Joint Forecasts of Dow Jones Stocks Under General Multivariate Loss Function

Tansel Alp [a,b], Matei Demetrescu [c]

[a] *Statistics and Econometric Methods, Goethe University Frankfurt, Germany*
[b] *Applied Research, Metzler Investment GmbH, Germany*
[c] *Applied Econometrics, Goethe University Frankfurt, Germany*

Abstract

When forecasts are assessed by a general loss (cost-of-error) function, the optimal point forecast is not, in general, the conditional mean, and depends on the conditional volatility – which, for stock returns, is time-varying. Our aim is to provide forecasts of daily returns of 30 DJIA stocks under a general multivariate loss function. The paper's contributions are as follows. We discuss what conditions define a multivariate loss function, and suggest a simple class of multivariate loss functions. Based on suitable combinations of univariate loss functions, these are convenient for practical applications with many variables. To keep the computational aspect tractable, we employ a flexible multivariate GARCH model to estimate forecast distributions. It easily copes with large number of series while allowing for skewness, fat tails, non-ellipticity, and tail-dependence. Based on Engle's DCC GARCH, the model uses multivariate affine generalized hyperbolic distributions as conditional probability law, and the number of parameters to be estimated *simultaneously* does not depend on the number of series. We fit our model with daily data from 2002 to 2007 (keeping data from 2008 for out-of-sample forecasts), and use a bootstrap procedure to derive point forecasts under several multivariate loss functions.

Key words:
Asymmetric loss function, loss optimization, GARCH model, generalized hyperbolic distribution, normal inverse Gaussian, bootstrap

Email addresses: tansel.alp@gmx.de, talp@metzler.com (Tansel Alp), deme@wiwi.uni-frankfurt.de (Matei Demetrescu).

Working Paper *This version: 12 September 2008*

1 Motivation

Point forecasts of stock returns series based solely on the history of the series have usually been regarded as being uninteresting, since, according to the martingale difference hypothesis for stock returns, the mean squared error-optimal conditional forecast equals the unconditional expectation of the respective returns series. But when forecast errors incur costs, the forecaster should use the relevant loss (cost-of-error) function, which would typically differ from the squared-error loss. Indeed, it has been argued for a long time (see Granger, 1969) that the squared-error loss function is not able to handle such situations. And, despite there being a tendency to ask, "the higher the return, the better; where's the loss?", costs do arise from misforecasting stock returns. E.g. if a financial intermediary sells a certificate with some risk-sharing scheme (such as a guaranteed minimal return, or even more complicated schemes), he has to forecast returns under the resulting loss function in order to correctly price the certificate. Moreover, asymmetric preferences may be of relevance for financial forecasts too.

Under a general (asymmetric) loss function,[1] the optimal forecast depends on the whole shape of the forecast distribution and not exclusively on the conditional mean. Moreover, stock returns exhibit conditional heteroscedasticity. This too is relevant for forecast optimality, since Christoffersen and Diebold (1996, 1997) show that the optimal forecast under asymmetric loss depends on the conditional variance, and would thus not be constant for stock returns. The history of the returns series is relevant for point forecasts under asymmetric loss, and not only for volatility forecasts.

This paper focuses on jointly forecasting 30 Dow Jones stocks under general multivariate loss functions. Most of the literature focusses on the univariate case. See Granger (1999) for a review of univariate point forecasting under a general loss; in related work, Demetrescu (2007) suggests a way to build forecast intervals under asymmetric loss, Elliott and Timmermann (2004) discuss combining forecasts from different models, and Elliott *et al.* (2005) deal with the subject of inferring about the unknown loss function used to produce a series of observed point forecasts.[2] In the case of multivariate forecasts the building blocks of forecasting procedures are, in principle, the same: on the one hand, one should use the relevant *multivariate* loss function in assessing forecast optimality, while, on the other hand, it is equally important to use a

[1] The most criticized aspect of the squared-error loss function is its symmetry; the literature hence uses the terms "general loss" and "asymmetric loss" interchangeably.

[2] Forecasting under a general loss takes us along the way to estimation under the relevant loss function, see Weiss and Andersen (1984) and Weiss (1996) for arguments in favor thereof, or Gonzáles-Rivera *et al.* (2007) for a recent application.

good estimate of the forecast distribution by modeling serial dependence and conditional distributions appropriately.

But the multivariate nature of our forecasting exercise poses some challenges. For instance, customers may purchase several certificates with risk-sharing from the financial intermediary, and require additional insurance for the value of their whole portfolio of certificates. This implies for the certificate seller that negative forecast errors cause a larger loss if all returns have been over-predicted, and, more generally, that forecast errors should be allowed to interact. By interactions it is understood that one component of the forecast error vector influences the shape of the loss generated by another component. Adding losses from univariate forecasts is a straightforward way to obtain a multivariate loss function, but one prohibiting such interactions. Minimizing a positive definite quadratic form of forecast errors is a tool often used in time series control, and such a quadratic form, if non-diagonal, does accommodate interactions – beside being the natural multivariate generalization of the univariate squared-error. Although a loss function could be derived in a concrete forecasting situation, it is not clear what general conditions multivariate loss functions, and in particular asymmetric ones, should fulfill; and there is no simple asymmetric multivariate loss function available in the literature, either. These are questions of practical, and not only theoretical, interest: although possible, it may not be feasible to derive an exact loss function for a given forecasting situation. [3]

Furthermore, modeling the forecast distribution is not an easy task if the number of variables is not small. For stock returns, multivariate GARCH models are the golden standard. But many of them do run into difficulties when the number of series is large.

Our contribution is twofold. First, we discuss what conditions a multivariate loss function must fulfill and suggest a flexible class of multivariate loss functions. These are based on suitable combinations of univariate loss functions and thus convenient for practical applications. Second, we employ a state-of-the-art multivariate GARCH model to estimate forecast distributions. Its basic structure is that of Engle's DCC model, but we replace the multivariate normal distribution of the innovations with multivariate affine generalized hyperbolic distributions. The model easily copes with large number of series while allowing e.g. for non-elliptical distributions and tail-dependence, and the number of parameters to be estimated *simultaneously* does not depend on the number of series. We then use a parametric bootstrap procedure to derive point forecasts under the relevant multivariate asymmetric loss.

The remainder of the paper is structured as follows. In Section 2, we review

[3] In such cases, loss functions of the type suggested in Section 2 can be used as an approximation and could be fitted from observed loss data.

some theoretical aspects of forecasting under asymmetric loss, discuss properties of multivariate loss functions, and suggest a flexible class of such functions. We then describe the used multivariate GARCH modeling strategy. We provide the forecasting exercise in Section 4, and the final section concludes.

2 Multivariate forecasts under general loss functions

Under a general loss function, the optimal point forecast for the time t minimizes the expected loss, where the expectation is taken with respect to the available information set. The decision-theoretic justification obviously applies in the case of multivariate forecasts, as it does in the case of univariate ones.

Denote $\mathbf{y}_t \in \mathbb{R}^N$ the multivariate time series modeling the quantities to be jointly forecast, and assume $\mathbf{y}_t$ to have conditional probability density function $f(\mathbf{y}_t | \mathcal{F}_{t-1})$. Here, the information set $\mathcal{F}_{t-1}$ contains past values of the process to be forecast, $\mathcal{F}_{t-1} = \{\mathbf{y}_{t-1}, \mathbf{y}_{t-2}, \ldots\}$. Assuming a multivariate loss function $\mathcal{L}$ (whose properties are discussed further below), the optimal point forecast, say $\widehat{\mathbf{y}}_t$, is given as:

$$\widehat{\mathbf{y}}_t = \arg\min_{\mathbf{y}^*} \mathbb{E}\left[\mathcal{L}(\mathbf{y}_t - \mathbf{y}^*) | \mathcal{F}_{t-1}\right]. \tag{1}$$

The forecast density $f(\mathbf{y}_t | \mathcal{F}_{t-1})$ must be of such nature that the expected loss is finite for all t.

The properties of univariate forecasts (see Granger, 1999, or Granger and Machina, 2006, for a review thereof) can be directly extended to the multivariate case, as stated in the following proposition. In what concerns instruments to assess the accuracy of the point forecasts, Demetrescu (2007) suggests an optimality criterion for interval forecasts under asymmetric loss: in the univariate case, an optimal interval should minimize the expected forecast loss conditional on a future realization within the interval, given $\mathcal{F}_{t-1}$ and the desired coverage probability $1 - \alpha$. This optimality criterion is also easily extended to the multivariate case, requiring the expected forecast loss conditional on a future realization within the desired region to be minimal; the way to build such regions in practice is given in the following proposition as well. The proof of Proposition 1 is a straightforward multivariate generalization of the arguments provided in Granger (1999) and Demetrescu (2007) and is not given here.

Proposition 1 *Following statements hold true:*

(1) Assume $\mathcal{L}$ is differentiable. Then, $\nabla\mathcal{L}|_{\mathbf{y}_t - \widehat{\mathbf{y}}_t}$ is a martingale difference

w.r.t. the filtration $\{\mathcal{F}_{t-1}, \mathcal{F}_{t-2}, \ldots\}$;[4]

(2) The optimal forecast $\widehat{\mathbf{y}}_t$ *depends additively on the conditional mean* $\boldsymbol{\mu}_t = \mathbb{E}[\mathbf{y}_t | \mathcal{F}_{t-1}]$;

(3) The so-called bias factor $\mathbf{b}_t = \widehat{\mathbf{y}}_t - \boldsymbol{\mu}_t$ *is a multivariate function of the conditional covariance matrix of* $\mathbf{y}_t$ *alone. The shape of this function is determined by the employed loss function and by the family of distributions* $\mathbf{y}_t$ *belongs to.*

(4) The boundary of the optimal region is described by following equation in $\mathbb{R}^N$*:*

$$\mathcal{L}(\mathbf{y} - \widehat{\mathbf{y}}_t) = c,$$

where c *is implicitly defined from* $P[\mathcal{L}(\mathbf{y}_t - \widehat{\mathbf{y}}_t) \leq c | \mathcal{F}_{t-1}] = 1 - \alpha$.

PROOF: *omitted.*

While, as seen, the principles of forecasting under asymmetric loss generalize in a straightforward manner to the multivariate setting, the loss function itself is more elusive. Any definition of multivariate loss functions should in some way formalize the intuition "the larger the forecast error, the larger the loss". In an univariate framework, this is not problematic; the following characterization, based on Granger (1969), is widely accepted.

Definition 1 *Any continuous function* $\mathcal{L}^*(u) : \mathbb{R} \to \mathbb{R}_+$ *satisfying following conditions*

(1) $\mathcal{L}^*(u)$ *is globally minimized at* $u = 0$*, and*
(2) $\mathcal{L}^*(u)$ *is increasing for* $u > 0$ *and decreasing for* $u < 0$*,*

is a univariate loss function.

An additional regularity condition, namely differentiability of the loss function, is sometimes required (e.g. for (1) in the proposition above).

