G}^\top\boldsymbol{\Omega}^{-\otimes 2}\mathbf{G} \\ &= \frac{\nu}{2}\mathbf{G}^\top\boldsymbol{\Omega}^{-\otimes 2}\mathbf{G}\end{aligned}$$

t-Riesz

We will need the expectation of

$$\begin{aligned}&\frac{\partial^2 \log(1+\text{tr}(\boldsymbol{\Omega}^{-1}\mathbf{R}))}{\partial \text{vech}(\boldsymbol{\Omega}^{-1})\partial \text{vech}(\boldsymbol{\Omega}^{-1})^\top} \\ &\qquad= \frac{\partial}{\partial \text{vech}(\boldsymbol{\Omega}^{-1})^\top}\left[(1+\text{tr}(\boldsymbol{\Omega}^{-1}\mathbf{R}))^{-1}\frac{\partial \text{tr}(\boldsymbol{\Omega}^{-1}\mathbf{R})}{\partial \text{vech}(\boldsymbol{\Omega}^{-1})}\right] \\ &\qquad= -(1+\text{tr}(\boldsymbol{\Omega}^{-1}\mathbf{R}))^{-2}\frac{\partial \text{tr}(\boldsymbol{\Omega}^{-1}\mathbf{R})}{\partial \text{vech}(\boldsymbol{\Omega}^{-1})}\frac{\partial \text{tr}(\boldsymbol{\Omega}^{-1}\mathbf{R})}{\partial \text{vech}(\boldsymbol{\Omega}^{-1})^\top} \\ &\qquad= -(1+\text{tr}(\boldsymbol{\Omega}^{-1}\mathbf{R}))^{-2}\mathbf{G}^\top\text{vec}^2(\mathbf{R})\mathbf{G}.\end{aligned}$$

For a better overview, denote the normalizing constant of the t-Riesz distribution by

$$c(\nu,\mathbf{n},\boldsymbol{\Omega}) = \frac{\Gamma\left((\nu+p\bar{\mathbf{n}})/2\right)}{\Gamma(\nu/2)\Gamma_p(\mathbf{n}/2)}|\boldsymbol{\Omega}|_{-\frac{\mathbf{n}}{2}}.$$

Then the desired expectation obtains from

$$
\begin{aligned}
&\mathbb{E}[(1+\operatorname{tr}(\mathbf{\Omega}^{-1}\mathbf{R}))^{-2}\operatorname{vec}^2(\mathbf{R})] \\
&= c(\nu,\mathbf{n},\mathbf{\Omega})\int_{\mathbf{R}>\mathbf{0}} \operatorname{vec}^2(\mathbf{R})|\mathbf{R}|_{\frac{\mathbf{n}-p-1}{2}}(1+\operatorname{tr}(\mathbf{\Omega}^{-1}\mathbf{R}))^{-\frac{\nu+p\bar{\mathbf{n}}+4}{2}}\mathrm{d}\mathbf{R} \\
&= \frac{c(\nu,\mathbf{n},\mathbf{\Omega})}{c(\nu+4,\mathbf{n},\mathbf{\Omega})}\mathbb{E}[\operatorname{vec}^2(\mathbf{R})] \text{ for } \mathbf{R}\sim t\mathcal{R}(\mathbf{n},\nu+4,\mathbf{\Omega}) \\
&= \frac{\Gamma((\nu+p\bar{\mathbf{n}})/2)}{\Gamma_p(\mathbf{n}/2)\Gamma(\nu/2)}\frac{\Gamma_p(\mathbf{n}/2)\Gamma((\nu+4)/2)}{\Gamma(((\nu+4)+p\bar{\mathbf{n}})/2)}\mathbb{E}[\operatorname{vec}^2(\mathbf{R})] \text{ for } \mathbf{R}\sim t\mathcal{R}(\mathbf{n},\nu+4,\mathbf{\Omega}) \\
&= \frac{(\nu+2)\nu}{(\nu+p\bar{\mathbf{n}})\,(\nu+p\bar{\mathbf{n}}+2)}\mathbb{E}[\operatorname{vec}^2(\mathbf{A})]\mathbb{E}[(\bar{b})^{-4}] \text{ for } \mathbf{A}\sim\mathcal{R}(\mathbf{\Omega},\mathbf{n}) \text{ and } \bar{b}\sim\chi_{\nu+4} \\
&= \frac{1}{(\nu+p\bar{\mathbf{n}})\,(\nu+p\bar{\mathbf{n}}+2)}(\operatorname{vec}^2(\mathbf{C}_\Omega\operatorname{dg}(\mathbf{n})\mathbf{C}_\Omega^\top)+2\mathbf{G}\mathbf{G}^+(\mathbf{C}_\Omega\operatorname{dg}(\mathbf{n})\mathbf{C}_\Omega^\top\otimes\mathbf{\Omega})\mathbf{G}\mathbf{G}^+),
\end{aligned}
$$

where we used the stochastic representation of the t-Riesz distribution, equation (2.39) and $(\bar{b})^{-2}$ following an Inverse Chi-squared distribution with[14]

$$
\mathbb{E}[(\bar{b})^{-2}] = (\nu+2)^{-1} \text{ and } \mathbf{Var}((\bar{b})^{-2}) = \frac{2}{(\nu+2)^2\nu},
$$

such that

$$
\mathbb{E}[(\bar{b})^{-4}] = \mathbb{E}[(\bar{b})^{-2}]^2 + \mathbf{Var}((\bar{b})^{-2}) = ((\nu+2)\nu)^{-1}. \tag{3.37}
$$

Now, using the derived expectation, we have

$$
\begin{aligned}
\mathcal{I}_{t\mathcal{R}}^{(\Omega^{-1})} &= -\mathbb{E}\left[\frac{\partial^2\log|\mathbf{\Omega}|_{-\frac{\mathbf{n}}{2}}}{\partial\operatorname{vech}(\mathbf{\Omega}^{-1})\partial\operatorname{vech}(\mathbf{\Omega}^{-1})^\top}+\frac{\nu+p\bar{\mathbf{n}}}{2}\frac{\partial^2\log\left(1+\operatorname{tr}(\mathbf{\Omega}^{-1}\mathbf{R})\right)}{\partial\operatorname{vech}(\mathbf{\Omega}^{-1})\partial\operatorname{vech}(\mathbf{\Omega}^{-1})^\top}\right] \\
&= \frac{1}{2}\mathbf{G}^\top(\mathbf{C}_\Omega\operatorname{dg}(\mathbf{n})\otimes\mathbf{I})\mathbf{F}^\top(\mathbf{G}^+(\mathbf{C}_\Omega\otimes\mathbf{I})\mathbf{F}^\top)^{-1}\mathbf{G}^+\mathbf{\Omega}^{\otimes 2}\mathbf{G} \\
&\quad - \frac{\nu+p\bar{\mathbf{n}}}{2(\nu+p\bar{\mathbf{n}})\,(\nu+p\bar{\mathbf{n}}+2)} \\
&\qquad \times\mathbf{G}^\top(\operatorname{vec}^2(\mathbf{C}_\Omega\operatorname{dg}(\mathbf{n})\mathbf{C}_\Omega^\top)+2\mathbf{G}\mathbf{G}^+(\mathbf{C}_\Omega\operatorname{dg}(\mathbf{n})\mathbf{C}_\Omega^\top\otimes\mathbf{\Omega})\mathbf{G}\mathbf{G}^+)\mathbf{G} \\
&= \frac{1}{2}\mathbf{G}^\top(\mathbf{C}_\Omega\operatorname{dg}(\mathbf{n})\otimes\mathbf{I})\mathbf{F}^\top(\mathbf{G}^+(\mathbf{C}_\Omega\otimes\mathbf{I})\mathbf{F}^\top)^{-1}\mathbf{G}^+\mathbf{\Omega}^{\otimes 2}\mathbf{G} \\
&\quad - \frac{1}{2(\nu+p\bar{\mathbf{n}}+2)}\mathbf{G}^\top(\operatorname{vec}^2(\mathbf{C}_\Omega\operatorname{dg}(\mathbf{n})\mathbf{C}_\Omega^\top)+2(\mathbf{C}_\Omega\operatorname{dg}(\mathbf{n})\mathbf{C}_\Omega^\top\otimes\mathbf{\Omega}))\mathbf{G}.
\end{aligned} \tag{3.38}
$$

Finally, use equation (3.34) to arrive at $\mathcal{I}_{t\mathcal{R}}^{\Omega}$.

14. Compare, e.g. Wikipedia.

t-Wishart

Starting form equation (3.38) and setting $n_1, \dots, n_p = n$, we have

$$\begin{aligned}
\mathcal{I}_{t\mathcal{W}}^{(\Omega^{-1})} &= \frac{n}{2}\mathbf{G}^\top \boldsymbol{\Omega}^{\otimes 2}\mathbf{G} - \frac{n}{2(\nu+pn+2)}\mathbf{G}^\top(n\mathrm{vec}^2(\boldsymbol{\Omega}) + 2\boldsymbol{\Omega}^{\otimes 2})\mathbf{G} \\
&= \frac{n}{2}\mathbf{G}^\top\left(\boldsymbol{\Omega}^{\otimes 2} - \frac{2}{\nu+pn+2}\boldsymbol{\Omega}^{\otimes 2} - \frac{n}{\nu+pn+2}\mathrm{vec}^2(\boldsymbol{\Omega})\right)\mathbf{G} \\
&= \frac{n}{2}\mathbf{G}^\top\left(\frac{\nu+pn}{\nu+pn+2}\boldsymbol{\Omega}^{\otimes 2} - \frac{n}{\nu+pn+2}\mathrm{vec}^2(\boldsymbol{\Omega})\right)\mathbf{G}.
\end{aligned}$$

Now, using equations (3.34), (3.26) and (3.23) and Lemma 3.6.3 we have

$$\begin{aligned}
\mathcal{I}_{t\mathcal{W}}^{\Omega} &= \frac{n}{2}(\mathbf{G}\mathbf{G}^+\boldsymbol{\Omega}^{-\otimes 2}\mathbf{G})^\top\left(\frac{\nu+pn}{\nu+pn+2}\boldsymbol{\Omega}^{\otimes 2} - \frac{n}{\nu+pn+2}\mathrm{vec}^2(\boldsymbol{\Omega})\right)\mathbf{G}\mathbf{G}^+\boldsymbol{\Omega}^{-\otimes 2}\mathbf{G} \\
&= \frac{n}{2}\mathbf{G}^\top\left(\frac{\nu+pn}{\nu+pn+2}\boldsymbol{\Omega}^{-\otimes 2} - \frac{n}{\nu+pn+2}\mathrm{vec}^2(\boldsymbol{\Omega}^{-1})\right)\mathbf{G}.
\end{aligned}$$

Inverse t-Riesz

Not derived since we don't need it in the paper. The derivation should be very similar to the one of the Inverse t-Wishart below, but at some point, we would need $\mathbb{E}[\mathrm{vec}^2(\mathbf{C}_\Omega^{-\top}\bar{\mathbf{B}}\bar{\mathbf{B}}^\top\mathbf{C}_\Omega^{-1})]$, which is probably very similar to derive as Lemma 2.7.3.

Inverse t-Wishart

This derivation is similar to the one of the t-Riesz distribution. We will need the expectation of

$$\begin{aligned}
\frac{\partial^2 \log(1+\mathrm{tr}(\boldsymbol{\Omega}\mathbf{R}^{-1}))}{\partial\mathrm{vech}(\boldsymbol{\Omega})\partial\mathrm{vech}(\boldsymbol{\Omega})^\top} &= \frac{\partial}{\partial\mathrm{vech}(\boldsymbol{\Omega})^\top}\left[(1+\mathrm{tr}(\boldsymbol{\Omega}\mathbf{R}^{-1}))^{-1}\frac{\partial\mathrm{tr}(\boldsymbol{\Omega}\mathbf{R}^{-1})}{\partial\mathrm{vech}(\boldsymbol{\Omega})}\right] \\
&= -(1+\mathrm{tr}(\boldsymbol{\Omega}\mathbf{R}^{-1}))^{-2}\frac{\partial\mathrm{tr}(\boldsymbol{\Omega}\mathbf{R}^{-1})}{\partial\mathrm{vech}(\boldsymbol{\Omega})}\frac{\partial\mathrm{tr}(\boldsymbol{\Omega}\mathbf{R}^{-1})}{\partial\mathrm{vech}(\boldsymbol{\Omega})^\top} \\
&= -(1+\mathrm{tr}(\boldsymbol{\Omega}\mathbf{R}^{-1}))^{-2}\mathbf{G}^\top\mathrm{vec}^2(\mathbf{R}^{-1})\mathbf{G}.
\end{aligned}$$

For a better overview, denote the normalizing constant of the Inverse t-Riesz distribution by

$$c(n,\nu,\boldsymbol{\Omega}) = \frac{\Gamma\left((n+p\nu)/2\right)}{\Gamma(n/2)\Gamma_p(\nu/2)}|\boldsymbol{\Omega}|^{\frac{\nu}{2}}.$$

Then the desired expectation obtains from

$$\begin{aligned}
&\mathbb{E}[(1+\text{tr}(\boldsymbol{\Omega}\mathbf{R}^{-1}))^{-2}\text{vec}^2(\mathbf{R}^{-1})] \\
&= c(n,\nu,\boldsymbol{\Omega})\int_{\mathbf{R}>\mathbf{0}} \text{vec}^2(\mathbf{R}^{-1})|\mathbf{R}|^{-\frac{\nu+p+1}{2}}(1+\text{tr}(\boldsymbol{\Omega}\mathbf{R}^{-1}))^{-\frac{n+p\nu+4}{2}}\text{d}\mathbf{R} \\
&= \frac{c(n,\nu,\boldsymbol{\Omega})}{c(n+4,\nu,\boldsymbol{\Omega})}\mathbb{E}[\text{vec}^2(\mathbf{R}^{-1})],\ \text{for } \mathbf{R}\sim it\mathcal{W}(\boldsymbol{\Omega},n+4,\nu) \\
&= \frac{\Gamma((n+p\nu)/2)}{\Gamma((n+4+p\nu)/2)}\frac{\Gamma((n+4)/2)}{\Gamma(n/2)}\mathbb{E}[(\underline{b})^{-4}]\mathbb{E}[\text{vec}^2(\mathbf{C}_\Omega^{-\top}\bar{\mathcal{B}}\bar{\mathcal{B}}^\top\mathbf{C}_\Omega^{-1})] \\
&= \frac{\nu}{(n+p\nu)(n+p\nu+2)}(2\mathbf{G}\mathbf{G}^+\boldsymbol{\Omega}^{-\otimes 2}(\mathbf{G}\mathbf{G}^+)^\top+\nu\text{vec}^2(\boldsymbol{\Omega}^{-1}))
\end{aligned}$$

where we used equation (3.37) and the fact that $\mathbf{C}_\Omega^{-\top}\bar{\mathcal{B}}\bar{\mathcal{B}}^\top\mathbf{C}_\Omega^{-1}\sim\mathcal{W}(\boldsymbol{\Omega}^{-1},\nu)$[15] and thus we used equation (2.39) with $\mathbf{n}=(\nu,\dots,\nu)$. Now, using the derived expectation we have, using Lemma 3.6.7 with $\mathbf{n}=(n,\dots,n)$,

$$\begin{aligned}
\mathcal{I}_{it\mathcal{W}}^{\Omega} &= -\mathbb{E}\left[\frac{\partial^2\log|\boldsymbol{\Omega}|^{\frac{\nu}{2}}}{\partial\text{vech}(\boldsymbol{\Omega})\partial\text{vech}(\boldsymbol{\Omega})^\top}\right]+\frac{n+p\nu}{2}\mathbb{E}\left[\frac{\partial^2\log\left(1+\text{tr}(\boldsymbol{\Omega}\mathbf{R}^{-1})\right)}{\partial\text{vech}(\boldsymbol{\Omega})\partial\text{vech}(\boldsymbol{\Omega})^\top}\right] \\
&= \frac{\nu}{2}\mathbf{G}^\top\boldsymbol{\Omega}^{-\otimes 2}\mathbf{G}-\frac{\nu}{2(n+p\nu+2)}\mathbf{G}^\top(2\mathbf{G}\mathbf{G}^+\boldsymbol{\Omega}^{-\otimes 2}(\mathbf{G}\mathbf{G}^+)^\top+\nu\text{vec}^2(\boldsymbol{\Omega}^{-1}))\mathbf{G} \\
&= \frac{\nu}{2}\mathbf{G}^\top\left(\boldsymbol{\Omega}^{-\otimes 2}-\frac{2}{n+p\nu+2}\boldsymbol{\Omega}^{-\otimes 2}-\frac{\nu}{n+p\nu+2}\text{vec}^2(\boldsymbol{\Omega}^{-1})\right)\mathbf{G} \\
&= \frac{\nu}{2}\mathbf{G}^\top\left(\frac{n+p\nu}{n+p\nu+2}\boldsymbol{\Omega}^{-\otimes 2}-\frac{\nu}{n+p\nu+2}\text{vec}^2(\boldsymbol{\Omega}^{-1})\right)\mathbf{G}.
\end{aligned}$$

<u>F-Riesz and Inverse F-Riesz</u>

It is not derived since we do not need it in the paper.[16]

<u>Matrix-F</u>

$$\begin{aligned}
&\frac{\partial^2\log p_F}{\partial\text{vech}(\boldsymbol{\Omega})\partial\text{vech}(\boldsymbol{\Omega})^\top}=\frac{\nu}{2}\frac{\partial^2\log|\boldsymbol{\Omega}|}{\partial\text{vech}(\boldsymbol{\Omega})\partial\text{vech}(\boldsymbol{\Omega})^\top}-\frac{n+\nu}{2}\frac{\partial^2\log|\boldsymbol{\Omega}+\mathbf{R}|}{\partial\text{vech}(\boldsymbol{\Omega})\partial\text{vech}(\boldsymbol{\Omega})^\top} \\
&=-\frac{\nu}{2}\mathbf{G}^\top\boldsymbol{\Omega}^{-\otimes 2}\mathbf{G}+\frac{n+\nu}{2}\mathbf{G}^\top(\boldsymbol{\Omega}+\mathbf{R})^{-\otimes 2}\mathbf{G} \\
&=-\frac{\nu}{2}\mathbf{G}^\top\boldsymbol{\Omega}^{-\otimes 2}\mathbf{G}+\frac{n+\nu}{2}\mathbf{G}^\top\left(\mathbf{C}_\Omega^{-\otimes 2}\right)^\top(\mathbf{I}+\mathbf{C}_\Omega^{-1}\mathbf{R}\mathbf{C}_\Omega^{-\top})^{-\otimes 2}\mathbf{C}_\Omega^{-\otimes 2}\mathbf{G}.
\end{aligned}$$

15. See Gupta and Nagar (2000).
16. Also, it seems very difficult.

Now, $\mathbf{Z}_{Kollo} = \mathbf{C}_\Omega^{-1}\mathbf{R}\mathbf{C}_\Omega^{-\top}$ follows a matrix F distribution with scale matrix $\mathbf{I}$ as defined in Theorem 2.4.9 of Kollo and von Rosen (2005). They derive on p. 265,

$$\begin{aligned}\mathbb{E}\left[\left(\mathbf{I}+\mathbf{C}_\Omega^{-1}\mathbf{R}\mathbf{C}_\Omega^{-\top}\right)^{-\otimes 2}\right] &= \left(c_3\mathbf{I}+c_4\mathbf{K}_{pp}+c_4\mathrm{vec}^2(\mathbf{I})\right)\\ &= \left((c_3-c_4)\mathbf{I}+c_4\left(\mathbf{I}+\mathbf{K}_{pp}\right)+c_4\mathrm{vec}^2(\mathbf{I})\right)\\ &= \left((c_3-c_4)\mathbf{I}+2c_4\mathbf{G}\mathbf{G}^+ +c_4\mathrm{vec}^2(\mathbf{I})\right),\end{aligned}$$

with, according to their p. 263,

$$\begin{aligned}c_4 &= \frac{n-p-1}{(n+\nu-1)(n+\nu+2)}\left(\left(n-p-2+\frac{1}{n+\nu}\right)c_2-\left(1+\frac{n-p-1}{n+\nu}\right)c_1\right),\\ c_3 &= \frac{n-p-1}{n+\nu}\left(\left(n-p-2\right)c_2-c_1\right)-\left(n+\nu+1\right)c_4,\\ c_2 &= \frac{n(\nu-p-2)+n^2+n}{(\nu-p)(\nu-p-1)(\nu-p-3)},\\ c_1 &= \frac{n^2(\nu-p-2)+2n}{(\nu-p)(\nu-p-1)(\nu-p-3)}.\end{aligned}$$

Thus

$$\begin{aligned}\mathcal{I}_F^\Omega &= -\mathbb{E}\left[\frac{\partial^2 \log p_F}{\partial\mathrm{vech}(\mathbf{\Omega})\partial\mathrm{vech}(\mathbf{\Omega})^\top}\right]\\ &= \frac{\nu}{2}\mathbf{G}^\top\mathbf{\Omega}^{-\otimes 2}\mathbf{G}-\frac{n+\nu}{2}\mathbf{G}^\top(\mathbf{C}_\Omega^{-\top})^{\otimes 2}((c_3-c_4)\mathbf{I}+2c_4\mathbf{G}\mathbf{G}^+ +c_4\mathrm{vec}^2(\mathbf{I}))\mathbf{C}_\Omega^{-\otimes 2}\mathbf{G}\\ &= \frac{\nu}{2}\mathbf{G}^\top\mathbf{\Omega}^{-\otimes 2}\mathbf{G}-\frac{n+\nu}{2}\mathbf{G}^\top((c_3-c_4)\mathbf{\Omega}^{-\otimes 2}+2c_4\mathbf{\Omega}^{-\otimes 2}+c_4\mathrm{vec}^2(\mathbf{\Omega}^{-1}))\mathbf{G}\\ &= \frac{1}{2}\mathbf{G}^\top((\nu-(n+\nu)(c_3+c_4))\mathbf{\Omega}^{-\otimes 2}-(n+\nu)c_4\mathrm{vec}^2(\mathbf{\Omega}^{-1}))\mathbf{G}.\end{aligned}$$

Fisher Information Matrices w.r.t. $\mathbf{\Sigma}$

Similar to equation (3.34), we can derive

$$\mathcal{I}_\mathcal{D}^\Sigma = \left(\frac{\partial\mathrm{vech}(\mathbf{\Omega})}{\partial\mathrm{vech}(\mathbf{\Sigma})^\top}\right)^\top \mathcal{I}_\mathcal{D}^\Omega \frac{\partial\mathrm{vech}(\mathbf{\Omega})}{\partial\mathrm{vech}(\mathbf{\Sigma})^\top}, \tag{3.39}$$

to get the Fisher information matrices w.r.t. $\mathbf{\Sigma}$.

