Central European Journal of Economic Modelling and Econometrics

Missing Observations in Daily Returns – Bayesian Inference within the MSF-SBEKK Model

Krzysztof Osiewalski*, Jacek Osiewalski†

Submitted: 26.10.2012, Accepted: 5.03.2013

Abstract

Often daily prices on different markets are not all observable. The question is whether we should exclude from modelling the days with prices not available on all markets (thus loosing some information and implicitly modifying the time axis) or somehow complete the missing (non-existing) prices. In order to compare the effects of each of two ways of dealing with partly available data, one should consider formal procedures of replacing the unavailable prices by their appropriate predictions. We propose a fully Bayesian approach, which amounts to obtaining the marginal posterior (or predictive) distribution for any particular day in question. This procedure takes into account uncertainty on missing prices and can be used to check validity of informal ways of "completing" the data (e.g. linear interpolation). We use the MSF-SBEKK structure, the simplest among hybrid MSV-MGARCH models, which can parsimoniously describe volatility of a large number of prices or indices. In order to conduct Bayesian inference, the conditional posterior distributions for all unknown quantities are derived and the Gibbs sampler (with Metropolis-Hastings steps) is designed. Our approach is applied to daily prices from six different financial and commodity markets; the data cover the period from December 21, 2005 till September 30, 2011, so the time of the global financial crisis is included. We compare inferences (on individual parameters, conditional correlation coefficients and volatilities), obtained in the cases where incomplete observations are either deleted or forecasted.

Keywords: Bayesian econometrics, hybrid MGARCH-MSV processes, forecasting unavailable data, financial markets, commodity markets

JEL Classification: C11, C32, C51, C58

*e-mail: krzysztof@osiewalski.eu
†Cracow University of Economics; e-mail: eeosiewa@cyf-kr.edu.pl

1 Introduction

Conceptually, the idea of analyzing the broadest possible data set appears sound. While modelling a multivariate portfolio on the basis of daily data, it may happen that not all assets are valued over a particular calendar day or some data are missing. Such a situation may result from several reasons: different national or religious holidays and different bank holiday schedule in different countries, lack of publicly available data, database incompleteness or just data errors. The most common way of handling such situations is deleting days with unavailable prices. This leads to loss of information and a smaller set of data. On top of that, financial time series are usually modelled by using autoregressive structures: both in conditional mean and variance. If we delete particular transactional days we modify the time axis in an uncontrolled manner, which may result in misleading conclusions in estimation as well as in forecasting.
The modern literature provides little information on this problem. For univariate basic autoregressive processes the problem was described by Tsay R. (2005), together with a straightforward formula for the posterior distribution of one missing value in the series. Whenever missing values occur in patches, Gibbs sampling is recommended. Kim J. (2005), Kim J. and Stroffer D. (2008) focused on the problem of unavailability of certain data points within univariate SV models. The methods they introduce for handling unavailable observations are based on particle filters and an expectation-maximization algorithm. However, most of the literature on multivariate time series models either implicitly assumes full data availability or removes the days with partial information only. The paper by Doman M. and Doman R. (2010, in Polish) is an exception, as the authors explicitly consider the problem of partially unavailable multivariate daily data. They simulate such data from diagonal BEKK(1,1) processes, and then they estimate diagonal BEKK(1,1) models using different approaches to partially available information. Doman M. and Doman R. (2010) are interested mainly in conditional correlations and they conclude that linear approximation of unavailable prices is a good strategy. Our research goal is somewhat different.
In this paper, we aim at empirical verification (on the basis of real data) of possible gains resulting from treating missing observations as unknown quantities of interest. We follow the formal Bayesian approach and construct the posterior distributions of unavailable data. Methodologically, the research is set within the hybrid models based on multivariate stochastic volatility and multivariate generalized conditional heteroscedasticity structures (MSV-MGARCH). Such hybrid models were proposed by Osiewalski J. (2009), Osiewalski J. and Pajor A. (2007, 2009), and further developed and expanded by Osiewalski J. and Pajor A. (2010), Pajor A. and Osiewalski J. (2012), Osiewalski J. and Osiewalski K. (2011a,b). Such a choice is strongly motivated by both good data explanatory abilities and relatively low computational burden of posterior sampling in these kinds of models. Hybrid models formally belong to the MSV class. However, we distinguish them from *pure* MSV models, in which the conditional covariance matrix does not contain a GARCH structure.

The above mentioned studies revealed that even the simplest hybrid structures have enough flexibility to describe data well and yet they are parsimonious enough to handle large portfolios and perform risk analysis successfully (using Value at Risk and Expected Shortfall). Thus, hybrid MSV-MGARCH models can be considered as a trade-off between model fit and computational burden of estimation. In this paper we use the simplest model, i.e. the MSF-SBEKK structure proposed by Osiewalski J. (2009) and Osiewalski J. and Pajor A. (2009), because the formal Bayesian treatment of unavailable data is computationally too demanding to consider more complex model specifications.

This paper contributes to the relevant literature in two ways. On the methodological side, we treat unknown values of certain assets as latent variables and provide the details of Bayesian inference about them, including the sampling scheme for drawing from the posterior distribution. On the empirical side, we show how the modification of the data set by removing days with missing observations influences posterior distributions of some model parameters and characteristics of returns volatility and correlation.

The paper is organized as follows. In the following section the Bayesian MSF-SBEKK model with missing data is introduced and the posterior distribution is presented. In Section 3 the generation of a pseudo-random sample from the posterior distribution is covered. Empirical study, presented in Section 4, is divided into two subsections. Firstly, we investigate the missing data issue for a chosen multivariate portfolio. Secondly, we present posterior characteristics of parameters and their selected functions together with the posterior distributions of missing data points. Finally, concluding remarks are presented in Section 5.

2 Bayesian MSF–SBEKK model with missing data

Let us denote by $0.01 \cdot x^*_{t,i}$ the natural logarithm of the unobserved price and by $0.01 \cdot x_{t,i}$ the natural logarithm of the observed price of asset i at time t. We consider the following multivariate specification for n individual assets. Let $r_t = (r_{t,1} \dots r_{t,n})'$ denote n-variate logarithmic return rates (in percentage points), i.e. $r_{t,i} = \widehat{x}_{t,i} - \widehat{x}_{t-1,i}$, where

$$\widehat{x}_{t,i} = \begin{cases} x_{t,i}, & \text{if price of asset } i \text{ is available at time } t, \\ x^*_{t,i}, & \text{otherwise.} \end{cases} \tag{1}$$

We model r_t using the basic VAR(1) framework:

$$r_t = \lambda + \Lambda r_{t-1} + \varepsilon_t, \qquad t = 1, \dots, T, \tag{2}$$

where T denotes the number of days when at least one asset is valued (and its price is observed). The error terms, ε_t $(t = 1, \dots, T)$, are assumed to follow the hybrid

MSF-SBEKK structure, as in Osiewalski J. (2009) and Osiewalski J. and Pajor A. (2009, 2010):

$$\varepsilon_t = G_t^{\frac{1}{2}} H_t^{\frac{1}{2}} \xi_t \quad t = 1, \ldots, T. \tag{3}$$

In (3) G_t represents the MSV (here MSF) component involving a latent AR(1) process:

$$G_t = g_t I_n, \quad \ln g_t = \phi \ln g_{t-1} + \zeta_t \tag{4}$$

with scalar $g_t>0$ and $|\phi|<1$, H_t represents the MGARCH – here SBEKK, i.e. scalar BEKK(1,1) – component (of the conditional covariance matrix):

$$H_t = (1 - \beta - \gamma)A + \beta \varepsilon_{t-1} \varepsilon_{t-1}' + \gamma H_{t-1} \tag{5}$$

with symmetric positive definite $n \times n$ matrix A, scalar parameters $\beta, \gamma > 0$ $(\beta + \gamma < 1)$, and

$$[\xi_t' \, \zeta_t]' \sim iiN^{(n+1)} \left(0_{[(n+1)\times 1]}, \begin{bmatrix} I_n & 0 \\ 0 & \tau^{-1} \end{bmatrix} \right). \tag{6}$$

The initial condition for H_t in (5) is taken as $H_0 = h_0 I_n$ with a scalar parameter $h_0 > 0$ and we assume $g_0 = 1$ to initialize (4). Note that under (2) – (6) the conditional distribution of $\widehat{x}_t = (\widehat{x}_{t,1} \ldots \widehat{x}_{t,n})'$ given g_t and the past of both processes, ψ_{t-1}, is n-variate Normal with mean $\widehat{x}_{t-1} + \lambda + \Lambda r_{t-1}$ and covariance matrix $g_t H_t$. In this specification the conditional variances (given ψ_{t-1}) are equal to $g_t h_{t,ii}$, thus they extend both the MSF and SBEKK cases. The conditional correlation coefficient (given ψ_{t-1}) does not depend on g_t and takes the SBEKK form. However, posterior inference on the conditional correlation coefficient is obviously influenced by the presence of the latent process and thus the final results on conditional correlation may not be the same as in the pure SBEKK model.

In order to separate available data from latent variables representing missing values, let us rewrite (2) in terms of hundreds of logarithms of prices:

$$\widehat{x}_t = \lambda + \widehat{x}_{t-1} + \Lambda(\widehat{x}_{t-1} - \widehat{x}_{t-2}) + \varepsilon_t. \tag{7}$$

In the Bayesian approach, all unknown quantities are treated as random variables. In the presence of latent variables (g_t and unavailable prices), we are usually interested in making inference on both the parameter vector, i.e. $\theta = (\lambda' \, (\mathrm{vec}\Lambda)' \, (\mathrm{vech}A)' \, \beta \, \gamma \, h_0 \, \phi \, \tau)'$ and the latent variable vectors: $g = (g_1 \ldots g_T)'$ and x^*, where x^* groups logarithms of all missing price values. The joint density of observations, $x = (x_1' \ldots x_T')'$, latent variables and parameters can be factorized as follows:

$$\begin{aligned} p(x, x^*, g, \theta) &= p(\theta) p(x, x^*, g|\theta) \\ &= p(\theta) \prod_{t=1}^{T} p(\widehat{x}_t|\psi_{t-1}, g_t, \theta) p(g_t|\psi_{t-1}, \theta). \end{aligned} \tag{8}$$

It is worth stressing here that the MSF–SBEKK structure is based on two basic conditional independence assumptions, which hold for any value of θ. Firstly, $\widehat{x}_t$ is independent of the past of g_t, given g_t itself and the past of $\widehat{x}_t$. Secondly, g_t is independent of the past of $\widehat{x}_t$, given the past of g_t. Thus in (8) we have

$$p(g_t|\psi_{t-1},\theta) = p(g_t|g_{t-1},\theta) = g_t^{-1} f_N\left(\ln g_t|\phi \ln g_{t-1}, \tau^{-1}\right), \tag{9}$$

which is a univariate log-normal density, and, for $\mu_t = \lambda + \widehat{x}_{t-1} + \Lambda(\widehat{x}_{t-1} - \widehat{x}_{t-2})$,

$$p(\widehat{x}_t|\psi_{t-1}, g_t, \theta) = f_N^n\left(\widehat{x}_t|\mu_t, g_t H_t\right), \tag{10}$$

which is a multivariate Normal density function. For the sake of simplicity, we assume that the initial conditions related to $\widehat{x}_t$ are known and constant, and we omit them in our notation. The joint distribution of observed and unobserved logarithms of prices need not be factorized further as we can jointly handle the existing and missing values. For sampling from the posterior distribution, the likelihood function (which takes a quite complicated form in this case) is not required.
In order to complete the Bayesian model, let us specify the prior structure of the parameter vector θ. We will subjectively set the distributions of interest, which will reflect our weak knowledge about model parameters (see Osiewalski J. and Pajor A. (2009)). Let us assume that:

$$p(\theta) = p(\lambda)p(\mathrm{vec}\Lambda)p(A^{-1})p(\beta,\gamma)p(h_0)p(\phi)p(\tau), \tag{11}$$

which means prior independence among blocks of parameters. Furthermore we take:

