11

Risk Prediction and Portfolio Optimization

Building on the framework from Chapter 10, we now consider some applications of univariate GARCH modeling when working with weekly, daily, or higher frequency financial asset returns data. Section 11.1 overviews their use in conjunction with prediction of value at risk (VaR) and expected shortfall (ES), along with a description of other methods designed for that purpose. Section 11.2 scratches the surface of multivariate GARCH modeling by presenting four such methods, all of which are such that estimation is primarily based on *univariate* GARCH, thus avoiding the curse of the dimensionality issue in estimation and other problems associated with some high-dimensional (and highly parameterized) multivariate GARCH models that have been proposed. The most basic one is the constant-conditional-correlation GARCH, referred to as CCC, and its popular extension, dynamic CC, or DCC. These are used in Section 11.3 to introduce the basics of portfolio optimization, where also the so-called univariate collapsing method for portfolio allocation is discussed, along with the concept of ES span.

11.1 Value at Risk and Expected Shortfall Prediction

The value-at-risk (VaR) and expected shortfall (ES) are among the most popular tail risk measures used in quantitative risk management. For continuous random variable X with finite expected value, the ξ-level ES of X, denoted $\mathrm{ES}(X,\xi)$, can be expressed as the tail conditional expectation

$$\mathrm{ES}(X,\xi) = \frac{1}{\xi}\int_{-\infty}^{q_{X,\xi}} u\, f_X(u)\,\mathrm{d}u = \mathbb{E}[X \mid X \leqslant q_{X,\xi}], \quad \xi \in (0,1), \tag{11.1}$$

where the ξ-quantile of X is denoted $q_{X,\xi}$ and is such that $\mathrm{VaR}(X,\xi) = q_{X,\xi}$ is the ξ-level value-at-risk corresponding to one unit of investment. In some presentations, VaR and ES are the negatives of the definitions above, so that the risk measures are positive numbers. Section III.A.8 provides a discussion of several important issues concerning VaR and ES, derives the ES for several common distributions used in empirical finance, and gives a large number of references to the literature.

One of the primary uses of GARCH modeling is for generating accurate short-term predictions of tail risk measures, based often on daily (or higher frequency) financial asset returns data. The NCT-APARCH(1,1) model from Section 10.4, and the MixN(k), MixN(k,g), and TV(u)-MixN(k) models from Section 10.6, perform very well in this regard. The first of these belongs to the class of models in which a parametric non-Gaussian distribution is coupled with a GARCH-type law of motion for the scale term, for which many variations exist. That class can be extended by further

Linear Models and Time-Series Analysis: Regression, ANOVA, ARMA and GARCH, First Edition. Marc S. Paolella.

allowing for dynamics in the shape parameters of the distribution; see Hansen (1994), Gerlach et al. (2013), and Gabrielsen et al. (2015). In addition to these fully parametric specifications, there are several other methods of VaR and ES prediction for univariate financial return series that were explicitly designed for this purpose. Some of these include:

1) The weighting method of Boudoukh et al. (1998).
   This treats the returns as i.i.d., thus ignoring, among other things, the volatility clustering, but places more weight on recent returns than ones further in the past. Boudoukh et al. (1998) do this by assigning weights that sum to one and decay with a geometric rate. The VaR forecast is determined from the empirical c.d.f. of the weighted returns, i.e., the appropriate sample quantile. This method is appealing because one can argue that the recent past is more important than events further back in time for generating a forecast for one, or a small number of, periods in the future, and thus there is a preference for shorter windows. When done without weighting or accounting for GARCH effects, resulting in what is called the method of **(simple) historical simulation**, past crisis and high-volatility periods are possibly not included in the window for estimation, resulting in the risk for the next period being highly under-estimated. See the following discussion on FHS.
   The idea of weighting observations through time, or, more generally, the assumed underlying i.i.d. sequence of innovations in the likelihood of the data, to account for the fact that the observed data are not i.i.d., or, more generally, that the assumed model is mis-specified, is elaborated upon in Chapter 13.
2) The use of **filtered historical simulation**, or FHS, from Hull and White (1998) and Barone-Adesi et al. (1999, 2002).[1]
   This method fits a GARCH model to (as with all models, a past window of specified length of) the time series of returns, such as (10.2) or (10.10), to generate the deterministic GARCH forecast of the scale term, $\hat{\sigma}_{t+1}$, and also the filtered innovation sequence $\{\widehat{Z}_t\}$. A VaR forecast is then given by $\hat{\sigma}_{t+1}$ times the relevant sample quantile, say q, from the $\{\widehat{Z}_t\}$, and an ES forecast can also be generated based on the $\{\widehat{Z}_t\}$ that exceed q. Pritsker (2006) and Kuester et al. (2006) demonstrate the viability of the method, notably compared to use of (simple) historical simulation.
   While the performance of *all* empirical methods for generating a VaR forecast are dependent on the choice of window size, this is particularly acute for (simple) historical simulation because it ignores the stochastic and highly changing nature of the volatility. In particular, if the recent past is "calm" and the chosen window length is such that the most recent high volatility period is not included, then the VaR forecast will tend to be too liberal, underestimating the risk. Or, imagine the window length is such that it just covers a highly volatile period in the past. As the window is progressed, the crisis period will exit the window, and the VaR prediction drops (in magnitude) severely from one day to the next. FHS is less sensitive to this issue because of the use of a GARCH filter.
   Observe how the non-parametric bootstrap can be applied to the $\{\widehat{Z}_t\}$ and thus used to generate confidence intervals of the VaR and ES. See also Gao and Song (2008), the textbook presentations in Dowd (2005) and Christoffersen (2011), and the references therein, for further information on FHS.

1 See the associated web site http://filteredhistoricalsimulation.com/. Use of FHS is also detailed on the Matlab help page for the topic: Using Bootstrapping and Filtered Historical Simulation to Evaluate Market Risk.

3) The **EVT-GARCH** approach from McNeil and Frey (2000) and the related considerations in Chavez-Demoulin et al. (2014).
   This method is similar to FHS in that it first uses a GARCH filter to (i) obtain the filtered innovations and (ii) generate a prediction of the scale σ_{t+1} based on the information set up to time t (this prediction being deterministic, recalling the discussion near the beginning of Section 10.2), and then fits a generalized Pareto distribution (GPD) to the tails of the filtered innovations (this being motivated by extreme value theory and the so-called **peaks-over-threshold** method, or POT), from which a VaR and ES forecast can be computed. See also Rocco (2014).
   As with FHS, the choice of GARCH filter and also the innovations assumption for the chosen GARCH model play a role in the accuracy of the forecasts, as detailed in Kuester et al. (2006). The fact that the GPD is fit to the filtered innovations to obtain the predictive quantile, as opposed to using the fully specified parametric structure of the GARCH model with a non-Gaussian innovations assumption, would seem to imply that the choice of innovation distribution used in the GARCH filter should not play a role. In fact, according to quasi maximum likelihood theory, the choice should be Gaussian: Recall the discussion in Section 10.3.2 and the findings of Fan et al. (2014) and Anatolyev and Khrapov (2015).
   However, if the (non-Gaussian)-GARCH model is just viewed as an approximate, mis-specified filter to the underlying d.g.p., then use of a flexible one accounting for all the stylized facts of the data should result in the filtered innovation sequence being closer to i.i.d. This appears to be the case, as shown in Kuester et al. (2006). The use of an GAt-APARCH(1,1) model in conjunction with the EVT method of McNeil and Frey (2000) results in excellent out-of-sample performance of VaR forecasts.
4) The robustified semiparametric GARCH method of Mancini and Trojani (2011).
   This method is related to FHS and EVT-GARCH, but employs robust statistical methods for estimation of the filtered scale terms from the GARCH equation, as well as a robustified resampling scheme for the GARCH residuals that controls bootstrap instability due to outlying observations. This leads to improved VaR forecasts and also smoother prediction intervals for VaR over time.
5) Quantile regression methods, namely the so-called **CAViaR** method, initiated in Engle and Manganelli (2004).
   This method is notable because it directly models the quantity of interest, using various functional forms for the VaR. One of its strengths is that use of a GARCH filter is not required, though this method does not fair as well as other methods in horse-race comparisons. Some variations of the method are proposed in Kuester et al. (2006), and an extension allowing for incorporation of implied volatility estimates is considered in Jeon and Taylor (2013). The CAViaR framework has been extended to the multivariate setting in White et al. (2015).
6) Use of conditional autoregressive logit (CARL) models, from Taylor and Yu (2016).
   Several variations of the proposed CARL model class are used in Taylor and Yu (2016) for modeling and forecasting the exceedance probability, i.e., the probability that the realization at time $t + 1$ exceeds a specified value, in either the left or right tail. This is the opposite of VaR prediction, which is a quantile, for a given probability. Taylor and Yu (2016) also propose a time-varying POT method building on the CARL model for VaR and ES prediction, and demonstrate its strong forecasting performance.[2]

---

2 The authors provide code in GAUSS; see http://onlinelibrary.wiley.com/journal/10.1111/(ISSN)1467-985X/homepage/179_4.htm.

7) Use of so-called **expectiles** and the resulting conditional autoregressive expectile (CARE) models; see Taylor (2008), Kuan et al. (2009), Gerlach and Chen (2016), Bellini and Di Bernardino (2017), and the references therein.
8) The use of the Gaussian-GARCH model (10.2) in conjunction with the bootstrap and a bias correction adjustment for improved VaR prediction; see Christoffersen and Gonçalves (2005), Giamouridis (2006), Pascual et al. (2006), and Hartz et al. (2006).
   The benefit of this method is its simplicity, given the very low number of parameters and ease of estimation of the Gaussian-GARCH model compared to more elaborate formulations, as discussed in Section 10.2. See also Chen et al. (2011a).
9) Further non-parametric methods; see Chen and Tang (2005), Cai and Wang (2008), Martins-Filho et al. (2016), Wang and Zhao (2016), and the references therein.
10) Use of realized volatility.
    Based on high-frequency intra-day data, when available, daily realized volatility can be "observed" (i.e., independent of a model and essentially error free) and then used for daily prediction purposes; see Martens (2001), Giot and Laurent (2004), Galbraith and Kisinbay (2005), Koopman et al. (2005), and the references therein. Giot and Laurent (2004) demonstrate with a variety of data sets that the method does *not* lead to improvements in forecast quality when compared to use of a skewed-t A-PARCH model for daily returns.
11) Use of implied volatility induced from option prices.
    A detailed account of volatility prediction based on option prices is given in Poon and Granger (2003). From their review, there is favorable evidence that this model class produces competitive volatility forecasts. See also Cesarone and Colucci (2016), Barone-Adesi (2016), and the references therein.

While the modeling techniques in the above list have been demonstrated to yield competitive VaR forecasts, they do not deliver an entire parametric density forecast for the future portfolio return. Having this density is of value for at least two reasons. First, interest might center not just on prediction of a particular tail risk measure, but rather on the *entire* distribution. Density forecasting has grown in importance in finance and other areas of econometrics because of its added value when working with asymmetric loss functions and non-Gaussian data; see Timmermann (2000) and Tay and Wallis (2000) for surveys, and Amisano and Giacomini (2007) for some associated tests. The second reason for preferring models that deliver an entire (parametric) density forecast is that univariate density predictions for (what turns out to be linear combinations of) individual assets can be analytically combined to yield the density of a *portfolio* of such assets, thus allowing portfolio optimization; see the discussion below in Section 11.3.

Typically, when **backtesting** a model for VaR prediction, i.e., estimating it over moving windows of a large time-series sample and computing, for each window, an h-step-ahead VaR prediction, one computes the resulting sequence of indicator functions (0 or 1) representing whether or not the actual return at time $t + h$ exceeded the forecasted VaR based on the model and the information set up to and including time t. For VaR backtesting, the nonzero components of this sequence are sometimes referred to as **(VaR) violations** or **hits**. If a nominal probability for the VaR quantile of, say, $\xi = 0.01$ is chosen, then, based on a set of w moving windows, one hopes to obtain $w/100$ hits. The resulting

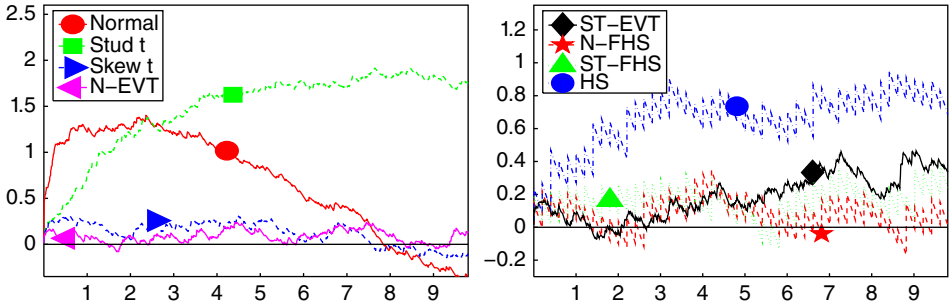

Figure 11.1 Examples of deviation plots for illustrating the unconditional coverage of VaR predictions. The *x*-axis is the VaR level (the tail probability) in percent, with 1, 2.5, and 5 being commonly checked values. The *y*-axis shows the deviation, so that a value of zero is ideal. Instead of showing tables of results for several VaR levels, such a graphic is more appealing and contains more information. The graphics are taken from Kuester et al. (2006), and pertain to VaR forecasts based on moving windows of 500 observations, from the log percentage returns of the daily closing prices of the NASDAQ composite index, from its inception on February 8, 1971, to June 22, 2001, yielding a total of 7,681 observations. The index itself is a market value-weighted portfolio of more than 5,000 stocks. The various models depicted are described in detail in Kuester et al. (2006).

number is obviously a random variable that follows a binomial distribution with parameters w and ξ under the null hypothesis that the model is correct.

The deviation of the actual number of hits compared to the expected number of $w\xi$ is a measure that, based on the $\text{Bin}(w, \xi)$ distribution, is used to assess the quality of the *unconditional* coverage probability associated with the model. While the proportion of hits can be tabulated for various models and probability levels, the use of a graphic is far more appealing and revealing. For example, Figure 11.1 shows the "deviation plots" for several VaR models and a large range of probabilities (from 0.001 to 0.10), as were initiated and used in Kuester et al. (2006). The VaR levels can be read off the horizontal axis, while the vertical axis depicts, for each VaR level, the excess of percentage violations over the VaR level. The goal here is not to compare the models for that particular data set (NASDAQ returns), but to illustrate the use of the deviation plot, and also demonstrate the varying performance of different models. The actual models used include some of the ones discussed above, and are described in detail in Kuester et al. (2006).

While the unconditional coverage is clearly very important, also of strong relevance is the *conditional* coverage, taking into account that the hits should be i.i.d. Bernoulli. For example, if a backtest based on a certain model and number of moving windows w results in precisely $w\xi$ hits, then the unconditional coverage is perfect, but if the hits tend to cluster together, then the model is clearly not generating i.i.d. realizations, and there would be predictability in them (and a sequence of severe losses close in time, which is highly undesirable for financial institutions and investors). For tests that address both unconditional and conditional coverage, see Christoffersen (1998, 2009), Haas (2005, 2009), Francioni and Herzog (2012), Abad et al. (2014), Pelletier and Wei (2016), and the numerous references therein.

Such tests are used in numerous empirical studies, as well as in comparisons of new and existing univariate and multivariate models for VaR prediction; see, e.g., Kuester et al. (2006), Bao et al. (2006, 2007), Santos et al. (2013), Slim et al. (2016), and Paolella and Polak (2017). As an example from the latter paper, Figure 11.2 shows the returns on the equally weighted portfolio through time based on the 30 stocks of the Dow Jones Industrial Index (DJIA), overlaid with the one-step-ahead VaR predictions and the realized hit sequences from several models. Use of an i.i.d. but non-Gaussian model results in both too many hits (though less than obtained with an i.i.d. Gaussian model) and also clustering of the hits, while use of the Gaussian DCC model results in too many hits, but less clustering. The use of the non-Gaussian GARCH so-called COMFORT model, discussed in 11.2.4, results in the best unconditional performance (number of hits) as well as the best conditional performance, i.e., less clustering of realized hits compared to the other models.

Backtesting the performance of the predicted ES is less straightforward, and is an ongoing research topic at the time of writing. See Section III.A.8 for references on backtesting ES amid the fact that it is not **elicitable**.