Recall, that for the Wishart-type distributions, i.e. $\mathcal{D} \in (\mathcal{W}, \mathcal{W}, t\mathcal{W}, it\mathcal{W}, F)$, we have $\mathbf{\Omega} = m_\mathcal{D}^{-1}\mathbf{\Sigma}$, with distribution-specific scalar $m_\mathcal{D}$ and thus

$$\frac{\partial\mathrm{vech}(\mathbf{\Omega})}{\partial\mathrm{vech}(\mathbf{\Sigma})^\top} = \mathbf{I}m_\mathcal{D}^{-1}.$$

Since, for all Wishart-type distributions, $\mathbf{\Omega}^{-1}$ enters in quadratic form in $\mathcal{I}_{\mathcal{D}}^{\Omega}$, applying formula (3.39) with $\mathcal{I}_{\mathcal{D}}^{\Omega}$ rewritten in terms of $\mathbf{\Omega}$, amounts to just replacing $\mathbf{\Omega}$ with $\mathbf{\Sigma}$. For example,

$$\begin{aligned}\mathcal{I}_{\mathcal{W}}^{\Sigma} &= m_{\mathcal{D}}^{-2}\frac{n}{2}\mathbf{G}^{\top}\mathbf{\Omega}^{-\otimes 2}\mathbf{G} \\ &= m_{\mathcal{D}}^{-2}\frac{n}{2}\mathbf{G}^{\top}(m_{\mathcal{D}}^{-1}\mathbf{\Sigma})^{-\otimes 2}\mathbf{G} \\ &= \frac{n}{2}\mathbf{G}^{\top}\mathbf{\Sigma}^{-\otimes 2}\mathbf{G}.\end{aligned}$$

Thus, the expressions in Table 3.2 follow.

3.6.3 Proofs of Theorems in Paper

Proof of Theorem 3.3.1

Proof. We omit the subscripts t. Note (Table 3.2) that the $\mathcal{I}^{\Sigma}$ of all Wishart-type have the form

$$\mathcal{I}_{\mathcal{D}}^{\Sigma} = \mathbf{G}^{\top}(\alpha_{\mathcal{D}}\mathbf{\Sigma}^{\otimes 2} + c_{\mathcal{D}}\,\mathrm{vec}^2(\mathbf{\Sigma}))\mathbf{G},$$

where the scalars $\alpha_{\mathcal{D}}$ and $c_{\mathcal{D}}$ only depend on the d.o.f. parameter(s) of the respective distribution. Thus, using Lemma 3.6.1 we have

$$\begin{aligned}(\mathcal{I}^{\Sigma})^{-1} &= \alpha_{\mathcal{D}}(\mathbf{G}^{\top}(\mathbf{\Sigma}^{\otimes 2} + \alpha_{\mathcal{D}}^{-1}c_{\mathcal{D}}\mathrm{vec}^2(\mathbf{\Sigma}))\mathbf{G})^{-1} \\ &= \alpha_{\mathcal{D}}\mathbf{G}^{+}\left(\mathbf{\Sigma}^{-\otimes 2} + \frac{\alpha_{\mathcal{D}}^{-1}c_{\mathcal{D}}}{1+\alpha_{\mathcal{D}}^{-1}c_{\mathcal{D}}p}\mathrm{vec}^2(\mathbf{\Sigma}^{-1})\right)\mathbf{G}^{+} \\ &= \mathbf{G}^{+}\left(\alpha_{\mathcal{D}}\mathbf{\Sigma}^{-\otimes 2} + \frac{\alpha_{\mathcal{D}}c_{\mathcal{D}}}{\alpha_{\mathcal{D}}+c_{\mathcal{D}}p}\mathrm{vec}^2(\mathbf{\Sigma}^{-1})\right)\mathbf{G}^{+}.\end{aligned}$$

Define $\beta_{\mathcal{D}} = \frac{\alpha_{\mathcal{D}}c_{\mathcal{D}}}{\alpha_{\mathcal{D}}+c_{\mathcal{D}}p}$,

$$\begin{aligned}&\mathbf{G}^{+}(\alpha_{\mathcal{D}}\mathbf{\Sigma}^{\otimes 2} + \beta_{\mathcal{D}}\mathrm{vec}^2(\mathbf{\Sigma}))(\mathbf{G}^{+})^{\top}\mathbf{G}^{\top}\mathrm{vec}(\triangle_{\mathcal{D}}^{\Sigma}) \\ &\quad= \mathbf{G}^{+}(\alpha_{\mathcal{D}}\mathbf{\Sigma}^{\otimes 2} + \beta_{\mathcal{D}}\mathrm{vec}^2(\mathbf{\Sigma}))\mathrm{vec}(\triangle_{\mathcal{D}}^{\Sigma}) \\ &\quad= \alpha_{\mathcal{D}}\mathbf{G}^{+}\mathrm{vec}(\mathbf{\Sigma}\triangle_{\mathcal{D}}^{\Sigma}\mathbf{\Sigma}) + \beta_{\mathcal{D}}\mathrm{tr}(\mathbf{\Sigma}\triangle_{\mathcal{D}}^{\Sigma})\mathbf{G}^{+}\mathrm{vec}(\mathbf{\Sigma}) \\ &\quad= \alpha_{\mathcal{D}}\mathrm{vech}(\mathbf{\Sigma}\triangle_{\mathcal{D}}^{\Sigma}\mathbf{\Sigma}) + \beta_{\mathcal{D}}\mathrm{tr}(\mathbf{\Sigma}\triangle_{\mathcal{D}}^{\Sigma})\mathrm{vech}(\mathbf{\Sigma})\end{aligned}$$

where we used equations (3.25), (3.12), and (3.14). Now simply apply the *ivech* operator. □

Proof of Theorem 3.3.1

Proof. We omit the subscripts t in the proof. We have

$$
\begin{aligned}
(\mathcal{I}_{\mathcal{R}}^{\Sigma})^{-1}\nabla_{\mathcal{R}}^{\Sigma} &= \mathbb{E}\left[\left(\frac{\partial p_{\mathcal{R}}}{\partial \mathrm{vech}(\boldsymbol{\Omega})^{\top}}\frac{\partial \mathrm{vech}(\boldsymbol{\Omega})}{\partial \mathrm{vech}(\boldsymbol{\Sigma})^{\top}}\right)^{\top}\frac{\partial p_{\mathcal{R}}}{\partial \mathrm{vech}(\boldsymbol{\Omega})^{\top}}\frac{\partial \mathrm{vech}(\boldsymbol{\Omega})}{\partial \mathrm{vech}(\boldsymbol{\Sigma})^{\top}}\right]^{-1} \\
&\quad\times\left(\frac{\partial p_{\mathcal{R}}}{\partial \mathrm{vech}(\boldsymbol{\Omega})^{\top}}\frac{\partial \mathrm{vech}(\boldsymbol{\Omega})}{\partial \mathrm{vech}(\boldsymbol{\Sigma})^{\top}}\right)^{\top} \\
&= \left(\left(\frac{\partial \mathrm{vech}(\boldsymbol{\Omega})}{\partial \mathrm{vech}(\boldsymbol{\Sigma})^{\top}}\right)^{\top}\mathbb{E}\left[\left(\frac{\partial p_{\mathcal{R}}}{\partial \mathrm{vech}(\boldsymbol{\Omega})^{\top}}\right)^{\top}\frac{\partial p_{\mathcal{R}}}{\partial \mathrm{vech}(\boldsymbol{\Omega})^{\top}}\right]\frac{\partial \mathrm{vech}(\boldsymbol{\Omega})}{\partial \mathrm{vech}(\boldsymbol{\Sigma})^{\top}}\right)^{-1} \\
&\quad\times\left(\frac{\partial \mathrm{vech}(\boldsymbol{\Omega})}{\partial \mathrm{vech}(\boldsymbol{\Sigma})^{\top}}\right)^{\top}\left(\frac{\partial p_{\mathcal{R}}}{\partial \mathrm{vech}(\boldsymbol{\Omega})^{\top}}\right)^{\top} \\
&= \left(\frac{\partial \mathrm{vech}(\boldsymbol{\Omega})}{\partial \mathrm{vech}(\boldsymbol{\Sigma})^{\top}}\right)^{-1}(\mathcal{I}_{\mathcal{R}}^{\Omega})^{-1}\nabla_{\mathcal{R}}^{\Omega} = \frac{\partial \mathrm{vech}(\boldsymbol{\Sigma})}{\partial \mathrm{vech}(\boldsymbol{\Omega})^{\top}}(\mathcal{I}_{\mathcal{R}}^{\Omega})^{-1}\nabla_{\mathcal{R}}^{\Omega} \\
&= \mathbf{G}^{+}(\mathbf{C}_{\Omega}\mathrm{dg}(\mathbf{n})\otimes\mathbf{I})\mathbf{F}^{\top}(\mathbf{G}^{+}(\mathbf{C}_{\Omega}\otimes\mathbf{I})\mathbf{F}^{\top})^{-1} \\
&\quad\times(\mathbf{G}^{\top}\boldsymbol{\Omega}^{-\otimes 2}\mathbf{G}\mathbf{G}^{+}(\mathbf{C}_{\Omega}\,\mathrm{dg}(\mathbf{n})\otimes\mathbf{I})\mathbf{F}^{\top}(\mathbf{G}^{+}(\mathbf{C}_{\Omega}\otimes\mathbf{I})\mathbf{F}^{\top})^{-1})^{-1} \\
&\quad\times\mathbf{G}^{\top}\mathrm{vec}(\boldsymbol{\Omega}^{-1}\mathbf{R}\boldsymbol{\Omega}^{-1}-\mathbf{C}_{\Omega}^{-\top}\mathrm{dg}(\mathbf{n})\mathbf{C}_{\Omega}^{-1}) \\
&= (\mathbf{G}^{\top}\boldsymbol{\Omega}^{-\otimes 2}\mathbf{G})^{-1}\mathbf{G}^{\top}\mathrm{vec}(\boldsymbol{\Omega}^{-1}\mathbf{R}\boldsymbol{\Omega}^{-1}-\mathbf{C}_{\Omega}^{-\top}\mathrm{dg}(\mathbf{n})\mathbf{C}_{\Omega}^{-1}) \\
&= \mathbf{G}^{+}\boldsymbol{\Omega}^{\otimes 2}(\mathbf{G}\mathbf{G}^{+})^{\top}\mathrm{vec}(\boldsymbol{\Omega}^{-1}\mathbf{R}\boldsymbol{\Omega}^{-1}-\mathbf{C}_{\Omega}^{-\top}\mathrm{dg}(\mathbf{n})\mathbf{C}_{\Omega}^{-1}) \\
&= \mathbf{G}^{+}\boldsymbol{\Omega}^{\otimes 2}\mathrm{vec}(\boldsymbol{\Omega}^{-1}\mathbf{R}\boldsymbol{\Omega}^{-1}-\mathbf{C}_{\Omega}^{-\top}\mathrm{dg}(\mathbf{n})\mathbf{C}_{\Omega}^{-1}) \\
&= \mathrm{vech}(\mathbf{R}-\mathbf{C}_{\Omega}\mathrm{dg}(\mathbf{n})\mathbf{C}_{\Omega}^{\top}) = \mathrm{vech}(\mathbf{R}-\boldsymbol{\Sigma}),
\end{aligned}
$$

where $\mathcal{I}_{\mathcal{R}}^{\Omega}$ is given in equation (3.35) and we used Lemma 3.6.5, equations (3.27) and (3.26). Thus, the theorem follows from

$$\mathbf{S}_{\mathcal{R}}^{\Sigma} = \mathbf{R} - \boldsymbol{\Sigma}.$$

□

3.6.4 Additional Material

Assets:	Random	Mining	Random	Finance	Random	Manuf.
#Assets:	5	6	10	15	25	25
			$\hat{a}_1 \times 100$			
Wishart	1.704	1.209	0.902	0.515	0.348	0.354
Riesz	1.193	0.824	0.601	0.325	0.239	0.227
Inv.Wishart	0.112	0.556	0.332	0.119	0.161	0.173
Inv.Riesz	0.110	0.478	0.301	0.112	0.143	0.155
t-Wishart	1.038	0.777	0.669	0.439	0.352	0.341
t-Riesz	0.838	0.660	0.535	0.286	0.249	0.241
Inv.t-Wishart	0.708	0.585	0.464	0.311	0.221	0.214
Inv.t-Riesz	0.635	0.522	0.396	0.255	0.191	0.185
F	0.890	0.713	0.538	0.337	0.251	0.239
F-Riesz	0.470	0.366	0.300	0.203	0.168	0.156
Inv.F-Riesz	0.546	0.412	0.332	0.215	0.168	0.165
			$\hat{a}_2 \times 100$			
Wishart	0.619	0.516	0.263	0.206	0.086	0.083
Riesz	0.818	0.631	0.356	0.256	0.104	0.101
Inv.Wishart	0.319	0.397	0.212	0.142	0.065	0.068
Inv.Riesz	0.357	0.435	0.238	0.149	0.074	0.076
t-Wishart	4.247	4.361	3.278	3.501	2.082	2.194
t-Riesz	4.287	4.505	3.497	4.163	2.261	2.367
Inv.t-Wishart	3.237	3.399	2.713	3.103	1.653	1.733
Inv.t-Riesz	3.412	3.394	2.857	3.343	1.673	1.793
F	0.594	0.454	0.256	0.160	0.083	0.080
F-Riesz	1.152	0.902	0.489	0.290	0.141	0.137
Inv.F-Riesz	1.098	0.871	0.464	0.272	0.131	0.128

Table 3.11: Estimated scaled score parameters $\hat{a}_1 \times 100$ and $\hat{a}_2 \times 100$ for our restricted GAS model where d.o.f. parameters are restricted to be constant. All estimates are highly significant. The median (smallest) t-statistic of $\hat{a}_1 \times 100$ is 240 (61) and of $\hat{a}_2 \times 100$ is 151 (39).

Assets:	Random	Mining	Random	Finance	Random	Manuf.
#Assets:	5	6	10	15	25	25
Wishart	0.983	0.988	0.991	0.997	0.995	0.988
Riesz	0.989	0.992	0.996	0.999	0.997	0.993
Inv.Wishart	0.998	0.996	0.997	1.000	0.998	0.997
Inv.Riesz	0.998	0.996	0.998	1.000	0.999	0.997
t-Wishart	0.992	0.994	0.996	0.999	0.996	0.991
t-Riesz	0.994	0.995	0.996	0.999	0.997	0.994
Inv.t-Wishart	0.994	0.996	0.997	0.999	0.998	0.996
Inv.t-Riesz	0.995	0.996	0.998	1.000	0.998	0.997
F	0.993	0.995	0.996	0.999	0.998	0.996
F-Riesz	0.997	0.998	0.999	1.000	0.999	0.998
Inv.F-Riesz	0.997	0.997	0.999	1.000	0.999	0.998

Table 3.12: Persistence parameter $\hat{b}_1 + \hat{b}_2 + \hat{b}_3$ for our restricted GAS model where d.o.f. parameters are restricted to be constant. All estimated parameters are highly significant, with the median (minimum) t-statistic being 248 (72), 41 (10), and 76 (11), respectively.

Assets:	Random	Mining	Random	Finance	Random	Manuf.
#Assets:	5	6	10	15	25	25
Wishart	0.983	0.989	0.990	0.997	0.995	0.988
Riesz	0.989	0.993	0.996	0.999	0.997	0.994
Inv.Wishart	0.997	0.996	0.997	0.999	0.998	0.997
Inv.Riesz	0.998	0.996	0.998	1.000	0.999	0.997
t-Wishart	0.992	0.994	0.996	0.999	0.997	0.992
t-Riesz	0.994	0.995	0.997	0.999	0.998	0.996
Inv.t-Wishart	0.994	0.996	0.997	0.999	0.998	0.996
Inv.t-Riesz	0.996	0.996	0.998	1.000	0.999	0.997
F	0.994	0.995	0.997	0.999	0.998	0.997
F-Riesz	0.997	0.998	0.999	1.000	0.999	0.998
Inv.F-Riesz	0.997	0.997	0.999	1.000	0.999	0.998

Table 3.13: Persistence parameter $\hat{b}_1 + \hat{b}_2 + \hat{b}_3$. All estimated parameters are highly significant, with the median (minimum) t-statistic being 142 (75), 45 (10), and 75 (11), respectively.

Assets:	Random	Mining	Random	Finance	Random	Manuf.
#Assets:	5	6	10	15	25	25
			$\bar{\boldsymbol{\xi}}^{\mathbf{n}}$, ξ^n			
Wishart	18	21	26	34	46	47
Riesz	17	20	24	31	41	42
Inv.Wishart	-	-	-	-	-	-
Inv.Riesz	-	-	-	-	-	-
t-Wishart	27	32	35	44	53	53
t-Riesz	24	29	31	39	46	46
Inv.t-Wishart	33	32	39	31	48	49
Inv.t-Riesz	33	33	37	30	50	49
F	100	110	156	194	175	256
F-Riesz	80	86	108	125	144	155
Inv.F-Riesz	157	2480	1756	796	976	2539
			$\bar{\boldsymbol{\xi}}^{\boldsymbol{\nu}}$, ξ^ν			
Wishart	-	-	-	-	-	-
Riesz	-	-	-	-	-	-
Inv.Wishart	31	28	35	52	59	59
Inv.Riesz	21	26	32	41	53	53
t-Wishart	23	24	25	21	31	28
t-Riesz	24	23	27	22	34	31
Inv.t-Wishart	29	35	40	51	61	62
Inv.t-Riesz	27	33	37	47	56	57
F	30	36	44	57	86	81
F-Riesz	22	26	31	41	56	55
Inv.F-Riesz	32	39	45	54	66	67

Table 3.14: (Mean of) estimated intercept of d.o.f. parameters specification, $\bar{\boldsymbol{\xi}}^{\mathbf{n}}$ and $\bar{\boldsymbol{\xi}}^{\boldsymbol{\nu}}$ (ξ^n and ξ^ν for Wishart-type distributions), that is, (mean of) estimated unconditional mean of d.o.f parameters. All are significant at the 1 % significance level.

Assets:	Random	Mining	Random	Finance	Random	Manuf.
#Assets:	5	6	10	15	25	25
			$\widehat{b^{\mathbf{n}}}$, $\widehat{b^{n}}$			
Wishart	0.998	0.996	0.998	0.997	0.998	0.998
Riesz	0.997	0.992	0.996	0.996	0.998	0.998
Inv.Wishart	-	-	-	-	-	-
Inv.Riesz	-	-	-	-	-	-
t-Wishart	0.983	0.976	0.979	0.817	0.983	0.978
t-Riesz	0.996	0.993	0.998	0.990	0.998	0.998
Inv.t-Wishart	0.966	0.953	0.939	0.958	0.968	0.953
Inv.t-Riesz	0.913	0.948	0.944	0.959	0.965	0.952
F	0.671	1.000	0.545	0.743	1.000	0.851
F-Riesz	0.940	0.946	0.934	0.943	0.962	0.980
Inv.F-Riesz	0.991	0.961	0.998	0.713	0.964	0.977
			$\widehat{b^{\boldsymbol{\nu}}}$, $\widehat{b^{\nu}}$			
Wishart	-	-	-	-	-	-
Riesz	-	-	-	-	-	-
Inv.Wishart	1.000	0.939	0.908	0.999	0.997	0.980
Inv.Riesz	0.997	0.695	0.993	0.996	0.997	0.998
t-Wishart	0.947	0.953	0.969	0.902	0.953	0.936
t-Riesz	0.952	0.969	0.964	0.951	0.976	0.960
Inv.t-Wishart	0.963	0.923	0.930	0.791	0.945	0.939
Inv.t-Riesz	0.982	0.951	0.978	0.990	0.997	0.999
F	0.706	0.864	0.756	0.830	0.999	1.000
F-Riesz	0.002	0.225	0.972	0.924	0.993	0.988
Inv.F-Riesz	0.903	0.798	0.524	0.925	0.969	0.973

Table 3.15: Estimated GARCH parameters $\widehat{b^{\mathbf{n}}}$ and $\widehat{b^{\boldsymbol{\nu}}}$ ($\widehat{b^{n}}$ and $\widehat{b^{\nu}}$ for Wishart-type distributions). No background color indicates significance at the 1% level, light gray and mid-light gray indicate significance at the 5% and 10% level, dark gray indicates insignificance.

Chapter 4

Dynamic Principal Component CAW Models for High-Dimensional Realized Covariance Matrices[1]

Introduction 4.1

The modeling and forecasting of covariance matrices of asset returns is central to financial decision making since it provides a measurement of the risk involved in different investment allocations. It is specifically used in option pricing, risk management and portfolio allocation.

