

Model complexity and model performance

Stanislav Anatolyev*
New Economic School

Andrey Shabalin
University of North Carolina, Chapel Hill

Abstract

By experimenting with real financial data we analyze the dependence of performance measures of dynamic time-series models on the presence of certain features in the model, such as mean persistence, volatility clustering, leverage effects, time-varying risk premia, heavy tails and skewness.

Key words: Dynamic models, model selection, goodness of fit, forecasting, financial returns.

JEL classification: C51

*Corresponding author. Address: Stanislav Anatolyev, New Economic School, Nakhimovsky Prospekt, 47, Moscow, 117418 Russia. E-mail: sanatoly@nes.ru.

1 Introduction

Applied researchers routinely construct dynamic time-series models to describe evolution and make forecasts for series of interest. In doing this, a researcher is guided, on the one hand, by a need to incorporate certain stylized features of data into the model. On the other hand, a model should fit the data well, both in- and out-of-sample, and it can be deemed successful only if it passes a battery of diagnostic and other tests.

In this paper, we investigate this trade-off between model complexity and model performance using real financial data. To this end, we run estimation and forecasting exercises on various types of data to answer the question: for a particular performance measure, what are the critical features that should be incorporated in the model? We try to shed light on this issue by looking at the dependence of various performance measures on the presence or absence of certain features in the model. We run panel regressions of these measures on dummy variables representing model features, also including fixed individual effects to control for heterogeneity across different series. The coefficients in such panel regression, together with their significance, indicate which model features are more and which are less important for a particular performance measure. The model features considered allow for mean persistence, volatility clustering, leverage effects, time-varying risk premia, heavy tails, and skewness.

The data types considered are individual stock prices, stock market indices, and exchange rates, popular in the empirical finance literature. The data are weekly, and extend for 20–30 years totaling to about a thousand observations. The results indicate that the tendencies in complexity–performance trade-off are quite different across the three data types, but are quite uniform across different series of the same type.

The paper is organized as follows. Section 2 describes our technology of constructing dynamic models. Section 3 lists criteria used for judgement about model performance. The description of data is given in Section 4. The results are analyzed in Section 5.

2 Dynamic models

We consider models where each feature f from a set of F features to be described shortly is either present or not. Some of the features may be present only if other features are or are not, hence there are fewer than 2^F dynamic models in total. Each model has the structure

$$y_t = \mu_t + e_t,$$

where μ_t is (roughly) the mean, and e_t is (roughly) the error term. By default, $\mu_t = \bar{\mu}$, a constant, and $e_t = \varepsilon_t$, a martingale difference (relative to past data) innovation. Let

$$\varepsilon_t = \sqrt{\sigma_t^2} \eta_t, \quad \eta_t \sim i.i.d. D(0, 1, \tau),$$

where σ_t^2 is a conditional variance, η_t is a standardized innovation, D is a conditional distribution, τ is a vector of additional parameters of this distribution. By default, $\sigma_t^2 = \omega$, a constant, and D is standard normal so that there is no τ .

The set of model features f contains, feature names appearing first in brackets:

- (AR) Autoregressive component in μ_t (may be in effect only when there is no feature MA). When this feature is present, $\mu_t = \bar{\mu} + \phi y_{t-1}$.
- (MA) Moving average component in e_t (may be in effect only when there is no feature AR). When this feature is present, $e_t = \varepsilon_t + \theta \varepsilon_{t-1}$.
- (GARCH) ARCH effect in ε_t . When this feature is present, $\sigma_t^2 = \omega + \alpha \varepsilon_{t-1}^2 + \beta \sigma_{t-1}^2$.
- (ArchM) ARCH-in-mean term in μ_t (may be in effect only when there is feature GARCH). When this feature is present, μ_t in addition contains $\delta \sigma_t^2$.
- (Lev) Leverage in σ_t^2 (may be in effect only when there is feature GARCH). When this feature is present, σ_t^2 in addition contains $\gamma \varepsilon_{t-1}^2 \mathbb{I}[\varepsilon_{t-1} < 0]$.
- (Stud) Conditional fat-tailedness of η_t . When this feature is present, $D(0, 1, \tau)$ is Student with $\tau = \nu$, where ν is the number of degrees of freedom.
- (Skew) Conditional skewness of η_t (may be in effect only when there is feature Stud). When this feature is present, $D(0, 1, \tau)$ is skewed Student with $\tau = (\nu, \lambda)'$, where λ measures the degree of skewness.

Thus, in total there are $F = 7$ features resulting in 45 dynamic models. The features we have included are, in our view, most important and popular in building a model for financial data that would serve general purposes of describing the dynamics and making short-run forecasts. The choice of a particular way to represent each feature when there is a variety of choices is made in favor of a simple and empirically popular model.

