

The Dynamics of Commodity Prices: A Clustering Approach

Özge Savaşçın
Department of Economics
University of North Carolina, Chapel Hill

Keywords: Commodity Prices, Comovement,
Endogenous Clustering, Dynamic Factors, Metropolis-Hastings
within Gibbs sampling.

November 14, 2011

Abstract

This paper uses an endogenously clustered dynamic factor model to gain a better understanding of commodity price comovements and their determinants. From a large dataset of commodity prices I extract the fundamental sources behind the price dynamics and find that commodity price comovements are mostly the result of sparse cluster factors that represent correlations of distinct group of commodities. Endogenous clustering of these groups does not represent the standard narrow classifications (indexes) of commodity prices as defined by statistical agencies (e.g IFS, BLS). Characterization analysis on these factors identifies a wide range of macroeconomic variables like crude oil prices, fertilizer prices, and the federal funds rate as possible sources of commodity price comovements.

1 Introduction

In recent years the world has witnessed a commodity boom that has raised several questions and various explanations about the characteristics of commodity prices. Prices of grains such as corn, soybeans, wheat, and rice more than doubled during the 2006 – 2008 peak period. Crude oil prices reached \$147 per barrel in July 2008 almost five times higher than what it was in 2003. The surge in prices created worldwide concern over energy costs and food security. The economic and social aspects of these price increases have led many researchers to speculate about the fundamentals of commodity prices. Commonly suggested determinants have been easy monetary policies, the devaluation of the dollar, excess liquidity, speculation in commodity markets, and high world demand.

The debates about what drives commodity price comovements had just commenced when the world was hit by the recent global downturn. With a sudden reversal around summer 2008, the soaring energy and food prices fell back to their 2006 values, signaling for a moment that this sudden price upsurge was nothing but a short-run phenomenon. But not long after, a second wave of rapidly increasing commodity prices came about. Beginning in May 2009, another surge in commodity prices is under way, reminding us what Frankel and Rose (2009) had pointed out: “...it cannot be a coincidence that almost all commodity prices rose together during much of the past decade...”

A rise in price of a single commodity would usually reflect something specific to that commodity and would not be informative about the overall economy. However, the synchronized movements of several commodities have different implications that a researcher should care about. These kinds of simultaneous movements could affect headline and core inflation of open economies or create concerns about food security for developing countries. Assume policy makers know which particular group of commodities share cycles and exhibit synchronized inflation and are also aware of the type of factors behind these dynamics. This would give them informational advantage in terms of which variables to carefully watch. If, say, oil prices, world demand, and interest rates are responsible for the upsurge in most of the commodity prices; then during expansionary periods when oil prices are trending up, contractionary monetary policy could dampen the price surge and prevent likely spillovers to core inflation.

The synchronicity of commodity prices is not new in macroeconomics. In a seminal paper, Pindyck and Rotemberg (1990) argue that there is “excess comovement” of seemingly unrelated commodities that cannot be explained by macroeconomic determinants such as interest rates or oil prices. They conclude that it is actually the herding behavior of market participants that causes the comovement in prices. Excessive or not, the consensus is that commodity prices do comove, whether it is through common macroeconomic fundamentals or through complementarity or substitutability in production and consumption or through a set of possible factors.

Most of the existing empirical work takes for granted that *all* commodity prices (or at least the ones defined under specific categories such as food and metals) move together (Baffes (2009), Hockman et al. (2010), Lombardi et al. (2010)). However, none consider how likely some group of commodities comove. If there are multiple factors driving primary commodities, different groups of commodities will share cycles due to different sources. We need to first identify the comovements of the commodities before we begin to talk about the determinants.

In light of above arguments, this paper addresses several questions. Which groups of commodities are likely to share cycles? Is there a common source behind the price comovements or are there multiple forces affecting different groups of commodities? How important are these possible factors or determinants behind the price dynamics and can we characterize them? To answer these questions we need to systematically decipher the correlation structure into its determinants, preferably with an empirical model suited for such an analysis.

The empirical model selection is important when it comes to analyzing the interrelations of many macroeconomic variables. Bernanke et al. (2005) suggest that VAR techniques suffer from a degrees of freedom problem, which puts restrictions on the number of variables that can be included in the system. They further emphasize the importance of dynamic factor models that can summarize the information from a large number of time series by a small set of indexes, or factors.¹ They propose a Factor Augmented VAR model to understand the common dynamics of many variables. In particular they apply a two-step approach: they first uncover the common space spanned by the factors of the data and then run a VAR of these estimated factors on possible determinants as a second step.

¹Using dynamic factor models also provides advantages compared with the simpler cross correlation analysis that has been selected as a tool to investigate synchronous cycles.

Likewise, in this paper I use a dynamic factor model to extract information from all the available non-energy commodity prices. I recognize that some comovements may not be simply due to a global factor like world demand but may also be a result of more sector-specific factors like droughts, floods or biofuel production that affect only smaller groups of commodities.² Novel to the paper is the use of an endogenous clustering procedure on a large data set of primary commodities to study these price dynamics. This approach was first introduced by Francis, Owyang, and Savascin (hereafter FOS, 2011) to study international business cycles.³ Such an approach allows the data to freely choose from a set of possible unobservable factors and define their own groups. The empirical model will allow commodities to share similar cycles beyond that driven by a common global factor – an avenue overlooked by the literature. In particular, the empirical model includes a global factor and several group-specific factors where the groups are not defined a priori. After successfully extracting the fundamentals (factors) behind commodity price movements, I then try to characterize these underlying factors using Bayesian auxiliary least square regressions.⁴

The endogenous clustering analysis reveals that unrelated commodities that belong to the metal, agricultural materials and food families share cycles not only through global determinants but also through cluster factors. The global factor is the most important determinant for only vegetable oils (excluding olive oil), while the cluster factors carry greater importance for the rest of the commodities. Even though the commodity world seems to be coupling overall, there is still a considerable amounts of decoupling of particular commodities. In particular some commodity clusters show similarities in ways identified by specific product characteristics. For example, timber industry isolates itself from the rest of the agricultural raw materials and form a separate cluster. Likewise coffee forms another. Grains and vegetable oils decouple from the rest of the food category products and share most of their correlations through their cluster factor. Overall, commodity clusters found in this paper are not representative of standard narrow classifications (indexes) of commodity prices as defined by statistical agencies like International Financial Statistics (IFS).

²The estimation procedure applied in this paper requires that the determinants of the factors should be outside of the sample when conducting the factor model. Therefore, I focus only on non-energy products and use the crude oil prices in the ex-post analysis to determine their effects on the model factors and look for the validity of the claims made in the studies that list oil prices as the fundamental source of the price correlations.

³Factor models are extensively used to study various topics from international business cycles to regional analysis. Examples include: Kose et al. (2003, 2008), Hamilton and Owyang (2010), Neely and Rapach (2009),

⁴In a work similar to this, Vansteenkiste (2009) uses a clustered approach where she defines 4 groups exogenously for only 11 commodities. She tries only to link the global factor to possible macro factors while neglecting the cluster factors.

This implies that these narrower definitions, or subgroups, by these agencies do not consist of homogeneously moving products.

Bayesian regression analysis reveals that the world demand proxied by industrial production growth of many economies is a determinant of the global shock affecting all commodities. While vegetable oils and grain prices react to crude oil and fertilizer prices, other foodstuff are affected by the same variables as metals and materials. Through the cluster factor, metals, materials, and some food products seem to be reacting to a combination of many potential factors discussed in the literature: namely, interest rates, world demand, oil prices, and stock market indexes. Simple examination of these group-specific commodity comovements implies that in times of high oil prices and high growth in world production, low interest rates can amplify the increases in their price levels and further quantitative easing may indicate higher inflation in commodity prices which could make the economy more vulnerable to inflationary pressures.

The remainder of the paper is as follows: Section 2 provides motivation for the paper. Section 3 presents a review of the literature. Section 4 describes the empirical model and the estimation procedure. Section 5 describes the data. Section 6 reports the findings and finally section 7 concludes.

2 Motivation

Commodity prices carry great importance, with their potential impact on aggregate output and the balance of payments and transmission of business cycle disturbances across countries by connecting commodity exporters and importers from developed to developing countries (Borenzstein and Reinhart, 1994). Changes in commodity price levels can create inflationary pressures on an economy that could make monetary policies harder to conduct. If commodity production constitutes a larger percentage of aggregate output, then their price movements should be taken into account in the design of policy. The same is true even for the monetary authorities that target the core inflation rate, which excludes volatile food and energy prices. For instance, the Fed pays attention to and targets the core inflation rate, claiming that it has resulted in better forecasts than the headline inflation rate over the past 25 years.⁵ It is true that the recent commodity price boom has not been reflected in core inflation largely because of the recent economic downturn, which resulted in strong disinflationary pressures, as the FOMC members expected. But what if the Fed is wrong about the expected moderation in global growth and high commodity prices *do* spillover the core inflation?

Commodities are used as inputs of production in many industries. For example, cotton is a major input for textile industry, which accounts for 4.6 percent of core personal consumption expenditure (PCE) inflation. Again, "Shelter" for the U.S. accounts for around 30 percent of core CPI and "Vehicles" around 6 percent. These groups (shelters, vehicles) include housing materials, equipment, and automobiles that are produced with extensive use of basic commodities such as copper, iron (used in steel production), rubber, timber, and lead. The price surge of these commodities is expected to alter the cost structure of many industries and, hence, create high prices that can heat up an economy's inflation rates. That could make targeting inflation difficult and create an environment in which easy monetary policy could overheat inflation, like back in 1970's.

Identifying comovements of prices allows for diversification of inflationary risk not just for monetary authorities. For example, if economic agents in a commodity-exporting country were to know which commodities are likely to experience price increases and the degree to which commodities

⁵Targeting core inflation has its own debate. Any central bank that wants to reconnect with households and businesses, which care more about food and energy price changes than bankers and hedge funds, should target headline inflation as suggested by Bullard (2011).

comove, then these agents (households and/or financial institutions) could diversify some of the risk by expanding the range of export commodities they invest in, sell, or hold. They could diversify by trading in commodities that have weak linkages and do not share common cycles with the commodities they currently export – a point stressed by Cashin et al. (1999). In financial markets, participants can settle their portfolio and investment decisions securely if information about comovements of commodities is known. Moreover, Lu and Neftci (2008) examine the use of commodity options to hedge against the vagaries of international commodity prices for developing nations.

Given the importance of commodity prices, any kind of theory that aims to investigate the policy implications of commodity price dynamics should rely on detailed empirical investigations. Without a diagnosis of the cause of price peaks, we cannot talk about policies that may alleviate the costs of price increases or take precautionary actions to prevent large fluctuations in prices that may result in a crisis.

Recent literature has looked for possible explanations of what has been driving the synchronized commodity price movements. Several factors are considered, from global factors such as high global demand to more market-specific factors such as the rise in biofuel production.⁶

The widely accepted view is that the correlations across commodities are solely a result of common factor(s) (Byrne et al. (2011), Vansteenkiste (2009), Cashin et al. (2002), Lescaroux (2009)).⁷ This may seem plausible at first, but it is incomplete at best. As Foerster et al. (2011) argue, additional cross correlations of any kind could contaminate the global factor, and if not taken into account, can lead to overestimation of the true nature and the contribution of common factors in explaining cross-product comovements. Using disaggregate industrial production data, Foerster et al. (2011) show that the common factors are contaminated by the unmodelled sectoral linkages. Likewise, common factors behind commodity price dynamics may reflect not only global shocks but also the propagation of idiosyncratic shocks within particular groups, usually by way of less pervasive factors.

⁶To cite a few: Krugman (2008), Wolf (2008), Frankel (2005, 2008), Calvo (2008), Lombardi et al. (2010), Baffes and Haniotis (2010), Lescaroux (2009), Cashin, McDermott and Scott (2002), and Vansteenkiste (2009). Hockman (2010) and Mueller (2011) look also at the biofuel effect on food commodities.

⁷Oil prices have been cited as the classic example of a common factor. As almost all industries are energy dependent (even when oil is not used directly in production, it is used in transportation), oil prices feed into the cost functions of almost all commodities.

The sparse factors can be thought of first as reflecting the different market properties across commodities. Shocks related to those specific markets may not spill over to other industries. In particular, shocks that generally emerge from climatic conditions or adverse weather such as floods and droughts directly affect agricultural products while their propagation into the mining industry is less likely. Similarly, for metals and even for some agricultural raw materials, such as timber or rubber, fertilizer costs may not be as relevant.

The traditional way to introduce the sparse industry factors is to exogenously group similar commodities; for example, one could ex-ante define “Food” and “Energy” clusters. While plausible, this may not be the best practice. Even within the same ex-ante categories we might have different underlying driving factors behind the commodity movements. In other words, assuming one “Food” sector will not allow for possible within-sector heterogeneity of particular commodities. Besides, seemingly unrelated commodities are argued to exhibit excess comovement.⁸ Therefore if the ex-ante grouped commodities are "closely related", such groupings would not allow for unrelated commodities to share cycles other than through the global factor.

Due to the characteristics of the food commodities, natural disasters like floods and droughts might only affect some small group of products. Droughts in grain-producing regions over the last years have helped lower the world grain supply, which was thought to have significant impact on the grain price levels (Trostle, 2008). Australia has been suffering from a severe drought since 2004, which considerably reduced its production of agricultural products. Figure 3 presents the growth rate in total supply in metric tons for Australia during the drought period from 2004 to 2007. While total meat, vegetables, and corn supply showed big fluctuations, fish supply was relatively more stable compared with the other food commodities. One gets a similar graph for China, which has been experiencing the worst drought of their recent history. These observations suggest that drought may not have a significant effect on countries’ seafood supply but severely influence the grains and livestock production. Therefore we may want to avoid grouping seafood with other grain and meat products.

Simple examination of the nature of commodities also advises against taking an ex-ante stance on commodity clusters. In particular, consider corn and rice. These two commodities experienced high prices during the price boom and their price rise was argued to be related to the same factor.

⁸Pindyck and Rotemberg (1994)

Krugman (2008) argued that biofuel production caused farmers to expand the portion of their land used to cultivate more corn as it has become more profitable for farmers to invest the corn proceeds in ethanol production. This reduced the hectares of land used for other grain plantings (e.g., soybeans). Since climatological and land conditions are different for corn and rice, farmers are unlikely to substitute land between them. While biofuel production might have had a direct effect on some grain products, we may not argue the same thing for rice crops.

From a more analytic view, FOS (2011) document the consequences of possible grouping (clustering) misspecifications in the Monte Carlo analysis they conduct. The idea of their simulation is to emphasize what may happen if the researcher unknowingly puts a time series in the wrong group and uses the traditional exogenously defined block factor approach in estimation.⁹ They show that even small degrees of misspecification can cause a huge reduction in the model's overall fit. Specifically the entropy measure and mean square errors for the estimated factors increase with misspecifications, causing inconsistent model estimates.

Given these arguments I choose not to apply the traditional ways of defining groups of commodities; instead, I employ the FOS (2011) endogenously clustered dynamic factor model that places no initial restrictions on the groupings to study the synchronous movements of commodity prices. With the aid of this unrestricted model, the commodity groups can be formed based on any one or combination of any possible factors, such as those discussed above, without the fear of model misspecifications.

3 Literature Review

After a stable phase of commodity price inflation for over two decades, the late 2000's have seen price increases reaching record high levels and causing the world to experience one of the longest and broadest post World War II price booms. The previous price boom happened in the early 1970's and was followed by a period of low levels in the 1980's. While commodity price levels maintained stability in the 1990's, nominal prices for grains (such as corn, soybean, palm oil, wheat, and rice), energy, and metals more than doubled during the 2006-2008 boom (see figure 2). The consequences for some developing nations were more severe than others. Riots and violent demonstrations over

⁹The time series in the simulation can represent many economic variables, such as a country's GDP or a commodity's price level, a city-specific housing price, or industrial production.

the soaring costs of basic food have been reported in many countries including Bangladesh, Haiti, Yemen, Egypt, Morocco, and Mexico. Due to the severe social aspect of high food and fuel prices, organizations around the world held meetings, and discussed possible coordinated policy actions and interventions in order to aid the societies that could not maintain sufficient dietary requirements. (Examples include the recent G-20 meeting, UNICEF Food Prices Increases/Nutrition Security: Action for Children and Food and Agriculture Organization's High-Level Conference on World Food Security.¹⁰ Right after the crises, International Fund for Agricultural Development made available up to US\$200 million to provide support for farmers). By the end of 2008, energy and food prices significantly declined in the wake of the financial crises and the global economic downturn. However, another surge in prices started in May 2009 and the rise continues as of the 3rd quarter of 2011.