Obviously, the multivariate loss function must be globally minimized at $\mathbf{0}$, but some care is needed in translating the formulation "the larger the forecast error, the larger the loss". Namely, we need to take into account the fact that there is no natural order relation in multidimensional spaces, hence Definition 1 lacks a straightforward generalization.

One could take this formulation literally by using some length measure for the vector of forecast errors. Unfortunately, this has the disadvantage of imposing rotational symmetry on the loss function, since it implies that vectors of the same length lead to equal loss, irrespective of their direction. This is not compatible with the purpose of allowing for interactions among forecast errors,

[4] The quantity $\nabla\mathcal{L}|_{\mathbf{y}_t - \widehat{\mathbf{y}}_t}$ is the multivariate counterpart of the so-called generalized forecast error, see Patton and Timmermann (2007) for a discussion.

which do not exclude different behavior of the loss function along different directions in the forecast errors space.

A more appealing approach is to define the distance to the origin of a vector of forecast errors in terms of the loss itself. Then, regions of lower loss have to lie closer to the origin, or, more precisely, have to be included in regions of higher loss. We formalize this in the following definition:

Definition 2 *Let* $\mathcal{D}(l) = \{\mathbf{u} : \mathbf{u} \in \mathbb{R}^N, \mathcal{L}(\mathbf{u}) \leq l\}$*, the closed set bounded by the level curve of loss* l*. Any continuous function* $\mathcal{L}(u) : \mathbb{R}^N \to \mathbb{R}_+$ *satisfying*

(1) $\mathcal{L}(\mathbf{u})$ *is globally minimized at* $\mathbf{u} = \mathbf{0}$*, and*
(2) $\forall l_1 < l_2 \in \mathcal{L}\left(\mathbb{R}^N\right) \Rightarrow \mathcal{D}(l_1) \subset \mathcal{D}(l_2)$*,*

is a multivariate loss function.

As expected, Definition 1 is recovered if $N = 1$. However, this property may be a bit too general to lead to a simple characterization of loss functions. Since loss interactions are, by their very nature, direction-specific, one is arguably better off if requiring the loss to be larger for larger forecast errors *in a given direction*. This leads us to following characterization of multivariate loss functions:

Definition 3 *Any continuous function* $\mathcal{L}(\mathbf{u}) : \mathbb{R}^N \to \mathbb{R}_+$ *such that the function* $\mathcal{L}^*(u) = \mathcal{L}(u \cdot \mathbf{u}_0)$ *is a univariate loss function for any* $\mathbf{u}_0 \in \mathbb{R}^N$*, is a multivariate loss function.*

It is immediately seen that this more restrictive definition implies that $\mathcal{L}(\mathbf{u})$ is globally minimized at $\mathbf{u} = \mathbf{0}$. A loss function in the sense of Definition 3 implies being one in the sense of Definition 2, but Definition 3 suggests there might be simple ways to build multivariate loss functions by resorting to univariate ones. Moreover, Definition 3 may be a more tractable tool in checking whether a particular function is a loss function.

Let us turn our attention to the practical problem of setting up a multivariate loss function. While many univariate loss functions (linex, asymmetric linear, asymmetric quadratic) have been used in the forecasting literature, these do not possess a multivariate generalization – apart from the interaction-free sum-of-single-losses. Furthermore, it is not straightforward to obtain a closed-form expression for multivariate loss functions based on either of the definitions proposed above. It can be shown that a continuous, globally convex, function minimized at $\mathbf{u} = \mathbf{0}$ is a multivariate loss in the sense of Definition 2; but this still does not help practitioners in obtaining a loss function appropriate for their needs.

We suggest to simplify the problem at hand by resorting to univariate loss functions which are readily available in the literature. One way to do this

is revealed by an examination of the quadratic form loss function, which, as pointed out in the Introduction, allows for interactions. It is well-known that a positive definite quadratic form can be reduced to a diagonal form by rotation. Therefore, it suggests itself to linearly transform the vector of forecast errors before applying univariate loss functions to its transformed components and adding the resulting losses (possibly with different weights).

Proposition 2 *A multivariate function given by*

$$\mathcal{L}(\mathbf{u}) = \sum_{i=1}^{M} \mathcal{L}^* \left(\sum_{k=1}^{N} a_{ik} u_k \right), \tag{2}$$

where $\mathcal{L}^*$ is a univariate loss function and $A = [a_{ik}]_{i=1,\ldots,M,k=1,\ldots,N}$ has rank N, is a multivariate loss function in the sense of Definition 3.

PROOF: *see the Appendix.*

The case $A = I_N$ recovers the no-interaction situation. Obviously, one can use more than one univariate loss function if additional flexibility is required. This leads to

$$\mathcal{L}(\mathbf{u}) = \sum_{i=1}^{M} \mathcal{L}_i^* \left(\sum_{k=1}^{N} a_{ik} u_k \right), \tag{3}$$

for which the result of Proposition 2 holds as well.

If one does not restrict A to orthonormality, a slightly more general transformation than a rotation is allowed for.[5] In fact, we prefer this form because it may be easier for the practitioner to work with unrestricted matrices A. Moreover, the form in (2) (or (3), for that matter) has a straightforward interpretation tying up to the motivation of the paper. Let a financial intermediary manage several portfolios for different customers; the portfolios are typically based on the same assets, but with different weights. The intermediary is required to forecast the asset returns to price his services. The relevant forecasting loss for the intermediary consists of aggregated portfolio-level losses, which leads to a multivariate loss function having the form in (3). The matrix A must have rank N, otherwise the optimal forecasts cannot be identified individually, case in which only linear combinations of the forecasts are identifiable. This implies, of course, that M must be larger than N.

Any univariate loss function can be used with Proposition 2. Recently, Elliott *et al.* (2005) proposed a simple class of univariate loss functions given by

$$\mathcal{L}^*(u) = (\alpha + (1 - 2\alpha) \cdot 1(u < 0)) |u|^p, \tag{4}$$

[5] One could also choose a nonlinear transformation, as long as its Jacobian has rank N.

with $\alpha \in (0,1)$, $p \in \mathbb{N}\setminus\{0\}$ and $1(\cdot)$ the usual indicator function. The parameter α regulates the asymmetry, while p controls the tail behavior of $\mathcal{L}^*$. Elliott *et al.* point out that these are more suitable for practical applications than, say, the popular linex loss, because of weaker moment requirements on the forecast error distribution. We can drop the requirement of an integer p, since it was only imposed for technical reasons by Elliott *et al.*; $p \in \mathbb{R}_+ \setminus \{0\}$ is sufficient for conditions in Definition 1 to be satisfied. Moreover, different tail behavior can be allowed for $u < 0$ and $u > 0$ by setting $p = p_1 + (p_2 - p_1) \cdot 1(u < 0)$, $p_1, p_2 > 0$. Note that $\mathcal{L}^*$, and thus $\mathcal{L}$, is smooth for $p > 1$ and has smooth derivative for $p > 2$.

As an example, Figure 1 plots for $\alpha = 0.2$ the level curves of loss functions built with $p = 1$ and $p = 2$. We consider the no-interaction case, $A_1 = [(1,0)', (0,1)']$, a case with a non-diagonal matrix A, $A_2 = [(1,0)', (1,1)']$ and the case $A_3 = [(1,1)', (0,1)']$, i.e. the transpose of A_2. For the no-interaction case, the loss generated by one of the forecast errors is influenced only additively, meaning that the optimal point forecast for that component is not influenced by the other component. In contrast, in the cases with interaction, the shape of the loss for one component changes as well, implying that the optimal prediction for one component depends on the optimal prediction for the other component.

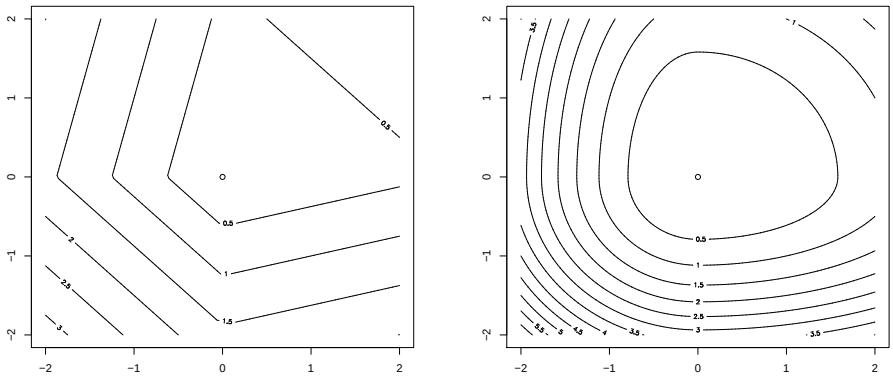
3 Estimating the forecast distribution

Having chosen a loss function, all that is needed to obtain an optimal forecast is to solve the optimum problem in Equation (1). To this end, one has to model the forecast distribution.

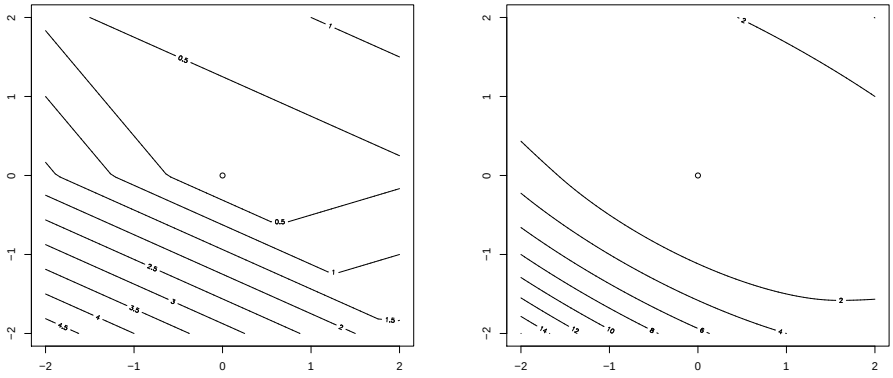
The dominant feature of stock returns series to be considered when modeling is of course volatility clustering. Shape of conditional densities is an important issue as well: one must account for possible non-ellipticity, fat tails and tail dependence while maintaining the model complexity at a low level to allow for reliable estimation of model parameters. This section presents a multivariate GARCH model fulfilling these conditions even for a very large number of components.