- $p(\lambda) = f_N^n(\lambda|0, I_n)$ – the n-variate Normal density with mean 0 and covariance matrix I_n,
- $p(\mathrm{vec}\Lambda) \propto f_N^{n^2}(\mathrm{vec}\Lambda|0, I_{n^2})\mathbf{1}_{\{M:\ \rho(M)<1\}}(\Lambda)$, – a multivariate Normal truncated by the restriction that all eigenvalues of Λ lie inside the unit circle, where $\rho(M)$ is the spectral radius of matrix M and $\mathbf{1}_A(x)$ is the indicator function of the set A: $\mathbf{1}_A(x) = \begin{cases} 1, & x \in A \\ 0, & x \notin A \end{cases}$,
- $p(A^{-1}) = f_{Wishart}(A^{-1}|I_n, n+2)$ – the Wishart distribution with mean I_n and $n+2$ degrees of freedom,
- $p(\beta,\gamma) \propto \mathbf{1}_{\{(x,y)\in[0,1]^2:\ x+y<1\}}(\beta,\gamma)$ – a uniform distribution over a unit simplex,
- $p(h_0) = f_{Exp}(h_0|1)$ – the Exponential distribution with mean 1,
- $p(\phi) \propto f_N(\phi|0,100)\mathbf{1}_{\{x:\ |x|<1\}}(\phi)$,
- $p(\tau) = f_{Exp}(\tau|200)$ – the Exponential distribution with mean 200.

Finally we can write the joint density function:

$$\begin{aligned} p(x, x^*, g, \theta) &= p(\lambda)p(\mathrm{vec}\Lambda)p(A^{-1})p(\beta, \gamma)p(h_0)p(\phi)p(\tau)\cdot \\ &\cdot \prod_{t=1}^{T} g_t^{-1} f_N\left(\ln g_t | \phi \ln g_{t-1}, \tau^{-1}\right) f_N^n\left(\widehat{x}_t | \mu_t, g_t H_t\right) \end{aligned} \tag{12}$$

that represents the Bayesian MSF-SBEKK model with missing (non-existing) data. The posterior distribution of all unobservable quantities (i.e. missing data, latent variables and parameters) is characterised by the conditional density function $p(x^*, g, \theta|x)$, which is proportional to $p(x, x^*, g, \theta)$ in (12).

3 Sampling from the posterior distribution

The joint posterior distribution, represented by the density $p(x^*, g, \theta|x)$, is highly dimensional and too complicated to obtain any analytical results. In this case, numerical methods must be applied in order to generate a (pseudo) random sample from the posterior distribution and to obtain estimates of posterior characteristics. Following Osiewalski J. (2009) and Osiewalski J. and Pajor A. (2009), we use a hybrid Markov Chain Monte Carlo method: the Gibbs sampler with Metropolis and Hastings steps. The algorithm is based on the conditional posterior distributions resulting from (12) with a natural block partition of all unknown quantities in the model. The conditional posterior distributions used to construct the sampler are described below, together with the resulting sampling scheme.

i) The VAR(1) parameters λ and Λ have the following conditional densities:

$$p(\lambda|x, x^*, g, \Lambda, A, \beta, \gamma, h_0, \phi, \tau) \propto p(\lambda) \prod_{t=1}^{T} f_N^n\left(\widehat{x}_t | \mu_t, g_t H_t\right), \tag{13}$$

$$p(\mathrm{vec}\Lambda|x, x^*, g, \lambda, A, \beta, \gamma, h_0, \phi, \tau) \propto p(\mathrm{vec}\Lambda) \prod_{t=1}^{T} f_N^n\left(\widehat{x}_t | \mu_t, g_t H_t\right). \tag{14}$$

We cannot directly sample from these conditional posterior distributions (as in pure MSV models) due to the presence of the MGARCH (SBEKK) structure, which implies that the VAR(1) parameters from conditional means have an impact on conditional variances at every time point. Thus, the Metropolis and Hastings step is implemented. The choice of the proposal distribution is quite arbitrary – it is set to be a Normal distribution centered at the previous state of the Markov chain (Random Walk MH). The covariance matrix of the proposal distribution is set to the sample covariance matrix (multiplied by a factor of 4) obtained from initial cycles, which are performed to calibrate the sampling mechanism. The resulting acceptance rate oscillated between 3 and 7 percent in the empirical example presented in the next section.

ii) For the MGARCH (SBEKK) parameters A, β, γ and h_0 we have:

$$p(A|x, x^*, g, \lambda, \Lambda, \beta, \gamma, h_0, \phi, \tau) \propto p(A) \prod_{t=1}^{T} f_N^n \left(\widehat{x}_t|\mu_t, g_t H_t\right), \tag{15}$$

$$p(\beta, \gamma, h_0|x, x^*, g, \lambda, \Lambda, A, \phi, \tau) \propto p(\beta, \gamma) p(h_0) \prod_{t=1}^{T} f_N^n \left(\widehat{x}_t|\mu_t, g_t H_t\right). \tag{16}$$

Here again MH steps enable sampling from the conditional distributions. For the matrix A^{-1}, we sample candidate states from a Wishart distribution. The proposal scale matrix is set to $\frac{1}{k}$ times the previously drawn A^{-1} (so that the proposal expectation equals the previously drawn A^{-1}) and degrees of freedom k are set to 200. The value of k was chosen after a number of initial cycles and resulted in acceptance rate around 5%. The candidate draws for parameters (β, γ) are generated from a bivariate Normal distribution truncated by the restrictions $\beta > 0$, $\gamma > 0$ and $\beta + \gamma < 1$. As in the case of the VAR(1) parameters, the Normal proposal density parameters are: the previous Markov chain state for the mean and the sample covariance from previous chains multiplied by a factor of 2. For the initial condition, h_0, we draw candidates from a Normal distribution truncated to $\mathbb{R}_+$ (with parameters chosen in the same manner as for β and γ). The acceptance rate was about 3-4%.

iii) For the parameters ϕ and τ, describing the latent process g_t, we can use the pure Gibbs step as in the MSF model (see Pajor A. (2010)), because their conditional posteriors are standard distributions:

$$\begin{aligned} p(\phi|x, x^*, g, \lambda, \Lambda, A, \beta, \gamma, h_0, \tau) &\propto p(\phi) \prod_{t=1}^{T} f_N \left(\ln g_t|\phi \ln g_{t-1}, \tau^{-1}\right) \\ &\propto f_N \left(\phi|\phi^*, {s^2}^*\right) \mathbf{1}_{(-1,1)}(\phi), \end{aligned} \tag{17}$$

$$\begin{aligned} p(\tau|x, x^*, g, \lambda, \Lambda, A, \beta, \gamma, h_0, \phi) &\propto p(\tau) \prod_{t=1}^{T} f_N \left(\ln g_t|\phi \ln g_{t-1}, \tau^{-1}\right) \\ &\propto f_G \left(\tau|\tfrac{T}{2} + 1, \beta^*\right), \end{aligned} \tag{18}$$

where

$${s^2}^* = \left[0.01 + \tau \sum_{i=1}^{T} (\ln g_{t-1})^2\right]^{-1}, \tag{19}$$

$$\phi^* = {s^2}^* \tau \sum_{i=1}^{T} \ln g_t \ln g_{t-1}, \tag{20}$$

$$\beta^* = \left[0.005 + \frac{1}{2} \sum_{i=1}^{T} (\ln g_t - \phi \ln g_{t-1})^2\right]^{-1}, \tag{21}$$

and $f_G(\cdot|a,b)$ denotes the density of the Gamma distribution with mean $\frac{a}{b}$ and variance $\frac{a}{b^2}$.

iv) The latent variables g_t can be drawn in the following manner, similarly as in Pajor (2010), i.e. for $t = 1, \ldots, T-1$ we would like to sample from:

$$\begin{aligned} &p(g_t^{-1}|x, x^*, g_1, \ldots, g_{t-1}, g_{t+1}, \ldots, g_T, \lambda, \Lambda, A, \beta, \gamma, h_0, \phi, \tau) \\ &\quad \propto f_N\left(\ln g_t|\phi \ln g_{t-1}, \tau^{-1}\right) f_N\left(\ln g_{t+1}|\phi \ln g_t, \tau^{-1}\right) f_N^n\left(\widehat{x}_t|\mu_t, g_t H_t\right) \\ &\quad \propto f_N\left(\ln g_t|\phi \ln g_{t-1}, \tau^{-1}\right) f_N\left(\ln g_{t+1}|\phi \ln g_t, \tau^{-1}\right) \cdot \\ &\qquad \cdot f_G\left(g_t^{-1}|\tfrac{n}{2}, \tfrac{(x_t-\mu_t)' H_t^{-1}(x_t-\mu_t)}{2}\right). \end{aligned} \tag{22}$$

In the case of g_t^{-1}, the Metropolis and Hastings step with a gamma distribution of candidate draws is used (following Jacquier E., Polson N. and Rossi P. (1994)):

$$p_c(g_t^{-1}|\cdot) = f_G\left(g_t^{-1}|\varphi_t, \eta_t\right), \tag{23}$$

where

$$\varphi_t = \frac{1 - 2\exp\left(\sigma^2\right)}{1 - \exp\left(\sigma^2\right)} + \frac{n}{2}, \tag{24}$$

$$\sigma^2 = \begin{cases} \tau^{-1}\left(1+\phi^2\right)^{-1}, & 1 \le t \le T-1 \\ \tau^{-1}, & t = T \end{cases}, \tag{25}$$

$$\eta_t = \left(\varphi_t - \frac{n}{2} - 1\right) \exp\left(s_t + \frac{\sigma^2}{2}\right) + \frac{1}{2}\left(x_t - \mu_t\right)' H_t^{-1}(x_t - \mu_t), \tag{26}$$

$$s_t = \begin{cases} \frac{\phi}{1+\phi^2}\left(\ln g_{t-1} + \ln g_{t+1}\right), & 1 \le t \le T-1 \\ \phi \ln g_{T-1}, & t = T \end{cases}. \tag{27}$$

The outlined MH sampling scheme is very efficient – the acceptance ratio did not fall below 90%.

v) The unobserved price values are also sampled using the Metropolis and Hastings steps, specially designed here to deal with this case. Assume that the first m, $m \in \{1, \ldots, n-1\}$, elements of the vector $\widehat{x}_t$ are not available. If the missing values are spread irregularly, we can always rearrange the assets. Then:

$$p(x_t^*|x, x_{(-t)}^*, g, \lambda, \Lambda, A, \beta, \gamma, h_0, \phi, \tau) \propto \prod_{j=t}^{T} f_N^n\left(\widehat{x}_j|\mu_j, g_j H_j\right), \tag{28}$$

where $x_{(-t)}^*$ denotes the missing values vector without unobserved prices from time t. In the presence of the autoregressive structure, each of the missing values

will have an impact on the whole future of the process. However, for simplicity we base the distribution of candidate draws on time t information only. Thus, as the proposal density we use:

$$p_c(x_t^*|x_1,\ldots,x_t,x_1^*,\ldots,x_{t-1}^*,g_1,\ldots,g_t,\theta) = f_N^m\left(x_t^*|\mu^*,\Omega^*\right), \tag{29}$$

where

$$\mu^* = \mu_{t,1:m} + (g_tH_t)_{(m+1):n,1:m}(g_tH_t)^{-1}_{(m+1):n,(m+1):n}\left(x_{t,(m+1):n} - \mu_{t,(m+1):n}\right), \tag{30}$$

$$\Omega^* = (g_tH_t)_{1:m,1:m} - (g_tH_t)_{(m+1):n,1:m}(g_tH_t)^{-1}_{(m+1):n,(m+1):n}(g_tH_t)_{1:m,(m+1):n} \tag{31}$$

and $M_{a:b,c:d}$ denotes the block composed of matrix M rows from a to b and matrix M columns from c to d. This proposal resulted in MH acceptance ratio between 20 and 70% (depending on the time index t), which is very satisfactory.