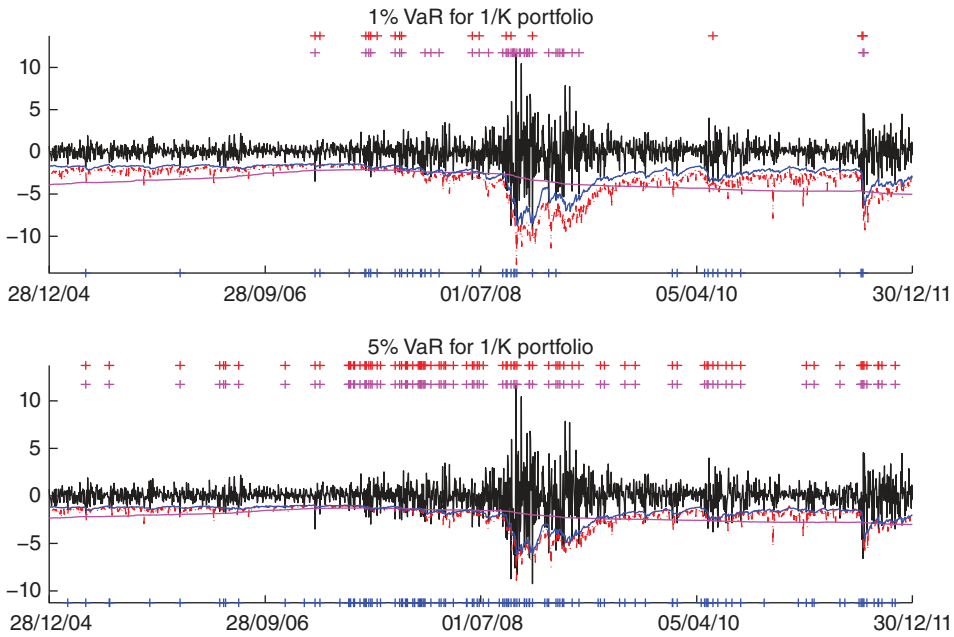

Figure 11.2 Center black lines are the returns on the equally weighted portfolio constructed from the 2,767 daily returns of $K = 30$ components of the DJIA from January 2, 2001, to December 30, 2011 (based on the index composition as of June 8, 2009). Overlaid as colored lines are the associated one-day-ahead 1% (top) and 5% (bottom) VaR forecasts, using: (i) one of the non-Gaussian GARCH COMFORT models (dashed red line), (ii) a non-Gaussian but i.i.d. model (solid magenta line), and the Gaussian DCC model (solid blue line). Further overlaid are the VaR violations, depicted by + signs on the top and bottom of the graphs, using the same color as corresponds to the lines for the VaR predictions.

11.2 MGARCH Constructs Via Univariate GARCH

11.2.1 Introduction

While direct extensions of (10.2) are possible, giving rise to various types of multivariate GARCH (hereafter MGARCH) models, the proliferation of parameters and thus the ensuing estimation problems, for even modest number of assets d, renders many such constructions virtually useless for applications in risk assessment or portfolio management (asset allocation). Several alternative formulations for MGARCH have been proposed that either substantially reduce the number of parameters that require numeric optimization, or, possibly while embodying a potentially large number of parameters, are such that the number of parameters to be *simultaneously* estimated by a generic optimization routine is very small. One fruitful and popular avenue in this latter direction is to build an MGARCH model by use of univariate GARCH models applied to each of the constituent series, sometimes referred to as **equation by equation** modeling, followed by a subsequent step that models the joint correlation structure. This estimation framework is explicitly considered in Francq and Zakoïan (2016),who prove its strong consistency and asymptotic normality in a general framework, including DCC-type models. The subsequent sections illustrate several methods of doing so, though before proceeding, we provide some important remarks.

Remarks

a) Other popular multivariate models include the so-called VEC model of Bollerslev et al. (1988); the BEKK model of Engle and Kroner (1995), so named after the authors of an earlier version, namely Baba, Engle, Kraft, and Kroner (see also Caporin and McAleer, 2008, 2012); the model of Kroner and Ng (1998), which is a weighted average of the CCC and (diagonal) BEKK models; the GARCC random coefficient model of McAleer et al. (2008), which generalizes the BEKK; the factor-GARCH models of Engle et al. (1990), Alexander and Chibumba (1996), Chan et al. (1999), Alexander (2001), Vrontos et al. (2003), and Santos and Moura (2014); and the generalized orthogonal GARCH, or GO-GARCH, models of van der Weide (2002), Lanne and Saikkonen (2007), Zhang and Chan (2009), Broda and Paolella (2009a), Boswijk and van der Weide (2011), and Ghalanos et al. (2015). The GO-GARCH construction is related to the so-called class of rotated ARCH models of Noureldin et al. (2014), which include a variant of the DCC model discussed below, and such that there are only $2d$ or even only $d + 1$ parameters requiring numeric optimization.
   A multivariate extension of the Q-GARCH model (10.11) is given in Sentana (1995), while, as mentioned above, multivariate generalizations of the univariate model in Section 10.6 have been proposed and investigated by Bauwens et al. (2007) and Haas et al. (2009). While these yield models that allow for a very rich dynamic structure, because of parameter proliferation, they are useful for only a small number of assets (though they could be used to drive the factors in a factor-GARCH setup). See the survey articles of Bauwens et al. (2006a) and Silvennoinen and Teräsvirta (2009) for discussions of many of these, and further multivariate model constructions.
b) Asymptotic properties of the variance targeting estimator (VTE) in the multivariate setting have been studied by Pedersen and Rahbek (2014) for the BEKK-GARCH model, and in Francq et al. (2016) for the CCC-GARCH model, while Burda (2015) uses covariance targeting in the general BEKK-GARCH model.

c) All multivariate time-series models share the problem that historical prices for some of the current assets of interest may not be available in the past, such as bonds with particular maturities, private equity, new public companies, merger companies, etc.; see Andersen et al. (2007, p. 515) for discussion and some resolutions to this issue. As our concern herein is on the statistical methodology, we skirt this important issue by considering only equities from major indexes (such as the components of the DJIA). We also ignore the issue of **survivorship bias**, whereby, based on the current date, we obtain past stock prices of the firms in the index, ignoring the fact that, in the past, some companies exited the index (and possibly went bankrupt), and new ones entered. See, e.g., Shumway (1997) and the references therein. This is a form of hindsight bias, and can result in analyses of model performance being exaggerated. When used for actual investment purposes, forecasting applications should attempt to incorporate the probability of bankruptcy. ■

11.2.2 The Gaussian CCC and DCC Models

> …, joint distributions estimated over periods without panics will mis-estimate the degree of correlation between asset returns during panics. Under these circumstances, fear and disengagement by investors often result in simultaneous declines in the values of private obligations, as investors no longer realistically differentiate among degrees of risk and liquidity, and increases in the values of riskless government securities. Consequently, the benefits of portfolio diversification will tend to be overestimated when the rare panic periods are not taken into account.
>
> (Alan Greenspan, 1999)

Arguably the most popular method of generating an MGARCH model via univariate GARCH is the **constant conditional correlation**, or CCC, model of Bollerslev (1990). For each of the component series, a univariate GARCH model is fit and the filtered innovations are ordered as columns in a matrix. The sample correlation of this matrix is used to estimate the correlations. Thus, this MGARCH model fulfils the desired aspect of ease of estimation in two ways. First, with respect to the univariate GARCH models, these require only joint estimation of two (in the Gaussian IGARCH case) to five (Gaussian APARCH) parameters each, and the optimization could be parallelized across assets, for further time savings. Second, via use of the sample correlation estimator applied to the matrix of filtered innovations, this large set of correlation parameters is trivially and nearly instantaneously estimated. This assumes, however, that the correlations are constant through time. Observe that, via the time-varying volatility from the individual fitted GARCH recursions, the covariance matrix itself is changing over time. The **dynamic conditional correlation**, or DCC, model of Engle (2002, 2009), and the **varying correlation**, or VC, model of Tse and Tsui (2002), augments this basic structure with a simple, two-parameter addition that allows for motion also in the correlations, as will be shown below.

We limit our discussion herein to the Gaussian DCC model (of which CCC is a special case) and a semi-parametric variant of it. See Remark (b) below and the subsequent subsections for some discussion of the non-Gaussian CCC, DCC, and other GARCH-type model settings. Let $\mathbf{Y}_t = (Y_{t,1}, Y_{t,2}, \ldots, Y_{t,d})'$ be a d-dimensional vector of asset returns, equally spaced through time (where we use the letter $\mathbf{Y}$ instead of $\mathbf{R}$ for the asset returns because $\mathbf{R}$ will be used to designate a correlation matrix, mimicking the notation in Engle, 2002). The ith univariate series, $i = 1, \ldots, d$, is assumed to follow the Gaussian GARCH(1,1) model (10.2) with possibly unknown mean, given by

$$Y_{t,i} - \mu_i = Z_{t,i}\sigma_{t,i}, \quad \sigma^2_{t,i} = c_{0,i} + c_{1,i}(Y_{t-1,i} - \mu_i)^2 + d_{1,i}\sigma^2_{t-1,i}, \quad Z_t \overset{\text{i.i.d.}}{\sim} \text{N}(0,1). \tag{11.2}$$

Differing from the notation in (10.2) (and to be consistent with that used in Engle, 2002), let $\epsilon_t = Z_t$, with $\boldsymbol{\epsilon}_t = (\epsilon_{t,1}, \dots, \epsilon_{t,d})'$.

We abbreviate $\mathbf{Y}_{t|\Omega_{t-1}}$, where Ω_t is, as in Section 10.2.1, the information set at time t, as just $\mathbf{Y}_{t|t-1}$. The DCC model can then be expressed as

$$\mathbf{Y}_{t|t-1} \sim \mathrm{N}_d(\boldsymbol{\mu}, \mathbf{H}_t), \quad \mathbf{H}_t = \mathbf{D}_t \mathbf{R}_t \mathbf{D}_t, \tag{11.3}$$

in conjunction with (11.2), where $\boldsymbol{\mu} = (\mu_1, \dots, \mu_d)'$, $\mathbf{D}_t^2 = \mathrm{diag}([\sigma_{t,1}^2, \dots, \sigma_{t,d}^2])$, and $\{\mathbf{R}_t\}$ the set of $d \times d$ matrices of time-varying conditional correlations with dynamics specified by

$$\mathbf{R}_t := \mathbb{E}[\boldsymbol{\epsilon}_t \boldsymbol{\epsilon}_t' \mid \Omega_{t-1}] = \mathrm{diag}\,(\mathbf{Q}_t)^{-1/2} \mathbf{Q}_t\, \mathrm{diag}\,(\mathbf{Q}_t)^{-1/2}, \tag{11.4}$$

$t = 1, \dots, T$. Observe that

$$\boldsymbol{\epsilon}_t = \mathbf{D}_t^{-1}(\mathbf{Y}_t - \boldsymbol{\mu}). \tag{11.5}$$

The $\{\mathbf{Q}_t\}$ form a sequence of conditional matrices parameterized by

$$\mathbf{Q}_t = \mathbf{S}(1 - a - b) + a(\boldsymbol{\epsilon}_{t-1} \boldsymbol{\epsilon}_{t-1}') + b\mathbf{Q}_{t-1}, \tag{11.6}$$

with $\mathbf{S}$ the $d \times d$ unconditional correlation matrix (Engle, 2002, p. 341) of the $\boldsymbol{\epsilon}_t$, and parameters a and b are estimated via maximum likelihood conditional on estimates of all other parameters, as discussed next. Matrices $\mathbf{S}$ and $\mathbf{Q}_0$ can be estimated with the usual plug-in sample correlation based on the filtered $\boldsymbol{\epsilon}_t$; see also Bali and Engle (2010) and Engle and Kelly (2012) on estimation of the DCC model. Observe that the resulting $\mathbf{Q}_t$ from the update in (11.6) will not necessarily be precisely a correlation matrix; this is the reason for the standardization in (11.4). The CCC model is a special case of (11.3), with $a = b = 0$ in (11.6).

The mean vector, $\boldsymbol{\mu}$, can be set to zero (and considered to be an extreme shrinkage estimator, with target determined from the economic theory of efficient markets) as done, e.g., in Kroner and Ng (1998, Sec. 5), or estimated using the sample mean of the returns, as in Engle and Sheppard (2001) and McAleer et al. (2008). If estimation is to be used, then, in a more general non-Gaussian context, it is best estimated jointly with the other parameters associated with each univariate return series. This is particularly important amid heavy-tails, in which case the sample mean has relatively low efficiency compared to the m.l.e.; see Paolella and Polak (2017) for some details in this regard.

Let $\mathbf{Y} = [\mathbf{Y}_1, \dots, \mathbf{Y}_T]'$, and denote the set of parameters as $\boldsymbol{\theta}$. The log-likelihood of the remaining parameters, conditional on $\boldsymbol{\mu}$, is given by

$$\begin{aligned} \ell(\boldsymbol{\theta}; \mathbf{Y}, \boldsymbol{\mu}) &= -\frac{1}{2} \sum_{t=1}^{T} (d \ln(2\pi) + \ln(|\mathbf{H}_t|) + (\mathbf{Y}_t - \boldsymbol{\mu})' \mathbf{H}_t^{-1} (\mathbf{Y}_t - \boldsymbol{\mu})) \\ &= -\frac{1}{2} \sum_{t=1}^{T} (d \ln(2\pi) + 2 \ln(|\mathbf{D}_t|) + \ln(|\mathbf{R}_t|) + \boldsymbol{\epsilon}_t' \mathbf{R}_t^{-1} \boldsymbol{\epsilon}_t). \end{aligned} \tag{11.7}$$

Then, as in Engle (2002), adding and subtracting $\boldsymbol{\epsilon}_t' \boldsymbol{\epsilon}_t$, $\ell = \ell(\boldsymbol{\theta}; \mathbf{Y}, \boldsymbol{\mu})$ can be decomposed as the sum of volatility and correlation terms, say $\ell = \ell_V + \ell_C$, where

$$\ell_V = -\frac{1}{2} \sum_{t=1}^{T} (d \ln(2\pi) + 2 \ln(|\mathbf{D}_t|) + \boldsymbol{\epsilon}_t' \boldsymbol{\epsilon}_t), \tag{11.8}$$

and

$$\ell_C = -\frac{1}{2}\sum_{t=1}^{T}(\ln(|\mathbf{R}_t|) + \boldsymbol{\epsilon}_t'\mathbf{R}_t^{-1}\boldsymbol{\epsilon}_t - \boldsymbol{\epsilon}_t'\boldsymbol{\epsilon}_t). \tag{11.9}$$

In this way, a two-step maximum likelihood estimation procedure can be applied. First, estimate the GARCH model parameters for each univariate returns series as discussed in Section 10.2.3, and construct the standardized residuals. Second, maximize the conditional likelihood with respect to parameters a and b in (11.6) based on the filtered residuals from the previous step. Note that, in the CCC model, the correlation matrix is assumed to be constant over time, with $\mathbf{R}_t = \mathbf{R}$, and the standardization step in (11.4) is not necessary.

Remarks

a) See Caporin and McAleer (2013) for several critiques of the DCC construction (including the standardization step (11.4)), and Aielli (2013) for a modified DCC model, termed cDCC, with potentially better small-sample properties. Fermanian and Malongo (2017) provide conditions for the existence and uniqueness of strictly stationary solutions of the DCC model. An interesting alternative to the DCC model is discussed in Section 11.2.3.

b) One might argue that having only two parameters for modeling the evolution of an entire correlation matrix will not be adequate. While this is certainly true, the models of Engle (2002) and Tse and Tsui (2002) have two strong points: First, their use is perhaps better than no parameters (as in the CCC model) and, second, it allows for easy implementation and estimation. Matrix generalizations of the simple DCC structure that allow the number of parameters to be a function of d, and also introducing asymmetric extensions of the DCC idea, are considered in Engle (2002) and Cappiello et al. (2006), though with a potentially very large number of parameters, the usual estimation and inferential problems arise.

   Bauwens and Rombouts (2007a) consider an approach in which similar series are pooled into one of a small number of clusters, such that their GARCH parameters are the same within a cluster. A related idea is to group series with respect to their correlations, generalizing the DCC model; see, e.g., Vargas (2006), Billio et al. (2006), Zhou and Chan (2008), Billio and Caporin (2009), Engle and Kelly (2012), So and Yip (2012), Aielli and Caporin (2013), and the references therein.

