Research Article

Stanislav Anatolyev* and Filip Staněk

Unrestricted, Restricted, and Regularized Models for Forecasting Multivariate Volatility

https://doi.org/10.1515/sample-YYYY-XXXX
Received Month DD, YYYY; revised Month DD, YYYY; accepted Month DD, YYYY

Abstract: We perform an extensive investigation of different specifications of the BEKK-type multivariate volatility models for a moderate number of assets, focusing on how the degree of parametrization affects forecasting performance. Because the unrestricted specification may be too generously parameterized, often one imposes restrictions on coefficient matrices constraining them to have a diagonal or even scalar structure. We frame all three model variations (full, diagonal, scalar) as special cases of a ridge-type regularized estimator, where the off-diagonal elements are shrunk towards zero and the diagonal elements are shrunk towards homogeneity.

Our forecasting experiments with BEKK-type Conditional Autoregressive Wishart model for realized volatility confirm the superiority of the more parsimonious scalar and diagonal model variations. Even though sometimes a moderate degree of regularization of the diagonal and off-diagonal parameters may be beneficial for forecasting performance, it does not regularly lead to tangible performance improvements irrespective of how precise is tuning of regularization intensity. Additionally, our results highlight the crucial importance of frequent model re-estimation in improving the forecast precision, and, perhaps paradoxically, a slight advantage of shorter estimation windows compared to longer windows.

Keywords: multivariate volatility, covariance matrix, BEKK, CAW, regularization, ridge.

1 Introduction

The knowledge of a covariance matrix of multivariate returns distribution is essential for many tasks such as portfolio allocation, risk management, derivative pricing, analysis of financial contagion, and so on. This has led to development of a variety of models for forecasting covariance matrices such as the constant conditional correlation model (CCC) of Bollerslev (1990), the dynamic conditional correlation model (DCC) of Engle (2002), the BEKK model (Engle and Kroner, 1995) and extensions thereof. More recently, with the machinery allowing the estimation of low frequency volatility from high frequency data (Andersen et al., 2003; Barndorff-Nielsen and Shephard, 2004), ideas behind these models were translated to realized covariance matrices. The Conditional Autoregressive Wishart model (CAW henceforth) of Golosnoy, Gribisch, and Liesenfeld (2012) utilizes directly realized volatilities and co-volatilities and models them via a BEKK-style dynamic equation. Among extensions of BEKK/CAW are asymmetric BEKK (Caporin and McAleer, 2014), proximity-based structured GARCH (Caporin and Paruolo, 2015), threshold CAW (Anatolyev and Kobotaev, 2018), and others.

The dynamic equation in these models is constructed to simultaneously attain two, somewhat conflicting objectives – to accurately capture both temporal and cross-sectional dependencies in return (co-)volatilities, and to maintain a reasonable model parsimony. In this regard, both the BEKK and CAW models are available

***Corresponding author: Stanislav Anatolyev,** CERGE-EI, Politickych vězňů 7, 11121 Prague 1, Czech Republic, e-mail: stanislav.anatolyev@cerge-ei.cz
Filip Staněk, CERGE-EI, Politickych vězňů 7, 11121 Prague 1, Czech Republic, e-mail: filip.stanek@cerge-ei.cz

in three variations of different complexity: a full BEKK/CAW where the law of motion is parameterized by unrestricted parameter matrices, a diagonal BEKK/CAW where these matrices are set to be diagonal, and a scalar BEKK/CAW where these matrices are identity matrices multiplied by scalars. While the full model offers highest flexibility in capturing the data generating process (DGP), it suffers from a curse of dimensionality as the number of parameters determining the law of motion increases at the rate $O(n^2)$ in the number of assets n, which frequently results in imprecise estimation of model parameters and, as a consequence, in poor out-of-sample predictive performance. Due to this demerit, researchers frequently opt for the diagonal or even the scalar model (see e.g.: Caporin and McAleer, 2012; Laurent, Rombouts, and Violante, 2012; Zhipeng and Shenghong, 2018; Zolfaghari, Ghoddusi, and Faghihian, 2020) rather than the full model.

In this article, we address the curse of dimensionality present in BEKK/CAW models by allowing for a smooth transition among differently parameterized models. Namely, we frame all three existing model variations – full, diagonal, scalar – as special cases of a regularized full model. The regularized estimator applies a combination of the standard ridge regularization towards zero (Hoerl and Kennard, 1970) that drives the full model towards the diagonal model, and the ridge regularization towards homogeneity (Anatolyev, 2020) that drives the diagonal model towards the scalar model. Thus, the regularized estimator naturally nests all three benchmarks – scalar, diagonal, full – and is hence capable of optimally selecting among them, or between any of their combinations. This allows us to assess the optimal degree of cross-sectional dependence or non-homogeneity that helps forecasting performance.

We perform an extensive battery of out-of-sample forecast evaluations on Noureldin, Shephard, and Sheppard (2012) data-set of realized stock market co-volatilities of up to ten assets. We trace the influence of a number of assets n, as well as other factors affecting forecasting performance, such as a length of the estimation window and the recency of estimated coefficients relative to the forecasting period. In addition, we analyze the in-sample fit to assess the degree, to which additional flexibility of the diagonal and full models helps account for the volatility dynamics.

Our experiments confirm the general superiority of more restricted models. The performance of the full model deteriorates for higher n, while among the scalar, diagonal and optimally regularized models, the diagonal model seems to be preferred, thought the evidence is noisy and the margin is small. The regularization does not seem to bring perceptible improvements indicating that cross-sectional dependencies are of limited relevance. This can be attributed to a need to tune the regularization intensities, but even in experiments with non-feasible optimal regularization, one can see that the maximal achievable gains from regularization tend to be below 1%. Additionally, we observe that increasing the length of the estimation window does not translate to more precise predictions and that forecasting performance rapidly deteriorates as we increase the distance between the forecasted period and the window on which parameters are estimated. The superiority of more parsimonious scalar and diagonal models is also confirmed by analysis of specification tests. The additional flexibility of the full model delivers only a very modest reduction in correlatedness of the transformed residuals.

The remainder of the article is structured as follows. Section 2 describes the canonical models as well as the regularized estimator. Section 3 presents evaluation of both forecasting quality and in-sample fit, using a real data-set. Section 4 concludes. The Appendix contains tables and figures with supplementary results.

2 Methodology

2.1 Canonical Models

Let us consider n assets living through time periods $t = 1, \dots, T$, and let $\mathcal{F}_t$ denote observable information at t. The BEKK model (Engle and Kroner, 1995) describes evolution of conditional second moments of the n-vector of returns, while the CAW model (Golosnoy, Gribisch, and Liesenfeld, 2012) describes evolution of its conditional first moments of the $n \times n$-matrix of realized co-volatilities. However, the structure of

their dynamic equations for the object of interest is the same. This object of interest is a matrix $R_t \in \mathbb{R}^{n^2}$ that represents, in the case of BEKK, the outer product of demeaned returns, and in the case of CAW, the realized co-volatility matrix.

The BEKK/CAW model postulates the following law of motion for the conditional expectation of R_t denoted $S_t = \mathbb{E}[R_t|\mathcal{F}_{t-1}]$. The canonical BEKK$(p,q)$/CAW$(p,q)$ model reads

$$S_t = CC^\top + \sum_{i=1}^{q} A_i R_{t-i} A_i^\top + \sum_{i=1}^{p} B_i S_{t-i} B_i^\top. \tag{1}$$

Here, in the case of the BEKK model, $R_t = (r_t - \mu)(r_t - \mu)^\top$, and $r_t|\mathcal{F}_{t-1} \sim N(\mu, S_t)$ is a vector of n returns. In the case of the CAW model, R_t is a realized volatility matrix computed from high frequency data, $R_t|\mathcal{F}_{t-1} \sim W_n(v, S_t/v)$, where $W_n(\nu, \Sigma)$ represents an n-dimensional Wishart distribution with ν degrees of freedom and a scale matrix Σ.

The lower-triangular parameter matrix $C \in \mathbb{R}^{n^2}$ and general parameter matrices $A_i \in \mathbb{R}^{n^2}$ and $B_i \in \mathbb{R}^{n^2}$ determine the law of motion for volatility and are to be estimated along with the other parameters.[1] This specification offers several advantages. First, it guarantees, by construction, the positive semidefinitness and symmetry of volatility predictions S_t without requiring parameter restrictions on matrices $\left\{C, \{A_i\}_{i=1}^q, \{B_i\}_{i=1}^p\right\}$. Second, in its most general form, it allows one to model various dependencies between the current volatility and past innovations or past volatility predictions across different assets through the off-diagonal elements of matrices A_i or B_i, respectively.

In practice, however, restrictions on parameters are frequently made in order to reduce the estimation noise at the cost of more probable misspecification. The most commonly used model variations are the following (the abbreviations corresponding to the CAW class):

1. fCAW: the full model with general parameter matrices: $A_i \in \mathbb{R}^{n^2}$ for $1 \le i \le q$ and $B_i \in \mathbb{R}^{n^2}$ for $1 \le i \le p$, resulting in $n^2(p+q) = O(n^2)$ parameters in matrices A and B.
2. dCAW: the diagonal model with zero restrictions on the off-diagonal elements: $A_i = \mathrm{dg}\{a_{i,1}, a_{i,2}, \ldots, a_{i,n}\}$ for $1 \le i \le q$ and $B_i = \mathrm{dg}\{b_{i,1}, b_{i,2}, \ldots, b_{i,n}\}$ for $1 \le i \le p$, resulting in $n(p+q) = O(n)$ parameters in matrices A and B.
3. sCAW: the scalar model with zero restrictions on the off-diagonal elements and equality restrictions across the diagonal elements: $A_i = a_i I_n$ for $1 \le i \le q$ and $B_i = b_i I_n$ for $1 \le i \le p$, resulting in $p + q = O(1)$ parameters in matrices A and B.

While the full BEKK/CAW offers highest flexibility, it is rarely used in practice due to its excessive parametrization – for example, the full CAW$(1,1)$ model with 10 assets requires estimation of 256 parameters. Instead, researchers frequently opt for the diagonal or even scalar models.

