Journal of Data Science **7**(2009), 423-431

Super-Whiteness of Returns Spectra

Erhard Reschenhofer
University of Vienna

Abstract: Until the late 70's the spectral densities of stock returns and stock index returns exhibited a type of non-constancy that could be detected by standard tests for white noise. Since then these tests have been unable to find any substantial deviations from whiteness. But that does not mean that today's returns spectra contain no useful information. Using several sophisticated frequency domain tests to look for specific patterns in the periodograms of returns series we find nothing or, more precisely, less than nothing. Actually, there is a striking power deficiency, which implies that these series exhibit even fewer patterns than white noise. To unveil the source of this "super-whiteness" we design a simple frequency domain test for characterless, fuzzy alternatives, which are not immediately usable for the construction of profitable trading strategies, and apply it to the same data. Because the power deficiency is now much smaller, we conclude that our puzzling findings may be due to trading activities based on excessive data snooping.

Key words: Frequency domain tests, power deficiency, predictability of returns.

1. Introduction

There is abundant evidence that price changes are predictable, but there is also evidence that this predictability is getting smaller over time (see, e.g., Patro and Wu, 2004, Reschenhofer 2004a). Of course, predictability does not necessarily imply the existence of profitable trading strategies, if transaction costs are taken into account. Over the past decades, transaction costs have fallen dramatically. But at the same time, predictability has also decreased. The economic relevance of predictability must therefore always be evaluated in temporal context, regardless whether these phenomena are connected or not.

The simplest and best documented source of predictability is serial correlation. The hypothesis that returns are serially uncorrelated (white noise hypothesis)

can be tested either with time domain tests or with frequency domain tests. While time domain tests for white noise are based on the sample autocorrelations, frequency domain tests are based on the periodogram. The periodogram of a time series is the sample analog of its spectral density. It is obtained by first decomposing the time series into a sum of sinusoidal components with different amplitudes, frequencies, and phases and then plotting the squared amplitudes against the frequencies. The white noise hypothesis implies on the one hand that all nontrivial theoretical autocorrelations vanish and on the other hand that the "theoretical periodogram", i.e., the spectral density, is constant. It will therefore be rejected if some sample autocorrelations are too large in absolute vale or if some periodogram values (or clusters of periodogram values) are much larger than others.

Some tests for white noise have been specially tailored to guarantee robustness against conditional heteroskedasticity (Taylor 1984, Lo and MacKinlay 1988, Deo 2000). However, it is a priori not clear whether approximate tests that are valid under a large range of assumptions are really more useful and reliable than exact tests assuming normality (see Faust, 1992, Richardson and Stock, 1989). Durlauf (1991) proved the robustness of the asymptotics of various frequency domain tests against many forms of heteroskedasticity. Reschenhofer (2004b) introduced a test that is also robust against other peculiarities of financial time series such as nonstationarities and calendar anomalies.

In contrast to the predictability of a returns series by its own past, predictability by various financial variables (such as the dividend-price ratio) and macroeconomic variables (such as inflation) is much harder to establish in a rigorous manner (for conflicting findings see Lanne, 2002, on the one hand and Xu, 2004, on the other hand). This is simply due to the much greater risk of over-fitting

In this paper, we call into question the common belief that an apparently constant spectrum is completely uninformative. Using conventional frequency domain tests, which are most powerful in the case of distinct spectral patterns, and taking explicitly into account the possibility that there may be a lack of patterns, we find even fewer distinct patterns in the periodograms of returns series than we would expect in the case of a perfect white noise process. For a further investigation of this puzzling finding, we design a simple and robust frequency domain test for white noise against characterless, fuzzy alternatives and apply it to the same data. It turns out that this new test is more powerful than the other tests, which indicates that a spectrum exhibiting an extremely flat and wide peak is a more realistic alternative to a constant spectrum than a spectrum with a steep and narrow peak. While the latter alternative implies the presence of sinusoidal components with large amplitudes and frequencies within a narrow band, the former alternative implies cycles with small amplitudes and

fuzzy periods, which cannot immediately be used for the construction of trading strategies. The observed power deficiency of the conventional tests may therefore be due to trading activities based on excessive data snooping.