Traditionally multivariate GARCH (MGARCH) or stochastic volatility (MSV) models have been applied to estimate conditional covariance matrices from daily asset return vectors (see e.g. Bauwens, Laurent, and Rombouts, 2006 and Asai, McAleer, and Yu, 2006 for surveys). Nowadays, the increasing availability of intraday asset return information enables the computation of consistent ex-post measures of daily (co)variation of asset prices, so-called realized (co)variances (see e.g. Andersen et al., 2003 and Barndorff-Nielsen and Shephard, 2004). These realized measures can then be modeled directly in order to obtain forecasts of the covariance matrix of asset returns. The literature provides broad evidence that models for realized covariance matrices (RCs) provide more precise forecasts than MGARCH and MSV models (see e.g. Golosnoy, Gribisch, and Liesenfeld, 2012 and the references therein). Pioneering approaches are found in Gourieroux, Jasiak, and Sufana (2009), Chiriac and Voev (2011), Bauer and Vorkink (2011), Noureldin, Shephard, and Sheppard (2012), and Golosnoy, Gribisch, and Liesenfeld (2012).

1. This chapter is a version of the equally named article by Gribisch and Stollenwerk (2020) published in *Quantitative Finance*. It is joint work with Bastian Gribisch. The notation has been changed to match the one in previous chapters.

These models have in common that applications to high-dimensional covariance matrices (say, for more than ten assets) are complicated if not impossible and empirical applications typically do not exceed the ten-dimensional case.[2] Realistic portfolios, however, consist of a large number of assets which makes high-dimensional covariance matrix forecasting an important field of research. The development of models for high-dimensional applications is challenging since the dimension of the object of interest is proportional to the square of the number of assets. This results in a huge number of model parameters and renders one-step maximum likelihood (ML) estimation virtually impossible (the so-called *curse of dimensionality*). An important task is therefore, to develop multivariate volatility models which allow for feasible estimation in high-dimensional applications.

One strategy which has been proposed to overcome the curse of dimensionality is the use of sparsity assumptions like, e.g. sparse factor structures for the assets' covariance matrix (see e.g. Wang and Zou, 2010, Tao et al., 2011, Shen, Yao, and Li, 2020, Sheppard and Xu, 2019, Asai and McAleer, 2015, Jin, Maheu, and Yang, 2019). An alternative is to design multivariate volatility models such that their parameters can be iteratively estimated by multistep procedures. In particular, Bauwens, Storti, and Violante (2012) proposed the Realized DCC (Re-DCC) CAW model (see also Bauwens, Braione, and Storti, 2016 and Bauwens, Braione, and Storti, 2017 for applications and extensions), which resembles the DCC GARCH idea of Engle (2002) under the Conditional Autoregressive Wishart (CAW) setting of Golosnoy, Gribisch, and Liesenfeld (2012) for RCs. The model is applicable in high-dimensional settings via three-step estimation with correlation targeting, similar to the corresponding MGARCH model. Bauwens, Storti, and Violante (2012) provide an empirical application for 50 assets. While the DCC idea builds on decomposing the conditional covariance matrix in variances and correlations, which are then estimated independently, an alternative strand of literature constructs orthogonal components via a spectral decomposition (SD) of the covariance matrix. The most prominent model here is the orthogonal GARCH (OGARCH) model of Alexander and Chibumba (1997) and Alexander (2001), where the estimation output can be readily interpreted in terms of (conditional) principal component analysis. Aielli and Caporin (2015) introduce additional flexibility via allowing for dynamic loading matrices. The resulting model is then called Dynamic Principal Component (DPC) GARCH model. Similar to the DCC approach, the framework assumes the presence of an auxiliary process generating

2. In this paper, we follow the convention of labeling covariance matrices of up to ten assets as "small dimensional" and covariance matrices of up to 100 assets as "high-dimensional". We are not concerned with "vast-dimensional" or "large-dimensional" covariance matrices with more than 100 assets (compare e.g. Lunde, Shephard, and Sheppard, 2016, Sheppard and Xu, 2019 and Engle, Ledoit, and Wolf, 2019 for similar conventions).

orthonormal dynamic eigenvectors and allows for three-step estimation in order to be applicable in high-dimensional settings (the authors provide an application for up to 30 assets).

In this paper, we adapt the DPC-GARCH model of Aielli and Caporin (2015) to the modeling of high-dimensional RCs. The model structure is based on the CAW framework of Golosnoy, Gribisch, and Liesenfeld (2012), assuming a conditional central Wishart distribution for the RC. This particular distributional assumption allows for a convenient Quasi Maximum Likelihood (QML) interpretation implying consistency of one-step estimation even if the Wishart assumption is violated. We present a scalar version of the resulting DPC-CAW model and its estimation via a three-step approach similar to Aielli and Caporin (2015) in order to enable parameter estimation in high-dimensional settings. The three-step approach suffers from similar inconsistency problems as the DCC GARCH, the Re-DCC CAW, and the DPC-GARCH model. We therefore conduct an extensive simulation experiment which shows that biases are present but mainly affect the unconditional variances of lower order principal components which are of minor relevance for covariance forecasting. An out-of-sample forecasting experiment for 100-dimensional RCs finally shows that the DPC-CAW model has good forecasting properties in one-day, five-day, and ten-day-ahead forecasting and outperforms its competitors in particular in forecasting the correlation structure and the weights of the global minimum variance portfolio (GMVP), which are of high relevance in practical portfolio optimization. In particular, the DPC-CAW approach features significantly lower correlation and GMVP losses compared to up-to-date competitors like the flexible Factor-HEAVY approach of Sheppard and Xu (2019) and the Factor-CAW of Shen, Yao, and Li (2020).

The rest of the paper is organized as follows. In Section 4.2, we briefly review the concept of realized covariance matrices. Section 4.3 introduces the scalar-DPC-CAW model, including one-step and three-step ML estimation. Section 4.4 presents the results of a simulation experiment analyzing the bias and consistency of estimates obtained via the three-step approach. The empirical application to RCs for 100 NYSE traded stocks, including in-sample diagnostics and an extensive out-of-sample forecasting experiment, is presented in Section 4.5. Section 4.6 concludes.

4.2 Realized Covariance Matrices

Consider an p-dimensional vector of log-prices $\mathbf{p}(\tau)$, where $\tau \in \mathbb{R}_+$ represents continuous time. Assume that $\mathbf{p}(\tau)$ is a Brownian stochastic semimartingale with $(p \times p)$ spot covariance matrix $\mathbf{\Theta}(\tau)$. Without loss of generality restricting the

trading day to the unit interval, we obtain the "true" integrated covariance matrix at day t as $\mathbf{IC}_t = \int_{t-1}^{t} \mathbf{\Theta}(\tau)\, d\tau$.

Now assume that we observe $m+1$ uniformly spaced intraday log-prices. Then the j'th intraday return vector on day t $(t = 1, \ldots, T)$ is given by

$$\mathbf{r}_{j,t} = \mathbf{p}\left(t - 1 + \frac{j}{m}\right) - \mathbf{p}\left(t - 1 + \frac{j-1}{m}\right), \quad j = 1, \ldots, m.$$

Let the $(p \times p)$ matrix $\mathbf{R}_t$ denote an RC, i.e. a non-parametric ex-post estimate of $\mathbf{IC}_t$ exploiting high-frequency asset return information. A well-known example for $\mathbf{R}_t$ is the standard realized covariance matrix, which is defined as

$$\mathbf{R}_t = \sum_{j=1}^{m} \mathbf{r}_{j,t}\mathbf{r}_{j,t}^{\top}. \tag{4.1}$$

In the absence of market microstructure noise and discontinuous price jumps, it can be shown that $\mathbf{R}_t$ is a consistent estimator of $\mathbf{IC}_t$ as $m \to \infty$ (see Barndorff-Nielsen and Shephard, 2004). If the observed intraday price data contains microstructure noise, jumps, or non-synchronous trading, one can employ one of several alternatives to the standard realized covariance matrix, such as the multivariate realized kernel of Barndorff-Nielsen et al. (2011).

4.3 The DPC-CAW Model

We model the time-evolution of p-dimensional stochastic positive-definite RCs $\{\mathbf{R}_t\}_{t=1}^{T}$. Given the filtration $\mathcal{F}_{t-1} = \{\mathbf{R}_{t-1}, \mathbf{R}_{t-2}, \ldots\}$, $\mathbf{R}_t$ is assumed to follow a central Wishart distribution

$$\mathbf{R}_t | \mathcal{F}_{t-1} \sim \mathcal{W}(\frac{\mathbf{\Sigma}_t}{n}, n), \tag{4.2}$$

where $n \geq p$ is the scalar degrees of freedom, and $\mathbf{\Sigma}_t/n$ denotes the symmetric, positive definite $p \times p$ scale matrix,[3] such that

$$\mathbb{E}[\mathbf{R}_t | \mathcal{F}_{t-1}] = \mathbf{\Sigma}_t. \tag{4.3}$$

Furthermore, let

$$\mathbf{\Sigma}_t = \mathbf{L}_t \mathbf{D}_t \mathbf{L}_t^{\top} \tag{4.4}$$

3. In this chapter, the Wishart distribution is parameterized by the parameter (a.k.a. scale) matrix $\mathbf{\Omega}_t = \mathbf{\Sigma}_t/n$, where $\mathbf{\Omega}$ is the same as in Chapter 2.

denote the SD of the conditional mean of $\mathbf{R}_t$, where the diagonal elements of $\mathbf{D}_t = \mathrm{dg}(d_{1,t}, d_{2,t}, \ldots, d_{p,t})$ are the eigenvalues of $\mathbf{\Sigma}_t$ and the columns of $\mathbf{L}_t$ are the associated orthonormal eigenvectors (see e.g. Lütkepohl, 1996, p. 69). We are interested in building a forecasting model for $\mathbf{R}_t$ where both the eigenvalues and the eigenvectors are allowed to vary persistently over time and which allows for convenient sequential estimation in high-dimensional applications.

4.3.1 Eigenvector Driving Process

In order to obtain time-varying orthonormal eigenvectors $\mathbf{L}_t$ in equation (4.4), we introduce a matrix-variate auxiliary process $\{\mathbf{Q}_t\}$ from which the eigenvectors $\mathbf{L}_t$ are obtained via computing the conditional SD of $\mathbf{Q}_t$. The auxiliary process is defined as a scalar-BEKK-type recursion (see Engle and Kroner, 1995) for RCs:

$$\mathbf{Q}_t = (1 - a - b)\mathbf{\Xi} + a\mathbf{R}_{t-1} + b\mathbf{Q}_{t-1}, \tag{4.5}$$

$$\mathbf{Q}_t = \mathbf{L}_t \mathbf{G}_t \mathbf{L}_t^\top. \tag{4.6}$$

The scalars a and b and the intercept matrix $\mathbf{\Xi}$ are parameters to be estimated. Time-varying orthonormal eigenvectors $\mathbf{L}_t$ are generated by the conditional SD of $\mathbf{Q}_t$ in equation (4.6). The diagonal matrix of eigenvalues $\mathbf{G}_t$ obtains as a "residual", which is of no further interest.

We consider model (4.5)-(4.6) as the true data generating process (DGP) for the loading matrices. Note that we may generalize the scalar dynamics of equation (4.5) to full BEKK dynamics (see Aielli and Caporin, 2015 and Noureldin, Shephard, and Sheppard, 2014 for details). However, estimation of such a "complete model" in high-dimensional settings is practically impossible since the number of autoregressive parameters is of order $\mathcal{O}(p^2)$ (the *curse of dimensionality*). We therefore restrict the model to feasible scalar dynamics similar to the popular DCC-GARCH approach.

The SD in equation (4.6) is not uniquely identified. Following Aielli and Caporin (2015) we therefore impose on all SDs within the model except equation (4.4) that the eigenvalues are arranged in strictly decreasing order.

The sign of each eigenvector is still unidentified. However, within the model, the eigenvector matrix appears only in quadratic form. Hence there is no need for imposing a sign restriction. The implicit assumption that the eigenvalues of $\mathbf{Q}_t$ are distinct holds almost surely and is thus mild.

In order to ensure that $\mathbf{Q}_t$ is always positive definite we furthermore impose that $0 \leq a,\ 0 \leq b,\ a + b < 1$ and $\mathbf{\Xi}$ and $\mathbf{Q}_0$ are positive definite.

We require an additional constraint on $\mathbf{\Xi}$ in order to ensure a unique sequence of eigenvectors. To see this intuitively, multiply equation (4.5) by some positive

constant c. Given the data $\{\mathbf{R}_t\}_{t=1}^T$ this would produce the same eigenvector matrix series $\{\mathbf{L}_t\}_{t=1}^T$ since $c\mathbf{Q}_t = c\mathbf{L}_t\mathbf{G}_t\mathbf{L}_t^\top = \mathbf{L}_t c\mathbf{G}_t\mathbf{L}_t^\top = \mathbf{L}_t\tilde{\mathbf{G}}_t\mathbf{L}_t^\top$. Identification can be ensured by restricting the magnitude of the intercept matrix $\boldsymbol{\Xi}$ as detailed in Section 4.3.2 below.

4.3.2 Eigenvalue Driving Process

The previous section discussed the eigenvector generating process for the SD in equation (4.4). What remains in order to define the covariance matrix forecast (4.4) is a model for the dynamics of the eigenvalues in $\mathbf{D}_t$. We employ p independent GARCH-type recursions in order to capture the dynamics of the diagonal elements of $\mathbf{D}_t$. Let

$$d_{i,t} = (1 - \alpha_i - \beta_i)\gamma_i + \alpha_i g_{i,t-1} + \beta_i d_{i,t-1}, \tag{4.7}$$

where $g_{i,t} = \mathbf{e}_i^\top\mathbf{L}_t^\top\mathbf{R}_t\mathbf{L}_t\mathbf{e}_i$ with $\mathbf{e}_i$ being an $p\times 1$ vector of zeros with a 1 at the i'th position. That is, $g_{i,t}$ is the i'th diagonal element of the random matrix $\mathbf{L}_t^\top\mathbf{R}_t\mathbf{L}_t$. Generalizations of model (4.7) obtained by increasing the lag order or e.g. including HAR-type dynamics (see Corsi, 2009) are straightforward to implement.

Note that

$$\mathbb{E}[\mathbf{L}_t^\top\mathbf{R}_t\mathbf{L}_t|\mathcal{F}_{t-1}] = \mathbf{L}_t^\top\mathbb{E}[\mathbf{R}_t|\mathcal{F}_{t-1}]\mathbf{L}_t = \mathbf{L}_t^\top\mathbf{L}_t\mathbf{D}_t\mathbf{L}_t^\top\mathbf{L}_t = \mathbf{D}_t, \tag{4.8}$$

such that

$$\mathbb{E}[g_{i,t}|\mathcal{F}_{t-1}] = \mathbb{E}[\mathbf{e}_i^\top\mathbf{L}_t^\top\mathbf{R}_t\mathbf{L}_t\mathbf{e}_i|\mathcal{F}_{t-1}] \tag{4.9}$$

$$= \mathbf{e}_i^\top\,\mathbb{E}[\mathbf{L}_t^\top\mathbf{R}_t\mathbf{L}_t|\mathcal{F}_{t-1}]\,\mathbf{e}_i = \mathbf{e}_i^\top\mathbf{D}_t\mathbf{e}_i = d_{i,t}. \tag{4.10}$$

Under the usual stationarity condition we then obtain

$$\mathbb{E}[d_{i,t}] = \gamma_i. \tag{4.11}$$

We now employ the SD of the intercept matrix $\boldsymbol{\Xi}$ of the eigenvector generating auxiliary process $\mathbf{Q}_t$ in equation (4.5), $\boldsymbol{\Xi} = \underline{\mathbf{L}}\,\underline{\mathbf{D}}\,\underline{\mathbf{L}}^\top$, where $\underline{\mathbf{D}} = \mathrm{dg}(\{\underline{d}_i\}_{i=1}^p)$, and set

$$\gamma_i = \underline{d}_i, \quad i = 1, \ldots, p. \tag{4.12}$$

That is $\{\underline{d}_i\}_{i=1}^n$ are the eigenvalues of the intercept matrix in the eigenvector driving recursion (see equation 4.5). In summary, we impose that $\gamma_i = \underline{d}_i,\ 0 \leq \alpha_i,\ 0 \leq \beta_i,\ \alpha_i + \beta_i < 1,\ 0 < d_{i,0}\ \ \forall i$. Since all parameters are restricted to be positive, this assumption also ensures that $d_{i,t}$ is always positive and consequently $\boldsymbol{\Sigma}_t$ is always positive definite.

The targeting-like constraint of setting $\gamma_i = \underline{d}_i$ solves the problem of identifying a unique $\mathbf{L}_t$-sequence via the $\mathbf{Q}_t$ auxiliary process since it implicitly imposes

$$\text{tr}(\mathbb{E}[\mathbf{R}_t]) = \text{tr}(\boldsymbol{\Xi}). \tag{4.13}$$

Hence the magnitude of the intercept matrix $\boldsymbol{\Xi}$ is restricted, which precludes the possibility of scaling the $\{\mathbf{Q}_t\}$ sequence by a constant c.

Proof.

$$\begin{aligned}\text{tr}(\mathbb{E}[\mathbf{R}_t]) &= \text{tr}(\mathbb{E}[\mathbb{E}[\mathbf{R}_t|\mathcal{F}_{t-1}]]) = \text{tr}(\mathbb{E}[\boldsymbol{\Sigma}_t]) = \mathbb{E}[\text{tr}(\mathbf{L}_t\mathbf{D}_t\mathbf{L}_t^\top)] = \mathbb{E}[\text{tr}(\mathbf{D}_t\mathbf{L}_t^\top\mathbf{L}_t)] \\ &= \mathbb{E}[\text{tr}(\mathbf{D}_t)] = \text{tr}(\underline{\mathbf{D}}) = \text{tr}(\underline{\mathbf{D}}\,\underline{\mathbf{L}}^\top\underline{\mathbf{L}}) = \text{tr}(\underline{\mathbf{L}}\,\underline{\mathbf{D}}\,\underline{\mathbf{L}}^\top) = \text{tr}(\boldsymbol{\Xi}),\end{aligned} \tag{4.14}$$

where we used the trace property $\text{tr}(\mathbf{ABC}) = \text{tr}(\mathbf{CAB}) = \text{tr}(\mathbf{BCA})$ and orthonormality of $\mathbf{L}_t$ and $\underline{\mathbf{L}}$. □

While this is not the only way to achieve identification of the eigenvector sequence $\{\mathbf{L}_t\}$, it does entail an appealing interpretation of the model. Specifically, if $a = b = 0$, the eigenvector driving process collapses to the constant matrix $\mathbf{Q}_t = \boldsymbol{\Xi}$, such that $\mathbf{L}_t = \underline{\mathbf{L}}$. The resulting specification resembles the popular orthogonal GARCH model of Alexander and Chibumba (1997) and Alexander (2001), such that the DPC-CAW is regarded as being a dynamic generalization of the OGARCH to the modeling of RCs.

Recall that we assumed that the diagonal elements of $\underline{\mathbf{D}}$ are arranged in decreasing order in order to identify the model. This implies that

$$\mathbb{E}[d_{1,t}] > \mathbb{E}[d_{2,t}] > \ldots > \mathbb{E}[d_{n,t}]. \tag{4.15}$$

This, however, does not imply that individual elements of $\mathbf{d}_t$ themselves are arranged in decreasing order since the random variables $g_{i,t}$ are not bounded above (recall that we did not need to impose identifying restrictions on the eigenvalue ordering for the SD of the covariance matrix forecast $\boldsymbol{\Sigma}_t$ in equation 4.4). A situation where $d_{i,t} < d_{i-1,t}$ happens particularly often in high dimensional applications where the elements of $\underline{\mathbf{d}}$ are close to each other. Note that this is not a drawback but merely reflects the fact that the conditional ordering of the eigenvalues may deviate from their unconditional ordering. As an example, unconditionally, the second principal component explains a lower fraction of the total volatility than the first. But conditionally, there may exist periods where the second component dominates the first. Such situations are well-known in the context of factor analysis (see also similar argumentations in Aielli and Caporin, 2015). In fact, it is possible to restrict the ordering of the conditional eigenvalues by, e.g. modeling their

positive increments. However, this would impose unnecessary and overidentifying restrictions on the model.

