Open Journal of Statistics, 2012, 2, 504-511
http://dx.doi.org/10.4236/ojs.2012.25065 Published Online December 2012 (http://www.SciRP.org/journal/ojs)

Bayesian Factorized Cointegration Analysis

Kai Cui[1], Wenshan Cui[2]
[1]Department of Statistical Science, Duke University, Durham, USA
[2]School of Science and Information, Qingdao Agricultural University, Qingdao, China
Email: kc52@stat.duke.edu, wshcui@qau.edu.cn

Received October 25, 2012; revised November 30, 2012; accepted December 8, 2012

ABSTRACT

The concept of cointegration is widely used in applied non-stationary time series analysis to describe the co-movement of data measured over time. In this paper, we proposed a Bayesian model for cointegration test and analysis, based on the dynamic latent factor framework. Efficient computational algorithms are also developed based on Markov Chain Monte Carlo (MCMC). Performance and efficiency of the the model and approaches are assessed by simulated and real data analysis.

Keywords: Cointegration; Bayesian; Dynamic Factor; Non-Stationary; Root Structure; MCMC

1. Introduction

Many macroeconomic and financial time series are non-stationary [1], characterized by the existence of stochastic trends or unit roots. Thus, modeling multivariate non-stationary time series via linear regression, while has been widely used on stationary time series analysis, has been shown to be a dangerous approach that could produce spurious regression [1]. Alternatively, if linear combinations of non-stationary or unit-root variables are stationary, then cointegration is said to occur, which is often observed in practice. Since the development of the concept of cointegration [2], there has been a rich literature on testing cointegration and inference using such models. However, most of the work adopted a classical econometric perspective based on the linear cointegrated error correction models. In this study, the Bayesian factorized cointegration analysis proposed here as a different perspective shed new lights on many key properties of cointegration models.

To establish the notion and illustrate the basic ideas underlying classical cointegration analysis, let $\{y_t\}_{t=1}^{T}$ be a realization of Q dimensional Vector Autoregressive (VAR) process of lag order p:

$$y_t = \sum_{i=1}^{p} \Gamma_i y_{t-i} + \Phi + \varepsilon_t$$

where $\varepsilon_t \sim N_Q(0, \Sigma)$. Φ denotes the autoregression intercept. This model can be written in the Vector Error Correction Model (VECM) form as:

$$\Delta y_t = \Pi y_{t-1} + \sum_{i=1}^{p-1} \Psi_i \Delta y_{t-i} + \Phi + \varepsilon_t$$

where the matrix Π of rank r can be written as $\Pi = \alpha\beta'$, where α and β are both full rank $Q \times r$ matrices, with $0 \le r \le Q$ denoting the number of cointegration relationships. Note that as a special case, when Π is zeros matrix with $r = 0$, then all time series are non-stationary with unit roots, but no cointegration relationship exists. β is called the cointegration vector with $\beta' y_t$ being stationary.

Classical cointegration tests have been developed based on the VECM models to test the rank of Π. For example, Johansen test [3] is considered as a generally applicable test for multivariate non-stationary time series that allows more than one cointegration relationship. While in practice when cointegration is often used for two time series, the Engle-Granger two-step method [2] has also been widely used. However, all the classical methods have their clear limitations when applied in practice. Johansen method, as a more general cointegration test method than the other two, is a complicated model with many degrees of freedom. Also, it has to model all the variables at the same time without a clear and straightforward interpretation in terms of exogenous and endogenous variables, which are all less favorable with multivariate cases especially if the relation for some variable is flawed. On the other hand, as the most well-known test for cointegration of two time series, the Engle-Granger test simply tests the stationarity of a given linear combination of two time series based on unit root tests (e.g. ADF test), where the cointegration ratio is obtain by running a static regression of the two time series. However, the two-step procedure suffers from many problems (e.g. multiple testing) and fails in many cases to be an effective approach to identify cointegrated pairs in practice.

In this paper, we propose the framework of Bayesian factorized cointegration analysis as a straightforward and computationally efficient approach for cointegration test and modeling. Although the dynamic latent factor model has been proposed for time series analysis for decades, the idea of incorporation of dynamic structures that can accommodate possible non-stationarity flexibly enough via the Bayesian modeling framework and computational strategies, which we proposed as an alternative Bayesian cointegration test, is novel. The model and approaches are described in in Section 2, followed by Bayesian computation and posterior analysis via Markov Chain Monte Carlo (MCMC) methods in Section 3. Section 4 applies the model to simulated data sets to gauge the performance of the models. In Section 5 the model is applied to detecting and analyzing cointegration between real non-stationary time series, with conclusions and future directions in Section 6.