   An alternative approach is to assume a Markov switching structure between two (or more) regimes, each of which has a CCC structure, as first proposed in Pelletier (2006), and augmented to the non-Gaussian case in Paolella et al. (2018a). Such a construction implies many additional parameters, but their estimation makes use of the usual sample correlation estimator, thus avoiding the curse of dimensionality, and shrinkage estimation can be straightforwardly invoked to improve performance. The idea is that, for a given time segment, the correlations are constant, and take on one set (of usually two, or at most three sets) of values. This appears to be better than attempting to construct a model that allows for their variation at every point in time. The latter, notably with the aforementioned matrix asymmetric DCC extensions, might be "asking too much of the data" and inundated with too many parameters requiring joint numeric optimization. Paolella et al. (2018a) demonstrate strong out-of-sample performance of their non-Gaussian Markov switching CCC model with two regimes, compared to the Gaussian CCC case, the Gaussian CCC switching case, the Gaussian DCC model, and the non-Gaussian single component CCC of Paolella and Polak (2015a).

c) CCC- and DCC-type MGARCH models that support non-Gaussian innovation processes have been proposed by various researchers. These include Aas et al. (2005), using the multivariate normal inverse Gaussian (NIG); Jondeau et al. (2007, Sec. 6.2) and Wu et al. (2015), using the multivariate skew-Student density; Santos et al. (2013) using a multivariate Student's t; Virbickaite et al. (2016) using a Dirichlet location-scale mixture of multivariate normals; and Paolella and Polak (2015b,c, 2017) using the multivariate generalized hyperbolic, the latter in a full maximum-likelihood framework applicable for large d because of the availability of an EM algorithm; see Section 11.2.4 below. ■

11.2.3 Morana Semi-Parametric DCC Model

Morana (2015) proposes a variation of the DCC model that incorporates a semi-parametric aspect, and denotes it SP-DCC. See also Morana (2017) and Morana and Sbrana (2017) for further details, applications, and simulation results. Similar to (11.3), let

$$\mathbf{Y}_t = \boldsymbol{\mu} + \boldsymbol{\epsilon}_t, \quad \text{and} \quad \boldsymbol{\epsilon}_t = \mathbf{H}_t^{1/2}\mathbf{Z}_t, \tag{11.10}$$

where $\boldsymbol{\epsilon}_t = (\epsilon_{t,1}, \ldots, \epsilon_{t,d})'$ and $\mathbf{Z}_t$ is i.i.d. with first two moments $\mathbb{E}[\mathbf{Z}_t] = \mathbf{0}_d$ and $\mathbb{V}(\mathbf{Z}_t) = \mathbf{I}_d$. Observe the difference between $\boldsymbol{\epsilon}_t$ in (11.5) as used for the DCC construction, and $\boldsymbol{\epsilon}_t$ as used here, this arising as it was deemed desirable to keep the notations used in the original works. They are related by $\boldsymbol{\epsilon}_t = \mathbf{D}_t^{-1}\boldsymbol{\epsilon}_t$. The mean term $\boldsymbol{\mu}$ is estimated as in the DCC model, namely via the sample averages of the return series.

Denote the (ij)th element of $\mathbf{H}_t$ by $h_{t,ij}$, $i, j = 1, \ldots, d$, and assume that the conditional variances $h_{t,i} := h_{t,ii} = \mathbb{V}(Y_{t,i} \mid \Omega_{t-1})$ respectively follow the strictly stationary GARCH(1,1) process (11.2) (using the notation of Morana, 2015, which is the same as shown in (10.3), but now for the ith series)

$$h_{t,i} = \omega_i + \alpha_i \epsilon_{t-1,i}^2 + \beta_i h_{t-1,i}, \quad i = 1, \ldots, d. \tag{11.11}$$

Differing from the usual DCC construction, the conditional covariances are determined by use of the **polarization identity**

$$4 \cdot \mathrm{Cov}(A, B) = \mathbb{V}(A + B) - \mathbb{V}(A - B), \tag{11.12}$$

arising from the simple fact given in (III.A.62) that $\mathbb{V}(A \pm B) = \mathbb{V}(A) + \mathbb{V}(B) \pm 2\mathrm{Cov}(A, B)$, for any two random variables A and B with existing second moments. The off-diagonal elements of $\mathbf{H}_t$, $h_{t,ij} = \mathrm{Cov}(Y_{t,i}, Y_{t,j} \mid \Omega_{t-1})$, can then be represented as

$$4 \cdot h_{t,ij} = \mathbb{V}_{t-1}(Y_{t,i} + Y_{t,j}) - \mathbb{V}_{t-1}(Y_{t,i} - Y_{t,j}), \quad i, j = 1, \ldots, d, \quad i \neq j, \tag{11.13}$$

where $\mathbb{V}_{t-1}(Y_{t,i})$ is shorthand for $\mathbb{V}(Y_{t,i} \mid \Omega_{t-1})$. Next, define the aggregate variables

$$Y_{t,ij}^+ := Y_{t,i} + Y_{t,j}, \quad Y_{t,ij}^- := Y_{t,i} - Y_{t,j}, \quad \epsilon_{t,ij}^+ := \epsilon_{t,i} + \epsilon_{t,j}, \quad \epsilon_{t,ij}^- := \epsilon_{t,i} - \epsilon_{t,j}, \tag{11.14}$$

and assume the conditional variance processes $h_{t,ij}^+ := \mathbb{V}_{t-1}(Y_{t,ij}^+)$ and $h_{t,ij}^- := \mathbb{V}_{t-1}(Y_{t,ij}^-)$ are given, respectively, by the GARCH(1,1) specifications

$$h_{t,ij}^+ = \omega_{ij}^+ + \alpha_{ij}^+ \epsilon_{t-1,ij}^{+2} + \beta_{ij}^+ h_{t-1,ij}^+, \quad i, j = 1, \ldots d, \quad i \neq j, \tag{11.15}$$

and

$$h_{t,ij}^- = \omega_{ij}^- + \alpha_{ij}^- \epsilon_{t-1,ij}^{-2} + \beta_{ij}^- h_{t-1,ij}^-, \quad i, j = 1, \ldots, d, \quad i \neq j. \tag{11.16}$$

By substituting (11.15) and (11.16) into (11.13), the implied parametric structure for the conditional covariance $h_{ij,t}$ can be expressed as

$$\begin{aligned} 4 \cdot h_{t,ij} &= \omega_{ij}^{+} + \alpha_{ij}^{+}\varepsilon_{t-1,ij}^{+2} + \beta_{ij}^{+}h_{t-1,ij}^{+} - \omega_{ij}^{-} - \alpha_{ij}^{-}\varepsilon_{t-1,ij}^{-2} - \beta_{ij}^{-}h_{t-1,ij}^{-} \\ &= \omega_{ij}^{+} - \omega_{ij}^{-} + \alpha_{ij}^{+}(\varepsilon_{t-1,i} + \varepsilon_{t-1,j})^2 - \alpha_{ij}^{-}(\varepsilon_{t-1,i} - \varepsilon_{t-1,j})^2 \\ &\quad + \beta_{ij}^{+}h_{t-1,ij}^{+} - \beta_{ij}^{-}h_{t-1,ij}^{-}. \end{aligned} \tag{11.17}$$

Note that, by assuming constant GARCH parameters across aggregate series, i.e., $\alpha_{ij}^{+} = \alpha_{ij}^{-} =: \alpha$ and $\beta_{ij}^{+} = \beta_{ij}^{-} =: \beta$, and rearranging (11.17) with $\omega_{ij} := (\omega_{ij}^{+} - \omega_{ij}^{-})/4$,

$$h_{t,ij} = \omega_{ij} + \alpha\varepsilon_{t-1,i}\varepsilon_{t-1,j} + \beta h_{t-1,ij},$$

showing how the SP-DCC model is more flexible than the usual DCC construct.

The log-likelihood can be expressed as in (11.7), decomposed similarly, as $\ell = \ell_V + \ell_C$, and a two-step procedure can be used. The volatility part of the likelihood is the same as in (11.8), namely (and recalling $\boldsymbol{\epsilon}_t = \mathbf{D}_t^{-1}\boldsymbol{\varepsilon}_t$)

$$\ell_V = -\frac{1}{2}\sum_{t=1}^{T}(d\ln(2\pi) + 2\ln(|\mathbf{D}_t|) + \boldsymbol{\varepsilon}_t'\mathbf{D}_t^{-1}\mathbf{D}_t^{-1}\boldsymbol{\varepsilon}_t). \tag{11.18}$$

Differing from the DCC model, SP-DCC does not maximize (11.9), but rather the sum of individual GARCH likelihoods for the aggregate series $Y_{t,ij}^{+}$ and $Y_{t,ij}^{-}$, i.e., $\ell_{\text{SP}} = \ell_{\text{SP}}^{+} + \ell_{\text{SP}}^{-}$, where

$$\ell_{\text{SP}}^{+} = -\frac{1}{2}\sum_{t=1}^{T} 2\sum_{i=1}^{d}\sum_{j>i}^{d}\left(\ln(2\pi) + \ln h_{t,ij}^{+} + \frac{\varepsilon_{t,ij}^{+2}}{h_{t,ij}^{+}}\right),$$

and similarly for ℓ_{SP}^{-}. This is jointly maximized by separately maximizing each term. Hence, the conditional variances for the aggregates $h_{t,ij}^{+}$ and $h_{t,ij}^{-}$, $i, j = 1, \dots, d$, $i \neq j$, are estimated equation by equation by means of quasi-maximum likelihood using the aggregated conditional mean residuals $\varepsilon_{t,ij}^{+}$ and $\varepsilon_{t,ij}^{-}$ from (11.14).

Through the polarization identity, the $h_{t,ij}$, $i, j = 1, \dots, d$, $i \neq j$, are estimated non-parametrically via $4 \cdot \hat{h}_{t,ij} = \hat{h}_{t,ij}^{+} - \hat{h}_{t,ij}^{-}$, $i, j = 1, \dots, d$, $i \neq j$. Finally, the estimator of the conditional correlation matrix $\mathbf{R}_t$ is given by $\widehat{\mathbf{R}}_t = \widehat{\mathbf{D}}_t^{-1}\widehat{\mathbf{H}}_t\widehat{\mathbf{D}}_t^{-1}$, where $\widehat{\mathbf{D}}_t^2 = \text{diag}([\hat{h}_{1,t}, \dots, \hat{h}_{d,t}])$.

As in Morana (2015), an ex-post correction to ensure that $\widehat{\mathbf{R}}_t$ is positive definite at each point in time can be implemented as follows. First, if required, the estimated conditional correlations in $\widehat{\mathbf{R}}_t$, $\hat{\rho}_{ij,t}$, $i \neq j$, are bounded to lie within the range $-1 \leqslant \hat{\rho}_{t,ij} \leqslant 1$ by applying the so-called **sign-preserving bounding transformation**

$$\hat{\rho}_{t,ij}^{*} = \hat{\rho}_{t,ij}(1 + \hat{\rho}_{t,ij}^{k})^{-1/k}, \quad k \in \{2, 4, \dots\}, \tag{11.19}$$

where k is selected optimally by minimizing the sum of squared Frobenious norms over the temporal sample

$$\arg\min_{k}\sum_{t=1}^{T}\|\widehat{\mathbf{R}}_t - \widehat{\mathbf{R}}_t^{*}\|_F^2 = \arg\min_{k}\sum_{t=1}^{T}\sum_{i=1}^{d}\sum_{j=1}^{d}|\hat{\rho}_{ij,t} - \hat{\rho}_{ij,t}^{*}|^2. \tag{11.20}$$

Second, if required, positive definiteness is enforced by means of nonlinear shrinkage of the negative eigenvalues of the $\widehat{\mathbf{R}}_t^{*}$ matrix toward their corresponding positive average values over the temporal

sequence in which they are positive. Denote the spectral decomposition as $\widehat{\mathbf{R}}_t^* = \widehat{\mathbf{E}}_t \widehat{\mathbf{V}}_t \widehat{\mathbf{E}}_t'$, where $\widehat{\mathbf{V}}_t$ is the diagonal matrix of sorted eigenvalues and the columns of $\widehat{\mathbf{E}}_t$ are the associated orthogonal eigenvectors, and let $\widehat{\mathbf{V}}_t^*$ be the diagonal matrix with adjusted eigenvalues. The adjusted estimators are then

$$\widehat{\mathbf{R}}_t^{**} = \widehat{\mathbf{E}}_t \widehat{\mathbf{V}}_t^* \widehat{\mathbf{E}}_t', \quad \text{and} \quad \widehat{\mathbf{H}}_t^{**} = \widehat{\mathbf{D}}_t \widehat{\mathbf{R}}_t^{**} \widehat{\mathbf{D}}_t. \tag{11.21}$$

An implementation of the SP-DCC method is available as part of the set of programs associated with the book.[3] Function `SPDCC1step` is such that, for an input set of returns data, the mean vector and variance-covariance matrix corresponding to the one-step-ahead predictive density are output, and thus can be used for portfolio optimization, as described below in Section 11.3.2.

11.2.4 The COMFORT Class

Recall from Section 10.3.1 that, in the univariate GARCH setting, when modeling daily (or higher frequency) financial asset returns with interest centering on density or VaR forecasting, the assumption on the innovations distribution almost always plays a more important role than does the functional form of the law of motion for the scale term. It is, unsurprisingly, also the case in the multivariate setting, notably amid non-ellipticity of the returns, and with portfolio allocation applications in mind. It thus suggests itself to use a CCC or DCC structure with a non-Gaussian distribution, though this is not so trivial in terms of estimation.

A common, simple way of attempting to address this has been to employ a two-step procedure, whereby first, via an appeal to quasi maximum likelihood (recall Section 10.3.2), a Gaussian CCC or DCC model is fit to the data, and, based on the ensuing residuals, a non-Gaussian distribution, such as the multivariate Student's t, is fit (see, e.g., Santos et al., 2013, and the references therein). Conveniently for applied researchers, both steps are available in numerous canned econometrics packages such as Eviews. However, use of this *ad hoc* method is certainly inferior to full m.l.e., and is not obvious if the resulting parameters are consistent. Its use is also compounded by the possibility of incorrectly accounting for how the dispersion matrix in the assumed non-Gaussian multivariate distribution is estimated; see Paolella and Polak (2017) for details.

The latter authors also show that use of joint maximum likelihood estimation, enabled by use of an EM algorithm developed in Paolella and Polak (2015b) for a CCC model with a multivariate generalized hyperbolic distribution (of which Student's t is a limiting case), results in superior out-of-sample density and value-at-risk forecasting performance. It also delivers impressive portfolio performance—far better than use of Gaussian DCC; see Paolella and Polak (2015c). The price to pay for using this **common market factor non-Gaussian returns model**, or COMFORT, is having to understand a more complicated stochastic process and the required estimation technique.

The starting point of the model is the **multivariate normal mean-variance mixture distribution** or MNMVM. The d-dimensional random vector $\mathbf{Y}$ is said to have such a distribution if $\mathbf{Y} = \mathbf{m}(G) + \mathbf{H}^{1/2}\sqrt{G}\mathbf{Z}$, where $\mathbf{Z} \sim \mathrm{N}(\mathbf{0}, \mathbf{I}_d)$, where $G \geqslant 0$ is a non-negative, univariate random variable, independent of $\mathbf{Z}$, $\mathbf{H}$ is a $d \times d$ symmetric and positive definite matrix, and $\mathbf{m} : [0, \infty) \to \mathbb{R}^d$ is a measurable function. The name MNMVM comes from the fact that $\mathbf{Y} \mid (G = g) \sim \mathrm{N}(\mathbf{m}(g), g\mathbf{H})$. The multivariate generalized hyperbolic (MGHyp) distribution, as introduced by Barndorff-Nielsen (1977),

3 The author is grateful to Claudio Morana and Matthias Hartmann for supplying their original codes, and to Marco Gambacciani for adapting them for use with the profile likelihood method of univariate GARCH estimation from Section 10.2.3 and generating the one-step-ahead forecasts, as required for the predictive density.

is a special case of MNMVM with $\mathbf{m}(G) = \mu + \boldsymbol{\gamma} G$, for $d \times 1$ vector $\boldsymbol{\gamma} \in \mathbb{R}^d$ and $G \sim \text{GIG}(\lambda, \chi, \psi)$, i.e., generalized inverse Gaussian. A highly detailed presentation in the univariate case, with links to the many special cases, and details on the GIG distribution, is given in Chapter II.9. (We will see this form again, and go into more detail, in Section 12.2, for the special case of the multivariate noncentral Student's t distribution, and further generalize the structure in Section 12.6.)

One of the benefits of use of the MGHyp for applications to portfolio optimization in finance is that, if $\mathbf{Y} \sim \text{MGHyp}(\boldsymbol{\mu}, \boldsymbol{\gamma}, \mathbf{H}, \lambda, \chi, \psi)$, where $\boldsymbol{\mu}$ is a location vector and $\mathbf{H}$ is a dispersion matrix, then the weighted sums of margins (the portfolio distribution), say $\mathbf{w}'\mathbf{Y}$, is univariate GHyp, i.e., $\mathbf{w}'\mathbf{Y} \sim \text{GHyp}(\mathbf{w}'\boldsymbol{\mu}, \mathbf{w}'\boldsymbol{\gamma}, \mathbf{w}'\mathbf{H}\mathbf{w}, \lambda, \chi, \psi)$. See, e.g., McNeil et al. (2015) for a proof. Different choices of shape parameters λ, χ, and ψ give rise to different tail behavior, from thin tails (the Gaussian and Laplace are limiting and special cases), to so-called semi-heavy tails such that the distribution is leptokurtic but still possesses a moment generating function, to genuinely heavy tailed (the Student's t being a limiting case). While the parameters of the MGHyp, notably the shape parameters λ, χ, and ψ, are identified, certain parameter restrictions are required; see McNeil et al. (2015) and Paolella and Polak (2015b) for details. Furthermore, use of all three shape parameters with typically sized data sets results in a rather flat likelihood, so one usually restricts one or two of them, giving rise to the numerous known special cases of the distribution.