The paper is organized as follows. The next section explains the different frequency domain tests that are used in our investigation. Section 3 reports the results obtained by applying these tests to financial data. Section 4 concludes.

2. Some Frequency Domain Tests for White Noise

Frequency domain tests for white noise are typically based on the periodogram of the observed time series. Since the normalized cumulative periodogram

$$0 \leq J_1 \leq \ldots \leq J_{m-1} \leq 1, \quad m = [(-1)/2] \tag{2.1}$$

where

$$J_k = \frac{I_1 + \cdots + I_k}{I_1 + \cdots + I_m} \tag{2.2}$$

and

$$I_k = \frac{1}{2\pi n} \left| \sum_{t=1}^{n} x_t e^{\frac{i2\pi kt}{n}} \right|^2, \quad k = 1, \ldots, m \tag{2.3}$$

of a sample $x_1, \ldots, x_n$ from white noise approximately has the same distribution as an ordered sample from a uniform distribution, common goodness-of-fit statistics such as the Kolmogorov-Smirnov statistic or the Anderson-Darling statistic can be used to test the white noise hypothesis

$$H_0: \quad Ex_s = Ex_t, \quad Var(x_s) = Var(x_t), Cov(x_t, x_{t+k}) = 0 \ \text{ if } \ k \neq 0.$$

Under the additional assumption that $x_1, \ldots, x_n$ are Gaussian, the tests are exact. In the next section, we will apply these tests to stock index returns $x_t = \log(y_t/y_{t-1})$, $t = 1, \ldots, n$, obtained from daily stock index data $y_t, t = 0, \ldots, n$.

Using a very general setting, Durlauf (1991) examined the convergence of the normalized cumulative periodogram as a random function and used the Continuous Mapping Theorem to establish the asymptotic behavior of certain statistics that map a random function into a scalar random variable. Durlauf's asymptotic theory is robust to many forms of heteroskedasticity and applies particularly also to tests based on common statistics such as the Kolmogorov-Smirnov statistic and the Anderson-Darling statistic. Durlauf (1991) also established the consistency of these tests against all MA alternatives.

In contrast, other popular tests focus only on a certain subset of frequencies and can therefore not be consistent against all alternatives. For example, it

follows from

$$\begin{aligned}\frac{1}{k+1}Var(x_1+\cdots+x_{k+1}) &= \frac{1}{k+1}(k\gamma_x(0)+2(k-1)\gamma_x(1)+\cdots+2\gamma_x(k)) \\ &= \gamma_x(0)+2\sum_{j=1}^{k}(1-\frac{j}{k+1})\gamma_x(j) \qquad (2.4)\end{aligned}$$

that the variance ratio statistic

$$V^k = \frac{\hat{Var}(x_1+\cdots+x_{k+1})}{(k+1)\hat{Var}(x_1)} \qquad (2.5)$$

can (apart from a multiplicative constant) be interpreted as an estimator for the normalized spectral density,

$$f_x(\omega) = \frac{1}{2\pi}(1+\sum_{j=1}^{\infty}\rho_x(j)\cos(\omega j)), \qquad (2.6)$$

at frequency zero (see, e.g., Cochrane, 1988, Lo and MacKinlay, 1988). Here γ_x and ρ_x denote the autocovariance function and the autocorrelation function, respectively, of the regular stationary process x_t. Thus, the use of this statistic can only be justified if there is a concrete suspicion that the most significant deviations from the null hypothesis occur in the very low frequency range.