2. Model Specification

2.1. Basic Dynamic Latent Factor Models

Let $\{y_t\}_{t=1}^{T}$ be a realization of Q dimensional non-stationary time series, we propose a Bayesian dynamic latent factor model to jointly model the multivariate non-stationary time series and detecting cointegration among the component time series via $q\left(q \le Q\right)$ common dynamic latent factors. Specifically, let $\eta_t = \left[\eta_{1t},\cdots,\eta_{qt}\right]'$ denote the vector of q dynamic latent factors, the model is of the form:

$$\begin{aligned} y_t &= \alpha + \Lambda\eta_t + \varepsilon_t \sim \varepsilon_t \sim N\left(0,\text{diag}\left(\sigma_1^2,\cdots,\sigma_Q^2\right)\right), \\ \eta_{it} &= \sum_{k=1}^{p_i}\beta_{i,k}\eta_{i,t-p_i} + \varepsilon_{\eta_{it}}, \varepsilon_{\eta_{it}} N\left(0,\sigma_{\eta_i}^2\right) \end{aligned} \quad (1)$$

where Λ is a lower triangular factor loading matrix. Each factor η_{it} is modeled as following a $AR\left(p_i\right)$ process that can be either stationary or non-stationary.

2.2. Model in the Matrix State-Space Form

Let $\zeta_{it} = \left[\eta_{i,t},\eta_{i,t-1},\cdots,\eta_{i,t-\left(p_i-1\right)}\right]'\left(i=1,\cdots,Q\right)$,

$\zeta_t = \left[\zeta_{1t}',\cdots,\zeta_{Qt}'\right]'$, $F = \begin{bmatrix} F_1' & & \\ & \ddots & \\ & & F_Q' \end{bmatrix}$, where

$F_i' = \left[1,0,0,\cdots\right]$ is a $1\times p_i$ vector, and

$A = \begin{bmatrix} A_1 & & \\ & \ddots & \\ & & A_Q \end{bmatrix}$, where $A_i = \begin{bmatrix} \beta_{i1},\cdots,\beta_{i,p_i-1} & \beta_{i,p_i} \\ I_{p_i-1} & 0 \end{bmatrix}$.

A_i is the transition matrix giving $\zeta_{it} = A_i\zeta_{i,t-1} + \varepsilon_{\zeta_i}$, with $\varepsilon_{\zeta_i} \sim N\left(0,W_i\right)$, where $W_i = \begin{bmatrix} \sigma_{\eta_i}^2 & \\ & 0_{p_i-1} \end{bmatrix}$. Then model (1) can be written in the matrix state-space form:

$$\begin{aligned} y_t &= \alpha + \Lambda F\zeta_t + \varepsilon_t, \quad \varepsilon_t \sim N\left(0,\text{diag}\left(\sigma_1^2,\cdots,\sigma_Q^2\right)\right) \\ \zeta_t &= A\zeta_{t-1} + \varepsilon_\zeta, \quad \varepsilon_\zeta \sim N\left(0,\text{diag}\left(W_1,\cdots,W_Q\right)\right) \end{aligned} \quad (2)$$

The matrix state-space form model (2) is primarily used in deriving efficient computational strategies for sampling latent factors as shown in the following section. As stated previously, although this dynamic latent factor model has been applied for time series analysis for decades, our idea of the incorporation of dynamic structures that can accommodate possible non-stationarity flexibly enough via the Bayesian modeling framework and computational strategies, which we proposed as an alternative Bayesian cointegration test, is novel and shown in the following sections.

2.3. Modeling Dynamic Latent Factors

The key idea of modeling multivariate nonstationary time series via independent latent factors requires flexible modeling of the dynamic factors. It is critical that the model is flexible in terms of two aspects. The values of $AR\left(p\right)$ coefficients of the dynamic structure of latent factors need to be able to flexibly model stationary and non-stationary (unit-root) processes. And secondly, the autoregressive order p should also be flexibly modeled.

To achieve these, we followed the $AR\left(p\right)$ process decomposition proposed by [4], with spike-and-slab priors placed on the root structures. Induced priors on the $AR\left(p\right)$ coefficients can be deduced. We further used half-cauchy prior for the innovation variances, as recommenced by [5] for hierarchical modeling.

In details: given that the i^{th} time-varying latent factor is modeled via an autoregressive process with (maximum) lag order p_i, we have that:

$$\phi_i\left(B\right)\eta_{it} = \varepsilon_{it}, \varepsilon_{it} \sim N\left(0,\sigma_{\eta_i}^2\right) \quad (3)$$

where B is the backshift operator, $\phi_i\left(\mu\right) = 1 - \sum_{k=1}^{p_i}\beta_{ik}\mu^k$ is the characteristic polynomial of the model, $\beta_i = \left(\beta_{i,1},\cdots,\beta_{i,p_i}\right)'$ is a $p_i\times 1$ vector of the autoregression coefficients of the i^{th} dynamic latent factor, and ε_{it} is white noise.

