6-2020

Essays on time series and financial econometrics

Yijie FEI
Singapore Management University, yijie.fei.2015@phdecons.smu.edu.sg

SINGAPORE MANAGEMENT UNIVERSITY

PHD DISSERTATION

Essays On Time Series and Financial Econometrics

Yiijie Fei

supervised by

Professor JUN YU

June 19, 2020

Essays On Time Series and Financial Econometrics

by

Yijie Fei

Submitted to School of Economics in partial fulfillment of

the requirements for the Degree of Doctor of Philosophy in Economics

Dissertation Committee:

Jun Yu (Supervisor/Chair)
Lee Kong Chian Professor of Economics and Finance
School of Economics and Lee Kong Chian School of Business
Singapore Management University

Peter C.B.Phillips
Yale University: Sterling Professor of Economics & Professor of Statistics
University of Auckland: Distinguished Professor
University of Southampton: Adjunct Professor of Economics
Singapore Management University: Distinguished Term Professor of Economics

Yichong Zhang
Assistant Professor of Economics
Singapore Management University

Xiaohu Wang
Assistant Professor
Department of Economics
The Chinese University of Hong Kong

Singapore Management University

2020

Abstract

In this dissertation, we make a few contributions to the literature of time series and financial econometrics.

In the second chapter, some asymptotic results are given for first-order autoregressive (AR(1)) time series with two features: (i). a nonzero constant intercept (ii). a root moderately deviating from unity. Both stationary and explosive sides are studied. It is shown that the inclusion of intercept will change drastically the large sample properties of the least-squares (LS) estimator obtained in Phillips and Magdalinos (2007, PM hereafter). For the near-stationary case, only an unusual convergence of a linear combination of intercept and AR coefficient can be derived. For the near-explosive case, on the other hand, the limiting distributions of two estimators will be independent and Gaussian, with the conventional t-test for both of them keeping valid. Due to its relevance in bubble detection, we also briefly discuss the limiting theory under the shrinking drift assumption for mildly explosive case. Empirical implication of the limit theory is also discussed.

The third chapter is concerned with the joint test of predictability and stability. The null hypothesis under investigation is that the potential predictors exhibit no predictability and meanwhile no structural break occurs during the sample period. We first show that the recently proposed IVX estimator provides better inference for the structural break of predictability than OLS. Based on this finding, we consider a new test combining IVX with the sup-Wald statistic. The limiting distribution of the test statistic is derived and shown to vary with the level of predictors' persistence. Though theoretically not applicable under very strong persistence, simulation results suggest that this test enjoys good finite sample size and power properties even in that case. Such robustness is informally explained through a simple simulation investigation. An application to the US stock return data is used to illustrate its empirical relevance.

The fourth chapter considers the potential impacts of predicted variable's level shifts on testing for predictive power, especially when persistent predictors are used. It is shown that the limiting distribution of the usual t-statistic will depend on the magnitude of break

size. In particular, if the shift is "large", it will dominate asymptotically and t-statistic will blow up, while if it is "moderate", it will coexist with the original limiting component. Moreover, applying IVX will not fully fix this problem. In all cases, a spurious relationship will be detected if the breaks are not taken into account. To alleviate this problem, we propose to base the inference on a sample-splitting procedure. We further discuss the finite sample issues caused by moderate shifts and propose a partial solution based on simulation results. Empirical applications to the prediction of stock return volatility and housing price index are provided.

In the last chapter, we make two contributions to the literature of volatility modeling. First, we consider a new multivariate stochastic volatility (MSV) model, applying a recently proposed novel parameterization of the correlation matrix. This modeling design is a generalization of Fisher's z-transformation to higher-dimensional cases and it is fully flexible as the validity of the resultant correlation matrix is guaranteed automatically, which allows us to separate the driving factors of volatilities and correlations. Second, we propose to use a different estimation tool. Like most existing literature on MSV, we work within a Bayesian framework and hence rely on Markov Chain Monte Carlo (MCMC) tool. However, when dealing with latent variables, conventional single-move or multi-move sampler is replaced by a novel technique called Particle Gibbs Ancestor Sampling (PGAS), which is built upon Sequential Monte Carlo (SMC) method. Extensive simulation studies are conducted to confirm the applicability of this method under the current setup and provide some guidance on the trade-off between estimation accuracy and computational cost. The new model is then implemented using two financial data sets and the comparison with existing models is discussed.

Contents

List of Figures

List of Tables

Acknowledgment

I would like to sincerely thank my supervisor, Professor Jun Yu, for his excellent guidance, support, patience, and maintaining a good research environment. I appreciate all his time and helpful discussion, which stimulates most of my research ideas. It would never have been possible for me to complete my work without his help and inspiration. My deepest gratitude also goes to Professor Yichong Zhang and Professor Phillips, as well as all the participants of the Econometrics workshop at SMU, from whom I learn a lot.

I also wish to express my thankfulness to my wife, Hu Die , for her love and encouragement. She has always been supportive during my long journey in pursuit of a Ph.D. degree, both a good time and a tough time. My thanks go to my parents and parents-in-law as well for their longstanding support and regretless giving. Last but not least, I dedicate this dissertation to my grandmother, who is not able to witness my graduation but watch me in heaven.

1 Overview

Many financial and macroeconomic time series exhibit non-stationarity. It is of paramount importance to understand the property of processes with such features. In next chapter, we consider a special nonstationary model, called mildly integrated autoregression. This model has been widely studied due to its relevance in bubble detection. Specifically, in Chapter 2 [1], we explore the consequences of a nonzero intercept on the least-squares estimator of model parameters. Both stationary and explosive cases are considered. We show that the presence of a constant intercept will dramatically alter the large sample properties derived in Phillips and Magdalinos (2007). For completeness, a brief discussion about the shrinking drift assumption is also included. We provide some intuition of these results and discuss their empirical implications as well.

Nonstationarity will also complicate the inference theory of predictive regressions. What makes things even worse is the presence of structural change. In the following two chapters, we consider problems related to the potential existence of structural breaks in a predictive regression model, with a focus on the recently popular methodology called IVX[2]. We provide a simple simulation as motivation for our study, which indicates that structural breaks in both intercept and slope coefficients hurt the test for the predictability. Specifically, when intercept experiences shifts, "spurious predictability" will be detected by a standard IVX procedure. Breaks in the slope coefficients, on the other hand, may result in a loss of power. In light of these deficiencies, we consider the following two issues.

The first issue we want to address is the joint test of predictability and stability. The null hypothesis under investigation is that the potential predictors exhibit no predictability, and meanwhile, no structural break occurs during the sample period. In Chapter 3, We first show that the recently proposed IVX estimator provides better inference for the structural break in predictability than OLS. Based on this finding, we consider a new test combining IVX with the well-known sup-Wald statistic. The limiting distribution of the

[1] This chapter is based on Fei, Yijie. "Limit theory for the mildly integrated process with intercept." Economics Letters 163 (2018): 98-101.

[2] See Chapter 3 for a brief review of IVX.

test statistic is derived and shown to vary with the degree of persistence in predictors. Though theoretically not applicable under extreme persistence, simulation results suggest that this test enjoys good finite sample size and power properties even in that case. Such robustness is explained through a simple simulation investigation.

The second issue we discuss is the potential impact of level shifts in the predicted variable on the predictability test when highly persistent predictors are used. It is shown in Chapter 4 that the limiting distribution of conventional t-statistic will depend on the magnitude of break size. In particular, if the break is "large", it will dominate asymptotically, and the t-statistic will blow up, while if it is "moderate", it will coexist with the original limiting component. Moreover, applying IVX will not fully fix this problem. In all cases, a spurious relationship will be detected if the breaks are ignored. To alleviate this problem, we propose to base the inference on a sample-splitting procedure. We further discuss the issues caused by moderate breaks and propose a partial solution based on simulation results.

Another topic in financial econometrics that has attracted long-standing attention is the modeling of dynamic volatility and correlation of asset returns. In the last chapter,[3], we intend to make two contributions to this already fruitful literature. First, we propose a novel multivariate stochastic volatility (MSV) model, which is based upon a new parameterization of the correlation matrix. It is a generalization of the well-known Fisher z-transformation to higher-dimensional cases, and it is sufficiently flexible as the validity of the resulting correlation matrix is guaranteed automatically. The new transformation allows us to model the driving factors of volatilities and correlations separately. Second, we propose to use a different estimation tool. Traditional single-move or multi-move sampler is replaced by a novel technique called Particle Gibbs Ancestor Sampling (PGAS), which is built upon Sequential Monte Carlo (SMC) method. Extensive simulation studies are conducted to confirm the applicability of this method under the current setup and provide some guidance on the trade-off between estimation accuracy and computational cost. The new model is then implemented using two financial data sets, and the comparison with existing models is discussed.

[3]This chapter is co-authored with Han Chen.

2 Limit Theory for Mildly Integrated Process with Intercept

2.1 Introduction

The past several decades have witnessed an enormous amount of effort in econometrics devoted to the autoregressive (AR) time series model. This simple yet powerful model has also been widely used in empirical researches. In the present chapter, we focus our attention on a particular case of first-order autoregression defined by

$$y_t = d + \rho_n y_{t-1} + u_t,\ t = 1, ..., n. \tag{2.1}$$

Throughout this chapter, we impose following assumptions on model (2.1).

Assumption 2.1 *For model (2.1) above, we assume that,*
(i). AR coefficient $\rho_n = 1 + \frac{c}{k_n}$ *depends on sample size* n *and* $k_n = o(n)$ *as* $n \to \infty$.
(ii). c *and* d *are both nonzero constant real numbers.*
(iii). The process is initialized at $y_0 = o_p(\sqrt{k_n})$ *independent of* $\sigma(u_1, ..., u_n)$.
(iv). u_t *is a sequence of independent and identically distributed (i.i.d.) random disturbances with* $E(u_1) = 0$, $E(u_1^2) = \sigma^2 \in (0, \infty)$. *When* $c < 0$, *we further assume that* $E|u_1|^{2+\delta} < \infty$ *for some* $\delta > 0$.

When $d = 0$, this model is rigorously studied in PM and often known as mildly integrated process in the literature . Specifically, when $c < 0$, it is termed mildly stationary process and when $c > 0$, mildly explosive process. Except for the nonzero intercept, all other assumptions made above are the same as in PM. [4] Assume that a set of observations $\{y_t\}_{t=1}^n$ is available. Let $\sum$ denote $\sum_{t=1}^n$ to simplify notation. When d is known to be zero a priori, the least-squares estimator of ρ_n based on the available sample is

$$\hat{\rho}_n = \frac{\sum y_t y_{t-1}}{\sum y_{t-1}^2} = \rho_n + \frac{\sum y_{t-1} u_t}{\sum y_{t-1}^2}$$

[4] For mildly explosive case, the proof suggests that we can relax the restriction on initial condition to $y_0 = o_p(k_n)$, without changing any results that will be obtained.

In this case, the limiting distribution of $\hat{\rho}_n$ is developed in PM. Specifically, they showed that when $c < 0$,

$$(nk_n)^{1/2}(\hat{\rho}_n - \rho_n) \Rightarrow N(0, -2c),$$

and when $c > 0$,

$$(k_n\rho_n^n/2c)(\hat{\rho}_n - \rho_n) \Rightarrow C,$$

where C is a random variable following standard Cauchy distribution. As is discussed in PM, the main attraction of this setup is that it provides a bridge between pure stationary or explosive process and unit-root (or local-to-unit-root) process.

If the value of d is unknown ex ante, however, the LS estimators of ρ and d will be, respectively,

$$\hat{\rho}_n = \frac{\sum(y_t - \bar{y})(y_{t-1} - \bar{y}_-)}{\sum(y_{t-1} - \bar{y}_-)^2} \text{ and } \hat{d} = \bar{y} - \hat{\rho}\bar{y}_-. \tag{2.2}$$

where $\bar{y} = \sum y_t/n$ and $\bar{y}_- = \sum y_{t-1}/n$. The potential impact of estimating the intercept, together with the AR coefficient on limiting behavior of LS estimator, is by far not well studied in the literature. Motivated by this incompleteness in theory, in this chapter, we extend the literature by developing the asymptotic distributions of $\hat{\rho}_n$ and $\hat{d}$ for the moderately integrated process with an unknown intercept. We first show that an unusual joint convergence result can be achieved for mildly stationary case. Then, the asymptotic normality, instead of Cauchy-type distribution, is proved for both intercept and AR coefficient under the mildly explosive assumption. All proofs could be found in Appendix A.

2.2 Models and Main Results

As is similar to the equation (2.1) in Wang and Yu (2015), an equivalent representation of y_t generated by model (1) is

$$y_t = \frac{d}{c}k_n(\rho_n^t - 1) + \rho_n^t y_0 + \sum_{j=1}^{t} \rho_n^{t-j} u_j, \tag{2.3}$$

where $\rho_n = 1 + \frac{c}{k_n}$ is utilized. This expression can be written more concisely as

$$y_t = \frac{d}{c} k_n(\rho_n^t - 1) + y_t^0, \tag{2.4}$$

where y_t^0 is a mildly stationary or mildly explosive process (depending on the sign of c) without intercept. Apparently, $\{y_t^0\}$ so defined is equivalent to the $\{y_t\}$ studied in PM. Decomposing y_t into these two components is helpful in terms of the derivation of asymptotic behaviors, because we can directly borrow some results for y_t^0 from that paper. In the following, we will discuss the $c < 0$ case first, and then move to the $c > 0$ part.

2.2.1 Limit Theory for Mildly Stationary Case

This subsection establish the limit properties of $\hat{\rho}_n$ and $\hat{d}$ when $c < 0$. First, we derive the limiting behavior of some components that will be involved in the LS estimators for both intercept and AR coefficient.

Theorem 2.1 *For model (1) with $c < 0$, we have, as $n \to \infty$,*
(a). $y_n = o_p(n)$*;*
(b). $n^{-1/2} \sum u_t \Rightarrow Z$*, where* $Z \sim N(0, \sigma^2)$*;*
(c). $n^{-1} {k_n}^{-1} \sum y_{t-1} \Rightarrow -d/c$*;*
(d). $n^{-1/2} {k_n}^{-1} \sum y_{t-1} u_t \Rightarrow (-d/c)Z$*;*
(e). $n^{-1} {k_n}^{-2} \sum y_{t-1}^2 \Rightarrow d^2/c^2$.

Note that the centered LS estimators of $\hat{\rho}_n$ and $\hat{d}$ are given by

$$\begin{bmatrix} \hat{d} - d \\ \hat{\rho}_n - \rho_n \end{bmatrix} = \begin{bmatrix} n & \sum y_{t-1} \\ \sum y_{t-1} & \sum y_{t-1}^2 \end{bmatrix}^{-1} \begin{bmatrix} \sum u_t \\ \sum y_{t-1} u_t \end{bmatrix}. \tag{2.5}$$

Therefore, after some manipulations of this equation and use the convergence results re-

ported in Theorem 2.1, one can easily show that

$$\begin{aligned}
&\begin{bmatrix} \sqrt{n}(\hat{d}-d) \\ \sqrt{n}k_n(\hat{\rho}_n-\rho_n) \end{bmatrix} \\
&= \begin{bmatrix} 1 & n^{-1}k_n^{-1}\sum y_{t-1} \\ n^{-1}k_n^{-1}\sum y_{t-1} & n^{-1}k_n^{-2}\sum y_{t-1}^2 \end{bmatrix}^{-1} \begin{bmatrix} n^{-1/2}\sum u_t \\ n^{-1/2}k_n^{-1}\sum y_{t-1}u_t \end{bmatrix} \\
&\Rightarrow \begin{bmatrix} 1 & -d/c \\ -d/c & d^2/c^2 \end{bmatrix}^{-1} \begin{bmatrix} 1 \\ -d/c \end{bmatrix} Z.
\end{aligned}$$

It is obvious that the first matrix in the second line has a zero determinant and is thus not invertible. If we pre-multiply this singular matrix on both sides of (6), we immediately achieve the following result.

Theorem 2.2 *For model (2.1) with $c < 0$, the following joint limit applies as $n \to \infty$,*

$$\sqrt{n}(\hat{d}-d) + \frac{d}{-c}\sqrt{n}k_n(\hat{\rho}_n-\rho_n) \Rightarrow Z.$$

Remark 2.1 This theorem suggests that a linear combination of $\hat{d}$ and $\hat{\rho}_n$ converges to normal distribution. Such a result, however, is not enough if we want to make inference about these two parameters, which relies on the individual limiting distribution of them.[5]

Remark 2.2 The joint convergence applies because drift and AR part are multicollinear. This leads to a singular signal matrix. Utilizing a transformation of regressor, Liu and Peng (2019) derive the convergence rate and limiting distribution of each estimator. They make additional assumption that

$$\lim_{n\to\infty} \frac{\sqrt{n}}{\max(\sqrt{n}, k_n)} = h_1 \in [0,1]$$

and

$$\lim_{n\to\infty} \frac{k_n}{\max(\sqrt{n}, k_n)} = h_2 \in [0,1].$$

[5]The Remark 2.3 in the published version of this paper, Fei (2018), is wrong as pointed out by Liu and Peng (2019) in their Remark 1.

Under this assumption, they manage to show that

$$\begin{pmatrix} \frac{\max(\sqrt{n},k_n)}{\sqrt{k_n}}(\hat{d}-d) \\ \sqrt{k_n}\max(\sqrt{n},k_n)(\hat{\rho}_n-\rho_n) \end{pmatrix} \to^d \begin{pmatrix} \frac{d^2h_2c^{-2}Z_1+dh_1c^{-1}Z_2}{d^2h_2^2/(-2c^3)+\sigma^2h_1^2/(-2c)} \\ \frac{dh_2c^{-1}Z_1+h_1Z_2}{d^2h_2^2/(-2c^3)+\sigma^2h_1^2/(-2c)} \end{pmatrix}, \tag{2.6}$$

where Z_1 and Z_2 are independent normal random variables.

2.2.2 Limit Theory for Mildly Explosive Case

This subsection considers the asymptotic behavior of $\hat{\rho}_n$ and $\hat{d}$ when $c > 0$. Mildly explosive time series has turned out to be important in the econometric analysis of bubbles; see for example, Phillips et al. (2015a, 2015b). Following PM, we define

$$X_n := \frac{1}{\sqrt{k_n}}\sum_{t=1}^{n}\rho_n^{-(n-t)-1}u_t \quad \text{and} \quad Y_n := \frac{1}{\sqrt{k_n}}\sum_{j=1}^{n}\rho_n^{-j}u_j.$$

To obtain the asymptotic distribution of the LS estimator without intercept, PM proved following results,[6] which are reproduced here because they will be used in the proof of limit theory under nonzero intercept later.