For the algorithm constructed in the manner above, the convergence is monitored via standardised CUMSUM plots; see Yu B. and Mykland P. (1998) and Pajor A. (2003). The chain length was set to 1 million states, in which the first 0.6 million were considered as *burn-in* period. The sampling speed was approximately 19.1 seconds per 1000 chain states on a regular desktop CPU (Intel Core2Duo E8600). The most time consuming steps are, however, the one related to drawing from the conditional distributions of missing observations. The sampler in the Bayesian MSF-SBEKK model with missing data modelled generates the chain states approximately 10 times slower than its equivalent without the days with unobserved values. All of the empirical results in the following section are based on the last 400,000 MCMC states, treated as a sample from the posterior distribution.

4 Joint analysis of volatility on different markets

In this section we try to examine how the way missing data are treated influences empirical results of joint modelling of several asset prices. We consider two methods. In the first one, we remove the days with missing data. In the second one, we treat unavailable data as latent variables and formally estimate them. We compare the posterior distributions of the model parameters in both cases and we check whether main conclusions about volatility and conditional correlation are affected by methods of handling missing price values. Finally, we present posterior distributions of some unavailable prices.

4.1 Data description

A six dimensional portfolio ($n = 6$) of assets from different markets is considered:

- stock exchanges, here represented by the main Polish index, WIG and American S&P500,
- noble metals markets – gold and silver prices (London Fix, USD/oz),
- energy commodities markets – crude oil and natural gas (WTI Spot Price, USD/Barrel and Henry Hub Gulf Coast Natural Gas Spot Price, USD/MMBTU).

We analyze daily data from December 21, 2005 till September 30, 2011. This results in 1492 days when at least one asset was valued.
Let us focus on a short period taken from the final part of the analyzed 1492 days; this period is presented in Table 1. There are three days with incomplete data. August 15 is a bank holiday in Poland (Assumption of the Blessed Virgin Mary), September 5 is a public holiday in the USA (Labor Day – first Monday of September) and August 29 is a bank holiday in the United Kingdom (last Monday of August, called Late Summer Bank Holiday). We could delete these days from our database - however, it would lead to a loss of some information and would implicitly modify the time axis (in our example there are 3 days in 3 adjacent weeks). In some cases, we might need to remove up to 10% of all data (in Table 1 it would be necessary to remove 12 valid data points out of 120). Deleting some data points is also connected with spuriously aggregating individual series where no missing data were found. It might result in artificial jumps (or omitting real jumps) and thus could have an impact on inference about unknown quantities of interest. In the analyzed time series there are 107 days with at least one missing price (7.2% of 1492) and 243 unavailable single data points (grouped in x^*). Details on pairwise missing prices are presented in Table 2.
The descriptive statistics of the logarithmic returns (calculated with or without the days characterized by missing values) can be found in Table 3. In the first case (with unavailable price values), linear interpolation was used to fill the missing values. The most visible difference is in the kurtosis of OIL returns: it has grown from 4.909 to 11.535 after time aggregation. Such a difference might result from deleting data from the days between December 23 and 29, 2008. On December 23, 24, 26 and 29 the oil price values were 30.28, 32.94, 37.58 and 39.89 USD per barrel, respectively. Thus if we delete two middle days when the stock exchanges in Poland and in the UK were closed (December 24 and 26), the return rate for OIL equals 27.564 percentage points – it is the maximum value in the whole series with removed missing observations. It means that by modifying the time axis we artificially created the biggest outlier on this coordinate. Other such examples will be discussed in the next subsection. In order to graphically present the artificial jumps (outliers) in the return series, we plot in Figure 1 the logarithmic returns for the case of omitted days with incomplete data.

Table 1: A part of the analyzed time series

date	WIG	S&P500	GOLD	SILVER	OIL	GAS
2011-8-10	37368.93	1120.76	1772	38.31	83.05	4.09
2011-8-11	38934.71	1172.64	1760	39.18	85.48	4.06
2011-8-12	39910.95	1178.81	1736	38.29	85.19	4.17
2011-8-15	X	1204.49	1739	39.18	87.88	4.05
2011-8-16	40883.9	1192.76	1782.5	39.36	86.65	4.03
2011-8-17	41125.87	1193.89	1790	40.02	87.58	3.98
2011-8-18	38697.56	1140.65	1824	40.32	82.38	3.98
2011-8-19	38749.61	1123.53	1848	41.98	82.33	3.99
2011-8-22	39549.67	1123.82	1877.5	43.49	84.42	3.97
2011-8-23	39556.82	1162.35	1876	42.88	85.35	4.01
2011-8-24	39588.91	1177.6	1770	42.08	84.99	4.1
2011-8-25	39715.27	1159.27	1729	39	85.15	4.01
2011-8-26	39774	1176.8	1788	41.06	85.37	3.96
2011-8-29	40715.85	1210.08	X	X	87.27	3.93
2011-8-30	41300.1	1212.92	1825	40.9	88.9	3.85
2011-8-31	42222.38	1218.89	1813.5	41.35	88.81	3.97
2011-9-1	41553.09	1204.42	1821	41.47	88.93	4.18
2011-9-2	40544.28	1173.97	1875.25	42.5	88.93	4.12
2011-9-5	38992.56	X	1895	42.71	X	X
2011-9-6	39189.1	1165.24	1895	41.85	85.99	3.93
2011-9-7	40418.3	1198.62	1810	40.98	89.34	3.96

Table 2: Missing prices: individually (diagonal) and pairwise (above diagonal)

date	WIG	S&P500	GOLD	SILVER	OIL	GAS
WIG	42	2	17	12	2	2
S&P500		41	9	8	38	38
GOLD			49	37	9	9
SILVER				37	8	8
OIL					40	40
GAS						40

Table 3: Descriptive statistics of analyzed data: with missing values removed (first line) and linearly interpolated (second line)

	WIG	S&P500	GOLD	SILVER	OIL	GAS
min	-10.186	-9.47	-7.972	-18.958	-12.827	-27.472
	-8.289	-9.47	-7.972	-18.693	-12.827	-25.529
max	6.084	10.957	6.841	18.279	27.564	28.391
	6.084	10.957	6.841	18.279	16.414	28.391
mean	0.007	-0.008	0.084	0.091	0.026	-0.095
	0.006	-0.007	0.078	0.085	0.024	-0.088
std. dev.	1.558	1.588	1.438	2.815	2.844	4.478
	1.481	1.516	1.362	2.670	2.602	4.166
skewness	-0.541	-0.324	-0.351	-0.696	0.858	0.373
	-0.431	-0.268	-0.417	-0.605	0.050	0.389
kurtosis	3.318	8.032	2.937	7.738	11.535	7.06
	2.784	8.804	3.382	7.598	4.909	6.171

Figure 1: Percentage return rates (days with incomplete date are removed)

