Application of sparse principal component analysis to high frequency trading data

Giovanni M. Merola
Xi'an Jioatong Liverpool University, Suzhou, China PR - giovanni.merola@xjtlu.edu.cn

Abstract

Principal Components Analysis (PCA) is frequently used for determining the most important factors that drive the variability of investment returns for a set of assets in a common market. In this paper we apply sparse PCA to intraday returns of futures traded in China to investigate the correlation between different futures segments. We show that commodity futures are mildly negatively correlated with stock futures and uncorrelated with interest rate futures. To our knowledge this is the first application of SPCA to high frequency data and the results are very promising.
Keywords: futures; intraday returns; realized returns.

1 Introduction

Future contracts (futures) are standardized forward contract in which the parties agree to buy and sell an asset for a price agreed upon today, with delivery and payment occurring at a future point, the delivery date. Since these contracts can be easily traded they are heavily used by investors as investment assets, as well as by actual consumers for hedging their risk. Futures are available for a large number of commodities and also for financial instruments, such as stock market indices, currency exchange rates and interest rates. Therefore they are usually divided into different *segments*, such as "Metals", "Agricultural", "Chemicals", "Financial" and so on.
Starting from the 2004 deregulations of the futures market, commodity futures have become a new widely traded assets class for portfolio investors. The intervention of speculators has increased the volume of trades, which is often referred to as the "financialization" of the commodity markets (e.g. Cheng and Wei Xiong, 2014). Consequently, the average returns of commodity markets are now comparable to those of equities (not from spot returns) and futures now represent a valid investment diversification instrument. Gorton and Rouwenhorst (2006) and Erb and Harvey (2006) found that the commodity futures market is negatively correlated with equity and bond and positively with inflation.
The characteristics of the returns of futures of different commodity segments were traditionally considered to be different, especially because of differences in seasonality patterns and volumes traded (therefore liquidity and volatilities). However, the large inflow of index investment registered after 2004 integrated the previously segmented commodity markets with each other and with outside financial markets. For example, it has been observed (Cheng and Xiong 2012) that the financialization of the commodity futures market:

- Largely increased correlations between individual commodities;
- Largely increased correlations between commodities and stocks;
- Created an affine structural model of commodity futures prices.

The points discussed above were taken up by some researchers who investigated the existence of common factors within commodity futures markets (among others Christoffersen et al. 2004, Vansteenkiste 2009 and Byrne et al. 2012) using principal component techniques to extract a latent factor either at global or specific sector level. These analyses were carried out using daily returns, specifically the logarithms of the realised returns.
In this paper we investigate the existence of factors for the commodity futures market within China using intraday with frequency 1-minute data. The current interest in high frequency data was largely spurred by Andersen and Bollerslev (1998b) who used the realized variance to show that standard volatility models deliver accurate forecasts. High-frequency transactions give an insight of the dynamics of trading which are otherwise lost in the daily close data. For example, Hansen and Lunde (2011) list six ways in which high frequency data can improve volatility estimation.

Intraday data measured at regular intervals (as opposed to being measured when ticks are reached) present several intervals where no trades took place, giving zero returns. In order to decrease the proportion of zero returns, we consider 5-minute returns and use the 1-minute prices to estimate their volatility. Following the most popular practice, in this preliminary analysis we analyse the logarithms of the realised returns, defined as defined as $r_t = log(P_t/P_{t-1})$. These are known to be approximately normally distributed and time uncorrelated after volatility adjustment.

A popular method used to compute market factors is to apply Principal Components Analysis (PCA) to the estimated stable (variance and) covariance matrix for the returns of several different commodities. However, it is difficult to use the principal components (PCs) for valuation, portfolio construction, and risk control because, since they are combination of all the observed variables, they cannot be easily interpreted because (e.g. Fabozzi et al. 2007). In recent years several sparse PCA (SPCA) methods have been proposed to estimate approximations of the PCs that are combination of only a few of the variables, which are easier to interpret. Merola (2015, 2017) shows that conventional SPCA methods do not achieve good approximations and include highly correlated variables in the solutions. Therefore, we compute sparse PCs using Projection SPCA (Merola 2017) which are guaranteed to give a good approximation to the PCs and the exclusion of highly correlated variables from the solutions.

We considered the prices of 36 futures contracts traded in China taken on 335 days between 15/05/2015 and 2/11/2016. The PCs computed for only the 31 commodity futures contract have only a mild negative correlation with futures of stock indices and are virtually uncorrelated with futures on interest rates. These last futures are also uncorrelated with those for the stock market.