Assume that the characteristic polynomial $\phi_i\left(\mu\right)$ has R_i non-zeros real roots and C_i pairs of non-zero conjugate complex roots (with $p_i = R_i + 2C_i$)[1], then the autoregressive process can be characterized in term so either the autoregressive coefficients β_i or the p cha-

[1]The roots are also the eigenvalues of A defined in (2).

racteristic roots. Let $\tau_j = r_j \exp(\pm i\omega)(j = 1, \cdots, C_i)$ denote the C_i pairs of complex roots and $\tau_j = r_j (j = 2C+1, \cdots, p)$ denote the real roots, then we place a class of hierarchical models on the root structures as introduced by [6], which indirectly induce models on the β_i s that either fall into the stationary or unit-root process space. Specifically, the models we used are:

1) For the complex roots $r_j \exp(\pm i\omega_j)(j = 1, \cdots, C_i)$, we model the modulus r_j and period $\lambda_j = 2\pi/\omega_j$ as follows:

$$\begin{aligned} r_j &\sim \pi_0 I_0(r_j) + \pi_1 I_1(r_j) + \pi_b \text{Beta}(r_j; \beta_0, 1) \\ \lambda_j &\sim \text{Unif}(2, T/2) \\ (\pi_0, \pi_1, \pi_b) &\sim \text{Dir}(1,1,1) \end{aligned} \tag{4}$$

where $I_0(\cdot)$ and $I_1(\cdot)$ are Dirac delta functions at 0 and 1 respectively. (π_0, π_1, π_b) are weights with Dirichlet prior distributions, and T is the number of time points.

2) For the real roots, similarly we have

$$\begin{aligned} r_j \sim\ & \pi^*_{-1} I_{-1}(r_j) + \pi^*_1 I_1(r_j) + \pi^*_0 I_0(r_j) \\ & + \pi^*_u \text{Unif}(r_j; -1, 1) \end{aligned} \tag{5}$$

Clearly, first of all, in modeling complex and real roots, by giving a probability of having root at 0, the uncertainly of lag orders p_i is incorporated in the model and thus can be simultaneously inferred from the number of non-zero real and complex roots. The framework allows specifying a very high-ordered autoregressive model without running into over-fitting problems, and thus can flexibly model the dynamic structure of latent factors. Secondly, stationarity is also simultaneously assessed, with the existence of at least one unit roots indicating the non-stationarity of a dynamic latent factor, which provides a useful link between the posterior distribution of root structures and testing cointegration.

2.4. Bayesian Cointegration Test via Latent Factors

We focus on the case where the all Q component time series of y_t are known to be non-stationary *a priori*. In model (1), if we partition η_t as a vector of q^* non-stationary (η_t^{NS}) and $(q - q^*)$ stationary (η_t^S) latent factors, then model (1) can also be written as:

$$y_t = \alpha + \Lambda_1 \eta_t^{NS} + \Lambda_2 \eta_t^S + \varepsilon_t \tag{6}$$

Proposition 1. *Given model* (6) *and that* $y_{tt=1}^T$ *is a* $Q \times 1$ *non-stationary time series, the followings are equivalent*:

1) y_t is cointegrated with rank *r*.

2) $\exists r$ linearly independent $\gamma \neq 0$, *s.t.* $\gamma' y_t$ is a stationary process. (existence of cointegration)

3) $\text{rank}(\Lambda_1) = r < Q$.

Therefore, testing the existence of cointegration can be achieved by testing whether the rank of Λ_1 is smaller than *Q*. Note that the cointegration vector γ may not be unique, if there exist *r* such linearly independent $\gamma_i, (i = 1, \cdots, r)$, then y_t is said to be cointegrated with cointegration rank *r*. It is worth mentioning that with two nonstationary time series, γ is unique if it exists and can be easily derived based on Λ_1. And we call the ratio between the two time series as cointegration ratio in the following studies.

3. Bayesian Computation and Posterior Analysis

We use Markov Chain Monte Carlo (MCMC) sampler for the Bayesian inference of unknown quantities in the model. The sampler is computationally efficient and mixes rapidly due to the conditionally multivariate normal matrix state-space representation of the model as shown by (2). In details:

Let Θ denote all model parameters in (1), so that

$$\Theta = \left\{ \alpha, \Lambda, \sigma_i^2; \beta_{i,k}, \sigma_{\eta_j}^2; (\pi_0, \pi_1, \pi_b); (\pi^*_{-1}, \pi^*_1, \pi^*_0, \pi^*_u) \right\} \tag{7}$$

where $i = 1, \cdots, Q; k = 1, \cdots, p_i; j = 1, \cdots, q$.

Also, recall that $\zeta_0 = (\zeta_{1,0}, \cdots, \zeta_{q,0})$ defined in (2), with $\zeta_{i,0} = (\eta_{i,0}, \eta_{i,-1}, \cdots, \eta_{i,p_i-1})'$ being a $p_i \times 1$ vector of initial values for the i^{th} dynamic latent factor. Posterior inference is based on a standard Gibbs sampler that iteratively simulate from posterior full conditional distributions

$$\begin{aligned} & f(\eta_{1:T} | \zeta_0, \Theta, \mathbf{y}_{1:T}), f(\Theta | \zeta_0, \eta_{1:T}, \mathbf{y}_{1:T}) \\ & \text{and } f(\zeta_0 | \Theta, \eta_{1:T}, \mathbf{y}_{1:T}) \end{aligned}.$$

Specifically,

1) To sample from $(\eta_{1:T} | \zeta_0, \Theta, \mathbf{y}_{1:T})$.