3 Performance criteria

We judge model performance by two sorts of criteria: in-sample and out-of-sample. The sample of y_t for $t = 1, \dots, T$ is divided into the estimation and prediction parts, the former running from 1 to R , the latter running from $R+1$ to $R+P(\equiv T)$. The proportion of R to P is 2 : 1.

The in-sample criteria are: the Ljung–Box statistics of order 10 applied to residuals $\hat{\varepsilon}_t$ (LB) and squared standardized residuals $\hat{\eta}_t^2$ (LB²); the BDS test statistic (BDS, see Brock, Dechert, Scheinkman and LeBaron, 1996) applied to standardized residuals $\hat{\eta}_t$ with the additional parameter 5 (recall that the BDS statistic is asymptotically normal under

the null of IID, and it is natural to use BDS test as one-sided); the Bayesian information criterion (BIC, see Schwarz, 1978). When making forecasts, we use parameter values estimated only once from the data from 1 to R . Let $\hat{y}_{t|t-1}$ denote a forecast of y_t made at $t-1$, and $\hat{\sigma}_{t|t-1}^2$ be a model-based volatility estimate. The out-of-sample criteria are: the mean squared (MSPE) and mean absolute (MAPE) prediction errors; the proportion of times the sign is correctly predicted (SIGN); and the mean squared prediction error for volatility (VSPE), i.e. average of $((y_t - \hat{y}_{t|t-1})^2 - \hat{\sigma}_{t|t-1}^2)^2$.

4 Data

We use three types of weakly (taken on Wednesdays) financial data that are typically fitted with dynamic models: stock market indices from developed markets, individual stock prices from the New York Stock Exchange, and exchange rates of currencies of industrialized countries versus the US dollar. The raw data are converted into the form of returns by taking log differences.

Individual stock returns are represented by 20 stocks that were included in the S&P100 index in 2001 and have been traded since 1971. The symbols of these stocks are: AA, AEP, DD, DIS, EK, GE, GM, HON, IBM, IP, JNJ, KO, MCD, MMM, MRK, PG, S, T, UTX, XOM, with 1548 observations in each series. The data for stock index returns are represented by 12 indices: DJIA, S&P500, Nasdaq, NYA, CRSP, FTSE-100, CAC-40, DAX, Nikkei-225, TSE-300, HSI, STI. These samples are more heterogeneous across series; the minimal sample size is 730, the maximal equals 1562. For all stock markets, data on 10.21.1987, 10.28.1987, 11.04.1987 were removed to avoid the influence of the 1987 stock market crash. Exchange rates returns are represented by 9 currencies: BEF, CAD, CHF, DEM, FRF, GBP, ITL, JPY, SEK; the sample period is 01.02.1974–12.27.2000, totaling to 1390 observations in each series.

5 Results

To sift out tendencies in a huge amount of output information, we aggregate the results in the following way which is reminiscent of the response surface methodology where linear regressions are used to make extrapolations. For every performance measure, we regress its values on dummy variables representing the seven model features, also including individual effects pertaining to different series. That is, we run linear panel data regressions with fixed “series effects”, but we do that separately for indexes, individual stocks and exchange rates. The output of interest is composed of regression coefficients and their significance showing a sign and impact of each model feature.

The results are presented in Table 1, where we have boldfaced coefficients whose

t-ratios are greater than 1.6 in absolute value. We regard the boldfaced numbers as deserving attention most, and corresponding features as most critical ones. Below, we briefly comment on general tendencies revealed by numerical results, both reported and unreported, for each performance measure separately.

LB The principal factor that plays a role for the LB characteristic is the presence of mean filtering, either of autoregressive or moving average type, while other features cannot seriously influence LB values. Interestingly, the AR or MA parameters (ϕ or θ), if present, may not be significant even at 10%.

LB² Most radically the LB² characteristic is influenced by the presence of volatility filtering by GARCH, and is absolutely insensitive to the presence of other features. The presence of pure GARCH alone is able to reduce LB² values from significant at 1% to insignificant at 10%. The GARCH parameters (α and β) are highly significant (an exception sometimes occurs in the presence of Lev when γ pulls significance away from β).

BDS The BDS statistic is able to point at no neglected nonlinearity after volatility filtering by GARCH, and it is rather insensitive to the presence of other features. However, even in the absence of a GARCH part BDS may not help detect nonlinearity. The presence of mean filtering and allowance for conditional thick tails may sometimes improve the BDS statistic.

BIC According to BIC, more parsimonious models containing one to two, rarely three features, are best, provided they contain a GARCH equation, while models not containing GARCH specifications are evident outsiders. Another important factor that significantly improves fit is the Stud feature that allows for conditionally heavy tails. For individual stocks, most often the best model is GARCH with conditional Student distribution (7 times out of 20), next comes conditionally normal GARCH with leverage (3 times). The latter model is best most often for stock indices (3 times out of 12). For exchange rates, *almost always* the best model is GARCH with conditional Student distribution (8 times out of 9).