2.2 Regularized Model

2.2.1 Maximum Likelihood Estimation

The parameters of equation (1) are estimated by the method of maximum likelihood. Let us denote the model parameters by θ, and the log-likelihood for one observation by $\ell\ell_t$. Then, $\theta = \left\{C, \{A_i\}_{i=1}^q, \{B_i\}_{i=1}^p, \mu\right\}$ for the BEKK, and $\theta = \left\{C, \{A_i\}_{i=1}^q, \{B_i\}_{i=1}^p, v\right\}$ for the CAW. The likelihood for observation t is

$$\ell\ell_t = \log\left((2\pi)^{-\frac{k}{2}} |S_t|^{-\frac{1}{2}}\right) - \frac{1}{2}\mathrm{tr}\left(S_t^{-1} R_t\right)$$

1 We maintain the standard assumption that the intercept matrix C is populated by $0.5n(n+1)$ parameters (Engle and Kroner, 1995; Golosnoy, Gribisch, and Liesenfeld, 2012). For unique parameter identification, it is conventional and convenient to restrict all the diagonal elements of C and the first diagonal elements of all matrices A_i and B_i to be positive (Engle and Kroner, 1995).

for the BEKK, and

$$\ell\ell_t = \log\left(\frac{|S_t v^{-1}|^{-\frac{v}{2}}|R_t|^{\frac{v-n-1}{2}}}{2^{\frac{vn}{2}}\pi^{\frac{n(n-1)}{4}}\prod_{i=1}^{n}\Gamma\left(\frac{v+1-i}{2}\right)}\right) - \frac{1}{2}\mathrm{tr}\left(vS_t^{-1}R_t\right)$$

for the CAW. The maximum likelihood estimator solves the optimization problem

$$\widehat{\theta}_{ML} = \arg\max_{\theta}\sum_{t=1}^{T}\ell\ell_t \tag{2}$$

subject to the evolution equation (1) and possibly additional parameter restrictions listed in subsection 2.1. The fCAW/fBEKK, dCAW/dBEKK and sCAW/sBEKK estimates emerge depending on which additional constraints are imposed.

2.2.2 Penalized Estimation

Regularization augments the log-likelihood in the optimization problem (2) by two ridge-type (i.e., relative to the L_2-norm) penalty terms.[2] The first penalty punishes for deviations of the off-diagonal elements from the zero value, providing regularization of fBEKK/fCAW towards dBEKK/dCAW corresponding to classical "ridging towards zero" (Hoerl and Kennard, 1970):

$$\tau_f = \sum_{i=1}^{q}\sum_{j=1}^{n}\sum_{k\neq j}^{n}A_{i,j,k}^2 + \sum_{i=1}^{p}\sum_{j=1}^{n}\sum_{k\neq j}^{n}B_{i,j,k}^2. \tag{3}$$

The second penalty punishes for deviations of the diagonal elements from the common value, providing regularization of dBEKK/dCAW towards sBEKK/sCAW corresponding to "ridging towards homogeneity" (Anatolyev, 2020):

$$\tau_d = \sum_{i=1}^{q}\sum_{j=1}^{n}\left(A_{i,j,j} - \frac{1}{n}\sum_{k=1}^{n}A_{i,k,k}\right)^2 + \sum_{i=1}^{p}\sum_{j=1}^{n}\left(B_{i,j,j} - \frac{1}{n}\sum_{k=1}^{n}B_{i,k,k}\right)^2. \tag{4}$$

This results in the following regularized maximum likelihood estimator:

$$\widehat{\theta}_{RML} = \arg\max_{\theta}\sum_{t=1}^{T}\left\{\ell\ell_t - \lambda_f\tau_f - \lambda_d\tau_d\right\} \tag{5}$$

subject to the evolution equation (1). The hyper-parameters λ_f and λ_d control the degree of regularization applied to the off-diagonal and diagonal matrix elements, respectively: λ_f is the intensity of ridging of fBEKK/fCAW towards dBEKK/dCAW, and λ_d is the intensity of ridging of dBEKK/dCAW towards sBEKK/sCAW.

This specification naturally nests the full BEKK/CAW ($\lambda_d = 0$, $\lambda_f = 0$), the diagonal BEKK/CAW ($\lambda_d = 0$, $\lambda_f \to \infty$), and the scalar BEKK/CAW ($\lambda_d \to \infty$, $\lambda_f \to \infty$). Apart from these three extreme cases, it also allows for intermediate states. We abbreviate this model and technique corresponding to the BEKK/CAW class by rBEKK/rCAW, where 'r' stands for 'regularized.'

2 In principle, other regularization norms are possible; however, they may not provide such natural framing of the three basic models within the regularized version, or may not be so attractive from the viewpoint of emerging restrictions. The popular L_1-norm, as an obvious alternative, would lead to setting most of the off-diagonal elements to zero shutting down dynamics for some realized covariances, while lassoing the diagonal elements would result in unmotivated equality of some elements and non-equality of others.

2.2.3 Feasible Penalization

In practice, the hyper-parameters λ_d and λ_f are not known ex-ante and need to be tuned. We follow the usual tradition and select them a via fixed scheme time-series cross-validation (see Clark and McCracken, 2013): given a pair of candidate values of hyper-parameters $\lambda_d \in \Lambda_d = \{\lambda_{d,1}, \lambda_{d,2}, \dots, \lambda_{d,k_d}\}$ and $\lambda_f \in \Lambda_f = \{\lambda_{f,1}, \lambda_{f,2}, \dots, \lambda_{f,k_f}\}$, the model is estimated on a window of data $\{1, 2, \dots, T_0\}$, and then evaluated on the remaining $\{T_0 + 1, T_0 + 2, \dots, T\}$ observations using a desired loss function; we use the Stein loss (see Section 3). The optimal values of hyper-parameters are selected so that they minimize the loss incurred in the validation sample $\{T_0 + 1, T_0 + 2, \dots, T\}$.

As computational complexity of large scale numerical optimization unfortunately makes an exhaustive grid search for optimal values of λ_d and λ_f impractical even for moderately sparse sets Λ_d and Λ_f, we opt for a sequential search for the optimal hyper-parameters, starting with Λ_d and only then proceeding to Λ_f. Furthermore, one is able to further reduce the run-time of numerical optimizations by sorting the sets Λ_d and Λ_f in descending order and then proceeding so that the parameters estimated for $\lambda_{d,i}$ (respectively, $\lambda_{f,i}$) form a starting point of numerical optimization for $\lambda_{d,i+1}$ (respectively, $\lambda_{f,i+1}$), effectively tracing the optimal $\widehat{\theta}$ as an approximately continuous function of λ_d (resp. λ_f) rather than performing a search along a long optimization path from a common starting point. This procedure is analogous to the practice of using estimates of more restricted models of BEKK as starting values for more flexible models (Engle and Kroner, 1995) but applied to a gradually changing degree of regularization rather than the extreme models themselves. These measures make the regularized estimator not too much more computationally expensive than the standard full BEKK/CAW.

Moreover, the fact that optimization on a less constrained parameter space utilizes solutions found in a more restricted space as starting points reduces the possibility of converging to a sub-optimal local maximum, an understandable concern given the scale of our optimization problems. This is so because, by design, optimization in the less constrained parameter space cannot ever attain a worse solution than the best solution already found in the more restricted space including that of the most parsimonious sBEKK/sCAW, effectively ruling out a whole range of potential sub-optimal local maxima.[3]

3 Empirical Evaluation

3.1 Evaluation Design

Throughout this section, we restrict our attention to volatility modeling via the CAW class of models for realized covariance matrices. The results for BEKK are qualitatively similar; these results are available upon request. The main reason of focusing on the CAW class is a wide availability of high frequency return data for modeling realized variances, whose use generally improves forecast precision relative to GARCH-type models (see, e.g., Andersen et al., 2003; Golosnoy, Gribisch, and Liesenfeld, 2012). In addition, the use of models for realized volatility allows higher quality evaluation of forecasts as realized volatility is observable. Furthermore, we set $q = p = 1$ as that represents arguably the most commonly used model specification and because it also reduces computational burden.

Let $\widehat{R}_t$ be a forecast of R_t made in the previous period. The Stein loss (James and Stein, 1961) that we exploit here for forecast evaluation,

$$L_S\big(R_r, \widehat{R}_t\big) = \operatorname{tr}\big(\widehat{R}_t R_r\big) - \log\Big(\det\big(\widehat{R}_t R_r\big)\Big) - n,$$

3 A MATLAB package building upon the MFE toolbox (Sheppard, 2013) implementing the regularized estimator for both BEKK and CAW is available at https://github.com/stanek-fi/RMV.

is coherent with the Wishart likelihood and is robust to volatility measurement in the sense of Patton (2011). Along with the Stein loss, we exploit the Frobenius loss function

$$L_F\big(R_r, \widehat{R}_t\big) = \operatorname{tr}\left(\big(R_r - \widehat{R}_t\big)\big(R_r - \widehat{R}_t\big)^\top\right).$$

This is (along with a very similar Euclidean loss, see Laurent, Rombouts, and Violante, 2012) a multivariate extension of the quadratic loss in the scalar case, and is also robust in the sense of Patton (2011). Among the two, the Stein loss generally exhibits less erratic behavior hence allowing us to draw finer conclusions.[4]

We use the popular realized stock market volatility data-set from Noureldin, Shephard, and Sheppard (2012) covering 10 stocks (BAC, JPM, IBM, MSFT, XOM, AA, AXP, DD, GE, KO) from 2001-02-01 to 2009-12-31. In order to attain maximal external validity and to explore different factors which might affect forecasting performance, we employ the following experimental design, which also reflects on practitioners' decisions to employ training samples of arbitrary sizes. For each of $n \in \{2, 3, \dots, 10\}$, 4 distinct combinations of stocks of size n are randomly selected (except the case $n = 10$ when only one combination is available). For each combination, training windows of sizes $T \in \{2, 2.5, 3, 3.5, 4, 4.5, 5\}$ years[5] are rolled through the sample with increment 0.5 year. For each position of the window, a day ahead forecasts for the following 0.5 year are produced via estimated scalar, diagonal, full, and regularized CAW (recall the acronyms sCAW, dCAW, fCAW and rCAW, respectively). Finally, the quality of the forecasts is measured via the Stein and Frobenius losses, and significance of the observed differences is tested via model confidence sets (MCS) (Hansen et al., 2011). This setting assures a substantial variation in both the estimation and evaluation data as well as in the assets under consideration. In the case of rCAW, the regularization hyper-parameters λ_d and λ_f are selected from $\Lambda_d = \Lambda_f = \{\kappa_1, \cdots \kappa_8\}$, where $\kappa_i = \big(\exp(2(i-1)) - 1\big)/10$ for $i = 1, \dots, 7$ and $\kappa_8 = \infty$,[6] by evaluation on the validation set – the last 0.5 year of data in the training window. The optimization itself is performed via a Newton-type optimizer with constraints accounted for with the use of an active-set method. The optimality tolerance is set to 10^{-6} to minimize the possibility of premature termination in almost flat regions of the objective function. To reduce computational requirements, the BFGS algorithm suitable for large scale problems such as this one is utilized in computation of Hessian updates.