3.1 The model

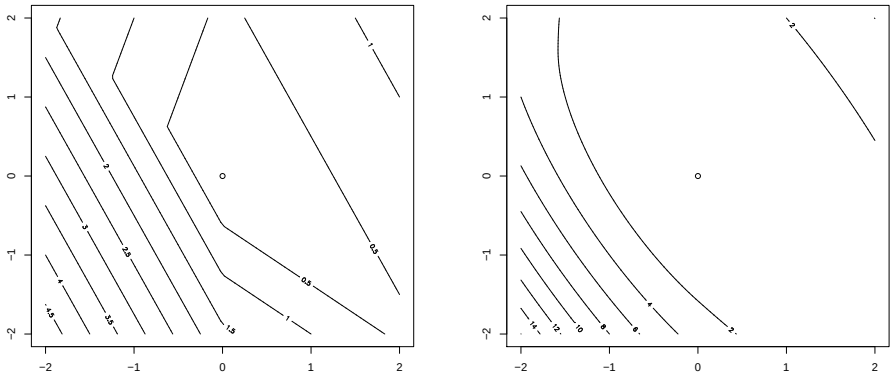
A number of non-elliptical multivariate distributions have been suggested as a means of modeling the conditional probability law of vector time series. To mention but a few, the families of skewed Student's t (see Bauwens and



$$A = [(1,0)', (0,1)']$$



$$A = [(1,0)', (1,1)']$$



$$A = [(1,1)', (0,1)']$$

Fig. 1. Contour plots of bivariate loss functions, $\alpha = 0.2$, $p = 1$ (left), $p = 2$ (right)

Laurent, 2005) and normal inverse Gaussian [NIG] distributions (see Aas *et al.*, 2006) have attracted much interest in the field of financial econometrics.

Although appealing from a statistical point of view and performing well with financial data, these distributions are less suitable in applications with dozens of variables. Even if the conditional volatilities and correlations were estimated in a preliminary stage by Gaussian quasi-maximum likelihood [QML], there would still be $O(N)$ parameters involved in *simultaneous* estimation of, say, a skewed Student's t model.

Thus, we follow Alp (2007) and take the multivariate affine generalized hyperbolic [MAGH] distribution of Schmidt *et al.* (2006) as the multivariate probability distribution of $\mathbf{y}_t$ conditional on the information set $\mathcal{F}_{t-1}$. More precisely, the model we use is a special case of the one due to Alp (2007), who allows for a Markov regime-switching framework to capture time-varying tail dependence. Thus, we assume

$$\begin{aligned} &\mathbf{y}_t - \boldsymbol{\mu}_t = \boldsymbol{\epsilon}_t \\ &\boldsymbol{\epsilon}_t = H_t^{1/2}\, \boldsymbol{v}_t \\ &\boldsymbol{v}_t \sim \text{ iid}(\mathbf{0}, I_N), \end{aligned} \tag{5}$$

where the lower triangular matrix $H_t^{1/2} = [h^*_{ik,t}]_{i,k=1,\ldots,N}$ is obtained by the Cholesky factorization of the conditional covariance matrix $H_t = \text{Cov}[\boldsymbol{y}_t|\mathcal{F}_{t-1}]$.[6] Following Bollerslev (1990) and Engle (2002), among others, a natural starting point in modeling the time evolution of the conditional covariances is to decompose $H_t = [h_{ik,t}]_{i,k=1,\ldots,N}$, as

$$H_t = \text{diag}(H_t)^{1/2}\; R_t\; \text{diag}(H_t)^{1/2}, \tag{6}$$

where $R_t = [\rho_{ik,t}]_{i,k=1,\ldots,N}$ is the time-varying conditional correlation matrix of $\boldsymbol{y}_t$ and diag$(\cdot)$ denotes the diagonal matrix with the same diagonal as its argument.

In what concerns the dynamics of the conditional variances $h_{ii,t}$, $i = 1, \ldots, N$, it is common practice to choose some univariate GARCH specification, typically GARCH(1,1). Note that additional features, such as unconditional volatility changes, can easily be added to the multivariate model through the univariate models for the volatilities $h_{ii,t}$.

For the conditional correlation matrix R_t we adopt the dynamic conditional correlation [DCC] model of Engle (2002). This class of multivariate GARCH models achieves additional flexibility compared to models with constant conditional correlations at the cost of only two additional parameters.

To specify the structure of the conditional correlation matrix R_t in the DCC

[6] Unless $\boldsymbol{v}_t$ is distributed according to an elliptically symmetric distribution, a different decomposition of H_t would result in a different class of distributions for $\boldsymbol{y}_t$.

model, it is natural to start with an auxiliary symmetric $N \times N$ matrix Q_t, which is positive definite and not necessarily a correlation matrix. We assume that Q_t obeys the following linear filter representation

$$\begin{aligned} \text{vech}(Q_t) &= \text{vech}(R) + \sum_{j=1}^{\infty} \Upsilon_j \, \text{vech}\Big(\boldsymbol{\epsilon}_{t-j}\boldsymbol{\epsilon}_{t-j}' - R\Big) \\ &= \text{vech}(R) + \Upsilon(B) \, \text{vech}\Big(\boldsymbol{\epsilon}_t\boldsymbol{\epsilon}_t' - R\Big), \end{aligned} \tag{7}$$

where B denotes the backshift or lag operator and the half-vec operator $\text{vech}(\cdot)$ transforms a symmetric matrix into a vector containing the lower triangular elements of the matrix. The sequence $\{\text{vech}(\boldsymbol{\epsilon}_t\boldsymbol{\epsilon}_t' - R)\}$ is a collection of zero-mean random vectors with the property that

$$\mathbb{E}\Big[\Big\{\epsilon_{i,t}\epsilon_{k,t} - \rho_{ik}\Big\}\Big\{\epsilon_{l,t-j}\epsilon_{m,t-j} - \rho_{lm}\Big\}\Big] = 0, \tag{8}$$

for all $i, k, l, m = 1, \ldots, N$ and $j = \pm 1, \pm 2, \ldots$, where $R = [\rho_{ik}]_{i,k=1,\ldots,N}$ denotes the unconditional correlation matrix of $\boldsymbol{\epsilon}_t$. For the process $\{\text{vech}(Q_t)\}$ to be covariance-stationary, we assume the Υ-weights in the infinite-order moving average representation in (7) to be absolutely summable, i.e. $\sum_{j=1}^{\infty} |\Upsilon_j| < \infty$. These weights are further restricted to be of such nature that Q_t is positive definite at all time.

In general, Q_t will not be a valid correlation matrix in the sense that it would be positive definite with a unit diagonal and off-diagonal elements between -1 and +1. Engle (2002) therefore suggested the following transformation

$$R_t = \text{diag}(Q_t)^{-1/2} \, Q_t \, \text{diag}(Q_t)^{-1/2} \tag{9}$$

in order to fully specify his DCC model.

A particular choice for the infinite-order polynomial $\Upsilon(B)$ is given by

$$\Upsilon(B) = \frac{\theta \, B}{1 - \kappa \, B} = \theta \sum_{j=0}^{\infty} \kappa^j \, B^{j+1} = \theta \, B \sum_{j=1}^{\infty} \kappa^{j-1} \, B^{j-1}, \tag{10}$$

provided that $|\kappa| < 1$. Thus, the linear filter representation of $\text{vech}(Q_t)$ becomes

$$\text{vech}(Q_t) = \text{vech}(R)(1 - \theta - \kappa) + \theta \, \text{vech}\Big(\boldsymbol{\epsilon}_{t-1}\boldsymbol{\epsilon}_{t-1}'\Big) + \kappa \, \text{vech}(Q_{t-1}) \tag{11}$$

or, equivalently, in matrix version

$$Q_t = R(1 - \theta - \kappa) + \theta \, \boldsymbol{\epsilon}_{t-1}\boldsymbol{\epsilon}_{t-1}' + \kappa \, Q_{t-1}, \tag{12}$$

which is recognized as the scalar specification of the DCC model introduced by Engle (2002). It is easily seen that the conditions for $\{\text{vech}(Q_t)\}$ to be

covariance-stationary amount to requiring $|\kappa| < 1$. Finally, the restrictions $\kappa, \theta > 0$ and $\kappa + \theta < 1$ are sufficient to ensure that Q_t will be positive definite at all time.

It should be noted that additional flexibility can be attained by considering a block-diagonal DCC structure, as suggested by Billio *et al.* (2006).

Let us now turn our attention to the innovation vector $\boldsymbol{\upsilon}_t = [\upsilon_{1,t}, \ldots, \upsilon_{N,t}]'$. Flexibility in modeling the innovations is the keystone to accounting for stylized facts such as skewness, fat-tails, non-ellipticity or tail dependence. We assume that $\boldsymbol{\upsilon}_t$ is an *iid* sequence of random vectors with independent components. To obtain the desired flexibility, let each $\upsilon_{i,t}$ admit the following stochastic decomposition

$$\upsilon_{i,t} = \phi_i + \beta_i \, \delta_i \, \zeta_{i,t} + \sqrt{\delta_i \, \zeta_{i,t}} \, Z_{i,t}, \quad i = 1, \ldots, N. \tag{13}$$

The quantities $\zeta_{1,t}, \ldots, \zeta_{N,t}$ are *iid* generalized inverse Gaussian random variables, $\zeta_{i,t} \sim \mathcal{GIG}(\lambda_i, 1, \gamma_i)$, distributed independently of $[Z_{1,t}, \ldots, Z_{N,t}]' \sim \mathcal{N}_N(\mathbf{0}, I_N)$.[7] The probability density corresponding to the $\mathcal{GIG}(\lambda_i, 1, \gamma_i)$ distribution is given by

$$\frac{\gamma_i^{\lambda_i}}{2K_{\lambda_i}(\gamma_i)} \exp\Big(-\tfrac{1}{2}\left(s^{-1} + \gamma_i^2 s\right)\Big) \, s^{\lambda_i - 1}, \qquad s > 0 \tag{14}$$

where K_{λ_i} denotes the modified Bessel function of the third kind with index $\lambda_i \in \mathbb{R}$. In fact, the normal variance-mean mixture in (13) defines the one-dimensional generalized hyperbolic [GH] distribution. Pseudo-random numbers are easily generated from the generalized inverse Gaussian distribution defined by (14), given suitable estimators $\widehat{\lambda}_i$ and $\widehat{\gamma}_i$. An algorithm that may be useful for these purposes is the one due to Michael *et al.* (1976).

In order to ensure H_t to be the conditional covariance matrix of $\boldsymbol{y}_t$, the coefficients δ_i and ϕ_i are re-parameterized as follows (see Alp, 2007)

$$\begin{aligned} \delta_i &= \begin{cases} \dfrac{-\mathbb{E}[\zeta_{i,t}] + \sqrt{\mathbb{E}[\zeta_{i,t}]^2 + 4\beta_i^2 \left(\mathbb{E}[\zeta_{i,t}^2] - \mathbb{E}[\zeta_{i,t}]^2\right)}}{2\beta_i^2 \left(\mathbb{E}[\zeta_{i,t}^2] - \mathbb{E}[\zeta_{i,t}]^2\right)}, & \text{if } \beta_i \neq 0 \\ \mathbb{E}[\zeta_{i,t}]^{-1}, & \text{if } \beta_i = 0 \end{cases} \\ \phi_i &= -\delta_i \, \beta_i \, \mathbb{E}[\zeta_{i,t}], \end{aligned} \tag{15}$$

[7] It is worth mentioning that Schmidt *et al.* (2006) were not the first to use an affine transformation of independent non-elliptical random variables for the purpose of modeling multivariate data. Nagahara (2004), for example, takes independent random variables from the Pearson distribution system and makes them subject to an affine transformation.

where the r^{th} order moments about the origin of the $\mathcal{GIG}(\lambda_i, 1, \gamma_i)$ distribution are given for $\gamma_i > 0$ and any integer r by

$$\mathbb{E}\Big[\zeta_{i,t}^r\Big] = \frac{K_{\lambda_i+r}(\gamma_i)}{\gamma_i^r\, K_{\lambda_i}(\gamma_i)}, \qquad i = 1, \ldots, N. \tag{16}$$

The parameters $\beta_i \in \mathbb{R}$ control the skewness of the GH distributions, which will be symmetric about zero if $\beta_i = 0$.