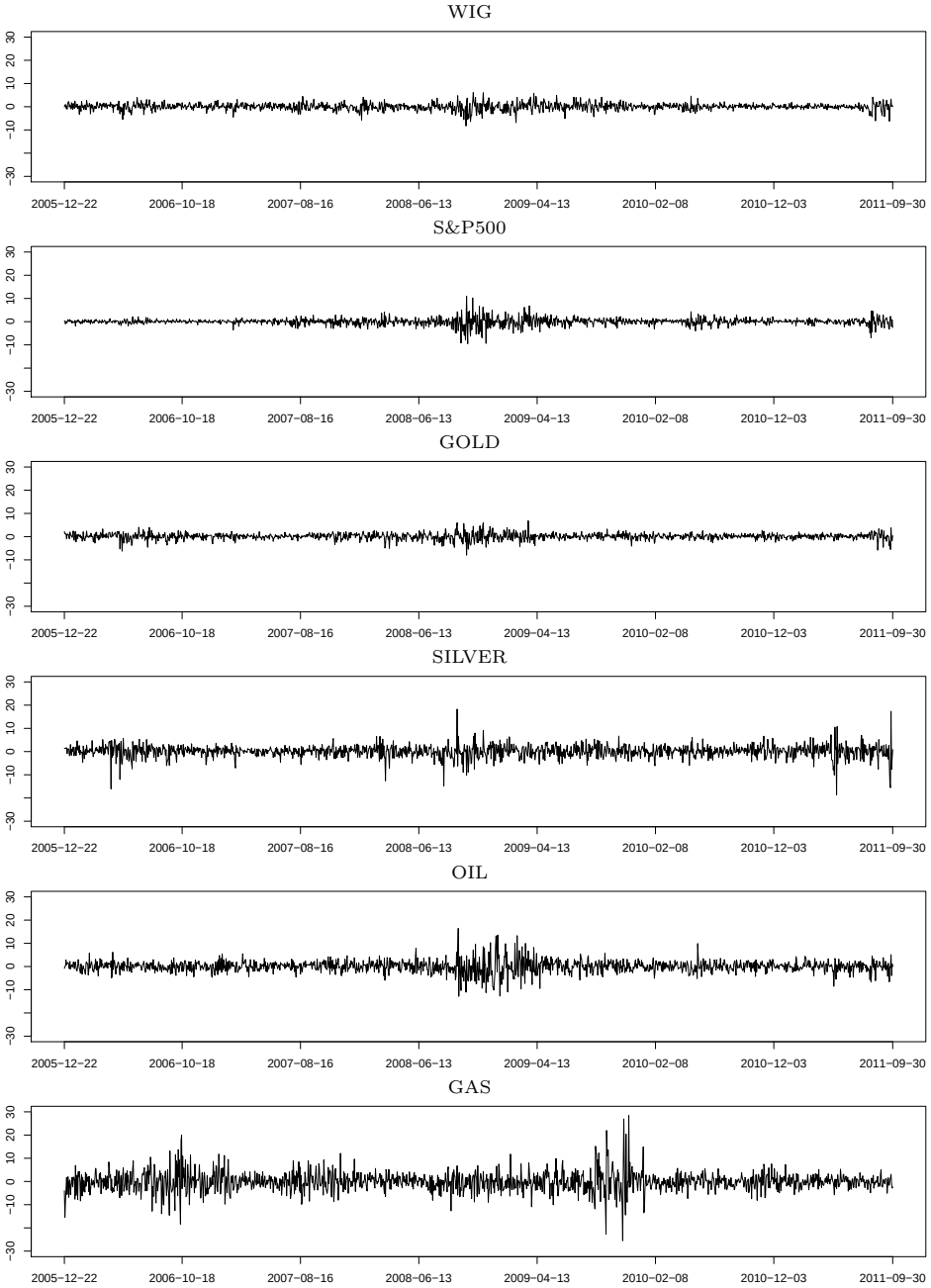
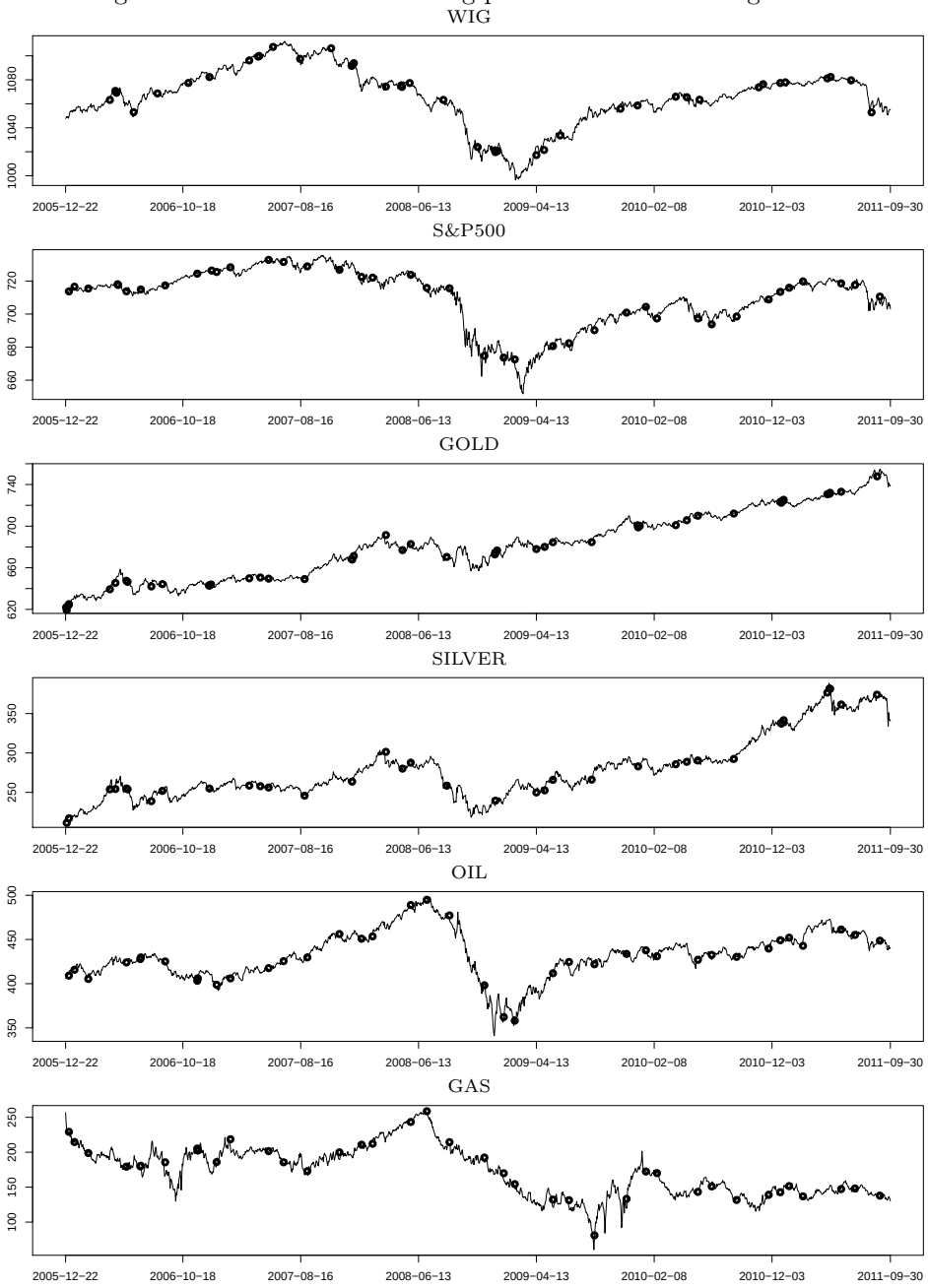


Figure 2: Price data with missing points – hundreds of logarithms



The series $\widehat{x}_t$, i.e. the logarithms of prices (in hundreds) with linear interpolations are plotted in Figure 2. The filled gaps are marked with bold circles. The analyzed period covers the 2008 *subprime* crisis and fuel price turbulences starting in the summer of 2008. Although the *subprime* explosion is well known and described, the OIL price movements in the summer of 2008 need more detailed explanation. Roesser R. (2009) suggests a number of factors which might have contributed to the fuel price spike in the summer of 2008:
1. Gulf of Mexico's Independence Hub was shut down due to a gas leak in early April 2008 – about 10 percent of total Gulf of Mexico production (900 million cubic feet per day) was lost, which resulted in supply shrinkage in this area and thus a price increase;
2. there were low storage volumes;
3. an active hurricane season was forecasted.
These factors contributed together to a spike of the oil price (145 USD/barrel) in July 2008. *"These record high oil prices, along with other emerging problems, such as subprime lending consequences, contributed to a global economic slowdown"* (Roesser R. (2009)). As the stock prices fell, both domestic (in the United States) and global demand for crude oil, mainly driven by production and transport at that time of the year, started to collapse. This affected the prices of oil in the second half of 2008 and early 2009.

4.2 Empirical results

First, we discuss the marginal posterior distributions of the model parameters. In most cases, the posterior distributions differ significantly from the prior ones in the sense that the posteriors are much sharper, much more informative. The data did not provide any strong information only about some parameters in the off-diagonal part of the matrix A. Important conclusions can be drawn from the comparison of marginal posterior distributions in the two estimated models (with incomplete data deleted or missing data forecasted). For the majority of parameters, the posterior histograms are indistinguishable. A small, yet visible, difference appears in the marginal posterior distributions of the parameters in the conditional mean that describe the impact of oil returns on metals and gas returns: Λ_{35}, Λ_{45}, Λ_{65}. In the model with the missing data forecasted, the linkages within the autoregressive part appear to be stronger. This may result from the lack of jumps artificially created in the oil series. The strongest differences are visible in the marginal posterior distributions of the parameters appearing in the specification of the latent process. The variance ($\sigma^2 = \tau^{-1}$) is significantly larger in the model in which the days with missing observations were removed: $p(\sigma^2_{\text{removed}} - \sigma^2_{\text{modelled}} < 0|x) = 0.015$. The parameter ϕ, responsible for autocorrelation in the latent process, is significantly higher when missing data are modelled: $p(\phi_{\text{removed}} - \phi_{\text{modelled}} < 0|x) = 0.923$.
In Figure 3 we only present these parameters, for which either the prior information is not dominated (A_{14}, A_{16}, A_{23}, A_{24}, A_{26}, A_{36}, A_{45}, A_{46}, A_{56}) or the way we treat

missing prices matters (Λ_{35}, Λ_{45}, Λ_{65}, ϕ, τ^{-1}). The posterior histograms are jointly plotted with the prior densities (dashed line). The light grey bars indicate histograms of the marginal posterior distributions obtained in VAR(1)-MSF-SBEKK model, in which missing data were modelled, while the dark grey bars – in the model with removed days. The bold triangles and squares mark the 0.025 and 0.975 quantiles of these distributions.

Let us discuss the posterior expectations of the parameters in the VAR(1)-MSF-SBEKK model with missing data modelled. Whenever zero was not within the 95% highest posterior density region, the values were marked with bold. The values in brackets are posterior standard deviations. At first, let us look at the VAR(1) parameters λ and Λ as well as their function $\alpha = (I_{n\times n} - \Lambda)^{-1}\lambda$, which is the unconditional mean in the case of a covariance stationary VAR(1) process:

$$E(\lambda'|x) = \begin{array}{c} \begin{array}{cccccc} \text{WIG} & \text{S\&P500} & \text{GOLD} & \text{SILVER} & \text{OIL} & \text{GAS} \end{array} \\ \left[\begin{array}{cccccc} \mathbf{0.068} & \mathbf{0.079} & \mathbf{0.084} & \mathbf{0.106} & \mathbf{0.131} & -0.105 \\ (0.028) & (0.022) & (0.027) & (0.048) & (0.048) & (0.087) \end{array}\right] \end{array},$$

$$E(\alpha'|x) = \begin{array}{c} \begin{array}{cccccc} \text{WIG} & \text{S\&P500} & \text{GOLD} & \text{SILVER} & \text{OIL} & \text{GAS} \end{array} \\ \left[\begin{array}{cccccc} \mathbf{0.080} & \mathbf{0.069} & \mathbf{0.089} & \mathbf{0.151} & \mathbf{0.127} & -0.102 \\ (0.030) & (0.021) & (0.026) & (0.048) & (0.046) & (0.083) \end{array}\right] \end{array},$$

$$E(\Lambda|x) = \begin{array}{c} \begin{array}{cccccc} \text{WIG}_{-1} & \text{S\&P500}_{-1} & \text{GOLD}_{-1} & \text{SILVER}_{-1} & \text{OIL}_{-1} & \text{GAS}_{-1} \end{array} \\ \left[\begin{array}{cccccc} -0.043 & \mathbf{0.242} & -0.021 & 0.002 & 0.007 & -0.008 \\ (0.027) & (0.031) & (0.031) & (0.015) & (0.016) & (0.008) \\ 0.019 & \mathbf{-0.091} & \mathbf{-0.062} & 0.017 & -0.013 & 0.006 \\ (0.022) & (0.029) & (0.026) & (0.012) & (0.014) & (0.006) \\ -0.033 & \mathbf{0.049} & -0.026 & 0.015 & \mathbf{0.077} & 0.006 \\ (0.026) & (0.028) & (0.031) & (0.015) & (0.015) & (0.007) \\ -0.001 & \mathbf{0.183} & \mathbf{0.687} & \mathbf{-0.350} & \mathbf{0.202} & 0.012 \\ (0.045) & (0.046) & (0.052) & (0.029) & (0.026) & (0.013) \\ -0.035 & \mathbf{0.146} & 0.031 & -0.015 & \mathbf{-0.048} & \mathbf{0.015} \\ (0.047) & (0.051) & (0.049) & (0.026) & (0.029) & (0.013) \\ -0.133 & 0.067 & -0.063 & -0.064 & \mathbf{0.174} & 0.016 \\ (0.075) & (0.064) & (0.089) & (0.043) & (0.039) & (0.028) \end{array}\right] \end{array} \begin{array}{c} \\ \text{WIG} \\ \\ \text{S\&P500} \\ \\ \text{GOLD} \\ \\ \text{SILVER} \\ \\ \text{OIL} \\ \\ \text{GAS} \end{array}$$

All posterior expectations of the elements of α, except for the last one representing the natural gas, seem to indicate that, although strong turbulences occurred, the significant growth of returns prevailed during the analyzed six years. In the autoregressive component, the only lagged factor having a significant impact on other assets returns (except natural gas) is the S&P return series; it is not surprising that the US economy plays the crucial and primary role on other markets. On the other hand, the most imitative and "dependent" asset is silver – with returns positively stimulated not only by their own past, but also by the lagged S&P500, gold and oil returns.