The COMFORT model uses the MGHyp distribution with a CCC or DCC augmentation of the dispersion matrix. That is, for a set of d financial assets, with associated (percentage log) return vector $\mathbf{Y}_t = (Y_{t,1}, Y_{t,2}, \ldots, Y_{t,d})'$, for time $t = 1, \ldots, T$, the model is given by

$$\mathbf{Y}_t = \boldsymbol{\mu} + \boldsymbol{\gamma} G_t + \boldsymbol{\varepsilon}_t, \quad \boldsymbol{\varepsilon}_t = \mathbf{H}_t^{1/2} \sqrt{G_t} \mathbf{Z}_t, \tag{11.22}$$

where $\mathbf{H}_t = \mathbf{S}_t \boldsymbol{\Gamma}_t \mathbf{S}_t$, such that $\mathbf{S}_t = \text{diag}(s_{1,t}, \ldots, s_{d,t})$ is a scale matrix, and $s_{k,t} > 0$, $k = 1, \ldots, d$, are the scale terms driven by the modified GARCH equation dynamics

$$s_{k,t}^2 = \omega_k + \alpha_k (\, y_{t-1,k} - \mu_k - \gamma_k G_{t-1})^2 + \beta_k s_{k,t-1}^2. \tag{11.23}$$

The dependency matrix $\boldsymbol{\Gamma}_t$ can be assumed time invariant, $\boldsymbol{\Gamma}_t = \boldsymbol{\Gamma}$, as in the CCC model, or structured analogous to the DCC model.

As mentioned above, despite the large parameterization as d increases, estimation by full (joint with all parameters) maximum likelihood estimation is feasible and fast via use of an EM algorithm; see Paolella and Polak (2015b) for details.

This modeling paradigm turns out to yield some remarkable results and insights:

1) The required incorporation of a sequence of univariate latent (positive, continuous) random variables, denoted $\{G_t\}$, can be endowed with the interpretation as a **common market factor**, and is able to account for information arrivals and jumps in such a way that, conditional on it, the returns distribution is Gaussian. This allows for two, essentially orthogonal, structures to model the data: A univariate "jump process" for modeling aberrations and news arrivals, and a GARCH structure for modeling the persistence in volatility. Even for very large d, as is typical for portfolios of major financial institutions, all model parameters are quickly and simultaneously estimated via joint maximum likelihood, enabled by an EM algorithm. This results in the $\{G_t\}$ sequence being imputed (filtered), and it can be plotted.
   As an example, the COMFORT model was fit to 11 years of daily data consisting of the $d = 30$ stocks that comprise the DJIA index. The top panel of Figure 11.3 plots the returns for Merck & Co. Inc., with the second and third rows showing the filtered $\{G_t\}$ sequence and the filtered scaled

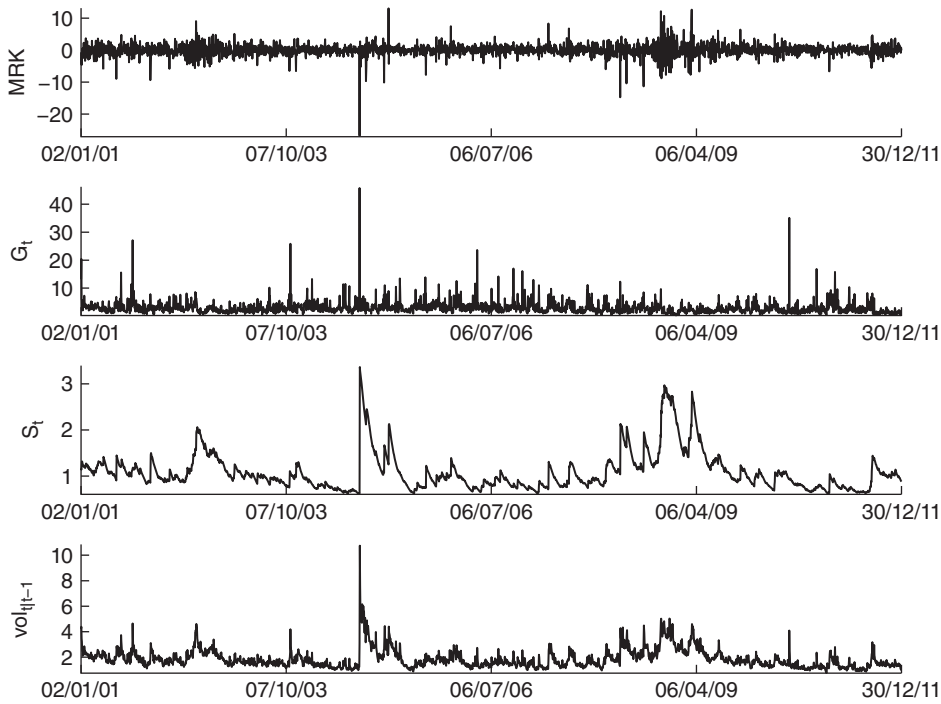

Figure 11.3 The (log percentage) returns on Merck & Co. Inc. for the dates indicated (top) and several filtered time series associated with the COMFORT model.

terms, denoted S_t, from the conditional Gaussian GARCH equation associated with Merck. The fourth panel shows the volatilities, as computed by appropriately combining the $\{G_t\}$ and $\{S_t\}$ (see Paolella and Polak, 2015b, for details). Observe how there is little relation between $\{G_t\}$ and $\{S_t\}$: As a first example, at the return below −20%, the filtered G_t value "picks this up", though (because of the bad news arrival) it was also the onset of a high volatility period, as seen in $\{S_t\}$.

As a second example, around the time of the global financial crisis in 2008 and 2009, the volatility of Merck, as seen from the returns in the top panel, is clearly relatively very high, as is, correspondingly, the $\{S_t\}$ around that period, *but the* $\{G_t\}$ *sequence is rather quiet* because, while the volatility is persistent and being adequately modeled by GARCH, there were no "major surprises" that needed to be caught by G_t. There are a handful of very large G_t "spikes" outside of the one associated with the over −20% drop, and these are not associated with any particular increase in the filtered S_t, but do influence the volatility via the combination of $\{G_t\}$ and $\{S_t\}$. The idea is that the latter two quantities are somewhat orthogonal and each is "doing its job". Without G_t, all there would be is the GARCH-induced volatility, and, from the visible enormous variation in $\{G_t\}$, it is clear that without $\{G_t\}$, the model would be rather mis-specified.

Figure 11.4 is similar, but shows all 30 series overlaid. The graphic emphasizes that $\{G_t\}$ is a univariate sequence, and also shows that the various $\{S_t\}$ are highly correlated through time, as are the COMFORT volatilities in the last panel, though they exhibit more variation than just the $\{S_t\}$ because of the influence of $\{G_t\}$.

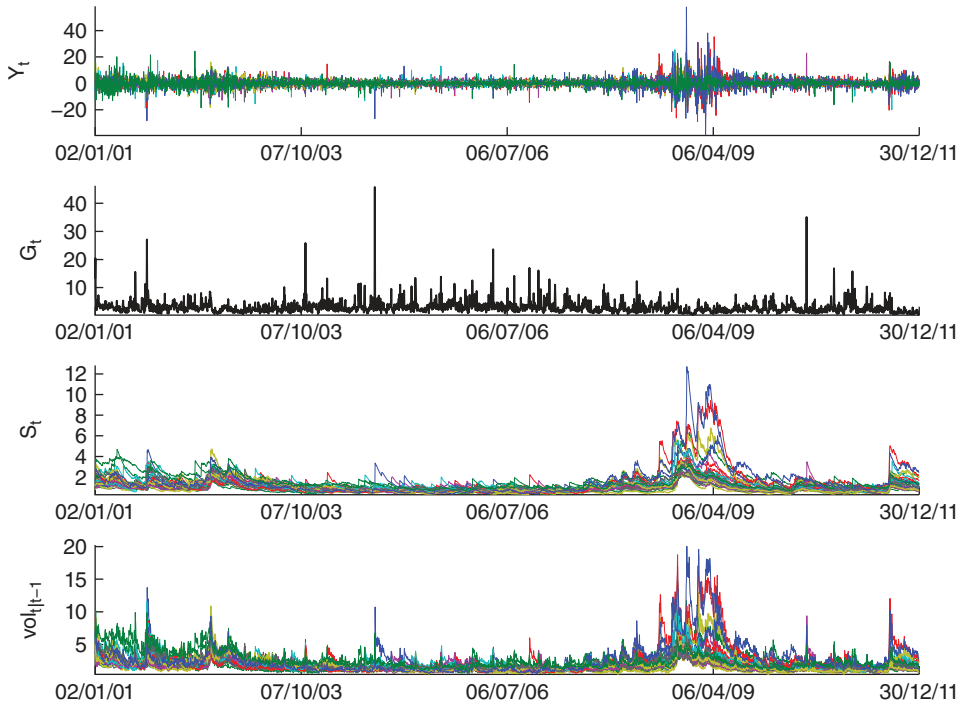

Figure 11.4 Similar to Figure 11.3, but overlaying the results for all 30 series associated with the DJIA.

2) It is worth contrasting the aforementioned *ad hoc* method of using the Gaussian CCC or DCC residuals to fit an i.i.d. multivariate Student's t distribution, with the COMFORT class of models: The former estimates the univariate series as Gaussian-GARCH, and then fits the degrees of freedom parameter of a multivariate Student's t, while COMFORT also fits univariate Gaussian-GARCH models to each margin, but *conditional* on the filtered $\{G_t\}$ sequence, in an iterative fashion, via the EM algorithm. This implies that there is feedback during estimation between the filtered $\{G_t\}$ sequence and the conditional Gaussian GARCH parameters. There is obviously no such feedback in the *ad hoc* method.
Somewhat fascinatingly, by disentangling these two effects, the estimated, conditionally Gaussian univariate GARCH processes from (11.23) yield essentially the *same parameters* across assets, i.e., the $\hat{\omega}_{0,i}$, when freely estimated for each of d assets, $i = 1, \ldots, d$, are virtually the same (and this having been confirmed by using numerous starting values in the estimation, and conducted also on numerous data sets), and similarly for $\hat{\alpha}_{1,i}$ and $\hat{\beta}_{1,i}$. This is far from the case in the Gaussian CCC or DCC setting and, thus, in the *ad hoc* Student's t DCC model. In light of the COMFORT model, this variation in GARCH parameters can be interpreted as the result of model mis-specification: The GARCH(1,1) structure is inadequate for capturing all the features of the individual series, as was discussed earlier, in Section 10.2.2, with the GARCH parameters moving towards the IGARCH border as the sample size increases.
This implies that, conditional on the latent sequence describing the common market factor, persistence in volatility, as captured by a GARCH(1,1) process, *is the same across all assets*. While of

great theoretical interest, this feature has the useful practical implication that the estimation of the univariate GARCH models (11.23) can be foregone (this being the most time-consuming part of the estimation process), replaced either by one joint estimation, or just fixing the three GARCH parameters to values that consistently arise for various sets of daily stock returns data. We term this Fast ReducEd Estimation, or FREE-COMFORT.

3) A third benefit of the COMFORT model class is that an extension to a pseudo type of **stochastic volatility** (SV) model is possible. Recalling the brief discussion at the beginning of Section 10.2.1, SV models are considered more realistic, as the volatility at time t is not simply a deterministic function of information up to time $t-1$. The price to pay for allowing a further source of randomness to enter into the volatility at time t is intractability of the likelihood and the requirement of more sophisticated methods for parameter estimation. In the COMFORT context, it was found that there is some predictability in the (otherwise i.i.d.) univariate latent sequence $\{G_t\}$, and a model extension was proposed that has some elements of an SV model, but such that the likelihood is still accessible, allowing for straightforward parameter estimation; see Paolella and Polak (2015b) for details on the formal connection to SV models, along with an application to multivariate option pricing. A different model in the univariate case that builds on the classic GARCH structure but allows for elements of an SV model and is such that it still is amenable to traditional likelihood optimization is proposed and studied in Smetanina (2017).
4) The incorporation of the univariate $\{G_t\}$ sequence allows one to move from the Gaussian CCC-GARCH model to the non-Gaussian COMFORT model, and its ensuing enhanced ability for risk assessment and portfolio allocation. The former can be thought of as the COMFORT model with constant G_t for all time periods. However, this single $\{G_t\}$ sequence is latent to all d (say) stocks, and that may not be realistic. One might argue that each industry sector should be endowed with its own such sequence. Such a construction no longer enjoys the convenient distribution theory associated with the MGHyp, and simulation from the predictive distribution would be required. Models that make use of this idea and reap benefits in terms of forecasting and portfolio construction ability are pursued in Näf et al. (2018b,a), with an introduction to the latter given in Section 12.6.

11.2.5 Copula Constructions

The development and use of copula-based models for various phenomena in finance continues to grow unabated. Part of the reason for their appeal is that they allow the (possibly time-varying, such as via GARCH) univariate series to be modeled separately, and independently of the copula structure that links them into a multivariate distribution.

Basic knowledge of copula theory, and some experience with the methodology, is now considered essential for financial econometricians. Informally, a copula is a multivariate distribution such that the univariate margins are Unif (0, 1), and fully describes the dependence among the margins. The copula (as its name suggests, for those skilled in Latin) can be viewed as a structure to tie, or join, a set of univariate marginal distributions. Their use and applicability have grown in various fields, particularly quantitative risk management and empirical finance. Detailed accounts can be found in Bradley and Taqqu (2003), Nelsen (2006), McNeil et al. (2015), Joe (2015), and Ibragimov and Prokhorov (2017), while Fermanian (2017) provides (at the time of writing) a recent overview of their use in financial econometrics. Surveys of the use of copula-based models for financial time series and volatility models are given by Patton (2009), Genest et al. (2009), and Heinen and Valdesogo (2012).

We will detail the use of one particular, and very straightforward, copula construction for modeling and predicting asset returns in Chapter 12.

11.3 Introducing Portfolio Optimization

11.3.1 Some Trivial Accounting

Assume a universe of financial assets, such as currencies, commodities, or stocks, that can be traded (i.e., have the required liquidity) at the desired frequency, such as monthly, weekly, daily, intra-day, etc. For illustration, we assume daily stock trading on companies $1, \dots, d$. Further, assume there is an investment amount of capital at time t, say C_t. We want to detail the evolution of the wealth through time.

It is useful to first do so without incorporating transaction costs; these are dealt with below. Further, we assume no short-selling, which is typical in many contexts and mandatory in others. Let

$$\mathbf{w}_t = (w_{1,t}, \dots, w_{d,t})' \in [0,1]^d, \quad \sum_{k=1}^{d} w_{k,t} = 1, \tag{11.24}$$

denote the non-negative weight vector summing to one that describes a portfolio for the d stocks at time t. Observe that this is another assumption, namely that we attempt to invest all the money available. More realistic settings could allow for a "risk fear strategy" such that, based on some calculated signal at time t, the investor wishes to exit the market (anticipating perhaps very high risk or a market downturn), and thus allowing for $\sum_{k=1}^{d} w_{k,t} = 0$.

Denote the price of stock k at time t as $P_{k,t} > 0$. In what follows, we assume time is measured in some discrete unit such as hours, days, weeks, etc. For our purposes, assume daily, and when we speak of time t, one can fix this to mean, say, at 4:00PM on day t.

Based on some investment strategy (such as use of an assumed stochastic process with parameters fit from a window of past data up to time t and a prediction for time $t+1$), we decide to hold the portfolio with weights $\mathbf{w}_t$, such that $C_t w_{k,t}$ is invested in the kth asset, $k = 1, \dots, d$. This will often not be possible because of discreteness, and thus entails buying

$$\alpha_{k,t} := \left\lfloor \frac{C_t w_{k,t}}{P_{k,t}} \right\rfloor$$

stocks of company k, where $\lfloor \cdot \rfloor$ denotes the floor function, i.e., $\lfloor 2.8 \rfloor = \lfloor 2.1 \rfloor = 2$. Thus, the amount of money invested in stock for company k is $\alpha_{k,t} P_{k,t}$, which has an upper bound of $C_t w_{k,t}$ and will be close to this bound when $C_t / P_{k,t}$ is large. We denote the total amount invested (wealth) as $\mathcal{W}$; in particular, at the beginning of the investment process, this is $\mathcal{W}_1 := \mathcal{W}_{1|1}$, where we define

$$\mathcal{W}_{s|t} := \sum_{k=1}^{d} \alpha_{k,t} P_{k,s},$$

i.e., $\mathcal{W}_{s|t}$ is the *portfolio wealth* at time s, given the prices at time s and the number of shares at time t. Observe that, as the weights sum to one,

$$\mathcal{W}_{t|t} = \sum_{k=1}^{d} \alpha_{k,t} P_{k,t} \approx \sum_{k=1}^{d} C_t w_{k,t} = C_t,$$

with C_t being the upper bound on the amount invested, $\mathcal{W}_{t|t} \leqslant C_t$. When $\mathcal{W}_{t|t} = C_t$, we will call this the *full investment approximation*.