Lemma 2.1 *For each $c > 0$, the sequences $(X_n)_{n\in N}$ and $(Y_n)_{n\in N}$ defined above satisfy,*
(a) $(X_n, Y_n) \Rightarrow (X, Y)$ as $n \to \infty$, where X and Y are independent $N(0, \sigma^2/2c)$ random variables.[7]
(b) As $n \to \infty$, we have $\rho_n^{-n}/k_n \sum y_{t-1}^0 u_t \Rightarrow XY$.

Taking advantage of these properties, we have following theorem which provides the limiting behavior we will need.

Theorem 2.3 *For model (2.1) with $c > 0$, we have, as $n \to \infty$,*
(a) $\rho_n^{-n}k_n^{-1}y_n \Rightarrow d/c$;
(b) $\rho_n^{-n}k_n^{-2}\sum y_{t-1} \Rightarrow d/c^2$;
(c) $\rho_n^{-n}k_n^{-3/2}\sum y_{t-1}u_t \Rightarrow (d/c)X$;

[6] See Lemma 4.2 and part (a) of Theorem 4.3 in that paper.
[7] As pointed out by the referee, this result is no longer necessary under the present setup, as we show that the normalized $\sum_{t=1}^{n} y_{t-1}^2$ will approach a constant rather than a random variable

(d) $(\rho_n^2-1)\rho_n^{-2n}k_n^{-2}\sum y_{t-1}^2 \Rightarrow d^2/c^2$.

Similar to the discussion in the last subsection, we need to examine the centered LS estimators of $\hat{\rho}_n$ and $\hat{d}$. Motivated by the results in Theorem 2.3, we can transform equation (2.5) into following representation

$$\begin{aligned}
&\begin{bmatrix} n^{1/2}(\hat{d}-d) \\ \rho_n^n k_n^{3/2}/2c(\hat{\rho}_n-\rho_n) \end{bmatrix} \\
&= \begin{bmatrix} 1 & \rho_n^{-n}n^{-1/2}k_n^{-3/2}\sum y_{t-1} \\ \rho_n^{-n}n^{-1/2}k_n^{-3/2}\sum y_{t-1} & (\rho_n^2-1)\rho_n^{-2n}k_n^{-2}\sum y_{t-1}^2 \end{bmatrix}^{-1} \\
&\quad \begin{bmatrix} n^{-1/2}\sum u_t \\ \rho_n^{-n}k_n^{-3/2}\sum y_{t-1}u_t \end{bmatrix} \\
&= \begin{bmatrix} 1 & o_p(1) \\ o_p(1) & (\rho_n^2-1)\rho_n^{-2n}k_n^{-2}\sum y_{t-1}^2 \end{bmatrix}^{-1} \\
&\quad \begin{bmatrix} n^{-1/2}\sum u_t \\ \rho_n^{-n}k_n^{-3/2}\sum y_{t-1}u_t \end{bmatrix} \\
&\Rightarrow \begin{bmatrix} 1 & o_p(1) \\ o_p(1) & d^2/c^2 \end{bmatrix}^{-1} \begin{bmatrix} Z \\ d/cX \end{bmatrix}.
\end{aligned} \tag{2.7}$$

Consequently, we obtain the following limiting distributions which extend PM's results to the mildly explosive process with nonzero intercept.

Theorem 2.4 *For model (2.1) with $c>0$, the following limits apply as $n\to\infty$,*
(a) $\sqrt{n}(\hat{d}-d)\Rightarrow N(0,\sigma^2)$*;*
(b) $\rho_n^n(\rho_n-1)^{-3/2}(\hat{\rho}_n-\rho_n)\Rightarrow N(0,2\sigma^2/d^2)$.

Remark 2.3 Compared with PM's result, the limiting behavior of $\hat{\rho}_n$ when intercept is unknown and nonzero is quite different.The convergence rate now is $\rho_n^n(\rho_n-1)^{-3/2}$, which is faster than the $\rho_n^n(\rho_n-1)^{-1/2}$ in PM. Meanwhile, the asymptotic distribution is Gaussian under the present model, while it is Cauchy in PM.

Remark 2.4 Note that when $d=0$, the asymptotic variance will be infinite and thus the

above theorem can not be applied. In that case, the Cauchy-type convergence derived in PM should be used.

Remark 2.5 Wang and Yu (2015) shows that, for pure explosive process (i.e. $k_n = 1$), when intercept and AR coefficient are estimated together, there exists invariance principle for $\hat{d}$. Their limiting distribution for $\hat{d}$ is exactly the same as reported in above theorem. However, they show that, without assuming Gaussianity, no invariance result can be achieved for $\hat{\rho}_n$, which is not the case when explosiveness is mild. This is not surprising, as a critical advantage of mildly explosive process over pure explosive one is that the former allows for invariance principle of LS estimator without imposing the assumption of normal errors.

Now that both estimators show asymptotic normality, we would like to further study the t-statistic of these two LS estimators. Since both $\hat{d}$ and $\hat{\rho}_n$ are consistent, we can consistently estimate the variance parameter σ^2 as well. That is, if we let $\hat{\sigma}^2 = n^{-1}\sum(y_t - \hat{d} - \hat{\rho}_n y_{t-1})$, then $\hat{\sigma}^2 \to^p \sigma^2$. With $\hat{\sigma}^2$ in hand, one can further construct the traditional t-statistic

$$t_d = \frac{(\hat{d} - d)[n\sum y_{t-1}^2 - (\sum y_{t-1})^2]^{1/2}}{[\sum y_{t-1}^2 \cdot \hat{\sigma}^2]^{1/2}}, \tag{2.8}$$

$$t_{\rho_n} = \frac{(\hat{\rho}_n - \rho_n)[n\sum y_{t-1}^2 - (\sum y_{t-1})^2]^{1/2}}{[n \cdot \hat{\sigma}^2]^{1/2}}. \tag{2.9}$$

It is not hard to prove the following corollary using the findings reported above.

Corollary 2.1 *For the t-statistics defined by equation (6) and (7), we have, as $n \to \infty$,*
(a) $t_d \Rightarrow N(0,1)$,
(b) $t_{\rho_n} \Rightarrow N(0,1)$,
(c) t_d and t_{ρ_n} are asymptotically independent.

Therefore, the conventional t-test remains valid for both parameters under this scenario. Similarly, the F-statistic for joint hypothesis of two parameters will have the same asymptotic distribution as in stationary case.

2.2.3 Mildly Explosive Case with Shrinking Drift[8]

As shown in the last subsection, assuming a constant intercept will make the limiting theory dominated by this drift term and thus in the literature of bubble detection, people usually consider a shrinking drift, which is asymptotically negligible. Here, we consider a case in which the intercept will coexist with the autoregressive component in the limit. To take into account the asymptotic effect of the initial value, we also assume it diverges at a faster rate. Specifically, suppose now that $d_n = \delta/\sqrt{k_n}$ and $y_{0n} = y_0\sqrt{k_n}$, where δ and y_0 are constants and we use subscript n to emphasize that both intercept and initial value depend on sample size. Then it is easy to show that

$$\frac{1}{\rho_n^n\sqrt{k_n}}y_n \to^d Y + \frac{\delta}{c} + y_0,$$

$$\frac{1}{\rho_n^n k_n}\sum y_{t-1}u_t \to^d X(Y + \frac{\delta}{c} + y_0),$$

$$\frac{\rho_n^2 - 1}{\rho_n^{2n} k_n}\sum y_{t-1}^2 \to^d (Y + \frac{\delta}{c} + y_0)^2,$$

where X and Y are two independent normal random variables. The proof of these results is omitted as it involves only simple algebra. Since we have

$$\begin{pmatrix} \sqrt{n}(\hat{d}_n - d_n) \\ \frac{\rho_n^n}{\rho_n^2-1}(\hat{\rho}_n - \rho_n) \end{pmatrix} = \begin{pmatrix} 1 & o_p(1) \\ o_p(1) & \frac{\rho_n^2-1}{\rho_n^{2n}k_n}\sum y_{t-1}^2 \end{pmatrix}^{-1} \begin{pmatrix} \frac{1}{\sqrt{n}}\sum u_t \\ \frac{1}{\rho_n^n k_n}\sum y_{t-1}u_t \end{pmatrix},$$

it follows that

$$\sqrt{n}(\hat{d}_n - d_n) \to^d N(0, \sigma^2),$$

$$\frac{\rho_n^n}{\rho_n^2 - 1}(\hat{\rho}_n - \rho_n) \to^d \frac{X}{Y + \frac{\delta}{c} + y_0}.$$

Under this shrinking drift assumption, both the intercept and the initial value enter the limiting distribution of the estimated slope coefficient, and it deviates from the standard Cauchy limit. On the other hand, the estimated intercept still has conventional Gaussian density. This is a special case of the model considered in Wang and Yu (2016), although we consider the conventional long-span paradigm while their motivation is from the dou-

[8] This part is not in the version published in Economicc Letters.

ble asymptotics.

2.3 Some Intuitions

The results reported in the above two subsections, though new, are not unexpected to a certain extent[9]. Note that in both cases we have considered, the convergence rate of the autoregressive coefficient is higher compared with PM. In time series models, it is not uncommon that including an intercept term would lead to an increase in the convergence rate of a persistent process. A famous example is the unit root model with a drift. Without a constant term, the convergence rate of the autoregressive parameter is T, while after including a drift, it increases to $T^{3/2}$. Hamilton (1994, P.407) pointed out that it is due to the fact "the regressor y_{t-1} is asymptotically dominated by the time trend. In large samples, it is as if the explanatory variable y_{t-1} were replaced by the time trend." In the present setup, the drift exclusively dominates the asymptotics as well. One can easily understand this by looking at the equation (4). If we assume $y_0 = 0$ for simplicity, it is clear that

$$\text{when } c < 0,\ E|y_t^0| = O(k_n^{1/2}) \text{ and } (d/c)k_n(\rho_n^t - 1) = O(k_n),$$

$$\text{when } c > 0,\ E|y_t^0| = O(\rho_n^t k_n^{1/2}) \text{ and } (d/c)k_n(\rho_n^t - 1) = O(\rho_n^t k_n),$$

uniformly in $t \in \{1, ..., n\}$. Indeed, in the large sample cases, the impact of the autoregressive part will gradually be dominated in the presence of the intercept term. Based on this result, we can straightforwardly obtain the following approximation of y_t uniformly in t:

$$y_t = \frac{d}{c}k_n(\rho_n^t - 1)\{1 + o_p(1)\} = \{1 + o_p(1)\} \times \begin{cases} \frac{d}{-c}k_n & c < 0 \\ \frac{d}{c}k_n\rho_n^t & c > 0 \end{cases}.$$

This property, then, immediately leads to the results in Theorem 2.2 and 2.4 above. It also explains why in our explosive case, different from the standard mildly explosive model, we can achieve some "ergodicity" that the normalized $\sum_{t=1}^n y_{t-1}^2$ converges in probability

[9] I am grateful to the referee for raising the issues in this part.

to a constant instead of a random variable. In that sense, the inclusion of intercept actually changes the nature of the model (1) because, in terms of the limit theory, it is no longer like a mildly integrated process.

2.4 Empirical Implication

Above theoretical derivation suggests that, under the mildly explosive assumption, an invariance principle exists for $\hat{d}$, and its conventional t-statistic still holds. Hence, based on available observations, we can consistently estimate the intercept and then test whether it is zero or not. This finding is important because, as can be seen from Remark 2.5, the limiting behavior of $\hat{\rho}_n$ is dramatically influenced by the appearance of intercept. Both the convergence rate and asymptotic distribution change. Therefore, before making inference about $\hat{\rho}_n$, one should first choose the proper limit theory according to the test result for $H_0 : d = 0$ vs. $H_A : d \neq 0$. Since asymptotically $\hat{d}$ and $\hat{\rho}_n$ are independent, this can be easily implemented by checking either the t-statistic or p-value of $\hat{d}$ after running an OLS regression of y_t on an constant and y_{t-1}.

The above-mentioned issue is especially relevant in the empirical study of various types of financial bubbles. As already mentioned, the mildly explosive process is prevalent in the modeling of bubble behavior. Usually, a data-generating process like

$$\begin{aligned} x_t = {} & x_{t-1}1\{t < \tau_e\} + \rho_n x_{t-1}1\{\tau_e \leqslant t \leqslant \tau_f\} \\ & + \left(\sum_{k=\tau_f+1}^{t} \epsilon_k + x^*_{\tau_f} \right) 1\{t > \tau_f\} + \epsilon_t 1\{t \leqslant \tau_f\}, \end{aligned} \tag{2.10}$$

where $\rho_n = 1 + \frac{c}{n^\alpha}$, $c > 0$, $\alpha \in (0, 1)$, will be used to describe the trajectory of a time series that is likely to experience explosive bubble period. Phillips et al. (2011) propose a recursive right-tailed unit root testing procedure to test explosive behavior and date stamp the origination and collapse of economic exuberance. They suggest that, if one rejects the null hypothesis that there is no explosive behavior, a valid asymptotic confidence interval for ρ_n could be constructed using PM's results, viz., a Cauchy-type distribution[10]. This

[10] See section 3.1 in that paper for more detailed discussion.

advice, which is based on (2.10), however, implicitly assumes that the bubble follows a mildly explosive process with an intercept known to be zero. As shown in Theorem 2.4, when the intercept is actually nonzero, the asymptotic distribution of estimator for the AR coefficient should be normal, and thus critical value based on standard normal distribution should be applied. Specifically, if a preliminary test on intercept term shows that we have $d \neq 0$, then a $100(1-\alpha)\%$ confidence interval for ρ_n, should be given by the region

$$\left(\hat{\rho}_n \pm \frac{\sqrt{2}\hat{\sigma}(\hat{\rho}_n - 1)^{3/2}}{\hat{d}\hat{\rho}_n^n} Z_\alpha \right),$$

where Z_α is the two-sided α percentile critical value of the standard normal distribution.

2.5 Conclusions

In this chapter, the limit theory for processes with a root moderately deviating from unity and a nonzero intercept is established. The results make several contributions to the literature. (i) We show that, with the mildly stationary assumption, as long as we allow a nonzero intercept, an unconventional joint convergence of estimators for intercept and AR coefficient will apply. (ii) It is shown that for the mildly explosive process with constant intercept, the LS estimator for the AR coefficient is not asymptotically Cauchy anymore. It will become Gaussian, and the convergence rate is faster than the no-intercept case. (iii) Based on the validity of t-tests for both estimators, suggestions are made regarding the inference for financial bubbles. In this research, however, only the i.i.d. error term is studied, and meanwhile, the deterministic time trend is excluded. How to incorporate a dependent error structure and time trend is beyond the scope of the present chapter.

3 Robust Joint Test Against Predictability and Structural Break

3.1 Introduction

Predictive regression is a popular research tool in empirical finance and macroeconomics. Potential usage of it includes mutual fund performance evaluation, tests of conditional CAPM, and optimal asset allocation; see Paye and Timmermann (2006) and reference therein. Concerning the stock market return, a large number of predictors have been uncovered in past literature, among which are dividend yield, term spread, default premia, and market sentiment, to name just a few.

In the early days, the routine procedure that an empiricist would apply was to assume a stable prediction relationship and regress target variables, such as equity index returns, on the lagged value of potential predictors and then check the standard t-statistic for significance. This common practice, however, is shown to be flawed in many aspects. Stambaugh (1999) first observed the finite sample bias in the OLS estimator, which is now known as Stambaugh bias, caused by the autoregressive property of predictors and proposed a first-order bias-corrected estimator. Based on this insight, Amihud and Hurvich (2004) further considered the second-order bias-correction and refined Stambaugh's estimator. Note that this problem is a finite sample issue that will eventually vanish when the sample size goes toward infinity. Under the assumption that predictors are stationary, the asymptotic property of the OLS estimator remains valid.

The stationarity assumption that underlies the asymptotic normality of the OLS estimator, unfortunately, appears to be implausible, given the strong persistence observed for most predictors. A lot of work has been dedicated to the non-standard inference for nonstationary predictors, among which the most notable contributions include: (i). the Bonferroni-type approach, as in Cavanagh et al. (1995) and Campbell and Yogo (2006), (ii). conditional likelihood method based on sufficient statistic, as in Jansson and Moreira (2006), and (iii). control function method proposed in Elliott (2011). All these tests, however, bear some undesirable properties that reduce the attraction to empiricists. For

instance, the first approach is shown to become invalid when the predictor is very far away from the unit root process. The first and second approaches are hard to be extended to the multivariate case. The validity of the third approach relies critically on the availability of a perfect orthogonalizing variable.

A novel approach that recently attracts much attention is the instrumental variable-based robust test proposed in Kostakis et al. (2015, KMS hereafter), which is built upon the theoretical advances in Phillips and Magdalinos (2009). The new method, which they call the IVX-Wald test, provides a new limit theory that keeps valid in predictive regression with predictors exhibiting very general characteristics by successfully removing the endogeneity. The limiting distribution of the IVX estimator is proven to be pivotal and (mixed) normal under all possible degrees of persistence. Hence, a self-normalized Wald statistic can be constructed that converges in distribution to a χ^2 random variable. In addition to its robustness against the persistence of predictors, the IVX approach enjoys other important advantages. For instance, it is very straightforward to implement and can be extended to multivariate predictive regressions easily. In fact, it can be even used to deliver a joint test for a system of multiple predictive regressions.

All the papers mentioned above maintain the assumption that predictive relationship is stable over the sample period, which could be as long as ninety years in many applications. It seems implausible to argue that no structural break happens in an extended period. Indeed, there is a growing body of literature that focuses on the instability of predictive regressions and suggests that predictability is time-varying. Paye and Timmermann (2006) applied a series of techniques developed in Bai (1997) and Bai and Perron (1998) to test the presence of structural break and estimate the break date if it exists. They also considered the fixed regressor bootstrap method proposed in Hansen (2000) and the optimal test proposed in Elliott and Muller (2006). Their empirical investigation on international markets shows evidence of breaks for the vast majority of countries in multivariate regression models for excess returns, and the predictive relationship may change dramatically after a break. In a similar study, Rapach and Wohar (2006) used the same set of econometric tools, as in Paye and Timmermann (2006), to examine the aggregate US stock returns. They also find evidence of instability and strong variation in

predictability over time. These results, however, are susceptible due to the particular test those authors chose, especially the machinery proposed by Bai and Perron (1998). The reason is that their tests are based on stationary data generating process and by no means valid for a nonstationary system, such as a predictive regression with persistent predictors. This invalidity is demonstrated by the simulation studies reported in Table 2 of Paye and Timmermann (2006). They show that when innovations are highly correlated, and predictors are persistent, the size distortion of Bai and Perron's supF and UDMax tests is large. The optimal test of Elliott and Muller shows great size control even under this situation, but its power appears to be unsatisfactory.