The conditional Wishart assumption for $\mathbf{R}_t$ in equation (4.2) implies a conditional Gamma distribution for $g_{i,t}$.

$$g_{i,t}|\mathcal{F}_{t-1} \sim \text{Gamma}(n/2, 2d_{i,t}/n). \tag{4.16}$$

Proof. Consider the following theorem of Rao (1965):

Theorem 4.3.1. *If an $p \times p$ random matrix $\mathbf{Y}$ has a central Wishart distribution with n degrees of freedom and scale matrix $\mathbf{\Omega}$, that is $\mathbf{Y} \sim \mathcal{W}_p(\mathbf{\Omega}, n)$, and $\mathbf{X}$ is a $q \times p$ matrix of rank q, then:*

$$\mathbf{XYX}^\top \sim \mathcal{W}_q(\mathbf{X\Omega X}^\top, n).$$

Set $\mathbf{X} = \mathbf{e}_i^\top \mathbf{L}_t^\top$, where $\mathbf{e}_i$ is defined as above, $\mathbf{\Omega} = \mathbf{\Sigma}_t/n$ and $\mathbf{Y} = \mathbf{R}_t$ to obtain

$$\begin{aligned}
\mathbf{XYX}^\top &= \mathbf{e}_i^\top \mathbf{L}_t^\top \mathbf{R}_t \mathbf{L}_t \mathbf{e}_i = g_{i,t} \\
\text{and } \mathbf{X\Omega X}^\top &= \frac{1}{n}\mathbf{e}_i^\top \mathbf{L}_t^\top \mathbf{\Sigma}_t \mathbf{L}_t \mathbf{e}_i = \frac{1}{n}\mathbf{e}_i^\top \mathbf{L}_t^\top \mathbf{L}_t \mathbf{D}_t \mathbf{L}_t^\top \mathbf{L}_t \mathbf{e}_i = \frac{1}{n}\mathbf{e}_i^\top \mathbf{D}_t \mathbf{e}_i = \frac{1}{n} d_{i,t}
\end{aligned}$$

such that

$$g_{i,t}|\mathcal{F}_{t-1} \sim \mathcal{W}_1\left(\frac{d_{i,t}}{n}, n\right). \tag{4.17}$$

Since the univariate Wishart resembles the Gamma density, $g_{i,t}$ follows a conditional gamma distribution with shape parameter $n/2$ and scale parameter $2d_{i,t}/n$:

$$g_{i,t}|\mathcal{F}_{t-1} \sim \text{Gamma}(n/2, 2d_{i,t}/n).$$

□

Equations (4.2) - (4.7) then constitute the scalar-DPC-CAW model, which is summarized by the distributional assumption $\mathbf{R}_t|\mathcal{F}_{t-1} \sim \mathcal{W}(\mathbf{\Sigma}_t/n, n)$ together with the following set of equations for $t = 1, \ldots, T$:

$$\begin{aligned}
\mathbf{\Sigma}_t &= \mathbf{L}_t \mathbf{D}_t \mathbf{L}_t^\top \\
\mathbf{Q}_t &= (1 - a - b)\,\mathbf{\Xi} + a\mathbf{R}_{t-1} + b\mathbf{Q}_{t-1} \\
\mathbf{Q}_t &= \mathbf{L}_t \mathbf{G}_t \mathbf{L}_t^\top \\
d_{i,t} &= (1 - \alpha_i - \beta_i)\gamma_i + \alpha_i g_{i,t-1} + \beta_i d_{i,t-1}, \quad \gamma_i = \underline{d}_i, \quad i = 1, \ldots, p \\
g_{i,t} &= \mathbf{e}_i^\top \mathbf{L}_t^\top \mathbf{R}_t \mathbf{L}_t \mathbf{e}_i, \quad i = 1, \ldots, p
\end{aligned}$$

where $\boldsymbol{\Xi} = \underline{\mathbf{L}}\,\underline{\mathbf{D}}\,\underline{\mathbf{L}}^\top$, $\underline{\mathbf{D}} = \mathrm{dg}(\{\underline{d}_i\}_{i=1}^p)$. The parameters of the DPC-CAW model are comprised in the parameter vector $\boldsymbol{\theta}$ with $\boldsymbol{\theta} = (\mathrm{vech}(\boldsymbol{\Xi})^\top, a, b, \{\alpha_i, \beta_i\}_{i=1}^p, n)^\top$ and $0 \leq a,\ 0 \leq b,\ a + b < 1$ and $\boldsymbol{\Xi}$ and $\mathbf{Q}_0$ are positive definite, $0 \leq \alpha_i,\ 0 \leq \beta_i,\ \alpha_i + \beta_i < 1,\ 0 < d_{i,0}\ \forall i$. In our empirical application below, we initialize the eigenvector and eigenvalue recursions by $\mathbf{Q}_0 = \frac{1}{T}\sum_{t=1}^T \mathbf{R}_t$ and $d_{i,0} = \frac{1}{T}\sum_{t=1}^T g_{i,t}$.

Maximum Likelihood Estimation 4.3.3

One-Step Estimation

Low-dimensional settings (say, up to five assets) allow for one-step estimation of the model parameters $\boldsymbol{\theta} = (\mathrm{vech}(\boldsymbol{\Xi})^\top, a, b, \{\alpha_i, \beta_i\}_{i=1}^p, n)^\top$ of the DPC-CAW model. Estimation can then be carried out by maximizing the log-likelihood function

$$\begin{aligned}\mathcal{L}(\boldsymbol{\theta}) = \sum_{t=1}^T \Bigg(& \frac{np}{2}\log\left(\frac{n}{2}\right) - \frac{p(p-1)}{4}\log(\pi) \\ & - \sum_{i=1}^p \log\Gamma\left(\frac{n+1-i}{2}\right) + \left(\frac{n-p-1}{2}\right)\log|\mathbf{R}_t| \\ & - \frac{n}{2}\Big(\log|\boldsymbol{\Sigma}_t(\boldsymbol{\psi})| + \mathrm{tr}(\boldsymbol{\Sigma}_t(\boldsymbol{\psi})^{-1}\mathbf{R}_t)\Big)\Bigg),\end{aligned} \tag{4.18}$$

where $\boldsymbol{\psi}$ summarizes the parameters for the $\mathbf{Q}_t$ and $d_{i,t}$ recursions, such that $\boldsymbol{\theta} = (\boldsymbol{\psi}, n)^\top$. The parameter n can be treated as a nuisance parameter due to its irrelevance for the matrix forecast (see equation 4.3). In fact, the first order conditions for the maximization of the log-likelihood with respect to $\boldsymbol{\psi}$ are proportional to n. Then

$$\hat{\boldsymbol{\psi}} = \underset{\boldsymbol{\psi}}{argmax}\ \mathcal{L}^*(\boldsymbol{\psi}), \tag{4.19}$$

with

$$\mathcal{L}^*(\boldsymbol{\psi}) = \sum_{t=1}^T -\frac{1}{2}\left(\log|\boldsymbol{\Sigma}_t(\boldsymbol{\psi})| + \mathrm{tr}\left((\boldsymbol{\Sigma}_t(\boldsymbol{\psi}))^{-1}\mathbf{R}_t\right)\right). \tag{4.20}$$

The score vector of observation t obtains as

$$\mathbf{s}_t(\boldsymbol{\psi}) = \frac{1}{2}\left((\mathrm{vec}(\mathbf{R}_t)^\top - \mathrm{vec}(\boldsymbol{\Sigma}_t)^\top)(\boldsymbol{\Sigma}_t^{-1}\otimes\boldsymbol{\Sigma}_t^{-1})\frac{\partial\mathrm{vec}(\boldsymbol{\Sigma}_t)}{\partial\boldsymbol{\psi}^\top}\right)^\top. \tag{4.21}$$

Assuming a correctly specified mean $\mathbb{E}[\mathbf{R}_t|\mathcal{F}_{t-1}] = \boldsymbol{\Sigma}_t$, $\mathbf{s}_t(\boldsymbol{\psi})$ is a martingale difference sequence since

$$\mathbb{E}[\mathbf{s}_t(\boldsymbol{\psi})|\mathcal{F}_{t-1}] = \mathbf{0}. \tag{4.22}$$

Consequently, as noted by Bauwens, Storti, and Violante (2012) and Noureldin, Shephard, and Sheppard (2012), under the usual regularity conditions (see e.g. Wooldridge, 1994) $\hat{\boldsymbol{\psi}}$ is consistent and asymptotically normal even if the Wishart assumption is violated, provided that the conditional mean is correctly specified. Hence equation (4.20) can be interpreted as a quasi-log-likelihood (QL). From the QL function in equation (4.20) we obtain the period-t log-likelihood contribution

$$\begin{aligned}
\ell_t^* &= -\frac{1}{2}\left(\log|\boldsymbol{\Sigma}_t| + \operatorname{tr}\left(\boldsymbol{\Sigma}_t^{-1}\mathbf{R}_t\right)\right) \\
&= -\frac{1}{2}\left(\log\left|\mathbf{L}_t\mathbf{D}_t\mathbf{L}_t^\top\right| + \operatorname{tr}\left((\mathbf{L}_t\mathbf{D}_t\mathbf{L}_t^\top)^{-1}\mathbf{R}_t\right)\right) \\
&= -\frac{1}{2}\left(\log|\mathbf{D}_t| + \operatorname{tr}\left(\mathbf{L}_t\mathbf{D}_t^{-1}\mathbf{L}_t^\top\mathbf{R}_t\right)\right) \\
&= -\frac{1}{2}\left(\sum_{i=1}^{p}\log\left(d_{i,t}\right) + \operatorname{tr}\left(\mathbf{D}_t^{-1}\mathbf{L}_t^\top\mathbf{R}_t\mathbf{L}_t\right)\right) \\
&= -\frac{1}{2}\sum_{i=1}^{p}\left(\log\left(d_{i,t}\right) + \frac{g_{i,t}}{d_{i,t}}\right). \qquad (4.23)
\end{aligned}$$

Standard errors can be obtained by the well-known sandwich formula, e.g. provided in Bollerslev and Wooldridge (1992). However, initial investigation showed that the QL function is multi-modal, such that standard local gradient-based optimization algorithms fail if the realized covariance measure comprises more than a few assets. As an alternative, gradient-free global optimization methods like pattern search (direct search), genetic algorithms, and simulated annealing can be employed (see e.g. Kelley, 1999 for details). Moreover and even more importantly, the parameter vector $\boldsymbol{\psi}$ quickly becomes large if the number of assets p increases, in particular since the parameter matrix $\boldsymbol{\Xi}$ of the $\mathbf{Q}_t$ process in equation (4.5) comprises $p(p+1)/2$ model parameters, e.g. 465 intercepts for, say, $p = 30$ assets (the so-called *curse of dimensionality*). This causes additional problems in numerically optimizing the likelihood and makes high-dimensional applications (say, for $p > 10$ assets) practically impossible.

The following section therefore proposes a three-step estimation approach, which solves the curse of dimensionality via multi-step estimation and *covariance targeting* (see e.g. Bauwens, Laurent, and Rombouts, 2006) where the intercept matrix $\boldsymbol{\Xi}$ is replaced by an ex-ante estimate of the unconditional mean of the covariance process, similar to the DCC framework of Engle (2002). The three-step approach can therefore be applied in empirically realistic settings with $p > 10$ assets and also provides a convenient solution to the numerical optimization problems arising for one-step estimation (see above).

Three-Step Estimation

In high-dimensional scenarios, the curse of dimensionality precludes one-step estimation of the $p(p+1)/2+2(p+1)+1$ parameters of the DPC-CAW model via the corresponding Wishart likelihood. Aielli and Caporin (2015) propose a three-step estimation technique called the *DPC estimator*, which is easily adapted to the CAW framework. The procedure works as follows:

1. Estimate $\boldsymbol{\Xi} = \underline{\mathbf{L}}\,\underline{\mathbf{D}}\,\underline{\mathbf{L}}^\top$ via $\hat{\boldsymbol{\Xi}} = T^{-1}\sum_{t=1}^{T}\mathbf{R}_t$ (*covariance targeting*);

2. Conditional on step 1. estimate $(a,b)^\top$ by fitting a scalar CAW model (see Golosnoy, Gribisch, and Liesenfeld, 2012) to the sequence of RCs, essentially (wrongly) assuming $\mathbf{R}_t|\mathcal{F}_{t-1} \sim \mathcal{W}(\mathbf{Q}_t/n, n)$, where $\mathbf{Q}_t$ is given by equation (4.5) and $\boldsymbol{\Xi} \overset{!}{=} \hat{\boldsymbol{\Xi}}$. Recover $\{\hat{\mathbf{Q}}_t\}_{t=1}^T$ as the $\mathbf{Q}_t$-sequence computed at the CAW parameter estimates in order to calculate $\{\hat{g}_{i,t}\}_{t=1}^T$ for $i = 1,\ldots,p$, where $\hat{g}_{i,t} = \mathbf{e}_i^\top\hat{\mathbf{L}}_t^\top\mathbf{R}_t\hat{\mathbf{L}}_t\mathbf{e}_i$ with $\hat{\mathbf{L}}_t$ being the matrix of eigenvectors of $\hat{\mathbf{Q}}_t$;

3. Conditional on 1. and 2. estimate $\{\alpha_i,\beta_i\}_{i=1}^n$ via univariate QML based on equations (4.7) and (4.16) separately $\forall i$ with $g_{i,t}$ replaced by $\hat{g}_{i,t}$ from estimation step 2. The i'th log-likelihood is given by

$$\mathcal{L}_i(\alpha_i,\beta_i) = \sum_{t=1}^{T}\Big[(n/2-1)\log(\hat{g}_{i,t}) - \log(\Gamma(n/2)) - (n/2)\log(2d_{i,t}/n) - 0.5n\hat{g}_{i,t}/d_{i,t}\Big]. \tag{4.24}$$

   Analogous to the Wishart, $\mathcal{L}_i(\alpha_i,\beta_i)$ features a QML interpretation given the previously estimated $\{\hat{g}_{i,t}\}_{t=1}^T$.

Steps 1 and 2 estimate the parameters of the eigenvector driving $\mathbf{Q}_t$-process of the DPC-CAW model by fitting a scalar CAW model to the sequence of RCs and employing covariance targeting in order to alleviate the curse of dimensionality. Under the assumption that the true data-generating process for the RCs $\mathbf{R}_t$ is a DPC-CAW, it is clear from equations (4.2) and (4.4) that the conditional mean of $\mathbf{R}_t$ is $\boldsymbol{\Sigma}_t = \mathbf{L}_t\mathbf{D}_t\mathbf{L}_t^\top$ and not $\mathbf{Q}_t = \mathbf{L}_t\mathbf{G}_t\mathbf{L}_t^\top$ (see equation 4.6). Hence the first two estimation steps, which essentially assume that the conditional mean of $\mathbf{R}_t$ is $\mathbf{Q}_t$ rather than $\boldsymbol{\Sigma}_t$, introduce bias and inconsistency in estimation since ML estimators of misspecified mean models are inconsistent (see e.g. Bollerslev and Wooldridge (1992)). However, since standard CAW models (which in the given context can be interpreted as misspecified conditional volatility models for data generated by the DPC-CAW) typically provide good approximations to the stochastic behavior of the RCs $\mathbf{R}_t$ (compare e.g. the results of Golosnoy, Gribisch, and Liesenfeld

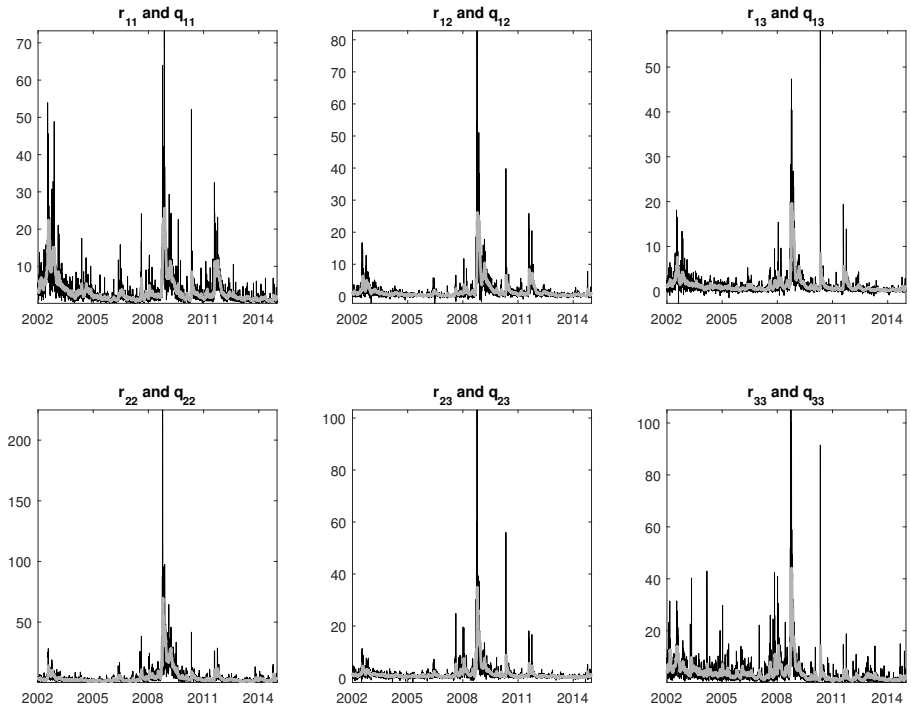

Figure 4.1: Black line: Realized variances and covariances $r_{ij} = (\mathbf{R})_{ij}$ of A ($i = 1$), AA ($i = 2$) and AAPL ($i = 3$); Gray line: estimates of the individual $\mathbf{Q}_t$ elements obtained via one-step QML estimation of the DPC-CAW model as specified in Section 4.3 to the according set of three-dimensional RCs.

(2012)), we can argue that the eigenvectors of $\mathbf{Q}_t$-sequences obtained as estimated conditional means from CAW recursions are expected to be rather close to the eigenvectors of $\mathbf{\Sigma}_t$, namely $\mathbf{L}_t$. This result is illustrated in Figure 4.1, which shows consistent estimates of the individual $\mathbf{Q}_t$ elements obtained via one-step QML estimation of the DPC-CAW model to a 3-dimensional realized covariance subset of the data discussed in Section 4. The $\mathbf{Q}_t$-dynamics closely follow the pattern of the realized (co)variance data, which is effectively approximated by the CAW in estimation steps 1 and 2.

Recall that steps 1 and 2 result in biased and inconsistent estimates of the parameters a, b, and $\mathbf{\Xi}$ since the scalar CAW likelihood in step 2 is not correctly specified (the matrix $\mathbf{Q}_t$ under the DPC-CAW is not the conditional mean of $\mathbf{R}_t$, as discussed above). Subsequently, conditional on Steps 1 and 2, the parameters of

the eigenvalue driving processes are estimated. This last estimation step does not add to the inconsistency due to the QML interpretation of the according likelihoods given in equation (4.24). Notice that the intercept parameters γ_i were fixed in step 1, such that step 3 essentially corresponds to univariate GARCH estimation with variance targeting.

The quasi-likelihood functions in steps 2 and 3 are smooth, hence standard gradient-based optimization procedures can be applied. However, estimation of standard errors becomes complicated due to the aforementioned misspecification. In fact, we may obtain asymptotic standard errors under the assumption of consistency by applying the GMM-framework of Engle (2009) for the three-step approach (for details, see also Aielli and Caporin, 2015). However, these standard errors are not valid if the model is inconsistently estimated since the moment conditions (the likelihood scores) are not valid in this case. In fact, our simulation experiments in Section 4.4 show signs of inconsistency. However, note that standard errors are of minor importance for forecasting applications, which are in the focus of the present paper.

The three-step approach is simple and intuitive but comes with the disadvantage of introducing bias and inconsistency in parameter estimation. This inconsistency is unavoidable in high-dimensional applications of the DPC approach and has already been encountered by Aielli and Caporin (2015) in the corresponding GARCH framework. Section 4.4 analyzes the properties of obtained estimates in an extensive simulation experiment. The results suggest that bias is present but acceptably small or of reduced impact, especially given the huge dimension of the estimation problem. The forecasting application of Section 4.5 furthermore shows that these issues do not negatively affect the out-of-sample performance, which is typically in the focus of empirical applications of multivariate volatility models.

The three-step estimator enables quick estimation in high-dimensional settings. The according CAW likelihoods are well-behaved as already found by Golosnoy, Gribisch, and Liesenfeld (2012), resulting in fast convergence of Quasi-Newton based numerical optimizers. Three-step estimation of a 100-dimensional DPC-CAW model with $T = 2500$ takes at most 100 seconds using an Intel Core i7 2.60 GHz processor under Matlab 2018b. However, the estimation of the DPC-CAW model in large/vast-dimensional settings (>>100 assets) quickly becomes challenging mainly due to the inversion of the conditional mean $\mathbf{\Sigma}_t$ in each single likelihood evaluation of estimation step 2. This shortcoming is well-known in the literature on multivariate volatility modeling (see e.g. Hafner and Reznikova, 2012). One possibility to overcome this problem is to apply composite likelihood techniques, as e.g. illustrated by Pakel et al. (2021) and Engle, Ledoit, and Wolf (2019) under the DCC GARCH framework. Here the composite log-likelihood is computed by summing up the log-likelihoods of pairs of assets. This greatly

reduces the dimension of the estimation problem and enables inference for vast-dimensional covariance matrices. We, however, note that - in contrast to GARCH-type models - the DPC-CAW model is based on observed RCs, which have been computed from synchronized intraday asset return observations. The computation of vast-dimensional RCs is challenging, involves sparsity assumptions like factor structures, trimming, or eigenvalue cleaning procedures (see e.g. Tao et al., 2011 and Lunde, Shephard, and Sheppard, 2016), which impose ex-ante structures on the RCs. As an alternative way of dealing with vast-dimensional applications of the DPC-CAW model, this rather suggests to impose a factor structure on the intraday asset returns and apply the DPC-CAW framework to the factor covariance matrix, which is typically of low dimension (e.g. 5 to ten-dimensional). See e.g. Asai and McAleer (2015) and Shen, Yao, and Li (2020) for similar applications of CAW models to vast-dimensional frameworks. We leave the application of factor structures in the DPC-CAW framework for future research.

4.4 Simulation Experiment

We conduct an extensive simulation experiment in order to assess the finite sample properties of the DPC estimator. Since we focus on high-dimensional applications, the cross-sectional size is set to $p = 100$.

The following parameter setup is used: The intercept matrix $\mathbf{\Xi} = \mathbf{L}\mathbf{D}\mathbf{L}^\top$ of the $\mathbf{Q}_t$ process is set equal to the average RC of the data employed in the empirical application of Section 4.5.1. We consider nine distinct eigenvector recursion parameter set-ups, where the ARCH parameter a is set equal to 0.025, 0.035, 0.05 or 0.1 and the GARCH parameter is chosen such that the persistence $(a + b)$ equals 0.9, 0.95, 0.99, or 0.997. Note that the setting $(a, b) = (0.035, 0.997)$ corresponds to our empirical results from Table 4.6.

In order to achieve some variability in the eigenvalue recursion parameters, they are drawn from uniform distributions according to

$$\alpha_i \sim U(0.22, 0.3), \qquad \beta_i|\alpha_i \sim U(0.94 - \alpha_i, 0.99 - \alpha_i). \tag{4.25}$$

Consequently the persistence parameters $(\alpha_i + \beta_i) \in [0.94, 0.99]$. The degrees of freedom parameter n is set to $n = 100$. This parameter setup is inspired by parameter estimates obtained in the empirical application of Section 4.5.1. The whole experiment covers 500 independent simulations for each of the four time-series lengths $T = 1000$, $T = 2500$, $T = 5000$, $T = 10000$, and each of the $4 \times 4 = 16$ parameter constellations.