All of the aforementioned changes in commodity prices raised interest in the determinants of such changes and a great deal of research has been devoted to understanding the comovements across commodities. Along the line of these studies, Calvo (2008) suggests excess liquidity and low interest rates as the cause of the recent price boom. Wolf (2008) blames it on increased world demand. Krugman (2008) argues that the increase in oil prices caused governments to support biofuel production, which provides incentives for farmers to supply corn to be used in bio-ethanol production.¹¹ Farmers also switched land between corn and other grains, which reduced the overall supply for grains, which led to the increase in food prices.

In an attempt to summarize the studies about commodity price dynamics, Frankel and Rose (2009) list three competing theories explaining the recent boom. The first one is "global demand growth", which accelerated with the inclusion of high-demand countries such as China and India, causing the observed high prices.¹² Yet, this line of argument is criticized by researchers looking at the early effects of the sub-prime mortgage crisis that hit the U.S. in 2007. Economic growth downgraded for many countries lowering the production (hence the demand) for commodities glob-

¹⁰G-20 meeting: Communiqué - Meeting of Finance Ministers and Central Bank Governors, Washington DC, 14-15 April 2011.

Unicef link can be found at http://www.unicef.org/eapro/Food_Prices_Technical_Note_-july_4th.pdf.

FAO Conference tried to make a strategy to deal with hunger and unrest resulted from soaring food and oil prices. Delegates of the conference also focused on increased biofuel production and how it relates to food production and prices. The conference however hit a snag over the debates about embargos and export restrictions.

¹¹Corn based ethanol is used to produce biofuels therefore ethanol production is usually used to proxy the biofuel production in empirical analysis.

¹²Wolf (2008), Svensson (2008).

ally during the time of crises, while commodity prices were still on the rise in the first 3 quarters of the recession, contradicting the obvious link between the two.

The second theory focuses on financial markets and argues that “speculation” was the main cause of the commodity boom.¹³ Given there are futures markets for commodities, when market participants expect high prices they may hold long positions. If there is no particular reason to expect higher prices but the financial agents continue to do so, the resulting buying behavior can inflate a speculative bubble that creates high stock prices in commodity markets. Opponents of this explanation of speculative buying of commodity futures draw attention to the low inventory levels of commodities. As stated by Krugman (2008): if there were a bubble then we should have seen high inventories, which were not evident. However, Frankel (2008) continues to acknowledge speculative explanations by claiming that inventories were not measured correctly. For example, the standard data exclude the amount of crude oil that still lies underground which is much larger than what has been measured as inventories.

The third and maybe the most popular explanation is “easy monetary policy.” Low interest rates reduce the cost of holding inventories, since it is no longer profitable for the producers to invest the proceeds. Hence, by keeping interest rates at low levels, the Fed indirectly and presumably unwillingly causes decreased production and high prices. Furthermore, low interest rates create excess liquidity that can find its way into commodity markets as more and more people switch from Treasury funds to commodity contracts, thereby causing prices to rise.¹⁴ Critiques of the interest rate channel use the same argument that was used against “speculation”: Where are the inventories?

Empirical investigations try to make theoretical links between commodity prices and their determinants. For instance, Baffes and Haniotis (2010) analyzed the effects of excess liquidity, speculation, food demand growth from emerging countries, and biofuel production on food prices. They found strong links between energy and non-energy commodities and less evidence for the effect of biofuel production on food prices. Instead they argue that it is the “new money,” the excess liquidity, which has found its way into the commodity markets and caused a speculative bubble – and hence the boom. Byrne et al. (2008) identify a common factor behind commodity

¹³Citations include Hamilton (2009), Wolf (2008), Baffes and Haniotis (2010), Frankel (2008).

¹⁴Frankel (2008), Wolf (2008), Akram (2009)

price comovements by applying non-stationary panel analysis. Then they relate this common factor to potential macroeconomic variables using a FAVAR approach and find evidence of interest rate influence on commodity prices. Lombardi et al. (2010) run separate VARs for each of 15 non-energy products to look for effects of global industrial production, the U.S. effective exchange rate, the U.S. interest rate and the price of crude oil. They support the link between exchange rate and commodity prices and reject the effects of interest rates and oil prices.

Using dynamic factor analysis, Vansteenkiste (2009) investigates the relative importance of common factors for the non-fuel commodity price dynamics of 32 commodities for the period 1957-2008. She finds evidence of a common factor that becomes increasingly important throughout the sample period. As a robustness check, she also includes group-specific factors in estimation and looks for the effects of the global factor for 11 commodities.¹⁵ She suggests that the global factor is more important than the group specific factors. However, her variance decomposition suggests this is true for only 3 (wheat, maize and cotton) out of the 11 products; for the rest the group specific factors seem more important. She later used IV regressions to test the potential effects of crude oil, fertilizer prices, dollar effective exchange rate, interest rate and global demand (proxied by the industrial production of OECD and 6 major non-OECD countries) on the extracted global factor component. She finds evidence that oil, exchange rate and interest rates are important. No attempt was made to characterize the group specific factors.

4 Empirical Model and Estimation Methodology

I employ a dynamic factor model where each commodity price inflation is affected by a global factor common to all and a block factor common to particular group to inflation series. Inflation rates across blocks can only be correlated through the global factor. There are in total J block-specific factors (clusters) and one global factor. Assumptions of the dynamic factor model will be that the factors are unobservable and orthogonal to each other, and all cross-correlation of the series comes only through the factors, i.e., the variance-covariance matrix of the factors is diagonal. There are thus, K ($K = 1 + J$) dynamic factors to determine the comovements of inflation rates. As for the

¹⁵The reason she reduced her sample for the exogenously defined factor analysis is that she only grouped the commodities she knows are in one way or another related. She avoids misspecification by this means. In this sense my analysis is the first one to introduce group-specific factors for the whole set of commodity price data without the fear of misspecification.

sector-specific factors, I follow the FOS (2011)'s endogenous clustering algorithm which gives the data the freedom to choose its own grouping.

Let I denote the number of goods and T denote the length of the time series. Then for an observable inflation rate denoted by $y_{i,t}$ for commodity i , we have

$$y_{i,t} = \alpha_i + \beta_{i,0}F_{0,t} + \sum_{j=1}^J \gamma_{i,j}\beta_{i,j}F_{j,t} + \varepsilon_{i,t}, \quad (1)$$

where α_i is a vector of intercepts; $\beta_{i,0}$ and $\beta_{i,j}$ are diagonal matrices of factor loadings; $F_{0,t}$ is the global factor affecting each series; $F_{j,t}$ are the group factors; and $\varepsilon_{i,t}$ is a series-specific idiosyncratic error term. As noted in FOS, $\gamma_{i,j} = \{0,1\}$ is a grouping indicator that defines whether series i belongs to group j . Further, each series is restricted to one single group obtained by the restriction $\sum_j \gamma_{i,j} = 1$. Factor loadings are specific to each series, which allows for different responses of inflation in response to the same shock.

The evolution of each factor and the idiosyncratic error term are determined by an autoregressive equation of order q^f and q^ε , respectively;

$$F_{k,t} = \phi_{k,1}F_{k,t-1} + \phi_{k,2}F_{k,t-2} + \dots + \phi_{k,p}F_{k,t-q^f} + e_{k,t} \quad \text{for } k \in \{1, \dots, K\}, \quad (2)$$

$$\varepsilon_{i,t} = \varphi_{i,1}\varepsilon_{i,t-1} + \varphi_{i,2}\varepsilon_{i,t-2} + \dots + \varphi_{i,q}\varepsilon_{i,t-q^\varepsilon} + \epsilon_{i,t} \quad \text{for } i \in \{1, \dots, I\}, \quad (3)$$

where $e_{k,t}$ is a factor-specific idiosyncratic error term with variance λ_k^2 , and $\epsilon_{i,t}$ is the idiosyncratic disturbance with variance σ_i^2 . The disturbance terms, e and ϵ , are uncorrelated and each distributed normally with zero mean and their respective variances. As the factors are unobservable the sign of the factors and the sign of the factors' loadings have to be separately identified. Following FOS, I normalize the first element of each factor to be positive to overcome the issue. Another identification problem is to identify the scale of the factors. Here I follow KOW (2003), Sargent and Sims (1997), and others in assuming that λ_k^2 is constant.

I estimate the model presented in equation (1) with Bayesian Markov Chain Monte Carlo (MCMC) techniques. To sample the factors, I follow Kim and Nelson (1999) and apply Kalman filters. Once the factors are known (or given) I follow Chib and Greenberg (1995) to sample the model parameters. Sampling iteratively from the conditional distribution of the model's parameters given

the factors and from the conditional distribution of the factors given the parameters is repeated many times. This is the essence of Gibbs sampling and under the regulatory assumptions (see Chip and Greenberg, 1995) these sequences of draws from the conditional distributions converge to the joint posterior density of the entire system, independent of initial values selected. The technical details of the estimation are provided in the appendix.

The endogenous clustering model represented here can be estimated using either Bayesian or classical techniques. The classical algorithm can solve the problem by forming every possible group and then employing a model selection criterion to determine the best clusters. However, with a large panel of data this grid search-like procedure would be inefficient and possibly infeasible.

The Bayesian approach offers several other advantages in estimation. First of all, Bayesian inference provides computational easiness for latent variable models like the one presented in this paper. As noted in Paap (2002) the likelihood function of classical estimations of these models includes many integrals and numerical integrations which make standard estimation models like maximum likelihood infeasible whereas the Bayesian MCMC approach only considers the likelihood function conditional on the simulated unobserved variables; therefore, it does not require computing the unconditional likelihood function of the model itself. This makes the estimation easier and much faster than most of the standard classical techniques. Paap also argues that Bayesian modeling allows a more convenient way of dealing with parameter uncertainty, which needs to be taken into account when dealing with unobserved variables.

Another important advantage of the Bayesian method concerns its small sample properties. It has been argued that MCMC computation works equally well for large and small samples.¹⁶ Recently, with the wide use of the disaggregated data, researchers have utilized dynamic panel data econometrics. However, these models have been documented to perform poorly in estimation and inference without correcting for the small sample biases if the sample size is small.¹⁷ The use of Bayesian methods offers an advantage in the sense that it does not require a correction when

¹⁶Western (1998), Martin (2005), Berger (1985)

¹⁷IMRR(2003), Chen and Engel (2005), Phillips and Sul (2007), and many others use several methods to account for small sample biases. The most commonly known small sample bias correction is Killian's bootstrap after bootstrap method. However this method has been proved to perform poorly with highly persistent series. The Andrews (1993) and Andrews and Chen (1994) median unbiased estimator is another way to correct for the bias. However this method does not work well if true AR(1) is near unity. Another method is by Pesaran Zhao (1999) who extends Killian's method for long-run coefficients that are nonlinearly dependent on the short-run ones.

The Bayesian methodology performs equally well for large and small samples and provides an estimation tool that does not need any correction for small samples.

dealing with small samples. As Berger (1985, page 125) says "Bayesian procedures are almost always equivalent to the classical large sample procedures when the sample size is very large and are likely to be more reasonable for moderate and (especially) small sample sizes (where many classical large sample procedures break down). Indeed, unless there have been extensive studies establishing the small and moderate sample size validity of a particular classical large sample procedure, almost any Bayesian analysis would seem preferable."

4.1 Model Selection

While cluster memberships are endogenously determined, the number of clusters still have to be exogenously selected. However we can also endogenize the number of clusters by computing the Bayes factors. This paper applies Chib (1995) for every Gibbs sample block and follows Chib and Jeliazkov (1995) where Metropolis Hastings is implemented. The basic marginal likelihood identity (BMI) of the model is given as:

$$\ln \hat{m}(\mathbf{Y}) = \ln f(\mathbf{Y}|\Theta^*) + \ln p(\Theta^*) - \ln \hat{p}(\Theta^*|\mathbf{Y}), \quad (4)$$

where Θ is the full set of model parameters. The expression requires the evaluation of the log likelihood function $\ln \hat{m}(\mathbf{Y})$, the prior $\ln p(\Theta^*)$, and an estimate of the posterior ordinate $\ln \hat{p}(\Theta^*|\mathbf{Y})$ evaluated at a high density point Θ^* (e.g., a modal point) of the posterior draws of the parameters.¹⁸ The log likelihood and the priors at the modal values of the parameters can be directly computed from the whole set of posterior draws gathered from the Gibbs estimation of the model (the Gibbs output). However, the posterior ordinate needs to be estimated with additional Gibbs sampling steps of the same model but with reduced samplers. Details for sampling each posterior ordinate as well as the model likelihood and priors are all supplied in the appendix.

In this framework, the models are distinguished by the number of clusters. In particular, the empirical model and the BMI is estimated assuming a different number of clusters one at a time. Then the marginal likelihoods are used to decide which model to select. The model that maximizes the marginal likelihood reveals the optimum number of clusters. In principle one needs to calculate

¹⁸Chib points out that the BMI holds for all values of Θ and the choice of the Θ^* is not critical but for the efficiency considerations Θ^* is selected to be a high density point so that the density can be accurately estimated.

the Bayes factors of two models l and h , using the BMI as

$$B_{l,h} = \exp(\ln \hat{m}(\mathbf{Y})_l - \ln \hat{m}(\mathbf{Y})_h).$$

If $B_{l,h} > 1$, then model "l" is favorable to model "h". Comparing the bivariate Bayes factors for all models and finding the superior model is equivalent to maximizing the BMI given all models. Hence in the results section, I only list the BMI with its components.

4.2 Bayesian Linear Estimation on Factors

Once the factors are carefully extracted, in the next step I try to characterize them with additional analysis. Let Ξ_t represent the set of the variables we want to test on the factors (global and clusters), $F_{k,t}$. Then we can estimate the linear regression of the form

$$F_{k,t} = \varpi_k \Xi_t + v_k \tag{5}$$

where the error term v is assumed to be normally distributed with mean zero and variance \oplus_k^2 . The estimation procedure is a simple Gibbs application with two sampler blocks of parameters, namely ϖ and \oplus^2 . The issue with such an estimation is that it makes use of the estimated factors as the regressand. The problems of using generated regressors are documented in Pagan (1984) and ideally one has to correct for the uncertainty coming from the generated regressor term for the inference to be correct as the posterior distribution of the model (5) depends on which factor $F_{k,t}$ is selected. .

If the factor is an observed variable, then the Gibbs application on (5) would converge to the unconditional posterior distribution of the parameters, i.e., $p(\varpi|\Xi)$. However if the factor is not observed and rather has a distribution, the Gibbs sampling yields a posterior distribution of the parameters conditional on the selected factor, say $p(\varpi|\Xi, \bar{F})$. Therefore to make inferences from $p(\varpi|\Xi)$, we can integrate $p(\varpi|\Xi, \bar{F})$ over the distribution of F :

$$p(\varpi|\Xi) = \int_F p(\varpi|\Xi, F)p(F|Y)dF, \tag{6}$$

where $p(F|Y)$ is the posterior distribution of the factor from the first step (factor analysis). Ana-

lytically we cannot solve this integral; instead we can approximate it by drawing large numbers of F from its posterior distribution and calculating the $p(\varpi|\Xi, F)$ by repeating the Gibbs sampling for each of these factor draws. This will result in an approximation of the unconditional posterior distribution of the parameters that we can make inferences from. Details are in the appendix.