The literature on rigorous tests for structural change aimed specifically at predictive regressions has recently mushroomed. Georgiev et al. (2018), for instance, extended the fixed regressor bootstrap approach in Hansen (2001) and proposed a fixed regressor wild bootstrap procedure to allow for heteroskedasticity. They show that this test is asymptotically valid and consider both SupF and LM test. Other available tests include Pitarakis (2017), who proposed a CUMSUM-type test, and Cai et al. (2014), who make use of an L_2-type statistic. All these tests are only concerned with the detection of potential instability and keep silent about the existence of predictability per se. To our best knowledge, the only existing papers that accommodates both are Demestrescu et al. (2020), who proposed to combine the predictability with subsampling technique.

The plan for the rest of the paper is as follows. In Section 3.2, we specify the model and review the recently proposed IVX method. Section 3.3 provides some preliminary simulation results serving as the motivation for our new test. Section 3.4 contains the construction of the test statistic, as well as its limiting properties. Simulation results are reported in Section 3.5, and an empirical illustration is presented in Section 3.6. Section 3.7 concludes the paper. All the proofs are contained in Appendix B.

3.2 Models and IVX methodology

In this section, we introduce the predictive regression model with break and also briefly review the state-of-the-art technique on statistical inference of this model, namely IVX-based robust Wald test. Our setup closely follows the one considered in KMS, and

thus we refer the readers to that paper for a more detailed description of the model as well as the construction of the IVX-Wald test statistic.

3.2.1 Predictive Regressions

In this chapter, we consider the following predictive regression model with a potential structural break at time t_0:

$$y_t = \alpha_t + \beta_t' x_{t-1} + u_{yt}, \quad [11] \tag{3.1}$$

$$x_t = \Phi_T x_{t-1} + u_{xt}, \tag{3.2}$$

$$\Phi_T = I_k - \frac{C}{T^{\gamma_x}}, \tag{3.3}$$

$$\alpha_t = \begin{cases} \alpha_1 & , 0 \leqslant t \leqslant t_0 \\ \alpha_2 & , t_0 \leqslant t \leqslant T \end{cases} \quad \text{and } \beta_t = \begin{cases} \beta_1 & , 0 \leqslant t \leqslant t_0 \\ \beta_2 & , t_0 \leqslant t \leqslant T \end{cases}, \tag{3.4}$$

where x_t is a k-dimensional vector of predictors with $x_0 = O_p(1)$ [12], while u_{yt} and u_{xt} are stationary. We will use T to denote the full sample size.

As presented in equation (3.1), we model the dynamics of predictors as an autoregression whose degree of persistence is determined by a matrix C and scalar γ_x. We assume that $C = diag(c_1, c_2, ...c_k)$, where all c_i's are non-negative real numbers. We also restrict γ_x to be a positive real number. As discussed in KMS, this setup is general enough to accommodate the cases from purely stationary predictors to integrated ones. Note that this formulation also encompasses the common local-to-unity assumption (e.g., Campbell and Yogo, 2006) as a special case. Such generality is empirically important because, in practice, it is hard for researchers to accurately assign the predictors to one particular case using a finite sample. Thus they may wish to apply an approach robust to all possible cases. We impose the following assumptions on the innovations driving the system.

[11] In KMS, predictand y_t is allowed to be multivariate as well. Here, we focus on the univariate case in accordance with the most literature on predictive regressions.

[12] In a recent working paper, Xu and Guo (2019) point out that existing predictability tests are problematic when k becomes large. We do not consider this effect here and pay attention only to a small set of predictors.

Assumption 3.1 *The innovation to x_t is a linear process and can be represented by*

$$u_{xt} = \sum_{j=0}^{\infty} C_j \epsilon_{t-j},$$

where $\{C_j\}_{j=0}^{\infty}$ is a sequence of absolutely summable constant matrices such that $\sum_{j=0}^{\infty} C_j$ has full rank and $C_0 = I_k$. Let $u_t = (u_{yt}, \epsilon_t')'$, then u_t is a martingale difference sequence with constant conditional variance matrix Σ and satisfies the moment condition that $\sup_t E||u_t||^{2s} < \infty$ for some $s > 1$.

We also make the following assumption, which is standard in the literature on structural breaks.

Assumption 3.2 *The fraction of structural break defined as $\tau_0 = t_0/T$ is within the interior of $(0, 1)$.*

When α_t and β_t are both constant over the whole sample period, that is, $\alpha_1 = \alpha_2 = \alpha$ and $\beta_1 = \beta_2 = \beta$, this model reduces to the widely used predictive regressions, and there is a vast volume of literature on the inference for such a model. Within this framework, the most critical hypothesis that people wish to test is $H_0 : \beta = 0$, whose rejection implies the existence of predictability. It has been proven that, if the ordinary least square (OLS) is used to estimate the model parameters, the limiting theory will be dependent on predictors' degree of persistence and become nonstandard when the predictors are (nearly-) integrated. Specifically, when the covariance of u_{xt} and u_{yt} is nonzero and $\gamma_x \geqslant 1$, the OLS estimator will converge to a non-mixed-Gaussian distribution and thereby suffer an asymptotic bias. Even worse, since this bias is determined by the location parameters c_1, ..., c_k, which are not consistently estimable, it can not be corrected. The inference for predictability, therefore, will be misleading if the traditional t-statistic is used to check for the significance. Although a battery of methods have been proposed to make better inference for predictive regressions, in this article, we focus solely on the IVX-based approach outlined in the next subsection due to its robustness and straightforward extension to multivariate cases.

3.2.2 Methodology of IV-based Inference and IVX-Wald test

The idea of the IVX approach was first proposed in Phillips and Magdalinos (2009) for a cointegration system to design test robust to arbitrary persistence in variables. KMS adopted this method to the predictive regressions and proposed a similar robust Wald test to detect the presence of predictability. In both cases, the intuition is to "construct an instrumental variable whose degree of persistence we explicitly control and avoid the inference problems arising due to the uncertainty regarding the persistence of the original regressors" (KMS, P1510). Moreover, the construction of IV requires no external information, thus justifying the terminology IVX.

In particular, the IVX approach proceeds by first specifying an artificial coefficient matrix

$$R_T = I_k - C_z/T^{\delta_z},$$

where C_z is a diagonal matrix with positive entries and $\delta_z < 1$. In practice, a researcher needs to make the decision on the choice of these tuning parameters. In this paper, we follow KMS to set $C_z = -I_k$ and $\delta_z = 0.95$. [13] After this specification, instrumental variables are constructed by

$$\tilde{z}_t = R_T \tilde{z}_{t-1} + \Delta x_t,$$

and if we impose zero initialization, can be written as

$$\tilde{z}_t = \sum_{j=1}^{t} R_T^{t-j} \Delta x_j.$$

One can treat $\tilde{z}_t$ as a filtered version of original predictors x_t and it requires no external information at all. Critically, it is mildly integrated and correlated with x_t. This instrumental variable $\tilde{z}_t$ is subsequently used to obtain an estimator of β, which we denote as $\tilde{\beta}^{IVX}$. If no break happens, the formula for estimation will be

$$\tilde{\beta}^{IVX} = \left[\sum_{t=1}^{T} \tilde{z}_{t-1}(x_t - \bar{x}_{T-1})' \right]^{-1} \sum_{t=1}^{T} \tilde{z}_{t-1}(y_t - \bar{y}_T), \tag{3.5}$$

[13] See section 4 in Phillips and Lee (2016) for a detail discussion in this regard.

where $\bar{x}_{T-1} = \frac{1}{T-1}\sum_{j=1}^{T-1} x_j$ and $\bar{y}_T = \frac{1}{T-1}\sum_{j=2}^{T} y_j$ are sample means.

As proved in KMS, when sample size T goes to infinity, $\tilde{\beta}^{IVX}$ will correctly center around the true value of β and the limiting distribution will be (mixed-) Gaussian, regardless of the degree of predictors' persistence. Hence, a natural statistic to test $H_0 : \beta = 0$ is the following IVX-Wald statistic

$$W = \tilde{\beta}^{IVX\prime}\tilde{Q}^{-1}\tilde{\beta}^{IVX},$$

where $\tilde{Q}$ is a consistent estimator of the asymptotic variance-covariance matrix of $\tilde{\beta}^{IVX}$ that accommodates both long-run endogeneity caused by the correlation between u_{xt} and u_{yt}, and finite-sample distortion results from the removing of intercept. Specifically,

$$\begin{aligned}
\tilde{Q} &= \left(\sum_{t=1}^{T} \tilde{z}_{t-1}x'_{t-1}\right)^{-1} M \left(\sum_{t=1}^{T} x_{t-1}\tilde{z}'_{t-1}\right)^{-1}, \\
M &= \hat{\sigma}_y^2 \sum_{t=1}^{T} \tilde{z}_{t-1}\tilde{z}'_{t-1} - T\bar{z}_{T-1}\bar{z}'_{T-1}\hat{\sigma}_{FM}^2, \\
\hat{\sigma}_{FM} &= \hat{\sigma}_y^2 - \hat{\Omega}_{yx}\hat{\Omega}_{yy}^{-1}\hat{\Omega}'_{yx},
\end{aligned} \tag{3.6}$$

where $\bar{z}_{T-1} = \frac{1}{T-1}\sum_{j=1}^{T-1} \tilde{z}_j$ and $\hat{\sigma}_y^2$ is a consistent estimator for σ_y^2. $\hat{\Omega}_{yx}$ and $\hat{\Omega}_{yy}$ are estimated long-run covariance between u_{yt} and u_{xt} and long-run variance of u_{yt} respectively. For a more detailed discussion on the construction of $\tilde{Q}$, we refer the reader to KMS (P1514-1515). Under the null hypothesis, this statistic will converge to the standard χ^2 distribution with degree of freedom equal to the number of predictors. Extensive simulations reported in KMS and the online appendix of that paper suggest this test has excellent finite sample properties.

3.3 Performance of IVX-Wald Test in the Presence of Structural Break

Although the IVX-Wald test provides an excellent remedy to the inference for predictability, it still suffers the drawback that all parameters are assumed to be constant over time, and no structural break is allowed. Given the fact that this assumption is highly

implausible for an extended period, in this section, we will explore the effects brought by the failure of constancy restriction. To this end, a simple simulation exercise is undertaken with a single predictor and one structural break to shed some light on the distortion brought by the presence of breaks.

3.3.1 Simulation Design

The data generating process that we consider in this section is a special case of equation (3.1)-(3.4). It is represented by

$$y_t = \alpha_t + \beta_t x_{t-1} + u_{yt}, \tag{3.7}$$

$$x_t = \phi x_{t-1} + u_{xt}, \tag{3.8}$$

$$\phi_T = 1 - \frac{c}{T_x^{\gamma}}, \tag{3.9}$$

$$\alpha_t = \begin{cases} \alpha_1 & , 0 \leqslant t \leqslant t_0 \\ \alpha_2 & , t_0 \leqslant t \leqslant T \end{cases} \text{ and } \beta_t = \begin{cases} \beta_1 & , 0 \leqslant t \leqslant t_0 \\ \beta_2 & , t_0 \leqslant t \leqslant T \end{cases}. \tag{3.10}$$

In this toy model, the sample size is chosen to be 100, and the persistence of the predictor is determined by setting $c = 10$ and $\gamma_x = 1$. The disturbance to the system, $(u_{yt}, u_{xt})'$, is generated by an independent multivariate normal distribution with zero mean and variance matrix $\begin{pmatrix} 1 & -0.95 \\ -0.95 & 1 \end{pmatrix}$. The choice of these parameters is justified by its empirical relevance. To address the structural break issue more clearly, we set the intercept and slope to be $\alpha_t = \begin{cases} -a & , t \leqslant 50 \\ a & , t > 50 \end{cases}$ and $\beta_t = \begin{cases} -b & , t \leqslant 50 \\ b & , t > 50 \end{cases}$, respectively, where $a = \{0, 0.2, 0.4, ..., 2\}$ and $b = \{0, 0.02, 0.04, ..., 0.2\}$. Note that under this parametrization, the break size is just twice of a for α and twice of b for β.

Figure 1 plots the rejection rates of the IVX-Wald test using the full sample with 2000 replications, under a 10% nominal size. It is clear from this figure that, when no structural break occurs for both intercept and slope coefficient, the rejection rates are close to the nominal size, as claimed by KMS. In general, when the break magnitude of either α or β increases, the rejection rate also rises. To better understand the size and power performance of IVX-Wald under the structural break, we need to further concentrate on

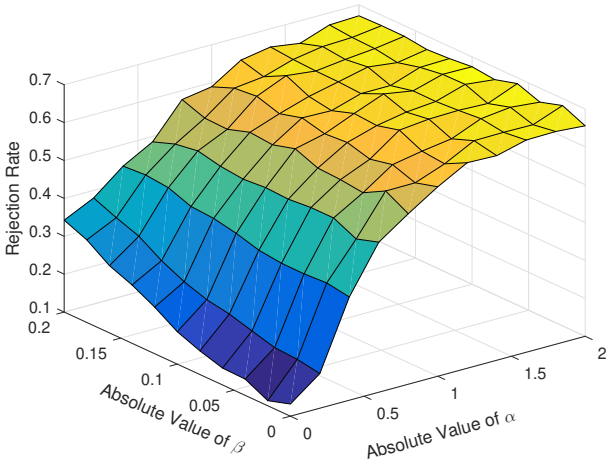


Figure 1: Rejection rate of IVX-Wald test under different structural break magnitude

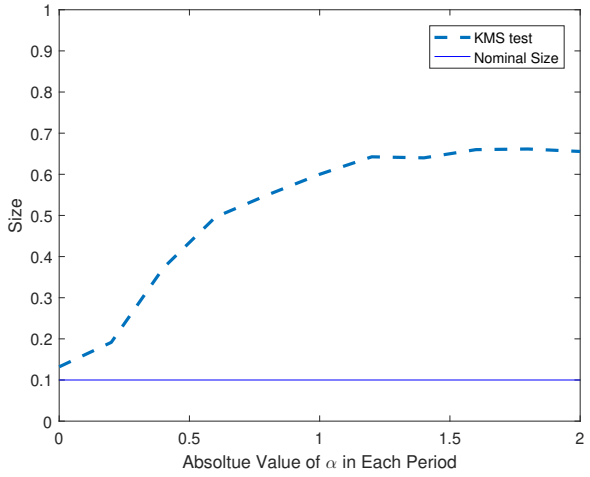


Figure 2: Size distortion of IVX-Wald test under break in α

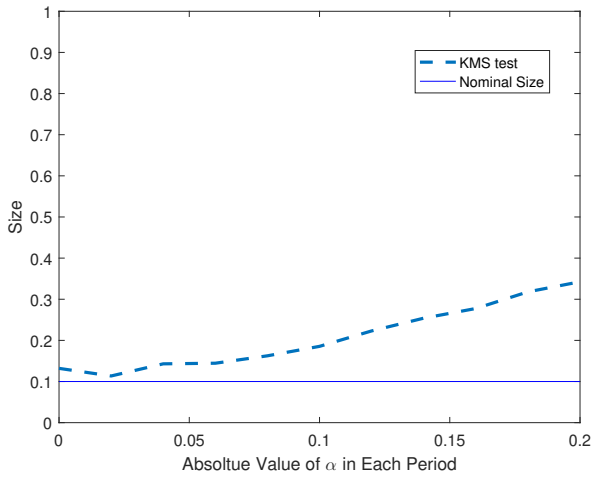


Figure 3: Power loss of IVX-Wald test under break in β

the following two cases.

3.3.2 Size Distortion and Power Loss

The first case we look into is that only α experiences structural break, and β remains constant. Figure 2 plots the rejection rates when $b = 0$. Since, in this case, there is no predictability throughout the sample period, an ideal test should reject the null hypothesis with a probability no greater than the nominal size. However, as suggested by Figure 2, the IVX-Wald test becomes remarkably over-sized when the value of a grows. For example, a break of magnitude 1 will lead to a null rejection rate above 40%. The terrible size control will lead to great concern in empirical research, as it may imply that the predictability confirmed by the IVX-Wald test is spurious and generated by the change in the intercept α. Actually, we can show that, when β is a constant over time, $\tilde{\beta}^{IVX} - \beta$, after proper normalization, will converge to a (mixed) Gaussian distribution, plus $(\alpha_1 - \alpha_2)\begin{cases} O_p(1) & ,\delta_z \leqslant \gamma_x \\ O_p(n^{\frac{\gamma_x - \delta_z}{2}}) & ,\delta_z > \gamma_x \end{cases}$. Therefore, when the predictors are more persistent than the IVX we choose, the break in intercept will result in an asymptotically non-negligible term and severe size distortion as a consequence. We will discuss this issue further in detail in the next chapter.

The second case we examine is that α keeps unchanged throughout, but β experiences a jump in value. The result is presented in Figure 3. For this case, β is nonzero within

the whole sample period, and thus, predictability exists. However, if one uses IVX-Wald, the rejection rate is very low. Essentially, when the break in β alters the direction of predictability, the power of the IVX-Wald test deteriorates dramatically. This is by no means surprising, as in this case, $\tilde{\beta}^{IVX}$ will converge asymptotically to the weighted average of β_1 and β_2, which could be very close to zero, even if both of them are far from zero. The empirical implication of this finding is that, when the IVX-Wald test rejects the presence of predictive power, it could be a consequence of the change in predictive direction before and after a structural break.

In summary, the above findings through simulation remind us that using the IVX-Wald test to detect predictability with a long span of data will incur a couple of potential problems. Specifically, both size control and power against the alternative are harmed by the occurrence of structural breaks. A natural question to ask given this discouraging result is whether it is possible to design a test that can be used under the presence of a break. We attempt to propose partial solutions in the current and the following chapter.

3.3.3 Preliminary Comparison of OLS against IVX For Testing Break

Before introducing the new test, we consider a simple simulation exercise in this subsection as a motivation. The results here may shed some light on the difference between OLS and IVX estimators in terms of testing the existence of structural break. The data generating process is a simple univariate predictive regression without constant term, $y_t = \beta x_{t-1} + u_{yt}$. The sample size is set to be 200. The predictor follows a local-to-unit-root process, $x_t = (1 - \frac{c}{T})x_{t-1} + u_{xt}$, where non-centrality parameter $c \in \{100, 50, 10, 5, 1, 0\}$. Both u_{yt} and u_{xt} are assumed to have unit variance, while the correlation between them is $\rho \in \{-0.9, -0.7, -0.5, 0, 0.5, -0.7, 0.9\}$. We apply the Andrews (1993) sup-Wald test, based on OLS and IVX estimators, respectively, to the generated data set. Specifically, we first compute the Wald statistic using OLS and IVX at all partitions of the sample and then take supreme over them. The corresponding variances are computed in a standard way. Note that here, to focus on the slope coefficients, we assume the intercept is known to be zero and thereby not estimated. The critical values for the rejection are taken from Andrews (1993).

