The Quarterly Review of Economics and Finance 51 (2011) 292–302

Contents lists available at ScienceDirect

The Quarterly Review of Economics and Finance

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

Independent component analysis for realized volatility: Analysis of the stock market crash of 2008

Andrew Kumiega[a,b,1], Thaddeus Neururer[a,2], Ben Van Vliet[b,*]

[a] *Infinium Capital Management, LLC, 600 W., Chicago, IL 60654, United States*
[b] *Stuart School of Business, Illinois Institute of Technology, 565 W. Adams, Chicago, IL 60661, United States*

ARTICLE INFO

Article history:
Received 13 April 2010
Received in revised form 10 February 2011
Accepted 11 March 2011
Available online 21 March 2011

JEL classification:
C32
G17

Keywords:
Independent component analysis
EGARCH
Equity index volatility

ABSTRACT

This paper investigates the factors that drove the U.S. equity market returns from 2007 to early 2010. The period was highlighted by volatile energy and commodity prices, the collapse of insurance and banking firms, extreme implied volatility and a subsequent rally in the overall market. To extract the driving factors, we decompose the returns of the S&P500 sector ETFs into statistically independent signals using independent component analysis. We find that the generated factors have interesting financial interpretations and are consistent with the major economic themes of the period. We find that there are two sets of general market betas during the period along with a dominant factor for energy and materials sector. In addition, we find that the EGARCH model which accommodates asymmetric responses between returns and volatility can plausibly fit the high levels of variance during the crash. Finally, estimated correlations dropped when commodity prices moved higher, but then spiked when the S&P500 crashed in late 2008.

1. Introduction

Independent component analysis (ICA) is a technique for extracting factors from a set of mixed sources of variation. It does so by maximizing the statistical independence of the components. We chose to apply ICA to data on S&P500 index sector exchange traded funds (ETF) to see if ICA could separate out key drivers of performance over the (extremely volatile) period 2007 to early 2010. *A priori* we did not expect the procedure to produce factors that relate back to market structure. Nevertheless, the extracted signals show commodity prices as the key driving economic signals.

Over the period 2007 through early 2010, several macroeconomic shocks hit the global economy, causing the stock market to fall over 50% from its all-time high in late 2007. In addition, implied volatilities – sometimes referred to as fear indices – spiked to levels unseen since the crash of 1987. The volatility indices of both the S&P500[3] (VIX) and Nasdaq 100 (VXN) peaked around 80 in the fall of 2009. These extreme levels imply an expectation of daily volatility of around 5% for a full month hence.[4] By comparison, the VIX was around 12 (a daily volatility of .75%) at the beginning of 2007, and by the end of February 2010 had dropped back to near 20 (a daily volatility of 1.25%).

During this period, commodity prices were also highly volatile relative to their historical levels. Hard red spring wheat rose from around $5 per bushel in 2007 to over $25 in early 2008, only to fall back to the $5 range by the end of 2009. Brent crude reached over $145 US per barrel in July of 2008 then quickly fell back to between $40 and $50 by the end of that year. Natural gas prices spiked near $14 per mmBtu, but by mid-2009 dropped to $3 per mmBtu. Silver fell from over $20 per ounce to less than $10 in late 2008. The price of copper fell over 62%, from over $4 per pound to the $1.50 range.

In this paper we also attempt to fit general autoregressive conditional heteroskedasticity (GARCH) models utilizing these ICA factors to replicate the volatility of the period. We find that the standard GARCH(1,1) cannot fit the variances without parameters that imply unstable variance processes. However, Nelson (1991) asymmetric exponential GARCH (EGARCH) model *does* fit the fac-

* Corresponding author at: Stuart School of Business, Illinois Institute of Technology, 565 W. Adams, Chicago, IL 60661, United States. Tel.: +1 312 906 6513.
E-mail addresses: akumiega@iit.edu (A. Kumiega), Thaddeus.Neururer@InfiniumCM.com (T. Neururer), bvanvliet@stuart.iit.edu (B. Van Vliet).
1 Tel.: +1 630 484 3301.
2 Tel.: +1 312 770 1712.

3 S&P500, Select Sector SPDRs, and S&P GSCI are registered trademarks of the McGraw-Hill Companies, Inc.
4 These volatility index values represent the square root of a thirty-day constant maturity variance swap.

1062-9769/$ – see front matter
doi:10.1016/j.qref.2011.03.002

tor variances while still providing finite stationary estimates. Using the EGARCH estimates for each factor, we are able to estimate the sector volatilities as well as the correlations between the sectors during the sample period.

The remainder of this paper is organized into five sections. The first section provides an overview of the ICA technique. The second describes the various GARCH processes. The third section, presents the empirical data used in the study. The fourth section presents fitting results. The fifth section concludes.

2. Independent component analysis

ICA is a method of extracting factors from sources of variation that are linearly mixed, as in blind signal separation. Unlike principal component analysis (PCA), which only removes correlations in a data set, ICA finds factors by maximizing the statistical independence of the components.[5] Through this process it produces a mixing matrix which allows the factors to be recombined to reproduce the observed signals.

For our analysis the observed signals are the log returns of the sector ETFs – namely, the Select Sector SPDRs (SSSs) – of the S&P500. To extract the statistical factors that explain the signal variations, ICA requires no information about the factors themselves, only requiring that at most one of the sources be Gaussian, and that at least as many mixed (observed) signals as factors exist. We limit ourselves to using fewer factors than sectors to focus on the main factors of the period. Further, because we are using daily frequency equity returns, our factors should meet ICA's non-Gaussianity requirements (see Mandelbrot (1963) or Fama (1965) for early papers rejecting the Gaussian hypothesis).