All estimation results in this section and the upcoming empirical application in Section 4.5 are obtained under the three-step DPC estimator. The CAW-

125.16	16.39	10.98	9.26	7.99	6.47	6.42	5.57	5.31	5.31
5.22	4.84	4.80	3.97	3.83	3.79	3.69	3.43	3.35	3.22
3.10	3.00	2.95	2.83	2.74	2.71	2.66	2.61	2.55	2.48
2.37	2.28	2.27	2.19	2.17	2.13	2.11	2.07	2.01	1.94
1.93	1.87	1.85	1.82	1.79	1.77	1.72	1.69	1.64	1.62
1.59	1.57	1.54	1.49	1.47	1.47	1.42	1.41	1.37	1.34
1.32	1.29	1.29	1.28	1.25	1.25	1.24	1.23	1.23	1.20
1.19	1.18	1.12	1.12	1.06	1.05	1.05	1.03	1.02	1.00
0.99	0.98	0.97	0.93	0.91	0.88	0.85	0.83	0.83	0.82
0.81	0.81	0.79	0.72	0.68	0.65	0.54	0.51	0.47	0.45

Table 4.1: Sorted eigenvalues obtained from $\hat{\mathbf{\Xi}} = T^{-1} \sum_{t=1}^{T} \mathbf{R}_t$ for the dataset described in Section 4.5.

and Gamma likelihoods are maximized via Quasi-Newton methods with BFGS-updating of the Hessian under Matlab 2018b. The eigenvector and eigenvalue recursions are initialized by $\mathbf{Q}_0 = \frac{1}{T} \sum_{t=1}^{T} \mathbf{R}_t$ and $d_{i,0} = \frac{1}{T} \sum_{t=1}^{T} g_{i,t}$ for $i = 1, \ldots, p$. As starting values for the numerical optimization we choose $(0.05, 0.9)$ for (a, b) and (α_i, β_i). The numerical optimization, however, appeared robust to the choice of the starting values.[4]

Estimation Step 1 Note that the symmetric 100×100 parameter matrix $\mathbf{\Xi}$ comprises 5050 distinct model parameters. For this reason, we focus on the analysis of the 100 eigenvalues $\underline{d}_i$, i.e. the diagonal elements of $\underline{\mathbf{D}}$, which are of particular importance since they determine the level of the eigenvalue recursions in estimation step 3. Moreover, the $\underline{d}_i$ estimates can be interpreted as the unconditional variances of the asset returns' principal components, with decreasing fraction of explained asset return variation for decreasing i. Figure 4.2 shows the average relative (percentage) estimation errors for the $\underline{d}_i$ estimates, $\frac{1}{500} \sum_{j=1}^{500} 100 \cdot (\hat{\underline{d}}_{ij} - \underline{d}_{ij})/\underline{d}_{ij}$, over the 500 simulated datasets for the 16 parameter set-ups and four sample sizes outlined above. Note that the setting $(a = 0.035, a + b = 0.997)$ corresponds to our empirical findings of Section 4.5. The true parameter values for the $\underline{d}_i$'s are reported in Table 4.1 and show a sharply decreasing pattern with 50% (90%) of the total asset return variation explained by the first 7 (60) $\underline{d}_i$'s. For the ten highest, and therefore most important, eigenvalues, the biases are ranging between -35% and 2% with rather low values between -2% and +2% for settings with low a and low eigenvalue persistence. For the setting which comes closest to our em-

4. The Matlab estimation files for the DPC-CAW model are available under `https://github.com/mstollenwerk/dpc_caw`.

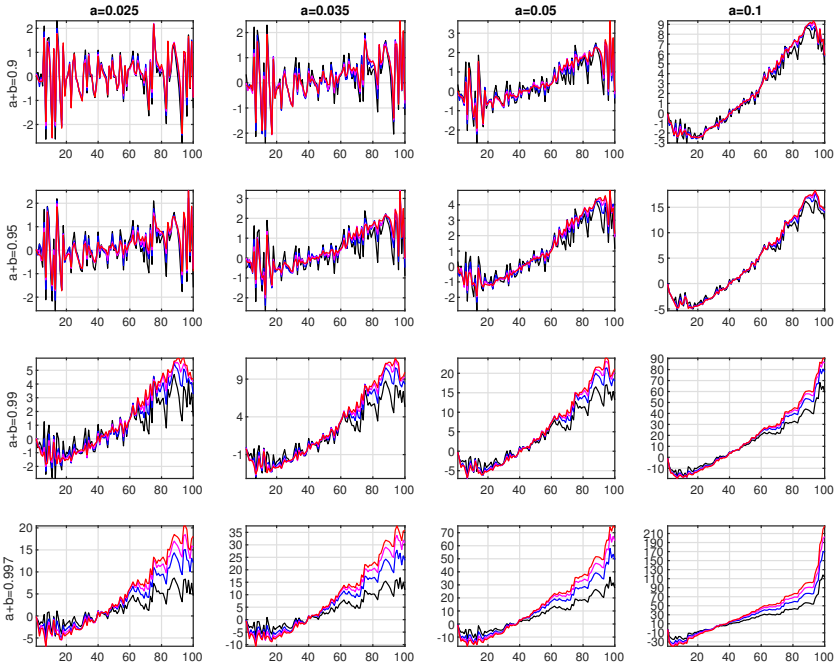

Figure 4.2: Average relative estimation errors for $\underline{d}_i$ $(i = 1, \ldots, 100)$ computed as arithmetic mean $\frac{1}{500}\sum_{j=1}^{500} 100 \cdot (\hat{\underline{d}}_{ij} - \underline{d}_{ij})/\underline{d}_{ij}$ over the estimates obtained for the 500 simulations. Black: $T = 1000$; blue: $T = 2500$; magenta: $T = 5000$; red: $T = 10000$. The results are obtained from the DPC estimator for the simulation experiment of Section 4.4. The DGP parameter values are reported at the top of the panel for a and on the left side of the panel for $(a + b)$. Each line comprises 100 data points, one for each $\underline{d}_i$ in descending order, with $\underline{d}_1$ being displayed on the left.

pirical results, $(a = 0.035, a + b = 0.997)$, the biases of the first ten eigenvalues range from -0.5 to -10%. For persistencies of 0.99 or higher accompanied with comparatively large values for the a parameter, we obtain biases of up to 220%. This trend is particularly obvious for the high ARCH, high persistence parameter set $(a = 0.1, (a + b) = 0.997)$, which, however, does not appear to be relevant in practice (see the empirical results of Section 4.5). Also, note that these values correspond to $\underline{d}_i$'s of very low level (compare Figure 4.2) whose contribution to the overall asset return volatility is very low. A significant impact of these relative biases on the forecasting performance of the DPC-CAW model is therefore not

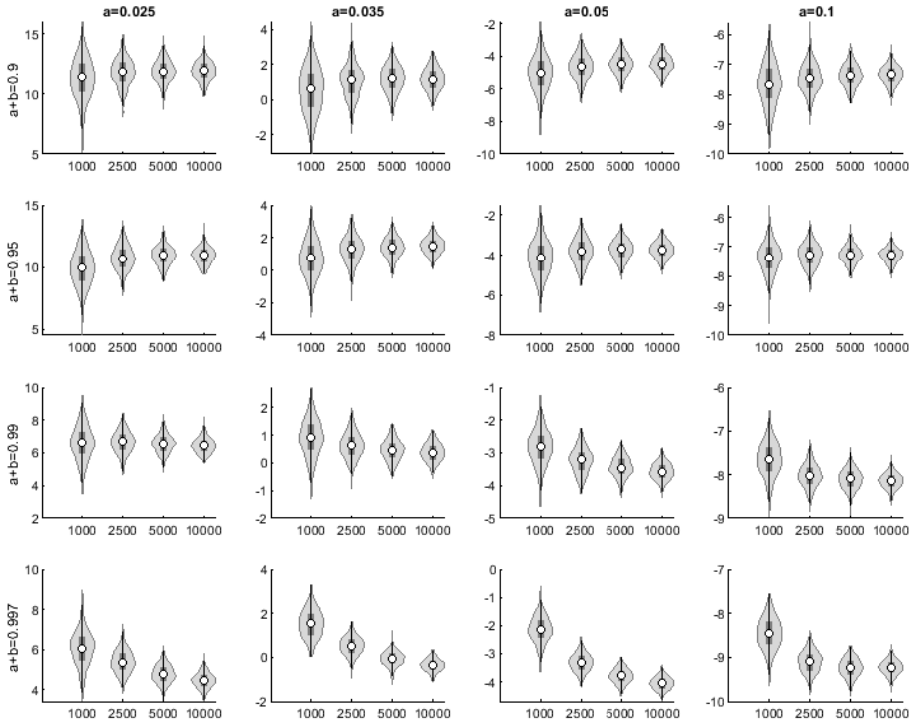

Figure 4.3: Violin plots of relative estimation errors $100 \cdot (\hat{a} - a)/a$ obtained from the DPC estimator for the simulation experiment of Section 4.4. The DGP parameter values are reported at the top of the panel for a and on the left side of the panel for $(a + b)$. The first violin plot in each subplot comprises results for $T = 1000$, the second for $T = 2500$, the third for $T = 5000$, and the fourth for $T = 10000$. The white dot within the box indicates the median.

to be expected (see also our further discussion below on the additional results in Figures 4.6 and 4.7).

Estimation Step 2 Figure 4.3 reports Violin Plots of relative estimation errors for the ARCH parameter a. The biases range from -9.2% to 12.2% and are positive for $a = 0.025$, close to zero for $a = 0.035$, and negative for high ARCH environments with $a > 0.035$. While the biases appear mostly stable for increasing sample sizes, we observe a decrease in the biases for the $(a = 0.035, a + b = 0.997)$ setting but also an increasing relative bias pattern for high persistence / high ARCH set-

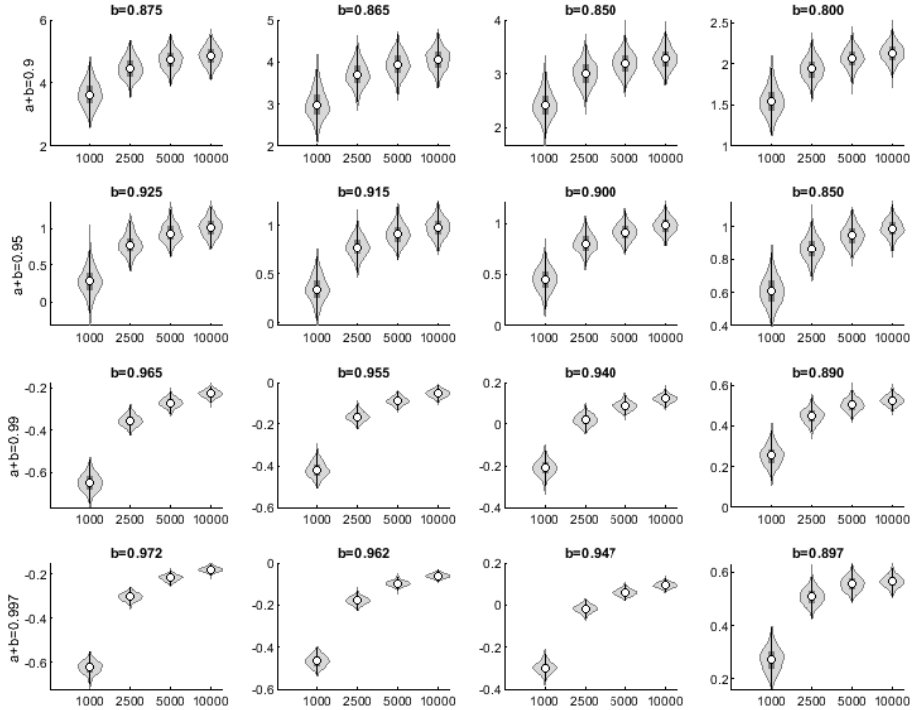

Figure 4.4: Violin plots of relative estimation errors $100 \cdot (\hat{b} - b)/b$ obtained from the DPC estimator for the simulation experiment of Section 4.4. The DGP parameter values are reported at the top of the panel for b and on the left side of the panel for $(a + b)$. The first violin plot in each subplot comprises results for $T = 1000$, the second for $T = 2500$, the third for $T = 5000$, and the fourth for $T = 10000$. The white dot within the box indicates the median.

tings. The smallest biases are obtained for the $(a = 0.035, a + b = 0.997)$ setting, which corresponds to our empirical results of Section 4.5. The largest biases are obtained for the low persistence / low ARCH setting $(a = 0.025, a + b = 0.9)$ and amount to 12% on average.

Figure 4.4 depicts the distribution of relative estimation errors for b. The biases appear small, ranging from -0.61% to 5.1%, and decrease with increasing persistence $a+b$. For lower values of the GARCH parameter b, the biases tend to increase for high-persistence scenarios. For increasing sample sizes, the biases are partly diverging but come close to zero for our empirical estimates $(b = 0.962, a+b = 0.997)$.

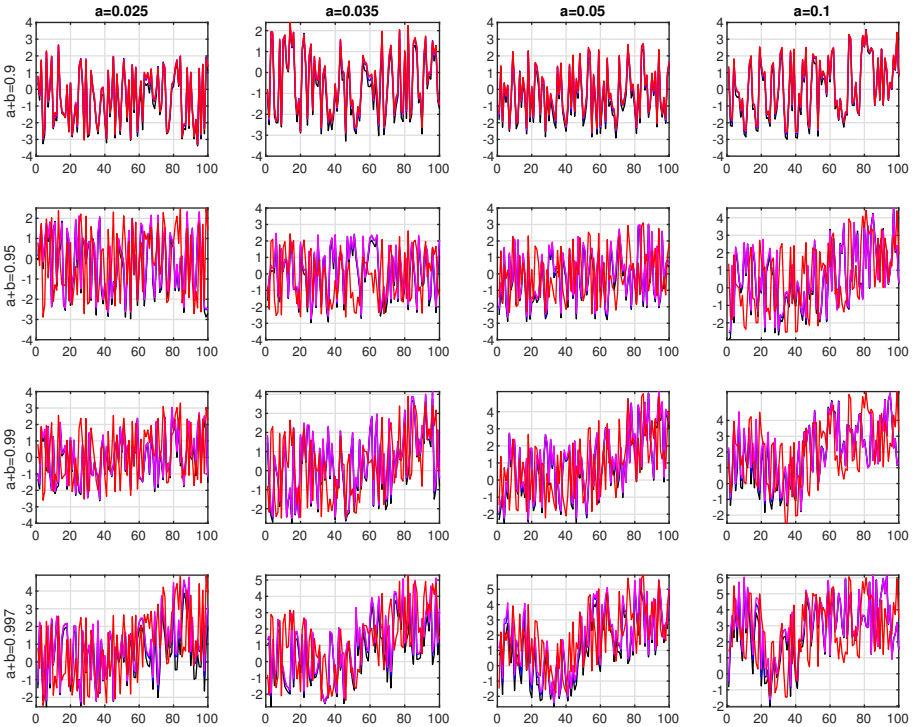

Figure 4.5: Average relative estimation errors for the eigenvalue persistence $\alpha_i + \beta_i$ computed as arithmetic mean $\frac{1}{500} \sum_{j=1}^{500} 100 \cdot (\hat{\alpha}_{ij} + \hat{\beta}_{ij} - (\alpha_{ij} + \beta_{ij}))/(\alpha_{ij} + \beta_{ij})$ over the estimates obtained for the 500 simulations. Black: $T = 1000$; blue: $T = 2500$; magenta: $T = 5000$; red: $T = 10000$. The results are obtained from the DPC estimator for the simulation experiment of Section 4.4. The DGP parameter values are reported at the top of the panel for a and on the left side of the panel for $(a + b)$. Each line comprises 100 data points, one for each $\alpha_i + \beta_i$ in descending order with $\alpha_1 + \beta_1$ corresponding to the highest eigenvalue being displayed on the left.

Estimation Step 3 Figure 4.5 shows relative biases for the eigenvalue persistencies $\alpha_i + \beta_i$ for $i = 1, \ldots, 100$, which appear overall low with values ranging from -3% to 6%. For the individual α_i and β_i we find negative and compensating positive biases between -60% and 10% (α) and -10% and 30% (β) (not reported here).

In order to investigate the effect of the biases on the estimated asset return (co)variances, we compute for each of the simulated datasets the sequence of es-

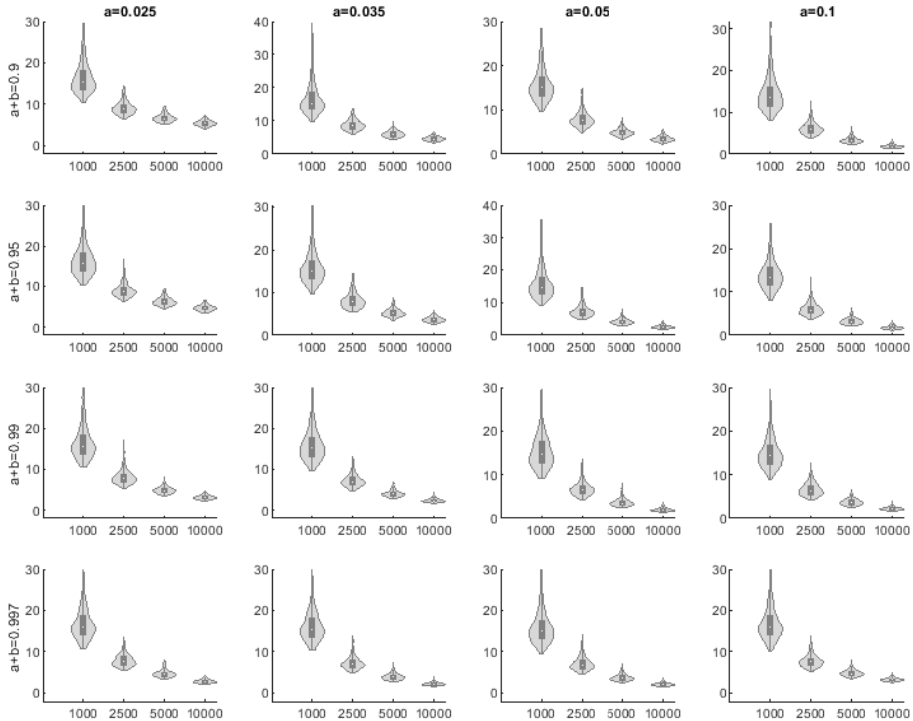

Figure 4.6: Violin plots of average Euclidean norms $(1/T)\sum_t \text{vech}(\hat{\boldsymbol{\Sigma}}_t - \boldsymbol{\Sigma}_t)^\top \text{vech}(\hat{\boldsymbol{\Sigma}}_t - \boldsymbol{\Sigma}_t)$ over the 500 simulations. The DGP parameter values are reported at the top of the panel for a and on the left side of the panel for $(a+b)$. The first violin plot in each subplot comprises results for $T = 1000$, the second for $T = 2500$, the third for $T = 5000$, and the fourth for $T = 10000$. The white dot within the box indicates the median. The results are obtained from the DPC estimator for the simulation experiment of Section 4.4.

timated covariance forecasts $\hat{\boldsymbol{\Sigma}}_t(\hat{\boldsymbol{\theta}})$ for $t = 1, \ldots, T$ as a function of the parameter estimates $\hat{\boldsymbol{\theta}}$ discussed above, and compare these estimates to the true simulated forecasts $\boldsymbol{\Sigma}_t(\boldsymbol{\theta})$. Figure 4.6 reports Violin Plots of the average Euclidean norms $(1/T)\sum_t \text{vech}(\hat{\boldsymbol{\Sigma}}_t(\hat{\boldsymbol{\theta}}) - \boldsymbol{\Sigma}_t(\boldsymbol{\theta}))^\top \text{vech}(\hat{\boldsymbol{\Sigma}}_t(\hat{\boldsymbol{\theta}}) - \boldsymbol{\Sigma}_t(\boldsymbol{\theta}))$ over the 500 simulations, and Figure 4.7 shows Violin Plots of the corresponding relative (percentage) biases computed as $(1/T)\sum_t (100/5050)\sum_{i>j}(\hat{\boldsymbol{\Sigma}}_{ij}(\hat{\boldsymbol{\theta}}) - \boldsymbol{\Sigma}_{ij}(\boldsymbol{\theta}))/\boldsymbol{\Sigma}_{ij}(\boldsymbol{\theta})$. The results show a pattern of decreasing Euclidean norms for increasing sample sizes, coming close to zero for $T = 10000$ and all considered parameter settings. The

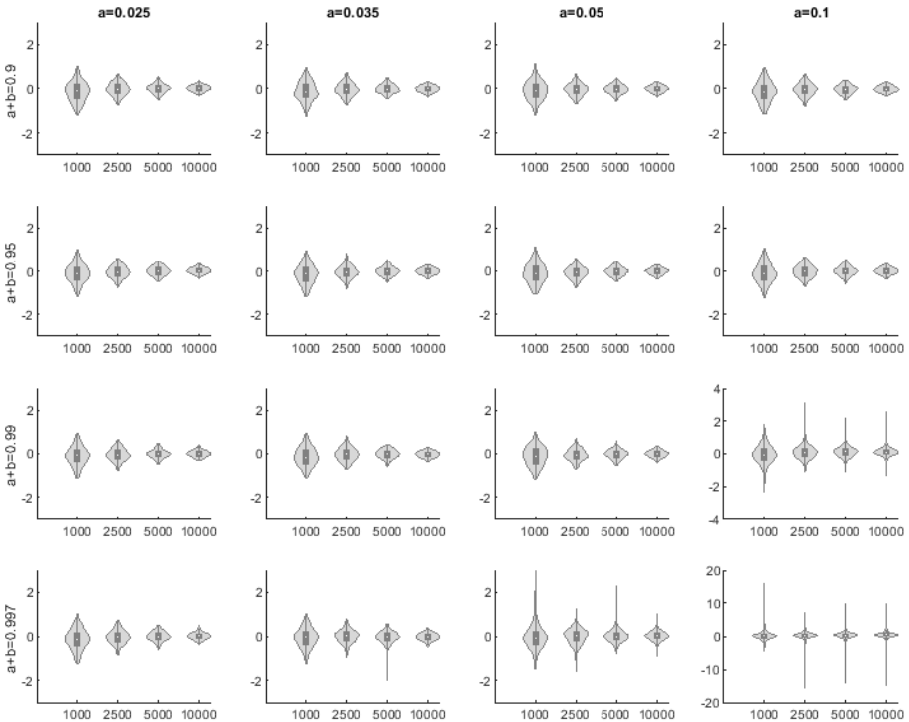

Figure 4.7: Violin plots of average relative biases $(1/T)\sum_t(1/5050)\sum_{i>j}(\hat{\mathbf{\Sigma}}_{ij}-\mathbf{\Sigma}_{ij})/\mathbf{\Sigma}_{ij}$ over the 500 simulations. The DGP parameter values are reported at the top of the panel for a and on the left side of the panel for $(a+b)$. The first violin plot in each subplot comprises results for $T = 1000$, the second for $T = 2500$, the third for $T = 5000$, and the fourth for $T = 10000$. The white dot within the box indicates the median. The results are obtained from the DPC estimator for the simulation experiment of Section 4.4.

distribution of relative biases shows low dispersion around zero, and the variation appears to be decreasing with increasing sample size.