3.2 Forecasting Performance

As can be seen in Figure 1 displaying the average volatility alongside the average performance of individual models and the selected λ_d and λ_f, the data-set spans two volatile periods – the US stock market downturn of 2002 and the US housing crisis – as well a relatively calm period from 2003 to 2007. Clearly, the out-of-sample performance of all models as measured by the Stein loss deteriorates during the volatile housing crisis. This is especially true for the fCAW whose relative performance, as measured by the ratio of out-of-sample Stein loss (with the rCAW being the benchmark) and the average ranking among the models, deteriorates when trained or evaluated on highly volatile periods.

Figure 1 here

This fact is also mirrored by the optimal hyper-parameters λ_d and λ_f selected, as we can see that more stringent regularization is being chosen during the US housing crisis whereas less stringent regularization

4 We report the results for the Stein loss in the main text, and those for the Frobenius loss in the Appendix.

5 Due to space considerations, we choose not to display separate results for estimation windows of non-integer sizes $T \in \{1.5, 2.5, 3.5, 4.5\}$ years, thou they still do enter aggregate computations. Results for these intermediate window lengths are, as expected, intermediate, and are available upon request.

6 This scale was purposefully chosen so that it allows for very mild up to stringent penalties. The cases $\lambda_d = \infty$ or $\lambda_f = \infty$ are implemented via corresponding restrictions on the likelihood maximization problem.

corresponding to lower λ_d (respectively, λ_f) is being chosen whenever dCAW (respectively, fCAW) performs notably well (e.g., during the 2003-2007 tranquil period). This variability over time indicates that possible cross-sectional dependencies might not be stable. Overall, however, we see that complete (κ_8) or very stringent regularization (κ_7) are most frequent for both the diagonal and off-diagonal elements.

Figures 2A and 2B displays in-sample and out-of-sample Stein losses for individual models relative to the loss of rCAW. As to be expected, more parameterized models achieve better in-sample performance, especially pronounced when the estimation window is short (which presumably reflects higher within-window stability). However, the magnitude of the observed differences between individual models is relatively small; for estimation windows larger than or equal to three years, it is generally within 1%. As for the rCAW, which optimally selects the degree of parameterization, the in-sample fit is comparable to that of the dCAW. With regard to the out-of-sample Stein losses, the sCAW, dCAW, and rCAW perform similarly (well within 1%), outperforming the much more parameterized fCAW on average by 1%–6%, depending on a length of the training window.

To be able to discern these relatively small differences between the sCAW, dCAW, and rCAW, Tables 1 and 2 provide a more detailed assessment of the out-of-sample performance broken down to individual combinations of the number of assets n and the training window length T. Table 1 depicts the average ranking of individual models as measured by the Stein loss. The average ranking of all the three models is in the neighborhood of 2.1, though slightly favoring the dCAW over the rCAW and sCAW, with the difference being somewhat more sizable for longer training windows. Table 2 depicts the frequency of rejection via the MCS at the 5% level for individual models, thus focusing on how often the models perform notably badly. According to this measure, the dCAW seems to be a slightly safer option relative to the sCAW or rCAW for situations with more assets and longer estimation windows, though it should be noted that the evidence is again not entirely conclusive. The most flexible fCAW model is, on average, rejected in more than half of the cases.

The fact that rCAW largely fails to outperform dCAW, merely matching its performance, can be attributed to two factors. First, it appears that sCAW and dCAW perform similarly without much gain obtained via optimal selection of λ_d. With respect to λ_f, the performance appears to quickly decline as we move towards less stringent regularization with no apparent intermediate region of improved performance relative to dCAW. In order to see this, we have performed an additional experiment. Figure 3A depicts the performance of sCAW and dCAW relative to the performance of *non-feasible* optimally diagonally regularized dCAW, with λ_d chosen to be *ex-post optimal* in terms of loss. The maximal achievable gains from diagonal regularization are generally below 1%. Furthermore, the potential gains from regularization of off-diagonal elements relative to dCAW are well bellow 0.3% as can be seen from Figure 3B depicting the performance of dCAW and fCAW relative to the performance of the unfeasible optimally off-diagonally regularized fCAW. Clearly, the off-diagonal regularization, no matter how stringent, fails to substantially improve dCAW even when we abstract from tuning the optimal λ_f.

Second, the quickly changing nature of the process (see Figure 1) makes it difficult to choose the optimal λ_d and λ_f and hence reap the already small improvements stemming from regularization as displayed in Figures 3A and 3B.

Figure 2 here

Figure 3 here

Table 1 here

Table 2 here

The factorial design of our experiments also allows us to address questions regarding the optimal length of the estimation window and frequency with which the model needs to be re-estimated. This is especially relevant because researchers, possibly due to computational limitations associated with high dimensional optimization, frequently opt for a fixed forecasting scheme (see eg. Asai, Gupta, and McAleer, 2020; Lucheroni, Boland, and Ragno, 2019), in which case a single estimated model is used throughout the whole out-of-sample period.

With respect to the former question (the optimal length of the estimation window), Figure 4A depicts the out-of-sample Stein losses for different estimation window lengths relative to the loss, which would be achieved if the estimation window length was equal to a benchmark of 2 years for the given out-of-sample part of the data and the combination of assets. In line with expectations, for the fCAW, the ratios of losses are generally below 1 implying that increasing the estimation window length leads to a more precise forecast. However, for the dCAW and especially for the rCAW and sCAW, increasing the estimation window length does not lead to a better forecasting performance; in fact, it somewhat worsens it (in the case of the sCAW, on average by 0.27% per additional year of the estimation window). This counter-intuitive behavior may be attributed to changes of the DGP, which penalize utilization of more distant data points during estimation.

With respect to the latter question (the importance of re-estimation), Figure 4B depicts the average Stein loss for individual out-of-sample observations with different distance from the end of the estimation window relative to the average Stein loss over the whole out-of-sample period.[7] Clearly, one-day-ahead forecasts made for the periods immediately following the estimation window are substantially better compared to day-ahead forecasts made for more distant periods. This effect is sizable averaging to 22% per year, completely dwarfing any differences that are observed among the sCAW, dCAW and rCAW. The effect is stronger for more parameterized models such as the fCAW and for larger numbers of assets n. Again, this indicates that the DGP may be changing, and/or that CAW-type models are likely to merely locally approximate the process rather than to correctly describe it globally.

Figure 4 here

3.3 Specification Tests

Our target is seeking an optimal degree of parameterization among the CAW models with a different structure of A- and B-matrices, which entails deliberate misspecification for the sake of a better forecasting performance. Here, the celebrated in-sample/out-of-sample tradeoff is resolved fully in favor of quality of predictions. Thus, the models under considerations are not expected to describe the true DGP (on top of conceptual meaning of a model as simplification of reality), so that hardly can even the most parameterized model, fCAW, be expected to perfectly describe the data generation in all its complexity. Nonetheless, the assessment of the in-sample performance can still serve a valuable role: comparison of individual models offers insights about poor forecasting performance of the fCAW and incapability of the rCAW to deliver a superior performance relative to the sCAW and dCAW. The more flexible fCAW is supposed to explain the DGP notably better than the least flexible sCAW to justify hundreds of additional parameters. Similarly, allowing for non-zero, albeit shrunk, off-diagonal parameters in the rCAW model is beneficial only to the extent, to which these parameters can genuinely help explain the actual volatility dynamics.

7 The outlier at 2007-02-27 was removed from this analysis, as its value exceeds the sample standard deviation by approximately ten times, heavily skewing the results.

To assess the in-sample fit, we employ the extended Bartlett decomposition method for Wishart processes as proposed in Alfelt, Bodnar, and Tyrcha (2020). In particular, we compute

$$U_t = \left(\underbrace{S_t^{-\frac{1}{2}} R_t \left(S_t^{-\frac{1}{2}}\right)^\top}_{=Q_t} \right)^{\frac{1}{2}}, \tag{6}$$

where Q_t is the re-scaled realized volatility matrix, and U_t is its lower-triangular Cholesky root. Then, we compute the transformed errors

$$e_{t,i,j} = \begin{cases} U_{t,i,j} \text{ for } j < i, \\ \Phi^{-1}\left(F_{\Gamma(\frac{v-i+1}{2},2)}(U_{t,i,j}^2)\right) \text{ for } j = i, \end{cases} \tag{7}$$

and collect them to a single error vector

$$e_t = \left(e_{t,1,1}, ..., e_{t,n,1}, e_{t,2,2}, ..., e_{t,n,n}\right)^\top. \tag{8}$$

As shown in Alfelt, Bodnar, and Tyrcha (2020), if the model is correctly specified, the errors are IID standard normal, i.e., $e_t \sim \mathcal{N}(\mathbf{0}_k, I_k)$ with $k = n(n+1)/2$. This convenient decomposition allows one to separately assess how well CAW models account for different features of the DGP; autocorrelation in e_t signals misspecification of the law of motion for S_t, whereas violation of $e_t \sim \mathcal{N}(\mathbf{0}_k, I_k)$ signals deviations from the assumed Wishart distribution and/or systematic prediction errors.

Table 3 displays average rejection rates across different window locations for selected tests performed on e_t. In this exercise, the estimation window is set to $T = 5$ years to enjoy the largest possible power, and the number of stocks is set to $n = 10$ to maximize a potential difference between sCAW, dCAW, and fCAW. We omit the rCAW results as the regularization is justified only from the out-of-sample perspective. For sCAW, dCAW and fCAW alike, all tests (with the exception of the t-test for $H_0 : \mathbb{E}[e_t] = \mathbf{0}_k$) reject the hypothesis that the data are consistent with the CAW$(1,1)$ model; the e_t exhibits an excess variance, non-normality, and temporal correlations. Considering the long estimation window and thus maximized power of these tests, these findings are not surprising. The CAW models rarely perfectly explains observed data; in-sample errors are often auto-correlated (Alfelt, Bodnar, and Tyrcha, 2020; Golosnoy, Gribisch, and Liesenfeld, 2012), and inconsistent with the assumption of Wishart distribution (Alfelt, Bodnar, and Tyrcha, 2020), with only the unbiasedness test usually exhibiting rejection rates close to 0 (see Alfelt, Bodnar, and Tyrcha, 2020). Moreover, similar results are likely to be observed when considering higher order CAW models; in Golosnoy, Gribisch, and Liesenfeld (2012), increasing the model order from CAW$(1,1)$ to CAW$(3,3)$ leads to a non-rejection of only one additional error auto-correlation (out of 15).