We assume that the set of multivariate signals $y_i(t), i = 1, 2, \ldots, n$ are generated by a linear mixture of factors which we denote $s_i(t)$. In matrix notation, at each t we have:

$$y(t) = As(t) \tag{1}$$

Here $y(t)$ is the set of observed signals at time t, A is the assumed static mixing matrix, and $s(t)$ is the set of factors at time t. Given that we have collected the time series of y_i's, the goal is to un-mix the time series of signals along with estimating the mixing matrix, so that:

$$x(t) = Wy(t) \tag{2}$$

If we attempt to estimate all the source signals, then the W and A matrices will be $[n \times n]$ where n is the original number of observed multivariate signals. The goal is to have $x(t)$ be close to $s(t)$ and, if it is done perfectly, then $A = W^{-1}$. In our case we will be seeking a subset of the factors, so the estimated W and $x(t)$'s will be chosen so as to maximize the explained variance of the observed signals $y(t)$.

There are two points of ambiguity in ICA. First, given the set up presented in (1) and (2), the factors generated have no ordering information. In execution, this does not usually present a problem. Second, ICA is not capable of determining the variances of the independent components. This is because, for estimation purposes, both the variances of the factors and the elements of the mixing matrix are unknown. For any factor one could multiply it by a scalar, divide its associated column by that same scalar, and end up with the same observed signals. Thus, most ICA signals produced by the estimation algorithms will have their variances set to one, and rely on the mixing matrix to control the absolute levels of variance (for more detailed explanations of ICA, see Hyvärinen, Karhunen, & Erkki (2001) and Stone (2004)).

To estimate the ICA factors and the mixing matrix, we use the FastICA algorithm (see Hyvärinen (1999) or Hyvärinen and Oja (2000)). FastICA first centers the data and then whitens it. Thus, the observed signals are linearly transformed into a new set of signals whose components are uncorrelated and their variances equal to one. This is done by projecting the data onto its principal component directions.[6] Next, the algorithm attempts to find both the un-mixing matrix and factor signals by maximizing the negentropy[7] approximation along with other constraints. By maximizing estimated negentropy, we maximize the non-Gaussianity of the factors. As noted by Hyvärinen and Oja, by optimizing the non-Gasussiantity, the algorithm will find the statistically independent factors and weighting matrix that correspond to the solution of the ICA problem. There are other algorithms for ICA estimation and we do not investigate the difference between them when applied to our data set, but García-Ferrer, González-Prieto, and Peña (2008) showed that the FastICA method works well with financial and artificial data.

For our tests, the log cosh function is used for the estimation of negentropy where $\alpha = 1$:

$$G(\mu) = \frac{1}{\alpha} \cdot \log \quad \cosh(\alpha\mu) \tag{3}$$

ICA has only been developed in the last twenty years with some of the estimation procedures only being developed in the last ten years. However, there have been some studies that have used ICA to study multivariate time series in finance and economics. For early examples, see Back and Weigend (1997), Kiviluoto and Oja (1998), and Cha and Chan (2000). In addition, there has been some recent work using ICA combined with stochastic volatility or GARCH models. At times, the term ICA-GARCH has been applied to differentiate it from other multivariate GARCH methods. (For example, see and García-Ferrer et al. (2008) and Xu and Wirjanto (2009) for ICA-GARCH.)

While ICA does make a number of strong assumptions (such as the static mixing matrix and independent driving factors), which may be invalid in an economic setting, the method has been used with success in some other financial studies. Wu, Yu, & Li (2006) find ICA-GARCH outperforms similar methods in Value-at-Risk estimations. In addition, Chen, Härdle, & Spokoiny (2010) find that using ICA factors combined with adaptively fitted generalized hyperbolic distributions provides accurate Value-at-Risk estimates. While neither of the two papers mentioned is concerned with the economic interpretation of the independent factors or the modeling of stochastic (asymmetric) volatilities or correlations *per se*, the fact that each paper successfully uses ICA for factor extraction for risk management purposes lends support to our use of ICA for this volatile market period.

Other methods have been proposed for modeling the volatility of multivariate time series. The Diagonal–Vech model (Bollerslev,

[5] The variables y_1 and y_2 are independent if there is no information in the value of y_1 that gives any information on the value of y_2, and vice versa. This definition extends naturally for any number n of random variables, in which case the joint density must be a product of n terms. Let $p(y_1, y_2)$ be the joint probability density function (pdf) of y_1 and y_2 and, and let $p_1(y_1)$ and $p_2(y_2)$ be the marginal pdfs of y_1 and y_2 respectively. Then y_1 and y_2 are independent if and only if the joint pdf can be expressed as: $p(y_1, y_2) = p_1(y_1) \cdot p_2(y_2)$.

[6] Although whitening is commonly done in ICA analysis, there is an assumption that the source signals are not associated with the principal components which explain a smaller amount of the sample variance. See Green, Nandy, and Cordes (2002) for an example where discarding the smaller principal components leads to the wrong results.

[7] Negentropy is a measure of the distance to normality. A Gaussian signal has a normal distribution. But, negentropy is always non-negative. Therefore, it will vanish if the signal is Gaussian.