Summarizing the results, the simulation experiment indicates relatively low biases for the eigenvalue persistencies and the ARCH- and GARCH parameters of the eigenvector recursions while the biases of the $\underline{d}_i$-estimates in estimation step 1 amount to up to 220% for eigenvalues of very low level. However, these biases do not significantly affect the covariance forecasts as indicated by the Euclidean norms and biases for $\mathbf{\Sigma}_t$ reported in Figures 4.6 and 4.7. Hence the biases are not expected to significantly affect forecasting performance of the DPC-CAW model.

4.5 Empirical Application

4.5.1 Data

	Mean	Min.	Max.	Range	Std. dev.	Skewness	Kurtosis
Realized variances (100 time-series)							
Min.	1.01	0.02	48.39	48.35	1.89	3.75	24.30
Median	3.29	0.10	117.00	116.88	5.04	8.90	146.03
Max.	12.51	0.35	7727.54	7727.50	151.31	43.58	2126.80
Realized covariances (4950 time-series)							
Min.	0.20	−126.77	14.07	14.97	0.87	−1.09	33.86
Median	1.05	−3.32	63.42	68.12	2.61	10.08	169.49
Max.	3.90	−0.02	1262.30	1282.60	25.51	38.93	1851.31

Table 4.2: Descriptive statistics for the 5050 realized variance and covariance time-series of the 100-dimensional dataset described in Section 4.5.

We apply the scalar-DPC-CAW model introduced in Section 4.3 in order to capture the dynamics of 100-dimensional RCs. The data have been computed from one-minute intraday asset returns by the microstructure-noise and jump robust multivariate realized kernel method of Barndorff-Nielsen et al. (2011). The corresponding ticker symbols are shown in Table 4.3 (p. 150). Note that the choice of the particular type of realized measure is not an important issue since the model can be fitted to any series of positive-definite RCs. The sample period starts on 01 January 2002 and ends on 03 December 2014, covering 3271 trading days.

Figure 4.8 depicts exemplary time-series plots of variance and covariance series and according sample autocorrelation functions for four stocks included in the dataset. Descriptive statistics are provided in Table 4.2. The (co)variance processes are highly persistent, skewed to the right, leptokurtic, and tend to move parallel to each other.

4.5.2 In-Sample Estimation Results

We start with analyzing the in-sample fit of the DPC-CAW model for various model-order settings using the BIC information criterion. According to Geweke and Meese (1981) it can be shown that for linear ARMA-type models, the BIC is consistent in the sense that asymptotically the correct model order is chosen. This, however, does not hold for nonlinear models such as the CAW of Golosnoy,

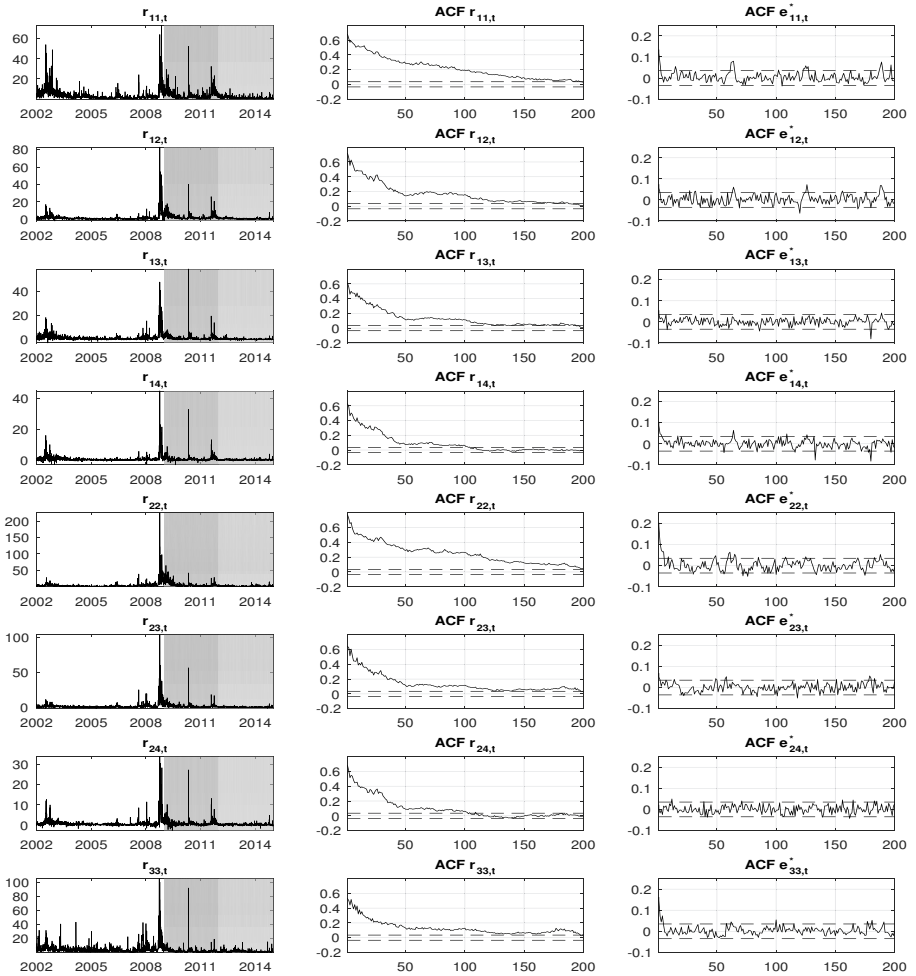

Figure 4.8: Realized (co)variance plots and sample autocorrelation functions (ACFs). Left panel: Sample of realized variances and covariances $r_{ij} = (\mathbf{R})_{ij}$ of A ($i = 1$), AA ($i = 2$), AAPL ($i = 3$) and ABT ($i = 4$). Gray shaded areas indicate the periods covered by the forecasting experiment of Section 4.5.3. Middle panel: Sample ACFs of realized (co)variances together with 95% confidence bounds under the null of zero serial correlation. Right panel: Sample ACFs and according 95% confidence bounds of standardized Pearson residuals obtained from the BIC selected DPC-CAW(3,4)-(1,1) model estimated by the DPC estimator for the 100-dim. dataset illustrated in Section 4.5.

Symbol	Company	Symbol	Company
a	Agilent Technologies Inc.	gild	Gilead Sciences Inc.
aa	Alcoa Inc.	glw	Corning Incorporated
aapl	Apple Inc.	gps	Gap, Inc.
abt	Abbott Laboratories	gs	Goldman Sachs Group, Inc.
abx	Barrick Gold Corporation	hal	Halliburton Company
adbe	Adobe Systems Incorporated	hd	Home Depot, Inc.
adi	Analog Devices Inc.	hig	Hartford Financial Services Group, Inc.
adp	Automatic Data Processing	hon	Honeywell International Inc.
aig	American International Group Inc.	hpq	Hewlett-Packard Company
all	Allstate Corporation	ibm	International Business Machines Corporation
altr	Altera Corporation	intc	Intel Corporation
amat	Applied Materials Inc.	intu	Intuit Inc.
amd	Advanced Micro Devices Inc.	ip	Internation Paper Company
amgn	Amgen Inc.	jcp	J.C. Penney Company, Inc.
amzn	Amazon.com, Inc.	jnj	Johnson & Johnson
apc	Anadarko Petroleum Corporation	jnpr	Juniper Networks, Inc.
axp	American Express Company	jpm	J P Morgan Chase & Co
ba	Boeing Company	klac	KLA-Tencor Corporation
bac	Bank of America Corporation	ko	Cocoa-Cola Company
bax	Baxter International Inc.	kr	Kroger Company
bbby	Bed Bath & Beyond Inc.	kss	Kohl's Corporation
bby	Best Buy Co., Inc.	lb	La Barge Inc.
bhi	Baker Hughes Incorporated	lltc	Linear Technology Corporation
bmy	Bristol-Myers Squibb Company	lly	Eli Lilly and Company
brcm	Broadcom Corporation	lmt	Lockheed Martin Corporation
c	Citigroup Inc.	low	Lowe's Companies, Inc.
cag	ConAgra, Inc.	luv	Southwest Airlines Company
cah	Cardinal Health Inc.	mas	Masco Corporation
cat	Caterpillar, Inc.	mcd	McDonald's Corporation
cbs	CBS Corporation new	mdt	Medtronic Inc.
cien	Ciena Corporation	met	MetLife, Inc.
cl	Colgate-Palmolive Company	mmc	Marsh & McLennan Companies, Inc.
cop	ConocoPhillips	mmm	3M Company
cost	Costco Wholesale Corporation	mo	Altria Group
csco	Cisco Systems, Inc.	mrk	Merck & Company, Inc.
ctxs	Citrix Systems, Inc.	ms	Morgan Stanley Dean Witter & Co
cvs	CVS Caremark Corp.	msft	Microsoft Corporation
cvx	Chevron Corporation	msi	Motorola Solutions, Inc.
dd	Dupont De Nemours Inc.	mu	Micron Technology, Inc.
de	Deere & Company	nem	Newmont Mining Corporation
dis	Walt Disney Company	nke	Nike, Inc.
dow	Dow Chemical Company	ntap	NetApp, Inc.
duk	Duke Energy Corporation new	nvda	NVIDIA Corporation
ea	Electronic Arts Inc.	orcl	Oracle Corporation
ebay	Ebay Inc.	oxy	Occidental Petroleum Corporation
emc	EMC Corporation MA	payx	Paychex, Inc.
emr	Emerson Electric Company	pep	Pepsico, Inc.
f	Ford Motor Company DEL	pfe	Pfizer, Inc.
fitb	Fifth Third Bancorp	pg	Procter & Gamble Company
ge	General Electric Company	qcom	QUALCOMM Incorporated

Table 4.3: Dataset of 100 stocks selected by liquidity from the S&P 500.

Gribisch, and Liesenfeld (2012) or, equivalently, the DPC-CAW - even if consistently estimated. Golosnoy, Gribisch, and Liesenfeld (2012) remark that there do

Order of eigenvalue process	Order of eigenvector process: (1,0)	(1,1)	(2,1)	(1,2)	(2,2)	(3,2)	(2,3)	(3,3)
(1,0)	-2.5515	-2.7671	-2.0818	-2.4645	-2.0505	-2.0298	-2.0460	-2.0998
(1,1)	-2.6289	-2.8342	-2.1779	-2.5363	-2.1531	-2.1402	-2.1502	-2.1892
(2,1)	-2.6287	-2.8340	-2.1778	-2.5361	-2.1531	-2.1404	-2.1503	-2.1890
(1,2)	-2.6304	-2.8361	-2.1793	-2.5377	-2.1546	-2.1418	-2.1517	-2.1905
(2,2)	-2.6303	-2.8360	-2.1793	-2.5376	-2.1546	-2.1420	-2.1518	-2.1904
(3,2)	-2.6301	-2.8358	-2.1791	-2.5374	-2.1544	-2.1418	-2.1516	-2.1902
(2,3)	-2.6311	-2.8371	-2.1801	-2.5384	-2.1553	-2.1428	-2.1525	-2.1911
(3,3)	-2.6309	-2.8370	-2.1799	-2.5382	-2.1552	-2.1426	-2.1524	-2.1909
(4,3)	-2.6308	-2.8368	-2.1798	-2.5380	-2.1550	-2.1425	-2.1522	-2.1908
(3,4)	-2.6314	**-2.8374**	-2.1801	-2.5384	-2.1554	-2.1428	-2.1526	-2.1911
(4,4)	-2.6313	-2.8372	-2.1800	-2.5383	-2.1553	-2.1427	-2.1524	-2.1910
(5,4)	-2.6312	-2.8371	-2.1798	-2.5381	-2.1551	-2.1426	-2.1523	-2.1908
(4,5)	-2.6313	-2.8373	-2.1800	-2.5382	-2.1553	-2.1428	-2.1525	-2.1910
(5,5)	-2.6312	-2.8372	-2.1799	-2.5381	-2.1552	-2.1426	-2.1524	-2.1908
HAR	-2.6280	-2.8341	-2.1773	-2.5357	-2.1525	-2.1397	-2.1497	-2.1883

Table 4.4: Bayes information criteria (BIC) for estimated DPC-CAW models with various lag-order constellations. BIC values: $\times 10e7$. Models are estimated using the three-step estimation approach.

not seem to exist published results on the consistency of the BIC for nonlinear time-series models, which would justify its use under the CAW framework. Nevertheless, the BIC is often applied as an "indicator", although not asymptotically valid in a strict sense. Here we follow Golosnoy, Gribisch, and Liesenfeld (2012) and align the BIC-based model selection by model diagnostics based on Ljung-Box residual autocorrelation tests for the fitted models.

All models are estimated by the three-step approach. We consider both order selection for the eigenvector- and for the eigenvalue processes given in equations (4.5) and (4.7), jointly. For the eigenvalues, we restrict the chosen order to be identical across the 100 assets. Table 4.4 shows the results of the BIC information criteria. We find a clear indication for the (1,1) specification of the eigenvector recursion, which corresponds to the typically chosen DCC-GARCH specification for correlations. The distribution of BIC values over the various eigenvalue order constellations is much more even and overall results in the preferred (3,4) model. For comparison, we also report the BIC obtained for a standard HAR specification of the eigenvalue dynamics (see Corsi, 2009). The model boils down to a restricted autoregressive specification of order 20. The HAR model, although very popular in empirical applications, is not preferred in any case.

Table 4.5 shows corresponding model diagnostic results on the in-sample fit to

Order of eigenvalue process	Order of eigenvector process (1,0)	(1,1)	(2,1)	(1,2)	(2,2)	(3,2)	(2,3)	(3,3)
(1,0)	472.14	229.53	229.64	229.17	230.23	230.02	230.19	229.99
(1,1)	463.15	229.00	230.11	229.31	229.44	228.57	228.70	227.98
(2,1)	410.37	216.57	216.88	216.40	217.04	216.74	216.82	216.59
(1,2)	411.70	216.89	216.90	216.70	216.96	216.89	216.94	216.75
(2,2)	410.12	216.50	216.24	216.24	216.66	216.33	216.57	216.23
(3,2)	413.72	222.43	222.82	222.30	223.76	223.47	223.52	221.87
(2,3)	408.66	216.43	216.28	216.19	216.43	216.68	216.61	216.40
(3,3)	409.83	217.14	216.77	216.74	217.02	216.68	216.71	216.65
(4,3)	417.61	219.02	217.56	218.43	218.08	217.56	218.09	217.61
(3,4)	409.43	216.20	216.27	216.06	216.43	216.30	216.40	216.17
(4,4)	408.87	215.96	216.41	**215.76**	216.32	216.20	216.29	216.10
(5,4)	409.95	216.47	216.66	216.12	216.82	216.28	216.75	216.24
(4,5)	409.37	216.32	216.47	216.20	216.41	216.25	216.36	216.08
(5,5)	408.90	216.35	216.37	216.06	216.29	216.14	216.26	215.99
HAR	408.75	218.19	224.40	218.02	218.04	223.19	217.90	218.98

Table 4.5: Average Ljung-Box autocorrelation test statistics of the $100 \times 101 = 5050$ time-series of standardized Martingale differences obtained from DCP-CAW model for the 100-dimensional dataset. The 5% critical value of the Ljung-Box test statistic at 200 lags is $\chi^2_{200} = 233.99$.

the autocorrelation structure of the underlying realized (co)variance data. The model diagnostics are based on Ljung-Box autocorrelation tests on the standardized Martingale differences ("Pearson residuals") in the $p(p+1)/2 = 5050$-dimensional vector

$$\begin{aligned} e_t^* &= \mathbf{Cov}(\text{vech}(\mathbf{R}_t)|\mathcal{F}_{t-1})^{-1/2}\ \text{vech}(\mathbf{R}_t - \mathbb{E}[\mathbf{R}_t|\mathcal{F}_{t-1}]) \\ &= \left(\frac{2}{n}\mathbf{G}_p^+(\mathbf{\Sigma}_t \otimes \mathbf{\Sigma}_t)(\mathbf{G}_p^+)^\top\right)^{-1/2} \text{vech}(\mathbf{R}_t - \mathbf{\Sigma}_t), \end{aligned}$$

where $\mathbf{G}_p^+ = (\mathbf{G}_p^\top \mathbf{G}_p)^{-1}\mathbf{G}_p^\top$ is an elimination matrix with $\mathbf{G}_p$ being the duplication matrix as defined by $\mathbf{G}_p\text{vech}(\mathbf{X}) = \text{vec}(\mathbf{X})$ for symmetric $p \times p$ $\mathbf{X}$ (see e.g. Lütkepohl, 2005, or Section 3.6.1). Under the null of correct model specification, these residuals are serially uncorrelated. The table reports averaged values of the Ljung-Box test statistics at 200 lags computed over the 5050 time-series of Pearson-residuals in the vector e_t^*. The 5% critical value of the Ljung-Box test statistic at 200 lags is $\chi^2_{200} = 233.99$. The results give an impression of the overall fit of the various model constellations to the (co)variance dynamics of the data since the average Ljung-Box test statistics represent an aggregate of the squared

	Eigenvalue Process							
	$\alpha_{i,1}$	$\alpha_{i,2}$	$\alpha_{i,3}$	$\beta_{i,1}$	$\beta_{i,2}$	$\beta_{i,3}$	$\beta_{i,4}$	$\sum_{\ell=1}^{r} \alpha_{i,\ell} + \sum_{\ell=1}^{q} \beta_{i,\ell}$
Median	0.311	0.074	0.000	0.132	0.068	0.135	0.135	0.978
Min.	0.025	0.000	0.000	0.000	0.000	0.000	0.000	0.947
Max.	0.492	0.180	0.116	0.519	0.348	0.373	0.378	0.987

Eigenvector Process		
a	b	$a+b$
0.035	0.962	0.997

Table 4.6: Summary of parameter estimates obtained by the DPC estimator for the 100-dimensional dataset described in Section 4.5 and the BIC selected model order (3,4)-(1,1).

residual autocorrelations for all variance and covariance series. The minimum average Ljung-Box test statistic is obtained for a (1,2) specification for the eigenvector dynamics combined with a (4,4) lag-order for the eigenvalue processes as the best-fitting model for the in-sample (co)variance dynamics. Compared to the BIC results, the residual diagnostics show a tendency for higher lag orders, which is explained by the absence of a penalty term for the number of model parameters. Similar to the BIC results, the HAR specification is not preferred in any case. Moreover, the minimum average Ljung-Box test statistic of 215.76 is very close to the one obtained for the BIC-preferred (1,1)-(3,4) specification (216.20). Hence the BIC - although not consistent - appears to be a reasonable advice for selecting the lag-order of the DPC-CAW.

Table 4.6 reports a summary of the obtained estimates for the BIC-preferred (3,4)-(1,1) DPC-CAW specification. The estimated persistence for the eigenvector- and eigenvalue recursions is very high with $(a + b) = 0.997$ and a median of $\sum_{\ell=1}^{p} \alpha_{i,\ell} + \sum_{\ell=1}^{q} \beta_{i,\ell}$ of 0.978. This corresponds to the findings in Aielli and Caporin (2015) and resembles analogous results for scalar DCC-GARCH applications with intercept targeting.

The right panel of Figure 4.8 (p. 149) shows sample autocorrelation functions of standardized Pearson residuals from the DPC-CAW(3,4)-(1,1) model for exemplary variance and covariance series of four stocks included in the 100-dimensional dataset. The results presented in the Figure are representative for the complete set of stocks. The ACFs are depicted together with 95% Bartlett confidence bands for the variance and covariance series separately and illustrate the overall good fit of the DPC-CAW approach. The model successfully reduces the serial depen-

dence to a minimum. We, however, observe some remaining predictability in the residual series: 441 of the 5050 series do not pass the Ljung-Box test for zero autocorrelation at the 1% level and 100 lags. The literature reports much worse fractions for applications of much lower dimension (see e.g. the model diagnostic results for the flexible CAW specifications in Golosnoy, Gribisch, and Liesenfeld (2012), for a five-dimensional application). The diagnostics therefore imply a good fit to the complex dynamics of 5050 distinct variance and covariance series. Also note that we may interpret some remaining residual predictability as a result of the sparse scalar model structure of the DPC-CAW, which enables applications to high-dimensional covariance matrices while avoiding overfitting and spuriously uncorrelated residuals. The residual ACFs in Figure 4.8 show that remaining predictability is typically found in variance residuals. This may be related to the direct modeling of principal component variances rather than return variances.

4.5.3 Out-of-Sample Forecasting

We now compare the out-of-sample one-day, five-day, and ten-day-ahead forecasting performance of the DPC-CAW specification to alternative forecasting models proposed in the literature on realized covariance modeling. We consider two out-of-sample windows: The first window starts on 01 January 2009 and ends on 31 December 2011, covering the subprime crisis period. The window exhibits a particularly high volatility level and pronounced volatility peaks. The second window covers a period of low to moderate volatility from 01 January 2012 until 31 December 2014, representing normal stock market fluctuations (see the left panel in Figure 4.8 (p. 149) for exemplary time-series plots). The models are re-estimated daily using a rolling window of the previous 1750 covariance measures, i.e. roughly seven years of data. New forecasts are generated based on the updated parameter estimates.