Table 3 here

To better assess the in-sample fit of different models, we also perform tests on individual elements $e_{t,i,j}$ rather than on the whole vector e_t; see Table 4. Despite its 198 additional parameters compared to sCAW, fCAW does not deliver a markedly better in-sample fit. It exhibits rejection rates comparable to those for sCAW, for all the tests with an exception of the univariate Box and Pierce (1970) test, where it offers a modest 8.6% reduction in the rejection rate. This shows that non-zero off-diagonal elements of the parameter matrices A and B generally fail to take care of error dependence. Interestingly however, the errors $e_{t,i,j}$ from the fCAW exhibit a smaller kurtosis relative to those from the sCAW (35.8 vs 42.2), as can be seen in Figure 7 in the Appendix. This is indicative of the fact that the off-diagonal elements of parameter matrices A and B are utilized to better accommodate outliers in the estimation window rather than to genuinely account for evolution of volatility. This also explains poor performance of fCAW and, by extension, also inability of the optimal regularized rCAW to outperform the sCAW and dCAW.

Table 4 here

4 Conclusions

We perform an extensive forecasting experiment examining the performance of the sCAW, dCAW, fCAW, and the regularized version thereof, rCAW, which nests all three via ridge-type regularization towards zero and towards homogeneity. The results confirm the poor predictive performance of the fCAW relative to more restricted models. The performance of the sCAW and dCAW is comparable, slightly favoring the dCAW. The optimal amount of regularization in the rCAW does not seem to bring any tangible improvements in terms of forecasting performance, irrespective of how precise is tuning of regularization intensity. This indicates that the cross-sectional volatility dependence is not a major factor, at least as far as the forecasting performance is concerned.

Further analysis shows that for the sCAW and dCAW, increasing the length of the estimation window typically does not lead to a better forecasting performance; oftentimes, the converse is true. Furthermore, we observe a very quick performance deterioration when one-day-ahead forecast are made using a model estimated on more distant segments of data, indicating possible model misspecification and/or changes in the DGP. Overall, based on the results, we would recommend to perform multivariate volatility forecasting via diagonal variants of volatility models estimated on a short rolling window, to achieve the best forecasting performance.

While we have performed all experiments with the canonical CAW(1,1) model, we conjecture that the tendencies we have discovered also carry over to other models with a similar structure of the dynamic volatility equation – more general CAW and BEKK models, various extensions thereof (Anatolyev and Kobotaev, 2018; Caporin and McAleer, 2014; Caporin and Paruolo, 2015), and extensions with a block structure, these tendencies being applicable for blocks of assets.

Acknowledgment: We thank an anonymous referee for the helpful comments and constructive remarks. Furthermore, we thank the audiences and discussants at the 40th International Symposium on Forecasting, XIt Workshop in Time Series Econometrics, and 5th International Workshop on Financial Markets and Nonlinear Dynamics.

Funding: Czech Science Foundation support under grant 20-28055S and Grantová agentura UK support under grant 264120 are gratefully acknowledged.

References

Alfelt, G., T. Bodnar, and J. Tyrcha 2020. "Goodness-of-Fit Tests for Centralized Wishart Processes." *Communications in Statistics–Theory and Methods* 49 (20): 5060–5090.

Anatolyev, S. 2020. "A Ridge to Homogeneity for Linear Models." *Journal of Statistical Computation and Simulation* 90 (13): 2455–2472.

Anatolyev, S., and N. Kobotaev 2018. "Modeling and Forecasting Realized Covariance Matrices with Accounting for Leverage." *Econometric Reviews* 37 (2): 114–139.

Andersen, T. G., T. Bollerslev, F. X. Diebold, and P. Labys 2003. "Modeling and Forecasting Realized Volatility." *Econometrica* 71 (2): 579–625.

Anderson, T. W., and D. A. Darling 1952. "Asymptotic Theory of Certain "Goodness of Fit" Criteria Based on Stochastic Processes." *Annals of Mathematical Statistics* 23 (2): 193–212.

Asai, M., R. Gupta, and M. McAleer 2020. "Forecasting Volatility and Co-Volatility of Crude Oil and Gold Futures: Effects of Leverage, Jumps, Spillovers, and Geopolitical Risks." *International Journal of Forecasting* 36 (3): 933–948.

Barndorff-Nielsen, O. E., and N. Shephard 2004. "Econometric Analysis of Realized Covariation: High Frequency Based Covariance, Regression, and Correlation in Financial Economics." *Econometrica* 72 (3): 885–925.

Bollerslev, T. 1990. "Modelling the Coherence in Short-Run Nominal Exchange Rates: A Multivariate Generalized ARCH Model." *Review of Economics and Statistics* 72 (3): 498–505.

Box, G. E. P., and D. A. Pierce 1970. "Distribution of Residual Autocorrelations in Autoregressive-Integrated Moving Average Time Series Models." *Journal of American Statistical Association* 65 (332): 1509–1526.

Caporin, M., and M. McAleer 2012. "Do We Really Need Both BEKK and DCC? A Tale of two Multivariate GARCH Models." *Journal of Economic Surveys* 26 (4): 736–751.

Caporin, M., and M. McAleer 2014. "Robust Ranking of Multivariate GARCH Models by Problem Dimension." *Computational Statistics and Data Analysis* 76: 172–185.

Caporin, M., and P. Paruolo 2015. "Proximity-Structured Multivariate Volatility Models." *Econometric Reviews* 34 (5): 559–593.

Clark, T., and M. McCracken 2013. "Advances in Forecast Evaluation." *Handbook of Economic Forecasting* 2: 1107–1201.

Engle, R. F. 2002. "Dynamic Conditional Correlation: A Simple Class of Multivariate Generalized Autoregressive Conditional Heteroskedasticity Models." *Journal of Business & Economic Statistics* 20 (3): 339–350.

Engle, R. F., and K. F. Kroner 1995. "Multivariate Simultaneous Generalized ARCH." *Econometric Theory* 11 (1): 122–150.

Golosnoy, V., B. Gribisch, and R. Liesenfeld 2012. "The Conditional Autoregressive Wishart Model for Multivariate Stock Market Volatility." *Journal of Econometrics* 167 (1): 211–223.

Hansen, P. R., A. Lunde, and J. M. Nason 2011. "The Model Confidence Set." *Econometrica* 79 (2): 453–497.

Hoerl, A. E., and R. W. Kennard 1970. "Ridge Regression: Biased Estimation for Nonorthogonal Problems." *Technometrics* 12 (1): 55–67.

James, W., and C. Stein 1961. "Estimation with Quadratic Loss." *Proceedings of the Fourth Berkeley Symposium on Mathematical Statistics and Probability* 1961: 443–460.

Laurent, S., J. V. Rombouts, and F. Violante 2012. "On the Forecasting Accuracy of Multivariate GARCH Models." *Journal of Applied Econometrics* 27 (6): 934–955.

Lucheroni, C., J. Boland, and C. Ragno 2019. "Scenario Generation and Probabilistic Forecasting Analysis of Spatio-Temporal Wind Speed Series with Multivariate Autoregressive Volatility Models." *Applied Energy* 239: 1226–1241.

Noureldin, D., N. Shephard, and K. Sheppard 2012. "Multivariate High-Frequency-Based Volatility (HEAVY) Models." *Journal of Applied Econometrics* 27 (6): 907–933.

Patton, A. J. 2011. "Volatility Forecast Comparison Using Imperfect Volatility Proxies." *Journal of Econometrics* 160: 246–256.

Sheppard, K. 2013. "MFE Toolbox, https://www.kevinsheppard.com/code/matlab/mfe-toolbox."

Zhipeng, Y., and L. Shenghong 2018. "Hedge Ratio on Markov Regime-Switching Diagonal BEKK–GARCH Model." *Finance Research Letters* 24: 49–55.

Zolfaghari, M., H. Ghoddusi, and F. Faghihian 2020. "Volatility Spillovers for Energy Prices: A Diagonal BEKK Approach." *Energy Economics* 92: 104965.