To obtain samples of latent factors $(\eta_{1:T})$ from the full conditional distribution, we applied a efficient sampler based on the Forward-Filtering-Backward-Sampling algorithm. Basically, forward filtering follows (2) and is performed in terms of ζ_t. For backward sampling, ζ_t is first sampled, giving samples of $\eta_{(T-p+1):T}$. Then $\eta_{1:(T-p)}$ are sequentially sampled backwardly based on the backward sampling distributions derived from model (1). The detailed filtering and sampling algorithm is shown in Appendix I.

2) To sample from $f(\Theta | \zeta_0, \eta_{1:T}, \mathbf{y}_{1:T})$.

Given latent factors $\eta_{1:T}$ and the initial values ζ_0, root structures τ_j are sampled following [6], from which samples of autoregression coefficients β_i can be derived from the following equation:

$$\left[\prod_{j=1}^{p}\left(1-\tau_j B\right)\right]\eta_t = \eta_t - \sum_{j=1}^{p}\beta_j \eta_{t-j}$$

where B is the backshift operator.

Given the half-cauchy priors for the innovation variances, posterior distributions of $\sigma_{\eta_i}^2\ (j=1,\cdots,q)$ do not have closed form, and thus are sampled using Adaptive-Rejection Sampler (ARS). Other model parameters generally have conjugate priors and have closed-form posteriors.

3) To sample from $f\left(\zeta_0 \middle| \Theta, \eta_{1:T}, \boldsymbol{y}_{1:T}\right)$.

We take advantage of the reversibility of $AR(p)$ process to sample initial values ζ_0. So,

$$\eta_{i,t} = \sum_{k=1}^{p_i}\beta_{i,k}\eta_{i,t+k} + \tilde{\varepsilon}_{i,t}, \text{ where } \tilde{\varepsilon}_{i,t} \sim N\left(0,\sigma_{\eta_i}^2\right) \quad (8)$$

Therefore, for the initial values $\zeta_{i,0} = \left(\eta_{i,0}, \eta_{i,-1}, \cdots, \eta_{i,p_i-1}\right)'$ of the i^{th} latent factor, $\eta_{i,t}$ are sequentially sampled backwardly from $t=0$ to $t=-\left(p_i-1\right)$.

4. Simulation Examples

4.1. Study 1: Simulation from Factor Model

We first look at the case of detecting cointegration between two time series, where the time series y_t are generated from two latent dynamic latent factors $\left(\eta_t = \left[\eta_{1t}, \eta_{2t}\right]'\right)$, where η_{1t} follows random walk and η_{2t} follows a stationary $AR(2)$ process, as shown in (9).

where

$$\begin{aligned} &y_t = \left[y_{1t}, y_{2t}\right]', \\ &y_t = \alpha + \Lambda\eta_t + \epsilon_t, \quad \epsilon_t \sim N\left(0,\Sigma_\varepsilon\right), \\ &\eta_{1t} = \eta_{1,t-1} + \epsilon_{1t}, \quad \epsilon_{1t} \sim N\left(0,\sigma_{\eta_1}^2\right), \\ &\eta_{2t} = 0.5\eta_{2,t-1} + 0.24\beta_{2,t-2} + \epsilon_{2t}, \\ &\epsilon_{2t} \sim N\left(0,\sigma_{\eta_2}^2\right). \end{aligned} \quad (9)$$

In the simulation study, we generated the times series of $T=800$, and used $\alpha = \left[1,2\right]'$, $\Lambda = \begin{bmatrix} 1 & 0 \\ -0.2 & 1.5 \end{bmatrix}$, $\sigma_{\eta_1} = \sigma_{\eta_2} = 1$ and $\sigma_\varepsilon = 0.3$. Clearly, in this case, the two time series cointegrate, with the cointegration ratio between y_{1t} and y_{2t} being 5, meaning that $z_t = y_{1t} + 5y_{2t}$ is a stationary time series. To test the model we proposed for cointegration analysis, we modeled the simulated time series using a two-factor dynamic factor model as shown in (1), specifying the maximum number of pairs of complex roots to be 2, and that of real roots to be 4 for both dynamic latent factors. Thus, the maximum autoregressive order for both latent factors are 8. For efficient MCMC sampling, priors are placed on parameters of the Parameter-eXpanded (PX) model as described by [7] and [8]. 30,000 posterior samples are drawn for model parameters, latent initial values and latent factors via MCMC after a 5000 iteration burn-in.