5 Data

Monthly time series data for 42 non-energy commodity prices spanning from 1980 to 2011 are gathered from the International Monetary Fund (IMF)'s International Financial Statistics (IFS) Database. The commodities are selected on the basis of availability for the entire sample period. Fertilizer and energy prices are excluded so that they could be used in auxiliary regressions to see if energy prices are the main fundamental driving force behind the commodity price movements as argued in several studies. The details of the data can be found in the appendix.

Data are first seasonally adjusted (Census X12 multiplicative adjustment) and then converted to quarterly frequency mainly to increase the signal-to-noise ratio and to save some computational time. Since the empirical model requires stationary series, I log-difference the data, thereby computing the inflation rates for each commodity. Finally, I follow the factor model literature and normalize these inflation rates by demeaning each commodity price and dividing it by each series' standard deviation.¹⁹

Figure 3 plots the pairwise cross correlations of the commodity sample. Since simple cross correlations are static and cannot represent joint moves of many commodities these results should only serve as a preliminary check of possible linkages. The nominal prices of the product sample exhibits high positive cross correlations. Measures used in the literature such as concordance defines comovements as the same direction synchronized movements. However, the pairwise correlations suggest there are some products that exhibit negative relationships. Just because commodities move in the opposite direction does not necessarily mean that they cannot share the same source. The factor analysis presented in this paper do not exclude these kind of inverse movements and will recover commonalities, positive or negative.

The data for the auxiliary regressions come from several sources. The interest rate is proxied by

¹⁹There were some questions raised about standardization. I compared the results for both standardized and non-standardized data and found variance decomposition for both cases to be similar.

the federal funds rate extracted from the Board of Governors of the Federal Reserve System. The exchange rate, Dow Jones stock market index, and U.S. house prices come from the St. Louis Fed's Federal Reserve Economic Data (FRED). The IFS database also provides crude oil and fertilizer prices (measured by phosphate rock). The federal funds rate and the U.S. effective exchange rate are deflated using U.S. consumer price index. The world demand is proxied by the industrial production of 30 countries from the IFS database where the countries are selected based on data availability. Ethanol production that accounts for biofuels is gathered from the Renewable Fuels Association. To measure climate changes, I use the global surface temperature anomalies from National Climatic Data Center (NCDC) of National Oceanic and Atmospheric Administration (NOAA) database.²⁰

6 Empirical Results

6.1 The Optimum Number of Clusters via Bayesian Model Selection

The optimal model is the one that maximizes the marginal likelihood, which results in 4 clusters. Table 1 presents the details of the Bayesian Model selection outcome. Intuitively the posterior ordinates can be thought as a penalty of having additional clusters; and as one can see, the posterior ordinates are decreasing as the number of factor increases, thereby validating its purpose. The results also look consistent in the sense that the optimal model maximizes the marginal likelihood as well as the likelihood.

6.2 Results for the Optimal Model

6.2.1 Inclusion Probabilities

Table 2 lists the commodities with their probability range across clusters. For each commodity the algorithm produces a posterior distribution of its indicator function. This means that each commodity has a probability, whether strong or weak, of belonging to each cluster. In table 2, I report the highest probability of belonging to one cluster for each commodity. Logs and wood inflation rates are strongly correlated through the first cluster. Lamb also belongs to this cluster

²⁰As stated in the database, the anomalies are observed temperature departures from the 20th century (1901-2000) average of global temperature. An increase in these departures is evidence of global warming.

with a weaker probability of 0.54.²¹ However, since it constitutes only 20 percent of the cluster size and since it selects the cluster only half of the time, the corresponding factor should be dominated by the timber industry, reflecting the industry-specific properties of logs and wood.²²

Consistent with observations in the commodity price boom, the second cluster consists of vegetable oils and grains. These products were responsible for the food price index spike more than any other commodities (Mueller et al., 2010), and it would contradict many related studies if they were not grouped together.

The third cluster consists of food, metal, and agricultural materials. The clustering analysis shows evidence that metals such as aluminium, copper, uranium, and zinc are strongly correlated with food products such as olive oil, fish, fishmeal, and sugar and weakly correlated with food products such as beef, lamb, oranges, bananas, and rice. Comovements of seemingly unrelated commodities are nothing new to the literature; for example, copper and wheat are found to be correlated by Pindyck and Rotemberg (1990). All metal prices in this cluster are highly correlated with other metals except iron. Cuddington and Jerrett (2008) offer empirical support for super cycles (long cycles for more than 15 years) of metal prices and posit the recent Chinese industrialization and urbanization as a likely cause. Therefore it would be interesting to check for this claim and test for Chinese demand on the cluster-3 factor.

Coffee products and iron belong to the last cluster. In particular, iron is the only metal that does not share strong linkages with other metals. It is out of the scope of this paper to understand why iron shares cycles with coffee rather than with copper; however, this finding may open up an interesting avenue of research about these commodities. Moreover, the variance share of iron attributable to this cluster factor is only 4 percent (see appendix for the complete variance decompositions listed under table 5); therefore, it would not be incorrect to claim that cluster-4 is defined by mostly the coffee industry and can be labeled as "Coffee Cluster".

Another interesting conclusion is the case of rice. Even though rice is listed under "grain products" in commodity price indices (along with barley, wheat, soybeans and corn) it does not share cycles with other grains – not even through the global factor.²³ What is more rice has the

²¹The case of lamb is rather interesting since it also belongs to the cluster-3 almost equally likely (with a probability of 0.46).

²²Lamb meat is 1 out of 5 commodities of cluster-1. Therefore it occupies 20 percent of the cluster size.

²³The explained variance attributable to the global factor for rice is 0. See the next section or appendix for the variance-decompositions tables.

weakest probability of belonging to any cluster in the all the commodity sample used in this paper (it only achieves a maximum of 0.34 for cluster-3). What makes rice decouple and stand alone? This might stem from the fact that rice goes through different agricultural processes with specific needs for rainfall. Wheat needs a dry, mild climate to grow. Soybean fields should be well drained for its cultivation. And corn is a warm weather drought-resistant crop that requires relatively less moisture when developing toward maturity. Whereas, rice needs extreme humidity, and prolonged sunshine, it requires standing water throughout its growing period and is best suited for regions with high amounts of water supply. Other than these agricultural differences, country-specific effects (which are not accounted for in this analysis) may also be responsible for this "rice decoupling" since, apart from palm oil, cluster-2 products come from North America and United Kingdom while rice prices are taken from Thailand.

Overall, looking at the strongly correlated ($p(\gamma_{i,j}) > 0.9$) commodity cluster formations we can define 4 distinct categories (and I will refer to them as such hereafter): "Timber", "Coffee", "Grains & Oils"(except olive oil) and a "Mixture" of agricultural raw products (e.g., wool, hides, rubber) metals (e.g., copper, lead) and food commodities (e.g., sugar, olive oil, salmon). This cluster formation provides evidence against distinct industrial categorization of commodities. In other words, these clusters are not representative of standard narrow classifications (indexes) of commodity prices as defined by statistical agencies. In particular, food products are spread across all clusters (weather weakly or strongly): While some of them share cycles with metals and agricultural raw materials, some of them decouple from the rest of the food sector (coffee and rice).

In related work, Vansteenkiste (2011) sets up an exogenously defined clustered factor model. She pools jointly produced or consumed commodities together into groups and defines 4 clusters: (1) Coffee and cocoa, (2) cotton-maize-sugar-wheat, (3) palm oil and soybean oil, and (4) copper-zinc-lead. This paper provides evidence against this kind of cluster formation. In particular, once the data are free to form their own clusters, cocoa and coffee fall into different groupings and sugar does not find its way into the same cluster as grain commodities.

6.2.2 Variance Decompositions

This section reports the variance decompositions where the clusters are constructed with the posterior values of the indicator function, $\gamma_{i,j}$. Since an observation may change clusters over the

Gibbs iterations, we need a fixed estimate for the indicator function for the variance calculations. Table 3 reports the "weak probability variance decompositions," where an observation is assumed to belong to a particular cluster if it picks that cluster for the majority of the Gibbs run, i.e., $i \in j$ if $p(\gamma_{i,j} = 1) > p(\gamma_{i,k} = 1) \forall k$. In this case all the commodities are matched to one cluster and we get the cluster memberships exactly as listed in table 2.²⁴

Looking at the average variance decomposition percentages of table 3, the global factor is not playing a significant role in explaining the bulk of the commodity sample developments except for cluster-2 ("Grains & Oils). The cluster factors on average explain about 27 percent of commodity price variations, which dominates the effect of the global factor, which is only 7.2 percent. In particular, for cluster-1 ("Timber") and cluster-4 ("Coffee") the global factor is negligible; the market and production processes of these products may be too specific and closed to global developments. Overall Table 3 suggests that the dominant source behind commodity price comovements is the interrelations that come through more sparse cluster factors. This finding contradicts many studies that assume only one or two common factors behind commodity price dynamics.²⁵

The simultaneous move in prices of grains (corn, soybeans, wheat, rice) oils and meat have led many studies to agree on the existence of a single shared source for the food sector (Byrne et al. 2010; Vansteenkiste, 2009). To have a better insight into this claim, Table 4 highlights the variance decompositions of these products. Surprisingly, the source of the fluctuations seems to be different even for similar grain products that are in the very same cluster-2 ("Grains & Oil") such as corn and wheat. Corn, soybeans, and soybean meal are highly dominated by the cluster factors, while vegetable oils are mainly driven by the global factor. However for wheat, rice, and meat, it is idiosyncratic shocks that matter the most.²⁶

In summary, the world factor does not seem to have a strong effect on corn, rice, wheat, meat, soybeans, and soybean meals prices, which invalidates explanations such as those that assert food commodities move together mainly due to, for example, increased world demand. Rice in particular is dominated by market-specific factors rather than aggregate factors as discussed in the

²⁴One of the drawbacks of such a calculation is that some of the interrelations among commodities are rather weak. Weakly correlated products share smaller portion of their cycles through factors which reduce the variance decomposition values for each cluster. Once the commodities that have probability of belonging to a cluster below 0.9 are excluded from the data sample, the effects of global and cluster factors get increasingly large, explaining 46 percent of the whole sample variations compared with 34 percent of "weak probability variance decomposition".

²⁵Byrne et al. (2010), Vansteenkiste (2010), Cashin et al. (2010), Lombardi et al. (2010).

²⁶Rice and meat belong to cluster-3 decoupling from the rest of the food commodities investigated in table 5.

previous section. The global factor does not have a significant effect on rice price fluctuations, which contradicts Vansteenkiste's (2009) findings where she finds 12 percent of rice price variations resulting from the global factor.²⁷

Overall, the premise is that there is no single common factor driving the major commodity prices over time; instead commodities are interrelated through a set of cluster factors which contribute to the recent price peak more than the common factor. "Timber" and "Coffee" decouple from the rest of the sample, exhibiting different and probably more product-specific sources. However, more in-depth analysis is needed on the global and cluster factors to validate such claims.

6.2.3 Characterizing Factors

Figures 4 to 8 plot the factors along with NBER recessions dates.²⁸ The downturns in the commodity factors coincides with the U.S. recessions. In particular the global, second, and third cluster factors show a great slump during the Great Recession of the late 2000's, which suggests all of them were affecting the price fluctuations of food, metals, and materials during the commodity price burst. The big fall that corresponds to year 1994 in cluster-4 ("Coffee", figure 8) is consistent with the Brazilian coffee plantation expansion and Vietnam's entry into the market, which put pressures on the supply and lowered coffee prices.²⁹

The variance decompositions of the previous section provided some evidence of multiple important factors behind the commodity price comovements. This section focuses on identifying these factors to see if any macroeconomic variables are correlated with the estimated factors.

In order to highlight the sources of these factors, I run Bayesian auxiliary regressions of the estimated factors on possible determinants that have been mentioned, argued, or strongly supported in the literature. In a related paper, Bryne et al. (2011) use a two-step FAVAR approach as

²⁷The difference is likely to originate from empirical model specifications. She uses a dynamic factor approach with one global factor and her global factor is suffering from overestimation due to the additional correlation of commodities that are not accounted for in her analysis.

²⁸The factor loadings are almost all positive for commodity prices except a few commodities. Namely, swine which belongs to cluster-4 is negatively related to the cluster-4 factor and hard logs, shrimp, soft Logs, soft sawnwood, soybean meal and again swine are negatively related to the global factor. Looking at the individual variance decompositions (see appendix) the average explained variation of these commodities attributable to the global factor is only 1.85. Which implies that the average inflation for those commodities due to the global factor during recession is negligible. Besides, once combined with the cluster factor effect (which explains 35 percent of their variations) the overall impact will be positive.

²⁹The "Timber" cluster factor (figure 4) seem to be recovered from big fluctuations of the 80's and early 90's and is relatively stable during the rest of the 2000's.

described in Bernanke et al. (2005) and relate the common factor to the real U.S. short-run interest rate, global demand as proxied by U.S. real GDP growth, real crude oil prices, and risk measured by standard deviations of closing value of Dow Jones average. Vansteenkiste (2009) also employs a dynamic factor approach and test the global factor on possible determinants. In particular I use variables similar to those of Vansteenkiste (2009) – namely; the federal funds rate, U.S. dollar effective exchange rate, fertilizer prices, industrial production, and stock market index. In addition to these variables, I also test for U.S. housing prices, biofuel production, Chinese demand, and climate changes. Detailed description of the variables used in this paper are listed below.

1. Deflated Effective Federal Funds Rate (FFR), Quarterly: The nominal rate is deflated by the Consumer price index. Given the arguments in the literature, it is expected to have a negative impact on the commodity prices.
2. U.S. dollar Real Effective Exchange Rate (EER), Quarterly: Devaluated dollar (represented as a fall in the EER) causes the commodities to get cheaper in terms of foreign currencies, which in turn puts on a positive pressure on the prices.
3. Dow Jones Stock Market Index, Quarterly: It is used to measure the speculation bubble effects on the commodity price. It should result in higher commodity prices implying an expected positive sign in the regressions.
4. World Industrial Production, Quarterly: I use the quarterly industrial production of 30 countries, which includes developed economies as well as emerging and underdeveloped countries to proxy for world demand.³⁰
5. Crude oil Prices, Dubai, Quarterly: Oil price increases cause a cost increase and higher commodity prices.
6. Fertilizer Prices, Quarterly: This should result in higher prices for many food and some agricultural materials (such as wool) due to cost increases.
7. Housing Prices, Quarterly: Recent subprime mortgage crises have spread around the globe and initiated the latest Great Recession. The burst in housing prices may not directly have

³⁰The countries include Australia, Austria, Barbados, Belgium, Canada, Denmark, Finland, France, Germany, Hungary, India, Ireland, Israel, Italy, Japan, Jordan, Republic of Korea, Luxembourg, Malaysia, Mexico, Netherlands, Norway, Portugal, Senegal, Spain, Sweden, Switzerland, Turkey, United Kingdom, and United States.

caused the commodity price boom; however, it may have reduced the demand for its basic inputs: logs, metals, and materials. Hence, we can expect to see a positive relationship between house prices and their input prices.

8. Ethanol Production, Annual: To account for the increase in biofuel production, I use its main ingredient – corn-based ethanol production. High ethanol production growth could cause high food prices, especially for grains and oils, due to reasons described previously in this paper.
9. China Volume of Exports and Imports (China Trade), Quarterly: Emerging countries, especially China, took on a larger role in world trade while increasing the demand for commodities as well as their prices. The widely used measures for a country’s demand are its industrial production or its gross domestic product. However, both of these variables for China are not available in quarterly frequency in 1980-2011 time period. Therefore, as an alternative, I use the total volume of exports and imports to account for quarterly Chinese demand.
10. Chinese Gross Domestic Product (China GDP), Annual: As discussed above, this variable is included to account for the increased Chinese demand. The cross correlation between volume of trade and GDP is 0.95; therefore, I do not use these two variables in the same regression to avoid multicollinearity issues.
11. Weather Anomalies, Quarterly and Annual: Climate changes could cause adverse weather conditions that could affect crop growth and reduce agricultural supplies. These anomalies are provided as departures from the 20th century average (1901-2000) and can be used as measures of adverse weather. An increase in temperature anomaly is an harbinger of global warming, which indicates the possibility of adverse weather reactions. To construct quarterly data, I aggregate monthly values for these global temperature anomalies across seasons.