Of course, nice asymptotic properties such as the consistency against all MA alternatives do not always imply high power in finite samples. Because of the weak performance of the ordinary Kolmogorov-Smirnov test in the case of multimodal alternatives, Reschenhofer (1997) introduced generalized Kolmogorov-Smirnov tests KS^j, $j = 2,3,4,\ldots$, which are designed for alternatives with $1,2,3,\cdots$ peaks. The test KS^j rejects the null hypothesis whenever the maximum sum of $j-1$ local deviations from uniformity is too large. In order to obtain a test that performs reasonably well in a wide range of alternatives, Choi (1999) proposed to combine the p-values obtained from k different, one-sided tests into the single test statistic

$$T = -2\sum_{j=1}^{k}\log(p_j) \qquad (2.7)$$

and Reschenhofer (2008) proposed to combine the $k-1$ generalized Kolmogorov-Smirnov tests $\mathrm{KS}^2, \ldots, \mathrm{KS}^k$ via their p-values into the single test statistic

$$\mathrm{KS}^{2-k} = \min_{2\le j\le k} p_j$$

Although the combined tests are extremely competitive in a wide range of distinctive, possible multimodal alternatives, it may be expected that they are not

useful in the case of returns spectra, which are either constant or at best characterless and fuzzy. In such a case we do not need sophisticated tests that look for specific spectral patterns. Simple tests that are based on vague alternatives may be more appropriate. Perhaps the weakest suspicion that we can have about returns spectra is that lower frequencies play a more important role than higher frequencies. A matching test statistic is given by

$$F^* = \frac{I_1 + \cdots + I_{[m/2]}}{I_{m-[m/2]+1} + \cdots + I_m}. \tag{2.8}$$

Under the null hypothesis of Gaussian white noise, this statistic has an $F(2[m/2], 2[m/2])$ distribution. Of course, if there is no excessive power in the low frequency range, the test F^*, which rejects the null hypothesis whenever the statistic F^* is too large, will be totally insensitive.

3. Application to Financial Data

Experience has taught us that neither a good performance in simulation studies nor nice theoretical properties such as consistency and robustness can ensure success in a concrete application. It is therefore a priori not clear how the tests discussed in the previous section will perform when they are applied to financial data. In general, it makes more sense to use stock indices rather than individual stocks to investigate market efficiency, because index futures can be bought or sold in large volumes without affecting the price. Perhaps the most important stock index is the S&P 500 index, which represents a broad set of stocks and has a long history. For our analysis we downloaded the daily S&P 500 index from January 3rd 1950 to March 30th 2007 from Yahoo! Finance. Because the characteristics of this time series, $y_0, \ldots, y_N$, change over time, we compare the performance of the different frequency domain tests for white noise in a rolling analysis using overlapping segments of $n = 41$ ($m = 20$) and $n = 101$ ($m = 50$) returns. For each n, the first segment contains the returns $x_t = \log(y_t/y_{t-1}), t = 1, \ldots, n$, the second segment contains the returns $x_t = \log(y_t/y_{t-1}), t = 2, \ldots, n+1, \ldots$, and the last segment contains the returns $x_t = \log(y_t/y_{t-1})$, $t = N-(n-1), \ldots, N$. We apply the ordinary Kolmogorov-Smirnov test KS^2, the generalized Kolmogorov-Smirnov tests KS^3 and KS^4, the combined test KS^{2-4}, the Anderson-Darling test AD, and the F-test F^* to each segment at the 5% level of significance. For each test T and each segment S_j let R_j^T denote the outcome. R_j^T has the value one if the null hypothesis is rejected and zero if the null hypothesis is not rejected. Because of the non-stationarity of the time series we do not simply report the total number of rejections for each test. Instead, we determine for each $t = 1, \ldots, N-(n-1)$

the cumulative number of rejections up to that point, i.e.,

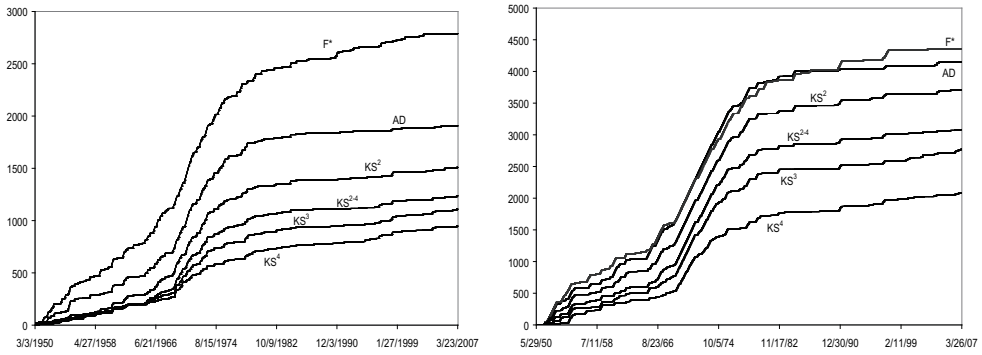
$$C_j^T(t) = \sum_{j=1}^{t} R_j^T \tag{3.1}$$