Given that the parameter λ_i is rather difficult to estimate, it suggests itself to fix its value and thereby consider a particular subclass of the GH distribution. The multivariate affine normal inverse Gaussian [MANIG] subclass of the MAGH distribution, for example, stems from setting $\lambda_1 = \cdots = \lambda_N = -1/2$. In this special case, the expressions for δ_i and ϕ_i in (15) simplify considerably:

$$\begin{aligned} \delta_i &= \begin{cases} \frac{\gamma_i^2}{2\beta_i^2}\left(\sqrt{1 + 4\beta_i^2/\gamma_i} - 1\right), & \text{if } \beta_i \neq 0 \\ \gamma_i^{-1}, & \text{if } \beta_i = 0 \end{cases} \\ \phi_i &= -\frac{\delta_i\, \beta_i}{\gamma_i}, \qquad i = 1, \ldots N. \end{aligned} \tag{17}$$

As discussed in Schmidt *et al.* (2006) and Alp (2007), the MAGH distribution can handle a wide variety of departures from normality in the marginal distributions as well as the copula. To name but a few, the marginal distributions of the variables can be modeled as semi-heavy tailed with one tail being heavier than the other and the dependence among the variables can be characterized by the property of multivariate tail dependence. The upper tail dependence coefficient of the MAGH distribution, given the available information $\mathcal{F}_{t-1}$, is defined for the set $M = \{1, \ldots, N\}$ as

$$\lim_{v \to 0^+} \mathbb{P}\Big[F_{m,t}(y_{m,t}|\mathcal{F}_{t-1}) > 1-v,\, \forall m \in M \backslash \{n\} \,\Big|\, F_{n,t}(y_{n,t}|\mathcal{F}_{t-1}) > 1-v,\, n \in M\Big], \tag{18}$$

where $F_{i,t}$, $i = 1, \ldots, N$, denotes the conditional probability distribution of $y_{i,t}$ given $\mathcal{F}_{t-1}$. It can be shown that the limit in (18) does exist; the conditions for the limit to be positive are given in Alp (2007). In the MANIG special case, these conditions read as follows: $h^*_{ji,t} > 0$ and

$$\left(\sqrt{\beta_i^2 + \gamma_i^2/\delta_i} - \beta_i\right)\rho_{j1}\sqrt{h_{jj,t}} \;\geq\; \left(\sqrt{\beta_1^2 + \gamma_1^2/\delta_1} - \beta_1\right)h^*_{ji,t}, \tag{19}$$

for all $i > 1$ and $j = i, \ldots, N$. If at least one of the inequalities in (19) is violated, the limit in (18) will be zero.

3.2 Estimation

The estimation of our time series model from Section 3.1 can be carried out by maximum likelihood. Under the assumption that the vector return series $\boldsymbol{y}_t$ follows the MANIG distribution conditionally, the log-likelihood function for a sample with T observations is

$$L_T(\Phi) = -\sum_{t=1}^{T}\Big[\log\Big(\det(R_t^{1/2})\Big) + \sum_{i=1}^{N}\Big(\tfrac{1}{2}\log(h_{ii,t}) - \log\left(g_i(\upsilon_{i,t})\right)\Big)\Big], \qquad (20)$$

where Φ comprises all unknown parameters in the model, g_i is the probability density corresponding to the right-hand side of (13) (for $\lambda_i = -1/2$) and $\det(\cdot)$ denotes the determinant of a square matrix. The argument of g_i, $\upsilon_{i,t}$, is the i-th element of the vector $\boldsymbol{\upsilon}_t = H_t^{-1/2}(\boldsymbol{y}_t - \boldsymbol{\mu}_t)$.

Direct maximization of the log-likelihood function in (20) w.r.t. Φ may be computationally intensive, especially if there are more than a few variables to be modeled. But a further advantage of DCC-type models, in addition to their parsimony, is that its estimation can be split in several stages which can be addressed one at the time. Hence, we obtain our estimates of the model parameters in three steps:

(1) For each $i = 1, \ldots, N$, estimate a univariate GARCH(1,1) model under conditional normality, possibly using the variance targeting constraint, see Engle and Mezrich (1996); denote estimated means and conditional variances as $\widehat{\mu}_i$ and $\widehat{h}_{ii,t}$, respectively.
(2) Obtain the standardized, but correlated, returns $\widehat{\epsilon}_{i,t} = (y_{i,t} - \widehat{\mu}_i)/\sqrt{\widehat{h}_{ii,t}}$, from which the pairwise unconditional correlations are estimated for all $k > i$ as $\widehat{\rho}_{ik} = (\sum_t \widehat{\epsilon}_{i,t}\,\widehat{\epsilon}_{k,t})/\sqrt{(\sum_t \widehat{\epsilon}_{i,t}^2)(\sum_t \widehat{\epsilon}_{k,t}^2)}$. Conditional on these estimates, maximize the Gaussian log-likelihood $-\frac{1}{2}\sum_t \log(\det(R_t)) + \widehat{\boldsymbol{\epsilon}}_t R_t^{-1}\widehat{\boldsymbol{\epsilon}}_t$ w.r.t. the unknown parameters in (12), namely κ and θ. Let $\widehat{\upsilon}_{i,t}$ denote the orthogonalized, standardized returns.
(3) Finally, compute for each $i = 1, \ldots, N$ the estimates of γ_i and β_i individually as the solution to $\max_{\gamma_i,\beta_i} \sum_t \log(g_i(\widehat{\upsilon}_{i,t}; \gamma_i, \beta_i))$.

A potential drawback of using Gaussian QML estimators is that the conditions for establishing consistency results, not to mention those required for asymptotic normality, are hard to verify for the time series model examined here.

In the constant conditional correlation case ($\kappa = \theta = 0$), however, the requirements given by Lee and Hansen (1994) for consistent QML estimation of the parameters in a univariate, covariance-stationary GARCH(1,1) model

are easily met by our data-generating process. [8] As a direct consequence, the two-step sample correlation estimator $\widehat{\rho}_{ik}$ will be consistent and so too will be the three-step ML estimators $\widehat{\gamma}_i$ and $\widehat{\beta}_i$.

For the DCC-MANIG data generating process, the following subsection contains a Monte Carlo examination of the finite-sample properties of the multi-stage estimator.

3.3 Finite-sample behavior

This subsection deals with the issue of multi-stage Gaussian QML estimation of the DCC-MANIG model. To investigate the finite-sample properties of the above multi-stage estimator we conducted a small Monte Carlo experiment on the basis of the parameter values in Table 1. We have also experimented with other parameter constellations. Since we have obtained similar results, we do not report them all for the sake of brevity; the results are available upon request.

On each simulation run we generated a vector time series of 1500 observations from the bivariate DCC-MANIG model. For sample sizes of 700, 1000 and 1500 observations we then estimated the parameters of interest as described in Subsection 3.2. The QML estimates at each step were found by a sequential quadratic programming algorithm using numerical derivatives. To make sure that the found maximum of a quasi-likelihood function was also its global maximum, each optimization problem was started several times from randomly chosen parameter values. All computations in this section were done using the Ox programming language of Doornik (2006).

Table 1 reports the bias and the mean squared error [MSE] for the estimators coming from the successive steps of our estimation procedure, based on a total of 1000 Monte Carlo replications. Notice that the bias and the MSE decrease, although not always steadily, with increasing sample size, an indication of the consistency of the multi-stage QML estimators.

For comparison, we have repeated the Monte Carlo experiment in the same setup for the multi-stage ML estimator in the Gaussian DCC model of Engle (2002). Although the bias and the MSE in the latter case are in general less than in case of the DCC-MANIG model, [9] the Monte Carlo results indicate

[8] The conditions for consistency Lee and Hansen (1994) are fulfilled: the process $\{\sum_{k=1}^{i} \rho_{ik}^{*} \upsilon_{k,t}\}$, as a collection of *iid* random variables, is strictly stationary and ergodic, and, for $\rho_{ik}^{*} > 0$, all moments of the NIG variable $\rho_{ik}^{*} \upsilon_{k,t}$ do exist and are finite, see Alp (2007).

[9] This should not come as a surprise, since, under the Gaussian DCC data gener-

Table 1
Monte Carlo experiment
[a]

		DCC-MANIG model			DCC-Gaussian model		
Parameter	True	700	1000	1500	700	1000	1500
		GARCH parameters					
d_1	0.15	-0.006	-0.004	-0.004	-0.002	0.000	-0.001
		(0.003)	(0.002)	(0.001)	(0.001)	(0.001)	(0.001)
e_1	0.8	-0.020	-0.014	-0.010	-0.014	-0.011	-0.006
		(0.007)	(0.004)	(0.003)	(0.003)	(0.002)	(0.001)
d_2	0.2	-0.003	-0.006	-0.005	-0.001	-0.002	-0.002
		(0.004)	(0.003)	(0.002)	(0.002)	(0.001)	(0.001)
e_2	0.7	-0.029	-0.014	-0.008	-0.015	-0.008	-0.005
		(0.011)	(0.006)	(0.004)	(0.005)	(0.003)	(0.002)
		DCC parameters					
ρ_{12}	0.7	-0.011	-0.009	-0.009	-0.014	-0.013	-0.012
		(0.003)	(0.002)	(0.001)	(0.002)	(0.001)	(0.001)
θ	0.1	0.006	0.005	0.002	0.003	0.004	0.001
		(0.002)	(0.002)	(0.001)	(0.001)	(0.001)	(0.000)
κ	0.8	-0.051	-0.032	-0.023	-0.023	-0.018	-0.008
		(0.023)	(0.012)	(0.009)	(0.008)	(0.005)	(0.002)
		NIG parameters					
γ_1	0.8	0.113	0.078	0.054			
		(0.079)	(0.043)	(0.027)			
β_1	-0.2	-0.008	-0.008	-0.007			
		(0.008)	(0.005)	(0.003)			
γ_2	0.6	0.090	0.059	0.037			
		(0.039)	(0.022)	(0.012)			
β_2	-0.3	-0.011	-0.005	-0.003			
		(0.009)	(0.005)	(0.004)			

[a] *Note*: For simulated vector time series of 700, 1000 and 1500 observations this table reports the bias and the MSE (in brackets) resulting from multi-stage estimation of the DCC-MANIG and the DCC-Gaussian models. The conditional variances are generated as $h_{ii,t} = \overline{h}_{ii}(1 - d_i - e_i) + d_i\, r^2_{i,t-1} + e_i\, h_{ii,t-1}$, $i = 1, 2$, where $\overline{h}_{11} = 0.347$ and $\overline{h}_{22} = 0.246$. The number of Monte Carlo replications is 1000.

that for large sample sizes, available in most financial data sets, the differences are not striking.