Competing Models and Forecast Evaluation

The scalar Re-DCC model of Bauwens, Storti, and Violante (2012) represents the "natural" competitor for the DPC-CAW approach. The Re-DCC model decouples correlations and variances, which facilitates three-step estimation similar to the DPC estimator (see Bauwens, Storti, and Violante, 2012 for details). The model assumes a conditional central Wishart distribution for the realized covariance measure and decomposes the scale matrix $\boldsymbol{\Sigma}_t$ into

$$\boldsymbol{\Sigma}_t = \mathbf{V}_t \boldsymbol{\rho}_t \mathbf{V}_t, \tag{4.26}$$

where $\mathbf{V}_t = \mathrm{dg}(\sqrt{(\mathbf{\Sigma}_t)_{1\,1}}, \sqrt{(\mathbf{\Sigma}_t)_{2\,2}}, \ldots, \sqrt{(\mathbf{\Sigma}_t)_{pp}})$ and $\boldsymbol{\rho}_t$ is the correlation matrix implied by $\mathbf{\Sigma}_t$. We consider GARCH(r, q) recursions for the conditional variances:

$$(\mathbf{\Sigma}_t)_{ii} = \gamma_i + \sum_{k=1}^{r} \alpha_{k,i}(\mathbf{R}_{t-k})_{ii} + \sum_{l=1}^{q} \beta_{k,i}(\mathbf{\Sigma}_{t-l})_{ii}. \tag{4.27}$$

The correlation matrix $\boldsymbol{\rho}_t$ is parameterized as follows:

$$\boldsymbol{\rho}_t = (1 - a - b)\bar{\boldsymbol{\rho}} + a\mathbf{P}_{t-1} + b\boldsymbol{\rho}_{t-1}, \tag{4.28}$$

where $\mathbf{P}_t$ is the realized correlation matrix

$$\mathbf{P}_t = \mathrm{dg}(\mathrm{vecd}(\mathbf{R}_t))^{-1/2}\mathbf{R}_t\mathrm{dg}(\mathrm{vecd}(\mathbf{R}_t))^{-1/2}, \tag{4.29}$$

where the vecd operator stacks the diagonal elements of a square matrix into a column vector and the dg operator creates a diagonal matrix from vector input. $\bar{\boldsymbol{\rho}}$ is estimated by the sample mean of realized correlation measures ("correlation targeting").

We also consider a constant conditional correlation CAW (CCC-CAW) model since it represents a restricted Re-DCC specification where $a = b = 0$. In a similar fashion, we restrict the DPC-CAW model to $a = b = 0$ in order to obtain the CAW-analogue to the OGARCH model (O-CAW). Additionally, we consider the DPC-CAW$_{0f}$ model, which is obtained by restricting the eigenvalue dynamics of the DPC-CAW model to $\alpha_i = \alpha$ and $\beta_i = \beta\ \forall i = 1, \ldots, p$. This particular model restriction turned out to be favorable in forecasting applications.

We furthermore analyze an exponentially weighted moving average (EWMA) specification, called RiskMetrics (see Morgan, 1996), which boils down to exponential smoothing of RCs using a preset smoothing parameter λ. The forecast of the RC is then given by

$$\mathbb{E}[\mathbf{R}_t|\mathcal{F}_{t-1}] = (1 - \lambda)\mathbf{R}_{t-1} + \lambda\mathbb{E}[\mathbf{R}_{t-1}|\mathcal{F}_{t-2}], \tag{4.30}$$

where λ is set to its typical value for daily data, i.e. $\lambda = 0.94$.

As further forecasting models, we apply the Factor-HEAVY approach of Sheppard and Xu (2019), where we use realized variances of the S&P 500 for the market factor (see Section 1.1 and in particular equation 12 in Sheppard and Xu (2019) for details on the model setup), and the PCA based CAW factor approach of Shen, Yao, and Li (2020) (labeled *Factor-CAW*), with seven factors selected by the eigenvalue criterion as discussed by Shen, Yao, and Li (2020). While the Factor-HEAVY model allows for dynamic loadings and idiosyncratic variances, the Factor-CAW uses the PCA results of Tao et al. (2011), who restrict the loadings and the idiosyncratic covariance matrix to be constant over time. Shen, Yao, and Li (2020)

first compute the realized factor covariance matrix using PCA techniques and then employ diagonal CAW(r, q) processes in order to forecast the factor covariances.

Multi-step-ahead forecasts are obtained by iterating the model recursions and replacing unknown future dependent variables by their forecasts (see the according papers for details). The first two columns of Table 4.7 provide an overview of all considered model specifications. The (r, q) column describes the number of GARCH lags in the conditional variance specifications (Re-DCC and CCC-CAW) or in the eigenvalue recursions (DPC-CAW and O-CAW), respectively. For the Factor-HEAVY model the order refers to the factor-, beta- and idiosyncratic variance dynamics, and for the Factor-CAW approach to the CAW(r, q) process for the factor covariance dynamics.

We now turn to the evaluation of the forecasting performance. Let $L(\hat{\mathbf{X}}, \mathbf{X})$ denote the Euclidean distance of the half-vectorization of the forecast error matrix given by

$$L(\hat{\mathbf{X}}, \mathbf{X}) = \text{vech}(\hat{\mathbf{X}} - \mathbf{X})^\top \text{ vech}(\hat{\mathbf{X}} - \mathbf{X}), \tag{4.31}$$

where $\hat{\mathbf{X}}$ represents a particular matrix forecast and $\mathbf{X}$ the according realization. We apply five different loss functions in order to evaluate the forecasting performance of the considered models:

(i) MSE of predicted covariance matrix: $L(\hat{\mathbf{R}}_t, \mathbf{R}_t)$;

(ii) MSE of predicted variances: $\text{vecd}(\hat{\mathbf{R}}_t - \mathbf{R}_t)^\top \text{vecd}(\hat{\mathbf{R}}_t - \mathbf{R}_t)$;

(iii) MSE of predicted correlation matrix: $L(\hat{\boldsymbol{\rho}}_t, \boldsymbol{\rho}_t)$;

(iv) Variance of predicted global minimum variance portfolio (GMVP): $V_{GMPV,t}$;

(v) QLIKE: $QLIKE_t = \log|\hat{\mathbf{R}}_t| + \text{vec}(\hat{\mathbf{R}}_t^{-1}\mathbf{R}_t)^\top \mathbf{1}$.

The model-specific forecast of the covariance matrix $\mathbf{R}_t$ is given by $\hat{\mathbf{R}}_t = \mathbb{E}[\mathbf{R}_t|\mathcal{F}_{t-1}]$ and accordingly $\hat{\boldsymbol{\rho}}_t = \text{dg}(\text{vecd}(\hat{\mathbf{R}}_t))^{-1/2}\hat{\mathbf{R}}_t\text{dg}(\text{vecd}(\hat{\mathbf{R}}_t))^{-1/2}$. We use the realized kernel estimate $\mathbf{R}_t$ as unbiased proxy for the true covariance matrix at period t.

Loss function (i) considers whole covariance matrix forecasts, while (ii) and (iii) focus on variances and correlations instead. These quantities are of particular interest since DCC frameworks model variance and correlation dynamics separately. Loss function (iv) considers economic losses via computing realized variance of the forecasted global minimum variance portfolio (GMVP) given by $V_{GMPV,t} = \hat{\mathbf{w}}^\top \mathbf{R}_t \hat{\mathbf{w}}$, with $\hat{\mathbf{w}} = \hat{\mathbf{R}}_t\mathbf{1}/(\mathbf{1}^\top\hat{\mathbf{R}}_t^{-1}\mathbf{1})$, where $\mathbf{1}$ is an p-dimensional vector of ones. See e.g. Patton (2011) for a discussion of the properties of the QLIKE loss function (v), which is known to be robust to noisy volatility proxies.

We compute sample averages of the obtained losses over the respective forecasting windows and assess the significance of differences in losses via the model

confidence set (MCS) approach of Hansen, Lunde, and Nason (2011). At a given confidence level $(1 - \alpha)$, the MCS contains the single model or the set of models with the best forecasting performance. We select $\alpha = 0.1$ as suggested by Hansen, Lunde, and Nason (2011) and compute the confidence sets using the stationary bootstrap method with window lengths determined by the maximum number of significant parameters obtained by fitting an AR(r) process on the loss differences and $5,000$ bootstrap replications.

Forecasting Results

The one-day-ahead forecasting results are summarized in Tables 4.7 and 4.8, the five-day-ahead results in Tables 4.9 and 4.10 and the ten-day-ahead results in Tables 4.11 and 4.12 (see pp. 159 – 164).

The DPC-CAW approach (DPC-CAW and DPC-CAW$_{0f}$) provides an overall good forecasting performance over all considered subperiods and forecasting horizons. Moreover, the DPC-CAW model provides the overall best forecasts w.r.t the correlation- and the economically important GMVP loss functions (with the exception of ten-day-ahead predictions in the calm period). In particular, the DPC-CAW approach features significantly lower correlation and GMVP losses compared to up-to-date competitors like the flexible Factor-HEAVY approach of Sheppard and Xu (2019) and the Factor-CAW of Shen, Yao, and Li (2020) across all subperiods and horizons. The DPC-CAW approach with HAR dynamics for the eigenvalue recursions also appears to be particularly strong in forecasting the whole covariance matrix, as indicated by the respective MSE results. For the variance loss, however, the DPC-CAW models are significantly outperformed by the EWMA approach in calm periods and forecasting horizons of five and ten trading days ahead. In turbulent periods the MCS for the variance- and covariance matrix losses are very wide due to a huge dispersion of the Euclidean distances in (4.31), such that differences in MSE losses across models are not significant in most cases. We also note that EWMA appears to be a serious competitor in multi-period forecasting during calm market phases, while the Factor-HEAVY approach is typically preferred under the QLIKE and turbulent market conditions. The Factor-CAW, in contrast, does not appear to be a serious competitor in any case, which may be explained by the model's strong restrictions on the covariance dynamics via imposing time-constant idiosyncratic variances and factor loadings. The Re-DCC-CAW, O-CAW, and CCC-CAW models do not show an overall remarkable performance and are typically outperformed by their competitors.

Taken all together, the results confirm the presumption that the independent modeling of principal component variances with time-varying eigenvectors offers a precise description of covariance and correlation dynamics. In particular, the

GMVP forecasting results highlight the importance of capturing correlation dynamics in the portfolio context. We conclude that the DPC-CAW approach has particularly good forecasting properties and notably outperforms its competitors, especially in covariance-, correlation- and GMVP forecasting.

4.6 Conclusion

In this paper, we propose a Dynamic Principal Component (DPC) CAW model for time-series of high-dimensional realized covariance matrices of asset returns. The model performs a spectral decomposition of the scale matrix of a central Wishart distribution and assumes independent dynamics for the principal components' variances and the eigenvector processes. A three-step estimation procedure similar to the DCC framework for asset returns makes the model applicable to high-dimensional realized covariance matrices.

We analyze the finite sample properties of the three-step estimation approach in an extensive simulation experiment and provide an empirical application to realized covariance matrices for 100 assets traded at the NYSE. The DPC-CAW model has particularly good forecasting properties and outperforms its competitors, including DCC-CAW, Factor-HEAVY and Factor-CAW specifications for RCs.

Volatile Market: 01.01.2009 – 31.12.2011						
Model	(r,q)	Cov matrix	Var	Corr	GMVP $\times 10^2$	QLIKE
DPC-CAW	(1,1)	32289	8322	195.9	37.86	150.8
	(2,2)	32036	8296	195.5	37.74	150.4
	(3,3)	31943	8316	195.1	37.70	149.8
	HAR	**31682**	8275	194.5	37.72	149.5
DPC-CAW$_{0f}$	(1,1)	32596	8511	193.5	37.76	147.0
	(2,2)	32262	8466	193.2	37.67	146.7
	(3,3)	32063	8450	**192.9**	37.65	146.2
	HAR	31981	8415	193.0	**37.64**	146.4
Re-DCC-CAW	(1,1)	32346	8222	229.1	39.72	201.5
	(2,2)	32196	8155	228.8	39.59	200.4
	(3,3)	32335	8239	228.6	39.53	199.3
	HAR	33874	8137	235.9	40.37	186.0
O-CAW	(1,1)	38834	10116	211.6	49.99	148.3
	(2,2)	38626	10112	211.2	49.97	148
	(3,3)	38528	10140	211.3	49.99	147.5
	HAR	38328	10139	210.4	49.91	147.1
CCC-CAW	(1,1)	34359	8222	273.2	41.62	225.2
	(2,2)	34223	8155	273.2	41.54	224.7
	(3,3)	34371	8239	273.2	41.51	223.9
	HAR	35583	8137	273.2	42.40	204.7
EWMA		37178	9823	204.9	38.90	162.9
Factor-HEAVY	(1,1)	32374	8056	208.7	55.11	**118.3**
	(2,2)	32191	**7997**	208.0	54.88	120.3
	(3,3)	31968	8047	207.5	54.76	120.8
	HAR	31935	8053	206.8	55.00	119.9
Factor-CAW	(1,1)	35079	9238	245.4	61.22	216.5
	(2,2)	34658	9157	243.8	60.68	212.3
	(3,3)	34554	9199	243.5	60.71	211.9
	HAR	34237	9118	242.9	60.23	209.6

Table 4.7: Average one-day-ahead forecasting losses for 01.01.2009 – 31.12.2011. The loss functions are defined in Section 5.3.1. The smallest value is shown in bold. Grey shaded values indicate that the 90% model confidence set includes the respective model.

Calm Market: 01.01.2012 – 31.12.2014						
Model	(r,q)	Cov matrix	Var	Corr	GMVP $\times 10^2$	QLIKE
DPC-CAW	(1,1)	1586	540.4	226.2	15.48	73.96
	(2,2)	1580	539.7	226.3	15.49	74.09
	(3,3)	1578	540.1	226.4	15.48	74.32
	HAR	**1571**	538.6	227.0	15.49	74.90
DPC-CAW$_{0f}$	(1,1)	1600	542.4	225.4	15.48	**66.76**
	(2,2)	1590	540.7	225.5	15.49	66.99
	(3,3)	1585	540.2	**225.4**	**15.48**	67.05
	HAR	1577	**538.1**	225.7	15.49	67.17
Re-DCC-CAW	(1,1)	1742	597.8	242.8	16.60	94.67
	(2,2)	1732	593.3	242.7	16.59	94.37
	(3,3)	1729	592.8	242.6	16.59	94.10
	HAR	1799	586.4	244.6	16.92	74.21
O-CAW	(1,1)	1756	609.4	243.9	20.08	110.16
	(2,2)	1750	608.2	244.2	20.08	110.43
	(3,3)	1749	607.6	244.5	20.08	110.59
	HAR	1743	605.5	245.2	20.11	111.14
CCC-CAW	(1,1)	1840	597.8	265.3	18.11	96.27
	(2,2)	1828	593.3	265.3	18.12	96.18
	(3,3)	1824	592.8	265.3	18.16	96.23
	HAR	1911	586.4	265.3	18.69	72.33
EWMA		1715	556.3	239.8	15.74	82.57
Factor-HEAVY	(1,1)	1642	595.9	237.5	22.91	68.84
	(2,2)	1639	596.8	237.3	22.93	67.42
	(3,3)	1640	598.4	237.4	22.91	67.66
	HAR	1624	583.5	237.5	23.00	67.00
Factor-CAW	(1,1)	2446	770.9	299.5	41.80	296.47
	(2,2)	2455	771.9	297.4	41.19	282.06
	(3,3)	2455	772.0	296.5	41.09	277.03
	HAR	2435	769.6	299.7	41.52	292.59

Table 4.8: Average one-day-ahead forecasting losses for 01.01.2012 – 31.12.2014. The loss functions are defined in Section 5.3.1. The smallest value is shown in bold. Grey shaded values indicate that the 90% model confidence set includes the respective model.

Volatile Market: 01.01.2009 − 31.12.2011						
Model	(r,q)	Cov matrix	Var	Corr	GMVP $\times 10^2$	QLIKE
DPC-CAW	(1,1)	41745	10876	214.3	39.34	162.8
	(2,2)	41309	10828	213.1	39.26	162.9
	(3,3)	40643	10774	211.8	39.21	164.6
	HAR	39061	**10642**	**207.6**	39.06	164.8
DPC-CAW$_{0f}$	(1,1)	40177	10741	210.3	39.21	160.3
	(2,2)	39928	10701	209.8	39.16	160.4
	(3,3)	39719	10671	209.5	39.12	161.0
	HAR	**38883**	10643	207.8	**38.93**	162.4
Re-DCC-CAW	(1,1)	44523	14599	234.6	41.92	209.7
	(2,2)	43588	13916	234.3	41.63	206.2
	(3,3)	42829	13403	234.0	41.30	202.8
	HAR	44485	11301	242.2	44.15	176.9
O-CAW	(1,1)	46821	11917	228.7	50.93	158.3
	(2,2)	46446	11866	227.8	50.86	158.4
	(3,3)	45727	11800	226.6	50.87	159.6
	HAR	44256	11737	221.8	50.36	159.2
CCC-CAW	(1,1)	46200	14599	274.2	43.68	232.4
	(2,2)	45294	13981	274.2	43.52	231.0
	(3,3)	44509	13403	274.2	43.34	227.0
	HAR	45776	11301	274.2	46.38	193.4
EWMA		39319	10726	213.0	40.10	179.1
Factor-HEAVY	(1,1)	46365	15265	240.5	58.82	133.2
	(2,2)	45124	14320	238.5	57.62	143.3
	(3,3)	43859	13690	236.7	57.43	142.0
	HAR	42659	12376	239.5	62.98	**120.4**
Factor-CAW	(1,1)	43528	11866	254.7	62.09	217.7
	(2,2)	42886	11713	252.2	61.10	211.0
	(3,3)	42600	11630	251.2	60.50	207.2
	HAR	41240	11302	248.4	59.02	197.4

Table 4.9: Average five-day-ahead forecasting losses for 01.01.2009 − 31.12.2011. The loss functions are defined in Section 5.3.1. The smallest value is shown in bold. Grey shaded values indicate that the 90% model confidence set includes the respective model.

Calm Market: 01.01.2012 – 31.12.2014						
Model	(r,q)	Cov matrix	Var	Corr	GMVP $\times 10^2$	QLIKE
DPC-CAW	(1,1)	1978	595.6	245.0	16.47	69.76
	(2,2)	1933	592.4	243.6	16.39	71.55
	(3,3)	1879	585.4	242.6	16.29	75.90
	HAR	1876	583.2	**241.4**	**16.24**	75.02
DPC-CAW$_{0f}$	(1,1)	1927	592.5	243.4	16.39	68.13
	(2,2)	1909	589.6	243.1	16.37	68.52
	(3,3)	1895	587.2	242.9	16.34	69.12
	HAR	1957	590.6	245.5	16.35	68.18
Re-DCC-CAW	(1,1)	2121	694.8	247.4	17.57	90.19
	(2,2)	2066	670.6	247.2	17.51	88.57
	(3,3)	2029	651.7	247.2	17.43	86.91
	HAR	2347	640.3	249.0	17.79	73.29
O-CAW	(1,1)	2055	651.5	252.3	20.09	102.06
	(2,2)	2024	645.9	252.3	20.18	104.34
	(3,3)	1975	639.7	254.7	20.18	110.69
	HAR	1990	643.7	251.1	20.05	106.14
CCC-CAW	(1,1)	2241	694.8	265.5	19.10	87.23
	(2,2)	2194	675.5	265.5	19.12	86.32
	(3,3)	2140	651.7	265.5	19.12	85.22
	HAR	2516	640.3	265.5	19.71	**67.64**
EWMA		**1798**	**568.3**	248.2	16.31	90.37
Factor-HEAVY	(1,1)	2001	710.2	258.9	24.89	82.52
	(2,2)	1962	683.4	258.2	24.40	85.38
	(3,3)	1921	661.5	257.6	24.31	84.34
	HAR	1849	610.8	256.9	26.15	74.36
Factor-CAW	(1,1)	2659	789.2	298.8	41.41	260.71
	(2,2)	2684	791.2	294.7	40.16	239.35
	(3,3)	2693	791.5	293.2	39.85	230.35
	HAR	2694	789.2	295.3	40.54	234.65

Table 4.10: Average five-day-ahead forecasting losses for 01.01.2012 – 31.12.2014. The loss functions are defined in Section 5.3.1. The smallest value is shown in bold. Grey shaded values indicate that the 90% model confidence set includes the respective model.

Volatile Market: 01.01.2009 − 31.12.2011						
Model	(r,q)	Cov matrix	Var	Corr	GMVP $\times 10^2$	QLIKE
DPC-CAW	(1,1)	42407	11254	221.5	40.48	171.3
	(2,2)	42291	11228	220.6	40.36	172.0
	(3,3)	42256	11194	220.0	40.38	175.3
	HAR	**40828**	11211	**214.7**	40.08	173.6
DPC-CAW$_{0f}$	(1,1)	41765	11145	218.7	40.35	169.4
	(2,2)	41672	11126	218.2	40.28	169.6
	(3,3)	41618	**11123**	217.8	40.20	170.2
	HAR	41011	11263	215.6	**39.80**	170.1
Re-DCC-CAW	(1,1)	48946	17292	239.7	45.08	221.7
	(2,2)	47227	15863	239.4	44.57	218.1
	(3,3)	45779	14767	239.1	44.00	214.1
	HAR	48323	12257	247.8	46.78	187.9
O-CAW	(1,1)	46903	12044	233.0	50.63	164.3
	(2,2)	46752	11996	232.3	50.61	164.6
	(3,3)	46622	11937	231.9	50.69	167.5
	HAR	45254	11974	226.5	50.13	165.3
CCC-CAW	(1,1)	50519	17292	274.8	46.29	243.8
	(2,2)	48971	16108	274.8	46.01	243.3
	(3,3)	47281	14767	274.8	45.52	237.4
	HAR	49293	12257	274.8	48.49	201.3
EWMA		41217	11347	223.3	40.92	197.2
Factor-HEAVY	(1,1)	53993	21772	263.0	62.15	149.8
	(2,2)	50542	18477	259.3	60.19	160.6
	(3,3)	48442	16832	256.0	60.35	158.6
	HAR	46241	14408	255.0	64.94	**136.1**
Factor-CAW	(1,1)	44877	12192	257.5	64.27	221.0
	(2,2)	44265	12067	254.5	62.69	212.6
	(3,3)	44012	12008	253.3	61.67	207.1
	HAR	42949	11845	253.9	60.70	203.1

Table 4.11: Average ten-day-ahead forecasting losses for 01.01.2009−31.12.2011. The loss functions are defined in Section 5.3.1. The smallest value is shown in bold. Grey shaded values indicate that the 90% model confidence set includes the respective model.