Posterior inference of the existence of cointegration in this case is to test whether the rank of the factor loading matrix of non-stationary factors is smaller than 2. Therefore, we first looked at the posterior samples of the characteristic roots of both η_{1t} and η_{2t} and the factor loading matrix. Two results can be obtained immediately from this. First of all, based on 30,000 posterior samples, the marginal probability of η_{1t} having unit roots is 99.39% and the joint probability of at least one of the latent factors having unit roots is 99.39%, which are consistent with the known facts that both y_{1t} and y_{2t} are non-stationary. Secondly, the probability that the rank of the factor loading matrix of non-stationary factors is smaller than the number of non-stationary component time series is 99.99%, indicating that the two time series are cointegrated at the 99.99% confidence level.

We also specifically focused on obtaining the conditional posterior distribution of the cointegration ratio (R) given the two cointegrate, which is an important quantity in cointegration analysis. This, in this case, can be derived from the posterior samples of roots and the factor loading matrix Λ, as shown in **Algorithm 1**. A histogram of the posterior samples of R is shown in **Figure 1**. R has a posterior mean of 5.14, posterior mode of 5.03, and 95% HPD interval of [4.47,5.79], which is consistent with the true values 5.

1) For the MCMC sampler at the i^{th} iteration, check if η_{1t} has unit roots and η_{2t} does not have unit roots.

2) If 1 satisfies, let $\Lambda^{(i)}$ denote the sample of the factor loading matrix of the i^{th} iteration, and a sample of R can be obtained by: $R = -\frac{\Lambda_{11}^{(i)}}{\Lambda_{21}^{(i)}}$, where $\Lambda_{kl}^{(i)}$ denotes the element of $\Lambda^{(i)}$ at row k and column l.

Algorithm 1. To derive posterior samples of the cointegration ratio (R) from Λ.

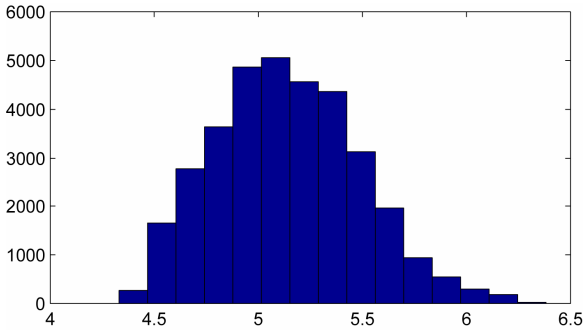

Figure 1. Histogram of posterior samples of cointegration ratio (R) shown, with true value being 5.

Furthermore, we looked at the inference of the latent factors structures. As shown in **Table 1**, the autoregression coefficients are all correctly inferred with posterior mean closed to and posterior C.I. covering the true values. Posterior distributions of autoregressive lag orders of both latent factors are also derived based on the non-zero posterior roots and the histograms are shown in **Figure 2**, which show good inference that the lag order of η_{1t} spikes at 1 and that of η_{2t} peaks at 2.

Overall, the framework provides very good inference of the latent structures and unknown quantities. Cointegration relationship is also found correctly between the two non-stationary time series with good recovery of the cointegration ratio.

4.2. Study 2: Simulation from VECM Model

Out of model studies are shown here. We applied our model and algorithm to simulations generated from the Vector Error Correction Model (VECM). As stated earlier, the VECM representation has been commonly used for modeling and testing cointegration among I(1) nonstationary time series in both classical and Bayesian models.

Taking advantage of the VECM model, an example of second-order nonstationary vector autoregressive model where cointegration exists is first considered here:

$$y_t = \begin{pmatrix} -0.2 & 0.1 \\ 0.5 & 0.2 \end{pmatrix} y_{t-1} + \begin{pmatrix} 0.8 & 0.7 \\ -0.4 & 0.6 \end{pmatrix} y_{t-2} + \varepsilon_t \quad \text{where} = \begin{pmatrix} 100 & 0 \\ 0 & 100 \end{pmatrix} \text{ and } y_0 = 0 \tag{10}$$

This process gives the following VECM(2) representation, indicating that the two time series are co-integrated, and the cointegration vector is $[1,-2]$ (with cointegration ratio being –2).

$$\Delta y_t = \begin{pmatrix} -0.4 \\ 0.1 \end{pmatrix} (1,-2)\, y_{t-1} - \begin{pmatrix} 0.8 & 0.7 \\ -0.4 & 0.6 \end{pmatrix} \Delta y_{t-1} + \varepsilon_t \tag{11}$$

Our goal is to test whether our proposed framework can correctly detect the cointegration relationship between the time series and provide inference of other unknown quantities. We simulated the time series up to $T = 800$, with 8 time points of y_{1t} randomly missing. Two-factor model is fitted to the data, and set the maximum number of complex pairs and real roots to be 5 and 10 respectively, making the maximum autoregressive lag order being 20. Missing data are interpolated during MCMC. Similarly, PX-model are used for prior specification and posterior sampling, and 15,000 posterior samples of missing values, model parameters, root structures and latent factors are obtained via MCMC after a 5000 burn-in. First of all, the posterior marginal probability of having unit roots for the two latent factors are 83.24% and 0.17%, and the posterior joint probability of at least one of them have unit roots is 83.27%, indicating that we cannot reject that both of the time series are nonstationary. Secondly, the existence of cointegration is tested. The posterior probability that the rank of the factor loading matrix of non-stationary factors is smaller than the number of unit-root processes is 99.83%, indicating that cointegration exists at the 99% confidence level.