Thus, the multi-stage approach to estimating the DCC-MANIG model is slightly biased in finite samples; on the other hand, it is a natural alternative to a one-step ML estimator, very much in the same way as Engle's multi-step DCC estimator.

ating process, these are ML estimators, unlike in the case of the DCC-MANIG data generating process.

4 Forecasting Dow Jones stock returns

4.1 Data and model estimates

The employed data set consists of daily closing prices for the components of the Dow Jones Industrial Average [DJIA] Index. The prices have been obtained from Yahoo! Finance (http://finance.yahoo.com) and are adjusted for dividends and splits. We use more than six years of data for the period ranging from December 31st, 2001 to March 5th, 2008; the returns are computed as log-price differences.

When fitting the model, we omit the last 44 days from the sample (i.e. all observations from 2008) for purposes of out-of-sample forecasting. The parameter estimates based on the above multi-stage procedure are given, for a total of 1510 observations, in Table 2. The corresponding standard errors are obtained semi-parametrically by the bootstrap methodology (see Efron and Tibshirani, 1986). We therefore draw 100 bootstrap samples, each having the same size as the original sample, from the estimated, standardized and orthogonalized returns. With the parameter estimates and the bootstrap samples at hand, we recursively generate the bootstrap series of interest and estimate for each the parameters of the DCC-MANIG model by the above multi-stage procedure. Each bootstrap series is constructed by conditioning on the first observation in the original sample. The standard error of an estimator is then computed as its sample standard deviation over the bootstrap replications.

The statistical significance of the parameter estimates of the GARCH and DCC processes is assessed on the basis of the Wald statistic, which, under the null, exceeds a given critical value $c > 0$ with probability $\frac{1}{2}\mathbb{P}[\chi^2(1) \geq c]$. The reason for the classical asymptotic theory to break down is that the parameters defining the null hypothesis lie at the boundary of the admissible parameter space, violating a standard regularity condition. We will return to this aspect later on. The significance of most of the estimated GARCH parameters at conventional critical levels confirms the presence of conditional heteroscedasticity in the data.

Further, both of the parameters θ and κ in the DCC process are found to be significantly different from zero, providing evidence for time-varying conditional correlations between the studied time series.

But the perhaps most interesting question about our estimated time series model is whether the MANIG law is to be preferred over its Gaussian counterpart in terms of describing the conditional probability distribution of the returns. In this respect, we follow Alp (2007) and conduct a test of the null

Table 2. Multi-stage parameter estimates of DCC-MANIG model

[a]

Stocks	GARCH parameters						NIG parameters						DCC parameters		
MMM	d_1	0.060 *	(0.038)	e_1	0.828 ***	(0.164)	ξ_1	1.225 ***	(0.298)	β_1	-0.004	(0.062)	θ	0.003 ***	(0.001)
AA	d_2	0.027 ***	(0.013)	e_2	0.964 ***	(0.077)	ξ_2	0.682 ***	(0.222)	β_2	0.000	(0.091)	κ	0.965 ***	(0.022)
AIG	d_3	0.114 ***	(0.030)	e_3	0.869 ***	(0.036)	ξ_3	0.873 ***	(0.282)	β_3	0.023	(0.066)			
AXP	d_4	0.081 ***	(0.017)	e_4	0.910 ***	(0.021)	ξ_4	0.657 ***	(0.204)	β_4	0.226 **	(0.103)			
T	d_5	0.035 ***	(0.009)	e_5	0.963 ***	(0.010)	ξ_5	0.462 ***	(0.113)	β_5	0.058	(0.087)			
BAC	d_6	0.031 ***	(0.015)	e_6	0.962 ***	(0.027)	ξ_6	0.972 ***	(0.359)	β_6	-0.119 *	(0.063)			
BA	d_7	0.036 ***	(0.011)	e_7	0.956 ***	(0.018)	ξ_7	0.369 ***	(0.152)	β_7	0.268 **	(0.129)			
CAT	d_8	0.010	(0.020)	e_8	0.977 ***	(0.277)	ξ_8	1.135 ***	(0.362)	β_8	-0.050	(0.066)			
CVX	d_9	0.072 ***	(0.017)	e_9	0.895 ***	(0.032)	ξ_9	0.177	(0.159)	β_9	-0.460	(0.331)			
C	d_{10}	0.072 ***	(0.015)	e_{10}	0.923 ***	(0.017)	ξ_{10}	0.995 ***	(0.276)	β_{10}	0.013	(0.057)			
KO	d_{11}	0.052 ***	(0.018)	e_{11}	0.938 ***	(0.027)	ξ_{11}	0.825 ***	(0.322)	β_{11}	0.069	(0.089)			
DD	d_{12}	0.044 ***	(0.014)	e_{12}	0.944 ***	(0.023)	ξ_{12}	0.918 ***	(0.269)	β_{12}	0.108	(0.081)			
XOM	d_{13}	0.061 ***	(0.018)	e_{13}	0.920 ***	(0.028)	ξ_{13}	0.430 ***	(0.141)	β_{13}	0.017	(0.084)			
GE	d_{14}	0.031 ***	(0.010)	e_{14}	0.967 ***	(0.016)	ξ_{14}	0.673 ***	(0.170)	β_{14}	0.234 **	(0.107)			
GM	d_{15}	0.054 ***	(0.024)	e_{15}	0.927 ***	(0.038)	ξ_{15}	1.487 ***	(0.382)	β_{15}	0.068	(0.052)			
HPQ	d_{16}	0.015	(0.015)	e_{16}	0.981 ***	(0.174)	ξ_{16}	1.420 ***	(0.354)	β_{16}	0.023	(0.055)			
HD	d_{17}	0.032 ***	(0.009)	e_{17}	0.964 ***	(0.014)	ξ_{17}	0.733 ***	(0.177)	β_{17}	-0.014	(0.073)			
IBM	d_{18}	0.078 ***	(0.019)	e_{18}	0.905 ***	(0.027)	ξ_{18}	1.344 ***	(0.372)	β_{18}	-0.021	(0.052)			
INTC	d_{19}	0.027 **	(0.015)	e_{19}	0.969 ***	(0.026)	ξ_{19}	1.256 ***	(0.486)	β_{19}	-0.023	(0.069)			
JNJ	d_{20}	0.090 ***	(0.022)	e_{20}	0.896 ***	(0.026)	ξ_{20}	1.023 ***	(0.203)	β_{20}	0.102	(0.066)			
JPM	d_{21}	0.065 ***	(0.014)	e_{21}	0.932 ***	(0.019)	ξ_{21}	0.707 ***	(0.199)	β_{21}	0.020	(0.076)			
MCD	d_{22}	0.048 ***	(0.011)	e_{22}	0.945 ***	(0.017)	ξ_{22}	0.751 ***	(0.190)	β_{22}	0.005	(0.076)			
MRK	d_{23}	0.009	(0.075)	e_{23}	0.929 ***	(0.389)	ξ_{23}	3.035 **	(1.711)	β_{23}	-0.003	(0.043)			
MSFT	d_{24}	0.034 *	(0.024)	e_{24}	0.955 ***	(0.037)	ξ_{24}	1.764 ***	(0.743)	β_{24}	0.098	(0.073)			
PFE	d_{25}	0.105 ***	(0.038)	e_{25}	0.836 ***	(0.062)	ξ_{25}	1.381 ***	(0.466)	β_{25}	-0.001	(0.064)			
PG	d_{26}	0.055 ***	(0.024)	e_{26}	0.902 ***	(0.072)	ξ_{26}	0.826 ***	(0.222)	β_{26}	0.086	(0.080)			
UTX	d_{27}	0.065 ***	(0.014)	e_{27}	0.920 ***	(0.023)	ξ_{27}	0.805 ***	(0.224)	β_{27}	0.028	(0.077)			
VZ	d_{28}	0.045 ***	(0.012)	e_{28}	0.951 ***	(0.017)	ξ_{28}	0.423 ***	(0.118)	β_{28}	0.101	(0.097)			
WMT	d_{29}	0.026 ***	(0.007)	e_{29}	0.969 ***	(0.015)	ξ_{29}	0.889 ***	(0.202)	β_{29}	0.051	(0.071)			
DIS	d_{30}	0.081 ***	(0.022)	e_{30}	0.903 ***	(0.028)	ξ_{30}	0.879 ***	(0.275)	β_{30}	0.045	(0.068)			

[a] *Note*: This table reports the parameter estimates of the DCC-MANIG model. For the stocks in the DJIA index, the returns are computed as 100 times the daily log-price changes. The standard errors, in parenthesis, are obtained by semi-parametric bootstrap. The column labeled "Stocks" contains the symbols of the index components in the same order as they enter the returns vector. *** denotes significance at the 1 percent level; ** denotes significance at the 5 percent level; * denotes significance at 10 percent level. See the text for further details.

hypothesis:

$$\mathcal{H}_0 : \nu_{i,t} \sim \mathcal{N}(0,1),\ \forall i = 1,\ldots,30,$$

against the alternative that the ν_{it}'s are distributed as mutually independent NIG variables. To do so, we use the fact that the standard normal distribution arises as a limiting case of the NIG distribution, as the parameter γ_i goes to infinity or equivalently $\xi_i = 1/\gamma_i \to 0^+$. For testing purposes, we employ the following Wald statistic

$$\frac{\left(\sum_{i=1}^{N} \widehat{\xi}_i\right)^2}{\mathrm{Var}(\sum_{i=1}^{N} \widehat{\xi}_i)}. \tag{21}$$

Under the null, the asymptotic distribution of this test statistic will be affected by the non-regularity that the parameters defining $\mathcal{H}_0$ lie at the boundary of the admissible parameter space. According to Fiorentini *et al.* (2003), it follows from the fact that the estimator $\widehat{\xi}_i$ cannot be negative that, under the null, $\sqrt{T}\sum_{i=1}^{N} \widehat{\xi}_i$ will converge in distribution to a normal variable which is truncated at zero. As a result, the probability that (21) exceeds a given critical value under $\mathcal{H}_0$ will be just half of the probability that a $\chi^2(1)$ distributed variable does so. On the basis of the parameter estimates in Table 2 and the associated bootstrap covariance matrix, the obtained value of the above Wald statistic equals 157.6. Thus, the null of normality can be easily rejected for the standardized and orthogonalized returns, $\nu_{1,t},\ldots,\nu_{30,t}$.

In what concerns the coefficients of asymmetry in the NIG distributions, we see that most of the estimated β_i's are not significantly different from zero. At the 5 percent level, the estimates of β_4, β_7 and β_{14} are found to be significant, while β_6 is estimated to be negative and only significant at the 10 percent level.