The posterior means (and standard deviations) of the SBEKK-related parameters of

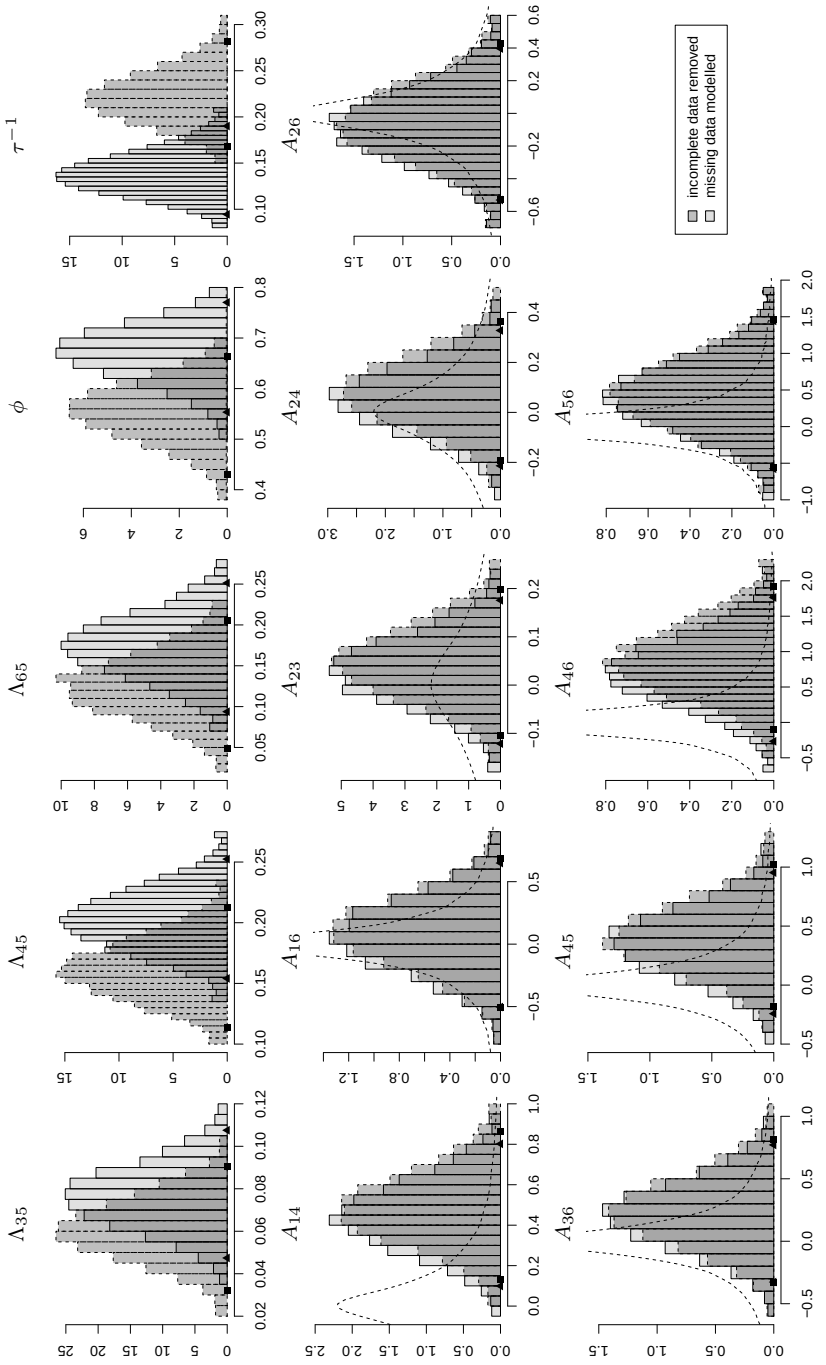


Figure 3: Histograms of chosen posterior marginal distributions of model parameters

the model with missing data forecasted are as follows:

$$E(A|x) = \begin{array}{c} \begin{array}{cccccc} \text{WIG} & \text{S\&P500} & \text{GOLD} & \text{SILVER} & \text{OIL} & \text{GAS} \end{array} \\ \left[\begin{array}{cccccc} \underset{(0.181)}{\mathbf{1.377}} & \underset{(0.089)}{\mathbf{0.348}} & \underset{(0.101)}{\mathbf{0.238}} & \underset{(0.181)}{\mathbf{0.445}} & \underset{(0.185)}{\mathbf{0.503}} & \underset{(0.297)}{0.073} \\ & \underset{(0.107)}{\mathbf{0.804}} & \underset{(0.076)}{0.029} & \underset{(0.138)}{0.056} & \underset{(0.141)}{\mathbf{0.400}} & \underset{(0.233)}{-0.079} \\ & & \underset{(0.158)}{\mathbf{1.225}} & \underset{(0.215)}{\mathbf{1.256}} & \underset{(0.173)}{\mathbf{0.414}} & \underset{(0.278)}{0.208} \\ & & & \underset{(0.527)}{\mathbf{3.950}} & \underset{(0.303)}{0.356} & \underset{(0.514)}{0.718} \\ & & & & \underset{(0.513)}{\mathbf{3.887}} & \underset{(0.502)}{0.401} \\ & & & & & \underset{(1.429)}{\mathbf{9.986}} \end{array}\right] \end{array} \begin{array}{c} \text{WIG} \\ \text{S\&P500} \\ \text{GOLD} \\ \text{SILVER} \\ \text{OIL} \\ \text{GAS} \end{array}$$

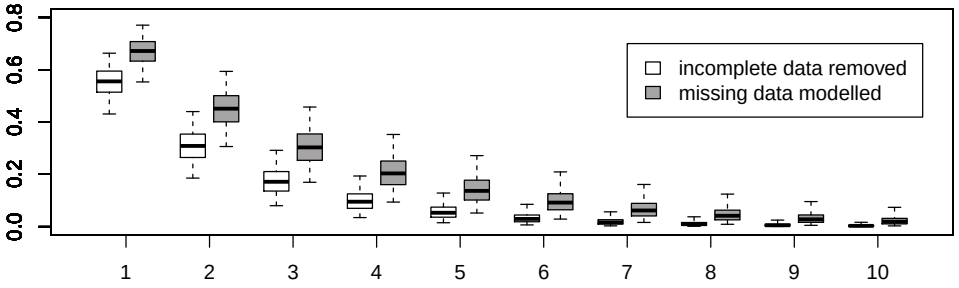
$$E(\beta|x) = \underset{(0.002)}{\mathbf{0.031}}, \quad E(\gamma|x) = \underset{(0.004)}{\mathbf{0.953}}, \quad E(h_0|x) = \underset{(0.405)}{\mathbf{1.190}}.$$

Again, the natural gas returns seem to follow a different volatility pattern than the returns on the remaining assets. Incidentally, the pure SBEKK stationarity condition is supported by the data, as the posterior distribution of $\beta + \gamma$ is separated from 1 (0.999 quantile equals 0.988).

The posterior means (and standard deviations) of the MSV-related parameters of the model with missing data forecasted are as follows:

$$E(\phi|x) = \underset{(0.055)}{\mathbf{0.669}}, \quad E(\tau^{-1}|x) = \underset{(0.024)}{\mathbf{0.139}},$$

Figure 4: Posterior autocorrelation function of $\ln g_t$ – mean with 0.025 and 0.975 posterior quantiles



The posterior distribution of ϕ indicates that the autocorrelation function (plotted in Figure 4) of the latent process $\ln g_t$ is nonzero only for a few lags. The half life of shock to the latent process, i.e. $HL = \frac{\ln 0.5}{\ln \phi}$, is centered at 1.20 in the model with incomplete data removed and at 1.78 in the model where missing data points are modelled;

however, the posterior distributions of these quantities are mostly overlapping. The posterior expectations of the latent process in both models are plotted in Figure 5. At the very beginning of the series the results differ most – which was caused by an artificially created jump in the GAS prices. After December 22, 2005, when GAS was valued at 13.03$ per MMBTU, its price declined on December 23 and 27, taking values of 11.17$/MMBTU and 10.22$/MMBTU, respectively, to end at 9.9$/MMBTU on December 28, 2005. When removing two middle days (as the GOLD was not valued), the growth rate takes an extreme value of -27.42, which was explained by g_t (marked with a filled square in Figure 5). Another such case occurred between April 28, 2011, and May 3, 2011. When the Polish stock market was closed due to the national holiday memorizing the declaration of the Constitution of May 3, 1791 (called the Constitution Day), the SILVER return rate was equal to -11.04. The following one (on May 4) was equal to -7.92, which means that aggregating them into one return rate (due to lack of availability of the value of WIG on May 3) led to an artificial jump of -18.96, again affecting the g_t series (and marked with a filled triangle) in the case when days with incomplete data were removed. The artificial jump created in the OIL series (outlined in the previous subsection) was also marked in Figure 5 with a filled circle. Although it was the most spectacular one, it did not affect the latent process much due to the already high volatilities captured in the MGARCH part – as this outlier fell in the time of perturbances related to the financial crisis.

In order to summarize our inferences on the latent variables g_t we calculated the time averages and standard deviations of the posterior means $E(g_t|x)$; they are equal 1.048 and 0.244 (respectively) in the case of forecasting unavailable data, and 1.038 and 0.207 in the case of removing incomplete data. This means that in both cases the average level and dispersion of the estimates of g_t are very close. Also, the correlation coefficient between the two series of posterior means of g_t is 0.907, indicating very similar dynamics. However, removing days with incomplete data led to much higher kurtosis of $E(g_t|x)$, namely 14,116, than in the case of forecasting unavailable data points, which resulted in kurtosis equal to 4,981 only. Posterior inferences on volatility of the returns were somewhat different. In Figures 7a-7c the posterior means of the conditional standard deviations $\sigma_{t,i} = D\left(r_{t,i}|\psi_{t-1}, g_t, \theta\right)$ (measuring individual volatilities) are presented for both cases: with missing data modelled and some days removed. The conditional standard deviation of the oil returns tends to peak higher when the days with missing data are removed. Summary statistics of our Bayesian volatility estimates for each t, i.e. $E\left(\sigma_{t,i}|x\right)$, are presented in Table 4. First, volatility of each asset is on average higher and more dispersed in the case of removing incomplete data. Second, the correlation coefficients between volatility estimates in both cases (incomplete data removed or forecasted) are very high (above 0.93), except for the OIL series, where the correlation coefficient is below 0.93 (and equal to 0.864). Another interesting aspect is the analysis of cross-market effects, hereafter measured by the posterior means of the conditional correlation coefficients; see Figure 8 for 6 (out of 15) pairs of assets. These posterior means do not differ by more than 5%

Figure 5: Posterior means (and posterior means plus one standard deviation) of latent variables g_t