Given that the test shows that the two time series cointegrate, we specifically analyze the cointegration ratio (R) between the two time series. Posterior distribution of the cointegration ratio are derived from posterior samples and plotted in **Figure 3**, with posterior mean at -2.060 and 95% C.I. $[-2.130,-1.997]$, This is very much consistent with the true value -2 (obtained from model (11)).

To confirm that the right posterior distribution of the cointegration ratio is found, we specifically focused on

Table 1. Posterior summary of selected autoregression coefficients of latent factors, as compared to true values.

βs	Mean	Std	95% C.I.	True
$\beta_{1,1}$	0.9991	0.0179	[0.9533,1.0310]	1
$\beta_{1,2}$	0.0014	0.0216	[−0.0405,0.0641]	0
$\beta_{2,1}$	0.4543	0.0292	[0.3962,0.5164]	0.5
$\beta_{2,2}$	0.2826	0.0320	[0.2242,0.3468]	0.24

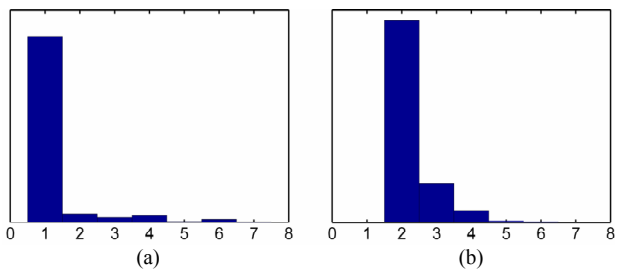

Figure 2. Histogram of posterior samples of autoregressive lag orders of latent factors η_{1t} and η_{2t}.

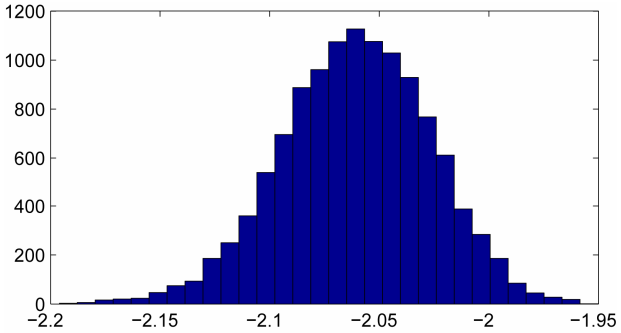

Figure 3. Histogram of posterior samples of cointegration ratio (*R*) shown.

the posterior mean $R = -2.060$, and let $z_t = y_{1t} + Ry_{2t}$. z_t is shown in **Figure 4** to illustrate the mean-reverting property, the stationarity of which is further confirmed by Augmented Dickey-Fuller (ADF) test [9], which rejects the null hypothesis that z_t has unit roots with all p-values < 0.01 for lag orders up to 10.

On the other hand, in a parallel study using the VECM model, two time series that are not co-integrated can be simulated from the following VAR model

$$y_t = \begin{pmatrix} 0.2 & -0.7 \\ 0.4 & 0.4 \end{pmatrix} y_{t-1} + \begin{pmatrix} 0.8 & 0.7 \\ -0.4 & 0.6 \end{pmatrix} y_{t-2} + \varepsilon_t \quad \text{where } \Sigma = \begin{pmatrix} 100 & 0 \\ 0 & 100 \end{pmatrix} \text{ and } y_0 = 0 \tag{12}$$

where the corresponding VECM representation is

$$\Delta y_t = -\begin{pmatrix} 0.8 & 0.7 \\ -0.4 & 0.6 \end{pmatrix} \Delta y_{t-1} + \varepsilon_t \tag{13}$$

We also tested whether our model and algorithms can correctly detect the non-existence of the cointegration relationship between the time series, among inference of other unknown quantities. In this study, we simulated the two-dimensional time series up to $T = 800$, fit the data with two-factor model, and set the maximum number of complex pairs to be 10 and that of real roots to be 15 for both latent factors, suggesting that the maximum autoregressive lag order is 35. Based on 15,000 MCMC posterior samples of root structures and factor loading matrix after a 5000 burn-in, the posterior probability that the rank of the factor loading matrix of non-stationary factors is smaller than the number of unit-root processes is 4.44%, indicating that there is no existence of cointegration between the two time series at the 95% confidence level. Secondly, the posterior marginal probabilities of having unit roots for the two latent factors are 83.94% and 95.56%, and the posterior joint probability of at least one of them having unit roots is 99.36%, which means that we cannot reject that y_{1t} is nonstationary, and y_{2t} is non-stationary at 99% confident level.

Therefore, our simulation studies clearly show the efficacy of the model proposed to both capture the true (non)stationary strutures of the multivariate time series and identify the cointegration relationship correctly.