The estimated factors are measured at a quarterly frequency. As a result, I conduct two separate analyses. First I look for the contemporaneous relationship of the quarterly factors with the interest rate, exchange rate, Dow Jones stock market index, world industrial production, U.S. housing prices, fertilizer prices, and climate anomalies. Additionally, to test for the biofuel effect (which is not available in quarterly frequency), I estimate annual regressions using annualized factors along with annual variates of everything listed above while substituting China Trade with China GDP.

The caveat of aggregating the factors is that the regression results can potentially lose short-run information and can suffer from aggregation bias.

Quarterly Regression Results Table 6 shows the results where each column represents a separate regression of the determinants listed in the rows.

The global factor looks like it is capturing the world industrial production. The intuition is straightforward. When industrial production increases, demand for metals and materials accelerates. Higher income due to higher production tends to increase the demand for food, thereby spreading around the effects of the high global demand to almost all sectors.

Chinese volume of trade seems to have a significant effect on the factor that drives the correlation structure of the cluster-1 ("Timber"), which includes wood and logs from two countries: U.S. and Malaysia. So how can we link Chinese trade with these commodities? China is one of the top five importers of Malaysian timber. Chinese buyers also turned to the U.S. and Canada for wood after 2007 when Russia imposed higher tariffs on its logs. Chinese lumber imports from North America more than quadrupled from 4 percent in 2005 to 18 percent in 2010.³¹ This revived the U.S. timber industry back from a depressed state since the subprime mortgage crisis. The Wallstreet Journal reports 10 to 15 percent expected increase in log harvests from big U.S. timber companies due to the recent export surge from China. These may be the reasons why we see a significant Chinese demand on cluster-1 ("Timber") factor.

Crude oil prices are found to be significant for cluster-2 ("Grains & Oils"). The farming sector is highly energy intensive; therefore, oil prices affect its cost structure. For example, Murray (2005) draws attention to the high use of fossil fuels in the U.S. farming industry with this comparison: "The U.S. food system uses over 10 quadrillion Btu (10,551 quadrillion Joules) of energy each year, as much as France's total annual energy consumption." Growing food without packaging, storage, or transportation accounts for 20 percent of this total amount. Besides, food travels from farms to distributors around the world, which amplifies the industry's dependence on fuel use. The conclusion is simple: fossil fuel reliance can alter grain commodity prices and hence can be reflected as this cluster's factor.

Most of the variables discussed in the literature as potential determinants of the commodity

³¹Source: International Centre for Trade and Sustainable Development Bridges Trade BioRes, Volume 11, Number 18, 17th October 2011, pp 14.

prices; namely, federal funds rate, speculation, world demand, and crude oil prices are found to be a part of "Mixed" cluster commodity commovements. Mining and manufacturing metals are energy intensive which help to link oil prices to this cluster-3 factor. From a monetary policy perspective, the federal funds rate can affect the dynamics of a large group of commodities of metals, materials, and some foodstuff. Intuitively this means that if the Fed keeps its quantitative easing policies in effect at times of high oil prices and high world demand, it would amplify the increases in commodity price inflation.

Lastly, cluster-4 ("Coffee") factor fails to highlight a significant presence of any of the variables tested in this analysis. This could be due to missing important macro variables or simply because this cluster represents shocks specific to the coffee industry that cannot be accounted for easily. For example, coffee markets are highly controlled by the International Coffee Organization (ICO), which was formed in 1963 in an attempt to stabilize prices through international cooperation. With members that account for 97 percent of world coffee exports and 80 percent of world coffee imports, ICO claims to achieve a balanced and sustainable world coffee economy and promotes coffee consumption. For example, it launched CoffeeClub Network in 2008 and implemented the Coffee Quality Improvement Programme in 2002 in order to stimulate demand through better standards of quality. The effects of these acts and agreements are likely to have an impact on the coffee industry, but are hard to measure.

Annual Regression Results This section adds the remaining regressors – namely, ethanol production and China demand as measured by GDP – to the regression analysis and lists the findings in Table 7. The significant variables in each regression are consistent with the findings from the quarterly regressions. Additionally, speculation, fertilizer prices, and the U.S. dollar-effective exchange rates become significant to clusters 1, 2, and 3, respectively.

The global factor helps to feed the effect of global demand into the commodity prices as also shown in quarterly analysis. Cluster-1 ("Timber") factor now adds speculation to its possible determinants. Cluster-2 ("Grains & Oils") is affected by fertilizer prices, which is not a surprising result as fertilizers are used in cultivation to improve plant growth. In particular, corn, soybeans, and wheat are the three major crops associated with high consumption of fertilizers.

Cluster-3 "Mixed" factor is now explained by the changes of the exchange rate along with the

federal funds rate. The devaluation of the dollar may reduce the price competitiveness of non-U.S. countries and diminish production of exporting goods for these countries. On the demand side, a devalued dollar can cause prices of a product to become cheaper in foreign currencies, this may increase the demand and alter the price dynamics. The combination of both supply and demand effects can accelerate the increase in its price level.

Finally, world demand affects the last cluster ("Coffee"). This could be linked to successful attempts of the ICO's coffee demand stimulation acts described previously.

Looking for the impact of biofuels on commodity prices, I cannot provide evidence in favor of Krugman's (2008) argument that increased biofuel production is one of the main causes of the grain price surge.

7 Conclusion

The dynamics of commodity prices have been changing over the last half of the decade. No economy is immune to inflation, and if price increases are synchronized and remain persistent enough they can pass through to the core inflation rate, creating a need for action by the monetary authorities. This paper investigates the comovements of commodity prices and what drives them from a statistical point of view. Summarizing information from a large panel set of commodity prices, I find that commodity cluster compositions do not represent the standard narrow classifications (indexes) of commodity prices as defined by statistical agencies like International Financial Statistics (IFS). For example, timber products isolate itself from other agricultural raw materials, and form a separate cluster. Likewise coffee forms another. I also find another cluster of commodities consisting of seemingly unrelated products, such as metals, agricultural materials, and some food products. Additional analysis to characterize these correlations indicates the importance of the federal funds rate, high world demand, high crude oil prices, fertilizer prices, Chinese demand, and speculation in financial markets in driving these products' common movements.

References

- [1] Akram Q. F., "Commodity prices, interest rates and the dollar", *Energy Economics*, Volume 31, Issue 6, November 2009, Pages 838-851
- [2] Andrews, D.W.K., "Exactly Median-Unbiased Estimation of First Order Autoregressive/Unit Root Models," *Econometrica* 61: (1993), 139-165.
- [3] Andrews, D.W.K., and H.-Y. Chen, "Approximately Median-Unbiased Estimation of Autoregressive Models," *Journal of Business & Economic Statistics*, Vol. 12, No. 2 (Apr., 1994), pp. 187-204
- [4] Bai, J. "Inference on factor models of large dimensions." *Econometrica* 71(1), 2003, pp. 135-172.
- [5] Bai, J. and Ng, S. "Determining the number of factors in approximate factor models." *Econometrica* 70(1), 2002, pp. 191-221.
- [6] Bernanke, B.S., Boivin, J., P. Elias, "Measuring the effects of monetary policy: A factor-augmented vector autoregressive (FAVAR) approach", *Quarterly Journal of Economics* 120, (2005) , 387-422
- [7] Baffes, J., "More on the Energy/Non-Energy Commodity Price Link" *World Bank Policy Research Working Paper Series* (2009)., No. 4982
- [8] Baffes, J. and T. Haniotis, "Placing the 2006/08 Commodity Price Boom into Perspective", *World Bank Policy Research Working Paper* 5371 (2010)
- [9] Borensztein E. and C. M. Reinhart, "The Macroeconomic Determinants of Commodity Prices", *IMF Staff Papers*, Vol. 41 (1994), No. 2, 236-258.
- [10] Byrne, J. P., G. Fazio, and N. M. Fiess, "Primary Commodity Prices: Co-Movements, Common Factors and Fundamentals", *World Bank Policy Research Working Paper Series* (2011), Vol. , pp.
- [11] Calvo, G. "Exploding Commodity Prices, Lax Monetary Policy, and Sovereign Wealth Funds". *Vox EU*. (2008)
- [12] Carter, C., G. Rausser and A. Smith, "Commodity Booms and Busts", *Annual Review of Resource Economics* (2011).

- [13] Casella, George and George, Edward I. "Explaining the Gibbs Sampler." *American Statistician*, August 1992, 46(3), pp. 167-74.
- [14] Cashin, P., C. J. McDermott, and A. Scott, "The Myth of Co-Moving Commodity Prices", Bank of New Zealand Discussion Paper (1999) No. G99/9.
- [15] Cashin, P., C. J. McDermott and A. Scott, "Booms and Slumps in World Commodity Prices", *Journal of Development Economics* (2002), vol. 69, pp. 277-296
- [16] Cashin P., H. Liang, C. and J. McDermott, "How Persistent Are Shocks to World Commodity Prices?", *IMF Staff Papers*, Vol. 47, No. 2 (2000), pp. 177-217
- [17] Chen, Shu-Ling, J. D. Jackson, H. Kim, and P. Resiandini, "What Drives Commodity Prices?", (2010) No auwp2010-05, Auburn Economics Working Paper Series, Department of Economics, Auburn University.
- [18] Chen, Shiu-Sheng, and C. Engel, "Does "Aggregation Bias" Explain the PPP Puzzle?" *Pacific Economic Review* 10, (February 2005), 49-7.
- [19] Chib, Siddhartha. "Marginal Likelihood from the Gibbs Output", *Journal of the American Statistical Association*, (1995), 90(432), pp. 1313-21.
- [20] Chib, S., and E. Greenberg, "Bayes Inference in Regression Models with ARMA (p, q) Errors," *Journal of Econometrics* 64 (1994), 183-206.
- [21] Chib, S., and E. Greenberg, "Understanding the Metropolis-Hastings Algorithm," *American Statistician* 49 (1995), 327-335.
- [22] Chib, S. and I. Jeliazkov, "Marginal Likelihood from the Metropolis-Hastings Output", *Journal of the American Statistical Association* (2001), 96(453), pp. 270-81.
- [23] Cuddington, J. T. and D. Jerrett, "Super Cycles in Real Metals Prices?" , *IMF Staff Papers* (2008), Vol. 55, Issue 4, pp. 541-565.
- [24] Foerster, A. T., P. G. Sarte, and M. W. Watson, "Sectoral versus aggregate shocks: A structural factor analysis of industrial production." *Journal of Political Economy* (2011), 119,1-38

- [25] Francis, N. R., M. T. Owyang and O. Savascin, "An Endogenously Clustered Factor Approach to International Business Cycles " (2011)
- [26] Frankel, J.A. "Why are Commodity Prices so High? Don't Forget Low Interest Rates", *Financial Times*, (2005). 4/15/05
- [27] Frankel, J.A. "The Effect of Monetary Policy on Real Commodity Prices", *Asset Prices and Monetary Policy* (2008), University of Chicago Press
- [28] Frankel J. A. and A. K. Rose, "Determinants of Agricultural and Mineral Commodity Prices", Working Paper, Kennedy School of Government, Harvard University, (2009).
- [29] Frühwirth-Schnatter, S. and S. Kaufmann, "Model-Based Clustering of Multiple Times Series." *Journal of Business and Economic Statistics*, January 2008, 26(1), pp. 78-89.
- [30] Hamilton, J. D. "Causes and Consequences of the Oil Shock of 2007-2008", NBER Working Paper Series (2009), 15002
- [31] Hamilton, J.D. "Time Series Analysis", Princeton University Press (1994).
- [32] Hamilton, J. D. and M. T. Owyang, "The Propagation of Regional Recessions." *Review of Economics and Statistics*, forthcoming.
- [33] Holmes, Chris C. and Held, Leonhard. "Bayesian Auxiliary Variable Models for Binary and Multinomial Regression." *Bayesian Analysis*, 2006, 1(1), pp. 145-168.
- [34] Hochman, G., D. Rajagopal and D. Zilberman, "Are Biofuels the Culprit? OPEC, Food, and Fuel". *American Economic Review*, Vol.100, No.2, (May 2010), pp. 183-18
- [35] Imbs, J., H. Mumtaz, M. O. Ravn, and H. Rey, "PPP Strikes Back: Aggregation and the Real Exchange Rate," NBER Working Paper 9372, (2003).
- [36] Kilian, Lutz, "Small-Sample Confidence Intervals for Impulse Response Functions," *Review of Economics and Statistics* LXXX (II) (1998), 218-30.
- [37] Kim, C.J. and C.R. Nelson, "State-Space Models with Regime Switching", The MIT Press (1999), Cambridge, MA.

- [38] Kose, M. Ayhan, C. Otrok, and C. H. Whiteman, "International Business Cycles: World, Region, and Country Specific Factors," *American Economic Review*, Vol. 93, (2003), pp. 1216–39.
- [39] Kose, M. Ayhan; C. Otrok, and C. H. Whiteman, "Understanding the Evolution of World Business Cycles." *Journal of International Economics*, May 2008, 75(1), pp. 110-30.
- [40] Krugman, P. "The Oil Nonbubble". *The New York Times*. (2008) 12 May.
- [41] Lescaroux, F. "On the excess co-movement of commodity prices-A note about the role of fundamental factors in short-run dynamics", *Energy Policy*, Volume 37, Issue 10 (2009), pp. 3906-3913
- [42] Lu, Y. and S. N. Neftci, "Financial Instruments to Hedge Commodity Price Risk for Developing Countries", *International Monetary Fund Working Papers* (2008), 08/6.
- [43] Lombardi, Marco J., Osbat, Chiara and Schnatz, Bernd, "Global Commodity Cycles and Linkages: A FAVAR Approach", *European Central Bank Working Paper* (2010), No. 1170.
- [44] Neely, C. J. and D. Rapach, "International Comovements in Inflation Rates and Country Characteristics" (June 13, 2011). *Federal Reserve Bank of St. Louis Working Paper No. 2008-025*
- [45] Martin, A. Bayesian Analysis. In J. Box-Steffensmeier & D. Collier (Eds.), *The Oxford Handbook of Political Methodology*, Oxford University Press, (2008).
- [46] Moench, Emanuel; S. Ng, and S. Potter, "Dynamic Hierarchical Factor Models." *Federal Reserve Bank of New York Staff Report No. 412*, December 2009.
- [47] Mueller, S. A., J. E. Anderson and T. J. Wallington, "Impact of biofuel production and other supply and demand factors on food price increases in 2008", *Biomass and Bioenergy* (2011), Volume 35, Issue 5, Pages 1623-1632
- [48] Otrok, C. and C. H. Whiteman, "Bayesian Leading Indicators: Measuring and Predicting Economic Conditions in Iowa" *International Economic Review*, November (1998), 39(4), pp. 997-1014.
- [49] Paap, R., "What are the advantages of MCMC based inference in latent variable models?" *Statistica Neerlandica* 56, (2002), 2-22.

- [50] Pesaran, M. H. and Z. Zhao, "Bias Reduction in Estimating Long-run Relationships from Dynamic Heterogeneous Panels," in *Analysis of Panels and Limited Dependent Variables*, Cambridge University Press, (1999), chapter 12, 297-32
- [51] Phillips, Peter and Donggyu Sul, "Bias in dynamic panel estimation with fixed effects, incidental trends and cross section dependence," *Journal of Econometrics*, Volume 137, Issue 1, (March 2007), 162-188.
- [52] Pindyck, R. S. and J. J. Rotemberg. "The excess co-movement of commodity prices", *Economic Journal*, Vol. 100, (1990) pp. 1173-89.
- [53] Sargent, T.J. and C.A. Sims, "Business Cycle Modeling Without Pretending to Have Too Much A Priori Economic Theory," in Christopher A. Sims et al., eds., *New Methods in Business Cycle Research*, (1977), pp. 45-108).
- [54] Svensson, L.E.O., "The effect of monetary policy on real commodity prices: Comment", In *Asset Prices and Monetary Policy*, Ed. John Y. Campbell, NBER Working Paper (2008) 12713.
- [55] Tanner, M., and W.H. Wong, "The Calculation of Posterior Distributions by Data Augmentation," *Journal of the American Statistical Association* 82 (1987), 84-88.
- [56] Trostle R. "Global Agricultural supply and demand: factors contributing to the recent increase in food commodity prices", USDA Economic Research Service (July 2008), Report WRS-0801
- [57] Vansteenkiste, I., "How important are common factors in driving non-fuel commodity prices? A dynamic factor analysis", *European Central Bank Working Paper* (2009), no. 1072.
- [58] Western, B., "Causal Heterogeneity in Comparative Research: A Bayesian Hierarchical Modelling Approach," *American Journal of Political Science*, (1998), 42:1233–1259.
- [59] Wolf, M. "Life in a tough world of high commodity prices", *Financial Times*, (2008), March 4.