At time $t+1$, we know the prices $P_{k,t+1}$, $k = 1, \ldots, d$, and the portfolio is worth

$$\mathcal{W}_{t+1|t} := \sum_{k=1}^{d} \alpha_{k,t} P_{k,t+1} = \sum_{k=1}^{d} \alpha_{k,t} \frac{P_{k,t}}{P_{k,t}} P_{k,t+1} \approx C_t \sum_{k=1}^{d} w_{k,t} \frac{P_{k,t+1}}{P_{k,t}}. \tag{11.25}$$

From the last expression in (11.25), it is clear that we wish to have chosen a zero weight in stocks such that the price change from time t to $t+1$ is negative, and ideally a weight of one in the stock whose relative price increase is the largest.

Remark Observe that (11.25) is not valid for negative weights. This is because a negative weight implies short selling, which means the stocks are borrowed at time t (for a fee, just like you pay interest when you borrow money from a bank), immediately sold, and purchased at a future date (for which you hope the price has gone down) to return to the lender. Expression (11.25) could be augmented to support short selling by taking

$$\mathcal{W}_{t+1|t} \approx C_t \sum_{k=1}^{d} |w_{k,t}| \left(\frac{P_{k,t+1}}{P_{k,t}} \right)^{\operatorname{sgn}(w_{k,t})},$$

though this will not be necessary to compute the returns below in (11.28) because there the relative price *difference* is used. ■

The percentage return of the portfolio at time $t+1$ based on starting time t, denoted $R_{P,t+1|t}$, is given by, with the full investment approximation,

$$R_{P,t+1|t} := 100 \frac{\mathcal{W}_{t+1|t} - \mathcal{W}_{t|t}}{\mathcal{W}_{t|t}} \approx 100 \frac{C_t \sum_{k=1}^{d} w_{k,t} \frac{P_{k,t+1}}{P_{k,t}} - C_t}{C_t} \tag{11.26}$$

$$= 100 \left(\sum_{k=1}^{d} w_{k,t} \frac{P_{k,t+1}}{P_{k,t}} - \frac{1}{d} \sum_{k=1}^{d} \frac{P_{k,t}}{P_{k,t}} \right) = 100 \sum_{k=1}^{d} \frac{1}{P_{k,t}} \left(w_{k,t} P_{k,t+1} - \frac{P_{k,t}}{d} \right). \tag{11.27}$$

Observe in (11.27) how, if the weights $w_{k,t}$ were chosen to be equal, i.e., the equally weighted portfolio, then (11.27) reduces to (with the full investment approximation)

$$\text{(equal weights) } R_{P,t+1|t} \approx \frac{1}{d} \sum_{k=1}^{d} 100 \left(\frac{P_{k,t+1} - P_{k,t}}{P_{k,t}} \right) = \frac{1}{d} \sum_{k=1}^{d} R_{k,t+1|t},$$

where $R_{k,t+1|t} := 100(P_{k,t+1} - P_{k,t})/P_{k,t}$ is the (simple) percentage return on asset k at time $t+1$ calculated with respect to its price at time t. As the portfolio weights sum to one, we can also write (11.26) as

$$R_{P,t+1|t} \approx 100 \left(\sum_{k=1}^{d} w_{k,t} \frac{P_{k,t+1}}{P_{k,t}} - \sum_{k=1}^{d} w_{k,t} \right) = 100 \sum_{k=1}^{d} w_{k,t} \frac{P_{k,t+1} - P_{k,t}}{P_{k,t}}$$

$$= \sum_{k=1}^{d} w_{k,t} R_{k,t+1|t} = \mathbf{w}_t' \mathbf{R}_{t+1|t}, \tag{11.28}$$

where $\mathbf{R}_{t+1|t} := (R_{1,t+1|t}, \ldots, R_{d,t+1|t})'$, generalizing the equally weighted case.

Now consider the multi-step returns. We first illustrate the returns with the full investment approximation. With the new price information at time $t+1$, the weights are updated to vector $\mathbf{w}_{t+1}$, as calculated by the investment method, and the wealth that can be invested is $\mathcal{W}_{t+1|t} = C_{t+1}$. We thus buy and sell shares of the d assets such that we have $\alpha_{k,t+1} = \lfloor C_{t+1} w_{k,t+1} / P_{k,t+1} \rfloor$ shares in company k, which, under the full investment approximation, implies a wealth in company k of $\alpha_{k,t+1} P_{k,t+1} = C_{t+1} w_{k,t+1}$. At time $t+2$, the prices $P_{k,t+2}$ are realized,

$$\mathcal{W}_{t+2|t+1} = \sum_{k=1}^{d} \alpha_{k,t+1} P_{k,t+2} = \sum_{k=1}^{d} \alpha_{k,t+1} \frac{P_{k,t+1}}{P_{k,t+1}} P_{k,t+2} \approx C_{t+1} \sum_{k=1}^{d} w_{k,t+1} \frac{P_{k,t+2}}{P_{k,t+1}},$$

and

$$R_{P,t+2|t} = 100 \frac{\mathcal{W}_{t+2|t+1} - \mathcal{W}_{t|t}}{\mathcal{W}_{t|t}}. \tag{11.29}$$

Consider now the return for h periods ahead, $h \geqslant 1$, i.e., based on starting time t, we want the return at time $t+h$. For the initial purchase at time t, it is algebraically convenient to let $\mathcal{W}_{t|t-1} := \mathcal{W}_{t|t}$. From (11.29) for time $t+h$,

$$\begin{aligned} \frac{1}{100} R_{P,t+h|t} + 1 = \frac{\mathcal{W}_{t+h|t+h-1}}{\mathcal{W}_{t|t}} &= \prod_{\tau=t+1}^{t+h} \left(\frac{\mathcal{W}_{\tau|\tau-1}}{\mathcal{W}_{\tau-1|\tau-2}} \right) \\ &= \prod_{\tau=t+1}^{t+h} \left(1 + \frac{\mathcal{W}_{\tau|\tau-1} - \mathcal{W}_{\tau-1|\tau-2}}{\mathcal{W}_{\tau-1|\tau-2}} \right) = \prod_{\tau=t+1}^{t+h} \left(1 + \frac{1}{100} R_{P,\tau|\tau-1} \right), \end{aligned}$$

or

$$R_{P,t+h|t} = 100 \left(\prod_{\tau=t+1}^{t+h} \left(1 + \frac{1}{100} R_{P,\tau|\tau-1} \right) - 1 \right). \tag{11.30}$$

Recall the Taylor series $\log(1+x) = \sum_{i=1}^{\infty} (-1)^{i+1} x^i / i$ for $|x| < 1$, with first-order approximation $\log(1+x) \approx x$. Thus, return $R_{P,t+2|t}$ satisfies $R_{P,t+2|t} \approx 100 \log(\mathcal{W}_{t+2|t+1} / \mathcal{W}_{t|t})$ (see, e.g., page I.152), and

$$\begin{aligned} R_{P,t+2|t} &\approx 100 \log \left(\frac{\mathcal{W}_{t+2|t+1}}{\mathcal{W}_{t|t}} \right) \\ &= 100 \log \left(\frac{\mathcal{W}_{t+2|t+1}}{\mathcal{W}_{t+1|t}} \frac{\mathcal{W}_{t+1|t}}{\mathcal{W}_{t|t}} \right) = 100 \log \left(\frac{\mathcal{W}_{t+2|t+1}}{\mathcal{W}_{t+1|t}} \right) + 100 \log \left(\frac{\mathcal{W}_{t+1|t}}{\mathcal{W}_{t|t}} \right) \\ &= R_{P,t+2|t+1} + R_{P,t+1|t} = \mathbf{w}'_{t+1} \mathbf{R}_{t+2|t+1} + \mathbf{w}'_t \mathbf{R}_{t+1|t}. \end{aligned}$$

In general, at time $t+h$, using both the log and full investment approximations,

$$R_{P,t+h|t} \approx \sum_{i=1}^{h} \mathbf{w}'_{t+i-1} \mathbf{R}_{t+i|t+i-1}. \tag{11.31}$$

This is the commonly used approximation for illustrating and comparing the performance of investment methods. It is conservative compared to (11.30) because $\log(1+x) \leqslant x$ for $|x| < 1$ (see, e.g., Lang, 1997, p. 87). The difference between (11.30) and (11.31) can also be exemplified as follows. For $h = 3$, (11.30) is

$$\frac{1}{100} R_{P,t+3|t} + 1 = \left(1 + \frac{1}{100} R_{P,t+1|t} \right) \left(1 + \frac{1}{100} R_{P,t+2|t+1} \right) \left(1 + \frac{1}{100} R_{P,t+3|t+2} \right)$$

or, multiplying out,

$$R_{P,t+3|t} = R_{P,t+1|t} + R_{P,t+2|t+1} + R_{P,t+3|t+2} + \frac{1}{100}(R_{P,t+1|t}R_{P,t+2|t+1} + R_{P,t+1|t}R_{P,t+3|t+2} + R_{P,t+2|t+1}R_{P,t+3|t+2}) + \frac{1}{100^2}R_{P,t+1|t}R_{P,t+2|t+1}R_{P,t+3|t+2}. \tag{11.32}$$

Thus, (11.31) ignores the higher-order terms in (11.32), which are clearly very small, but become visible over long investment horizons.

Now consider relaxing the full investment approximation. Assume the investor starts with capital C_t and invests in portfolio $\sum_{k=1}^{d} \alpha_{k,t}P_{k,t} \leqslant C_t$, and let the "savings" be that amount that cannot be invested because of the discreteness, i.e.,

$$S_t := C_t - \sum_{k=1}^{d} \alpha_{k,t}P_{k,t}.$$

At time $t+1$, the portfolio is worth $\mathcal{W}_{t+1|t} = \sum_{k=1}^{d} \alpha_{k,t}P_{k,t+1}$ and imagine the investor sells everything, obtaining $C_{t+1} = S_t + \mathcal{W}_{t+1|t}$. (We also assume the interest on S_t from time period t to $t+1$ is zero, which currently is not so unrealistic, though is easily accommodated.) She then purchases the portfolio with $\alpha_{k,t+1}$ shares from company k, at price $P_{k,t+1}$, $k = 1, \ldots, d$, where $\alpha_{k,t+1} := \lfloor C_{t+1}w_{k,t+1}/P_{k,t+1} \rfloor$. Naturally, in practice, one does not sell everything and then repurchase the new portfolio because of transaction costs and the fact that there will be overlap between the two portfolios. Instead, one buys or sells the shares of company k to adjust $\alpha_{k,t}$ to $\alpha_{k,t+1}$, $k = 1, \ldots, d$. Without transaction costs and assuming a zero bid-ask spread, this is equivalent to imagining selling everything and then purchasing the updated portfolio anew.

It follows that

$$S_{t+1} = C_{t+1} - \sum_{k=1}^{d} \alpha_{k,t+1}P_{k,t+1}.$$

At time $t+2$, the portfolio is worth $\mathcal{W}_{t+2|t+1} = \sum_{k=1}^{d} \alpha_{k,t+1}P_{k,t+2}$, and selling gives

$$\begin{aligned} C_{t+2} &= S_{t+1} + \mathcal{W}_{t+2|t+1} = C_{t+1} - \sum_{k=1}^{d} \alpha_{k,t+1}P_{k,t+1} + \sum_{k=1}^{d} \alpha_{k,t+1}P_{k,t+2} \\ &= C_{t+1} + \sum_{k=1}^{d} \alpha_{k,t+1}(P_{k,t+2} - P_{k,t+1}) = S_t + \sum_{k=1}^{d} \alpha_{k,t}P_{k,t+1} + \sum_{k=1}^{d} \alpha_{k,t+1}(P_{k,t+2} - P_{k,t+1}) \\ &= C_t + \sum_{k=1}^{d} \alpha_{k,t}(P_{k,t+1} - P_{k,t}) + \sum_{k=1}^{d} \alpha_{k,t+1}(P_{k,t+2} - P_{k,t+1}). \end{aligned}$$

Continuing, she purchases $\alpha_{k,t+2}$ shares from company k, at price $P_{k,t+2}$, $k = 1, \ldots, d$, and $S_{t+2} = C_{t+2} - \sum_{k=1}^{d} \alpha_{k,t+2}P_{k,t+2}$. At time $t+3$, the portfolio is worth

$$\mathcal{W}_{t+3|t+2} = \sum_{k=1}^{d} \alpha_{k,t+2}P_{k,t+3}, \quad \text{and selling gives} \quad C_{t+3} = S_{t+2} + \mathcal{W}_{t+3|t+2},$$

which reduces to

$$C_t + \sum_{k=1}^{d} \alpha_{k,t}(P_{k,t+1} - P_{k,t}) + \sum_{k=1}^{d} \alpha_{k,t+1}(P_{k,t+2} - P_{k,t+1}) + \sum_{k=1}^{d} \alpha_{k,t+2}(P_{k,t+3} - P_{k,t+2}).$$

The pattern should now be clear, so that, at period $t + h$,

$$C_{t+h} = C_t + \sum_{\tau=t}^{t+h-1} \sum_{k=1}^{d} \alpha_{k,\tau}(P_{k,\tau+1} - P_{k,\tau}), \tag{11.33}$$

and the percentage return is

$$R_{P,t+h|t} = 100\frac{C_{t+h} - C_t}{C_t}, \quad \text{or} \quad C_{t+h} = C_t\left(1 + \frac{R_{P,t+h|t}}{100}\right). \tag{11.34}$$

We now address how to account for transaction costs. To do so, we could adjust C_t at each t, but it is easier to imagine having a separate account (without interest) to pay the costs, and these costs are removed from the return calculated at time $t + i$, $i = 1, \dots, h$. To this end, let T_1 be the initial startup cost for buying the $\sum_{k=1}^{d} \alpha_{k,1}$ shares. This can be seen as imagining the existing portfolio at time $t = 0$ to consist of $\alpha_{k,0} = 0$, $k = 1, \dots, d$, and thus we take $T_{1|0} := T_1$. Let $T_{t+1|t}$ be the total induced transaction cost for buying or selling $\alpha_{k,t+1} - \alpha_{k,t}$ shares on company k, $k = 1, \dots, d$, at the price at time $t + 1$, and similarly for $T_{t+2|t+1}, \dots, T_{t+h-1|t+h-2}$.

A typical approximation uses so-called **proportional transaction costs**, and does not account for the bid-ask spread, i.e., the different buying (ask) and selling (bid) prices. This method will be subsequently discussed. Assuming the investment procedure stops at time $t + h$, all shares are sold at time $t + h$, at cost T_{t+h}, and we define $T_{t+h|t+h-1} := T_{t+h}$. The *gross* percentage return, i.e., before paying transaction costs, is (11.34), which we now denote as

$$R^G_{P,t+h|t} = 100\frac{C_{t+h} - C_t}{C_t}. \tag{11.35}$$

The *net* percentage return, meaning after transaction costs, is then

$$R^N_{P,t+h|t} = 100\frac{C_{t+h} - \sum_{i=0}^{h} T_{t+i|t+i-1} - C_t}{C_t}. \tag{11.36}$$

The *proportional transaction cost* approximation, as in DeMiguel et al. (2013), assumes that transaction costs lead to a deduction of the total portfolio return proportional to the amount of portfolio rebalancing. It is defined as

$$(100 + R^N_{P,t+h|t}) = \left(1 - \kappa \sum_{i=0}^{h} \sum_{k=1}^{d} |w_{k,t+1+i} - w_{k,(t+i)^+}|\right)(100 + R^G_{P,t+h|t}), \tag{11.37}$$

where

1) $w_{k,t}$ from (11.24) is the portfolio weight of asset k, computed at time t, held from time t to $t + 1$.
2) w_{k,t^+} is the proportion of wealth that is held in asset k at time $t + 1$ just before rebalancing the portfolio, i.e.,

$$w_{k,t^+} = \frac{\alpha_{k,t}P_{k,t+1}}{\sum_{k=1}^{d} \alpha_{k,t}P_{k,t+1}}. \tag{11.38}$$

3) $\kappa > 0$ quantifies the level of proportional transaction cost, with values of 0.005 and 0.010 (five and ten basis points, respectively) being typical.

Observe how (11.37) and (11.38) account for the change in the proportion of wealth invested in asset k due to a change in asset prices from time t to $t+1$.

An important aspect of this method is how the equally weighted portfolio is treated. This is characterized by $w_{k,t} = 1/d$ for all assets $k = 1, \ldots, d$ and all time periods $t = 1, 2, \ldots$. When the relative prices of assets change from time t to $t+1$, $w_{k,t^+} \neq 1/d$, and the portfolio needs to be rebalanced (and incurs transaction costs).