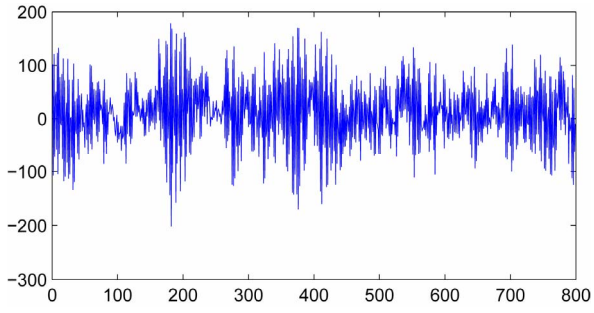

Figure 4. Plot of z_t to illustrate mean-reverting and stationarity. $z_t = y_{1t} + Ry_{2t}$, where R is the cointegration ratio, which is confirmed by formal ADF-test.

5. Real Data Analysis

5.1. GDX-GLD Pair vs PEP-KO Pair

Gold ETF GLD versus gold-miner ETF GDX have been reported as good candidates for real-world cointegrated pairs [10], because GLD reflects the spot price of gold, and GDX is a basket of gold-mining stocks. The cointegration relationship has been tested in empirical studies via frequentist cointegration test (e.g. Covariate-Augmented Dickey-Fuller (CADF) test), using the coefficient of the ordinary least squares regression as the cointegration ratio [11]. However, as mentioned earlier, most classical cointegration tests suffer from many problems, including multiple testing issues, how to choose lag order and whether deterministic terms (e.g. the intercept) should be included.

We applied our Bayesian cointegration analysis via dynamic factor models to the GDX and GLD daily closed price from Jun. 1, 2006 to Aug. 19, 2008 (T = 560). The original data (shown in **Figure 5**) is log-transformed, normalized and fitted with a two-factor dynamic factor model as (1), with priors placed on the corresponding PX-model. 15,000 posterior samples are obtained for each unknown quantity via MCMC and analyzed, after a 5000 burn-in.

Posterior samples show that the probability that the rank of the factor loading matrix of nonstationary factors is smaller than the number of unit-root process is 99.99%, indicating that based on this dataset the GDX and GLD

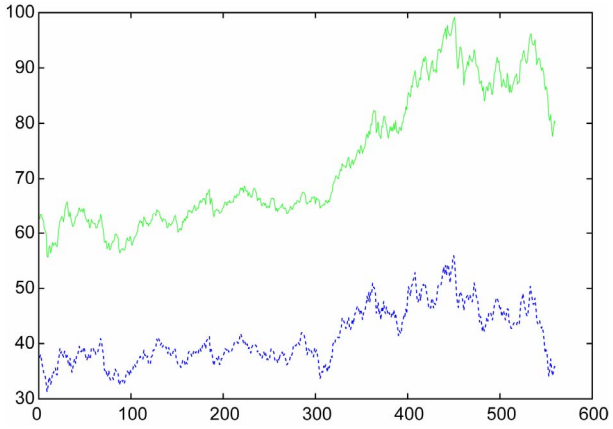

Figure 5. Original daily closed prices of GDX ETF (shown with the dotted line) and GLD ETF (shown with the solid line) from Jun. 1, 2006 to Aug. 19, 2008 are plotted. The original data is log-transformed and "normalized" before modeling.

cointegrate at the 99% confidence level. The marginal probability of the first latent factor having unit roots is 98.69%, and the joint probability of at least one factor having unit roots is 98.70%, showing that both time series are unit-root processes. Given that GDX and GLD cointegrate, the posterior samples of cointegration ratio (R) is calculated ($z_t = \mathrm{GDX}_t + R \times \mathrm{GLD}_t$ is stationary). R has the posterior mean of −0.8909 and 95% C.I. [−0.9290,−0.8524]. The histogram of R is shown in **Figure 6**.

To confirm that the right posterior distribution of R is found, we take the posterior mean $R = -0.8909$, let $z_t = \mathrm{GDX}_t + R \times \mathrm{GLD}_t$, and z_t is plotted in **Figure 7** to illustrate the stationarity and mean-reversion. The stationarity of z_t is further confirmed by testing the null hypothesis H_0 that the process z_t has unit roots, using the Augmented Dickey-Fuller (ADF) test [9], and the p-values are shown in **Table 2**.

Another candidate cointegrated pair could the stock price of PepsiCo Inc. (PEP) and Coca-Cola Co. (KO), because they operate in the same market and thus may be exposed to similar macro-economic and industry conditions. [11] specifically looked at this pair and showed

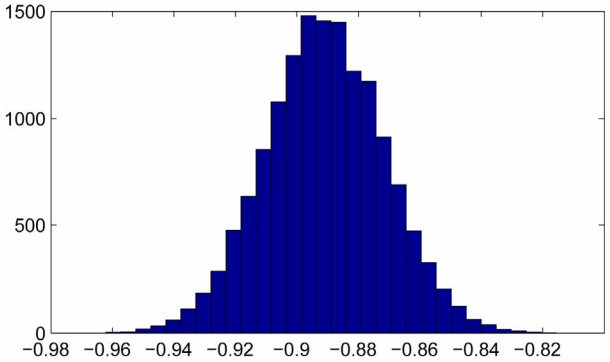

Figure 6. Histogram of posterior samples of cointegration ratio (*R*) between log-transformed GDX and GLD daily close price.