A Appendix

A.1 Estimation Details for Endogenous Clustering Framework

Following FOS (2011) and Kose et al. (2003), I estimate the model presented in equation (1) via Gibbs sampling. In particular, I utilize Metropolis-Hastings in Gibbs sampling to draw from the joint posterior distribution of the factors and the model's parameters. Given an initial draw of model's parameters, the factors can be extracted using Kalman filters based on Kim and Nelson (1999). In the next step, taking these factors as given, one can sample from the conditional density of the parameters. Once the parameters are known, Kalman filtering technique is applied again to extract the factors. Sampling from the conditional densities of the parameters and the factors is repeated many times. This is known as Gibbs sampling and under the regulatory assumptions (see Chip and Greenberg, 1995) these hese sequence of draws from the conditional distributions converge to the joint posterior density of the entire system, independent of the initial values selected.

To describe the sampler fully, let \mathbf{Y} represent the data, Θ represent the full set of model parameters and let \mathbf{F} represent the factors. We can define the set of blocks of parameters to be estimated in the sampler as: (1) the set of intercepts, α_i and global factor loadings, β_{i0} collected in the set $\rho = \{\alpha_i, \beta_{i0}\}$; (2) the set of innovation variances, $\sigma^2 = \{\sigma_i^2\}$; (3) the set of autoregressive parameters for the factors, $\phi = \{\phi_0, \dots, \phi_J\}$, (4) the sectoral factor loadings $\beta = \{\beta_{i,j}\}$ joint with the group probabilities $\gamma = \{\gamma_{i,j}\}$, (5), the set of factors, $\mathbf{F} = \{\mathbf{F}_0, \mathbf{F}_j\}$ and (6) the set of autoregressive parameters for the factors, $\varphi = \{\varphi_1, \dots, \varphi_I\}$

The steps of the Gibbs algorithm to sample from the joint distribution of Θ, \mathbf{F} are given as follows:

Step 1: Specify starting values $\Theta^{(0)}, \mathbf{F}^{(0)}$ and set $n = 0$.

Step 2: Simulate the unknown variables;

2.1: Sample $\rho^{(n+1)}$ from $p(\rho | (\sigma^2)^{(n)}, \phi^{(n)}, (\beta, \gamma)^{(n)}, \varphi^{(n)}, F^{(n)}, Y)$,

2.2: Sample $(\sigma^2)^{(n+1)}$ from $p(\sigma^2 | \rho^{(n+1)}, \phi^{(n)}, (\beta, \gamma)^{(n)}, \varphi^{(n)}, F^{(n)}, Y)$,

2.3: Sample $(\beta, \gamma)^{(n+1)}$ from $p(\beta, \gamma | \rho^{(n+1)}, (\sigma^2)^{(n+1)}, \phi^{(n)}, \varphi^{(n)}, F^{(n)}, Y)$,

2.4: Sample $\phi^{(n+1)}$ from $p(\phi | \rho^{(n+1)}, (\sigma^2)^{(n+1)}, (\beta, \gamma)^{(n+1)}, \varphi^{(n)}, F^{(n)}, Y)$,

2.5 Sample $\varphi^{(n+1)}$ from $p(\varphi | \rho^{(n+1)}, (\sigma^2)^{(n+1)}, (\beta, \gamma)^{(n+1)}, \phi^{(n+1)}, F^{(n)}, Y)$,

2.6 Apply Kalman Filter and sample $\mathbf{F}^{(n+1)}$.

Step 3: Set $n = n + 1$ and go to step 2.

This iteration loop is repeated 30,000 times and the initial 25,000 draws are discarded to allow for convergence. To initialize the sampler, I generate factors from a uniform normal distribution, and run the regressions of (1) and (2) separately. The coefficient estimates of factor loadings, factor AR parameters and variances for measurement errors are then used to start the sampler.

A.1.1 The Prior Distributions

The prior distributions and their corresponding hyperparameters are given below:

Priors for Estimation			
Parameter	Prior Distribution	Hyperparameters	
$\rho_i = [\alpha_i, \beta_{i,0}]'$	$N(\mathbf{r}, \mathbf{R})$	$\mathbf{r} = \mathbf{0}_2$; $\mathbf{R} = \mathbf{I}_2$	$\forall i$
$\beta_{i,j}$	$N(\mathbf{b}, \mathbf{B})$	$\mathbf{b} = \mathbf{0}$; $\mathbf{B} = \mathbf{1}$	$\forall i, j$
σ_i^2	$IG\left(\frac{\nu}{2}, \frac{\delta}{2}\right)$	$\nu = 6$; $\delta = 0.1$	$\forall i$
$\gamma_{i,j}$	Uniform (κ)	$\kappa_{ij} = \frac{1}{J}$	$\forall i, j$
ϕ	$N(\boldsymbol{\eta}, \boldsymbol{\Phi})$	$\boldsymbol{\eta} = \mathbf{0}_{q^f}$, $\boldsymbol{\Phi} = \frac{1}{2}\mathbf{I}_{q^f}$	$\forall j$
φ	$N(\bar{\boldsymbol{\mu}}, \bar{\boldsymbol{\Upsilon}})$	$\bar{\boldsymbol{\mu}} = \mathbf{0}_{q^\varepsilon}$, $\bar{\boldsymbol{\Upsilon}} = \frac{1}{2}\mathbf{I}_{q^\varepsilon}$	$\forall i$

Note that the cluster membership indicator has a uniform prior over all clusters – that is, a priori, a series is equally likely to belong to any cluster. Also, recall that the factor innovation variances, λ_k^2 , are constant and predetermined.

A.1.2 Notations

- The variance-covariance matrix for each factor k is $\lambda_k^2 \Sigma_k$, where

$$vec(\Sigma_k) = (I - \Phi_k^F \otimes \Phi_k^F)^{-1} vec(u_{q^f}' u_{q^f}),$$

and

$$\Phi_k^F = \begin{bmatrix} \phi_k & & \\ & I_{q^f-1} & \\ & & 0_{q^f-1 \times 1} \end{bmatrix}$$

is the companion matrix associated with autoregression (2). u_{q^f} is a $(q^f \times 1)$ vector with a 1 as the first element and zeros as the rest.

- Similarly, From (3) i^{th} idiosyncratic measurement error is given by;

$$\varepsilon_{it} = y_{it} - \alpha_i - \beta_{i,0}F_{0,t} - \sum_j^J \gamma_{i,j}\beta_{i,j}F_{jt}.$$

Then stack ε_{it} as a vector $\widehat{\varepsilon}_{iq} = (\varepsilon_{i,q^\varepsilon+1-q}, \dots, \varepsilon_{i,T-q})'$ and define

$$\boldsymbol{\varepsilon}_i = [\widehat{\varepsilon}_{i1}, \dots, \widehat{\varepsilon}_{iq^\varepsilon}].$$

- Let I be the total number of observations. $\gamma_i = (\gamma_1, \gamma_2 \dots \gamma_J)$ denotes the probability of belonging to clusters from 1 to J for each series i . By assumption only one of γ_i 's elements is 1 where all the others are zero.

A.1.3 The Sampler

Generating $\rho|\Theta_{-\rho}, \mathbf{F}, \mathbf{Y}$ Given previous draw of factors, the equations in (1) are just i independent regressions with serially correlated errors. Following Chib and Greenberg (1994) we need to account for the serially correlation in the error structure before writing down the distribution of the parameter block. This can be done by building the likelihood for the first q^ε observations and continue building the posterior distribution for the rest. To begin define $\mathbf{X}_i^* = [\mathbf{1}_T, \mathbf{F}_0]$, where $\mathbf{1}_T$ is a $(T \times 1)$ vector of ones and \mathbf{F}_0 is the last draw of the global for series i , respectively and let $\mathbf{Y}_i^* = Y_i - \sum_{j=1}^J \gamma_{ij}\beta_{ij}\mathbf{F}_{j,t}$. The following steps lists the process as in Chib and Greenberg (1994)

1. $\widetilde{\mathbf{X}}_{i,1}^* = \begin{bmatrix} 1 & \mathbf{F}_{0,1} \\ \dots & \dots \\ 1 & \mathbf{F}_{0,q^\varepsilon} \end{bmatrix}$ denote the first q^ε rows of \mathbf{X}_i^* ;
2. $\widetilde{\mathbf{Y}}_{i,1}^* = (\mathbf{Y}_{i,1}^*, \mathbf{Y}_{i,2}^*, \dots, \mathbf{Y}_{i,q^\varepsilon}^*)$ denote the first q^ε observations of \mathbf{Y}_i^* ;
3. $\widetilde{\mathbf{X}}_{i,1} = Q_i^{-1}\widetilde{\mathbf{X}}_{i,1}^*$ and $\widetilde{\mathbf{Y}}_{i,1} = Q_i^{-1}\widetilde{\mathbf{Y}}_{i,1}^*$;
4. $\widetilde{\mathbf{X}}_{i,2}$ be a $(T - q^\varepsilon) \times 2$ matrix with t^{th} row given by $\varphi_i(L)(\mathbf{X}_{i,t}^*)'$;
5. $\widetilde{\mathbf{Y}}_{i,2}$ be a vector of length $(T - q^\varepsilon)$ with t^{th} row given by $\varphi_i(L)(\mathbf{Y}_{i,t}^*)'$;

6. Finally, in stack form define $\tilde{\mathbf{X}}_i = \begin{bmatrix} \tilde{\mathbf{X}}_{i,1} \\ \tilde{\mathbf{X}}_{i,2} \end{bmatrix}$ and $\tilde{\mathbf{Y}}_i = \begin{bmatrix} \tilde{\mathbf{Y}}_{i,1} \\ \tilde{\mathbf{Y}}_{i,2} \end{bmatrix}$.

Then for each observation i , $\boldsymbol{\rho}_i = [\alpha_i, \beta_{i0}]'$ is drawn from

$$\boldsymbol{\rho}_i | \Theta_{-\boldsymbol{\rho}_i}, \mathbf{F}, \mathbf{Y} \sim N(\mathbf{r}_i, \mathbf{R}_i),$$

where $\mathbf{R}_i = (\mathbf{R}_0^{-1} + \sigma_i^{-2} \tilde{\mathbf{X}}_i' \tilde{\mathbf{X}}_i)^{-1}$ and $\mathbf{r}_i = \mathbf{R}_i (\mathbf{R}_0^{-1} \mathbf{r}_0 + \sigma_i^{-2} \tilde{\mathbf{X}}_i' \tilde{\mathbf{Y}}_i)$.

Generating $\sigma_i^2 | \Theta_{-\sigma_i^2}, \mathbf{F}, \mathbf{Y}$ σ_i^2 conditional on \mathbf{Y} and $\Theta_{-\sigma_i^2}$, can be drawn from the inverse gamma posterior;

$$\sigma_i^{-2} | \mathbf{Y}, \mathbf{X}, \Theta_{-\sigma_i^2} \sim \Gamma\left(\frac{\nu_0 + T}{2}, \frac{2}{d_0 + d_i' d_i}\right),$$

where $d_i = \tilde{\mathbf{Y}}_i - \tilde{\mathbf{X}}_i \boldsymbol{\rho}_i$.

Generating $\boldsymbol{\gamma}, \boldsymbol{\beta} | \Theta_{-\boldsymbol{\gamma}, \boldsymbol{\beta}}, \mathbf{F}, \mathbf{Y}$ This step samples the cluster probability and the cluster loadings jointly following FOS (2011). FOS simply utilize an algorithm similar to that of sections 2.5 and 2.6 in Holmes and Held (2006). The joint distribution we are interested in is:

$$p(\boldsymbol{\beta}, \boldsymbol{\gamma} | \Theta, \mathbf{F}) = p(\boldsymbol{\gamma} | \Theta, \mathbf{F}) p(\boldsymbol{\beta} | \Theta, \boldsymbol{\gamma}, \mathbf{F}).$$

As the closed form for the joint density is not available, this step requires a Metropolis-Hastings sampler within Gibbs draw. Following, Holmes and Held (2006) we can define a joint proposal density, $q(\beta_i^*, \gamma_i^*)$ as;

$$q(\beta_i^*, \gamma_i^*) = p(\beta_i^* | \gamma_i^*, \Theta, \mathbf{F}) q(\gamma_i^* | \gamma_i),$$

where β_i^* and γ_i^* are the candidates and β_i and γ_i are held over from the last draw.

The idea is to draw γ_i^* from a proposal density and to sample β_i^* from its full conditional distribution given this current draw of γ_i^* . The candidates β_i^* and γ_i^* are then accepted with an acceptance probability α . If the candidates are rejected, then the past draws are retained.

The proposal density for γ_i , is assumed to be symmetric in which one draws a random element $\gamma_{i,j}$ and set it equal to 1, while setting all other elements of γ_i to zero. Given this draw of cluster probability, one can sample the candidate β_i^* from the full conditional distribution. In order to compute it, similar to the draw for ρ_i , first define $\bar{\mathbf{X}}_i = \sum_j \gamma_{ij}^* \mathbf{F}_j$ and $\bar{\mathbf{Y}}_i = (\mathbf{Y}_i - \alpha_i \mathbf{1}_t - \beta_{0,i} \mathbf{F}_0)$, and let;

1. $\bar{\mathbf{X}}_{i,1}^* = \begin{bmatrix} \sum_j \gamma_{i,j}^* \mathbf{F}_{j,1} \\ \dots \\ \sum_j \gamma_{i,j}^* \mathbf{F}_{j,q^\varepsilon} \end{bmatrix}$ denote the first q^ε rows of $\bar{\mathbf{X}}_i$;
2. $\bar{\mathbf{Y}}_{i,1}^* = (\bar{\mathbf{Y}}_{i,1}, \bar{\mathbf{Y}}_{i,2}, \dots, \bar{\mathbf{Y}}_{i,q^\varepsilon})$ denote the first q^ε observations of $\bar{\mathbf{Y}}_i$;
3. $\hat{\mathbf{X}}_{i,1}^* = Q_i^{-1} \bar{\mathbf{X}}_{i,1}^*$ and $\hat{\mathbf{Y}}_{i,1}^* = Q_i^{-1} \bar{\mathbf{Y}}_{i,1}^*$;
4. $\hat{\mathbf{X}}_{i,2}^*$ be a vector of length $(T - q^\varepsilon)$ matrix with t^{th} row given by $\varphi_i(L)(\bar{\mathbf{X}}_{i,t})'$;
5. $\hat{\mathbf{Y}}_{i,2}^*$ be a vector of length $(T - q^\varepsilon)$ with t^{th} row given by $\varphi_i(L)\bar{\mathbf{Y}}_{i,t}$;
6. Finally, in stack form define $\hat{\mathbf{X}}_i = \begin{bmatrix} \hat{\mathbf{X}}_{i,1}^* \\ \hat{\mathbf{X}}_{i,2}^* \end{bmatrix}$ and $\hat{\mathbf{Y}}_i = \begin{bmatrix} \hat{\mathbf{Y}}_{i,1}^* \\ \hat{\mathbf{Y}}_{i,2}^* \end{bmatrix}$.