Appendix

Figure 5 here

Figure 6 here

Table 5 here

Table 6 here

Figure 7 here

Tables and Figures

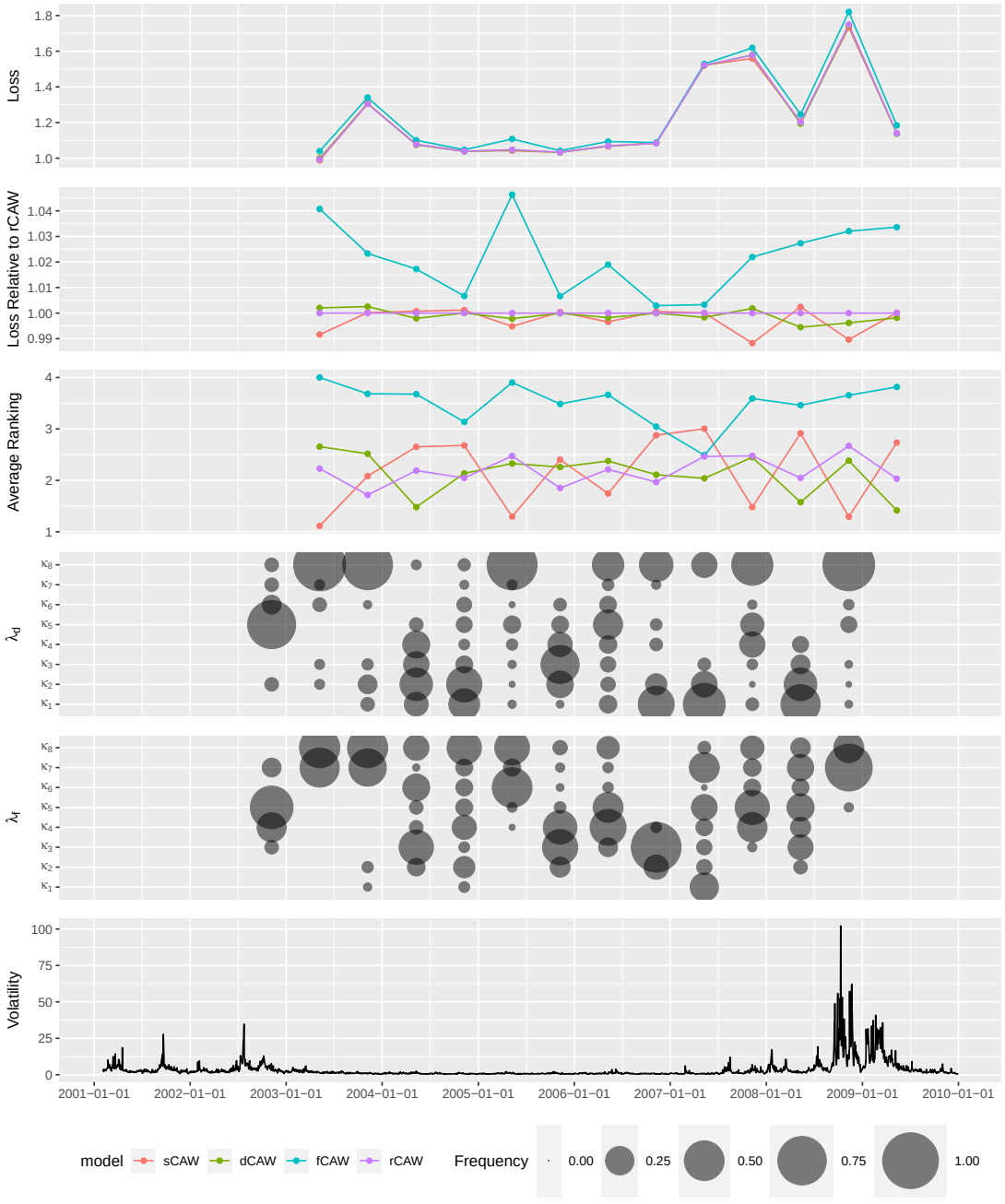
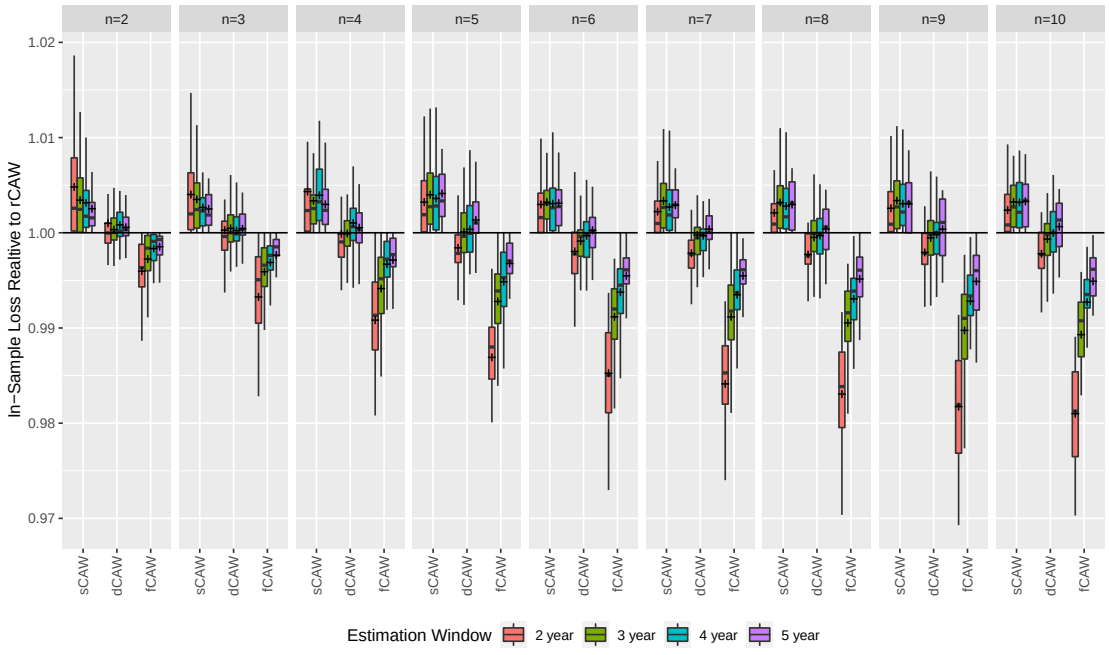
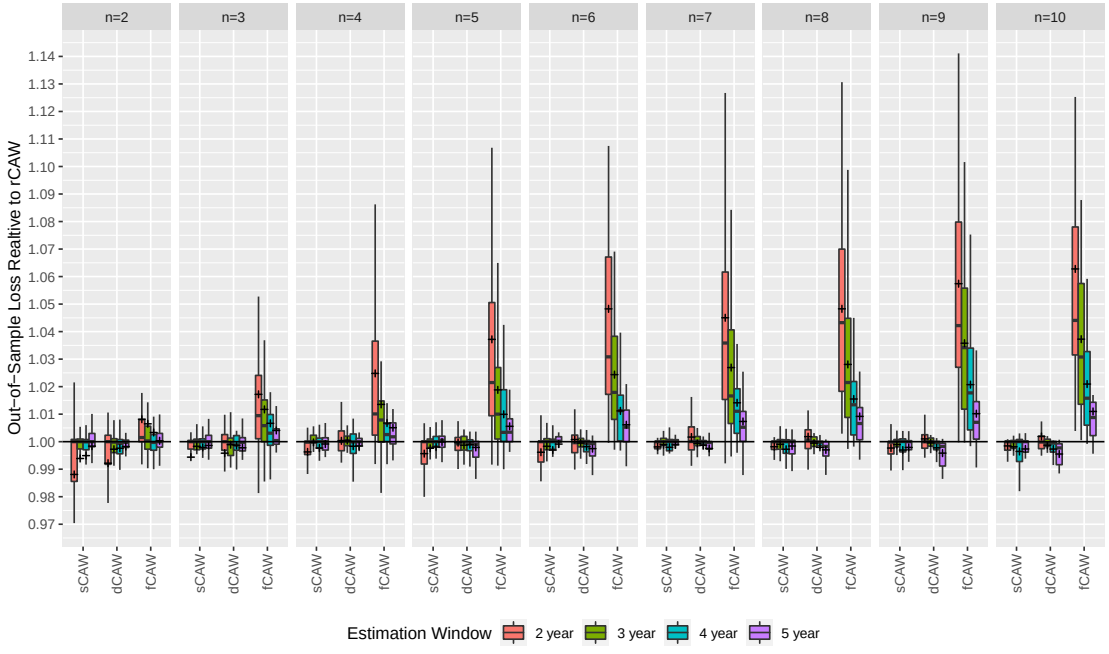


Fig. 1: The three upper panels display the out-of-sample Stein loss, ratio of Stein losses (where rCAW is the benchmark), average ranking (with 1 indicating the best model and 4 the worst model), for individual models as a function of time. The three lower panels display the optimal regularization hyper-parameters λ_d and λ_f selected, alongside with the average realized volatility, as a function of time.

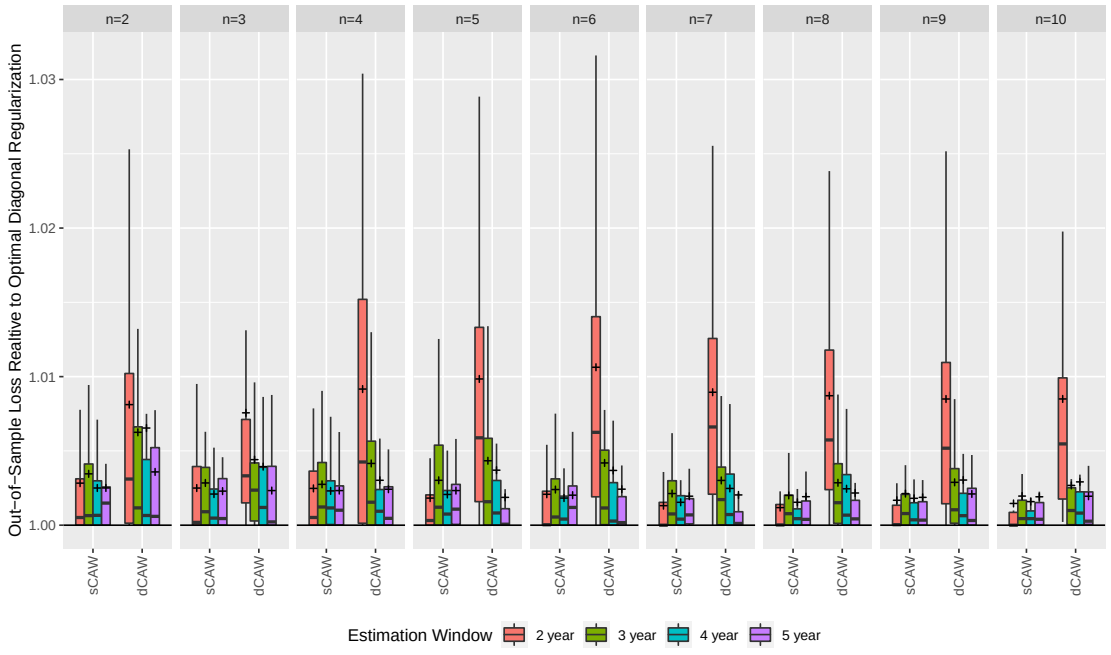


(A)

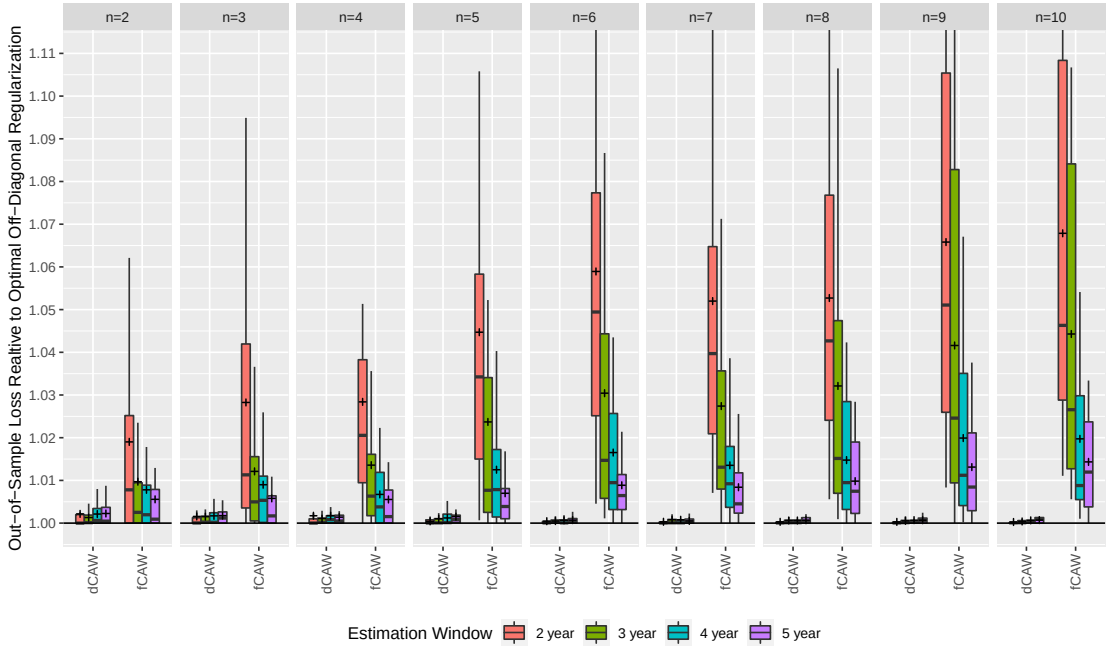


(B)

Fig. 2: Ratios of in-sample (A) and out-of-sample (B) Stein losses of individual models (rCAW is the benchmark represented by the horizontal line) plotted for different combinations of a number of assets n and length of the estimation window T.