4.2 Forecasts

To obtain the desired forecasts, we need to solve the minimum problem in (1). Even when the exact distribution is known, finding the solution may be cumbersome, so we resort to Monte Carlo methods to obtain optimal forecasts. I.e., draw W pseudo-random numbers $\mathbf{y}_t^w$, $w = 1,2,\ldots,W$, from the conditional distribution of $\mathbf{y}_t$. Then, the optimal forecast is given by

$$\widehat{\mathbf{y}}_t = \arg\min_{\mathbf{y}^*} \frac{1}{W}\sum_{w=1}^{W} \mathcal{L}\left(\mathbf{y}_t^w - \mathbf{y}^*\right) = \arg\min_{\mathbf{y}^*} \mathcal{Q}\left(\mathbf{y}^*\right). \tag{22}$$

Since we use a model-based estimate for the conditional distribution of $\mathbf{y}_t$, this is nothing else than a parametric bootstrap procedure. We set $W = 50\,000$ for obtaining optimal point forecasts from Equation (22) to make sure that sampling variability is negligible.

We use the BFGS algorithm with numerical gradient to solve (22). If N is very large, however, one needs a larger number of draws W, which may lead to time-consuming computation due to the sheer amount of replications. For such cases, it may be useful to have analytical expressions for the gradient and the Hessian of the target function in (22). They are in fact of rather simple structure for loss functions as proposed in Section 2, $\mathcal{L}(\mathbf{u}) = \sum_{i=1}^{M} \mathcal{L}_i^* \left(\sum_{k=1}^{N} a_{ik} u_k\right)$:

$$\begin{aligned}
\frac{\partial \mathcal{Q}}{\partial y_l^*} &= -\frac{1}{W} \sum_{w=1}^{W} \sum_{i=1}^{M} a_{il} \frac{\partial \mathcal{L}_i^*}{\partial u} \left(\sum_{k=1}^{N} a_{ik} \left(y_{tk}^w - y_k^* \right) \right) \\
\frac{\partial^2 \mathcal{Q}}{\partial y_l^* \partial y_{l'}^*} &= \frac{1}{W} \sum_{w=1}^{W} \sum_{i=1}^{M} a_{il} a_{il'} \frac{\partial^2 \mathcal{L}_i^*}{\partial u^2} \left(\sum_{k=1}^{N} a_{ik} \left(y_{tk}^w - y_k^* \right) \right).
\end{aligned}$$

As is often the case, the gradient and the Hessian have simplified expressions in matrix notation. Denoting $\overline{\mathcal{L}^{(1)}}$ the M-dimensional column vector having as i^{th} entry $\frac{1}{W} \sum_{w=1}^{W} \frac{\partial \mathcal{L}_i^*}{\partial u} \left(\sum_{k=1}^{N} a_{ik} \left(y_{tk}^w - y_k^* \right) \right)$, and $\overline{\mathcal{L}^{(2)}}$ the $M \times M$ diagonal matrix having as i^{th} entry on the main diagonal $\frac{1}{W} \sum_{w=1}^{W} \frac{\partial^2 \mathcal{L}_i^*}{\partial u^2} \left(\sum_{k=1}^{N} a_{ik} \left(y_{tk}^w - y_k^* \right) \right)$, we obtain

$$\begin{aligned}
\frac{\partial \mathcal{Q}}{\partial \mathbf{y}^*} &= -A' \, \overline{\mathcal{L}^{(1)}} \\
\frac{\partial^2 \mathcal{Q}}{\partial \mathbf{y}^* \left(\partial \mathbf{y}^* \right)'} &= A' \, \overline{\mathcal{L}^{(2)}} \, A.
\end{aligned}$$

The vector $\overline{\mathcal{L}^{(1)}}$ can be interpreted as the mean marginal loss at $\mathbf{y}^*$, and $\overline{\mathcal{L}^{(2)}}$ as the mean loss curvature at $\mathbf{y}^*$. [10]

The results of the one-step-ahead predictions for the out-of-sample period are given in Figures 2 through 6. For each stock, we plot the MSE-optimal forecast, and the point forecasts resulting from using the following loss functions: $\alpha = 0.2$ and $p = 1, 2$ with $A = I_N$, $A = A_1$ where $a_{ij,1} = 1(i \leq j)$, and $A = A_2 = A_1'$. These are the 30-dimensional counterparts of the bivariate loss functions whose contour lines are plotted in Figure 1.

Note how, due to asymmetry of the loss functions, the optimal forecasts under asymmetric loss functions are negative: negative forecast errors incur higher costs than positive ones, and thus the optimal forecast is shifted downwards. [11]

[10] By suitably approximating the second derivative of the univariate loss functions $\mathcal{L}_i^*$ (as suggested by Demetrescu, 2006), one can even formulate the solution of (22) as an iteratively reweighted least squares scheme.

[11] Choosing a parameter α larger than 0.5 would reverse the situation and lead to positive optimal forecasts.

Moreover, the forecasts for any given stock return differ when loss interactions are allowed for, although they are based on the same information.

5 Concluding remarks

We provided multivariate forecasts of 30 Dow Jones stock returns series under a general, multivariate loss function.

While properties of optimal point and interval forecasts have straightforward extensions to the multivariate case, it is more difficult to find a multivariate loss function suitable for the application at hand. We suggested and discussed defining properties of multivariate loss functions, and proposed a class of easy-to-handle multivariate loss functions based on combinations of univariate ones.

To estimate the forecast distributions of daily returns of 30 DJIA stocks, from which the optimal forecasts were derived, we used a DCC multivariate GARCH model with innovations following a multivariate affine generalized hyperbolic distribution. The model allows for a large number of components, and is flexible enough to capture non-ellipticity and tail dependence. Moreover, the number of parameters to be estimated simultaneously does not depend on the number of series to be jointly forecast.

In samples typical for financial time series, the multi-step quasi-ML estimators for our non-Gaussian model are found to behave similar to the multi-step estimators for the Gaussian DCC data generating process. We fitted our model with daily data observed over six years, and used a parametric bootstrap procedure to derive point forecasts under a number of different multivariate loss functions of the type we proposed.

Acknowledgements

The authors would like to thank Luc Bauwens, Nikolaus Hautsch, and Peter Winker for very helpful suggestions and comments. The research for this paper was carried out while the second author was a Max Weber Fellow at the European University Institute in Florence, whose hospitality is gratefully acknowledged.

Appendix

Proof of Proposition 1

The proof can be carried out with elementary arguments. Consider first the situation where $A = I_N$. Choose an arbitrary $\mathbf{u}_0 = (u_{10}, \ldots, u_{N0}) \in \mathbb{R}^N$, and note that $\mathcal{L}(u) = \sum_{k=1}^{N} \mathcal{L}^*(u \cdot u_{k0})$. Assume $u > 0$; increasing u will increase the magnitude of each argument $u \cdot u_{k0}$ and, due to properties of univariate loss functions, will increase each $\mathcal{L}^*(u \cdot u_{k0})$. Thus $\mathcal{L}(u)$ is itself increasing for positive u. Proceed analogously to show $\mathcal{L}(u)$ is decreasing for $u < 0$. This also implies $u = 0$ to minimize $\mathcal{L}(u)$. Extending this to the case $A \neq I_N$, where $\operatorname{rank} A \geq N$, is straightforward, since $A(u\mathbf{u}_0) = u(A\mathbf{u}_0)$. A and $\mathbf{u}_0$ being fixed, the same argumentation applies for $\mathbf{u}_0^* = A\mathbf{u}_0$ and the desired result is established (continuity of $\mathcal{L}(u)$ being obvious).

References

Aas K., I.K. Haff, and X.K. Dimakos (2006), Risk Estimation Using the Multivariate Normal Inverse Gaussian Distribution. *Journal of Risk* **8**, 39-60.

Alp, T. (2007), A New Multivariate Markov Switching Model of Changes in the Conditional Dependence Between Financial Time Series. *Working Paper*.

Barndorff-Nielsen, O.E. and R. Stelzer (2005), Absolute Moments of Generalized Hyperbolic Distributions and Approximate Scaling of Normal Inverse Gaussian Lévy Processes. *Scandinavian Journal of Statistics* **32**, 617-637.

Bauwens L. and S. Laurent (2005), A New Class of Multivariate Skew Densities, with Application to Generalized Autoregressive Conditional Heteroscedasticity Models. *Journal of Business and Economic Statistics* **23**, 346-353.

Billio, M., M. Caporin, and M. Gobbo (2006), Flexible Dynamic Conditional Correlation Multivariate GARCH Models for Asset Allocation. *Applied Financial Economics Letters* **2**, 123-130.

Bollerslev, T. (1990), Modelling the Coherence in Short-Run Nominal Exchange Rates: A multivariate generalized ARCH model. *The Review of Economics and Statistics* **72**, 498-505.

Christoffersen, P.F. and F.X. Diebold (1996), Further Results on Forecasting and Model Selection Under Asymmetric Loss. *Journal of Applied Econometrics* **11**, 561-571.

Christoffersen, P.F. and F.X. Diebold (1997), Optimal Prediction Under Asymmetric Loss. *Econometric Theory* **13**, 808-817.

Demetrescu, M. (2006), An Extension of the Gauss-Newton Algorithm for Estimation under Asymmetric Loss. *Computational Statistics & Data Analysis* **50**, 379-401.

Demetrescu, M. (2007), Optimal Forecast Intervals Under Asymmetric Loss. *Journal of Forecasting* **26**, 227-238.

Doornik, J. (2006), *Object-oriented matrix programming using Ox, 5th edition.* Timberlake Consultants Press, London.

Efron, B. and R. Tibshirani (1986), Bootstrap Methods for Standard Errors, Confidence Intervals, and Other Measures of Statistical Accuracy. *Statistical Science* **1**, 54-75.

Elliott, G., I. Komunjer, and A. Timmermann (2005), Estimation and Testing of Forecast Rationality Under Flexible Loss. *Review of Economic Studies* **72**, 1107-1125.

Elliott, G. and A. Timmermann (2004), Optimal Forecast Combinations Under General Loss Functions and Forecast Error Distributions. *Journal of Econometrics* **122**, 47-79.

Engle, R.F. (2002), Dynamic Conditional Correlation: A Simple Class of Multivariate Generalized Autiregressive Conditional Heteroscedasticity Models. *Journal of Business and Economic Statistics* **20**, 339-350.

Engle, R.F. and J. Mezrich (1996), GARCH for Groups. *Risk* **9**, 36-40.

Engle, R.F. and K. Sheppard (2001), Theoretical and Empirical Properties of Dynamic Conditional Correlation MVGARCH. Working Paper No. 2001-15, University of California, San Diego.

Fiorentini, G., E. Sentana, and G. Calzolari (2003), Maximum Likelihood Estimation and Inference in Multivariate Conditionally Heteroscedastic Dynamic Regression Models With Student t Innovations. *Journal of Business and Economic Statistics* **21**, 532-546.