Table 1: Finite sample size performance of OLS vs. IVX for testing structural break

		ρ						
		-0.9	-0.7	-0.5	0	0.5	0.7	0.9
OLS	100	0.08	0.08	0.08	0.08	0.07	0.07	0.07
	50	0.09	0.09	0.08	0.07	0.07	0.07	0.08
c	10	0.15	0.12	0.10	0.09	0.12	0.13	0.16
	5	0.21	0.15	0.12	0.10	0.14	0.16	0.20
	1	0.31	0.23	0.17	0.11	0.19	0.24	0.33
	0	0.35	0.26	0.18	0.11	0.20	0.27	0.38
IVX	100	0.06	0.08	0.09	0.10	0.09	0.08	0.07
	50	0.07	0.08	0.09	0.10	0.09	0.08	0.07
c	10	0.08	0.09	0.10	0.11	0.10	0.10	0.08
	5	0.08	0.09	0.10	0.11	0.10	0.09	0.08
	1	0.07	0.08	0.09	0.11	0.09	0.08	0.07
	0	0.07	0.08	0.09	0.11	0.09	0.08	0.07

The empirical rejection rates based on 2000 Monte Carlo simulations are reported in Table 1, where the nominal level is 10%. It can be seen that the traditional OLS-based sup-Wald test performs well when ρ is zero, even if the predictor is highly persistent. However, as long as the correlation between innovations deviates from 0, the size distortion is severe. Indeed, when the predictor follows a random walk, the empirical rejection rate could be as large as 0.38. Given the highly nonstandard limiting distribution of OLS estimator under such circumstances, this is by no means surprising. On the contrary, when we modify the sup-Wald statistic by using the IVX estimator and its corresponding variance, the performance becomes insensitive to the change in the level of both persistence and correlation. The robustness is precisely the purpose of constructing IVX in the first place. This simulation study indicates that testing for a structural break in predictive regressions can also benefit from using IVX. This motivates us to propose an IVX-based test in the predictive regression model with a potential structural break.

3.4 Testing for predictability and Structural Break

In this part, we propose a test that has power against the structural break in the parameters, as well as the existence of predictability. We start from the case in which the intercept term α is known to be stable. The null hypothesis remains the same as in the IVX-Wald test, but the alternative hypothesis we have in mind is now $H_A^\beta : \beta_1 \neq \beta_2$, or $\beta_1 = \beta_2 \neq 0$. Subsequently, we further consider the test valid in the presence of struc-

tural break in α. For that case, the alternative hypothesis is $H_A^{\alpha\beta} : \alpha_1 \neq \alpha_2, \beta_1 \neq \beta_2$, or $\beta_1 = \beta_2 \neq 0$.

3.4.1 Test when α is known to be stable

When α is known to be free of structural break, we can focus on the behavior of β. Encouraged by the simulation results reported in the last section, we propose to construct a IVX-based sup-Wald statistic to test against the potential existence of shift in β over time. Specifically, for every $t \in [t_L, t_U]$, where t_L and t_U are lower and upper bound for possible break date[14], we split the full sample $\{y_j, x_{j-1}\}_{j=1}^T$ into two parts, namely the one before time t, $\{y_j, x_{j-1}\}_{j=1}^t$ and the one after, $\{y_j, x_{j-1}\}_{j=t+1}^T$. Given each partition, two IVX estimates of β and also their asymptotic variances are computed, using the observations in each subperiod respectively. The calculation is based on the equation (3.5) and (3.6) above. Denote the resulting estimates as $\tilde{\beta}_1^{IVX}(t)$, $\tilde{Q}_1(t)$, $\tilde{\beta}_2^{IVX}(t)$ and $\tilde{Q}_2(t)$. Now, we can construct the test statistic as

$$W_\beta \equiv \sup_{t\in[t_L,t_U]} (W_{\beta=0} + W_{bt}),$$

where

$$W_{\beta=0} \equiv \tilde{\beta}^{IVX\prime}\tilde{Q}^{-1}\tilde{\beta}^{IVX},$$

is the IVX-Wald statistic using full sample and

$$W_{bt} \equiv \left(\tilde{\beta}_1^{IVX}(t) - \tilde{\beta}_2^{IVX}(t)\right)' \left[\tilde{Q}_1(t) + \tilde{Q}_2(t)\right]^{-1} \left(\tilde{\beta}_1^{IVX}(t) - \tilde{\beta}_2^{IVX}(t)\right)$$

is a Chow-type statistic for detecting break at time t. The following theorem presents the limiting distribution of W_{bt} under the null hypothesis, which depends on the persistence level of the predictors.

Theorem 3.1 *Consider model (3.1)-(3.4) under Assumption 3.1-3.2 and α is known to be*

[14] In this paper, we set t_L and t_U to be 20% and 80% of sample size, respectively.

stable a priori. Then, under $H_0 : \beta_1 = \beta_2$*, we have* $W_{bt} \Rightarrow H(\tau)'M(\tau)^{-1}H(\tau)$*, where*

$$H(\tau) = B(\tau) - R(\tau)B(1)$$

$$M(\tau) = \tau(I_k - R(\tau))(I_k - R(\tau))' + (1-\tau)R(\tau)R(\tau)'$$

$$R(\tau) = \begin{cases} (\tau I_k + \int_0^\tau \underline{B}dB')(I_k + \int_0^1 \underline{B}dB')^{-1} & \text{, if } \gamma_x > 1 \\ (\tau I_k + \int_0^\tau \underline{J}_C dJ_C')(I_k + \int_0^1 \underline{J}_C dJ_C')^{-1} & \text{, if } \gamma_x = 1 \\ \tau I_k & \text{, if } \gamma_x < 1 \end{cases}$$

$B(\cdot)$ *is a k-dimensional standard Brownian motion,* $J_C(\tau) = \int_0^\tau e^{C(\tau-s)}dB(s)$ *is an Ornstein-Uhlenbeck (OU) process and* $\underline{B}(\tau) = B(\tau) - \int_0^1 B(s)ds$ *and* $\underline{J}_C(\tau) = J_C(\tau) - \int_0^1 J_C(s)ds$ *are corresponding demeaned process.*

Theorem 3.1 makes it clear that if the predictors are integrated or nearly integrated, it goes to a nonstandard process, in the similar spirit of Hansen (2000). If, on the contrary, the predictors are stationary or mildly stationary, it will converge to the familiar squared tied-down Bessell process, as in Andrews (1993). This latter special case is reported in the following corollary.

Corollary 3.1 *When* $\gamma_x < 1$*, we have* $W_{bt} \Rightarrow \frac{BB_k(\tau)'BB_k(\tau)}{\tau(1-\tau)}$*, where* BB_k *is a k-dimensional standard Brownian bridge.*

As pointed out in Hansen (2000), the arguments in Andrews (1993) that lead to a standard distribution will fail if the marginal distributions of regressors are not time-invariant. This failure will occur in our setup if x_t is non-stationary, i.e. either nearly integrated or integrated. In particular, if we apply OLS to estimate β, its limiting distribution will depends on $\sum_{t=1}^{[\tau n]}(x_{t-1} - \bar{x}_{n-1})(x_{t-1} - \bar{x}_{n-1})'$ and $\sum_{t=1}^{[\tau n]}(x_{t-1} - \bar{x}_{n-1})(u_{yt} - \bar{u}_n)'$. The first part will converge to the integral of a squared demeaned OU process, while the second part will go to a stochastic integral of that demeaned OU process. Therefore, the linear growth of sample variance and the limiting Gaussianity of sample covariance are both violated, and we can not directly apply the traditional sup-Wald test in the current setup. Indeed, Georgiev et al. (2018) derive the limiting distribution of sup-Wald test statis-

tic under the local-to-unit-root assumption, which is highly nonstandard and non-pivotal. Therefore, in that paper, they propose to make inference using fixed regressor bootstrap.

If, on the contrary, we use IVX to estimate β, the asymptotics is determined by $\sum_{t=1}^{[\tau n]} \tilde{z}_{t-1}(x_{t-1} - \bar{x}_{n-1})'$ and $\sum_{t=1}^{[\tau n]} \tilde{z}_{t-1}(u_{yt} - \bar{u}_n)'$. As summarized in Lemma B.1 in Appendix, the former will grow linearly if predictors are stationary or mildly stationary, while still converge to a random limit if they are (nearly-) integrated. On the other hand, the latter will converge to a Brownian motion in all cases. Therefore, when $\gamma_x < 1$, we expect the conventional limiting theory applies, while when $\gamma_x \geqslant 1$, asymptotic distribution will still be nonstandard even if we resort to IVX. Theorem 3.1 just confirms above intuitive argument.

One subtle yet important point to note is that, even if we can not achieve standard Brownian bridge-type of limit with IVX under strong persistence, we expect that the deviation should be much smaller than the case of OLS. The reason is twofold. First, the $\sum_{t=1}^{[\tau T]} \tilde{z}_{t-1}(u_{yt} - \bar{u}_n)'$ part has a Gaussian process limit, instead of a stochastic integral as in OLS case. Second, $\sum_{t=1}^{[\tau T]} \tilde{z}_{t-1}(x_{t-1} - \bar{x}_{n-1})'$, though accumulates nonlinearly, is closer to a linearly growing second moments than $\sum_{t=1}^{[\tau T]}(x_{t-1} - \bar{x}_{n-1})(x_{t-1} - \bar{x}_{n-1})'$. To illustrate this latter claim, let us concentrate on the univariate case with a single local-to-unit-root predictor. According to the Theorem 3.1, the deviation of W_{bt} from the conventional limiting process when $\gamma_x \geqslant 1$ is due to the difference between $R(\tau)$ and τ. Therefore, we investigate the behavior of

$$R(\tau) - \tau = \frac{\tau + \int_0^\tau \underline{J}_C dJ_C}{1 + \int_0^1 \underline{J}_C dJ_C} - \tau$$

As it is challenging to obtain the analytic property of this object, we rely on simulation to provide some conducive insights. To this end, we simulate 10000 replication for $\tau \in [0.2, 0.8]$ and $C \in \{100, 5, 1\}$. The underlying Brownian motion that drives the OU process involved in the above object is approximated by a 1000-step random walk. The mean and 95% confidence interval are plotted in Figure 4 to 6. As a comparison, we also

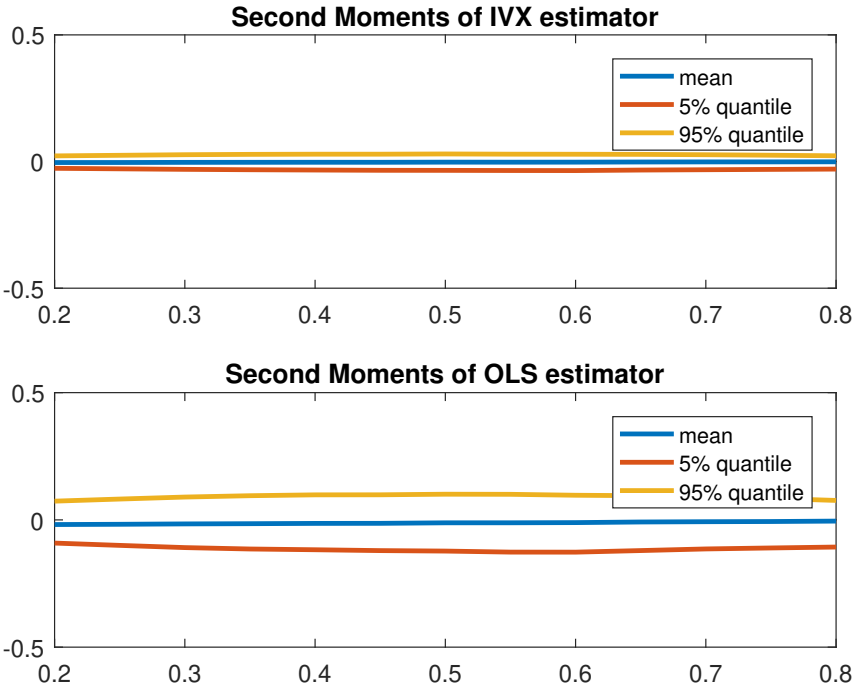


Figure 4: Accumulation of Second Moments of IVX vs. OLS when $c = -100$

plot the counterparts for OLS estimator, which is

$$\frac{\int_0^\tau \underline{J}_C^2 dJ_C}{\int_0^1 \underline{J}_C^2 dJ_C} - \tau.$$

It can be clearly seen that, when the persistence level is low, the second moments of both IVX and OLS will grow linearly as $R(\tau)$ is very close to τ for all $\tau \in [0.2, 0.8]$. However, for the OLS case, when C converges towards zero, the mean of $R(\tau) - \tau$ becomes far from zero and the variation increases dramatically. $R(\tau)$ for IVX, on the other hand, still remains close to τ with a tiny fluctuation and therefore ensures the limiting process of W_{bt} can be well approximated by the conventional limit theory. Indeed, our preliminary simulation reported earlier and more extensive results shown later both indicate that our test still controls size even when the predictors follow the unit root process. In this sense, the test we propose is robust for structural break detection, just like the IVX-Wald for predictability test.

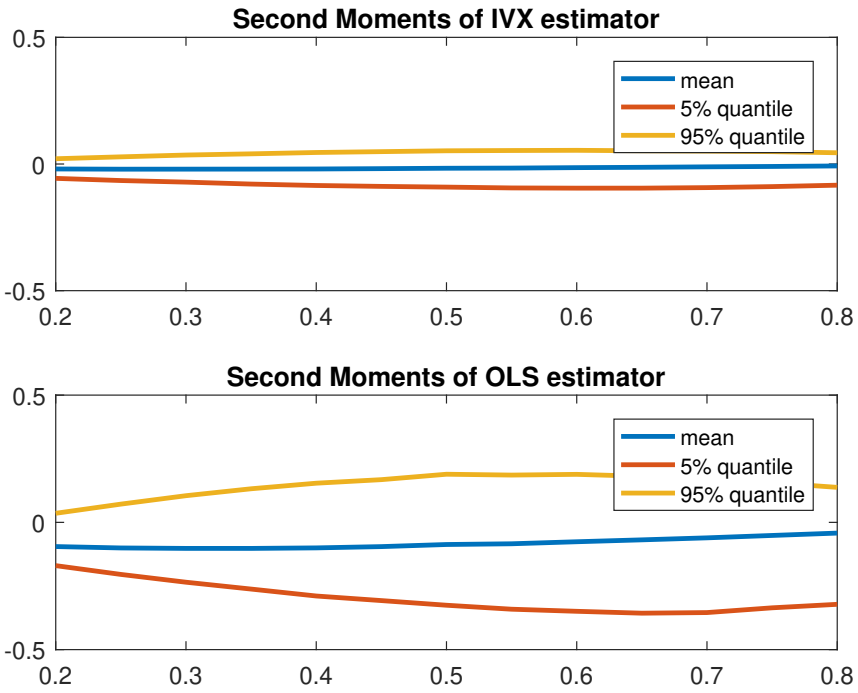


Figure 5: Accumulation of Second Moments of IVX vs. OLS when $c = -10$

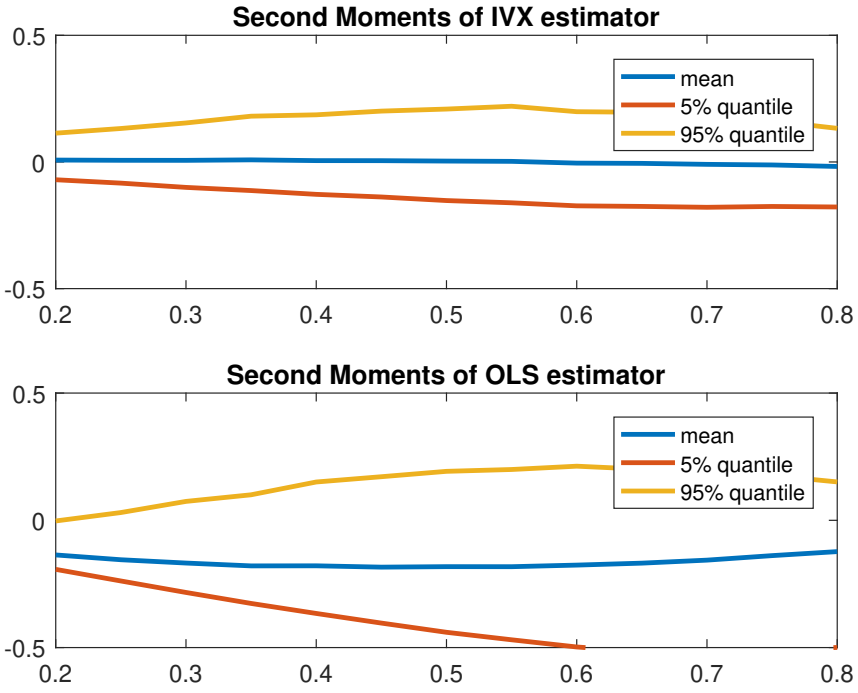


Figure 6: Accumulation of Second Moments of IVX vs. OLS when $c = 0$

Corollary 3.2 *(i) Under the conditions of Theorem 3.1, we have that when the sample size tends to infinity,*

$$W_\beta \Rightarrow B(1)'B(1) + \sup_{\tau\in[\tau_L,\tau_U]} H(\tau)'M(\tau)^{-1}H(\tau)$$

where $B(\cdot)$, $H(\tau)$ and $M(\tau)$ are defined in Theorem 3.1.

(ii) As a special case, if $\gamma_x < 1$, we have

$$W_\beta \Rightarrow \chi^2_k + \sup_{\tau\in[\tau_L,\tau_U]} \frac{BB_k(\tau)'BB_k(\tau)}{\tau(1-\tau)},$$

where BB_k is a k-dimensional standard Brownian bridge and χ^2_k is a random variable following χ^2 distribution with degree of freedom equal to k. Furthermore, these two components in the limiting distribution are independent of each other.