Observe that (11.37) utilizes the total returns, as opposed to the log returns (11.31), to guarantee proportionality of the transaction costs to the portfolio value. To see this, rewrite (11.37) using $C_{t+h} = C_t(1 + R^N_{P,t+h|t}/100)$ from the right-hand side of (11.34) to get

$$C_{t+h} = C_t \left(1 - \kappa \sum_{i=0}^{h} \sum_{k=1}^{d} |w_{k,t+1+i} - w_{k,(t+i)^+}|\right)\left(1 + \frac{R^G_{P,t+h|t}}{100}\right). \tag{11.39}$$

For $h = 1$, (11.39) appears in, among others, DeMiguel et al. (2009b). From this, the total transaction cost amount can be expressed as a fraction of the final wealth, proportional to the amount of rebalancing, i.e.,

$$C_{t+h}\kappa \sum_{i=0}^{h} \sum_{k=1}^{d} |w_{k,t+1+i} - w_{k,(t+i)^+}| = \sum_{i=0}^{h} T_{t+i|t+i-1}. \tag{11.40}$$

This approximation is implemented in the program in Listing 11.1.

```
function [pandl_net,pandl_gross] = netreturns(wmat,rmat,pmat,kap)
% Computes the portfolio returns net of transactions costs as
%   r_t =( 1 - kap sum_j=1^N |w_j,t - w_j,(t-1)+| ) * ( w_t .* r_t) where
%   w_j,(t-1)+ is the portfolio weight in asset j at time t before rebalancing;
%   w_j,t is the desired portfolio weight at time t after rebalancing;
%   kap is the proportional transaction cost;
%   w_t is the vector of portfolio weights; and
%   r_t is the vector of returns.
%   p_t is the vector of prices
pandl_gross = sum( wmat .* rmat ,2 ); pandl_net = zeros(size(pandl_gross));
pandl_net(1) = pandl_gross(1);
wmatplus =  zeros(size(wmat));
alpha = zeros(size(wmat));
for t=2:length(pandl_gross)
    alpha(t-1,:) = wmat(t-1,:)./ pmat(t-1,:);
    % without loss of generality the total wealth invested is 1
    wmatplus(t-1,:) = (alpha(t-1,:).*pmat(t,:)) ./ (sum(alpha(t-1,:).*pmat(t,:),2));
    pandl_net(t) =  ( ( 1 -  kap * sum( abs( wmat(t,:) - wmatplus(t-1,:) ) ,2 ) ) ...
        * ( 100 + pandl_gross(t) ) ) - 100;
end
```

Program Listing 11.1: Computes the returns net of transaction costs.

A further approximation involves use of only the returns on each asset. This is convenient and will often be enough. The implementation is given in Listing 11.2.

```
function [pandl_net,pandl_gross] = netreturns(wmat,rmat,kap)
pandl_gross = sum( wmat .* rmat ,2 );
pandl_net=zeros(size(pandl_gross));
pandl_net(1)=pandl_gross(1);
for t=2:length(pandl_gross)
  pandl_net(t) = ( ( ( 1 -  kap * sum( abs( wmat(t,:) - wmat(t-1,:) ) ,2 ) ) ...
   * (100 + pandl_gross(t))) - 100 );
end
```

Program Listing 11.2: Further approximation of accounting for transaction costs, requiring only the returns.

Remarks

a) To help reduce transaction costs without an appreciable effect on performance, one approach is to impose some form of constraint on the rebalancing of the portfolio weights; see DeMiguel et al. (2009a, 2014), Fan et al. (2012), Fastrich et al. (2015), and the references therein.
Another method is to "tame" the evolution of the covariance matrix, allowing for some dynamic variation, but not as much as induced by use of traditional multivariate GARCH models. One way of accomplishing this is by using principle components analysis (PCA), as investigated in Paolella et al. (2018b) (in a non-Gaussian context).
The use of PCA in this context is not new, with the seminal works being Ding (1994) and Alexander and Chibumba (1996), with subsequent embellishments by Alexander (2001, 2002, 2008). The idea is conceptually very simple: The spectral decomposition of the covariance matrix is computed and, instead of using univariate GARCH processes for all margin processes, only a small number of leading principle components of the covariance matrix are endowed with a GARCH structure (and the remaining eigenvalues are set to zero). Finally, the reader should know that the general methodology of PCA goes back to Pearson (1901) and Hotelling (1933); a good textbook starting point is Jolliffe (2002), though PCA now gains even more popularity via its applicability in machine learning, and is discussed in many such textbooks.
Other methods include shrinking the ex-post portfolio weights towards a constant target portfolio, as demonstrated in Suh (2016), and use of averaging of covariance matrix forecasts over subsequent rolling windows to stabilize portfolio weights and thus reduce transaction costs; see, e.g., Hautsch et al. (2015).

b) We are still not done—we have not accounted for paid dividends on the stocks. These can only increase returns (even after adjusting for the fact that dividends might be taxed as income), so a conservative measure of returns can ignore them. Often, one uses returns that are (split and) *dividend adjusted*, such that the dividend is added to the return, and one can proceed as above, with the returns automatically adjusted for dividends. In the case of nonzero-coupon bonds, one would need to incorporate coupon payments. ■

11.3.2 Markowitz and DCC

Consider a set of d financial assets, such as highly liquid stocks on major exchanges, for which returns are observed over a specified period of time and frequency (e.g., daily, ignoring the weekend effect for stocks), and assume (as is common in numerous real contexts) that short-selling will not be used. The set of valid portfolio weights is thus

$$\mathcal{A} = \{\mathbf{a} \in [0,1]^d \,:\, \mathbf{1}_d' \,\mathbf{a} = 1\}. \tag{11.41}$$

In the classic portfolio optimization framework going back to the seminal work of Markowitz (1952), the returns are assumed to be an i.i.d. multivariate normal process with mean $\boldsymbol{\mu}$ and variance $\boldsymbol{\Sigma}$. One wishes to determine the portfolio weight vector, say $\mathbf{a}^*$, that yields the lowest variance of the predictive portfolio percentage return at time $t+1$, given information up to time t, say $P_{t+1|t}$, conditional on its expected value being greater than some positive threshold τ_{daily}. That is,

$$\mathbf{a}^* = \arg\min_{\mathbf{a}\in\mathcal{A}} \mathbb{V}(P_{t+1|t,\mathbf{a}}) \quad \text{such that} \quad \mathbb{E}[P_{t+1|t,\mathbf{a}}] \geqslant \tau_{\text{daily}}, \tag{11.42}$$

where $\mathcal{A}$ is given in (11.41), and, with discrete compounding,

$$\tau_{\text{daily}} = 100\left(\left(1+\frac{\tau}{100}\right)^{1/250}-1\right), \quad \tau = 100\left(\left(1+\frac{\tau_{\text{daily}}}{100}\right)^{250}-1\right), \tag{11.43}$$

for $\tau = \tau_{\text{annual}}$ the desired annual percentage return, here calculated assuming 250 business days per year. In this i.i.d. Gaussian Markowitz setting, note that

$$\widehat{\mathbb{E}}[P_{t+1|t,\mathbf{a}}] = \mathbf{a}'\widehat{\boldsymbol{\mu}}, \qquad \widehat{\mathbb{V}}(P_{t+1|t,\mathbf{a}}) = \mathbf{a}'\widehat{\boldsymbol{\Sigma}}\mathbf{a}, \tag{11.44}$$

where $\widehat{\boldsymbol{\mu}}$ and $\widehat{\boldsymbol{\Sigma}}$ refer to the usual plug-in unbiased estimators of the Gaussian distribution parameters. When short selling is allowed in the i.i.d. Markowitz framework, there is a well-known closed-form solution to (11.42); see, e.g., Ruppert (2004, Sec. 5.5) for a clear exposition and also Matlab codes for calculating and plotting the ubiquitous efficiency frontier and calculation of the tangency portfolio. For the long-only case, numerical methods are required to determine it. This is very easy to set up in Matlab using their constrained optimization program `fmincon`, with bare-bones code given in Listing 11.3 for this i.i.d. case, as well as with using the predicted variance-covariance matrix from the DCC model.

```
function w = PortMNS(data, tau, DCC)
if nargin<3, DCC=0; end
if DCC, [mu,Sigma] = DCC1step(data); else mu = mean(data); Sigma = cov(data); end
DEDR=100*((tau/100 + 1)^(1/250)-1); feas=max(mu) <= DEDR;
if feas, w=meanvar(mu,Sigma,DEDR)'; else w=zeros(length(mu),1); end

function w = meanvar(mu, Sigma, tau)
opt=optimset('Algorithm','active-set','LargeScale', 'off','Display','off');
d=length(mu); A = -mu; B = -tau; LB = zeros(1,d); UB = ones(1,d); w0=UB/d;
Aeq = ones(1,d); Beq = 1; % sum(w) = 1
w = fmincon(@fv, w0, A, B,  Aeq, Beq, LB, UB, [], opt, Sigma);

function f = fv(w, Sigma), f = w * Sigma * w';
```

Program Listing 11.3: MNS stands for Markowitz No Short (selling). Delivers the long-only mean-variance pure Markowitz (i.i.d. model with plug-in estimators for mean and variance–covariance) portfolio weight vector or the long-only mean-variance portfolio weight vector based on the DCC density prediction of the mean and covariance matrix. It is based on a set of returns passed as `data` and for a given desired expected yearly return τ. Function `DCC1step` is from the author, and computes the predictive distribution (here, Gaussian, and thus characterized by the mean vector and covariance matrix) corresponding to the fitted DCC model of Section 11.2.2, having used the profile likelihood method of univariate Gaussian GARCH estimation discussed at the end of Section 10.2.

Remarks

a) In the more general elliptic setting, which nests the Gaussian distribution and can allow for heavy tails such as via a multivariate t distribution, (11.42) is still valid, provided that second moments exist.

b) The success of the method depends crucially on the estimates $\widehat{\boldsymbol{\mu}}$ and $\widehat{\boldsymbol{\Sigma}}$. For a particular length of data, say T, the number of parameters, and thus the estimation error, grow with the number of assets d, and is such that, for typical T, even a modest choice of d will lead to highly unreliable estimates and, thus, poor performance. This was emphasized in DeMiguel et al. (2009b), who provide an example showing that, in order to outperform the **equally weighted portfolio** (equal weights for each asset; see Section 11.3.3 below) in the case of monthly updating with 25 assets, about 3000 months (250 years) of past historical returns are required. Not only is that completely unrealistic, but this is all the more problematic if the data generating process of the returns is changing over time.

   Many studies have shown the deleterious effect of estimation error, and developed suggestions for mitigating the problem; see, e.g., Frankfurter et al. (1971), Kalymon (1971), Klein and Bawa (1976), Frost and Savarino (1986), Britten-Jones (1999), and Kolm et al. (2014), as well as Brandt (2010) for an overview. Shrinkage estimation is a key methodology in this pursuit. See, e.g., Jorion (1986), Ledoit and Wolf (2004, 2012), and Kan and Zhou (2007) for model parameter shrinkage, and DeMiguel et al. (2009a,b), Brown et al. (2013), and the references therein for portfolio weight shrinkage.

   Chen and Yuan (2016) propose to restrict the portfolio weight vector to a subspace determined by using only a subset of the spectral decomposition of the estimated covariance matrix. The idea is to offset the loss of efficiency from restricting the investment set by reduced estimation error.

   Another approach for determining the portfolio weights that avoids the pitfalls inherent in the estimation of the large number of parameters associated with first and second moments (let alone possible use of higher-order moments) is by Brandt et al. (2009), and briefly discussed in a remark below in Section 11.3.4. There, another method for avoiding the "curse of dimensionality" suited for daily (or higher frequency) data, called the univariate collapsing method, is presented.

c) One would think that use of a multivariate GARCH-type model such as CCC or DCC should result in superior portfolio performance at the daily level, given the blatant non-i.i.d. nature of the data. This is true if investment and portfolio updating can take place without transaction costs (see Section 11.3.1). As reality dictates, transaction costs are significant, particularly for individual investors, but also for financial institutions. When using GARCH-type models, the **turnover**, i.e., the sum of the absolute changes of the portfolio weights induced when re-balancing, tends to be vastly higher compared to use of an i.i.d. model. This is because of the far greater changes from period to period of the estimated covariance matrix. As such, it is often the case that an i.i.d. model can outperform the use of DCC with even modest transaction costs. ■

In the non-elliptic setting (elliptical distributions, and examples of non-elliptic ones being discussed in Section C.2), given the asymmetry of the portfolio distribution, the variance as the risk measure is not as desirable. Instead, left-tail risk measures such as value-at-risk (VaR) and expected shortfall (ES) are commonly used, and indeed lead to different allocations than with use of the variance; see, e.g., Embrechts et al. (2002) and Campbell and Kräussl (2007). In this case, the portfolio optimization problem can be expressed as follows: We want the portfolio weight vector $\mathbf{a}^*$ that yields the least expected shortfall (in magnitude—recall the ES will be negative, so, formally, we want the largest ES) of

the predictive portfolio return random variable $P_{t+1|t}$, conditional on its expected value being greater than τ_{daily}, i.e.,

$$\mathbf{a}^* = \arg\min_{\mathbf{a}\in\mathcal{A}} |\text{ES}(P_{t+1|t,\mathbf{a}}, \xi)| \quad \text{such that} \quad \mathbb{E}[P_{t+1|t,\mathbf{a}}] \geqslant \tau_{\text{daily}}, \tag{11.45}$$

where ξ is a pre-specified probability associated with the ES (for which we take 0.01). This will be used in the model discussed below in Section 11.3.4 and also in Chapter 12.

11.3.3 Portfolio Optimization Using Simulation

It is useful to think about what one can do if the function `fmincon`, as invoked, for example, by the code in Listing 11.3, or *any* black-box constrained optimizer, were not available. The first reason to entertain this idea is that, in more advanced model settings, particularly for large d, the numeric optimizer could encounter an inferior local minimum of (11.42) or (11.45), as well as exhibit other numeric problems, as discussed at length in Paolella (2014, Sec. 4.2). The second reason is that, when the objective function is not smooth in the parameters (as occurs when using a model such as the one discussed below in Section 11.3.4), `fmincon` will tend to fail, as it attempts to use gradient and Hessian information; see Paolella (2014) for a demonstration. In these cases, the evolutionary optimization algorithms discussed in Section III.4.4 would seem to be appropriate, though they are still subject to the possibility of returning an inferior local minimum, as well as having relatively slow convergence properties.[4]

The issue of avoiding potential inferior local minima (whether with use of gradient/Hessian-based methods or with evolutionary algorithms) can be mitigated by ensuring that the starting value passed to the optimization method is such that the associated objective function is "close enough" to the global minimum. This can be done with simulation, and is, in fact, a viable method in and of itself for locating a suitable portfolio vector, obviating any need for (traditional or evolutionary) optimization algorithms.

The method of simulation is extraordinarily simple: We randomly draw s portfolio weights (how to do so being subsequently discussed), where s will be a function of d and the desired accuracy (and possibly also depending on features of the data and the nature of the optimum; see below). Those draws that do not match all the required constraints (such as the mean constraint in (11.42), but potentially many others, as is typical in pension funds and financial institutional investors) are deleted. From the remaining, choose the portfolio vector that most closely satisfies (11.42) or (11.45). This vector could then be used as a starting value for the optimization methods, or, if s is high enough, the optimal portfolio vector (up to simulation discrepancy) is obtained, and use of optimization algorithms can be forgone.

Note that, as $s \to \infty$, the probability of locating $\mathbf{a}^*$, if it exists (its existence depending on the choice of τ), goes to one. Observe that, by the nature of simulation-based estimation with finite s, (11.42) or (11.45) will not be exactly obtained, but only approximated. We argue that this is not a drawback: All

4 Moreover, those heuristic optimization methods, as presented in Section III.4.4, were not designed to handle general constraints. If the only constraints are box constraints, i.e., fixed bounds on one or more parameters, then a straightforward transformation, as was done in Section III.4.3.2, can be employed. In our setting, we do have the simple fixed bound of $[0, 1]$ on each of the portfolio weights, but additionally we require that their sum is equal to one, and also that the constraint on the minimum expected return is met. The reader interested in the CMAES optimization algorithm is encouraged to explore how constraints can be embedded, possibly by appending the objective function with penalty terms to respect the desired constraints.

models are wrong w.p.1, are anyway subject to estimation error, and the portfolio delivered will depend on the chosen data set, in particular, how much past data to use and which assets to include, and, in the case of non-ellipticity, also depends on the choice of ξ (see, e.g., Rockafellar and Uryasev, 2000; Embrechts et al., 2002). As such, the method should be judged not on how well (11.42) or (11.45) can be evaluated *per se*, but rather on the out-of-sample portfolio performance, for a given model, given universe of assets, and conditional on all tuning parameters (such as τ and s).