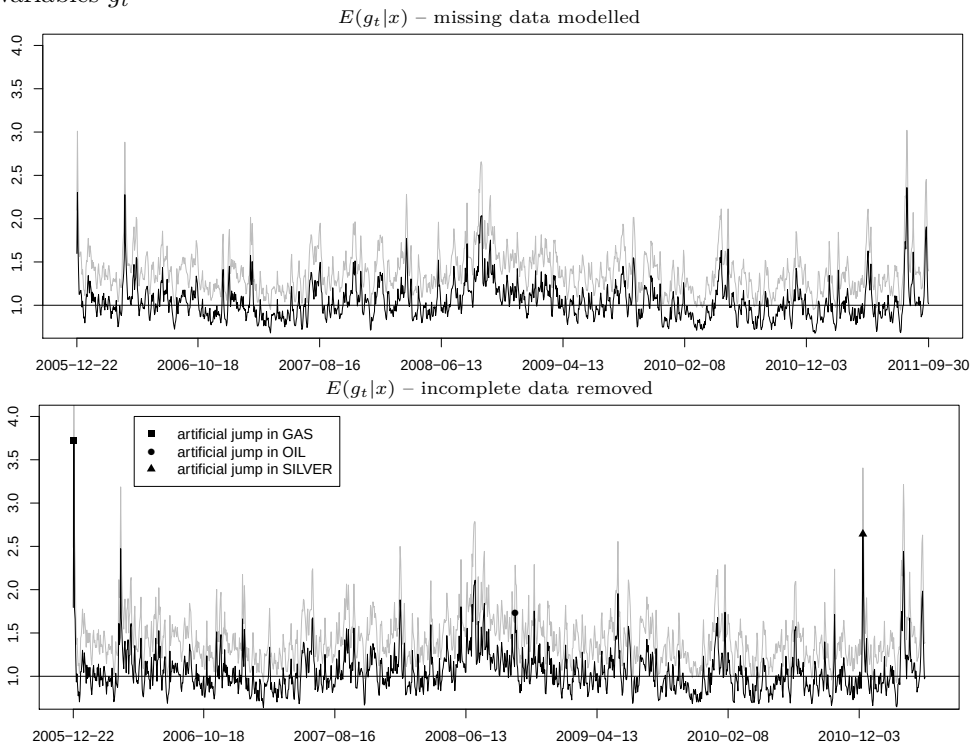


Table 4: Averages (and standard deviations) of posterior means of $\sigma_{t,i}$ in both models and correlations between posterior means

asset	WIG	S&P500	GOLD	SILVER	OIL	GAS
incomplete data removed	1.420 (0.556)	1.295 (0.842)	1.340 (0.556)	2.359 (0.965)	2.488 (1.302)	3.966 (1.549)
missing data forecasted	1.370 (0.515)	1.255 (0.809)	1.288 (0.514)	2.257 (0.881)	2.364 (1.148)	3.772 (1.423)
correlation	0.980	0.976	0.965	0.934	0.864	0.963

when we compare the results obtained in the two models – with incomplete data forecasted or deleted. It seems to be the confirmation of an initial belief that adding incomplete information into the data set with some days removed does not have a strong impact on posterior inference about assets' comovements. Summary statistics of our Bayesian estimates (i.e. posterior expectations) of the conditional correlation coefficients between returns on assets i and j, $\rho_{t,ij} = Corr\left(r_{t,i}, r_{t,j}|\psi_{t-1}, g_t, \theta\right)$ – for $t = 1, \ldots, T$ – are presented in Tables 5 and 6. Table 5 groups all averages and standard deviations of $E\left(\rho_{t,ij}|x\right)$ calculated over $t \in \{1, 2, \ldots, T\}$. Table 6 groups the

correlation coefficient calculated over common observations period between $E(\rho_{t,ij}|x)$ in the two models: with incomplete data removed and missing data forecasted. The way we treat incomplete data seems of little importance for inference on cross-market effects and contagion analysis. The lowest values in Table 6 appear for the dependencies between return on OIL prices and WIG, S&P500, GOLD and SILVER (correlation coefficients about 0.93). As it is visible on Figure 8 and as one might suppose from Section 4.1, the linkage of global crisis tightened the oil and stock markets in the United States for almost two years after the summer of 2008. Gold confirmed to have been a good security at the times of more volatile markets' perturbations. In general, the model confirmed what *subprime* crisis revealed - that diversifying portfolio by investing in noble metals is a good strategy.

Table 5: Averages (and standard deviations) of posterior means of $\rho_{t,ij}$ in the model with missing data forecasted (under diagonal) and incomplete data removed (above diagonal)

	WIG	S&P500	GOLD	SILVER	OIL	GAS
WIG		0.412 (0.124)	0.181 (0.162)	0.204 (0.128)	0.271 (0.156)	0.052 (0.094)
S&P500	0.400 (0.126)		0.043 (0.147)	0.077 (0.131)	0.265 (0.217)	0.023 (0.091)
GOLD	0.174 (0.166)	0.036 (0.152)		0.606 (0.080)	0.220 (0.127)	0.090 (0.116)
SILVER	0.188 (0.130)	0.067 (0.132)	0.594 (0.085)		0.168 (0.116)	0.157 (0.115)
OIL	0.260 (0.153)	0.256 (0.218)	0.214 (0.127)	0.144 (0.121)		0.095 (0.099)
GAS	0.045 (0.093)	0.005 (0.086)	0.084 (0.106)	0.140 (0.110)	0.081 (0.091)	

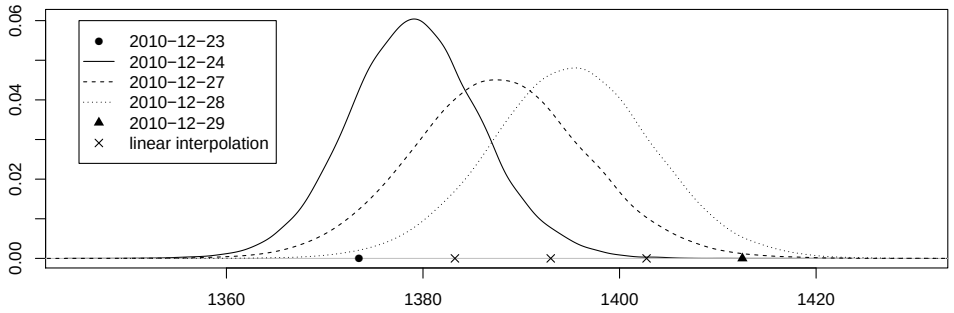
Table 6: Correlation between $E\left(\rho_{\text{modelled};t\in\{j:\widehat{x}_j=x_j\}}|x\right)$ and $E(\rho_{\text{removed}}|x)$

	WIG	S&P500	GOLD	SILVER	OIL	GAS
WIG	–	0.987	0.983	0.979	0.931	0.968
S&P500		–	0.969	0.986	0.934	0.985
GOLD			–	0.973	0.933	0.967
SILVER				–	0.933	0.958
OIL					–	0.968
GAS						–

Finally we shall move to the posterior distributions of the missing values themselves. The plots of daily GAS prices, together with whiskers representing the 95% highest posterior density regions for unavailable data, are presented in Figure 9. As we can see, in all cases the posterior mean falls near the linearly interpolated value presented in Figure 2. In other words, the results of linear interpolation have high marginal posterior density values. The posterior variance of any missing observation reacts to volatility (i.e. the conditional standard deviation) of prices of the individual

asset in question and other assets in the portfolio. The more volatile the market is, the less certain we are about the unobserved price. As an example, the posterior distributions of two latent variables were zoomed in and juxtaposed on the series in Figure 9. The first one represents a rather calm period (when the posterior conditional standard deviation oscillated around 2), while the second one is located where the posterior conditional standard deviation peaks even above 10. The posterior standard deviations of these two latent variables (missing data points) are equal to 0.016 and 0.058, respectively. We presented only the GAS prices as this series is most spectacular in terms of differences in posterior variances of missing data points. However, forecasting unavailable prices of other assets has lead to similar results.
In some cases, missing values appeared a few times in a row. The longest such period included 3 missing values in the gold price series in the period around Christmas 2010 (between December 23 and 29). The posterior distributions of these three missing prices are plotted in Figure 6. It is clearly visible that for these three days a broad range of possible prices might be fitted. However, values representing a smooth transition seem very likely. In particular, the values obtained by linear approximation (marked with "x" on the price axis) lie above the medians and close to the third quartiles of the marginal posterior distributions of unavailable prices. Instead of smooth transition (represented by linear approximation) we could consider keeping the missing prices at the constant level of the last observed price. It would be a reasonable solution for the first day with missing data, but clearly not for the third day and even not for the second day. Firstly, this would amount to using values of lower and lower posterior density. Secondly, it would create an artificially high return in the first period after price unavailability. So our conclusion that linear interpolation is a good strategy (at least in our particular empirical example) is in agreement with the general results obtained by Doman M. and Doman R. (2010) in a completely different setup.

Figure 6: Posterior densities of unobserved prices – three missing values in a row



The use of linear approximation or any other technique of replacing unavailable prices can be justified from the perspective of formal, purely Bayesian inference in

our VAR(1)–MSF-SBEKK model. Let κ denote all unknown quantities (parameters and latent variables), except missing prices denoted by x^*. We are interested in obtaining the marginal posterior $p(\kappa|x) = \int_{X^*} p(\kappa|x, x^*)\, p(x^*|x)\, dx^*$ and we are able to efficiently draw κ from its conditional posterior given x^*, but drawing x^* is very time consuming. So we replace the marginal posterior of x^*, $p(x^*|x)$, with a very sharp (degenerate) distribution concentrated at some preliminary estimate of x^* (say, $\widehat{x}^*$, which may correspond to linear interpolation). This results in using $p(\kappa|x, x^* = \widehat{x}^*)$, as if one conditioned on a data-based value of x^*, instead of using the true marginal posterior distribution $p(\kappa|x)$. The same argument was presented by Osiewalski J. and Pajor A. (2009) in order to justify replacing large parameter matrices by their OLS based counterparts in the context of Bayesian inference in MSF-SBEKK models for very large portfolios.

5 Concluding remarks

In this paper we consider the Bayesian MSF-SBEKK multivariate volatility model estimated on the basis of incomplete daily prices. We discuss two methods of missing data treatment: the first one is to remove all days with only partially available information and the second one is to include them with missing prices treated as latent variables. The MCMC algorithm is suitably adjusted to sample from the posterior distribution of missing prices. An empirical study on joint modelling of six time series from three different markets (stocks, noble metals and fuels) is presented. In this example, including latent variables that represent missing price values resulted in visible changes in the posterior distribution of the parameters of the latent process in the SV part of the model; consequently, posterior results on individual volatilities changed as well. No differences were found in the case of cross-market comovements of price changes and contagion analysis (examined on the basis of the conditional correlation coefficients).

For all forecasted prices, the results of linear interpolation had high marginal posterior density values. Simple linear approximation appeared a Bayesianly justified shortcut procedure. The exact Bayesian procedure seems too demanding as the MCMC sampler in the model with forecasting unavailable prices was approximately 10 times slower than in the model with such days removed. Because of that, it seems that including the broadest possible data set (with linear interpolation of unavailable data points) is the most practical solution. It prevents from loosing available information, modifying time axis, creating artificial outliers and thus changing volatility estimates within multivariate framework. Therefore we advocate not to delete days with incomplete data, even if such approach is harmless for some particular purposes, like inference on conditional correlation between assets returns.

As a by-product of our empirical study we also show that the use of the MSF-SBEKK modelling framework can provide substantial information about cross-market links.

We confirmed that during the period of *subprime* crisis not the energy commodities, but the noble metals market was the right place to seek for hedging against risk.

Acknowledgements

The authors would like to sincerely thank Anna Pajor, Błażej Mazur and Łukasz Kwiatkowski for their constructive remarks on the preliminary version of this study. Special thanks go to an anonymous referee and a co-editor for useful comments on the previous version of the paper.

References

[1] Doman M. and Doman R. (2010), Modelowanie zależności pomiędzy notowaniami giełdowymi o różnych wzorcach (Modelling dependencies between stock quotes of different patterns [in:], *Ekonometria i statystyka w procesie modelowania (Econometrics and Statistics in the Process of Modelling)*, edited by T.Walczak, Główny Urząd Statystyczny (Central Statistical Office), Warszawa.

[2] Jacquier E., Polson N. and Rossi P. (1994), Bayesian analysis of stochastic volatility models [with discussion], *Journal of Business & Economic Statistics*, vol. 12, s. 371-417.