At first glance, it seems that the interesting part of Corollary 3.2 is the special situation $\gamma_x < 1$, in which case the test statistic W_β is pivotal when the sample size goes to infinity. Due to the absence of a nuisance parameter, asymptotic critical values for testing the null hypothesis can be easily established under this assumption. Indeed, one can obtain the quantile of W_β through a direct Monte Carlo simulation given any particular number of predictors k. However, the actual merit of Corollary 3.2 lies in its robustness under very strong persistence, as discussed above for W_{bt}. Note that the robustness comes from two components making up W_β, as the detection of both predictability and structural break becomes insensitive to the level of persistence after replacing OLS with an IVX estimator.

Let us summarize the discussion in this subsection. If one wishes to test the null hypothesis that $\beta_1 = \beta_2 = 0$, imposing the restriction that constant term α does not experience any structural break, she can just rely on the test statistic W_β. Though theoretically only free of nuisance parameter with mildly stationary predictors, from a pragmatic viewpoint, that asymptotic pivotal distribution still works well enough when the predictors are highly persistent.

3.4.2 Test with Potential Break in α

For empirical analysis, usually it is not convincing to allow slope coefficient β to change over time while claim that intercept keeps stable. Hence, it is desirable to have a test that also shows power against the potential structural break in α. For that purpose, we need an estimator for α with good property so that we can detect the presence of break. A natural choice would be $\hat{\alpha} = \bar{y}_n - \tilde{\beta}^{IVX}\bar{x}_{n-1}$. This choice, however, leads to a complicated limiting theory and more importantly varies with the persistence level of predictors. The main reason is that the convergence rate of this estimator depends on the persistence level of x_t and when x_t is very persistent, it will converge to a highly nonstandard limit. To see this, notice that

$$\sqrt{T}(\hat{\alpha} - \alpha) = \sqrt{T}\bar{\epsilon} - \left[T^{\frac{1+\min[\gamma_x,\delta_z]}{2}}(\tilde{\beta} - \beta)\right]\left[T^{-\frac{\min[\gamma_x,\delta_z]}{2}}\bar{x}_{T-1}\right].$$

In this formula, it is easy to know that $\sqrt{T}\bar{\epsilon}$ and $T^{\frac{1+\min[\gamma_x,\delta_z]}{2}}(\tilde{\beta} - \beta)$ are both $O_p(1)$. The order of the rest term can been shown to depend on the persistence level of the predictors. Indeed, since we have

$$\sum_{t=1}^{T} x_{t-1} = \begin{cases} O_p(T^{-1/2}) & ,\gamma_x = 0 \\ O_p(T^{\frac{1}{2}+\gamma_x}) & ,0 < \gamma_x < 1 \\ O_p(T^{\frac{3}{2}}) & ,\gamma_x \geqslant 1 \end{cases}$$

it is straightforward too check that $T^{-\frac{min[\delta_z,\gamma_x]}{2}}\bar{x}$ will dominate in the limit if $\gamma_x > (\delta_z + 1)/2$, while vanish if the reverse holds. When $\gamma_x = (\delta_z + 1)/2$, all three terms will emerge in the asymptotic distribution, making it very complicated and non-standard.

To circumvent this complexity, we turn to estimate intercept by $\tilde{\alpha} = \bar{y}_T - \tilde{\beta}^{IVX}\bar{z}_{T-1}$, where $\bar{z}_{T-1} = \frac{1}{T-1}\sum_{j=2}^{T}\tilde{z}_{j-1}$. Since in this case the persistence level of $\tilde{z}_t$ is controlled by researcher, we are free from the varying persistence concern. Another surprising benefit of this choice is that, it will also bring some power against a nonzero β. To see this point,

note that the estimator so defined can be written as

$$\begin{aligned}\tilde{\alpha} =&\alpha + \beta\bar{x}_{T-1} + \bar{\epsilon}_T - \tilde{\beta}^{IVX}\bar{z}_{T-1} \\ =&\alpha + \bar{\epsilon}_n - (\tilde{\beta}^{IVX} - \beta)\bar{z}_{T-1} + \beta(\bar{x}_{T-1} - \bar{z}_{T-1})\end{aligned} \tag{3.11}$$

and therefore $\tilde{\alpha} - \alpha$ consists of three components. It is easy to see that the order of first term is $O_p(T^{-1/2})$. We show in the appendix that the second term is asymptotically dominated by the first term, while the last object will diverge when β is not too close to zero. It is this third term that results in power against the presence of predictability.

Analogous to the case for β, we compute $\tilde{\alpha}_1(t)$ and $\tilde{\alpha}_2(t)$ given the partition at time t, using data before and after that date. Using this IVX-based estimator for intercept, we can define the second component as

$$W_{at} \equiv (\tilde{\alpha}_1(t) - \tilde{\alpha}_2(t))' \left[\tilde{\Omega}_1(t) + \tilde{\Omega}_2(t)\right]^{-1} (\tilde{\alpha}_1(t) - \tilde{\alpha}_2(t)),$$

where

$$\tilde{\Omega}_1(t) = \hat{e}_1'\hat{e}_1/t^2 + \bar{z}_t'\tilde{Q}_1(t)\bar{z}_t$$

and

$$\tilde{\Omega}_2(t) = \hat{e}_2'\hat{e}_2/(T-t)^2 + \bar{z}_{T-t}'\tilde{Q}_2(t)\bar{z}_{T-t}.$$

Note that the second terms in the definition of $\tilde{\Omega}_1(t)$ and $\tilde{\Omega}_2(t)$ serve as second-order bias correction. Such correction is necessary because the second term in equation (3.11), though dominated asymptotically, may have a non-negligible influence in finite sample.

Theorem 3.2 *If Assumption 3.1 and 3.2 are satisfied, then under* $H_0 : \alpha_1 = \alpha_2$, $\beta_1 = \beta_2 = \beta$ *and* $||\beta|| = o_p(T^{\frac{\gamma_x+1}{2}})$,

$$W_{at} \Rightarrow \frac{BB_1(\tau)'BB_1(\tau)}{\tau(1-\tau)},$$

where BB_1 *stands for a one-dimensional standard Brownian bridge.*

With these two components obtained, our final test statistic is simply constructed by

$$W_{\alpha\beta} = \sup_{t\in[t_L,t_U]} (W_{\beta=0} + W_{at} + W_{bt}).$$

Since we have already derived the asymptotic distribution for each of the component, it is easy to obtain the following result for $W_{\alpha\beta}$.

Theorem 3.3 *If Assumption 3.1 and 3.2 are satisfied, then under* $H_0 : \alpha_1 = \alpha_2$, *and* $\beta_1 = \beta_2 = 0$,

(i). When sample size goes to infinity, we have

$$W_{\alpha\beta} \Rightarrow B(1)'B(1) + \sup_{\tau\in[\tau_L,\tau_U]} \tilde{H}(\tau)'\tilde{M}(\tau)^{-1}\tilde{H}(\tau),$$

where $\tilde{H}(\tau) = (BB_1(\tau), H(\tau))'$ *and* $\tilde{M}(\tau) = \begin{pmatrix} \tau(1-\tau) & \\ & M(\tau) \end{pmatrix}$. *Again,* $H(\tau)$ *and* $M(\tau)$ *are defined in Theorem 3.1.*

(ii). As a special case, if $\gamma_x \in (0,1)$,

$$W_{\alpha\beta} \Rightarrow \chi^2_k + \sup_{\tau\in[\tau_L,\tau_U]} \frac{BB_{k+1}(\tau)'BB_{k+1}(\tau)}{\tau(1-\tau)},$$

where BB_{k+1} *is a (k+1)-dimensional standard Brownian bridge and* χ^2_k *is a random variable following* χ^2 *distribution with degree of freedom equal to* k. *Furthermore, two components in the limiting distribution are independent of each other.*

Again, the limiting distribution is pivotal only when the predictors are stationary or mildly stationary. However, simulation results reported below shows that, even if we use the critical value of this pivotal distribution for (nearly-) integrated predictors, the size control is still satisfactory. Therefore, we do not need to worry about the unknown persistence level in practice and the part (ii) of above theorem is of greater relevance.

3.5 Simulation Results

We rely on Monte Carlo simulation to investigate the finite-sample performance of our new test. The data generating process we consider is

$$y_t = \alpha_t + \beta_t x_{t-1} + u_{yt}, \tag{3.12}$$

$$x_t = \phi x_{t-1} + u_{xt}, \tag{3.13}$$

$$\phi_T = 1 - \frac{c}{T}, \tag{3.14}$$

$$\alpha_t = \begin{cases} \alpha_1 & , 0 \leqslant t \leqslant t_0 \\ \alpha_2 & , t_0 \leqslant t \leqslant T \end{cases} \text{ and } \beta_t = \begin{cases} \beta_1 & , 0 \leqslant t \leqslant t_0 \\ \beta_2 & , t_0 \leqslant t \leqslant T \end{cases} \tag{3.15}$$

The disturbance to the system, $(u_{yt}, u_{xt})'$, is generated by independent multivariate normal distribution with zero mean and and variance $\begin{pmatrix} 1 & \rho \\ \rho & 1 \end{pmatrix}$. The results for empirical size and power shown below are computed using 10,000 Monte Carlo repetitions.

3.5.1 Empirical Size

A critical issue to explore is whether our test performs well when the autoregressive root of predictors is near unity. Given any finite sample size, it is impossible to precisely distinguish between a mildly stationary process and a local-to-unit-root process. Hence, Monte Carlo simulation is required to check the applicability of our test under empirically relevant circumstances. To exam the size control of the test under null hypothesis, we set $\alpha_1 = \alpha_2 = 0$ and $\beta_1 = \beta_2 = 0$. Three sample size, $T = 100$, $T = 200$ and $T = 400$ are considered. For a fixed sample size, the closeness to a unit root is captured by the parameter c, for which we consider six values, viz $c \in \{100, 50, 10, 5, 1, 0\}$. This set ranges from a very stationary case ($c = 100$) to unit root case ($c = 0$) and is thus representative enough. Another parameter that may have an impact on the testing performance is the correlation level ρ. In this regard, we also choose seven empirical relevant values, $\rho \in \{-0.95, -0.7, -0.5, 0, 0.5, 0.7, 0.95\}$.

In Table 2 and Table 3, we report the empirical rejection rates for two nominal levels, 0.1, and 0.05, respectively. It is clear from these tables that our test shows a good control

of size in general, except for the case with a very small sample size like $T = 100$. In particular, even with the extreme case that the predictor is highly local to the unit root process and the random shocks are strongly correlated, the rejection rate is still near and under the nominal level. An interesting finding from this study is that, for a very persistent predictor, when the absolute value of the correlation coefficient ρ is close to 1, our test seems to be undersized.

Since the easy application of IVX in multivariate cases is a great advantage of this technique, it is also critical to investigate the performance of our new test when multiple predictors are involved simultaneously. To this end, we consider a model with four predictors, all of which have the same persistence level captured by a different value of c as in univariate simulation. Two correlation structures are chosen. The first case, labeled 'uncorrelated case', assumes the covariance matrix of $(u_{yt}, u'_{xt})'$ is identity matrix and thus no long-run endogeneity exists. The second case, labeled 'correlated case', assumes the errors have the following covariance structure,

$$\begin{pmatrix} 0.0412 & 0.0347 & -0.0299 & 0.1088 & -0.0001 \\ 0.0347 & 0.7664 & 0.1072 & 0.9606 & -0.0008 \\ -0.0299 & 0.1072 & 1.4576 & 0.4638 & 0.0006 \\ 0.1088 & 0.9606 & 0.4638 & 11.4784 & 0.0011 \\ -0.0001 & -0.0008 & 0.0006 & 0.0011 & 0.0001 \end{pmatrix}$$

which is taken from Yang et al. (2020) for its empirical relevance. As can be seen in Table 4, our test has a good size control under both cases regardless of sample size. The only exception is when the predictors are very stationary, in which case our test seems to be over-sized a little bit. To conclude, the simulation results are encouraging, and our test has a good finite sample performance under various parameter settings.

3.5.2 Empirical Power

To address the structural break issue, we set intercept and slope to be

$$\alpha_t = \begin{cases} -a/\sqrt{T} & , t \leqslant \tau_0 \cdot T \\ a/\sqrt{T} & , t > \tau_0 \cdot T \end{cases}$$

Table 2: Finite sample size performance of joint test with nominal level 0.1

		ρ						
		-0.9	-0.7	-0.5	0	0.5	0.7	0.9
T=100	100	0.08	0.09	0.11	0.12	0.10	0.09	0.08
	50	0.08	0.10	0.11	0.13	0.11	0.10	0.08
c	10	0.10	0.13	0.14	0.14	0.14	0.13	0.10
	5	0.12	0.14	0.16	0.14	0.15	0.14	0.11
	1	0.12	0.16	0.17	0.14	0.16	0.15	0.12
	0	0.12	0.16	0.17	0.14	0.16	0.15	0.12
T=200	100	0.08	0.09	0.10	0.12	0.10	0.09	0.08
	50	0.08	0.09	0.10	0.12	0.10	0.09	0.08
c	10	0.08	0.10	0.12	0.12	0.12	0.10	0.09
	5	0.08	0.11	0.12	0.13	0.13	0.11	0.08
	1	0.08	0.11	0.13	0.12	0.13	0.11	0.08
	0	0.08	0.10	0.13	0.12	0.13	0.11	0.08
T=400	100	0.08	0.09	0.10	0.12	0.10	0.09	0.08
	50	0.08	0.09	0.10	0.11	0.10	0.09	0.08
c	10	0.07	0.09	0.11	0.12	0.12	0.10	0.07
	5	0.06	0.09	0.12	0.12	0.12	0.10	0.06
	1	0.06	0.09	0.11	0.12	0.12	0.10	0.06
	0	0.07	0.09	0.11	0.12	0.12	0.10	0.06

Table 3: Finite sample size performance of $W_{\alpha\beta}$ test with nominal level 0.05

		ρ						
		-0.9	-0.7	-0.5	0	0.5	0.7	0.9
T=100	100	0.04	0.04	0.05	0.06	0.05	0.04	0.04
	50	0.04	0.05	0.05	0.06	0.05	0.04	0.04
c	10	0.05	0.06	0.07	0.07	0.07	0.06	0.05
	5	0.05	0.07	0.08	0.07	0.08	0.07	0.05
	1	0.06	0.08	0.08	0.07	0.08	0.08	0.05
	0	0.06	0.08	0.08	0.07	0.08	0.08	0.05
T=200	100	0.03	0.04	0.04	0.05	0.05	0.04	0.03
	50	0.03	0.04	0.05	0.05	0.05	0.04	0.03
c	10	0.03	0.04	0.05	0.06	0.05	0.05	0.03
	5	0.03	0.05	0.06	0.06	0.06	0.05	0.03
	1	0.03	0.05	0.06	0.06	0.06	0.05	0.03
	0	0.03	0.05	0.06	0.06	0.06	0.05	0.03
T=400	100	0.03	0.04	0.04	0.06	0.04	0.03	0.03
	50	0.03	0.04	0.04	0.05	0.04	0.03	0.03
c	10	0.02	0.04	0.05	0.06	0.05	0.04	0.03
	5	0.02	0.04	0.05	0.06	0.06	0.04	0.02
	1	0.02	0.04	0.05	0.06	0.05	0.04	0.02
	0	0.02	0.04	0.05	0.06	0.05	0.04	0.02

Table 4: Finite sample size performance of $W_{\alpha\beta}$ test with nominal level 0.10 with multiple predictors

c	100	50	10	5	1	0
Uncorrelated Case						
T=200	0.13	0.10	0.10	0.10	0.09	0.09
T=400	0.09	0.09	0.09	0.09	0.09	0.08
Correlated Case						
T=200	0.14	0.10	0.10	0.10	0.10	0.09
T=400	0.10	0.09	0.09	0.09	0.08	0.08

and

$$\beta_t = \begin{cases} -b\sqrt{1-\rho^2}/T & , t \leqslant \tau_0 \cdot T \\ b\sqrt{1-\rho^2}/T & , t > \tau_0 \cdot T \end{cases}$$

respectively, where $a = \{0, 1, ..., 12\}$, $b = \{0, 1, ..., 30\}$. Three τ_0's are considered, namely 0.3, 0.5 and 0.7. Note that under this parametrization, the break size is just twice of $a/\sqrt{T}$ for α and twice of $b\sqrt{1-\rho^2}/T$ for β. All power functions are computed under three level of persistence, i.e. $c = 10$, $c = 5$ and $c = 0$. The sample size is fixed to be 200 and ρ is selected to be -0.7.

The power against the break in intercept is presented in Figure 7. It can be seen that the rejection rate of the test statistic quickly goes to 1 as the break size increases. This is the case for all three levels of persistence, and it is insensitive to the break fraction. Figure 8-10 exhibit the power against a break in the predictability. The power of the test is higher when the predictors are more persistent and the break occurs earlier in the data. The former result can be explained by the fact that when the predictors are more persistent, IVX estimators will have a higher rate of convergence and thus a smaller variance given a particular sample size, which suggests an easier detection of the shift in β. As an important comparison, we also plot the rejection rate of the corresponding IVX-Wald test for each scenario under consideration. As expected, the power of IVX is deficient, as it ignores the break. When the direction of predictability is opposite pre and post the break, the null hypothesis will be rejected by our test. Hence, the rejection of our joint test is more informative than the IVX-Wald test.

