Journal of Economic Dynamics & Control 90 (2018) 1–29

Contents lists available at ScienceDirect

Journal of Economic Dynamics & Control

journal homepage: www.elsevier.com/locate/jedc

Model Complexity and Out-of-Sample Performance: Evidence from S&P 500 Index Returns

Andreas Kaeck[a,1], Paulo Rodrigues[b,2], Norman J. Seeger[c,*]

[a] University of Sussex, United Kingdom
[b] Maastricht University, The Netherlands
[c] VU University Amsterdam, De Boelelaan 1105, 1081 HV Amsterdam, The Netherlands

ARTICLE INFO

Article history:
Received 10 April 2017
Revised 22 January 2018
Accepted 25 January 2018
Available online 5 February 2018

JEL classification:
G12
G15
C53

Keywords:
Out-of-sample specification tests
Jump-diffusion models
Lévy-jump models
Non-affine variance models
Forecasting

ABSTRACT

We apply a range of *out-of-sample* specification tests to more than forty competing stochastic volatility models to address how model complexity affects out-of-sample performance. Using daily S&P 500 index returns, model confidence set estimations provide strong evidence that the most important model feature is the non-affinity of the variance process. Despite testing alternative specifications during the turbulent market regime of the global financial crisis of 2008, we find no evidence that either finite- or infinite-activity jump models or other previously proposed model extensions improve the out-of-sample performance further. Applications to Value-at-Risk demonstrate the economic significance of our results. Furthermore, the out-of-sample results suggest that standard jump diffusion models are misspecified.

1. Introduction

In this paper, we analyze continuous-time and discrete-time models for S&P 500 index returns to study the relationship between model complexity and out-of-sample performance. The study of time-series dynamics of major stock market indices, such as the S&P 500, has previously attracted a large number of empirical studies, see e.g. Bates (2012); Christoffersen et al. (2010); Eraker et al. (2003), or Kou et al. (2013) and applied research today is faced with the challenge of selecting model dynamics from a huge number of alternative specifications.

Despite the importance of this research area, many papers in the continuous-time literature focus on *in-sample* specification tests. In this paper, we diverge from this approach and provide a range of different *out-of-sample* performance tests. In-sample studies are very helpful to learn about the structural building blocks required to produce stylized facts in the data. However, eventually the out-of-sample performance of a model is crucial for market participants using such a model in finance applications that are affected by uncertain future market scenarios. Our main aim is to understand to what extent

* Corresponding author.
E-mail addresses: a.kaeck@sussex.ac.uk (A. Kaeck), p.rodrigues@maastrichtuniversity.nl (P. Rodrigues), n.j.seeger@vu.nl (N.J. Seeger).
1 Tel.: +44 1273 678433.
2 Tel.: +31 43 388 36 33.

https://doi.org/10.1016/j.jedc.2018.01.040
0165-1889/

the superior performance of sophisticated stochastic models prevails when they are applied outside their estimation period. To this end, we first estimate more than forty different stochastic models that encompass the most widely used model features in the continuous-time literature. These features include affine vs non-affine models, single-factor vs multi-factor specifications, diffusion models vs jump models, finite activity vs infinite activity, discrete-time vs continuous-time models. The combination of these building blocks leads to a very comprehensive set of competing models. Although not the focus of this paper, we also study some new model specifications such as non-affine time-changed Lévy models. To the best of our knowledge, this paper is the first to provide comprehensive out-of-sample evidence for such a large set of stochastic models.[3]

Various model specification tests are then applied to an out-of-sample period of S&P 500 index returns, including the turbulent market regime during the onset of the financial market crisis in 2008. To compare performances of a very large number of model specifications, we employ the model confidence set estimation procedure of Hansen et al. (2011). We separate subsets of models that have a statistically indistinguishable performance according to various different out-of-sample loss functions. In doing so, we accept that one single best performing model might not exist but rather that different modeling approaches may be equally successful. First, we compare likelihood-based out-of-sample fit statistics, including sequential likelihoods as proposed by Johannes et al. (2009). This allows us to detect the time-periods during which particular model specifications out- or underperform. Secondly, we follow Gneiting and Ranjan (2011) in comparing models using the continuous ranked probability score (CRPS), a criterion that can be used to compare the out-of-sample forecasting performance. CRPS has the advantage that weighted versions of the statistic retain propriety, which is essential for comparing the performance in various areas of the forecasting distributions. It is often argued that jump models in particular provide a better fit to the tails of the return distribution, and weighted CRPS fit statistics are employed to study model performance in the tails (as well as the center of the return distribution). Thirdly, we test the economic significance of our results by applying the VaR loss function of González-Rivera et al. (2004). And fourthly, we use a range of absolute model tests suggested by Berkowitz (2001) and others.

Our empirical tests provide two main results. First, we find that no model is able to produce out-of-sample predictions in line with the true data-generating process. Using the test statistics developed in Berkowitz (2001) we find that all models analyzed are rejected when tested on the entire out-of-sample period. Second, we find that in terms of relative model performance more parsimonious stochastic volatility models outperform models that include a jump component. This is a surprising result, since numerous papers find that jump models outperform continuous stochastic volatility models *in-sample* (see Eraker, 2004; Eraker et al., 2003 or Ignatieva et al., 2015).

There are two possible explanations why jump models are outperformed. First, one may interpret this result as evidence for misspecification of the jump component (despite the fact that we use quite sophisticated jump modeling) and *not* as evidence against the importance of modeling jumps in equity returns. Our results may be driven by the fact that jumps are difficult to estimate and jump distributions and intensities may vary strongly over time. For instance, jump parameters may be very different during periods of crisis and this may cause model misspecification. This finding is related to results in Santa-Clara and Yan (2010) who find a weak connection between variance and the jump intensity when both processes are estimated independently.

Second, and more important, our results may provide useful insights of how the global financial crisis unfolded. High returns may either be driven by jumps or stochastic volatility. Jumps are crucial to explain a number of rare events such as the market crash of 1987 (see the discussion in Eraker et al. (2003)). On the contrary, periods of high market volatility may render jumps obsolete as stochastic volatility is sufficient to generate a sequence of large returns in times of prolonged high market volatility. The result of pure stochastic volatility models outperforming jump models implies that an increasing level of market volatility during our out-of-sample period was sufficient to model financial crisis returns from 2007 to 2009. The distinction between how shocks are created is important for many applications in finance as rare event models may have very different implications compared to models driven by stochastic volatility. This finding is related to Stroud and Johannes (2014) who draw similar conclusions using high frequency returns.

Finally, we provide several additional empirical exercises to corroborate our findings. First, we increase our parameter updating frequency to investigate the impact of the partitioning of the sample into one estimation and one forecasting period. Second, we investigate the effect of including additional data in the information set, namely realized variance and the VIX index. Third, we investigate the impact of time varying expected returns on our results.

2. Related Literature

Prior literature on testing continuous time models for stock returns are often interested in the in-sample performance of models. To tackle the challenge of estimating complex continuous time models a range of different estimation and filtering techniques has been developed. These include simulated methods of moments approaches, approximate maximum likelihood estimation, efficient methods of moments and Bayesian MCMC estimation algorithms (see Andersen et al., 2002; Bates, 2006; Eraker et al., 2003 or Johannes et al., 2009). At least partly driven by the differences in estimation methodology,

[3] Few papers consider the out-of-sample performance of continuous-time models. Yun (2014) conducts a range of density forecasting tests using affine one-factor jump-diffusion models, Shackleton et al. (2010) use similar model specifications. This paper differs substantially from the aforementioned papers as we focus on a much broader number of models and specification tests.