Then, candidate β_i^* can be drawn from the full conditional distribution below:

$$\beta_i | \Theta_{-\beta, \gamma}, \gamma_i^*, \mathbf{F}, \mathbf{Y} \sim N(\mathbf{b}_i^*, \mathbf{B}_i^*), \quad (7)$$

where $\mathbf{B}_i^* = (\mathbf{B}_0 + \sigma_i^{-2} \hat{\mathbf{X}}_i' \hat{\mathbf{X}}_i)^{-1}$, $\mathbf{b}_i^* = \mathbf{B}_i^* (\mathbf{B}_0^{-1} \mathbf{b}_0 + \sigma_i^{-2} \hat{\mathbf{X}}_i' \hat{\mathbf{Y}}_i)$.

The acceptance probability is written in the following form:

$$\alpha = \min \left\{ 1, \frac{f(y|\gamma^*, \beta^*, \Theta_{-\beta, \gamma}, F) p(\beta^*) p(\gamma^*) q(\gamma|\gamma^*) q(\beta|\beta^*)}{f(y|\gamma, \beta, \Theta_{-\beta, \gamma}, F) p(\beta) p(\gamma) q(\gamma^*|\gamma) q(\beta^*|\beta)} \right\}, \quad (8)$$

where the first term is the likelihood; the second term is the prior for γ evaluated at either the candidate or the past draw; the third term is the prior for β ; and the last two terms are the probability of a move. This acceptance probability can be simplified further. First off all, β^l s are drawn from the full conditional densities which cancels out the probabilities with β^l s from above. The choice of the symmetric proposal density for γ_n implies that the probability of moving from

γ_i^* to γ_i is exactly the same as moving from γ_i to γ_i^* , so that $q(\gamma_i^*|\gamma_i) = q(\gamma_i|\gamma_i^*)$. Given also that γ has a uniform prior, which implies $p(\gamma^*) = p(\gamma)$, equation(8) reduces to;

$$\alpha = \min \left\{ 1, \frac{f(y|\gamma^*, \beta^*, \Theta, F)}{f(y|\gamma, \beta, \Theta, F)} \right\}.$$

Finally, incorporating the normal likelihoods yields:

$$\alpha_i = \min \left\{ 1, \frac{|\mathbf{B}_i^*|^{1/2} \exp\left(\frac{1}{2}\mathbf{b}_i^* \mathbf{B}_i^{*-1} \mathbf{b}_i^*\right)}{|\mathbf{B}_i|^{1/2} \exp\left(\frac{1}{2}\mathbf{b}_i \mathbf{B}_i^{-1} \mathbf{b}_i\right)} \right\}, \quad (9)$$

where \mathbf{b}_i^* and \mathbf{B}_i^* are defined as above and \mathbf{b}_i and \mathbf{B}_i are calculated using the value for γ_i from the past draw. Note that, the draw of the indicator γ_i determines which series enter into the distribution of each group factor.

Generating $\phi|\Theta_{-\phi}, \mathbf{F}, \mathbf{Y}$ Since the conditional density of ϕ has an unknown form, it cannot be sampled directly. I apply Chib and Greenberg (1994) in drawing $\phi = [\phi_0, \phi_1, \dots, \phi_k]$ conditional on the factors, data, and remaining parameters using a Metropolis-Hastings algorithm. For each iteration, one generates a candidate draw ϕ^* from the proposal distribution below:

$$\phi_k^* \sim N\left(\hat{\phi}_k, \mathbf{V}_k^{-1}\right),$$

where

$$\mathbf{V}_k = \mathbf{\Phi}_k^{-1} + \lambda_k^2 \mathbf{e}_k \mathbf{e}_k'$$

and

$$\hat{\phi}_k = \mathbf{V}_k^{-1} (\mathbf{\Phi}_k^{-1} \boldsymbol{\mu}_k + \lambda_k^2 \mathbf{e}_k' \hat{e}_{k0}).$$

The candidate ϕ^* is then accepted with a probability that is determined by the likelihood of the data: $\alpha_k = \min\{\hat{\alpha}_k, 1\}$, where

³²Refer to the "notations" above for the equation of \mathbf{e}_k .

$$\widehat{\alpha}_k = \frac{\Psi(\phi_k^*)}{\Psi(\phi_k^{(n-1)})},$$

and

$$\Psi(\phi_k) = |\Sigma_k(\phi_k)|^{-1/2} \exp \left[-\frac{1}{2\lambda_k^2} \widehat{e}'_{k0} \Sigma_k^{-1}(\phi_k) \widehat{e}_{k0} \right],$$

with the superscript $n - 1$ reflecting the previous iteration. If the draw is less than the acceptance probability, the candidate is accepted. If not, the past draw is retained.

Overall, the draw works as follows:

1. First generate the candidate from the proposal density, ϕ^* ,
2. Draw from a standard uniform distribution,
3. If the draw is less than the acceptance probability, α_k , set $\phi^{(n)} = \phi^*$
4. Otherwise, retain past draw, $\phi^{(n)} = \phi$.

Generating $\varphi | \Theta_{-\varphi}, \mathbf{F}, \mathbf{Y}$ The draw for φ follows the same steps as the draw for ϕ . The autoregression coefficients for the innovation coefficients, $\varphi = [\varphi_1, \dots, \varphi_i]$, conditional on the factors, data, and remaining parameters are drawn from

$$\varphi_i^* \sim N(\boldsymbol{\mu}_i, \boldsymbol{\Upsilon}_i^{-1}),$$

where $\boldsymbol{\mu}_i$, $\boldsymbol{\Upsilon}_i$ and the pseudolikelihood $\Psi(\varphi_i)$, follows from above with the necessary change in notation. Here we would have $\boldsymbol{\Upsilon}_i = \bar{\boldsymbol{\Upsilon}}_i + \sigma_i^{-2} \boldsymbol{\varepsilon}_i \boldsymbol{\varepsilon}_i'$, $\boldsymbol{\mu}_i = \boldsymbol{\Upsilon}_i^{-1} (\bar{\boldsymbol{\Upsilon}}_i \bar{\boldsymbol{\mu}}_i + \sigma_i^{-2} \boldsymbol{\varepsilon}_i' \widehat{\boldsymbol{\varepsilon}}_{i0})$ and $\varphi_i = |\Omega_i(\varphi_i)|^{-1/2} \exp \left[-\frac{1}{2\sigma_i^2} \widehat{\boldsymbol{\varepsilon}}_{i0}' \Omega_i^{-1}(\varphi_i) \widehat{\boldsymbol{\varepsilon}}_{i0} \right]$.

Generating $\mathbf{F} | \Theta, \mathbf{Y}$ I follow Kim and Nelson (1999) in sampling from the conditional posterior density of the factors given the model's parameters, where the draw of the indicator γ_{ij} determines which series enter into the distribution of each cluster factor. Assume for simplicity that factors and observation errors have the same lag length ($q^f = q^\varepsilon$) and denote it by q . Let $Y_t = (y_{1,t}, y_{2,t}, \dots, y_{i,t})$, $F_t = (F_{0,t}, F_{1,t}, \dots, F_{k,t})$, $\boldsymbol{\varepsilon}_t = (\varepsilon_{1,t}, \varepsilon_{2,t}, \dots, \varepsilon_{i,t})$, $e_t =$

$(e_{0,t}, e_{1,t}, \dots, e_{k,t})$ and $\epsilon_t = (\epsilon_{0,t}, \epsilon_{1,t}, \dots, \epsilon_{k,t})$ then we can write the state space as by stacking the state variables (factors and observation error terms):

Measurement Equation:

$$[Y_t] = A + [H] \begin{bmatrix} F_t \\ \epsilon_t \\ F_{t-1} \\ \epsilon_{t-1} \\ \vdots \\ F_{t-q} \\ \epsilon_{t-q} \end{bmatrix}$$

where $A = (\alpha_1, \alpha_2, \dots, \alpha_i)'$ and H is $(i \times (k+i)q)$ matrix given below:

$$H = \begin{bmatrix} \beta_{1,0} & \gamma_{1,1}\beta_{1,1} & \cdots & \gamma_{i,f}\beta_{i,f} & 1 & \cdots & \cdots & 0 & 0 & \cdots & 0 \\ \beta_{2,0} & \gamma_{2,1}\beta_{2,1} & \cdots & \gamma_{i,f}\beta_{i,f} & 0 & 1 & \cdots & 0 & 0 & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots & \vdots & \ddots & \ddots & \vdots & \vdots & \ddots & \vdots \\ \beta_{i,0} & \gamma_{i,1}\beta_{i,1} & \cdots & \gamma_{i,f}\beta_{i,f} & 0 & 0 & 0 & 1 & 0 & \cdots & 0 \end{bmatrix}$$

with zero variance covariance matrix ($R = 0$), since we stacked all the observation error terms (ϵ_t) as state variables.

Transition Equation:

$$\begin{bmatrix} F_{t+1} \\ \epsilon_{t+1} \\ \vdots \\ F_{t+2-q} \\ \epsilon_{t+2-q} \end{bmatrix} = \mathcal{F} \begin{bmatrix} F_t \\ \epsilon_t \\ \vdots \\ F_{t-q} \\ \epsilon_{t-q} \end{bmatrix} + \begin{bmatrix} \epsilon_{t+1} \\ \epsilon_{t+1} \\ 0 \\ \vdots \\ 0 \end{bmatrix}$$

where \mathcal{F} is $(k+i)q \times (k+i)q$ matrix given by,

$$\mathcal{F} = \begin{bmatrix} \text{diag}(\phi_{f,1}, \psi_{i,1}) & \text{diag}(\phi_{f,2}, \psi_{i,2}) & \cdots & \text{diag}(\phi_{f,q}, \psi_{i,q}) \\ \mathbf{1}_{(k+i)(q-1) \times (k+i)(q-1)} & & & \mathbf{0}_{(k+i) \times (k+i)} \end{bmatrix}$$

with variance-covariance matrix;

$$Q = \begin{bmatrix} \text{diag}(\lambda_k, \sigma_i) & \mathbf{0}_{(k+i)q \times (k+i)(q-1)} \\ \mathbf{0}_{(k+i)(q-1) \times (k+i)q} & \end{bmatrix}$$

Then the standard Kalman filtering technique can be applied. Let the state vector represented as $\mathbf{S}_t = [F_t \ \varepsilon_t \ \dots \ F_{t-q} \ \varepsilon_{t-q}]'$

Given initial values for $\mathbf{S}_{1|0}$ and for the unconditional density of the state vector $P_{1|0}$ the Kalman filter is run from $t = 1$ to $t = N$ following the steps below:

The prediction Step:

$$\begin{aligned} \mathbf{S}_{t|t-1} &= \mathcal{F}\mathbf{S}_{t-1|t-1} \\ P_{t|t-1} &= \mathcal{F}P_{t-1|t-1}\mathcal{F}' + Q \end{aligned}$$

Update:

$$\begin{aligned} \mathbf{S}_{t|t} &= \mathbf{S}_{t|t-1} + P_{t|t-1}H'(H'P_{t|t-1}H + R)^{-1}(Y_t - A - H\mathbf{S}_{t|t-1}) \\ P_{t|t} &= P_{t|t-1} - P_{t|t-1}H'(H'P_{t|t-1}H + R)^{-1}HP_{t|t-1} \end{aligned}$$

Once the Kalman filtering step is over, the factors are sampled from a Gaussian distribution. Since the Q matrix is singular an additional step is required to modify the densities one samples from as in Kim and Nelson (1999). Shrink the size of Q to $(k+n) \times (k+n)$ by excluding all the zeros in the matrix. Let $*$'s denote the reduced rank matrices. Now one samples the factors

recursively from Normal distribution with mean $\mathbf{S}_{t|t}^*$ and the variance $P_{t|t}^*$

$$\begin{aligned}\mathbf{S}_{t|t}^* &= \mathbf{S}_{t|t} + P_{t|t} \mathcal{F}^{*'} (\mathcal{F}^* P_{t|t} \mathcal{F}^{*'} + Q^*)^{-1} (\mathbf{S}_{t+1}^* - \mathcal{F}^* \mathbf{S}_{t|t}) \\ P_{t|t}^* &= P_{t|t} - P_{t|t} \mathcal{F}^{*'} (\mathcal{F}^* P_{t|t} \mathcal{F}^{*'} + Q^*)^{-1} \mathcal{F}^* P_{t|t}\end{aligned}$$

where \mathbf{S}_{t+1}^* is the reduced $(k+n) \times 1$ state vector sampled from a Normal distribution with mean $\mathbf{S}_{t|t}$ and variance $P_{t|t}$ after the Kalman filter step.

B Implementation of Chib's Bayes factor algorithm

The method follows Chib (1995). The marginal likelihood of the model itself is given as:

$$\ln \hat{m}(\mathbf{Y}) = \ln f(\mathbf{Y}|\Theta^*) + \ln p(\Theta^*) - \ln \hat{p}(\Theta^*|\mathbf{Y}),$$

where Θ is the vector of model parameters, $\ln \hat{m}(\mathbf{Y})$ is the log marginal likelihood, $\ln f(\mathbf{Y}|\Theta^*)$ is the log likelihood evaluated at a given $\Theta = \Theta^*$, $\ln p(\Theta^*)$ is the log of the prior evaluated at Θ^* , and $\ln \hat{p}(\Theta^*|\mathbf{Y})$ is an approximation of the posterior ordinate. Θ^* need only be a high density value of Θ (e.g., a modal point). The posterior ordinates can be approximated using the Gibbs output of the full model run. In particular, the posterior ordinate for N sampler blocks that were previously defined is given as;

$$\hat{p}(\Theta^*|\mathbf{Y}) = p(\Theta_1^*|\mathbf{Y}) \times p(\Theta_2^*|\mathbf{Y}, \Theta_1^*) \times \dots \times p(\Theta_N^*|\mathbf{Y}, \Theta_1^*, \dots, \Theta_{N-1}^*),$$

where a typical term is written as;

$$\hat{p}(\Theta_n^*|\mathbf{Y}) = \frac{1}{G} \sum_{g=1}^G p\left(\Theta_n^*|\mathbf{Y}, \Theta_1^*, \dots, \Theta_{n-1}^*, \Theta_{n+1}^{(g)}, \dots, \Theta_N^{(g)}, F_0^{(g)}, F_1^{(g)}, \dots, F_J^{(g)}\right).$$

Excluding the latent factors, there are 6 blocks of parameters to determine the posterior ordinates for. Sections below describes each one of them.

B.1 Calculation of the posterior ordinate $\hat{p}(\rho^*|\mathbf{Y})$

Define $\Theta = \{\rho, \sigma^2, \varphi, [\gamma, \beta], \phi\}$ along with $F_0^{(g)}$, and $F_1^{(g)}, \dots, F_J^{(g)}$ where g denotes the number of Gibbs iterations. Let ρ^* be the posterior mode of ρ . Recall that prior of ρ is $N(\mathbf{r}_0, \mathbf{R}_0)$ and the posterior is $N(\mathbf{r}_i, \mathbf{R}_i)$ for each i , where and

$$\mathbf{R}_i = \left(\mathbf{R}_0^{-1} + \sigma_i^{-2} \tilde{\mathbf{X}}_i' \tilde{\mathbf{X}}_i \right)^{-1},$$

$$\mathbf{r}_i = \mathbf{R}_i \left(\mathbf{R}_0^{-1} \mathbf{r}_0 + \sigma_i^{-2} \tilde{\mathbf{X}}_i' \tilde{\mathbf{Y}}_i \right),$$

and $\tilde{\mathbf{X}}_i$ and $\tilde{\mathbf{Y}}_i$ are defined appropriately from above. Then, $p(\rho^*|\mathbf{Y})$ is approximated by

$$\hat{p}(\rho^*|\mathbf{Y}) = \frac{1}{G} \sum_{g=1}^G N\left(\rho^*|y, \Theta^{(g)}, \mathbf{F}^{(g)}\right),$$

where $N(\cdot)$ is the normal pdf with mean and variance defined by the full Gibbs run. As noted in Chib (1995) when calculating the posterior ordinate for the first block we do not need to resample a reduced Gibbs run. Instead the draws from the full Gibbs run should be used to evaluate the following posterior ordinate:

$$\hat{p}(\rho^*|\mathbf{Y}) = \frac{1}{G} \sum_{g=1}^G N\left(\rho^*, \mathbf{r}_i^{(g)}, \mathbf{R}_i^{(g)}\right), \quad (10)$$

where $\mathbf{r}_i^{(g)}$ and $\mathbf{R}_i^{(g)}$ as defined above are saved from the Full Gibbs run along with the values of $\mathbf{F}^{(g)}$ and $\Theta^{(g)}$.