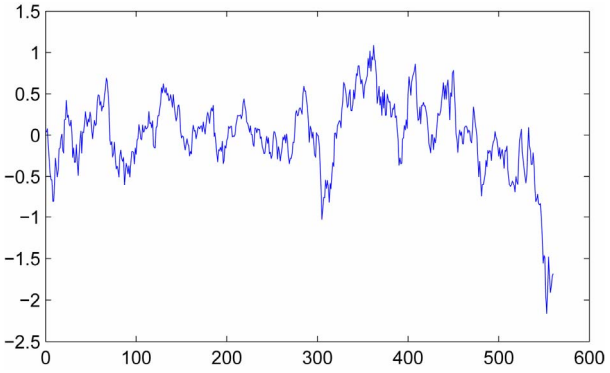

Figure 7. Plot of z_t to illustrate stationarity. z_t = GDX + *R* × GLD, where GDX and GLD are normalized log-transformed data, and R is the cointegration ratio.

Table 2. p-Values of the ADF test for H_0 with respect to different lag orders. H_0:z_t has unit roots.

Lag Orders:	1	2	3	4	5
p-values ($\times 10^{-2}$):	0.55	0.65	1.2	0.96	1.5

that they are indeed correlated $(\rho = 0.4849)$, but are not cointegrated, using the classical Augmented Dickey-Fuller (ADF) test. We also looked at the daily closed price of PEP and KO from Dec. 3, 1998 to Jan. 11, 2002 $(T = 780)$. The log transformed data are fitted with the two-factor model, setting C = 5 and R = 10 (maximum autoregressive lag order of latent factors is thus 20). 15,000 posterior samples are obtained after a 5000 burn-in, which shows that the probability that PEP and KO cointegrate is 1.6% and thus reject the existence of cointegration based on this dataset, which is also consistent to the finding of [11].

6. Conclusion

In this study, we showed the Bayesian factorized cointegration analysis for detecting cointegration among non-stationary time series. A strong message of this study is that, while testing cointegration of the original multivariate non-stationary time series via classical tests may have limitations both conceptually and computationally, studying the properties of independent latent factors in a Bayesian dynamic factor model framework serves as a powerful tool. As the majority of studies in the literature have used the traditional linear cointegrated error correction model, the models and methods we have described show accurate inference and efficient computation, and thus is a promising area for future research. One extension worth noting is to consider the more general case where the number of dynamic latent factors are unknown, in which the fundamental modeling and testing approaches discussed in this study are also applicable while some model selection criteria or penalization need to be further imposed to also incorporate the uncertainly of number of factors.

REFERENCES

[1] C. W. J. Granger and P. Newbold, "Spurious Regressions in Econometrics," *Journal of Econometrics*, Vol. 2, No. 2, 1974, pp. 111-120.

[2] R. F. Engle and C. W. J. Granger, "Co-Integration and Error Correction: Representation, Estimation, and Testing," *Econometrica*, Vol. 55, No. 2, 1987, pp. 251-276.

[3] S. Johansen, "Estimation and Hypothesis Testing of Cointegration Vectors in Gaussian Vector Autoregressive Models," *Econometrica*, Vol. 59, No. 6, 1991, pp. 1551-1580.

[4] M. West, "Time Series Decomposition," *Biometrika*, Vol.

84, No. 2, 1997, pp. 489-494.

[5] A. Gelman, "Prior Distributions for Variance Parameters in Hierarchical Models," *Bayesian Analysis*, Vol. 24, No. 1, 2006, pp. 1-19.

[6] G. Huerta and M. West. "Priors and Component Structures in Autoregressive Time Series Models," *Journal of the Royal Statistical Society Series B*, Vol. 61, No. 4, 1999, pp. 881-899.

[7] J. Ghosh and D. B. Dunson, "Default Prior Distributions and Efficient Posterior Computation in Bayesian Factor Analysis," *Journal of Computational and Graphical Statistics*, Vol. 18, No. 2, 2009, pp. 306-320.

[8] K. Cui and D. B. Dunson, "Generalized Dynamic Factor Models for Mixed-Measurement Time Series," *Journal of Computational and Graphical Statistics*, 2012.

[9] D. A. Dickey and W. A. Fuller, "Distribution of the Estimators for Autoregressive Time Series with a Unit Root," *Journal of the American Statistical Association*, Vol. 74, No. 366, 1979, pp. 427-431.

[10] K. Triantafyllopoulos and G. Montana, "Dynamic Modeling of Mean-Reverting Spreads for Statistical Arbitrage," *Computational Management Science*, Vol. 8, No. 1-2, 2011, pp. 23-49.

[11] E. P. Chan, "Quantitative Trading," John Wiley and Sons, Hoboken, 2008.