(A)



(B)

Fig. 3: Ratios of Stein losses of individual models and the loss which would be achieved under the ex-post optimal regularization parameter λ_d (A) and λ_f (B) (benchmark, represented by the horizontal line) plotted for different combinations of a number of assets n and length of the estimation window T.

	n=2	n=3	n=4	n=5	n=6	n=7	n=8	n=9	n=10	all
2y	2.15	2.14	1.97	2.15	1.92	1.86	1.86	1.88	1.89	1.98
3y	2.43	2.33	2.41	2.26	2.12	2.00	2.04	2.15	2.04	2.21
4y	2.53	2.42	2.38	2.43	2.35	2.29	2.22	2.27	2.11	2.35
5y	2.70	2.52	2.43	2.75	2.71	2.55	2.39	2.28	2.36	2.52
all	2.43	2.31	2.30	2.37	2.22	2.16	2.15	2.13	2.13	2.25

(A) sCAW

	n=2	n=3	n=4	n=5	n=6	n=7	n=8	n=9	n=10	all
2y	2.26	2.14	2.28	1.94	2.02	2.19	2.10	2.05	2.00	2.11
3y	2.26	2.23	2.25	2.11	2.07	2.23	2.09	2.02	1.91	2.14
4y	2.28	2.22	2.22	2.03	2.03	1.97	1.94	2.07	2.00	2.09
5y	2.14	2.14	2.21	1.89	1.82	1.68	1.68	1.58	1.57	1.87
all	2.28	2.25	2.29	2.03	1.95	2.07	1.99	1.96	1.94	2.09

(B) dCAW

	n=2	n=3	n=4	n=5	n=6	n=7	n=8	n=9	n=10	all
2y	3.30	3.46	3.56	3.68	3.92	3.85	4.00	3.96	4.00	3.74
3y	2.76	3.18	3.30	3.46	3.84	3.73	3.89	3.88	3.91	3.54
4y	2.39	2.94	3.11	3.28	3.42	3.64	3.69	3.56	3.67	3.28
5y	2.50	2.89	2.96	3.14	3.18	3.36	3.57	3.58	3.57	3.18
all	2.70	3.11	3.24	3.38	3.67	3.66	3.80	3.75	3.77	3.45

(C) fCAW

	n=2	n=3	n=4	n=5	n=6	n=7	n=8	n=9	n=10	all
2y	2.30	2.26	2.19	2.23	2.14	2.11	2.04	2.11	2.12	2.16
3y	2.55	2.26	2.04	2.17	1.97	2.04	1.98	1.96	2.14	2.11
4y	2.81	2.42	2.29	2.26	2.21	2.10	2.14	2.10	2.22	2.28
5y	2.66	2.45	2.39	2.21	2.29	2.41	2.36	2.55	2.50	2.42
all	2.59	2.33	2.16	2.22	2.16	2.11	2.06	2.16	2.16	2.22

(D) rCAW

Tab. 1: Average ranking of individual models in terms of out-of-sample Stein loss (1 indicates the best model, 4 indicates the worst model) for combinations of a number of assets n and length of the estimation window T.

	n=2	n=3	n=4	n=5	n=6	n=7	n=8	n=9	n=10	all
2y	0.09	0.16	0.12	0.14	0.18	0.17	0.19	0.14	0.15	0.15
3y	0.12	0.18	0.16	0.23	0.18	0.25	0.25	0.27	0.27	0.21
4y	0.19	0.14	0.14	0.17	0.17	0.19	0.25	0.28	0.11	0.19
5y	0.14	0.11	0.14	0.18	0.39	0.25	0.21	0.27	0.29	0.22
all	0.13	0.16	0.15	0.19	0.23	0.22	0.23	0.23	0.21	0.20

(A) sCAW

	n=2	n=3	n=4	n=5	n=6	n=7	n=8	n=9	n=10	all
2y	0.23	0.16	0.36	0.26	0.29	0.36	0.38	0.28	0.31	0.29
3y	0.19	0.11	0.20	0.23	0.11	0.14	0.16	0.18	0.18	0.17
4y	0.31	0.14	0.17	0.17	0.14	0.19	0.19	0.22	0.22	0.19
5y	0.18	0.04	0.11	0.00	0.11	0.04	0.07	0.00	0.00	0.06
all	0.25	0.14	0.21	0.18	0.14	0.19	0.20	0.19	0.19	0.19

(B) dCAW

	n=2	n=3	n=4	n=5	n=6	n=7	n=8	n=9	n=10	all
2y	0.47	0.52	0.64	0.76	0.88	0.88	0.88	0.92	0.92	0.76
3y	0.24	0.41	0.52	0.68	0.73	0.73	0.77	0.85	0.91	0.64
4y	0.25	0.31	0.33	0.50	0.67	0.67	0.83	0.78	0.89	0.56
5y	0.11	0.25	0.43	0.43	0.57	0.46	0.64	0.63	0.57	0.46
all	0.27	0.39	0.46	0.59	0.70	0.70	0.77	0.81	0.83	0.61

(C) fCAW

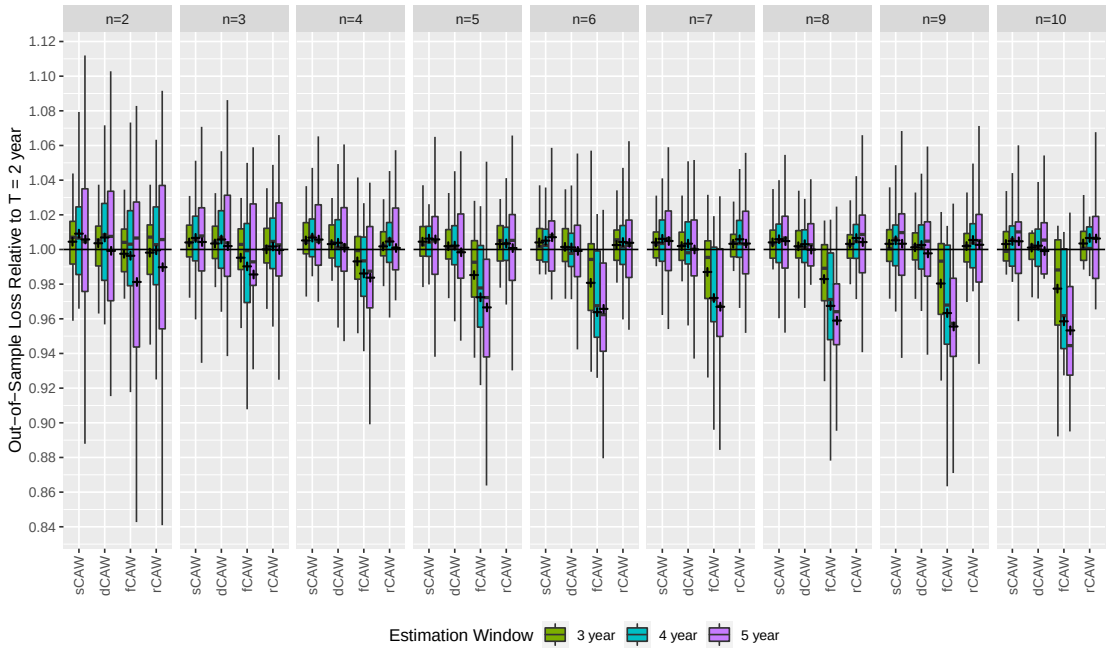
	n=2	n=3	n=4	n=5	n=6	n=7	n=8	n=9	n=10	all
2y	0.34	0.28	0.36	0.32	0.35	0.29	0.33	0.27	0.23	0.31
3y	0.21	0.20	0.11	0.20	0.20	0.16	0.18	0.21	0.36	0.19
4y	0.28	0.22	0.14	0.19	0.22	0.25	0.28	0.37	0.44	0.26
5y	0.14	0.14	0.14	0.29	0.36	0.25	0.36	0.42	0.43	0.28
all	0.25	0.22	0.19	0.22	0.26	0.22	0.25	0.28	0.31	0.24

(D) rCAW

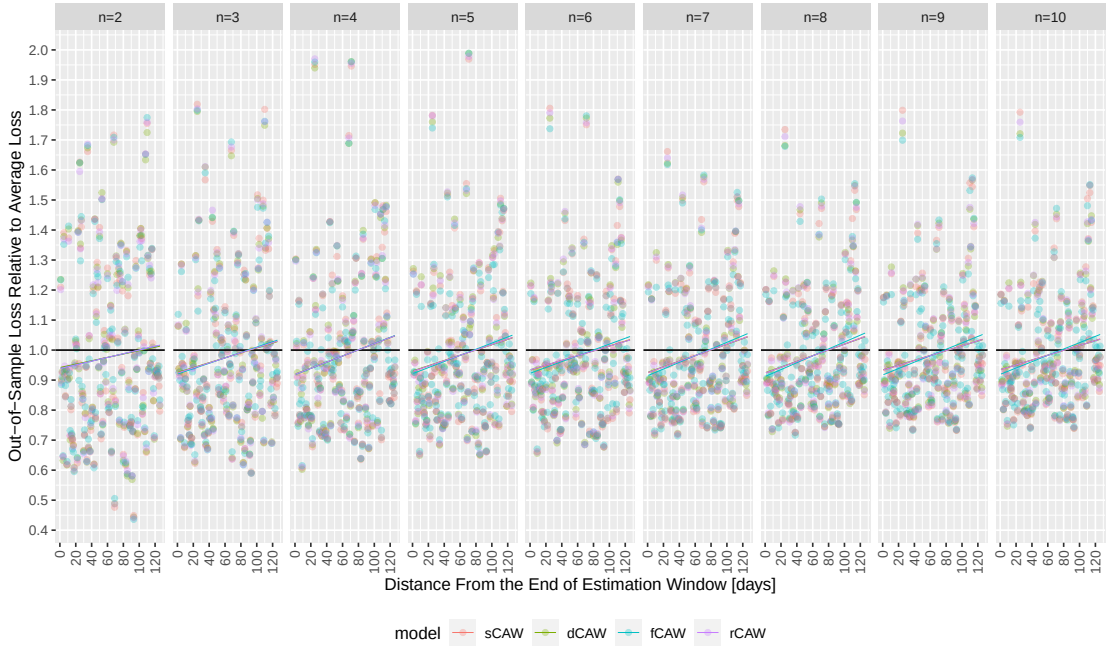
Tab. 2: Frequency of rejection of individual models via the MCS for the Stein loss at the 5% level for combinations of a number of assets n and length of the estimation window T.