Granger, C.W.J. (1969), Prediction With a Generalized Cost of Error Function. *Operational Research Quarterly* **20**, 451-468.

Granger, C.W.J. (1999), Outline of Forecast Theory Using Generalized Cost Functions. *Spanish Economic Review* **1**, 161-173.

Granger, C.W.J. and M.J. Machina (2006), Forecasting and Decision Theory. In Elliott, G., C.W.J. Granger, and A. Timmermann (Eds.), *Handbook of Economic Forecasting, Volume I*, North-Holland, Amsterdam, 81-98.

Gonzáles-Rivera, G., T.-H. Lee and E. Yoldas (2007), Optimality of the RiskMetrics VaR Model. *Finance Research Letters* **4**, 137-145.

Lee, S.-W. and B.E. Hansen (1994), Asymptotic Theory for the GARCH(1,1) Quasi-Maximum Likelihood Estimator. *Econometric Theory* **10**, 29-52.

Michael, J.R., W.R. Schucany, and R.W. Haas (1976), Generating Random Variates Using Transformations with Multiple Roots. *The American Statistician* **30**, 88-90.

Nagahara, Y. (2004), A Method of Simulating Multivariate Nonnormal Distributions by the Pearson Distribution System and Estimation. *Computational Statistics & Data Analysis* **47**, 1-29.

Newey, W.K. and D.L. McFadden (1994), Large Sample Estimation and Hypothesis Testing. In Engle, R.F. and D.L. McFadden (Eds.), *Handbook of Econometrics, Volume IV*, Elsevier Science, 2111-2245.

Patton, A.J. and A. Timmermann (2007), Properties of Optimal Forecasts

Under Asymmetric Loss and Nonlinearity. *Journal of Econometrics* **140**, 884-918.

Schmidt, R., T. Hrycej, and E. Stützle (2006), Multivariate Distribution Models With Generalized Hyperbolic Margins. *Computational Statistics & Data Analysis* **50**, 2065-2096.

Weiss, A.A. (1996), Estimating Time Series Models Using the Relevant Cost Function. *Journal of Applied Econometrics* **11**, 539-560.

Weiss, A.A. and A.P. Andersen (1984), Estimating Time Series Models Using the Relevant Forecast Evaluation Criterion. *Journal of the Royal Statistical Society A* **147**, 484-487.

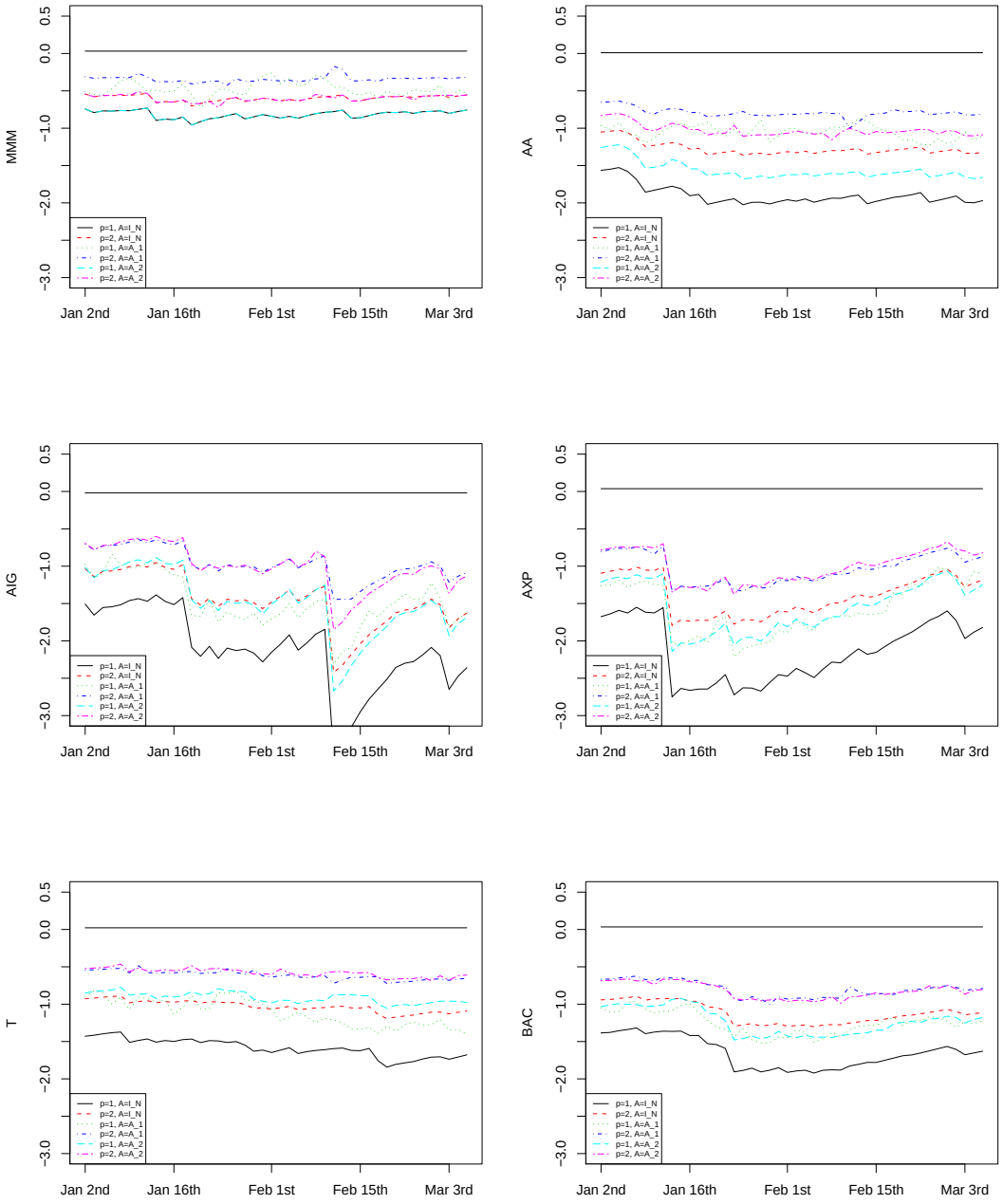


Fig. 2. Plots of stock returns forecasts (in %) using several multivariate loss functions

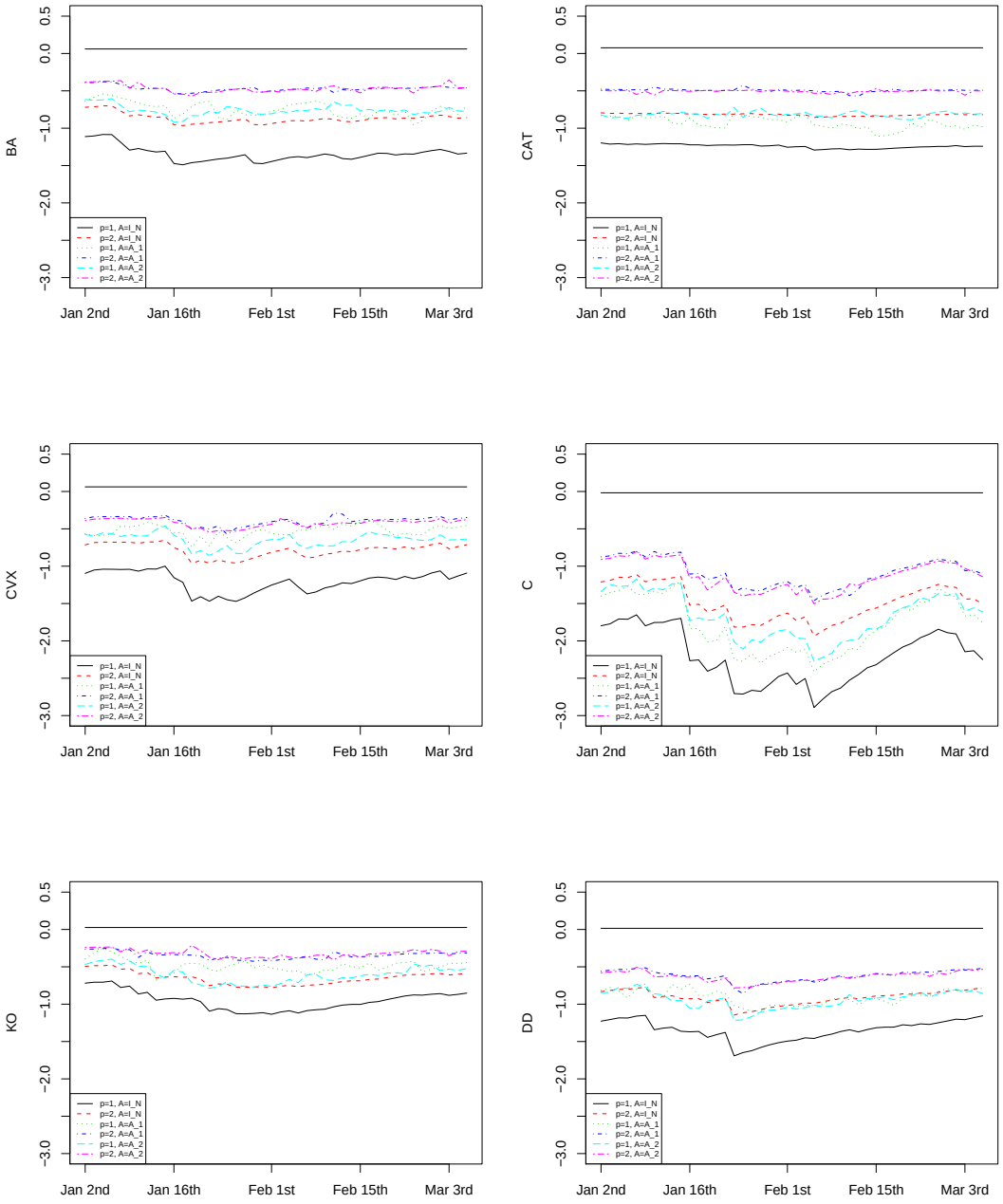


Fig. 3. Plots of stock returns forecasts (in %) using several multivariate loss functions