2 Sparse Principal Components Analysis

Given a matrix $\mathbf{X}$ containing n observations on p variables, the $d < p$ first PCs are linear combinations of the variables, defined by $\mathbf{u}_j = \mathbf{X}\mathbf{a}_j$, $j = 1, \ldots, d$, where the elements of the vector $\mathbf{a}$ are called *loadings*. The PCs are the mutually orthogonal components that sequentially minimise the least squares criterion

$$\min \sum_{j=1}^{d} ||\mathbf{X} - \mathbf{u}_j(\mathbf{u}_j{}^\intercal \mathbf{u}_j)^{-1}\mathbf{u}_j{}^\intercal \mathbf{X}||^2$$
$$\text{subject to } \mathbf{u}_i{}^\intercal \mathbf{u}_j = 0,\ i < j.$$

From the definition above it is easy to see that the PCs are obtained by maximising the squared norm of the projection onto the components, $||\widehat{\mathbf{X}}(\mathbf{T})||^2$. This quantity is referred to as *variance explained* by the d components and is equal to the sum of the first d largest eigenvectors of the covariance matrix of $\mathbf{X}$. PCA is more commonly known under Hotelling's definition by which the PCs are the linear combinations with unit norm coefficients that have maximal norm. As explained in Merola (2015 and 2017) this definition is misleading when sparsity constraints are applied.

Merola (2015) gives the solution that yields the best rank deficient approximation of $\mathbf{X}$ under sparsity constraints, under the name of least squares SPCA. Approximated least squares SPCA solutions can be efficiently computed with Projection SPCA (Merola, 2017) in which the sparse components $\mathbf{t}_j$ are the solutions of

$$\min \sum_{j=1}^{d} ||\mathbf{X} - \mathbf{t}_j(\mathbf{t}_j{}^\intercal \mathbf{t}_j)^{-1}\mathbf{t}_j{}^\intercal \mathbf{X}||^2$$
$$\text{subject to } \mathit{cardinality}(\mathbf{t}_j) < p \,\text{and}\, ||\hat{\mathbf{u}}_j(\mathbf{t}_j)||^2 > \alpha||\mathbf{u}_j||^2,$$

where $\hat{\mathbf{u}}_j(\mathbf{t}_j)$ is the projection of the PC $\mathbf{u}_j$ onto the sparse PC $\mathbf{t}_j$, the *cardinality* is the number of variables in the combination and $0 < \alpha \leq 1$ is a tuning parameter.

3 Data analysis

We had available 36 series of minute prices which, after merging, spanned the period from May 15, 2015 to November 2, 2016. The series present some gaps because the 4 existing commodity futures exchanges operating in China have different working hours. This is not a problem for us because we consider 5 minutes returns and exclude the returns over longer periods of time. All together we had 65325 prices with which we computed 13265 5 minute returns. The contracts are divided into 5 segments: Agriculture (11), Energy (1), Financial (5), Industrial (10) and Metals (10), as shown in Table 1.

Table 1: Future contracts considered in this study.

Symbol	Commodity	Sector	Symbol	Commodity	Sector
a	No.1 Soybeans	Agriculture	FG	Glass	Industrial
c	Corn	Agriculture	i	Iron Ore	Industrial
CF	cotton	Agriculture	j	Coke	Industrial
cs	Corn Starch	Agriculture	jm	Hard Coking Coal	Industrial
jd	egg	Agriculture	l	LDPE	Industrial
m	Soybean Meal	Agriculture	MA	Methanol	Industrial
OI	Rapeseed Oil	Agriculture	pp	Polypropylene	Industrial
p	RBD Palm Olein	Agriculture	ru	Natural Rubber	Industrial
RM	Rapeseed Meal	Agriculture	TA	PTA	Industrial
SR	white sugar	Agriculture	v	PVC	Industrial
y	Soybean Oil	Agriculture	IC	CSI 500 Stock Index Futures	Financial
bu	Bitumen	Energy	IF	CSI 300 Stock Index Futures	Financial
ag	Silver	Metals	IH	SSE 50 Stock Index Futures	Financial
al	Aluminum	Metals	T	10-year Treasury Bond Futures	Financial
cu	Copper	Metals	TF	5-year Treasury Bond Futures	Financial
hc	Hot Rolled Coils	Metals			
ni	Nickel	Metals			
pb	Lead	Metals			
rb	Rebar	Metals			
sn	Tin	Metals			
zn	Zinc	Metals			