H_0 :	$\mathbb{E}[e_t] = \mathbf{0}_k$	$var(e_t) = I_k$	$e_t \sim \mathcal{N}(\mu, \Sigma)$	$e_t \perp\!\!\!\perp e_{t'}$
sCAW	0.571	1.000	1.000	1.000
dCAW	0.714	1.000	1.000	1.000
fCAW	0.000	1.000	1.000	1.000

Tab. 3: Average rejection rates at $p = 0.01$ across different window locations. For each window location, the null hypotheses are tested via the t-test, chi-squared test, Anderson and Darling (1952) normality test, and multivariate Box and Pierce (1970) auto-correlation test (with $\lfloor \log(5 \cdot 252) \rfloor = 7$ lags), respectively.



(A)

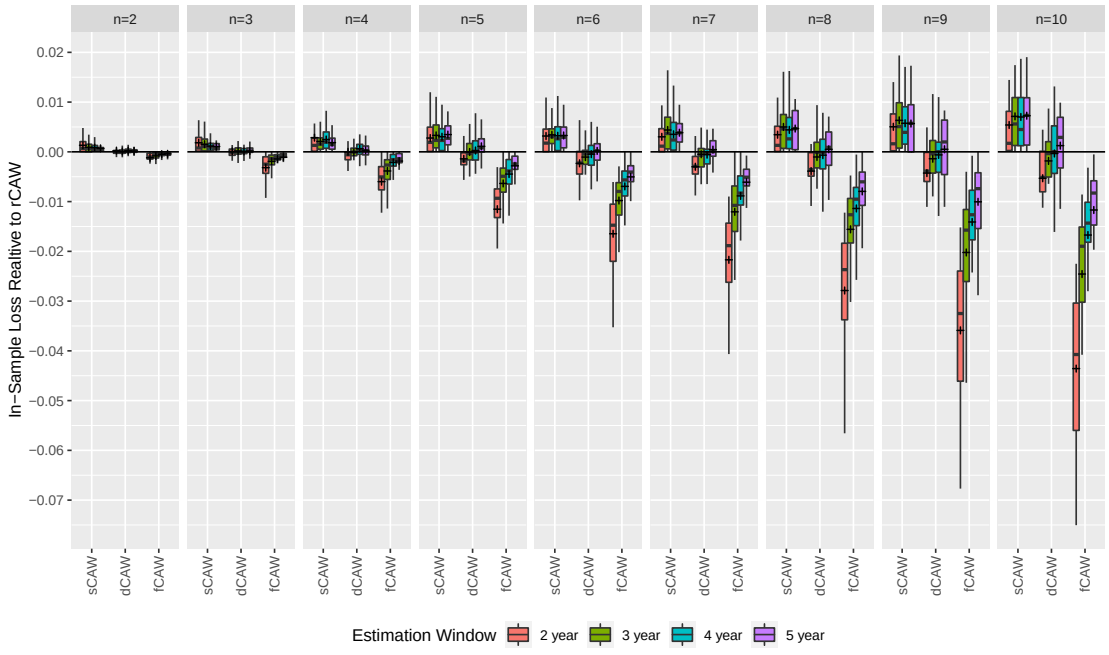


(B)

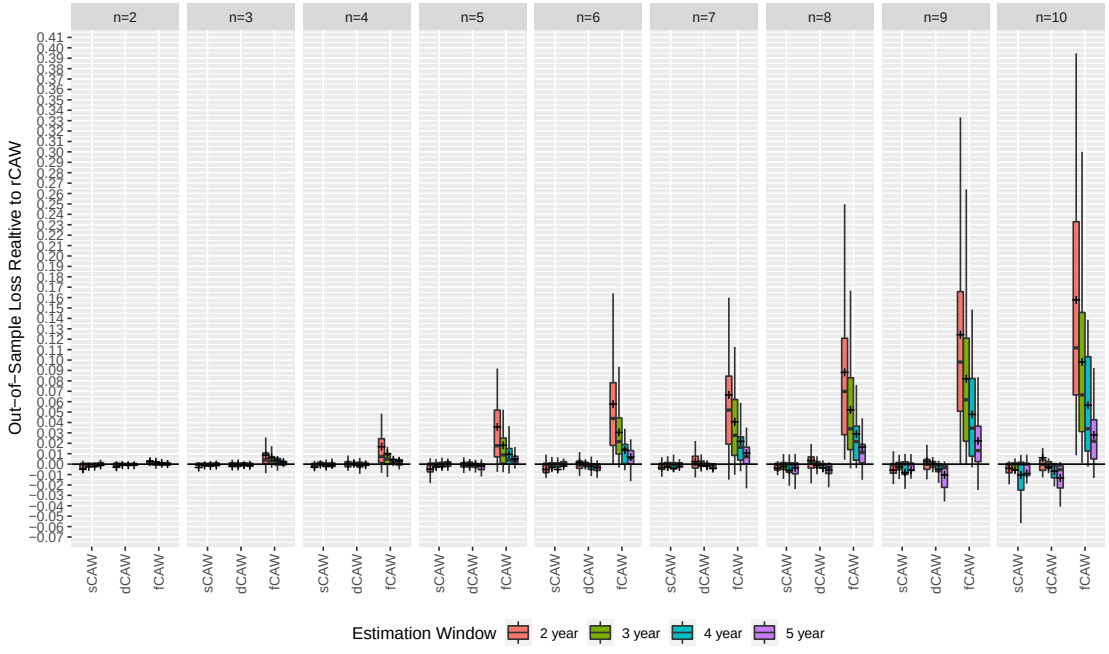
Fig. 4: Panel (A) displays ratios of out-of-sample Stein loss for lengths of estimation window $\{3, 4, 5\}$ years and the loss for the estimation window of only 2 years (benchmark represented by the horizontal line) plotted for different combinations of a number of assets n and individual models. Panel (B) displays ratios of the average Stein loss for individual out-of-sample observations with different distance from the end of the estimation window and the average loss over the whole out-of-sample period (benchmark represented by the horizontal line) plotted for different combinations of a number of assets n and individual models.

H_0 :	$\mathbb{E}[e_{t,i,j}] = 0$	$var(e_{t,i,j}) = 1$	$e_{t,i,j} \sim \mathcal{N}(\mu, \sigma^2)$	$e_{t,i,j} \perp\!\!\!\perp e_{t',i,j}$	$cov(e_{t,i,j}, e_{t,i',j'}) = 0$
sCAW	0.125	0.987	0.699	0.847	0.176
dCAW	0.109	0.987	0.714	0.875	0.177
fCAW	0.117	0.987	0.688	0.761	0.168

Tab. 4: Average rejection rates at $p = 0.01$ across different window locations and i, j. For each window location and combination of i and j, the null hypotheses are tested via the t-test, chi-squared test, Anderson and Darling (1952) normality test, univariate Box and Pierce (1970) auto-correlation test (with $\lfloor \log(5 \cdot 252) \rfloor = 7$ lags), and t-test for correlation, respectively.

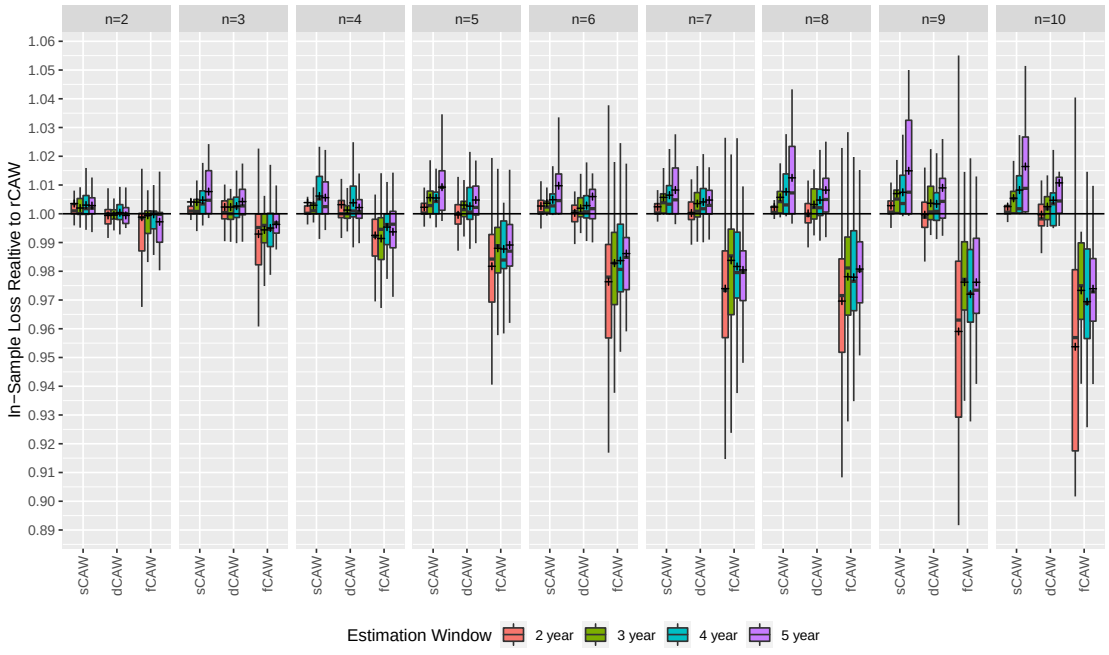


(A)