Calm Market: 01.01.2012 – 31.12.2014						
Model	(r,q)	Cov matrix	Var	Corr	GMVP $\times 10^2$	QLIKE
DPC-CAW	(1,1)	2368	640.9	258.4	17.17	**68.59**
	(2,2)	2250	630.9	254.7	17.04	70.50
	(3,3)	2107	614.2	250.8	16.84	76.54
	HAR	2104	612.0	**249.5**	16.78	76.46
DPC-CAW$_{0f}$	(1,1)	2235	632.1	253.5	16.94	69.15
	(2,2)	2189	625.4	252.8	16.91	69.44
	(3,3)	2143	618.6	251.9	16.87	69.91
	HAR	2220	623.4	254.6	16.87	69.67
Re-DCC-CAW	(1,1)	2528	761.7	250.7	17.98	93.11
	(2,2)	2402	722.2	250.5	17.93	90.65
	(3,3)	2320	694.5	250.5	17.84	88.76
	HAR	2685	687.7	252.1	18.34	79.24
O-CAW	(1,1)	2358	692.5	255.8	20.03	97.69
	(2,2)	2275	678.4	255.5	20.16	100.57
	(3,3)	2142	662.6	257.1	20.14	109.60
	HAR	2180	667.5	254.7	20.11	104.91
CCC-CAW	(1,1)	2662	761.7	265.7	19.37	87.89
	(2,2)	2558	732.8	265.7	19.39	86.99
	(3,3)	2438	694.5	265.7	19.38	84.93
	HAR	2859	687.7	265.7	19.90	73.43
EWMA		**1889**	**577.1**	255.0	**16.77**	98.88
Factor-HEAVY	(1,1)	2299	828.1	263.8	26.35	91.51
	(2,2)	2209	769.0	263.3	25.34	94.09
	(3,3)	2122	727.8	262.7	25.37	91.42
	HAR	2006	654.6	261.0	26.95	81.64
Factor-CAW	(1,1)	2818	804.0	292.7	40.13	224.23
	(2,2)	2867	807.5	287.9	38.04	201.12
	(3,3)	2882	807.9	286.3	37.52	192.07
	HAR	2821	800.4	292.4	39.52	216.50

Table 4.12: Average ten-day-ahead forecasting losses for 01.01.2012−31.12.2014. The loss functions are defined in Section 5.3.1. The smallest value is shown in bold. Grey shaded values indicate that the 90% model confidence set includes the respective model.

Bibliography

Aielli, Gian P., and Massimiliano Caporin. 2015. "Dynamic Principal Components: A New Class of Multivariate GARCH Models." Working Paper, ssrn.com.

Alexander, Carol O. 2001. "Orthogonal GARCH." *Mastering Risk* (2): 21–38.

Alexander, Carol O., and A. M. Chibumba. 1997. "Multivariate Orthogonal Factor GARCH." Working Paper, University of Sussex.

Andersen, Torben G., Tim Bollerslev, Peter F. Christoffersen, and Francis X. Diebold. 2006. "Volatility and Correlation Forecasting." In *Handbook of Economic Forecasting,* edited by G. Elliott, C. W. J. Granger, and A. Timmermann, 1:777–878. Elsevier.

Andersen, Torben G., Tim Bollerslev, Francis X. Diebold, and Heiko Ebens. 2001. "The Distribution of Realized Stock Return Volatility." *Journal of Financial Economics* 61 (1): 43–76.

Andersen, Torben G., Tim Bollerslev, Francis X. Diebold, and Paul Labys. 2003. "Modeling and Forecasting Realized Volatility." *Econometrica* 71 (2): 579–625.

Asai, Manabu, and Michael McAleer. 2015. "Forecasting Co-Volatilities via Factor Models with Asymmetry and Long Memory in Realized Covariance." *Journal of Econometrics* 189 (2): 251–262.

Asai, Manabu, Michael McAleer, and Jun Yu. 2006. "Multivariate Stochastic Volatility: A Review." *Econometric Reviews* 25 (2-3): 145–175.

Asai, Manabu, and Mike K. P. So. 2013. "Stochastic Covariance Models." *Journal of the Japan Statistical Society* 43 (2): 127–162.

Bai, Jushan, and Zhihong Chen. 2008. "Testing Multivariate Distributions in GARCH Models." *Journal of Econometrics* 143 (1): 19–36.

Baker, Malcolm, Brendan Bradley, and Jeffrey Wurgler. 2011. "Benchmarks as Limits to Arbitrage: Understanding the Low-Volatility Anomaly." *Financial Analysts Journal* 67 (1): 40–54.

Bandi, Federico M., Jeffrey R. Russell, and Chen Yang. 2008. "Microstructure Noise, Realized Variance, and Optimal Sampling." *The Review of Economic Studies* 75 (2): 339–369.

Barndorff-Nielsen, Ole E., Peter R. Hansen, Asger Lunde, and Neil Shephard. 2009. "Realized Kernels in Practice: Trades and Quotes." *The Econometrics Journal* 12 (3): C1–C32.

———. 2011. "Multivariate Realised Kernels: Consistent Positive Semi-Definite Estimators of the Covariation of Equity Prices with Noise and Non-Synchronous Trading." *Journal of Econometrics* 162 (2): 149–169.

Barndorff-Nielsen, Ole E., and Neil Shephard. 2004. "Econometric Analysis of Realized Covariation: High Frequency Based Covariance, Regression, and Correlation in Financial Economics." *Econometrica* 72 (3): 885–925.

Bauer, Gregory H., and Keith Vorkink. 2011. "Forecasting Multivariate Realized Stock Market Volatility." *Journal of Econometrics* 160 (1): 93–101.

Bauwens, Luc, Manuela Braione, and Giuseppe Storti. 2016. "Forecasting Comparison of Long Term Component Dynamic Models for Realized Covariance Matrices." *Annals of Economics and Statistics* (123/124): 103–134.

———. 2017. "A Dynamic Component Model for Forecasting High-Dimensional Realized Covariance Matrices." *Econometrics and Statistics* 1:40–61.

Bauwens, Luc, Sébastien Laurent, and Jeroen V. K. Rombouts. 2006. "Multivariate GARCH Models: A Survey." *Journal of Applied Econometrics* 21 (1): 79–109.

Bauwens, Luc, Giuseppe Storti, and Francesco Violante. 2012. "Dynamic Conditional Correlation Models for Realized Covariance Matrices." Working Paper, CORE.

Bellman, Richard. 1956. "A Generalization of some Integral Identities due to Ingham and Siegel." *Duke Mathematical Journal* 23 (4): 571–577.

Blasques, Francisco, Christian Francq, and Sébastien Laurent. 2022. "Quasi Score-Driven Models." *Journal of Econometrics.*

Blasques, Francisco, Siem J. Koopman, and Andre Lucas. 2015. "Information-Theoretic Optimality of Observation-Driven Time Series Models for Continuous Responses." *Biometrika* 102 (2): 325–343.

Blasques, Francisco, Andre Lucas, Anne Opschoor, and Luca Rossini. 2021. "Tail Heterogeneity for Dynamic Covariance-Matrix-Valued Random Variables: The F-Riesz Distribution." Working Paper, Tinbergen Institute Discussion Paper.

Blasques, Francisco, Janneke van Brummelen, Siem J. Koopman, and André Lucas. 2022. "Maximum Likelihood Estimation for Score-Driven Models." *Journal of Econometrics* 227 (2): 325–346.

Bollerslev, Tim, Andrew J. Patton, and Rogier Quaedvlieg. 2018. "Modeling and Forecasting (Un)Reliable Realized Covariances for More Reliable Financial Decisions." *Journal of Econometrics* 207 (1): 71–91.

Bollerslev, Tim, and Jeffrey M. Wooldridge. 1992. "Quasi-Maximum Likelihood Estimation and Inference in Dynamic Models with Time-Varying Covariances." *Econometric Reviews* 11 (2): 143–172.

Boudt, Kris, Christophe Croux, and Sébastien Laurent. 2011. "Outlyingness Weighted Covariation." *Journal of Financial Econometrics* 9 (4): 657–684.

Brownlees, Christian T., Robert F. Engle, and Bryan T. Kelly. 2012. "A Practical Guide to Volatility Forecasting Through Calm and Storm." *Journal of Risk* 14 (2): 3–22.

Chiriac, Roxana, and Valeri Voev. 2011. "Modelling and Forecasting Multivariate Realized Volatility." *Journal of Applied Econometrics* 26 (6): 922–947.

Corsi, Fulvio. 2009. "A Simple Approximate Long-Memory Model of Realized Volatility." *Journal of Financial Econometrics* 7 (2): 174–196.

Cox, David R. 1981. "Statistical Analysis of Time Series: Some Recent Developments." *Scandinavian Journal of Statistics,* 93–115.

Creal, Drew, Siem J. Koopman, and André Lucas. 2011. "A Dynamic Multivariate Heavy-Tailed Model for Time-Varying Volatilities and Correlations." *Journal of Business and Economic Statistics* 29 (4): 552–563.

———. 2013. "Generalized Autoregressive Score Models with Applications." *Journal of Applied Econometrics* 28 (5): 777–795.

Díaz-García, José A. 2013. "A Note on the Moments of the Riesz Distribution." *Journal of Statistical Planning and Inference* 143 (11): 1880–1886.

———. 2014. "On Riesz Distribution." *Metrika* 77 (4): 469–481.

NIST Digital Library of Mathematical Functions. http://dlmf.nist.gov/, Release 1.1.4 of 2022-01-15. F. W. J. Olver, A. B. Olde Daalhuis, D. W. Lozier, B. I. Schneider, R. F. Boisvert, C. W. Clark, B. R. Miller, B. V. Saunders, H. S. Cohl, and M. A. McClain, eds.

Engle, Robert F. 2002. "Dynamic Conditional Correlation: A Simple Class of Multivariate Generalized Autoregressive Conditional Heteroskedasticity Models." *Journal of Business and Economic Statistics* 20 (3): 339–350.

———. 2009. *Anticipating Correlations: A New Paradigm for Risk Management.* Princeton University Press.

Engle, Robert F., and Kenneth F. Kroner. 1995. "Multivariate Simultaneous Generalized ARCH." *Econometric Theory* 11 (1): 122–150.

Engle, Robert F., Olivier Ledoit, and Michael Wolf. 2019. "Large Dynamic Covariance Matrices." *Journal of Business and Economic Statistics* 37 (2): 363–375.

Faraut, Jacques, and Adam Korányi. 1994. *Analysis on Symmetric Cones.* Oxford University Press.

Francq, Christian, Lajos Horváth, and Jean-Michel Zakoïan. 2014. "Variance Targeting Estimation of Multivariate GARCH Models." *Journal of Financial Econometrics* 14 (2): 353–382.

Geweke, John, and Richard Meese. 1981. "Estimating Regression Models of Finite but Unknown Order." *International Economic Review* 22 (1): 55–70.

Golosnoy, Vasyl, Bastian Gribisch, and Roman Liesenfeld. 2012. "The Conditional Autoregressive Wishart Model for Multivariate Stock Market Volatility." *Journal of Econometrics* 167 (1): 211–223.

Gorgi, Paolo, Peter R. Hansen, Pawel Janus, and Siem J. Koopman. 2019. "Realized Wishart-GARCH: A Score-Driven Multi-Asset Volatility Model." *Journal of Financial Econometrics* 17 (1): 1–32.

Gourieroux, Christian, Joann Jasiak, and Razvan Sufana. 2009. "The Wishart Autoregressive Process of Multivariate Stochastic Volatility." *Journal of Econometrics* 150 (2): 167–181.

Gribisch, Bastian, and Jan P. Hartkopf. 2022. "Modeling Realized Covariance Measures with Heterogeneous Liquidity: A Generalized Matrix-Variate Wishart State-Space Model." *Journal of Econometrics.*

Gribisch, Bastian, and Michael Stollenwerk. 2020. "Dynamic Principal Component CAW Models for High-Dimensional Realized Covariance Matrices." *Quantitative Finance* 20 (5): 799–821.

Gupta, Arjun K., and Daya K. Nagar. 2000. *Matrix Variate Distributions.* Chapman and Hall/CRC.

Gupta, Arjun K., Tamas Varga, and Taras Bodnar. 2013. *Elliptically Contoured Models in Statistics and Portfolio Theory.* Springer.

Hafner, Christian M., and Olga Reznikova. 2012. "On the Estimation of Dynamic Conditional Correlation Models." *Computational Statistics & Data Analysis* 56 (11): 3533–3545.

Hansen, Peter R., and Elena-Ivona Dumitrescu. 2022. "How Should Parameter Estimation be Tailored to the Objective?" *Journal of Econometrics* 230 (2): 535–558.

Hansen, Peter R., Asger Lunde, and James M. Nason. 2011. "The Model Confidence Set." *Econometrica* 79 (2): 453–497.

Harvey, Andrew C. 2013. *Dynamic Models for Volatility and Heavy Tails: With Applications to Financial and Economic Time Series.* Vol. 52. Econometric Society Monographs. Cambridge University Press.

———. 2022. "Score-Driven Time Series Models." *Annual Review of Statistics and Its Application* 9 (1): 321–342.

Harville, David A. 1997. *Matrix Algebra From a Statisticians Perspective.* Springer.

Hassairi, Abdelhamid, Fatma Ktari, and Raoudha Zine. 2022. "On the Gaussian Representation of the Riesz Probability Distribution on Symmetric Matrices." *AStA Advances in Statistical Analysis* 106 (4).

Hautsch, Nikolaus, and Stefan Voigt. 2019. "Large-Scale Portfolio Allocation under Transaction Costs and Model Uncertainty." Big Data in Dynamic Predictive Econometric Modeling, *Journal of Econometrics* 212 (1): 221–240.

Hayashi, Takaki, and Nakahiro Yoshida. 2005. "On Covariance Estimation of Non-Synchronously Observed Diffusion Processes." *Bernoulli* 11 (2): 359–379.

Jin, Xin, Jia Liu, and Qiao Yang. 2021. "Does the Choice of Realized Covariance Measures Empirically Matter? A Bayesian Density Prediction Approach." *Econometrics* 9 (4).

Jin, Xin, and John M. Maheu. 2016. "Bayesian Semiparametric Modeling of Realized Covariance Matrices." *Journal of Econometrics* 192 (1): 19–39.

Jin, Xin, John M. Maheu, and Qiao Yang. 2019. "Bayesian Parametric and Semiparametric Factor Models for Large Realized Covariance Matrices." *Journal of Applied Econometrics* 34 (5): 641–660.

Kelley, Carl T. 1999. *Iterative Methods for Optimization.* SIAM.

Koev, Plamen, and Alan Edelman. 2006. "The Efficient Evaluation of the Hypergeometric Function of a Matrix Argument." *Mathematics of Computation* 75 (254): 833–846.

Kollo, Tõnu, and Dietrich von Rosen. 2005. *Advanced Multivariate Statistics with Matrices.* Springer.

Liu, Lily Y., Andrew J. Patton, and Kevin Sheppard. 2015. "Does Anything Beat 5-minute RV? A Comparison of Realized Measures Across Multiple Asset Classes." *Journal of Econometrics* 187 (1): 293–311.

Louati, Mahdi, and Afif Masmoudi. 2015. "Moment for the Inverse Riesz Distributions." *Statistics & Probability Letters* 102:30–37.

Lunde, Asger, Neil Shephard, and Kevin Sheppard. 2016. "Econometric Analysis of Vast Covariance Matrices Using Composite Realized Kernels and Their Application to Portfolio Choice." *Journal of Business and Economic Statistics* 34 (4): 504–518.

Lütkepohl, Helmut. 1989. "A Note on the Asymptotic Distribution of Impulse Response Functions of Estimated VAR Models with Orthogonal Residuals." *Journal of Econometrics* 42 (3): 371–376.

———. 1996. *Handbook of Matrices.* John Wiley and Sons.

———. 2005. *New Introduction to Multiple Time Series Analysis.* Springer.

Maaß, Hans. 1971. *Siegel's Modular Forms and Dirichlet Series.* Springer.

Magnus, Jan R. 2010. "On the Concept of Matrix Derivative." *Journal of Multivariate Analysis* 101 (9): 2200–2206.

Magnus, Jan R., and Heinz Neudecker. 1980. "The Elimination Matrix: Some Lemmas and Applications." *SIAM Journal on Algebraic Discrete Methods* 1 (4): 422–449.

———. 2019. *Matrix Differential Calculus with Applications in Statistics and Econometrics.* John Wiley and Sons.

McAleer, Michael, and Marcelo C. Medeiros. 2008. "Realized Volatility: A Review." *Econometric Reviews* 27 (1-3): 10–45.

Mittelhammer, Ron C. 2013. *Mathematical Statistics for Economics and Business.* Springer.

Morgan, J. P. 1996. "RiskMetrics." Technical Document.

Murray, Ian. 2016. "Differentiation of the Cholesky Decomposition." Working Paper, arXiv.org.

Noureldin, Diaa, Neil Shephard, and Kevin Sheppard. 2012. "Multivariate High-Frequency-Based Volatility (HEAVY) Models." *Journal of Applied Econometrics* 27 (6): 907–933.

———. 2014. "Multivariate Rotated ARCH Models." *Journal of Econometrics* 179 (1): 16–30.

Olkin, Ingram. 1959. "A Class of Integral Identities with Matrix Argument." *Duke Mathematical Journal* 26 (2): 207–213.

Opschoor, Anne, Pawel Janus, André Lucas, and Dick van Dijk. 2018. "New HEAVY Models for Fat-Tailed Realized Covariances and Returns." *Journal of Business and Economic Statistics* 36 (4): 643–657.

Opschoor, Anne, and André Lucas. 2022. "Time-Varying Variance and Skewness in Realized Volatility Measures." *International Journal of Forecasting.*

Pakel, Cavit, Neil Shephard, Kevin Sheppard, and Robert F. Engle. 2021. "Fitting Vast Dimensional Time-Varying Covariance Models." *Journal of Business and Economic Statistics* 39 (3): 652–668.

Patton, Andrew J. 2011. "Volatility Forecast Comparison using Imperfect Volatility Proxies." *Journal of Econometrics* 160 (1): 246–256.

Rao, Calyampudi R. 1965. *Linear Statistical Inference and Its Applications.* John Wiley and Sons.

Shen, Keren, Jianfeng Yao, and Wai K. Li. 2020. "Forecasting High-Dimensional Realized Volatility Matrices using a Factor Model." *Quantitative Finance* 20 (11): 1879–1887.

Sheppard, Kevin. 2012. "Forecasting High Dimensional Covariance Matrices." In *Handbook of Volatility Models and Their Applications,* edited by L. Bauwens, C. Hafner, and S. Laurent, 103–125. John Wiley and Sons.

Sheppard, Kevin, and Wen Xu. 2019. "Factor High-Frequency-Based Volatility (HEAVY) Models." *Journal of Financial Econometrics* 17 (1): 33–65.

Sucarrat, Genaro, and Steffen Grønneberg. 2020. "Risk Estimation with a Time-Varying Probability of Zero Returns." *Journal of Financial Econometrics* 20 (2): 278–309.

Sutradhar, Brajendra C., and Mir M. Ali. 1989. "A Generalization of the Wishart Distribution for the Elliptical Model and its Moments for the Multivariate t Model." *Journal of Multivariate Analysis* 29 (1): 155–162.

Tao, Minjing, Yazhen Wang, Qiwei Yao, and Jian Zou. 2011. "Large Volatility Matrix Inference via Combining Low-Frequency and High-Frequency Approaches." *Journal of the American Statistical Association* 106 (495): 1025–1040.

Veleva, Evelina. 2009. "Testing a Normal Covariance Matrix for Small Samples with Monotone Missing Data." *Applied Mathematical Sciences* 3 (54): 2671–2679.

Walck, Christian. 2007. *Hand-book on Statistical Distributions for Experimentalists.* University of Stockholm.

Wang, Yazhen, and Jian Zou. 2010. "Vast Volatility Matrix Estimation for High-Frequency Financial Data." *The Annals of Statistics* 38 (2): 943–978.

Wooldridge, Jeffrey M. 1994. "Estimation and Inference for Dependent Processes." Chap. 45 in *Handbook of Econometrics,* 4:2639–2738. Elsevier.

Yu, Philip L. H., Wai K. Li, and F. C. Ng. 2017. "The Generalized Conditional Autoregressive Wishart Model for Multivariate Realized Volatility." *Journal of Business and Economic Statistics* 35 (4): 513–527.

Zhang, Lan. 2011. "Estimating Covariation: Epps Effect, Microstructure Noise." *Journal of Econometrics* 160 (1): 33–47.

Zhang, Lan, Per A. Mykland, and Yacine Aït-Sahalia. 2005. "A Tale of Two Time Scales." *Journal of the American Statistical Association* 100 (472): 1394–1411.

Zhou, Jiayuan, Feiyu Jiang, Ke Zhu, and Wai K. Li. 2019. "Time Series Models for Realized Covariance Matrices based on the Matrix-F Distribution." Working Paper, arXiv.org.