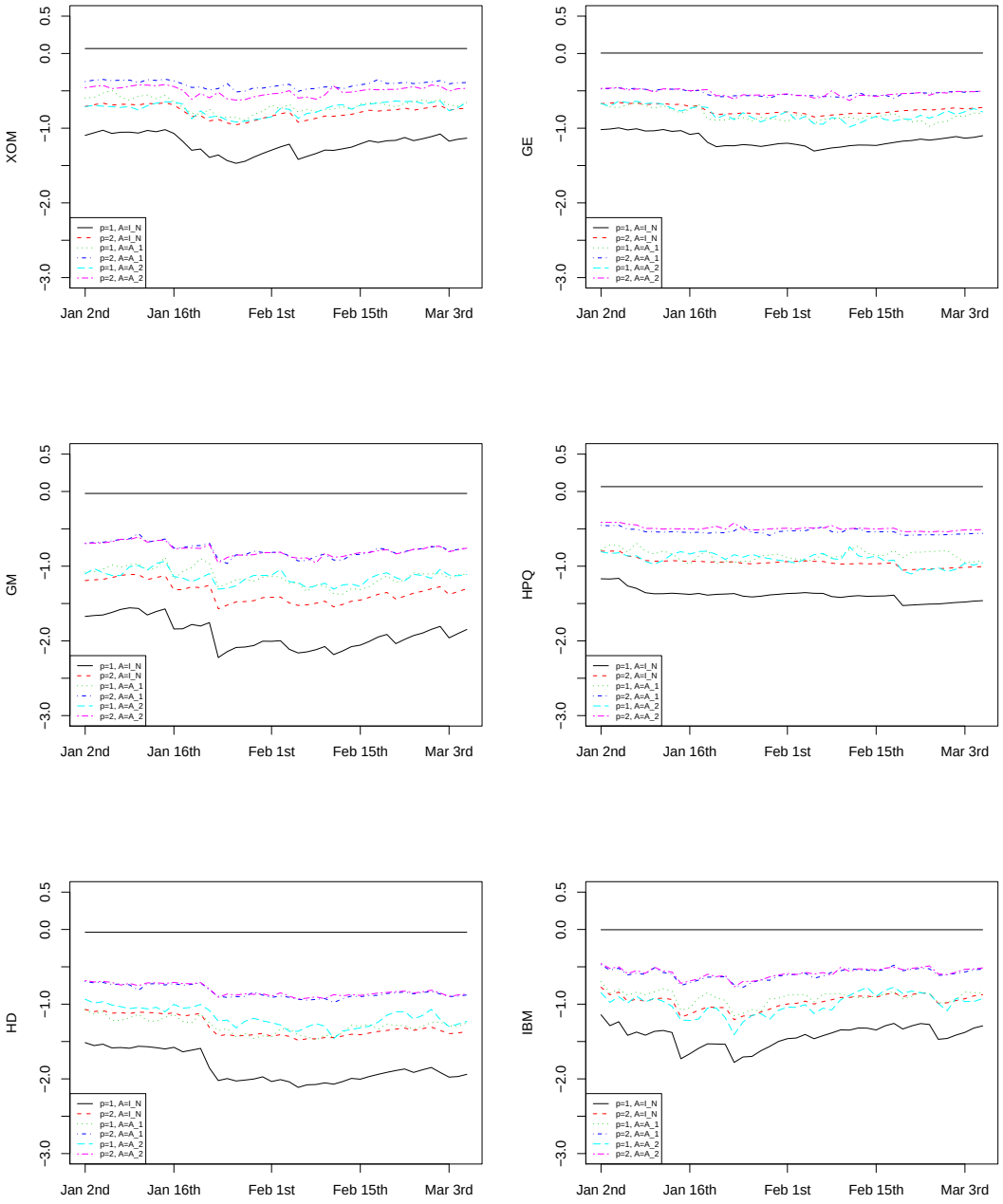


Fig. 4. Plots of stock returns forecasts (in %) using several multivariate loss functions

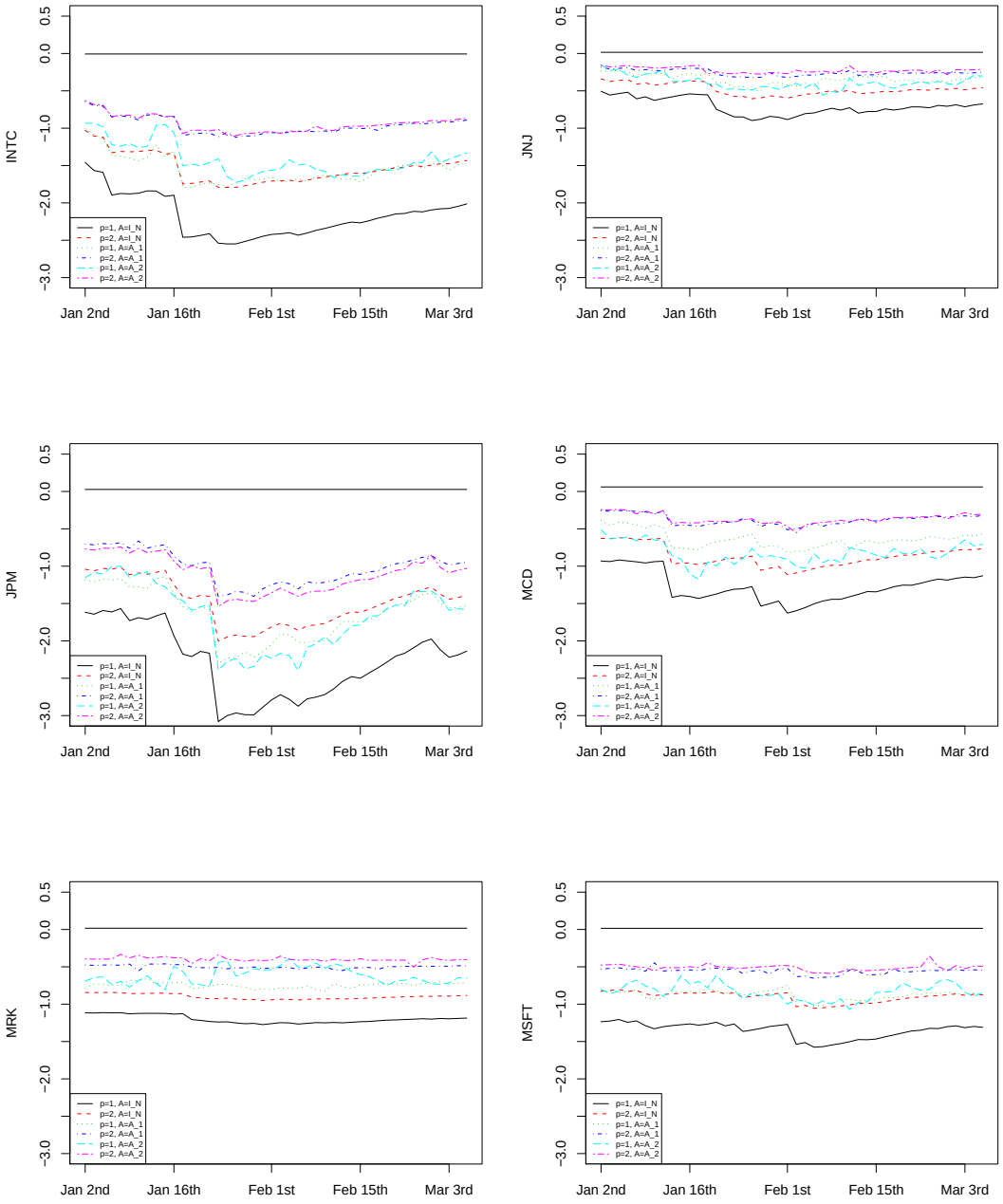


Fig. 5. Plots of stock returns forecasts (in %) using several multivariate loss functions

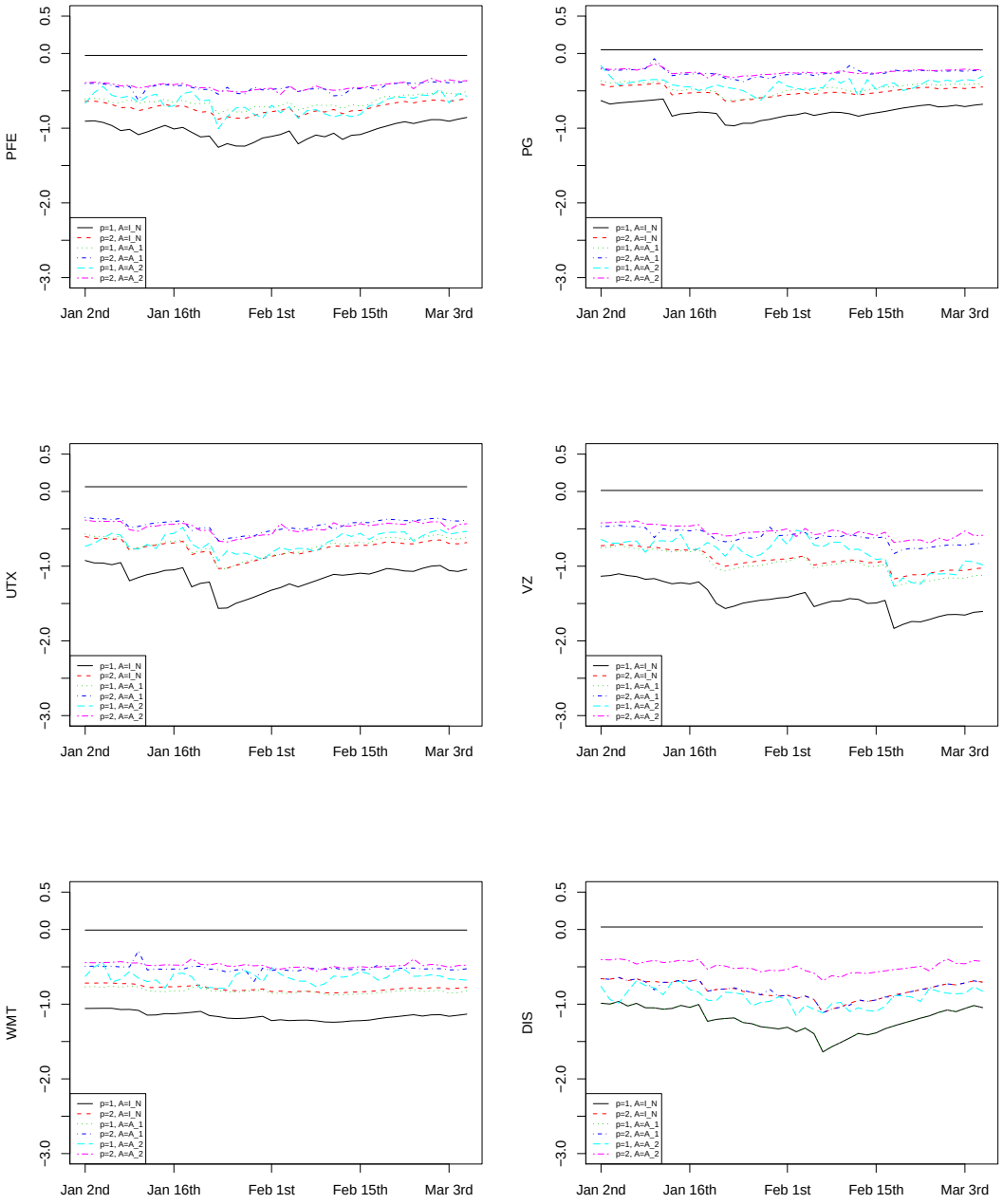


Fig. 6. Plots of stock returns forecasts (in %) using several multivariate loss functions