The rolled prices were converted into log-realised 5-minute returns which, following a pragmatic approach, were winsorised with a threshold of 3.5 IQR units away from the median. The volatilities of the five minute returns was first estimated as the standard deviations of the 1-minute log-realised returns, which were then smoothed using a robust version of an exponentially weighted moving average with parameter 0.93 (this last approach is suggested in Fabozzi et al., 2007). Hence, the winsorised log-returns were adjusted by dividing them by the smoothed volatilities. We checked the distribution of the adjusted series and we were satisfied that they were sufficiently well behaved, with the exception that many presented, as expected, a greater proportion of zero values than normal for a Gaussian or t distribution.

The pairwise correlations between series shown in Figure (1) show that there is a separation between financial and commodity futures. The interest rates futures are virtually uncorrelated with the other contracts while the stock indices futures show a mild negative correlation with commodities.

Figure (2) shows the scatter plot and the densities of the first two sparse and full cardinality PCs. The sparse components were computed so as to explain at least 95% of the variance explained by the PCs. The correlation between sparse and non-sparse PCs is about 0.97 for both pairs and the densities are approximately normal, where the full cardinality PCs show a larger variance due to the inclusion of redundant variables that account for more noise.

The loadings of the first two full cardinality and sparse PCs are shown in Figure (3). Clearly, the sparse PCs are more useful for identifying important futures. The only financial contract in the first sparse PC is the CSI 300 Stock Index. Instead, the second SPC is dominated by financial contracts.

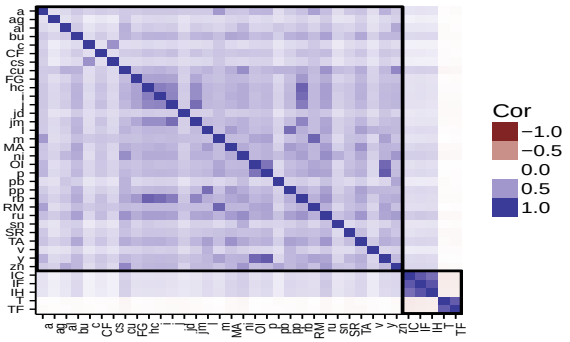

Figure 1: Correlation between log-realised returns.

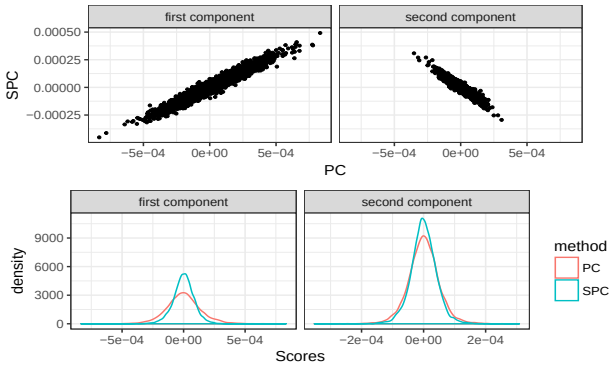

Figure 2: Scatter plots and densities of the first two full cardinality and sparse PCs for all contracts.

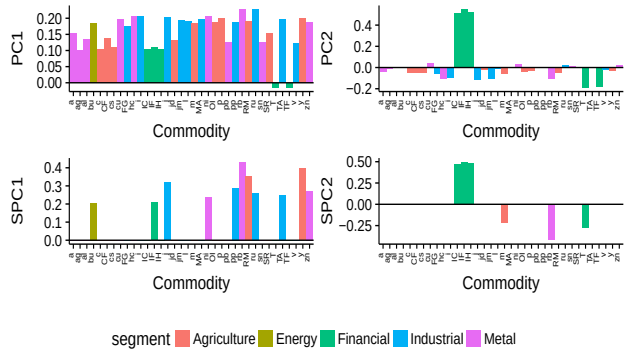

Figure 3: Loadings of the first two full cardinality and sparse PCs of all contracts.

Table 2 shows the correlation of the financial contracts with the first two SPCs.

Table 2: Correlation of the financial contracts with the first two SPCs of all contracts

	IC	IF	IH	T	TF
first SPC	0.342	0.368	0.342	-0.040	-0.043
second SPC	0.787	0.830	0.795	-0.302	-0.254

The PCs computed only for the non-financial contracts can be used to determine whether there exists correlation between commodity and financial contracts. The loadings of the full cardinality PCs and those of the sparse PCs (also computed so as to explain at least 95% of the variance explained by the PCs) are shown together in Figure (4).