[3] Kim J. (2005), *Parameter Estimation in Stochastic Volatility Models with Missing Data Using Particle Methods and The EM Algorithm*. PhD dissertation, University of Pittsburgh.

[4] Kim J. and Stroffer D. (2008), Fitting Stochastic Volatility Models in the Presence of Irregular Sampling Via Particle Methods And The EM Algorithm. *Journal of Time Series Analysis* Vol. 29, No. 5, p. 811–833.

[5] Osiewalski J. (2009), New Hybrid Models of Multivariate Volatility (a Bayesian Perspective). *Przegląd Statystyczny (Statistical Review)*, vol. 56(1).

[6] Osiewalski J. and Osiewalski K. (2011a), Modele hybrydowe MSV-MGARCH z dwoma procesami ukrytymi (Hybrid MSV-MGARCH models with two latent processes), *Zeszyty Naukowe Uniwersytetu Ekonomicznego w Krakowie, seria Finanse*, nr 895 (forthcoming)

[7] Osiewalski J. and Osiewalski K. (2011b), Modele hybrydowe MSV-MGARCH z trzema procesami ukrytymi w badaniu zmienności cen na różnych rynkach (Hybrid MSV–MGARCH models with three latent processes in examining price volatility on different markets), *Folia Oeconomica Cracoviensia*, vol. LII (2011), 71–85.

[8] Osiewalski J. and Pajor A. (2007), Flexibility and Parsimony in Multivariate Financial Modelling: A Hybrid Bivariate DCC-SV Model [in:] *Financial Markets: Principles of Modelling, Forecasting and Decision-Making*, FindEcon Monograph Series No 3 [eds.] Milo W. and Wdowiński P., Łódź University Press.

[9] Osiewalski J. and Pajor A. (2009), Bayesian Analysis for Hybrid MSF-SBEKK Models of Multivariate Volatility. *Central European Journal of Economic Modelling and Econometrics* vol 1 issue 2, 179–202.

[10] Osiewalski J. and Pajor A. (2010), Bayesian Value-at-Risk for a portfolio: multi– and univariate approaches using MSF–SBEKK models. *Central European Journal of Economic Modelling and Econometrics* vol 2 issue 4, 253–277.

[11] Pajor A. (2003), *Procesy zmienności stochastycznej w bayesowskiej analizie finansowych szeregów czasowych (Stochastic Variance Processes in Bayesian analysis of Financial Time Series)*, Monografie: Prace Doktorskie Nr 2, Wydawnictwo Akademii Ekonomicznej w Krakowie, Kraków.

[12] Pajor A. (2010), *Wielowymiarowe procesy wariancji stochastycznej w ekonometrii finansowej. Ujęcie bayesowskie (Multivariate Stochastic Variance Processes in Financial Econometrics. Bayesian approach)*, Wydawnictwo Uniwersytetu Ekonomicznego w Krakowie, Kraków.

[13] Pajor A. and Osiewalski J. (2012), Bayesian Value-at-Risk and Expected Shortfall for a Large Portfolio (Multi- and Univariate Approaches). *Acta Physica Polonica A* vol 121 issue 2B, 101–109.

[14] Roesser R. (2009), Natural Gas Price Volatility, *California Energy Commission*. CEC-200-2009-009-SD.

[15] Tsay R. (2005), *Analysis of Financial Time Series,* 2^{nd} *ed.*, Wiley.

[16] Yu B. and Mykland P. (1998), Looking at Markov samplers through cusum path plots: a simple diagnostic idea, *Statistics and Computing* 8, 275–286.

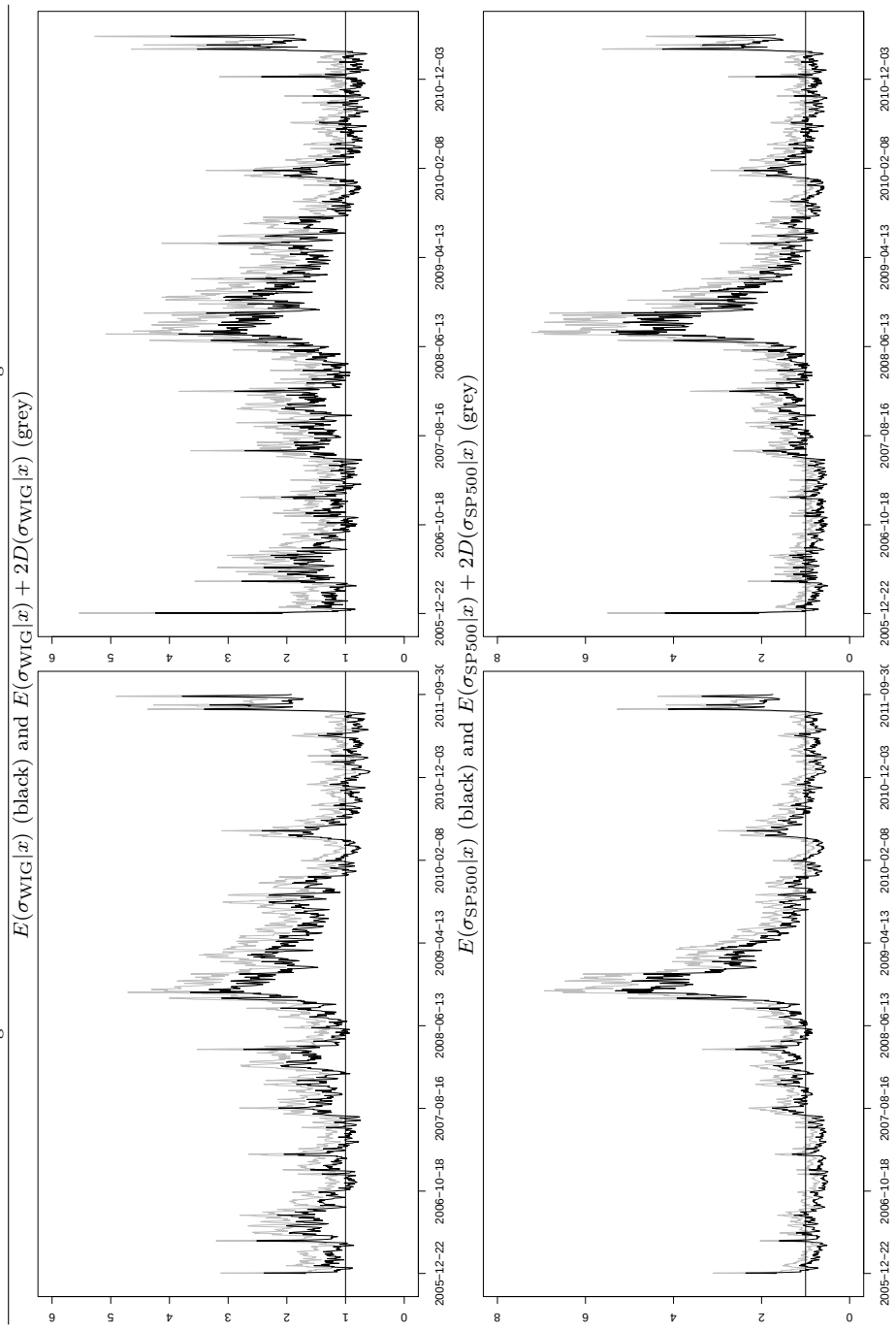


Figure 7a: Conditional standard deviations (posterior mean and posterior mean plus two standard deviations)

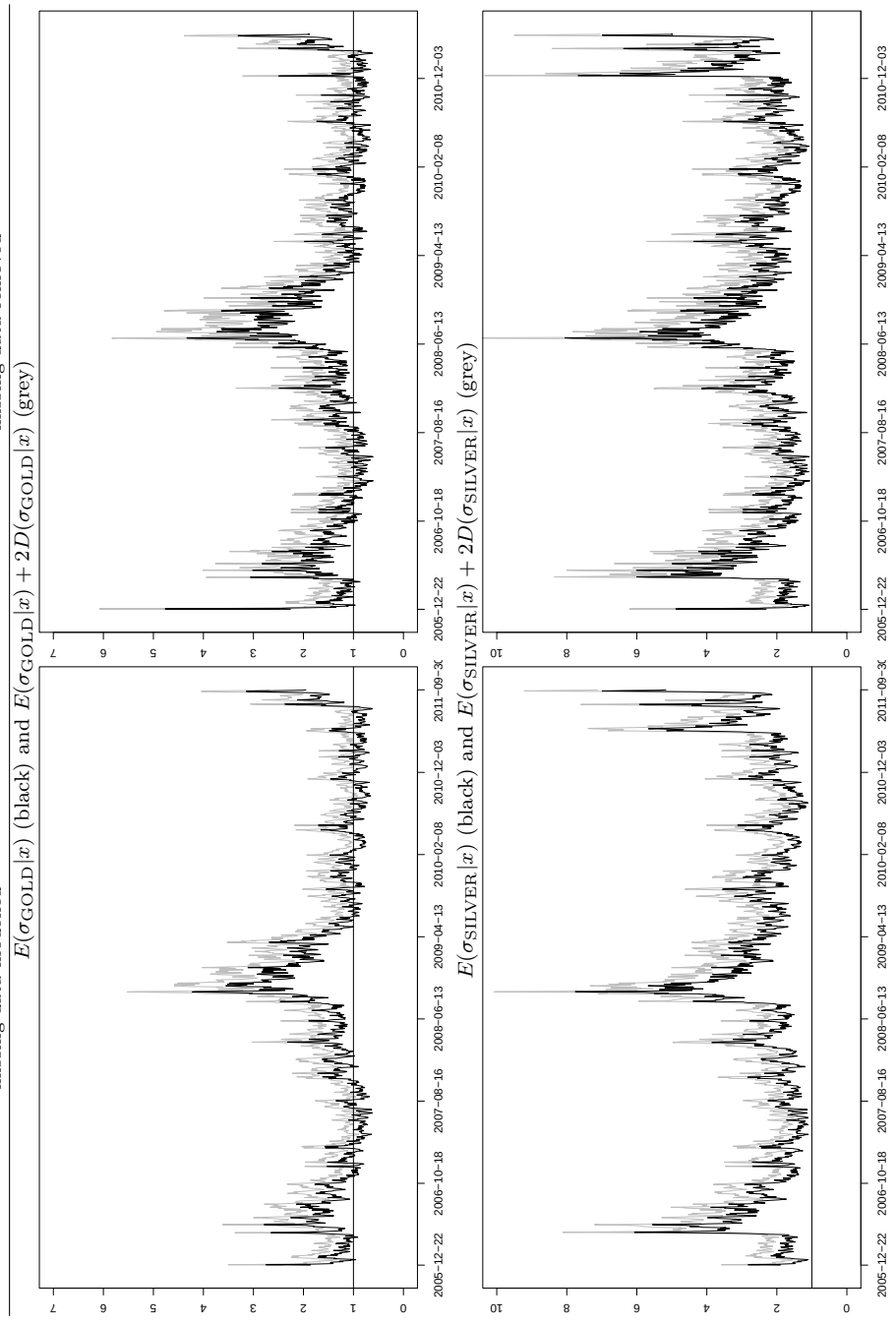


Figure 7b: Conditional standard deviations (posterior mean and posterior mean plus two standard deviations)