B.2 Calculation of the posterior ordinate $\hat{p}(\sigma^{2*}|\mathbf{Y}, \rho^*)$

Next, we require $\hat{p}(\sigma^{2*}|\mathbf{Y}, \mu^*)$, which is obtained from an additional (reduced) G runs of the Gibbs sampler holding the previous block fixed at its modal values and sampling all the other model parameters including σ^2 and latent factors, i.e. $\{\sigma^{2(g)}, \phi^{(g)}, \varphi^{(g)}, \beta^{(g)}, \gamma^{(g)}, \mathbf{F}^{(g)}\}$. Then the posterior estimate is calculated using these draws from the reduced conditional Gibbs run as;

$$\hat{p}\left(\sigma^{2*}|\mathbf{Y}, \rho^*\right) = \frac{1}{G} \sum_{g=1}^G \Gamma^{-1}\left(\sigma^{2*}|\mathbf{Y}, \rho^*, \Theta_{-\rho}^{(g)}, \mathbf{F}^{(g)}\right),$$

where $\Gamma^{-1}(\cdot)$ is the pdf of the inverted gamma distribution. This is operationalized by recalling that σ_i^2 is assumed to have a prior distribution $\sigma_i^2 \sim \Gamma^{-1}\left(\frac{\nu_0}{2}, \frac{\Upsilon_0}{2}\right)$. The posterior distribution from the reduced Gibbs run is saved for each iteration. Then the posterior pdf is evaluated at the modal value of σ^{2*} . The average across iterations yields the posterior distribution as;

$$\hat{p}\left(\sigma^{2*}|\mathbf{Y}, \rho^*\right) = \frac{1}{G} \sum_{g=1}^G \Gamma^{-1}\left(\sigma^{2*}, \frac{\nu_0 + T}{2}, \frac{\Upsilon_0 + \tilde{\boldsymbol{\varepsilon}}_{iT}^{(g)'} \tilde{\boldsymbol{\varepsilon}}_{iT}^{(g)}}{2}\right). \quad (11)$$

B.3 Calculation of the posterior ordinate $\hat{p}(\phi^*|\mathbf{Y}, \rho^*, \sigma^{2*})$

Next, we require $\hat{p}(\phi^*|\mathbf{Y}, \rho^*, \sigma^{2*})$, which – because the parameter is drawn via an MH-in-Gibbs step – is obtained from the method of Chib and Jeliazkov (2001). Their method requires us to save the original draws from the full run and to resample additional G draws of the Gibbs sampler denoted by $\{\phi^{(g)}, \varphi^{(g)}, \beta^{(g)}, \gamma^{(g)}, \mathbf{F}^{(g)}\}$ for the numerator holding the previous blocks fixed at ρ^* and σ^{2*} . The denominator needs an additional M reduced Gibbs run for $\{\varphi^{(g)}, \beta^{(g)}, \gamma^{(g)}, \mathbf{F}^{(g)}\}$ holding the aforementioned previous blocks as well as the current block fixed (ϕ^*) at their corresponding modal values. We can then compute

$$\hat{p}\left(\phi^*|\mathbf{Y}, \rho^*, \sigma^{2*}\right) = \frac{\frac{1}{G} \sum_g \hat{\alpha}\left(\phi^{(g)}, \phi^*|\boldsymbol{\Theta}_{-\rho, \sigma^2}^{(g)}, \mathbf{F}^{(g)}\right) \hat{q}\left(\phi^{(g)}, \phi^*|\boldsymbol{\Theta}_{-\rho, \sigma^2}^{(g)}, \mathbf{F}^{(g)}\right)}{\frac{1}{M} \sum_j \hat{\alpha}\left(\phi^*, \phi^{(m)}|\boldsymbol{\Theta}_{-\rho, \sigma^2, \phi^*}^{(m)}, \mathbf{F}^{(m)}\right)}$$

where $\hat{q}\left(\phi^{(g)}, \phi^*\right) = N\left(\hat{\boldsymbol{\phi}}_i^*, \mathbf{V}_i^{*-1}\right)$ and the acceptance probabilities are defined above. A similar procedure follows for the posterior ordinate of φ .

B.4 Calculation of the posterior ordinate $\hat{p}(\beta^*, \gamma^*|\mathbf{Y}, \rho^*, \sigma^{2*}, \phi^*)$

Next, we require $\hat{p}(\beta^*, \gamma^*|\mathbf{Y}, \rho^*, \sigma^{2*}, \phi^*)$, which is obtained from both the retained full run and an additional G runs of the Gibbs sampler. Define $\varrho = [\beta, \gamma]$. This step follows from Chib and Jeliazkov (2001) as described in the previous section. The posterior ordinate estimate is calculated with additional G runs for the numerator and additional M runs for the denominator given as below:

$$\hat{p}\left(\varrho^*|\mathbf{Y}, \mu^*, \sigma^{2*}, \phi^*\right) = \frac{\frac{1}{G} \sum_g \alpha\left(\varrho^{(g)}, \varrho^*|F^{(g)}\right) q\left(\varrho^{(g)}, \varrho^*|\mathbf{F}^{(g)}\right)}{\frac{1}{M} \sum_j \alpha\left(\varrho^*, \varrho^{(m)}|\mathbf{F}^{(m)}\right)},$$

where the proposal density, $q(\cdot, \cdot)$, and the acceptance probability, $\alpha(\cdot, \cdot)$ are defined above.³³

B.5 Calculation of the log likelihood evaluated at Θ^*

The log likelihood evaluated at the modal point, Θ^* , can be computed by Monte Carlo integration from the average of the likelihoods for draws of the underlying latent variables:

$$\ln f(\mathbf{Y}|\Theta^*) = \frac{1}{G} \sum_{g=1}^G \ln f(\mathbf{Y}|\Theta^*, \mathbf{F}^{(g)}). \quad (12)$$

To compute this, we would set the model parameters at the mode and use factors sampled from the full Gibbs run to compute the likelihood at each point. The log-likelihood would then be the average of these likelihoods across iterations.

B.6 Calculation of the prior evaluated at Θ^*

The second term in (4) represents the prior distributions evaluated at their modal values and can be evaluated as

$$\ln p(\Theta^*|y) = \ln p(\Theta_1^*) + \ln p(\Theta_2^*) + \dots + \ln p(\Theta_i^*).$$

C Application of Bayesian Linear Regression

We are interested in approximating equation (6). To do so, we can save every 50th draw from the full posterior distribution of the model factors, and run Bayesian linear regression on each of the saved factor draws. Ξ_t represents the set of the variables we want to test on the factors $F_{k,t}$. Then we can estimate the linear regression of the form

$$F_{k,t} = \varpi_k \Xi_t + v_k$$

where the error term v is assumed to be normally distributed with mean zero and variance \oplus_k^2 . The estimation is a simple Gibbs application with two sampler blocks of parameters, namely

³³Note that the notation here is a move from the first to the second.

ϖ and \oplus^2 . Let the prior distributions for the coefficients and the variance be represented with $N(a, b)$ and $IG\left(\frac{c}{2}, \frac{d}{2}\right)$ respectively ($a = 0, b = 2, c = 6, d = 0.1$).

Generating $\varpi|\oplus^2\mathbf{F}, \Xi$ The conditional distribution can be drawn from

$$\varpi_k|\oplus^2, \mathbf{F}, \Xi \sim N(\mathbf{A}_k, \mathbf{B}_k),$$

where $\mathbf{A}_k = (\mathbf{b}^{-1} + \oplus_k^{-2}\Xi'\Xi)^{-1}$ and $\mathbf{B}_k = \mathbf{A}_k(\mathbf{b}^{-1}\mathbf{a} + \oplus_k^{-2}\Xi'F_k)$.

Generating $\oplus^2|\varpi, \mathbf{F}, \Xi$ \oplus^{-2} conditional on \mathbf{F}, Ξ and ϖ , can be drawn from the gamma posterior;

$$\oplus_k^{-2}|\varpi, \mathbf{F}, \Xi \sim \Gamma\left(\frac{c+T}{2}, \frac{2}{d + D_k'D_k}\right),$$

where $D_k = \mathbf{F}_k - \varpi_k\Xi_t$.

Recall that for each factor k , we saved every 50th draw. We apply the steps above to each saved draw of $F_{k,t}$ for 1000 iterations while burning in the first 500. This gives us many posterior distributions for each block ϖ_k and \oplus_k^{-2} . Then the posterior distributions of these parameter blocks are pooled together from which we can make inferences. From these pooled posterior distributions I report the mean of ϖ along with the Bayesian confidence intervals. The confidence interval is the 5th and 95th percentile interval endpoints of the pooled distribution.

D Data

The list of primary commodities and their explanations are directly taken from IFS database.

Wheat: United States, No.1 Hard Red Winter, ordinary protein, FOB Gulf of Mexico, US\$ per metric tonne.

Maize (corn): United States. No.2 Yellow, FOB Gulf of Mexico, U.S. price, US\$ per metric tonne.

Rice: Thailand, 5 percent broken milled white rice, Thailand nominal price quote, US\$ per metric tonne.

Barley: Canada, no.1 Western Barley, spot price, US\$ per metric tonne.

Soybean Meal: United States, Chicago Soybean Meal Futures (first contract forward) Minimum 48 percent protein, US\$ per metric tonne.

Soybean Oil: United States, Chicago Soybean Oil Futures (first contract forward) exchange approved grades, US\$ per metric tonne.

Soybeans: United States., Chicago Soybean futures contract (first contract forward) No. 2 yellow and par, US\$ per metric tonne.

Fishmeal: Peru, Fish meal/pellets 65% protein, CIF, US\$ per metric tonne.

Sunflower oil: United Kingdom, US export price from Gulf of Mexico, US\$ per metric tonne.

Olive Oil: United Kingdom, extra virgin less than 1% free fatty acid, ex-tanker price U.K., US\$ per metric tonne.

Palm oil: Malaysia, Palm Oil Futures (first contract forward) 4-5 percent FFA, US\$ per metric tonne.

Rapeseed (referred as Canola) oil: United Kingdom, crude, fob Rotterdam, US\$ per metric tonne

Groundnuts (peanuts): Nigeria, 40/50 (40 to 50 count per ounce), US\$ per metric tonne.

Beef: Australia and New Zealand, 85% lean fores, CIF U.S. import price, US cents per pound.

Lamb: New Zealand, frozen carcass Smithfield London, US cents per pound.

Swine (pork): United States, 51-52% lean Hogs, US cents per pound.

Poultry (chicken): United States, Whole bird spot price, Ready-to-cook, whole, iced, Georgia docks, US cents per pound.

Fish (salmon): Norway, Farm Bred Norwegian Salmon, export price, US\$ per kilogram.

Shrimp: United States, No.1 shell-on headless, 26-30 count per pound, Mexican origina, New York port, US cents per pound.

Sugar: World, Free Market, Coffee Sugar and Cocoa Exchange (CSCE) contract no.11 nearest future position, US cents per pound.

Sugar: United States, U.S. import price, contract no.14 nearest futures position, US cents per pound (Footnote: No. 14 revised to No. 16).

Oranges: France, miscellaneous oranges CIF French import price, US\$ per metric tonne.

Bananas: Central America and Ecuador, FOB U.S. Ports, US\$ per metric tonne.

Coffee: Africa not specified, Robusta, International Coffee Organization New York cash price, ex-dock New York, US cents per pound.

Coffee, Other Mild Arabicas, International Coffee Organization New York cash price, ex-dock New York, US cents per pound.

Cocoa beans: Ghana, International Cocoa Organization cash price, CIF US and European ports, US\$ per metric tonne.

Tea: Mombasa, Kenya, Auction Price, US cents per kilogram, From July 1998, Kenya auctions, Best Pekoe Fannings. Prior, London auctions, c.i.f. U.K. warehouses.

Hard Logs: Malaysia, Best quality Malaysian meranti, import price Japan, US\$ per cubic meter.

Soft Logs: United States, Average Export price from the U.S. for Douglas Fir, US\$ per cubic meter.

Hard Sawnwood: Malaysia, Dark Red Meranti, select and better quality, C&F U.K port, US\$ per cubic meter.

Soft Sawnwood: United States, average export price of Douglas Fir, U.S. Price, US\$ per cubic meter.

Cotton: United States, Cotton Outlook 'A Index', Middling 1-3/32 inch staple, CIF Liverpool, US cents per pound.

Wool coarse: United Kingdom, 23 micron, Australian Wool Exchange spot quote, US cents per kilogram.

Wool fine: United Kingdom, 19 micron, Australian Wool Exchange spot quote, US cents per kilogram.

Rubber: Malaysia, No.1 Smoked Sheet, Singapore Commodity Exchange, 1st contract, US cents per pound.

Hides: United States, Heavy native steers, over 53 pounds, wholesale dealer's price, Chicago, fob Shipping Point, US cents per pound.

Copper: United Kingdom, grade A cathode, LME spot price, CIF European ports, US\$ per metric tonne.

Aluminum: Canada, 99.5% minimum purity, LME spot price, US\$ per metric tonne.

Iron Ore: China import 62% FE spot (CFR Tianjin port), US cents per dry metric tonne unit.

Tin: United Kingdom, standard grade, LME spot price, US\$ per metric tonne.

Nickel: United Kingdom, melting grade, LME spot price, CIF European ports, US\$ per metric tonne.

Zinc: United Kingdom, high grade 98% pure, US\$ per metric tonne.

Lead: United Kingdom, 99.97% pure, LME spot price, CIF European Ports, US\$ per metric tonne.

Uranium: World, NUEXCO, Restricted Price, Nuexco exchange spot, US\$ per pound.

Crude Oil: Arab Emirates, Dubai, medium, Fateh 32 API, fob DubaiCrude Oil (petroleum), Dubai Fateh Fateh 32 API, US\$ per barrel.

E Tables and Figures

BMI Estimation – Model Selection

<i>No.of Clusters</i>	$\ln f(\mathbf{Y} \Theta^*)$	$\ln p(\Theta^*)$	$\ln \hat{p}(\Theta^* \mathbf{Y})$	$\ln \hat{m}(\mathbf{Y})$
2	-6123	-1036	308.6	-7468
3	-5954	-1048	278.5	-7287
4	-5797	-1051	269.5	-7118
5	-5802	-1056	278.0	-7137
6	-7689	-1068	222.5	-8979

Table 1: Notes: First column shows the likelihood of the model, Second column refers to the prior value at the modal points. Third column represents the sum of the posterior ordinates calculated from the reduced runs described in the appendix. And finally last column shows the model marginal likelihood.