The primary starting point for sampling portfolio weight vectors is to obtain values that are uniform on the simplex (11.41). This is achieved by taking $\mathbf{a} = (a_1, \ldots, a_d)'$ to be

$$\mathbf{a} = \mathbf{U}^{(\log)} / \mathbf{1}_d' \mathbf{U}^{(\log)}, \quad \mathbf{U}^{(\log)} = (\log U_1, \ldots, \log U_d)', \quad U_i \overset{\text{iid}}{\sim} \text{Unif}(0,1), \tag{11.46}$$

see, e.g., Devroye (1986). However, use of

$$\mathbf{a} = \mathbf{U}^{(q)} / \mathbf{1}_d' \mathbf{U}^{(q)}, \quad \mathbf{U}^{(q)} = (U_1^q, \ldots, U_d^q)', \quad U_i \overset{\text{iid}}{\sim} \text{Unif}(0,1), \tag{11.47}$$

is valuable for exploring other parts of the parameter space. In particular, the non-uniformity corresponding to $q = 1$ is such that there is a disproportionate number of values close to the **equally weighted portfolio**. As $q \to 0$, (11.47) collapses to equal weights. This is useful, given the well-studied ability of the seemingly naive equally weighted portfolio to outperform other allocation methods, as discussed in the subsequent remark.

Remark The equally weighted (or, commonly, "$1/N$") portfolio simply takes the portfolio weights to be equal. As the weights need to sum to one, the weight of each asset is, in our notation, $1/d$. This can be seen as an extreme form of shrinkage such that the choice of portfolio weights does not depend on the data itself, but only on the number of assets.

Studies of the high performance of the equally weighted portfolio relative to classic Markowitz allocation goes back at least to Bloomfield et al. (1977), with further analysis provided by DeMiguel et al. (2009b) and Brown et al. (2013). Fugazza et al. (2015) confirm that the "startling" performance by such a naive strategy indeed holds at the monthly level, but fails to extend to longer-term horizons, when asset return predictability is taken into account. This finding is thus very relevant for mutual- and pension-fund managers, and motivates the search for techniques to improve upon the $1/N$ strategy, particularly for short-term horizons such as monthly, weekly or daily.

See also Stivers (2018) for a possible explanation of why the $1/N$ portfolio can outperform traditional mean-variance approaches for asset allocation. ■

As $q \to \infty$ in (11.47), $\mathbf{a}$ will approach a vector of all zeroes, except for a one at the position corresponding to the largest U_i. Thus, large values of q can be used for exploring what we will refer to as **corner solutions**, or allocations such that only a small number of stocks have appreciable weight, and the remaining ones have weights close to or equal to zero. Figure 11.5 illustrates these sampling methods via scatterplots of a_1 versus a_2 for $d = 3$, using $s = 1{,}000$ points.

The sampling methods can be mixed: A data-driven heuristic is developed in Paolella (2017) for determining the proportions of portfolio vectors to be generated via uniform sampling (11.46) and from (11.47) for $q = 1$ and $q = 8$, resulting in a lower total number of replications required to reach an acceptable minimum of (11.42) or (11.45).

A natural way of checking the efficacy of the simulation method is to use it when the optimal portfolio can be easily obtained, such as with the i.i.d. Markowitz setting and use of (11.42). As our sampling

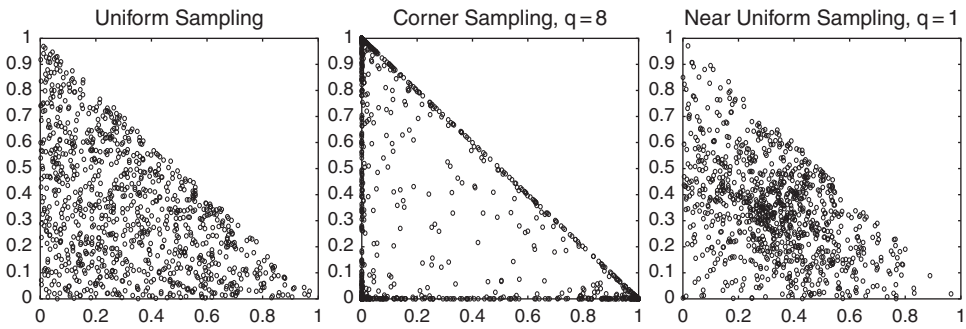

Figure 11.5 Scatterplot of the first two out of three portfolio weights, for different sampling schemes.

schemes are defined so far with only non-negative portfolio weights from (11.41) and (11.46), we use the long-only i.i.d. Markowitz setting.[5] The goal is to conduct a backtesting exercise over moving windows of stock return data, computing for each window the optimal portfolio corresponding to the i.i.d. long-only Markowitz framework, using the program in Listing 11.3, and the optimal portfolio obtained via simulation with s replications. For each of the s replications, a random portfolio vector is selected, and its mean and variance are computed from (11.44). Of these s results, those not fulfilling the mean criterion are discarded, and from the remaining, the one with the smallest variance is returned.

The idea is to repeat this exercise with ever-increasing numbers of simulated portfolios s, and confirm that, as s increases, so does the closeness between the numeric-optimized portfolio and the portfolio obtained via the simulation method. The optimal s is then chosen as the smallest value such that the results of the sampling method are "adequately close" to those obtained from the optimized Markowitz portfolio solution. An obvious measure is the average (over the moving windows) Euclidean distance between the analytical-optimized and sampling-based optimal portfolio.

For this exercise, we use the $T = 3{,}923$ daily (percentage log) returns on 29 stocks from the DJIA-30, from June 1, 1999 to December 31, 2014. (The DJIA-30 index consists of 30 stocks, but for the dates we use, the Visa company is excluded due to its late IPO in 2008.) With $\tau = 10\%$, observe that, for some time periods (for which we use moving windows of length $w = 250$), there might be no solution to the portfolio problem, and the portfolio vector consisting of all zeros is returned, i.e., trading is not conducted on that day. The left panel of Figure 11.6 shows the resulting cumulative returns, including those for the equally weighted portfolio, giving our first demonstration that the simple $1/N$ strategy can outperform the Markowitz allocation framework.

(The extent of this outperformance is understated: When transaction costs are accounted for, the $1/N$ strategy will be relatively even better, as it induces far lower turnover than other strategies.) The interested reader is encouraged to replicate these findings, as it serves as a good warmup to more advanced methods.

We see from the figure that, even for $s = 10{,}000$, the uniform sampling-based method is not able to fully reproduce the optimized portfolio vector results. They differ substantially only during periods for which the sampling method is not able to find a portfolio that satisfies the mean constraint.

---

5 To allow for negative weights in the sampling scheme, one could simply randomly assign a positive or negative sign to the simulated vector **a**, and then renormalize. However, in this case, each weight is no longer restricted to lie in [0, 1], and some constraints would need to be imposed on the lower and upper limits.

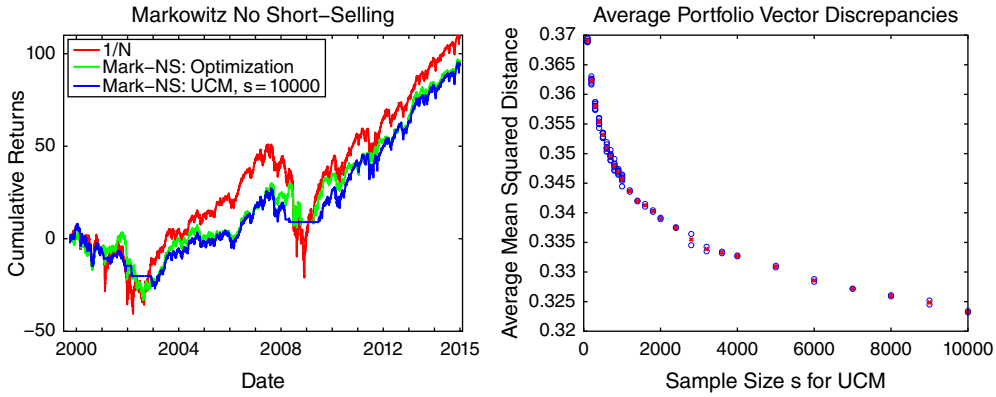

Figure 11.6 Left: Cumulative return sequences of the DJIA data using the equally weighted allocation and the Markowitz iid long-only framework (denoted Mark-NS), based on moving windows of $w = 250$ returns. **Right:** Circles indicate the average, over all the windows, of $\| \mathbf{w}^A - \mathbf{w}^U \|_2$, where $\mathbf{w}^A$ and $\mathbf{w}^U$ refer to the analytic (optimized) and UCM-based portfolio vectors, respectively. This was conducted $h = 8$ times per sample size s for $s \leqslant 1000$, and otherwise $h = 2$ times. Crosses indicate the average over the h-values.

The average portfolio vector discrepancies, as a function of s, are plotted in the right panel. The trajectory indicates that the concept works, but even $s = 10{,}000$ is not yet adequate, and that the primary issue arises during periods for which relatively few random portfolios will obtain the desired mean constraint. Based on this analysis, it is clear that brute-force sampling will not be appropriate, and more clever sampling strategies are required. Section 11.3.4 addresses this issue.

11.3.4 The Univariate Collapsing Method

With more sophisticated models and distributional assumptions, simple formulae such as those in (11.44) are often not available. We discuss a simple and general alternative method of calculating the mean, variance, and, in particular, the ES. Consider a set of d assets for which returns are observed at a particular frequency (such as daily) over a specified period of time. For a particular set of portfolio weights $\mathbf{a} = (a_1, a_2, \ldots, a_d)'$ (chosen either from the simulation-based methodology or by a numeric optimizer), a univariate time series, say $\mathbf{R}_P = (R_{P,1}, R_{P,2}, \ldots, R_{P,T})'$, which we will call the **constructed portfolio return series**, is computed from the d time series of past observed asset returns, $\mathbf{R}_1, \ldots, \mathbf{R}_d$, as $\mathbf{R}_P = a_1\mathbf{R}_1 + \cdots + a_d\mathbf{R}_d$. In our toy example under the i.i.d. assumption, the sample mean and variance of $\mathbf{R}_P$ replaces the analytic calculation in (11.44).

The idea of the constructed portfolio return series is to look at the past returns of the portfolio dictated by weight vector $\mathbf{a}$. These returns are "fictitious" in the sense that the particular portfolio designated by vector $\mathbf{a}$ was most likely not held by the active portfolio manager over the specified time period. It shows the returns that would have occurred if those portfolio weights were used and not changed over time. The use of the constructed portfolio series for risk *assessment*, whereby the portfolio weights are known, and only the risk of the position is required (typically VaR or ES), is within the scope of univariate GARCH modeling, and has been well-studied; see, e.g., Kuester et al. (2006) and the references therein. Its use for risk *management*, whereby active portfolio trading is engaged, is less common; see Manganelli (2004), Bauwens et al. (2006a, p. 143), Andersen et al. (2007, p. 541), Christoffersen (2009, Sec. 3), Paolella (2014), and the references therein. Our goal is to use

$\mathbf{R}_p$, in conjunction with a modeling technique and method for portfolio optimization, for (active) risk management.

Generalizing the toy example above, the—for daily returns data, rather untenable—i.i.d. assumption is replaced by assuming a GARCH-type time-series model for $\mathbf{R}_p$, and this is fit to $\{R_{P,t}\}_{t=1}^{T}$. Then, an h-step-ahead (univariate) density prediction is formed, from which any measurable quantity of interest, such as the mean, variance, VaR, and ES, can be (analytically or empirically) computed. We refer to this as the **univariate collapsing method**, or UCM. While very straightforward conceptually, the problem is the computational time required. In an MGARCH model such as DCC, estimation is performed once, the predictive mean and covariance matrix are determined, and then portfolio optimization is conducted based on the multivariate predictive density. With UCM, for every entertained portfolio vector (by simulation or an optimization algorithm), the constructed portfolio return series needs to be computed (that being computationally trivial) and, in particular, a univariate GARCH-type model needs to be estimated and, from its h-step-ahead density prediction, the mean and a risk measure (variance or VaR or ES) needs to be computed. The latter steps of GARCH model estimation and analytic computation of the ES are the severe bottlenecks of the otherwise useful and straightforward method, and partially explain why, compared to other methods, it has not received much (academic at least) attention.

One solution to this computational issue, as proposed in Paolella (2014), is to use the NCT-APARCH(1,1) model and the fast estimation technique discussed in Section 10.4. Moreover, as the predictive distribution is then NCT (with scale being determined from the APARCH update), the VaR (the left tail quantile) and the ES are also delivered instantaneously via the pre-tabulation method. This enormous gain in speed allows the performance of UCM to be investigated in backtest exercises and further developed.

Observe that, by using an asymmetric, heavy-tailed distribution as the innovation sequence of a flexible GARCH-type model that allows for asymmetric responses to the sign of the returns, UCM respects all the major *univariate* stylized facts of asset returns, as well as a *multivariate* aspect that many models do not address, namely non-ellipticity (see Section C.2), as induced, for example, by differing tail thicknesses and asymmetries across assets. This latter feature is accomplished *in an indirect way* by assuming that the conditional portfolio distribution can be adequately approximated by a noncentral Student's t distribution (NCT). If the underlying assets were to actually follow a location-scale multivariate noncentral t distribution, then their weighted convolution is also noncentral t. This motivation is unfortunately highly tempered, first by the fact that the scale terms are not constant across time, but rather exhibit strong GARCH-like behavior—and it is known that GARCH processes are not closed under summation; see, e.g., Nijman and Sentana (1996). Second, the multivariate NCT necessitates that each asset has the same tail thickness (degrees of freedom), this being precisely an assumption we wish to avoid, in light of evidence against it. Third, in addition to the fact that the underlying process generating the returns is surely not precisely a multivariate noncentral t-GARCH process, even if this were a reasonable approximation locally, it is highly debatable if the process is stationary, particularly over several years.

As such, and as also remarked by Manganelli (2004), UCM (i) relies on the fact that the pseudo-historical time series corresponding to a particular portfolio weight vector can be very well approximated by (as our choice) an NCT-APARCH process and (ii) uses shorter windows of estimation (say, 250 observations, or about one year of daily trading data) to account for the non-stationarity of the underlying process.

> The primary benefit of the UCM is that it avoids the ever-increasing complexity, implementation, numerical issues, and parameter estimation inaccuracy associated with multivariate models, particularly those that support differing tail thicknesses of the assets and embody a multivariate GARCH-type structure.

With the UCM there is no need to employ formal numeric optimization methods to obtain the desired portfolio, nor optimization of model parameters associated with an elaborate multivariate model for the return process. This avoids all their associated problems, such as initial values, local maxima, convergence issues, and specification of tolerance parameters. Moreover, while a multivariate model explicitly captures features such as the (possibly time-varying) covariance matrix, this often necessitates estimation of many parameters, and the curse of model mis-specification can be magnified, as well as the curse of dimensionality, in the sense that, the more parameters there are to estimate, the larger is the magnitude of estimation error. Of course, for the latter, shrinkage estimation is a notably useful method for error reduction. However, not only is the ideal method of shrinkage not known, but even if it were, the combined effect of the two curses can be detrimental to the multivariate density forecast.

Finally, note that, with the UCM, the objective function will not be differentiable in the portfolio weights. This, however, is irrelevant when used with the simulation-based methodology for determining the optimal portfolio.

Remark Observe how portfolio optimization usually first involves obtaining the multivariate predictive distribution of the returns at the future date of interest, and then, in a second step, based on that predictive distribution, the optimal portfolio weight vector is determined. The UCM method does not make use of this two-step approach, but rather uses only univariate information of (potentially thousands) of candidate portfolio distributions to determine the optimal portfolio. The idea of avoiding the usual two-step approach is not new. For example, Brandt et al. (2009) propose a straightforward and successful method that directly models the portfolio weight of each asset as a function of the asset's characteristics, such as market capitalization, book-to-market ratio, and lagged return (momentum), as in Fama and French (1993, 1996). In doing so, and as emphasized by those authors, they avoid the large dimensionality issue of having to model first- and, notably, second-order moments (let alone third-order moments to capture higher-order effects and asymmetries, in which case, the dimensionality explodes).

Based on their suggestion of factors, the method of Brandt et al. (2009) is particularly well-suited to monthly (as they used) or lower-frequency re-balancing, such as bi-monthly, quarterly, or yearly. Our goal is higher frequency re-balancing, such as daily, in which case GARCH-type effects become highly relevant. Fletcher (2017) has independently confirmed the efficacy of the Brandt et al. (2009) approach, using the largest 350 stocks in the United Kingdom. However, he finds that (i) the performance benefits are concentrated in the earlier part of the sample period and have disappeared in recent years, and (ii) there are no performance benefits from use of the methodology based on random subsets of those 350 largest stocks. ■

With respect to the UCM, it is shown in Paolella (2017) that use of naive sampling, as described above, is problematic in the sense that s needs to be very large, and even then performance is not much better than use of the simple $1/N$ strategy. Large improvements are gained by (i) using a data-driven heuristic for mixing the sampling schemes of (11.46) and (11.47), (ii) avoiding trading if a certain fraction of sampled portfolios do not meet the mean-constraint requirement, (iii) augmenting the

search for the optimal portfolio by accounting for characteristics of the individual stock returns, referred to there as the *performance ratio of individual time forecasts* or, amusingly, the PROFITS measure, and (iv) invoking a cutoff mechanism such that one of two portfolios is chosen based on the ES.