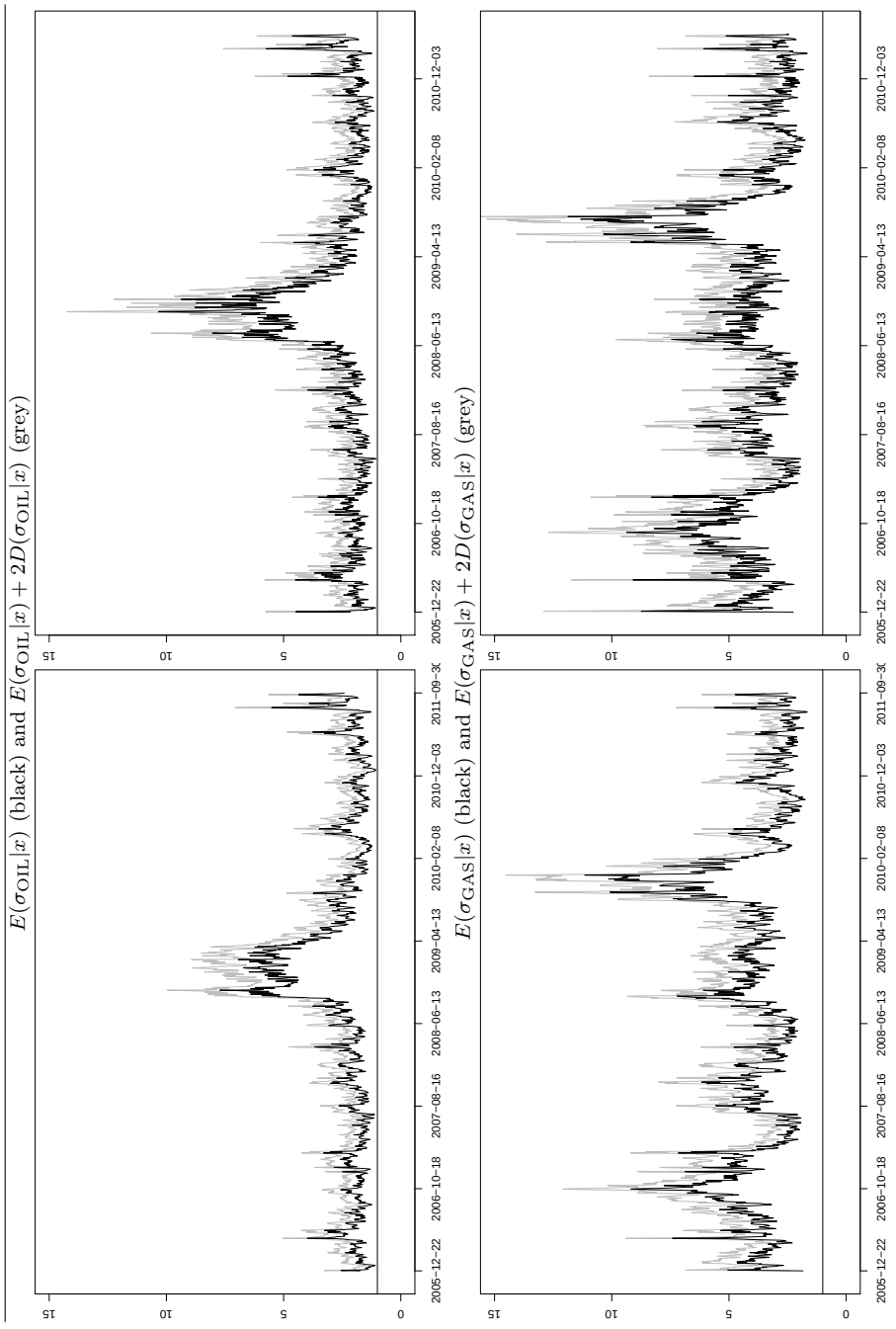


Figure 7c: Conditional standard deviations (posterior mean and posterior mean plus two standard deviations)

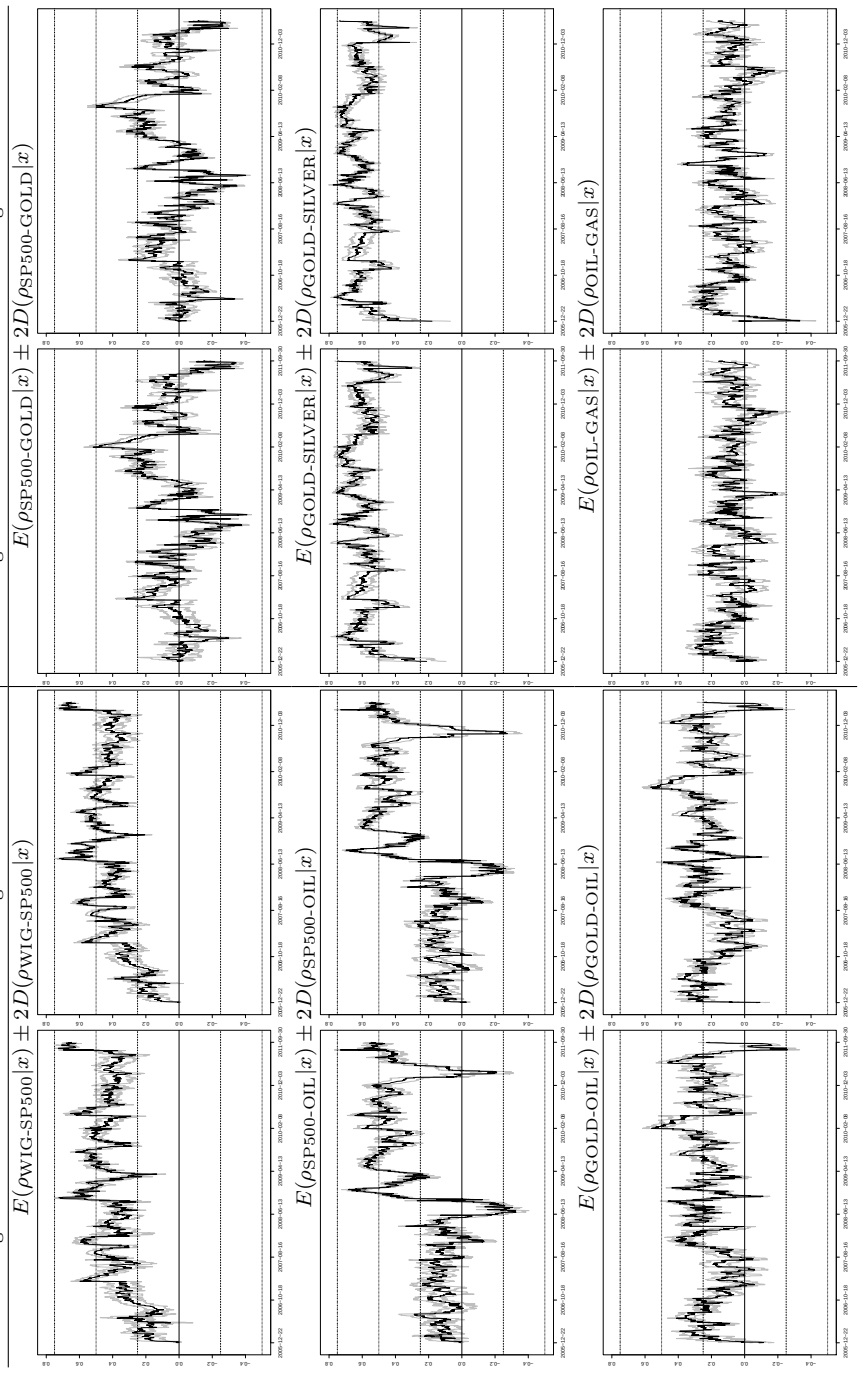


Figure 8: Conditional correlations (posterior mean $\pm$ two standard deviations)

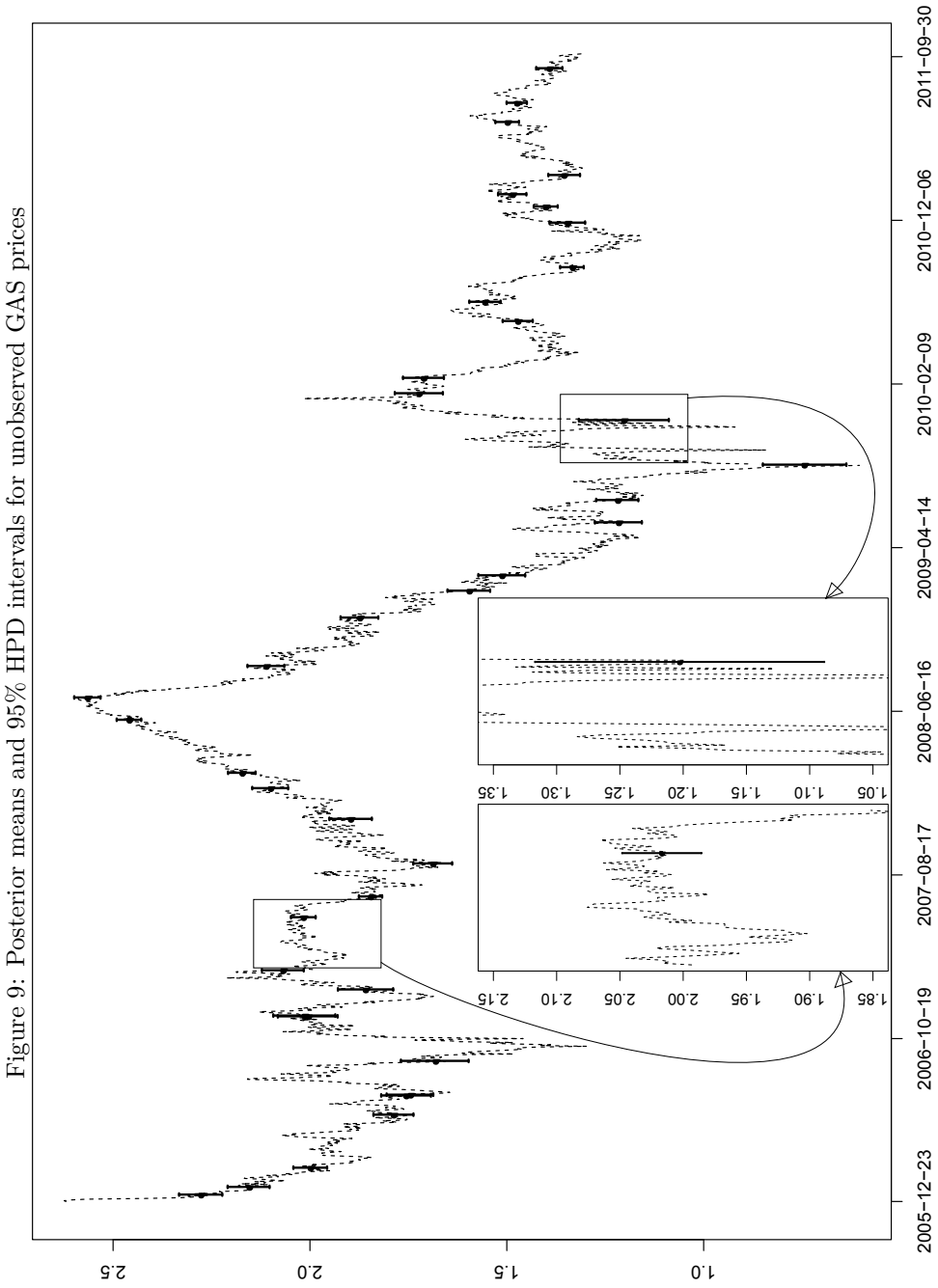


Figure 9: Posterior means and 95% HPD intervals for unobserved GAS prices