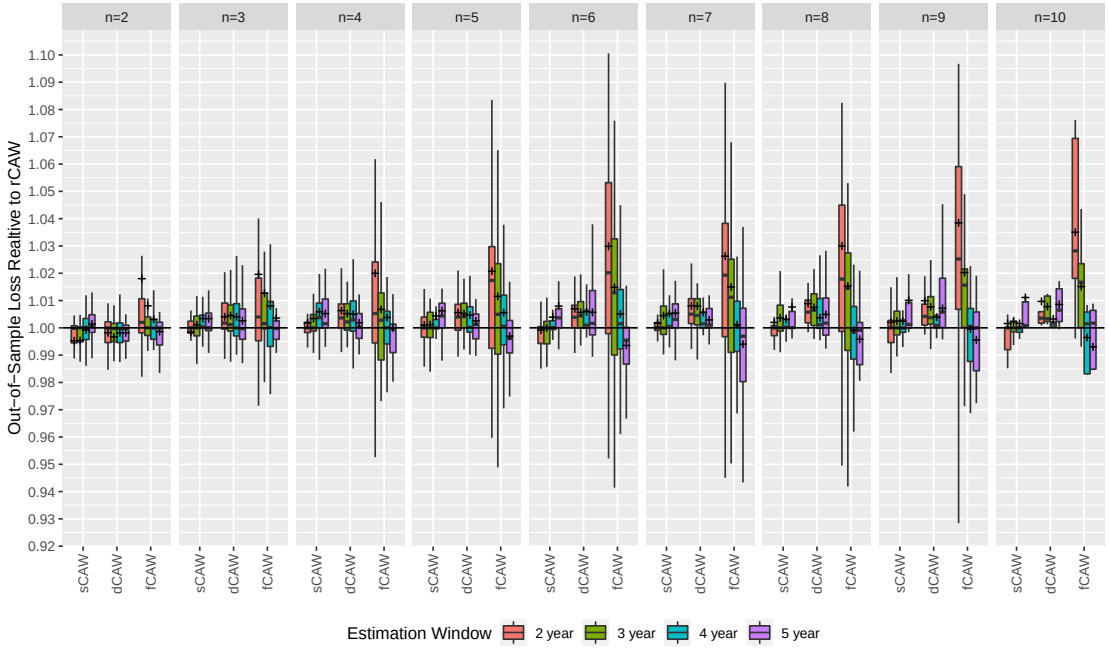


(B)

Fig. 5: Differences of the in-sample (A) and out-of-sample (B) Stein losses of individual models and rCAW (benchmark represented by the horizontal line) plotted for different combinations of a number of assets n and length of the estimation window T.



(A)



(B)

Fig. 6: Ratios of in-sample (A) and out-of-sample (B) Frobenius losses of individual models and rCAW (benchmark represented by the horizontal line) plotted for different combinations of a number of assets n and length of the estimation window T.

	n=2	n=3	n=4	n=5	n=6	n=7	n=8	n=9	n=10	all
2y	2.11	2.24	2.23	2.17	1.96	1.97	2.00	1.94	1.89	2.06
3y	2.21	2.38	2.46	2.26	2.10	2.18	2.11	1.99	2.14	2.20
4y	2.61	2.61	2.88	2.60	2.40	2.54	2.47	2.14	1.78	2.49
5y	2.88	2.77	3.07	3.18	3.18	3.12	3.00	2.94	3.07	3.01
all	2.42	2.47	2.60	2.49	2.35	2.39	2.35	2.19	2.16	2.39

(A) sCAW

	n=2	n=3	n=4	n=5	n=6	n=7	n=8	n=9	n=10	all
2y	2.47	2.72	2.84	2.78	2.88	2.87	2.96	2.92	3.00	2.82
3y	2.36	2.50	2.80	2.89	2.80	2.96	3.04	3.08	3.09	2.83
4y	2.44	2.53	2.86	2.56	2.81	2.94	2.94	3.11	3.22	2.81
5y	2.39	2.32	2.54	2.57	2.57	2.71	2.50	2.68	2.71	2.55
all	2.43	2.59	2.78	2.72	2.81	2.84	2.86	2.97	3.04	2.77

(B) dCAW

	n=2	n=3	n=4	n=5	n=6	n=7	n=8	n=9	n=10	all
2y	2.98	2.84	2.76	2.84	3.04	3.17	3.12	3.36	3.38	3.05
3y	2.81	2.93	2.48	2.77	3.02	2.93	2.91	3.06	3.00	2.88
4y	2.36	2.47	2.08	2.53	2.67	2.50	2.44	2.48	2.89	2.46
5y	2.21	2.36	2.11	2.00	1.82	2.07	2.21	2.34	2.29	2.16
all	2.63	2.64	2.41	2.59	2.69	2.74	2.76	2.83	2.91	2.68

(C) fCAW

	n=2	n=3	n=4	n=5	n=6	n=7	n=8	n=9	n=10	all
2y	2.45	2.20	2.17	2.21	2.12	1.99	1.92	1.78	1.73	2.07
3y	2.62	2.19	2.27	2.08	2.08	1.93	1.93	1.88	1.77	2.10
4y	2.58	2.39	2.18	2.32	2.12	2.01	2.14	2.27	2.11	2.25
5y	2.52	2.55	2.29	2.25	2.43	2.09	2.29	2.04	1.93	2.28
all	2.52	2.30	2.21	2.20	2.15	2.04	2.03	2.00	1.89	2.16

(D) rCAW

Tab. 5: Average ranking of individual models in terms of out-of-sample Frobenius loss (1 indicating the best model and 4 indicating the worst model) for combinations of number of assets n and length of the estimation window T.

	n=2	n=3	n=4	n=5	n=6	n=7	n=8	n=9	n=10	all
2y	0.11	0.12	0.14	0.14	0.02	0.14	0.15	0.12	0.08	0.12
3y	0.02	0.20	0.14	0.14	0.09	0.07	0.14	0.06	0.09	0.10
4y	0.11	0.17	0.14	0.06	0.03	0.06	0.03	0.04	0.00	0.07
5y	0.07	0.07	0.11	0.18	0.25	0.21	0.18	0.22	0.29	0.17
all	0.07	0.14	0.15	0.12	0.10	0.11	0.13	0.10	0.10	0.11

(A) sCAW

	n=2	n=3	n=4	n=5	n=6	n=7	n=8	n=9	n=10	all
2y	0.21	0.30	0.26	0.18	0.22	0.36	0.33	0.33	0.46	0.28
3y	0.07	0.20	0.14	0.16	0.16	0.14	0.16	0.20	0.09	0.15
4y	0.17	0.11	0.08	0.06	0.03	0.06	0.08	0.07	0.11	0.08
5y	0.11	0.04	0.07	0.11	0.14	0.18	0.04	0.10	0.14	0.10
all	0.14	0.19	0.16	0.13	0.15	0.19	0.17	0.19	0.20	0.17

(B) dCAW

	n=2	n=3	n=4	n=5	n=6	n=7	n=8	n=9	n=10	all
2y	0.23	0.24	0.28	0.26	0.37	0.31	0.29	0.36	0.38	0.30
3y	0.17	0.23	0.11	0.20	0.23	0.30	0.23	0.30	0.18	0.22
4y	0.17	0.17	0.06	0.08	0.17	0.06	0.08	0.09	0.11	0.11
5y	0.11	0.14	0.07	0.04	0.07	0.11	0.11	0.15	0.00	0.10
all	0.16	0.20	0.13	0.16	0.22	0.19	0.19	0.23	0.23	0.19

(C) fCAW

	n=2	n=3	n=4	n=5	n=6	n=7	n=8	n=9	n=10	all
2y	0.17	0.14	0.18	0.20	0.20	0.19	0.19	0.18	0.23	0.18
3y	0.10	0.16	0.14	0.11	0.09	0.07	0.07	0.08	0.00	0.10
4y	0.14	0.11	0.06	0.00	0.00	0.00	0.00	0.00	0.00	0.04
5y	0.14	0.14	0.11	0.11	0.11	0.14	0.07	0.15	0.14	0.12
all	0.14	0.15	0.13	0.11	0.10	0.11	0.10	0.10	0.10	0.12

(D) rCAW

Tab. 6: Frequency of rejection of individual models via the MCS for the Frobenius loss at the 5% level for combinations of a number of assets n and length of the estimation window T.

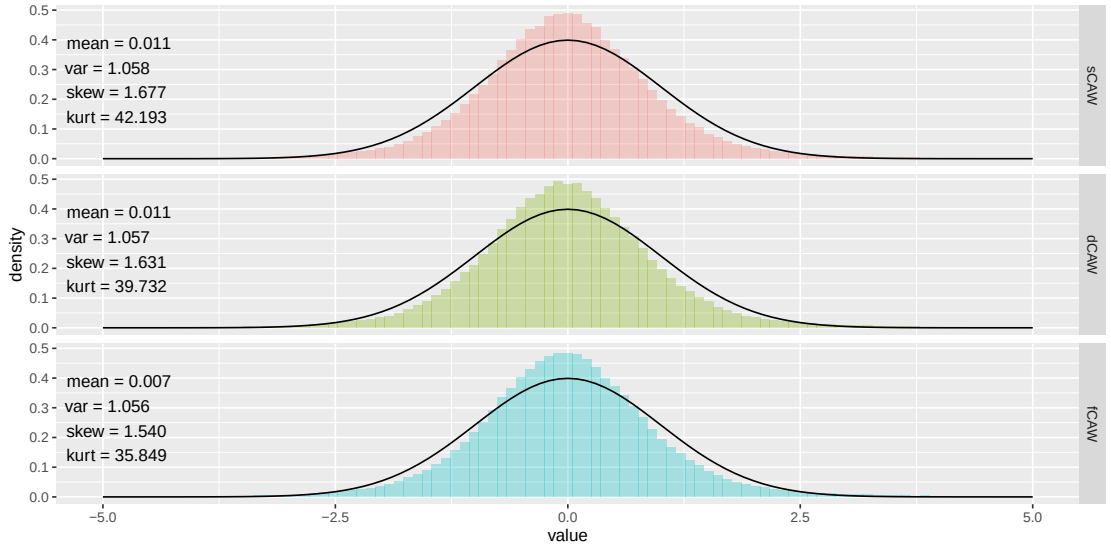


Fig. 7: Distributions of $e_{t,i,j}$ for individual models plotted against $\mathcal{N}(0,1)$ distribution function.