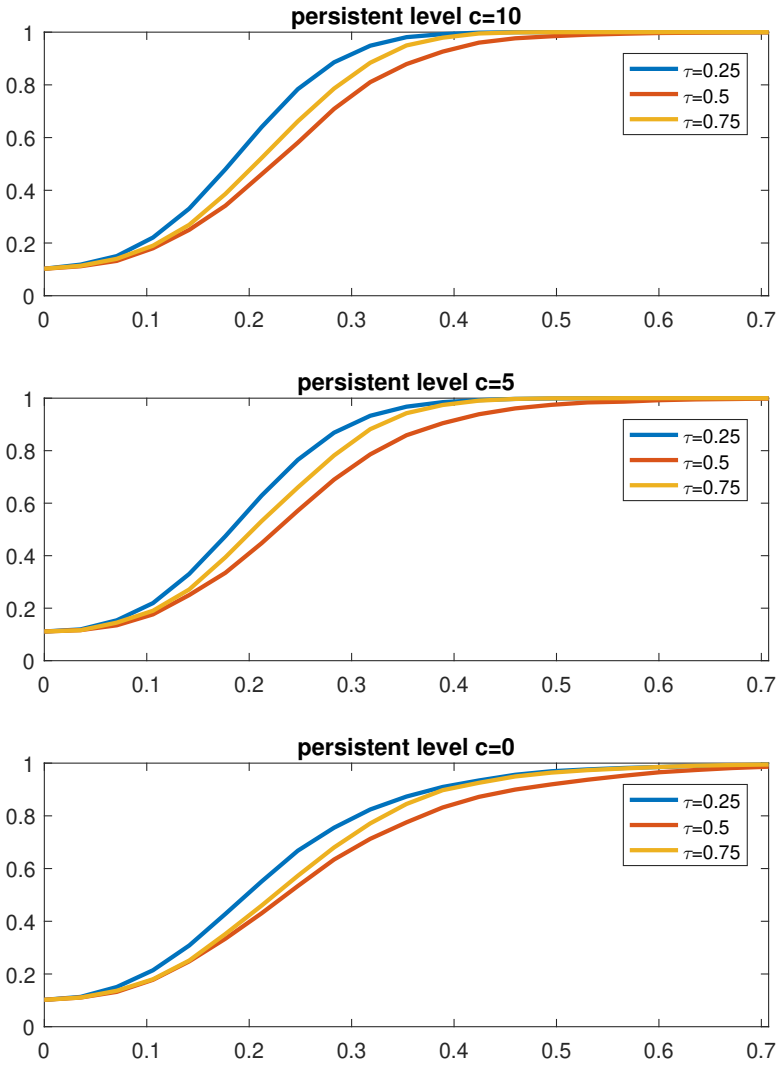


Figure 7: Power performance of joint test against break in intercept

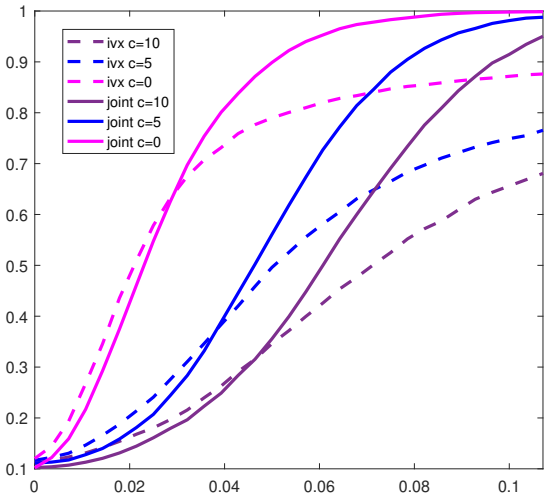


Figure 8: Power performance of joint test against non-zero slope when τ=0.3

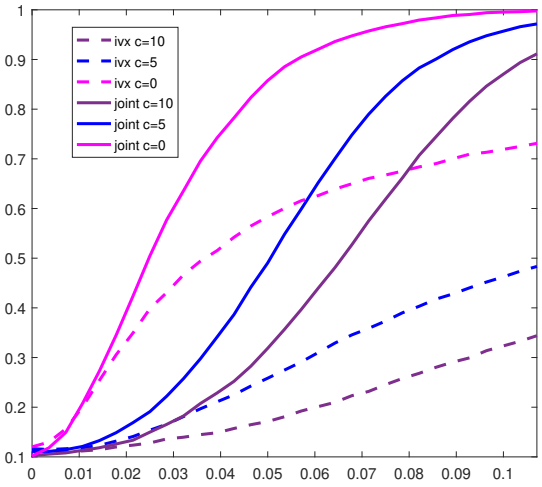


Figure 9: Power performance of joint test against non-zero slope when τ=0.5

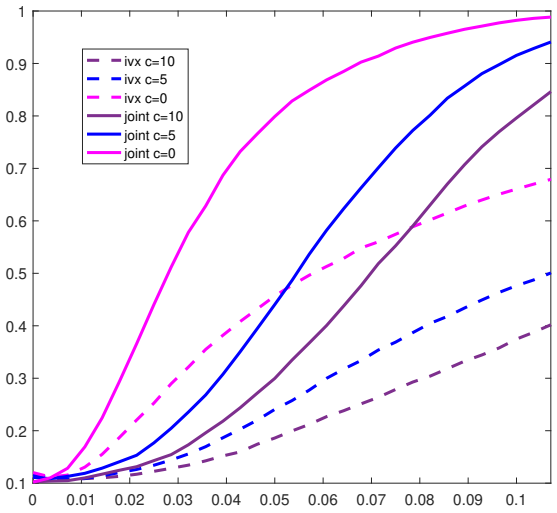


Figure 10: Power performance of joint test against non-zero slope when τ=0.7

3.6 Application to the US stock returns

To illustrate the empirical relevance of our proposed procedure, we implement the test to investigate the predictability of US stock returns. Following much of the existing literature, we focus on the well-known dataset from Welch and Goyal (2008)[15]. The variable to be predicted is S&P500 value-weighted log excess returns. To make a comparison with the results reported in KMS, we consider following 11 predictors: Tbill rate (tbl), long-term yield (lty), term spread (tms), default yield spread (dfy), dividend-price ratio (d/p), dividend yield (d/y), earnings-price ratio (e/p), dividend payout ratio (d/e), book-to-market value ratio (b/m), net equity expansion (ntis) and inflation rate (inf). For quarterly data, we also consider the consumption-wealth ratio (cay). For the detailed definition and construction of these variables, see the Online Appendix of KMS. Since our test is designed to have power only under a single break case, including the data from a very long time span is not reasonable. Hence, the sample period included in our empirical analysis is from 1952 to 2012.

We first consider the univariate predictor case. The results are reported in table 5. For most predictors in quarterly frequency case, our new test, in line with the corresponding

[15]The latest updated version of this dataset can be retrieved from Amit Goyal's website: http//www.hec.unil.ch/agoyal.

Table 5: Joint test with single predictor

predictors	quarterly IVX-Wald	$W_{\alpha\beta}$	monthly IVX-Wald	$W_{\alpha\beta}$
Dividend payout ratio	1.097	12.205	0.672	15.444**
Long-term yield	0.782	18.199**	1.396	26.175***
Dividend yield	2.235	9.445	1.425	5.888
Dividend-price ratio	1.525	4.795	1.142	3.649
T-bill rate	2.362	27.870***	3.537*	32.874***
Earnings-price ratio	0.518	4.617	0.588	5.862
Book-to-market value ratio	0.546	14.074*	0.174	7.453
Default yield spread	0.329	4.506	0.389	5.620
Net equity expansion	0.06	10.934	0.220	15.131**
Term spread	3.057*	17.477**	3.808*	19.986**
Inflation rate	2.356	18.265**	5.922**	19.743**
Consumption-wealth ratio	11.351***	17.360**	-	-

1. The critical value for joint test is obtained by simulation. *,** and *** denote rejection of null hypothesis at 10%, 5% and 1% level respectively.
2. For consumption-wealth ratio, only quarterly data is available.

IVX-Wald result, fails to reject the null hypothesis and thus implies no predictability. The variables within this group consist of d/p, d/y, d/p, e/p, b/m, dfy and ntis. More importantly, it also suggests that for these particular predictors, the absence of predictability is stable over time without any structural break. In this sense, it is more informative than the original IVX-Wald test, and we can safely argue that these variables are not valid predictors for stock returns. Notable exceptions are long-term yield, T-bill rate, and inflation rate. When tested using IVX-Wald, the results are highly insignificant, indicating no predictability. However, under our joint test, the null hypothesis is strongly rejected. This seems to be a strong signal supporting the existence of structural break during the sample period for these predictors. The situation for monthly data is similar, with a few rather minor differences. For example, d/p and ntis are significantly rejected by joint test when monthly data is used. Again, the predictive regression using long-term yield and T-bill rate as predictor exhibits stark difference under IVX-Wald and the joint test, potentially suffering a break in coefficients.

Now, let us turn to the predictive regressions with multiple predictors. Multivariate regressions are more informative when used to testing for market efficiency compared with univariate regressions, as predictability should be evaluated with respect to the entire information set. Besides, recent advances in financial economics also point out that using the dividend-price ratio alone is not enough for characterizing expected stock returns.

Table 6: Joint test with multiple predictors

predictors	quarterly		monthly	
	IVX-Wald	$W_{\alpha\beta}$	IVX-Wald	$W_{\alpha\beta}$
Ang and Cekaert (2007)	3.745	13.487	4.132	16.545*
Ferson and Schadt (1996)	6.880	27.850**	7.653	36.720***
Kothari and Shanken (1997)	1.883	11.4753	2.085	7.331
Lamont (1998)	1.954	12.042	1.326	15.264
Campbell and Vuolteenaho (2004)	4.574	19.786	5.42	20.180*
Lettau and Ludvigson (2001)	13.199***	28.437**	-	-
General-to-specific	23.985***	41.564***	8.160**	25.853**

1. The critical value for $supW$ test is obtained by simulation. *,** and *** denote rejection of null hypothesis at 10%, 5% and 1% level respectively.
2. For the model in Lettau and Ludvigson (2001), only data for quarterly version is available.

Hence, it is interesting to explore whether a set of variables jointly predict the future equity return movement and whether the predictive power is stable over time. Following KMS, seven combinations of potential predictors, all of which are either theoretically or empirically motivated, are considered: (1) d/p and tbl (Ang and Bekaert, 2007), (2) d/p, tbl, dfy and tms (Ferson and Schadt, 1996), (3) d/p and b/m (Kothari and Shanken 1997), (4) d/p and d/e (Lamont, 1998), (5)e/p, tms, and b/m (Campbell and Vuolteenaho, 2004), (6) d/p, d/e and cay (Lettau and Ludvigson, 2001) and (7) model selected through a general-to-specific approach. The best model suggested by the general-to-specific approach is e/p and tbl for the monthly data, and e/p, tbl, dfy and cay for the quarterly case. Results for multivariate regressions are reported in Table 6. For most models considered, the conclusions obtained by using the IVX Wald test are reaffirmed by our joint test, indicating no structural break occurs. The only major difference comes from the model proposed in Ferson and Schadt (1996), which is significantly rejected by our joint test, while not rejected by IVX Wald test.

3.7 Conclusion

In this chapter, we consider a novel testing procedure built upon the recently developed technique for inference in predictive regressions. The new test exhibits reasonable size control under the joint null hypothesis and also has excellent power against both predictability and the existence of structural break. In the same spirit of KMS, our test is easy to implement and robust to (i). various degrees of persistence and (ii). the strong corre-

lation between innovations to predictive regression and those driving predictors. Asymptotic properties of the test are established and extensive simulation studies suggest the excellent size and power performance in finite sample. The major drawback of this joint test is that when the null hypothesis is rejected, we do not know whether it is due to a structural break or the existence of predictability. In that case, one may wish to use the recently proposed structural break test for predictive regression and the IVX Wald statistic for the full sample to further investigate the reason for rejection. A few extensions of our test are possible. Examples include testing for multiple breaks and smooth structural change. These are left for future work.

4 Testing for Predictability under Level Shifts of Predicted Variable

4.1 Introduction

As is shown in the simulation study conducted in Section 3.3, when the intercept of predictive regression experiences structural breaks, the IVX-Wald test will suffer severe size distortion. This phenomenon can be viewed as a signal of the potential spurious regression effect and is worth a detailed study. In this chapter, we consider the following model

$$y_t = \begin{cases} \alpha_1 + \beta_1' x_{t-1+} + u_{yt} & , 0 \leqslant t < t_1 \\ \alpha_2 + \beta_2' x_{t-1+} + u_{yt} & , t_1 \leqslant t < t_2 \\ \vdots & \vdots \\ \alpha_{m+1} + \beta_{m+1}' x_{t-1+} + u_{yt} & , t_m \leqslant t < T \end{cases}, \tag{4.1}$$

$$x_t = \left(I_k - \frac{C}{T^{\gamma_x}} \right) x_{t-1} + u_{xt}, \tag{4.2}$$

where γ_x is a positive real number, C is a diagonal matrix and independent of sample size. Innovations u_{yt} and u_{xt} satisfy the same conditions as in the last chapter. In this model, we have in total m breaks and thus $m+1$ different regimes.

Assumption 4.1 *The innovation to x_t is a linear process and can be represented by $u_{xt} = \sum_{j=0}^{\infty} C_j \epsilon_{t-j}$, where $\{C_j\}_{j=0}^{\infty}$ is a sequence of absolutely summable constant matrices such that $\sum_{j=0}^{\infty} C_j$ has full rank and $C_0 = I_k$. Let $u_t = (u_{yt}, \epsilon_t')'$, then u_t is a martingale difference sequence with constant conditional variance matrix Σ and satisfies the moment condition that $\sup_t E||u_t||^{2s} < \infty$ for some $s > 1$. Denote σ_y as the first diagonal element of Σ.*

To accommodate multiple breaks, we modify Assumption 3.2 to

Assumption 4.2 *The fractions of structural breaks defined as $\tau_1 \equiv \frac{t_1}{T}$, ..., $\tau_m \equiv \frac{t_m}{T}$ satisfy that $0 = \tau_0 < \tau_1 < ... < \tau_m < \tau_{m+1} = 1$.*

4.2 Level Shifts and Spurious Predictive Power

In this section, we derive the limiting distribution of OLS and IVX-based t-statistic under the data generating process (4.1)-(4.2). For simplicity, we consider the univariate predictor case and thus set k to be one. It is shown that the results depend critically on the magnitude of the shifts, as well as the degree of persistence of the predictor.

Specifically, the OLS based t-statistic is computed as

$$t_{\beta}^{ols} = \frac{\hat{\beta}^{ols}}{\hat{\sigma}_y \sqrt{\sum_{t=1}^{T}(x_{t-1} - \bar{x})^2}},$$

where

$$\hat{\beta}^{ols} = \left[\sum_{t=1}^{T}(x_{t-1} - \bar{x})^2\right]^{-1} \sum_{t=1}^{T}(x_{t-1} - \bar{x})(y_t - \bar{y}),$$

and $\bar{x}$, $\bar{y}$ are the sample average of predictor and predicted variable, respectively. The standard deviation of the predicted variable, σ_y^2, is estimated by

$$\hat{\sigma}_y^2 = \frac{1}{T}\sum_{t=1}^{T}\left[y_t - \hat{\alpha} - x_{t-1}\hat{\beta}^{ols}\right]^2,$$

where $\hat{\alpha} = \bar{y} - \bar{x}\hat{\beta}^{ols}$ is the least-squares estimator of the intercept term. The IVX t-statistic, on the other hand, is defined as

$$t_{\beta}^{ivx} = \frac{\tilde{\beta}^{ivx}}{\hat{\sigma}_y \sqrt{\left[\sum_{t=1}^{T}\tilde{z}_{t-1}^2 - T\bar{z}^2(1-\hat{\rho}_{xy}^2)\right]\left[\sum_{t=1}^{T}\tilde{z}_{t-1}(x_{t-1} - \bar{x})\right]^{-2}}},$$

where

$$\tilde{\beta}^{ivx} = \left[\sum_{t=1}^{T}\tilde{z}_{t-1}(x_{t-1} - \bar{x})\right]^{-1} \sum_{t=1}^{T}\tilde{z}_{t-1}(y_t - \bar{y}),$$

and IVX is constructed by

$$\tilde{z}_t = (1 - \frac{1}{T^{\delta_z}})\tilde{z}_{t-1} + (x_t - x_{t-1}).$$

$\bar{z}$ is the sample average of the constructed IVX and $\hat{\rho}_{xy}$ is the estimated correlation co-

efficient between u_{xt} and u_{yt}. Note that the estimator for σ_y^2 is the same as in the least squares case.

In summary, we consider following cases:

- For the OLS estimator, three cases are considered:
  - 1. $\gamma_x = 0$, i.e. stationary predictors
  - 2. $0 < \gamma_x < 1$, i.e. mildly stationary predictors
  - 3. $\gamma_x \geqslant 1$, i.e. local-to-unit root or unit root
- For the IVX estimator, also three cases are considered:
  - 1. $\delta_z < \gamma_x$, which includes local-to-unit-root and unit-root predictors cases
  - 2. $\delta_z \geqslant \gamma_x$ and $\gamma_x + \delta_z \geqslant 1$
  - 3. $\delta_z \geqslant \gamma_x$ and $\gamma_x + \delta_z < 1$
- For all cases above with the nonstationary predictor, we discuss three types of break, whose precise definitions are case-dependent:
  - 1. small breaks that are asymptotically negligible
  - 2. moderate breaks that lead to a non-standard $O_p(1)$ t-statistic
  - 3. large breaks that drive the t-statistic going to infinity

Theorem 4.1 below shows that, when the predictor is not persistent, the existence of level shifts in the predicted variable will not distort the standard inference. The intuition of this result is straightforward. Taking the single break case with a stationary predictor as an example, it is straightforward to show that

$$\sqrt{T}(\hat{\beta}^{ols} - \beta) \to^d N\left(0, (\sigma_y^2 + (\alpha_1 - \alpha_2)\tau(1-\tau))E(x_1^2)^{-1}\right),$$

$$\hat{\sigma}_y^2 \to^p \sigma_y^2 + (\alpha_1 - \alpha_2)\tau(1-\tau),$$

and thus, the effect of breaks will cancel out, leaving the limiting distribution of the t-statistic unchanged.

Theorem 4.1 *If the predictor is stationary, i.e., $\gamma_x = 0$, then under the null hypothesis $H_0 : \beta_1 = \beta_2 = ... = \beta_{m+1} = 0$, $t_\beta^{ols} \Rightarrow N(0,1)$ regardless of the number and size of level shifts.*

Now, we consider the cases of non-stationary predictors, which are empirically more relevant. Before discussing IVX-based statistic, we first examine the influence of breaks on OLS t-statistic, which is still used in some applied research. We first concentrate on the case that the predictor is highly persistent, which includes local-to-unit-root and unit-root processes.

Theorem 4.2 *Let*

$$|\alpha_{i+1} - \alpha_i| = \frac{d_{i,T}}{\sqrt{T}}$$

be the break magnitude of i^{th} shift. As T goes to infinity, we call a break

- *'small' if $d_{i,T} = o(1)$,*
- *'moderate' if $d_{i,T} = d_i$ is independent of T,*
- *'large' if $d_{i,T} = O(T^\epsilon)$, for some $\epsilon \in (0, 1/2]$.*

Assume that the predictor is generated by a local-to-unit-root or unit root process, i.e. $\gamma_x \geqslant 1$, then under the null hypothesis $H_0 : \beta_1 = \beta_2 = ... = \beta_{m+1} = 0$,

- *(i) if all the level shifts are small,*

$$t_\beta^{ols} \Rightarrow \frac{\int_0^1 \underline{J}_c(r) dW(r)}{\sqrt{\int_0^1 \underline{J}_c^2(r) dr}};$$

- *(ii) if all the level shifts are moderate,*

$$t_\beta^{ols} \Rightarrow \frac{\int_0^1 \underline{J}_c(r) dW(r)}{\sqrt{\int_0^1 \underline{J}_c^2(r) dr}} + \frac{(\sum_{i=1}^m \tau_i d_i) \int_0^1 \underline{J}_c(r) dr - \left[\sum_{i=1}^{m+1} (\sum_{j=i}^m d_i) \int_{\tau_{i-1}}^{\tau_i} \underline{J}_c(r) dr\right]}{\sigma_y \sqrt{\int_0^1 \underline{J}_c^2(r) dr}};$$

- *(iii) if the size of any shift is large,* $t_\beta^{ols} \to \infty$.