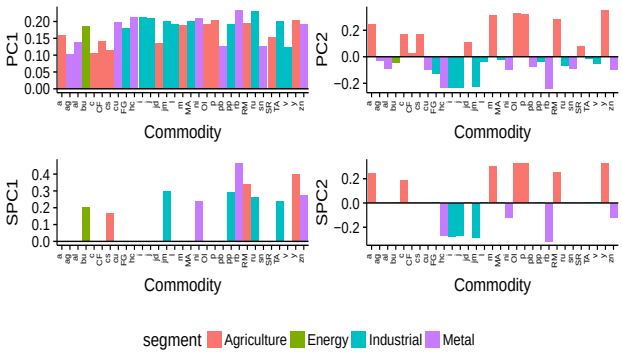

Figure 4: Loadings of the first two full cardinality and sparse PCs of only commodity contracts.

As shown by the plots in Figure (5), only the first PC of the commodity contracts is mildly negatively correlated with stock indices futures. The interest rate contracts are not correlated with any of the PCs.

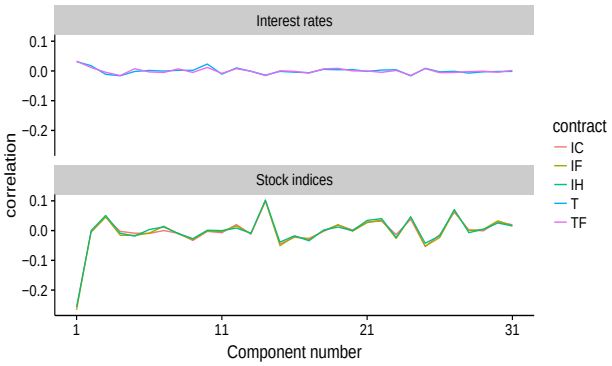

Figure 5: Correlations between the sparse PCs of only commodity futures with the financial futures.

4 Conclusions

We show that, in the Chinese market, high-frequency returns from commodity futures are not correlated with those from interest rate futures. Also, the returns from stock indices futures only present a mild negative correlation with those from commodity futures and are not correlated with those from interest rates futures. Sparse principal components analysis is a valid tool for identifying a few important assets that drive the variability of whole markets. To our knowledge this is the first application of SPCA to high frequency data; from the results that we present in this paper it seems to be very promising.

References

Andersen, T. G. and Bollerslev, T. (1998b), Answering the skeptics: Yes, standard volatility models do provide accurate forecasts, International Economic Review 39(4), 885905.

Byrne, J. P., Fazio G. and Fiess N. (2012). Primary commodity prices: Co-movements, common factors and fundamentals. Journal of Development Economics, 101, 1626.

Cheng I. and Xiong W.(2014). The financialization of commodity markets. The Annual Review of Financial Economics 2014. 6:41941 doi:10.1146/annurev-financial-110613-034432

Christoffersen P., Lunde A. and Olesen K.(2014) Rotman School of Management Working Paper No. 2495779 available at: `https://papers.ssrn.com/sol3/papers.cfm?abstract_id=2495779`

Erb C, Harvey C. 2006. The strategic and tactical value of commodity futures. Financ. Anal. J. 62(2):6997

Fabozzi F., Kolm P., Pachamanova D. and Focardi S. (2007). Robust Portfolio Optimization and Management. Wiley.

Hansen, P. R. and Lunde, A. (2011), Forecasting volatility using high-frequency data, in M. Clements and D. Hendry, eds, The Oxford Handbook of Economic Forecasting, Oxford: Blackwell, chapter 19, pp. 525556.

Hotelling, H. (1933). Analysis of a Complex of Statistical Variables with Principal Components. Journal of Educational Psychology 24, 498520.

Merola, G. M. (2015). Sparse Principal Component Analysis: a Least Squares Approach. Australian & New Zealand J. of Stats. 57(3). Preprint available at http://arxiv.org/abs/1406.1381

Merola, G. M. (2017). Projection Sparse Principal Component Analysis: an efficient least squares method. Accepted for publication Computational Statistics and Data Analysis.
Preprint available at https://arxiv.org/abs/1612.00939

Vansteenkiste, I. (2009). How important are common factors in driving non-fuel commodity prices? A dynamic factor analysis. Working paper, European Central Bank.