Inclusion Probabilities Across Clusters

	Cluster-1	Cluster-2	Cluster-3	Cluster-4
$p(\gamma) \geq 0.9$	Hard Logs Hard Sawnwood Soft Logs Soft Sawnwood	Barley Corn Soybean Oil Wheat Soybean Meal Soybeans Palm oil Canola oil	Aluminum Copper Fishmeal Hides Lead Nickel Olive Oil Rubber	Tin Wool,Coarse Wool,Fine Zinc Salmon Sugar,US Uranium Sugar,World
$0.8 \leq p(\gamma) < 0.9$		Cotton Groundnuts Poultry Sunflower Oil	Cocoa	
$0.5 \leq p(\gamma) < 0.7$	Lamb			Iron
$0.3 \leq p(\gamma) < 0.5$			Beef Oranges Tea Bananas Rice	Shrimp Swine

Table 2: Notes: The table summarizes the posterior inclusion probabilities for each cluster and lists the members. For each commodity highest probability of belonging to one cluster is reported.

$$Variance\ Decompositions - p(\gamma_{i,j} = 1) > p(\gamma_{i,k} = 1) \forall k$$

Factor	Cluster 1	Cluster 2	Cluster3	Cluster4	Sample Average
Global	1.04	18.69	3.64	0.86	7.2
Cluster	37.74	31.29	20.09	35.36	26.9
Global +Cluster	38.78	49.98	36.22	23.74	34.1
Idiosyncratic	61.2	49.98	76.24	63.8	65.9

Table 3: Notes: The table summarizes the variance decomposition in percentages where the clusters are estimated with the endogenous clustering algorithm. The clusters are constructed with the posterior values of the indicator function. An observation is assumed to belong to one cluster if the said observation picked that cluster more of the time over the Gibbs run than the other clusters.

$$Variance\ Decompositions\ for\ Grains\ \&\ Oils\ \&\ Meat$$

Product Name	Global	Group	Idiosyncratic
Corn (2)	8.8	46	45.2
Soybeans (2)	2.3	92.8	4.9
Soybean meal (2)	5.3	90.3	4.3
Soybean oil (2)	49.3	39.1	11.6
palm oil (2)	49.4	14	36.6
canola oil (2)	52.6	11.9	35.5
Wheat (2)	4.2	21.9	73.9
Rice (3)	0	7.7	92.2
Meat (3)	6.5	1.7	91.8

Table 4: Notes: The table summarizes the variance decomposition in percentages. Members are allocated to clusters that they picked the most over the Gibbs run. The paranthesis indicates each commodity's selected cluster. Bold values represents the highest variance decomposition for each product.

Variance Decompositions for All Commodities

Product Name	Global	Group	Idio.	Product Name	Global	Group	Idio.
Hard Logs (1)	1.2	72.6	26.2	Copper (3)	5.9	60.7	33.3
Hard Sawnwood (1)	0	68.8	32.2	Fish-Salmon (3)	2.1	11.2	86.7
Lamb (1)	0.4	10.1	89.5	Fishmeal (3)	1.1	18.8	80.1
Soft Logs (1)	2.9	25.4	71.6	Hides (3)	0.8	16.1	83.3
Soft Sawnwood (1)	0.7	11.8	87.5	Lead (3)	1.9	22.1	76.1
Barley (2)	14.8	31.1	54.1	Nickel (3)	3.4	40.9	55.7
Corn (2)	8.8	46	45.2	Olive Oil (3)	0	16.6	83.3
Soybeans (2)	2.3	92.8	4.9	Oranges (3)	0.3	1.4	98.3
Soybean meal (2)	5.3	90.3	4.3	Rubber (3)	9.5	45	45.5
Soybean oil (2)	49.3	39.1	11.6	Sugar, Free Market (3)	0.7	7.8	91.5
palm oil (2)	49.4	14	36.6	Sugar, US (3)	1.2	2.9	95.8
canola oil (2)	52.6	11.9	35.5	Tea (3)	0.8	2.9	96.2
Wheat (2)	4.2	21.9	73.9	Tin (3)	13.5	15.5	71
Cotton (2)	6.5	14.1	79.3	Uranium (3)	1.9	12.6	85.6
Groundnuts (2)	1.6	5.5	92.8	Wool, coarse (3)	7,2	23.9	68.8
Poultry (2)	0.1	4	95.9	Wool, fine (3)	9.5	23.5	67
Sunflower Oil (2)	29.4	5.6	64.9	Zinc (3)	2.6	41.6	55.8
Rice (3)	0	7.7	92.2	Coffee, Robusta (4)	0.7	80.9	13.4
Meat (3)	6.5	1.7	91.8	Coffee, Other (4)	0.2	85.5	14.3
Aluminium (3)	6	55.8	38.2	Iron (4)	2.4	4	93.6
Cocoa beans (3)	4.9	12.6	82.5	Shrimp (4)	0.7	1.8	97.5
Bananas (3)	0.4	0.7	98.9	Swine (4)	0.3	4.6	95.2

Table 5: Notes: The table summarizes the variance decomposition in percentages. The members are allocated to clusters given the modal value of the cluster probability. The paranthesis indicates each commodity's selected cluster.

Quarterly Regression Results

<i>Variable Name</i>	Global	Cluster-1	Cluster-2	Cluster-3	Cluster-4
Federal Funds Rate	-0.02 (-0.06 0.02)	-0.01 (-0.05 0.03)	-0.02 (-0.05 0.02)	-0.05* (-0.08 -0.02)	-0.02 (-0.05 0.02)
World IP	0.22* (0.04 0.39)	-0.16 (-0.35 0.03)	0.05 (-0.13 0.23)	0.34* (0.18 0.51)	0.18 (-0.02 0.38)
Dow	-0.02 (-0.04 0.01)	-0.002 (-0.03 0.03)	0.01 (-0.02 0.04)	0.04* (0.01 0.06)	-0.02 (-0.05 0.01)
Oil Price	0.01 (-0.00 0.02)	0.01 (-0.00 0.02)	0.01* (0.008 0.03)	0.02* (0.01 0.03)	-0.002 (-0.02 0.01)
Fertilizer Prices	0.01 (-0.02 0.00)	0.01 (-0.01 0.02)	0.005 (-0.01 0.02)	-0.01 (-0.02 0.002)	0.002 (-0.01 0.02)
US House Price	-0.06 (-0.22 0.09)	0.06 (-0.10 0.24)	-0.001 (-0.16 0.16)	0.02 (-0.11 0.16)	-0.16 (-0.34 0.02)
Exchange Rate	0.001 (-0.01 0.01)	-0.001 (-0.01 0.01)	0.000 (-0.01 0.01)	-0.001 (-0.01 0.00)	0.002 (-0.00 0.01)
China Trade	0.000 (-0.02 0.02)	0.04* (0.02 0.07)	-0.004 (-0.02 0.02)	-0.001 (-0.02 0.02)	0.02 (-0.01 0.04)
Climate Anomaly	0.35 (-0.28 0.98)	0.19 (-0.49 0.87)	0.25 (-0.39 0.89)	-0.12 (-0.67 0.42)	0.01 (-0.64 0.66)

* denotes statistical significance

Table 6: Notes: Each column represents a separate Bayesian Regression on the variables listed in rows. Constant is excluded as in estimation. China Trade is measured as the volume of exports and imports. Variables except FFR and Exchange Rate are all percentage growth rates. FFR and Exchange rate are in deflated levels. Credible Intervals that are measured by the 5th and 95th percentiles are shown below of each coefficient.

Annual Regression Results

<i>Variable Name</i>	Global	Cluster-1	Cluster-2	Cluster-3	Cluster-4
Federal Funds Rate	-0.001 (-0.06 0.05)	-0.037 (-0.02 0.09)	-0.009 (-0.05 0.036)	-0.06* (-0.11 -0.01)	-0.04 (0.1 -0.01)
World IP	0.14* (0.25 0.3)	-0.098 (-0.02 0.22)	-0.074 (-0.16 0.02)	0.037 (-0.06 0.136)	0.13* (0.02 0.24)
Dow	-0.02 (-0.04 -0.002)	0.02* (0.001 0.04)	0.002 (-0.01 0.02)	0.01 (-0.01 0.03)	-0.01 (-0.03 0.01)
Oil Price	-0.001 (-0.01 0.01)	0.01 (-0.001 0.02)	0.002 (-0.01 0.01)	0.0073 (-0.00 0.02)	-0.0085 (-0.02 0.001)
Fertilizer Prices	0.003 (-0.01 0.01)	-0.001 (-0.01 0.01)	0.01* (0.00 0.01)	-0.003 (-0.01 0.00)	-0.001 (-0.01 0.01)
US House Price	-0.017 (-0.09 0.06)	0.044 (-0.04 0.13)	0.028 (-0.03 0.09)	0.059 (-0.01 0.12)	-0.008 (-0.08 0.07)
Exchange Rate	-0.004 (-0.01 0.01)	-0.011 (-0.02 0.001)	-0.004 (-0.01 0.01)	-0.01* (-0.02 -0.002)	-0.01 (-0.02 0.00)
China GDP	0.0154 (-0.02 0.05)	0.04* (0.003 0.07)	0.026 (-0.00 0.05)	0.024 (-0.01 0.05)	0.03 (-0.06 0.003)
Bio Fuel	-0.001 (-0.02 0.02)	0.003 (-0.02 0.02)	-0.009 (-0.02 0.004)	0.01 (-0.00 0.03)	0.016 (-0.001 0.03)
Climate Anomaly	-0.21 (-0.97 1.4)	0.75 (-0.5 1.96)	-0.48 (-1.52 0.57)	-0.19 (-1.3 0.92)	-0.43 (-1.65 0.78)

* denotes statistical significance

Table 7: Notes: Each column represents a separate Bayesian Regression on the variables listed in rows. Constant is excluded as in estimation. Variables except FFR and Exchange Rate are all percentage growth rates. FFR and Exchange rate are in deflated levels. Credible Intervals that are measured by the 5th and 95th percentiles are shown below of each coefficient.

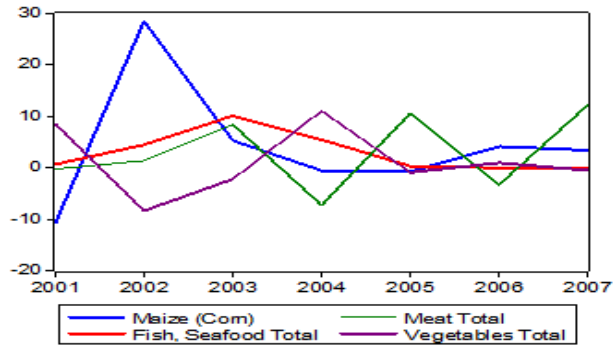


Figure 1: Notes: Total supply growth in metric tones of Maize, Meat (total), Seafood (total) and Vegetables (total) for Australia during the draught period. Annual data is gathered from Food and Agriculture Organization of the United Nations Statistics (FAOSTAT).

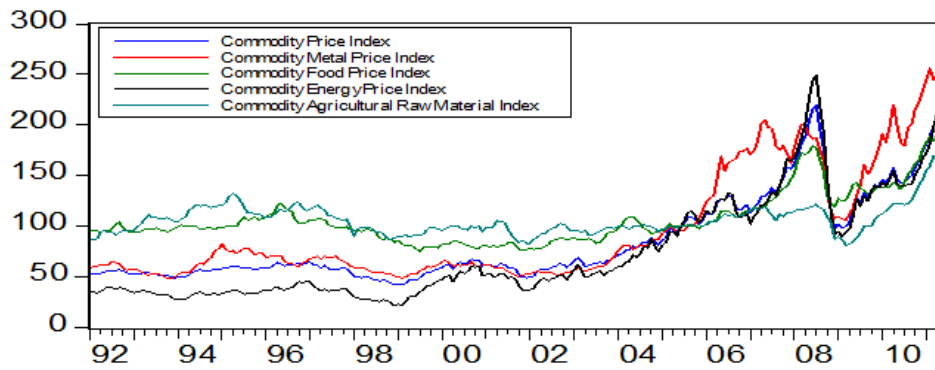


Figure 2: Notes: Nominal Commodity Price Index and Nominal Commodity Price indices for major subgroups, metals, food, energy and materials. The quarterly series extracted from IFS database.

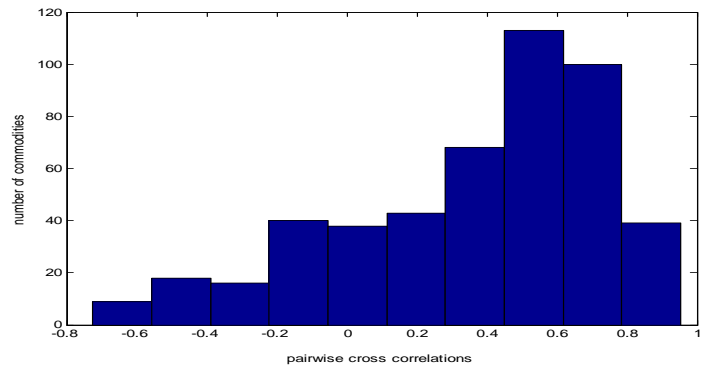


Figure 3: Notes: Histogram lists all the pairwise cross correlations across 42 non-energy nominal commodity prices.

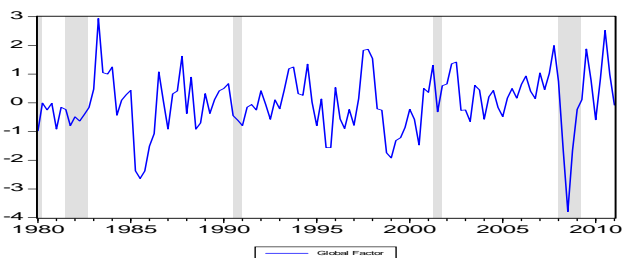


Figure 4: Notes: Estimated Median Global Factor. Shaded areas represent the NBER recessions.

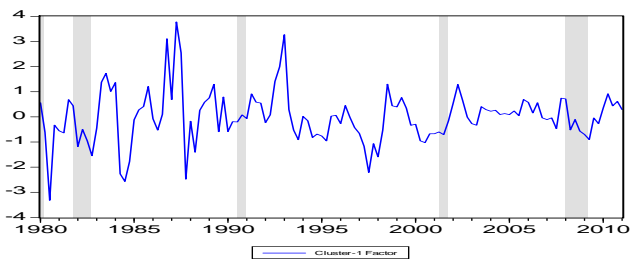


Figure 5: Notes: Estimated Median Cluster-1 ("Timber"). Shaded areas represent the NBER recessions. This cluster is dominantly made of logs and wood, lamb meat weakly belong to this cluster.

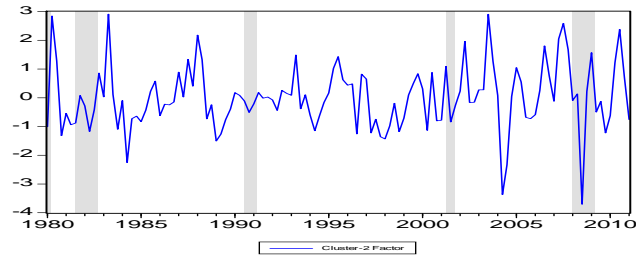


Figure 6: Notes: Estimated Median Cluster-2 ("Grains Oil"). Shaded areas represent the NBER recessions. This cluster is dominantly made of grains (except rice) and vegetable oils (except olive oil). Some other food products and cotton weakly belong to this cluster.

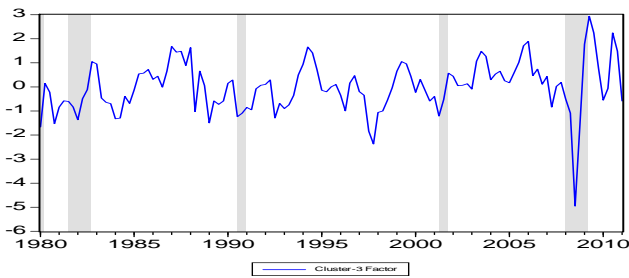


Figure 7: Notes: Estimated Median Cluster-3 ("Mixed"). Shaded areas represent the NBER recessions. This cluster consists of metals (except iron), agricultural raw materials (except cotton and timber) and some food products.

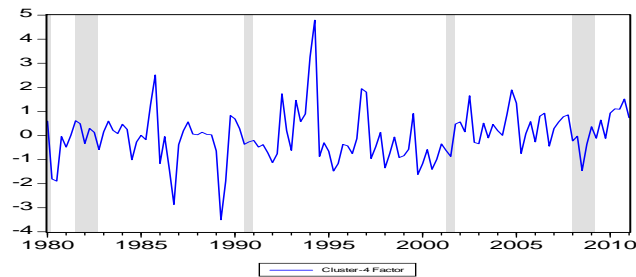


Figure 8: Notes: Estimated Median Cluster-4 ("Coffee"). Shaded areas represent the NBER recessions. This cluster is dominantly made of coffee. Iron, shrimp and swine also weakly belong to this cluster with the degrees in the written order.