Next, we derive the limiting behavior of the OLS t-statistic with a mildly stationary predictor. In this case, the limiting distribution is Gaussian, but with a larger asymptotic variance.

Theorem 4.3 *Let*

$$|\alpha_{i+1} - \alpha_i| = \frac{d_{i,T}}{T^{\gamma_x/2}}$$

be the break magnitude of i^{th} *shift. As T goes to infinity, we call a break*

- *'small' if* $d_{i,T} = o(1)$*;*
- *'moderate' if* $d_{i,T} = d_i$ *is independent of T;*
- *'large' if* $d_{i,T} = O(T^\epsilon)$*, for some* $\epsilon \in (0, \gamma_x/2]$.

Assume that the predictor is generated by a mildly stationary process, i.e., $0 < \gamma_x < 1$*, then the under the null hypothesis* $H_0 : \beta_1 = \beta_2 = ... = \beta_{m+1} = 0$*,*

- *(i) if all the level shifts are small,*

$$t_\beta^{ols} \Rightarrow N(0, 1);$$

- *(ii) if all the level shifts are moderate,*

$$t_\beta^{ols} \Rightarrow N\left(0, 1 + \frac{1}{\sigma_y^2} \sum_{i=1}^{m+1} \left[\sum_{i=1}^{m} \tau_i d_i - \sum_{j=i}^{m} d_j\right]^2 (\tau_i - \tau_{i-1})\right);$$

- *(iii) if the size of any shift is large,* $t_\beta^{ols} \to \infty$.

Now we turn to the instrumental variable-based test. Theorem 4.4 below shows that, even if the IVX-based t test replaces the OLS, the spurious effect is still present. For these results, we need to modify the classification of break magnitude. This is due to the fact that OLS and IVX estimator has a different convergence rate when the predictor is non-stationary.

Theorem 4.4 *Let*

$$|\alpha_{i+1} - \alpha_i| = \frac{d_{i,T}}{T^{\frac{\delta_z + \min[0,\gamma_x - 1]}{2}}}$$

be the break magnitude of i^{th} shift. As T goes to infinity, we call a break

- *'small' if $d_{i,T} = o(1)$;*
- *'moderate' if $d_{i,T} = d_i$ is independent of T;*
- *'large' if $d_{i,T} = O(T^{\epsilon})$, for some $\epsilon \in (0, \frac{\delta_z + \min[0,\gamma_x - 1]}{2}]$.*

Assume that the chosen persistence level for IVX is lower than that of the process generating the predictor, i.e. $\delta_z < \gamma_x$, then under the null hypothesis $H_0 : \beta_1 = \beta_2 = ... = \beta_{m+1} = 0$,

- *(i) if all the level shifts are small,*

$$t_{\beta}^{ivx} \Rightarrow N(0,1);$$

- *(ii) if all the level shifts are moderate,*

$$\begin{aligned} t_{\beta}^{ivx} &\Rightarrow N(0,1) \\ &+ \left(\sum_{i=1}^{m} \tau_i d_i\right) \frac{\underline{J}_c(1)}{\sigma_y/\sqrt{2}} - \left[\sum_{i=1}^{m} \left(\sum_{j=i}^{m} d_i\right) \frac{\underline{J}_c(\tau_{k+1}) - \underline{J}_c(\tau_k)}{\sigma_y/\sqrt{2}}\right]; \end{aligned}$$

- *(iii) if the size of any shift is large, $t_{\beta}^{ivx} \to \infty$.*

Theorem 4.5 *Let*

$$|\alpha_{i+1} - \alpha_i| = \frac{d_{i,T}}{T^{\frac{\delta_z + \gamma_x - 1}{2}}}$$

be the break magnitude of i^{th} shift. As T goes to infinity, we call a break

- *'small' if $d_{i,T} = o(1)$;*
- *'moderate' if $d_{i,T} = d_i$ is independent of T;*
- *'large' if $d_{i,T} = O(T^{\epsilon})$, for some $\epsilon \in (0, \frac{\delta_x + \gamma_z - 1}{2}]$.*

Assume that the chosen persistence level for IVX satisfies that $\delta_z \geqslant \gamma_x$ *and* $\gamma_x + \delta_z > 1$*, then under the null hypothesis* $H_0 : \beta_1 = \beta_2 = ... = \beta_{m+1} = 0$*,*

- *(i) if all the level shifts are small,*

$$t_\beta^{ivx} \Rightarrow N(0,1);$$

- *(ii) if all the level shifts are moderate,*

$$t_\beta^{ivx} = N(0,1) + O_p(1);$$

- *(iii) if the size of any shift is large,* $t_\beta^{ivx} \to \infty$*.*

Theorem 4.6 *Assume that for all* $i = 1, ..., m$*, we have* $|\alpha_{i+1} - \alpha_i| = \frac{d_i}{T^\epsilon}$*, where* d_i *is a constant and* $\epsilon \geqslant 0$*. Also assume that* $\gamma_x + \delta_z < 1$*. Then under the null hypothesis,* $t_\beta^{ivx} \Rightarrow N(0,1)$*.*

To visually demonstrate the consequences of level shifts on predictability test, we plot the empirical quantile of IVX-Wald and OLS-Wald statistic, together with that of a $\chi^2(1)$ random variable. For simplicity, we consider a single break that occurs in the middle of the sample. Figure 11 shows the case without a shift. In this case, as expected, the OLS-based statistic is far from being $\chi^2(1)$, while IVX-Wald is well approximated by this limiting distribution. Figure 12 presents the case with a break of magnitude $d = 5$. It can be seen that the empirical quantile of the IVX-Wald statistic moves rightward and thereby deviates from $\chi^2(1)$, although the deviation is less severe than OLS. When the break magnitude doubles to $d = 10$, as shown in Figure 13, the situation is even worse as the quantile of the IVX-Wald statistic shifts further rightward. In summary, when the level shift of the predicted variable occurs, both IVX and OLS-based inference will suffer distortion and this distortion deteriorates as the break magnitude increases.

As suggested by the simulation and theoretical results, there may exist 'spurious' predictability, if y is stationary but experiences some level shifts. This is because stationary series with mean shifts behaves like process with long-range dependence or nonstationarity. See, e.g., Granger and Hyung (2004). Therefore, regressing a stationary variable

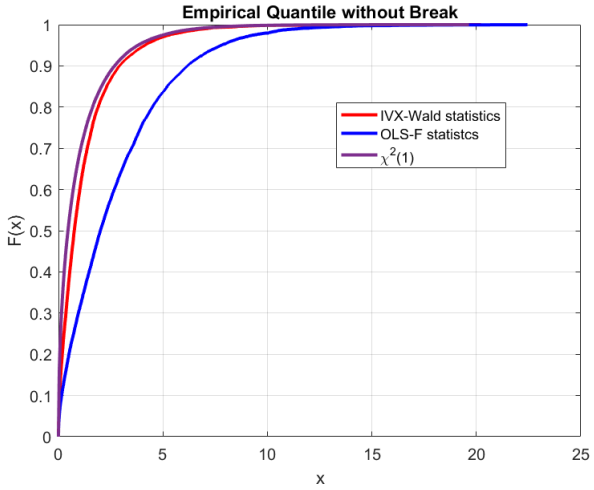


Figure 11: Empirical quantiles of OLS and IVX vs. Asymptotic Distribution (no break)

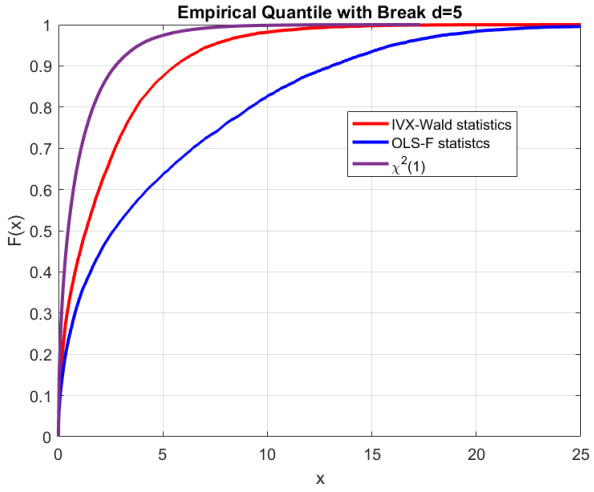


Figure 12: Empirical quantiles of OLS and IVX vs. Asymptotic Distribution (d=5)

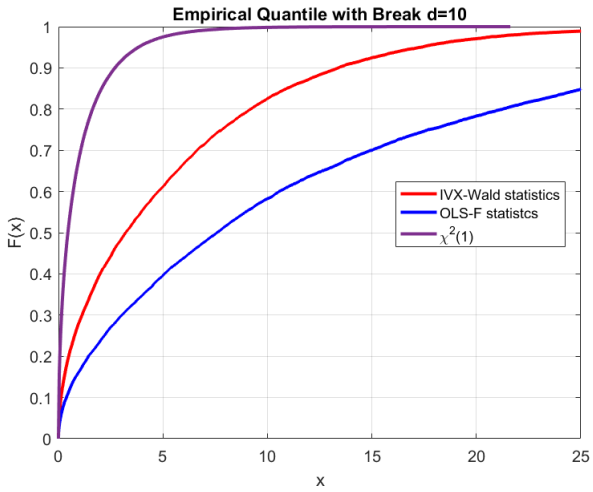


Figure 13: Empirical quantiles of OLS and IVX vs. Asymptotic Distribution (d=10)

with level shifts on a (nearly)-integrated variable may result in a spurious relationship, even if they are completely independent. This type of spurious regression is considered by, for instance, Noriega and Ventosa-Santaularia (2006). However, they do not derive the limiting distribution and only consider a fixed break magnitude.

4.3 Testing for Predictability With Large Level Shifts

4.3.1 Sampling-splitting-based IVX-Wald Test

Note that here, the hypothesis we wish to test is $H_0 : \beta_1 = \beta_2 = ... = \beta_{m+1} = 0$. Hence, under the null, the model reduces to a break-in-mean model

$$y_t = \begin{cases} \alpha_1 + u_{yt} & , 0 \leqslant t \leqslant t_1 \\ \alpha_2 + u_{yt} & , t_1 < t \leqslant t_2 \\ \vdots & \vdots \\ \alpha_{m+1} + u_{yt} & , t_m < t \leqslant T \end{cases}.$$

The inference theory for such a model is well studied in the literature, such as Bai and Perron (1998 BP hereafter). In this section, we consider the case with large shifts, which can be accurately detected by BP's method.

Assumption 4.3 *The break sizes can be expressed as* $|\alpha_i - \alpha_{i-1}| = \frac{d_{i,T}}{T^\epsilon}$, *where* $d_{i,T} = O(T^\epsilon)$, *for some* $\epsilon \in (0, 1/2]$.

Assumption 4.3 imposes that the break size either is independent of sample size or shrinking towards zero, but with a rate of convergence slower than $T^{-1/2}$. This assumption is used to ensure that we can consistently estimate the break fraction under the null hypothesis. The new procedure we propose, which is based on sample-splitting, contains the following three steps:

- 1. Estimate the break date(s) of the null break-in-mean model using the traditional method as described in Bai and Perron (1998).
- 2. Split the data set into disjoint regimes according to those estimated date(s).
- 3. Calculate the IVX-Wald test statistic in each individual subperiod and aggregate them to obtain the final statistic.

Denote the fraction of break date(s) estimated in the Step 1 as $(\hat{\tau}_0, ..., \hat{\tau}_{\hat{m}+1})$, where we set $\hat{\tau}_0 = 0$ and $\hat{\tau}_{\hat{m}+1} = 1$ following the convention. For a given number of breaks m, these break dates can be easily estimated by:

$$(\hat{\tau}_1, ..., \hat{\tau}_m) = \frac{1}{T}(\hat{t}_1, ..., \hat{t}_m),$$

$$(\hat{t}_1, ..., \hat{t}_m) = \underset{(t_1,...,t_m)}{\arg\min}\, S_n(t_1, ..., t_m),$$

where $S_n(t_1, ..., t_m)$ is the minimized sum of squared residuals

$$\sum_{i=1}^{m+1} \sum_{i=t_{i-1}+1}^{t_i} \left(y_i - \frac{1}{t_i - t_{i-1}} \sum_{i=t_{i-1}}^{t_i} y_i \right)^2$$

computed given a particular set of break dates. To consistently estimate the number of breaks, a few alternative methods can be applied. Here, for convenience, we adopt the sequential method, which is based on testing (l+1) breaks against l breaks one at time starting from some pre-specified maximum possible number of breaks.

Correspondingly, we will denote the IVX-Wald statistic calculated using data from $\hat{\tau}_i$ to $\hat{\tau}_{i+1}$ by $W^{ivx}(i)$, where i runs from 1 to $\hat{m} + 1$. Specifically,

$$W^{ivx}(i) = \tilde{\beta}^{ivx\prime}(i)\tilde{Q}(i)^{-1}\tilde{\beta}^{ivx}(i), \tag{4.3}$$

$$\tilde{\beta}^{ivx}(i) = \left[\sum_{t=\hat{t}_{i-1}+1}^{\hat{t}_i} \tilde{z}_{t-1}(x_t - \bar{x}_i)'\right]^{-1} \sum_{t=\hat{t}_{i-1}+1}^{\hat{t}_i} \tilde{z}_{t-1}(y_t - \bar{y}_i), \tag{4.4}$$

where $\bar{x}_i = \frac{1}{t_i - t_{i-1}} \sum_{t=t_{i-1}}^{t_i} x_t$ and $\bar{y}_i = \frac{1}{t_i - t_{i-1}} \sum_{t=t_{i-1}}^{t_i} y_t$.

The following theorem indicates that under the null hypothesis, the test statistic,

$$W_b = \sum_{i=0}^{\hat{m}} W^{ivx}(i),$$

will converge to a chi-squared distribution with degree of freedom equal to the number of subperiods times the number of predictors.

Theorem 4.7 *Under Assumption 4.1-4.3 and the null hypothesis that* $H_0 : \beta_1 = \beta_2 = ... = \beta_m = 0$*, as* $T \to \infty$*, we have* $W_b \to \chi^2((m+1)k)$*, where* m *is the number of breaks and* k *is the number of predictors considered.*

This result is intuitive. If the level shifts of the predicted variable are relatively large, we can consistently detect them and pin down their locations with high precision. In particular, the convergence rate of the break fraction estimator will be faster than that of the IVX estimator of β. Hence, asymptotically, it is as if we know the true dates of those breaks.

4.3.2 Size and Power under Large Level Shifts

In this section, we examine the finite sample performance of the sample-splitting-based test proposed above. For all simulation bellow, we consider a single predictor whose level of persistence is $c = 10$ and the correlation between u_{xt} and u_{yt} is $\rho = -0.9$. Three sample size, $T = 200$, $T = 400$ and $T = 800$, are considered. We first focus on the single break case, in which we assume

$$y_t = \begin{cases} \alpha_1 + u_{yt} & , 0 \leqslant t \leqslant t_1 \\ \alpha_2 + u_{yt} & , t_1 \leqslant t \leqslant T \end{cases}, \tag{4.5}$$

where $\alpha_1 = 0$, $\alpha_2 = \{12, 14, ..., 20\}/\sqrt{T}$ and $t_1 = \{0.3, 0.4, 0.5\} \times T$. The rejection rates with the nominal levels of 0.1, 0.05, and 0.01 are reported in Table 7. In general, the size control is satisfactory, especially when the sample size is large enough. Meanwhile, as expected, the performance of the test is better when the break is closer to the middle of the sample period. The finite sample size control under multiple breaks are reported in Table 8. The data generating process (DGP) for the predicted variable is now

$$y_t = \begin{cases} \alpha_1 + u_{yt} & , 0 \leqslant t < t_1 \\ \alpha_2 + u_{yt} & , t_1 \leqslant t < t_2 \\ \alpha_3 + u_{yt} & , t_2 \leqslant t \leqslant T \end{cases}, \tag{4.6}$$

and

$$y_t = \begin{cases} \alpha_1 + u_{yt} & , 0 \leqslant t < t_1 \\ \alpha_2 + u_{yt} & , t_1 \leqslant t < t_2 \\ \alpha_3 + u_{yt} & , t_2 \leqslant t < t_3 \\ \alpha_4 + u_{yt} & , t_3 \leqslant t \leqslant T \end{cases}, \tag{4.7}$$

For the DGP (4.6), two breaks occur and we consider $t_1 = 0.3 \times T$ and $t_2 = \{0.5, 0.7\} \times T$. For this two-break scenario, break magnitude are determined by $\alpha_1 = \alpha_3 = 0$ and $\alpha_2 = \{20, 22, ..., 30\}/\sqrt{T}$. As for the DGP (4.7), we set $t_1 = 0.3 \times T$, $t_2 = 0.5 \times T$ and $t_3 = 0.7 \times T$. For this three-break scenario, break magnitude are determined by $\alpha_1 = \alpha_4 = 0$, $\alpha_2 = \{30, 32, ..., 40\}/\sqrt{T}$ and $\alpha_3 = 2\alpha_2$. In all cases, again, the size control is good when T is large. An important conclusion drown from this study is that, in finite samples, the precise meaning of 'large' break depends on the true number of shifts. For a fixed sample size, when the number of breaks increases, we need level shifts to be larger for our procedure to be satisfactory.

We next turn to the power performance of the test against the departure from null hypothesis. The DGP for predicted variable is

$$y_t = \begin{cases} \alpha_1 + \beta_1 x_{t-1} + u_{yt} & , 0 \leqslant t < t_1 \\ \alpha_2 + \beta_2 x_{t-1} + u_{yt} & , t_1 \leqslant t \leqslant T \end{cases}. \tag{4.8}$$

Two different cases are considered. In the first case, no break in intercept occurs and thus $\alpha_1 = \alpha_2 = 0$. In another case, $\alpha_1 = 0$, $\alpha_2 = 30/\sqrt{T}$ and thus one large break is present.