MSPE and MAPE Confirming common wisdom, there is a general tendency that more parsimonious models tend to predict the mean more successfully. Further patterns are not very clear and are different for different types of data. On the whole, the MSPE and MAPE criteria are in consensus most of the time, but the two do not tend to agree on which models are best and which are worst. For individual stocks, the presence of ArchM is harmful for predicting the mean, while the presence of AR or MA filters has a favorable impact

in most cases, even though corresponding coefficients are rarely significant. For stock indices, the presence of a thick-tailed conditional distribution worsens mean prediction, but allowing for skewness acts in the opposite direction, and the net effect on mean prediction is favorable. For exchange rates, the decisive and favorable factor happens to be heavy-tailed conditional distribution, and, perhaps surprisingly, mean filtering tends to worsen mean predictability.

SIGN For individual stocks and stock indices, it is easy to exceed the coin toss sign prediction of 50% using a model that is not among best; with best models one can achieve 59% for some stocks (XOM) and 62% for indices (CRSP and NYA). In contrast, it is much harder to predict signs of exchange rate movements; even with best models one cannot exceed 50% appreciably. There is no clear-cut pattern of which features impact SIGN most, but mean filtering seems to have greatest effect, negative in case of stocks or indices, and positive in case of exchange rates.

VSPE In volatility predictions, the presence of GARCH is important, if not decisive. Strangely, however, that the GARCH factor has a favorable impact in case of individual stocks or stock indices, but an adverse impact in case of exchange rates.

Overall, we observe quite appreciable difference in performance of the same dynamic models when they are fit to exchange rates compared to when they are fit to stock returns and indexes. A possible explanation is that the behavior of exchange rates may not be temporally stable during long periods.

References

- Brock, W., D. Dechert, J. Scheinkman and B. LeBaron (1996) A test for independence based on the correlation dimension. *Econometric Reviews* 15, 197–235.
- Schwarz, G. (1978) Estimating the dimension of a model. *Annals of Statistics* 6, 461–464.

Features	In-sample				Out-of-sample			
	LB $\times 10^0$	LB ² $\times 10^2$	BDS $\times 10^0$	BIC $\times 10^{-2}$	MSPE $\times 10^0$	MAPE $\times 10^{-2}$	SIGN $\times 10^{-2}$	VSPE $\times 10^5$
Individual stocks								
AR	-3.501	0.004	-0.013	0.373	-3.713	-2.455	-0.181	-0.520
MA	-3.528	0.003	-0.013	0.367	-4.164	-2.583	-0.242	-0.552
GARCH	0.143	-1.087	-0.813	-3.057	-1.466	-1.663	-0.025	-4.190
ArchM	-0.234	0.003	0.000	0.537	12.561	8.248	-0.147	1.934
Lev	0.000	0.007	-0.002	0.163	2.779	2.790	0.003	0.117
Stud	-0.042	0.008	-0.013	-1.613	0.530	0.773	0.073	-1.048
Skew	-0.009	0.001	-0.001	0.448	-0.032	-0.465	-0.035	0.122
Stock indices								
AR	-4.939	-0.003	-0.019	0.473	0.186	2.267	-0.508	-0.885
MA	-4.434	-0.002	-0.018	0.501	0.348	2.074	-0.607	-0.806
GARCH	-0.060	-1.265	-1.240	-6.104	1.125	-0.584	0.204	-3.496
ArchM	0.024	0.002	0.002	0.780	3.000	2.946	-0.211	-0.351
Lev	-0.131	-0.004	-0.042	0.105	0.196	1.749	-0.231	-0.966
Stud	0.141	-0.010	-0.039	-1.880	2.314	0.133	-0.001	0.125
Skew	-0.039	0.000	0.010	-0.068	-3.329	-2.184	0.076	-0.487
Exchange rates								
AR	-4.167	0.007	-0.033	0.251	1.326	1.325	1.281	0.007
MA	-3.907	0.006	-0.028	0.317	1.062	1.236	1.110	0.001
GARCH	0.101	-0.682	-1.934	-7.235	0.588	1.582	-0.505	0.394
ArchM	0.069	-0.002	0.005	0.648	1.009	1.896	0.048	0.016
Lev	-0.097	-0.007	0.003	0.546	-0.149	-0.678	0.144	0.079
Stud	-0.467	-0.050	-0.311	-7.334	-1.491	-3.988	0.296	0.242
Skew	0.045	0.000	0.004	0.456	0.268	1.641	-0.379	0.021

Table 1. Panel regressions: marginal influence of model features.