The interested reader is referred to Paolella (2017) for a detailed account of this so-called UCM(τ, s):DDS+DONT(k_C)+PROFIT(k_S, k_{CS})+$\tau^*(k_{\text{ES}})$ method or, far more appealing, just UCM2 (perhaps standing for "You See Money Too"). The key to developing such a strategy is to avoid the pernicious trap of **backtest overfitting**, as discussed in Section 13.3 and detailed in Paolella (2017).

To avoid too many lines on the graph (especially without color), we first limit ourselves here to showing the performance of the UCM2 methodology compared to the use of the i.i.d. Markowitz, DCC-Markowitz, and equally weighted strategies, along with the method denoted MCD, based on an i.i.d. discrete two-component multivariate normal mixture model (MixN), as detailed in Chapter 14. Figure 11.7 shows the results. As the UCM is stochastic in nature, eight runs are depicted. We see that the overall best performer is the MCD method in terms of cumulative returns, and it is noteworthy that this model does not use any type of GARCH filter, but rather an i.i.d. framework based on short windows of estimation (to account for the time-varying scale) and use of shrinkage estimation.

Arguably of more interest than the cumulative returns themselves is a risk-adjusted performance measure. The most important (or at least the most common) is the **Sharpe ratio** (assuming a risk-free interest rate of zero), here computed simply as

$$\text{SR} = \frac{250\,\bar{\boldsymbol{r}}}{\sqrt{250}\,\text{std}(\boldsymbol{r})}, \tag{11.48}$$

where $\boldsymbol{r} = (r_1, \ldots, r_T)$ denotes the collection of observed one-step-ahead portfolio results.

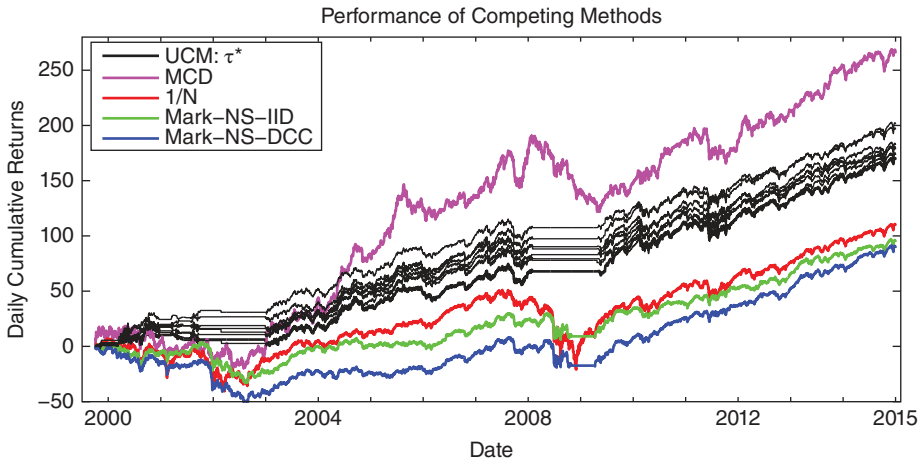

Figure 11.7 The same as the left panel of Figure 11.6, i.e., comparison of cumulative returns for the stocks listed on the DJIA, but with other methods. From top to bottom, the first is the i.i.d. two-component mixed normal distribution with parameters estimated via the MCD methodology, from Gambacciani and Paolella (2017) (in the color version, in purple). This is followed by eight runs of the UCM2 method based on 900 replications (black lines), the equally weighted method (red line), Markowitz (no short selling) based on the i.i.d. assumption (green line), and Markowitz (no short selling) but using the Gaussian DCC-GARCH model for computing the expected returns and their covariance matrix (blue line).

Remarks

a) Other, more recent measures of downside risk-adjusted return that have some advantages over the Sharpe ratio are the so-called **stable tail adjusted return ratio**, or STARR (Martin et al., 2003), and the **Sortino ratio** (Sortino and van der Meer, 1991). Gambacciani and Paolella (2017) show that these measures also favor the MCD model.

b) It is imperative to note that transaction costs were not accounted for in the comparisons illustrated in Figures 11.7 and 11.8. Besides necessarily lowering all the plotted returns (except the equally weighted, which is not affected by use of the simple proportional transaction cost approximations), it could change their relative ranking. For example, while the two Markowitz cases of i.i.d. and DCC-GARCH result in similar performance, inspection of the actual portfolio weights over time reveals that they are much more volatile for the latter, as is typical when GARCH-type filters are used, and presumably would thus induce greater transaction costs, making the DCC approach yet less competitive. ■

For $1/N$, Markowitz-IID, and Markowitz-DCC, we obtain Sharpe ratios of 0.38, 0.47, and 0.43, respectively. The mean Sharpe ratio over the eight UCM runs is 0.95, while that for the MCD method is 0.66. The reason for the superior (and quite good) performance of the UCM method in terms of Sharpe ratio is because the UCM method is unique among the methods shown in its ability to avoid trading during, and the subsequent losses associated with, crisis periods.

As such, it recommends itself to develop a similar methodology for incorporation into the MCD method. This results in the graph shown in Figure 11.8, with label "MixN-MCD-Enhanced" and results in a Sharpe ratio of 0.91. Overlaid is the analogous performance graphic using the method developed in Näf et al. (2018b), which augments the COMFORT model of Section 11.2.4 with more than one latent random variable sequence. It results in a Sharpe ratio of 0.62.

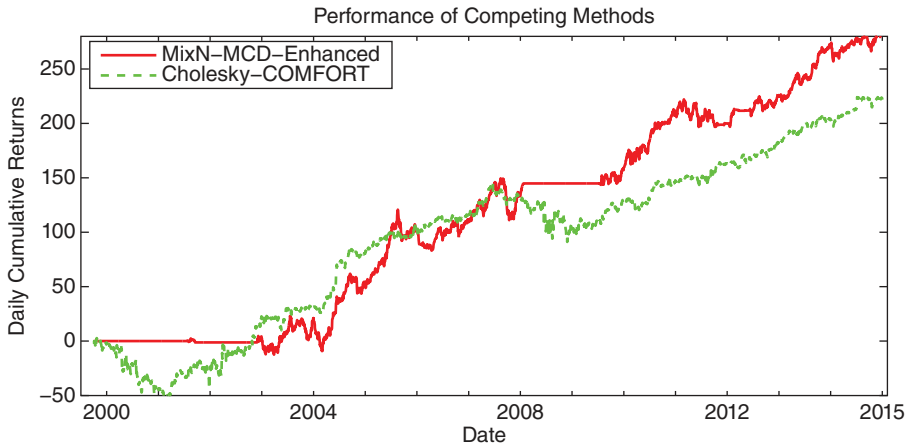

Figure 11.8 Similar to Figure 11.7 but using (i) a modified version of the mixed normal MCD method such that a signal, based on information up to time t, is used to determine if trading should take place at time $t + 1$ or not, and (ii) a new variant of the COMFORT method discussed in Section 11.2.4.

11.3.5 The ES Span

A benefit of the simulation-based approach to determining the optimal portfolio weight vector, as compared to use of direct (gradient/Hessian-based, or evolutionary) optimization algorithms, is that one obtains as a by-product the so-called **ES span**, as introduced in Paolella (2014). Based on a particular time segment of returns data consisting of d assets, denoted as $\mathbf{D}$, and a specified tail probability ξ, we define the distribution of possible values that the ES can take on, over the set of all $\mathbf{a}$, when $\mathbf{a}$ is uniformly chosen over the simplex (11.41), and conditional on a chosen model $\mathcal{M}$, to be $\text{span}_{\text{ES}}(\mathbf{D}, \mathcal{M}, \xi)$. The values obtained from simulation can be plotted as a histogram, and convey knowledge of the distribution of the ES corresponding to $\mathbf{D}$ (and $\mathcal{M}$ and ξ). Observe that use of optimization algorithms gives no such information—they just return a single value (also dependent on $\mathbf{D}$, $\mathcal{M}$ and ξ), which hopefully is the global optimum.

The spread of the ES values, measured as, say, the (sample) variance or interquartile range of $\text{span}_{\text{ES}}(\mathbf{D}, \mathcal{M}, \xi)$, or other measures, such as the distance from the minimal ES value to, say, the ES corresponding to the equally weighted portfolio (possibly scaled by the sample variance or interquartile range), contain information about the relative sensitivity of risk to changes in the portfolio vector, and might be of use in a trading strategy. As an example from Paolella (2014), Figure 11.9 shows approximations of the ES span based on two models. In particular, the $\text{span}_{\text{ES}}(\mathbf{D}, \mathcal{M}, 0.01)$ is depicted as a histogram based on 100,000 replications, drawn uniformly from the simplex, for $\mathbf{D}$ being the matrix of 252 log percentage returns of the 30 stocks in the DJIA, corresponding to years 2005 (left panels) and 2008 (right panels), based on the UCM (top panels) and the i.i.d. two-component multivariate normal mixture model, fit via m.l.e. with shrinkage, as detailed in Chapter 14 (bottom panels).

By contrasting the ES span for the same data set $\mathbf{D}$ based on two models, we see the dependency of the span on the choice of model, and by how much they can differ. Unexpectedly, for both models, the ES span differs remarkably between the two years 2005 and 2008, with 2005 having been chosen because it was a relatively quiet, low-volatility, low-risk period, and 2008 being in the midst of the liquidity crisis. Indicated on the plots as long vertical lines are the minimum ES based on 100,000 simulated portfolios, and the ES corresponding to use of constrained numeric optimization.[6] Also in the plots are short vertical lines, indicating the ES corresponding to putting all the weight on a single stock. (The x-axis was truncated for readability, so that not all 30 are shown.)

Also shown is the ES corresponding to the equally weighted portfolio, denoted "$1/N$". (Further shown is the ES corresponding to the Markowitz minimum variance portfolio. This is just for curiosity: it is not directly comparable to the rest of the graphic because it allows for short selling.) Observe that the ES of the equally weighted portfolio is very close to the center of the span for 2008, for both models, while for 2005 it is much more left of center. This seems logical: The data in 2005 are much less volatile than those in 2008 (during the unfolding of the liquidity crisis) and, more relevantly, also have much thinner tails. As such, the equally weighted portfolio for 2005 should, via the central limit theorem, yield a more Gaussian-like distribution (and, thus, a lower ES) than that in 2008.

We now wish to provide more detail on the claim that the conditional tails in 2008 are thicker than those in 2005. Before proceeding, we explicitly remind the reader what we mean by the

6 The constrained minimization was based on Matlab's `fmincon` function. For the UCM, the ES, being the objective function to minimize, is "close to", but not precisely continuous in the portfolio weights, so that the reported `fmincon` result (whose algorithm requires differentiability and, thus, continuity) is based on the best result obtained from 1,000 runs using different starting values, drawn randomly and uniformly from the simplex. This obviously takes a large amount of time, and was just done for illustration. It is not feasible in practice.

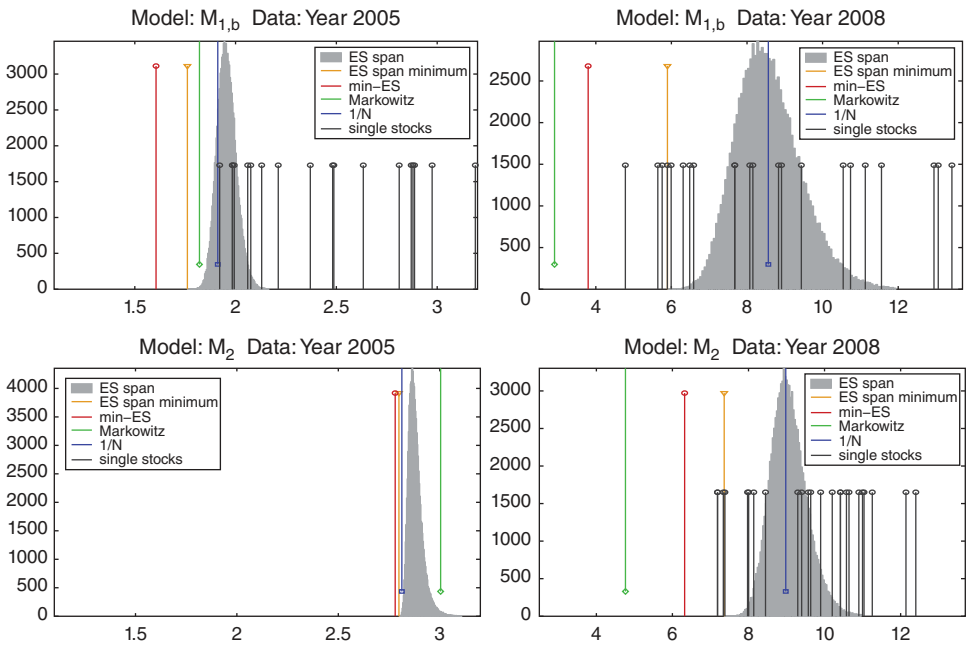

Figure 11.9 Top: The ES span, $\text{span}_{\text{ES}}(\mathbf{D}, \mathcal{M}, 0.01)$, based on the UCM NCT-APARCH(1,1) model of Section 11.3.4 (denoted in the title as $\mathcal{M}_{1,b}$), shown as a histogram from 100,000 random portfolio replications drawn uniformly from the simplex via (11.46), for **D** corresponding to the 252 trading days of years 2005 (left) and 2008 (right). Its minimum value is denoted by the (orange) line "ES span minimum", while the minimum ES obtained by constrained optimization (`fmincon`; repeated 1,000 times because of non-differentiability of the objective function) is indicated by the (red) line "min-ES". The (green) line "Markowitz" indicates the ES corresponding to the minimum variance portfolio allowing for short selling (i.e., negative portfolio weights). The short vertical (black) lines indicate the ES corresponding to putting a weight of one on a single asset (and the rest zero). The *x*-axis was truncated on the right to improve readability, so that some (or all, in the case of the lower left panel) of the ES values corresponding to individual assets are not shown. **Bottom:** Same as top, but based on the i.i.d. discrete two-component multivariate normal mixture model, as discussed in Chapter 14, fit via maximum likelihood with shrinkage (denoted in the title as $\mathcal{M}_2$).

just-mentioned "conditional" tails. This is conditioning on the changing volatility, captured by using a GARCH model for the time-varying scale term. The conditional distribution of asset returns is of far more relevance when interest centers on short-term forecasting and asset allocation. We estimated the NCT-APARCH model for each of the 30 series in 2005. This yields an *average* estimated degrees of freedom parameter of 15.6. The same exercise applied to the 30 stocks for the year 2008 resulted in (the shockingly low value of) 4.0. Now recall the discussions in Section III.9.1 regarding fallacious inference about the tail index when basing it on a fully specified parametric distribution. In particular, if the true underlying process is i.i.d. stable Paretian with tail index (obtained after a trivial bit of trial and error) 1.78, then the estimated tail index *under the erroneous location scale Student's t assumption* is, on average, 4.0 (and between 3.5% and 4.5 90% of the time when conducted using series of length $T = 5{,}000$). The point is: Not knowing the true distribution of the returns, it is very difficult to make reliable inference about the tail index, and the rather low values of the fitted degrees of freedom parameters in the year 2008 under the conditional (accounting for GARCH)

Student's t assumption leads us to question the existence of certainly fourth, but also third, and even second moments in many of the stocks (recall the *average* was 4.0, so approximately half the stocks have an estimated degrees of freedom below this value). It is important to emphasize that the aforementioned average degrees of freedom parameter of 4.0 is *not* referring to the unconditional estimated parametric tail indexes (which would be influenced by conditional heteroskedasticity), but rather the conditional (parametric) tail index (via the degrees of freedom parameter of the NCT) for an NCT-APARCH model, i.e., the varying volatility is accounted for.

Next, we look at the short vertical lines corresponding to investment in a single stock. For 2005, most such investments deliver a much higher ES than the minimum ES portfolio, whereas for 2008, a small number of stocks are such that their ES values are not relatively far from the minimum ES. This at first might appear to be a perverse anomaly (or a mistake), but it makes sense when we juxtapose the facts in the previous discussion about the conditional tail index, the fallacy of parametric-based tail index estimation, and remind ourselves that we cannot assume the conditional returns are literally Student's t. In particular, if one imagines a case in which most stocks have a *genuine* tail index below two (and we again emphasize that its determination is difficult and cannot be based on a parametric assumption such as Student's t or stable Paretian; recall Sections III.9.1 and III.9.2), then, as d increases and (obviously erroneously assuming they are independent), their convolution will be in the domain of attraction of a non-Gaussian stable law, and diversification may not be any better than use of a particular individual stock. Of course, stock returns are not independent; they are, unfortunately, usually all positively dependent. (One can speak of positively *correlated* if second moments exist.) If some stock returns were negatively correlated, then, clearly, diversification would help lower risk. As this is not the case, and given their very heavy tails in 2008, this explains how it can be that holding a single asset may not be much riskier than holding an equally weighted portfolio of all of them.