The cumulative numbers of rejections are plotted for $n = 41$ and $n = 101$ in Figures 1a and 1b, respectively. In the first three decades, in which the returns exhibit considerable positive autocorrelation, all tests are able to detect deviations from whiteness. The rejection rates clearly exceed the significance level. So what we observe is definitely power and not just the type I error. But there are also striking performance differences. Both the ordinary Kolmogorov-Smirnov test KS^2 and the Anderson-Darling test AD, which gives more weight to the tails than KS^2, outperform the more sophisticated tests KS^3, KS^4, and KS^{2-4}. In this application, it is definitely not worth looking for complex alternatives. The simplest test, F^*, is the most powerful. However, its competitive position worsens as n increases. For $n = 101$ ($m = 50$) the Anderson-Darling test is already almost as powerful as F^*.

The performance differences occur only up to the late 70's. Afterwards all tests except F^* are equally bad. Their power even falls below the level of significance. This anomaly becomes apparent when the cumulative net rejections of the null hypothesis are plotted (see Figures 2a and 2b). The net rejections are obtained by subtracting the rejections corresponding to the level of significance, i.e.,

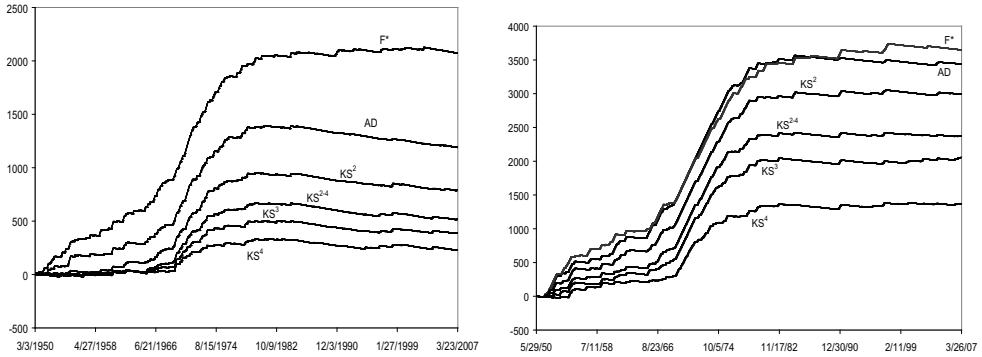
$$c_j^T(t) = \sum_{j=1}^{t} R_j^T - 0.005t \tag{3.2}$$

The fact that the rejection rates are now much smaller than the significance level implies that they must again be interpreted as power and not just as type I error. But which alternatives can produce so few rejections? Any deviation from the null hypothesis of a constant spectrum implies that the spectrum must be higher at some frequencies and lower at others. Of course, it should be much easier for a test to detect real differences rather than spurious ones. So we would hardly expect to observe rejection rates that are lower than the significance level. But if returns series exhibited even fewer distinct patterns than purely random series, the power could indeed fall below the significance level. This might, for example, happen if emerging spurious patterns are over-interpreted by financial analysts and subsequently affected by their trading activities. This line of argumentation is supported by the fact that we observe no comparable power deficiency for the test F^*, which has been designed to detect deviations from whiteness that occur in extremely broad frequency bands and are therefore of little practical value for the construction of profitable trading strategies.



(a) (b)

Figure 1: (a): Cumulative rejections of the null hypothesis at the 5% level for segments of $n = 41$ S&P500 returns by various tests. (b): Cumulative rejections of the null hypothesis at the 5% level for segments of $n = 101$ S&P500 returns by various tests



(a) (b)

Figure 2: (a): Cumulative net rejections of the null hypothesis at the 5% level for segments of $n = 41$ S&P500 returns by various tests. (b): Cumulative net rejections of the null hypothesis at the 5% level for segments of $n = 101$ S&P500 returns by various tests.

4. Concluding remarks

In order to examine the question whether returns are serially uncorrelated or not we must not rely on only one test. Even a test that is consistent against a wide class of alternatives can have low power in certain situations. In principle, the combination of several tests could help lessen this problem. Unfortunately, returns spectra do not exhibit distinctive features such as peaks and troughs. At best we can observe a vague tendency of the spectral densities of returns to be

higher at low frequencies and lower at high frequencies. The results presented in this paper show that such deviations from constancy can be detected much better by a simple test such as F^*, which divides the Fourier frequencies into two halves and compares the sum of the periodogram ordinates at the lower frequencies with the sum of the periodogram ordinates at the higher frequencies. In contrast, the power of more sophisticated tests, which look for more distinctive deviations from constancy, can even be smaller than the level of significance. This is particularly true for todays's returns.

A typical frequency domain test for white noise rejects the null hypothesis whenever the time series contains sinusoidal components with too large amplitudes. Under the null hypothesis of white noise, all amplitudes are roughly of the same size, no amplitudes can be systematically larger than others. There is a perfect uniformity. Amazingly, the amplitudes obtained from the S&P 500 returns appear even more uniform. This super-whiteness can not just be the result of random fluctuations. Something more must be at work besides chance.

A possible explanation is that even spurious patterns have an effect on the strategies of active traders. This explanation is corroborated by the fact that tests looking for fuzzy alternatives, which cannot immediately be used for the construction of trading strategies, do not show the same power deficiency. All in all, it seems that active traders contribute to market efficiency only up to a certain point. Beyond that point, their activities may lead to super-efficiency, which must on no account be misinterpreted as perfect efficiency but rather implies some special form of predictability.

References

Choi, I. (1999). Testing the random walk hypothesis for real exchange rates. *Journal of Applied Econometrics* **14**, 293-308.

Cochrane, J. H. (1988). How big is the random walk in GNP? *Journal of Political Economy* **96**, 893-920.

Deo, R. S. (2000). Spectral tests of the martingale hypothesis under conditional Heteroscedasticity. *Journal of Econometrics* **99**, 291-315.

Durlauf, S. N. (1991). Spectral based testing of the martingale hypothesis *Journal of Econometrics* **50**, 355-376.

Faust, J. (1992). When are variance ratio tests for serial dependence optimal? *Econometrica* **60**, 1215-1226.

Lanne, M. (2002). Testing the predictability of stock returns. *The Review of Economics and Statistics* **84**, 407-415.

Lo, A. W. and MacKinlay, A. C. (1988). Stock prices do not follow random walks: Evidence from a simple specification test. *The Review of Financial Studies* **1**, 41-66.

Patro, D. K. and Wu, Y. (2004). Predictability of short-horizon returns in international equity markets. *Journal of Empirical Finance* **11**, 553-584.

Reschenhofer, E. (1997). Generalization of the Kolmogorov-Smirnov test. *Computational Statistics and Data Analysis* **22**, 433-441.

Reschenhofer, E. (2004a). Unexpected features of financial time series: higher order anomalies and predictability. *Journal of Data Science* **2**, 1-15.

Reschenhofer, E. (2004b). Robust tests of the random walk hypothesis. *Quantitative Finance* **4,** 1-4.

Reschenhofer, E. (2008). Combining generalized Kolmogorov-Smirnov tests. *InterStat*, June #4.

Richardson, M. and Stock, J. (1989). Drawing inferences from statistics based on multiyear asset returns. *Journal of Financial Economics* **1**, 323-348.

Taylor, S. J. (1984). Estimating the variance of autocorrelations calculated from financial time series. *Applied Statistics* **33**, 300-308.

Xu, Y. (2004). Small levels of predictability and large economic gains. *Journal of Empirical Finance* **11**, 247-275.

Received December 27, 2007; accepted March 7, 2008.

Erhard Reschenhofer
Department of Statistics and Decision Support Systems
University of Vienna
Universitätsstr. 5, A-1010
Vienna, Austria
erhard.reschenhofer@univie.ac.at