Table 7: Finite sample size performance under single large break

	$T = 200$			$T = 400$			$T = 800$		
level	0.1	0.05	0.01	0.1	0.05	0.01	0.1	0.05	0.01
	$\tau = 0.3$								
d=12	0.13	0.07	0.02	0.12	0.06	0.01	0.12	0.06	0.02
d=14	0.12	0.06	0.02	0.11	0.05	0.01	0.11	0.06	0.01
d=16	0.12	0.06	0.02	0.11	0.05	0.01	0.11	0.06	0.01
d=18	0.12	0.06	0.02	0.11	0.05	0.01	0.11	0.06	0.01
d=20	0.11	0.06	0.02	0.11	0.05	0.01	0.11	0.06	0.01
	$\tau = 0.4$								
d=12	0.12	0.06	0.02	0.11	0.06	0.01	0.12	0.06	0.01
d=14	0.11	0.06	0.02	0.11	0.06	0.01	0.11	0.06	0.01
d=16	0.11	0.06	0.01	0.11	0.05	0.01	0.11	0.05	0.01
d=18	0.11	0.06	0.01	0.11	0.05	0.01	0.11	0.05	0.01
d=20	0.11	0.06	0.01	0.11	0.05	0.01	0.11	0.06	0.01
	$\tau = 0.5$								
d=12	0.12	0.06	0.02	0.12	0.06	0.01	0.12	0.06	0.01
d=14	0.12	0.06	0.01	0.11	0.06	0.01	0.12	0.06	0.01
d=16	0.12	0.06	0.01	0.11	0.06	0.01	0.12	0.06	0.01
d=18	0.11	0.06	0.01	0.11	0.05	0.01	0.12	0.06	0.01
d=20	0.11	0.06	0.01	0.11	0.05	0.01	0.11	0.06	0.01

Table 8: Finite sample size performance under multiple large breaks

	$T = 200$			$T = 400$			$T = 800$		
level	0.1	0.05	0.01	0.1	0.05	0.01	0.1	0.05	0.01
	$\tau_1 = 0.3, \tau_2 = 0.5$								
d=20	0.14	0.08	0.02	0.13	0.07	0.02	0.13	0.06	0.01
d=22	0.14	0.08	0.02	0.13	0.07	0.01	0.13	0.06	0.01
d=24	0.14	0.08	0.02	0.12	0.06	0.01	0.12	0.06	0.01
d=26	0.14	0.08	0.02	0.12	0.06	0.01	0.12	0.06	0.01
d=28	0.14	0.08	0.02	0.12	0.06	0.01	0.12	0.06	0.01
d=30	0.14	0.07	0.02	0.12	0.06	0.01	0.12	0.06	0.01
	$\tau_1 = 0.3, \tau_2 = 0.7$								
d=20	0.13	0.07	0.02	0.12	0.06	0.02	0.13	0.06	0.01
d=22	0.13	0.07	0.02	0.12	0.06	0.01	0.13	0.06	0.01
d=24	0.13	0.07	0.02	0.12	0.06	0.01	0.12	0.06	0.01
d=26	0.13	0.07	0.02	0.12	0.06	0.01	0.12	0.06	0.01
d=28	0.12	0.07	0.02	0.12	0.06	0.01	0.12	0.06	0.01
d=30	0.12	0.07	0.02	0.12	0.06	0.01	0.12	0.06	0.01
	$\tau_1 = 0.3, \tau_2 = 0.5, \tau_3 = 0.7$								
d=30	0.15	0.09	0.02	0.13	0.07	0.01	0.12	0.06	0.01
d=32	0.15	0.09	0.02	0.13	0.07	0.01	0.12	0.06	0.01
d=34	0.15	0.08	0.02	0.12	0.07	0.01	0.12	0.06	0.01
d=36	0.15	0.08	0.02	0.12	0.07	0.01	0.12	0.06	0.01
d=38	0.15	0.08	0.02	0.12	0.07	0.01	0.12	0.06	0.01
d=40	0.15	0.08	0.02	0.12	0.07	0.01	0.12	0.06	0.01

Table 9: Finite sample power of W_b test

	$d = 30$			$d = 0$		
	$\tau = 0.5$	$\tau = 0.3$	$\tau = 0.7$	$\tau = 0.5$	$\tau = 0.3$	$\tau = 0.7$
b=3	0.14	0.14	0.13	0.14	0.13	0.13
b=6	0.18	0.16	0.16	0.19	0.16	0.16
b=9	0.22	0.20	0.20	0.26	0.20	0.21
b=12	0.30	0.25	0.24	0.36	0.26	0.27
b=15	0.40	0.32	0.31	0.47	0.32	0.34
b=18	0.53	0.41	0.40	0.58	0.39	0.41
b=21	0.66	0.50	0.50	0.66	0.45	0.48
b=24	0.78	0.62	0.61	0.73	0.51	0.54
b=27	0.87	0.73	0.71	0.79	0.57	0.60
b=30	0.92	0.82	0.80	0.83	0.61	0.65

We assume t_1 to be $\{0.3, 0.5, 0.7\} \times T$. As for slope coefficient, we set $\beta_1 = b/T$ and $\beta_2 = -b/T$, where $b \in \{3, ..., 30\}$. Empirical rejection rates for all parameterization considered are contained in Table 9. In both cases, we can clearly find a progression toward 1 of empirical power as b increases. Since we assume the opposite direction of predictability before and after the shift, the power is better when α also experiences a large break. However, even if α is constant over time, splitting the sample does not sacrifice much power. Another finding is that, the power of the test is better when the break is in the middle of the sample period compared with early or late break.

4.4 Testing for Predictability with Moderate Shift: Single Break Case

4.4.1 Issues with Moderate Break

A critical assumption for the validity of the sample-splitting-based test discussed above is that the shift magnitude of all breaks must be large enough in the sense that they either keep constant when sample size goes to infinity or shrink toward zero at a rate slower than $T^{-1/2}$. Essentially, this restriction is imposed to guarantee the successful estimation of the number of breaks and their locations.

As suggested by limiting theory derived in Section 4.2, a break of order $T^{-1/2}$ is asymptotically negligible as it will not incur size distortion if IVX-Wald is used to test for the predictability. Indeed, since we must have $\delta_z + \min[\gamma_x - 1, 0] < 1$ by construction, a break of order $T^{-1/2}$ goes to zero faster than that of order $T^{-\frac{\delta_z + \min[\gamma_x - 1, 0]}{2}}$. This implies

Table 10: Fraction of break detected by BP's Method

	$\tau_0 = 0.2$	$\tau_0 = 0.3$	$\tau_0 = 0.4$	$\tau_0 = 0.5$
$d = 1$	2%	3%	3%	3%
$d = 2$	4%	5%	6%	6%
$d = 3$	7%	10%	12%	12%
$d = 4$	12%	18%	22%	23%
$d = 5$	21%	31%	37%	39%
$d = 6$	32%	47%	55%	57%
$d = 7$	46%	63%	72%	74%
$d = 8$	61%	77%	84%	86%
$d = 9$	74%	88%	93%	94%

that in large samples, even if we can not detect such a break, the performance of our procedure should be unimpaired. However, in finite sample, this is not the case.

In this section, we consider this issue under the assumption that the maximum number of breaks is known to be one. The reason for this choice is that, to the best of our knowledge, in the existing literature, there exists no discussion on the multiple moderate breaks problem. In fact, even in the single break case, the accurate detection of the break date is challenging enough. This is first shown in Elliott and Muller (2007), which provides the limiting distribution of break fraction for the single moderate break case. Critically, the estimated fraction is inconsistent as it converges to true value plus some $O_p(1)$ term. This result is reaffirmed by Jiang et al. (2018) for a simpler break-in-mean case.

To see the consequences of moderate break magnitude on the estimated number of break as well as the break location, some simulation studies are conducted. We generate a sequence of stationary variables with break in mean using DGP (4.5) with $t_0 = \tau_0 \cdot T$ and $\alpha_1 = 0$, $\alpha_2 = \frac{d}{\sqrt{T}}$. Four values of τ_0 are considered, namely 0.2, 0.3, 0.4 and 0.5. We set d to be 1,2,...,9. For each case, we first investigate the detection of break using BP's method. Table 7 reports the percentile of correct detection of a break for 10000 simulations. One can find that for the value of d we consider, BP's method is unreliable as a large fraction of moderate break is undetected. This is especially the case for break occurring far from the middle. For instance, when $d = 5$ and true break fraction being 0.2, the rate of correct detection is only 21%. That means, in finite sample, applying BP's method will not split the sample, which leads to the size distortion caused by the break undetected.

Next, we examine the consequence of moderate break magnitude on the estimation of

Table 11: Summary of Different Cases

	Break Magnitude		
	Small	Moderate	Large
Accurate Detection	No	No	Yes
Spurious Effect	No	Yes	Yes

break fraction. To this end, we choose $d = 3$ as an example and plot the kernel-smoothed finite-sample distribution of estimated break fraction for $\tau = 0.2, 0.3, 0.4, 0.5$. To remove the influence of the underestimated number of breaks, in this exercise, we estimate a break date without estimating the number of breaks. As can be seen from Figure 14, the finite sample distribution suffers issues like skewness and tri-modality. As argued in Jiang et al. (2018), these features lead to an estimator that is biased even asymptotically. Hence, for a moderate break case, even if the number of breaks is known, we still can not pin down them with high precision.

Table 11 summarizes the cases we have considered. When breaks are small, they are asymptotically negligible, and hence no special treatment is necessary. When the breaks are large, they will lead to spurious regression and severe size distortion. However, since they are easy to detect, we can use BP's method to split the sample first and construct a new statistic, which delivers valid inference. The most challenging situation is when the breaks are moderate. They will distort the distribution of the IVX-Wald test statistic in finite sample, while in the meantime, we can accurately pin down neither the number of breaks nor their locations.

To deal with this problem, we propose a partial solution based on the simulation results shown in the next subsection. The idea is that, compared with the underestimation of the break number, the uncertainty of the break date induced by a moderate break seems to be of secondary importance. Hence, we propose to fix the break number instead of estimating it. As will be shown, this greatly alleviates the size distortion when the break is not large enough.

One intuitive explanation of this result is the following. The biased estimation of break fraction is relevant only for d that is quite small. For those cases, however, the distortion of limiting distribution induced by the break is relatively marginal, which means even if we estimate an incorrect break date far from the true one, no serious problem occurs. This is

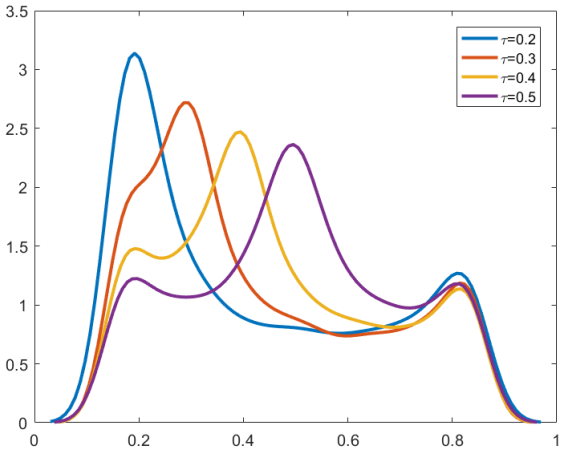


Figure 14: Finite sample distribution of estimated break fraction when $d = 3$

not the case for the underestimation of the break number. From Table 10, we see that even when $d = 7$, only half of the breaks are detected. Ignoring the break of this magnitude then causes a sever distortion. Hence, in finite sample with a moderate shift magnitude, pre-specifying the number of shifts is more important than estimating the break fraction accurately.

4.4.2 Size Control Under Single Moderate Break

In this section, we compare different estimators under the moderate break case. Here, again, we focus on the scenario in which the maximum number of the level shifts is known to be one. We compare the size control of sample-splitting estimator based on the estimated number of breaks as well as its counterpart with a pre-specified number of shifts. Three break fractions are considered, namely $\tau = 0.5$, $\tau = 0.2$ and $\tau = 0.8$. For $\tau = 0.5$ case, we also examine two persistence level, $c = 0$ and $c = 10$. To understand the consequences of break magnitude on size control, we compute the rejection rate with $d \in (0, 0.25, 0.5, ..., 10)$ for IVX-Wald, adjusted IVX with estimated break number, and adjusted IVX with pre-specified break number. The results are summarized in Figure 15-18. As expected, the IVX-Wald test over-rejects seriously quickly after the break size deviates from zero. Meanwhile, it can also be found that, in each case, for shift magnitude in some particular interval, the test statistic based on the data-dependent number

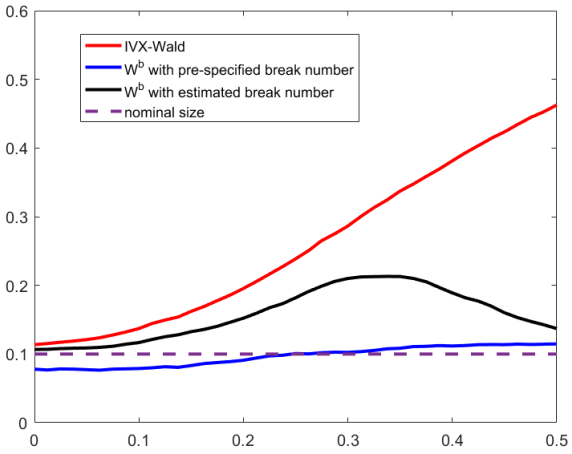


Figure 15: Rejection rate of three test statistic under different structural break magnitude when $c = 10$ and $\tau = 0.5$

of breaks will suffer size distortion. As discussed above, this is due to the fact that the number of breaks is underestimated, and the estimated break fraction is biased and inconsistent. However, as long as the number of shifts is fixed to be the true value, one, the size distortion almost disappears for all cases we consider. Hence, in practice, at least for the single break case, pre-specifying the number of breaks is a partial solution for the moderate break problem.

4.5 Empirical Illustrations

To illustrate the empirical relevance of the issues discussed above, we consider two examples in this section. The first example is concerned with the absolute returns of the stock market index, which serves as a proxy for volatility. The second one is on the predictability of macroeconomic variables on the Housing Price Index (HPI).

4.5.1 Predicting Absolute Equity Returns

The first empirical application that we consider is the predictability of a few financial and macroeconomic variables on the volatility of the S&P 500 index. A few candidate proxies for volatility are available, including the one inferred from the GARCH model or some realized measures. In this chapter, we opt for the absolute return. Taking the

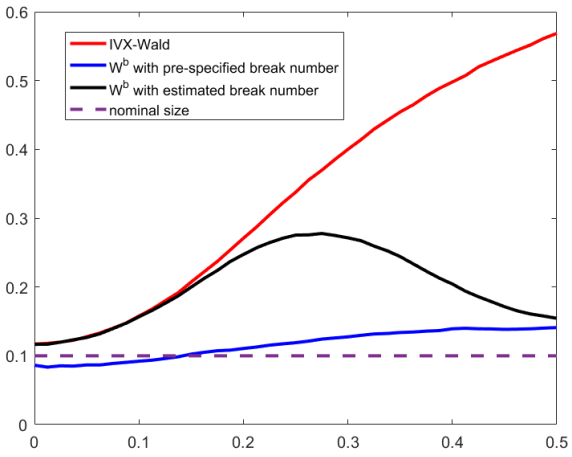


Figure 16: Rejection rate of three test statistic under different structural break magnitude when $c = 0$ and $\tau = 0.5$

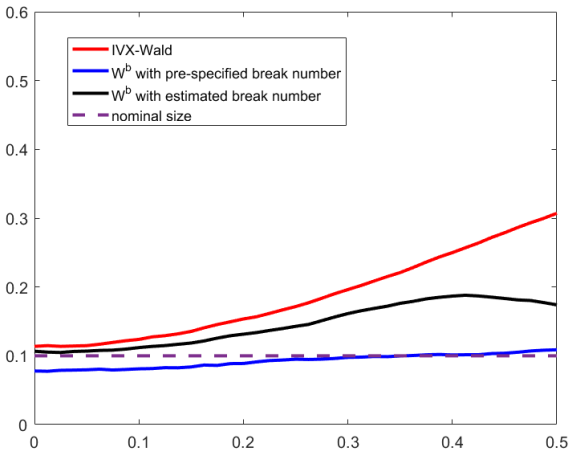


Figure 17: Rejection rate of three test statistic under different structural break magnitude when $c = 10$ and $\tau = 0.2$

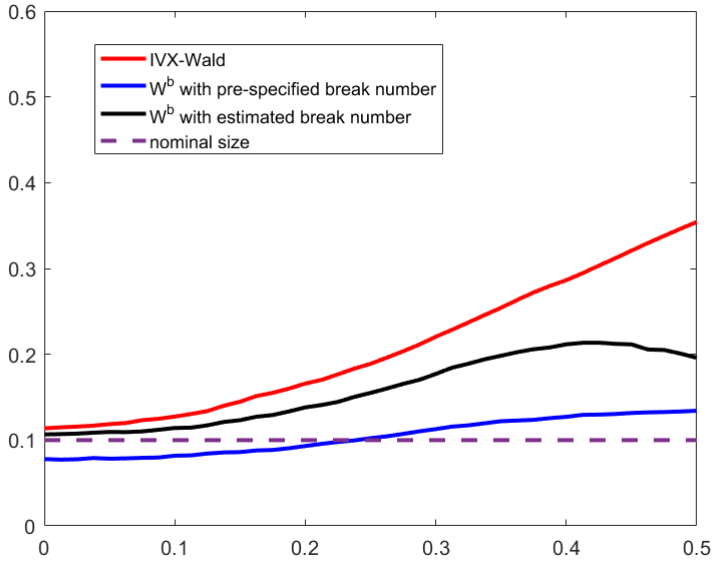


Figure 18: Rejection rate of three statistic under different structural break magnitude when $c = 10$ and $\tau = 0.8$

absolute return as a proxy for volatility is the basis of much-applied work in the literature. It has been used primarily in econometrics and econophysics research, e.g., Cizeau et al. (1997) and, in recent years, has been shown to be a better measurement of volatility, see Forsberg and Ghysels (2007).

It is well established that volatility of equity return is much more predictable than the return per se, and thus we expect that there exist some variables showing significant predictive power. In a recent contribution to predicting volatility of stock return, Paye (2012) considers dozens of macroeconomic variables. In the current work, we consider, for simplicity, the potential predictors mentioned in the last chapter. Recall that these include: Tbill rate (tbl), long-term yield (lty), term spread (tms), default yield spread (dfy), dividend-price ratio (d/p), dividend yield (d/y), earnings-price ratio (e/p), dividend payout ratio (d/e), book-to-market value ratio (b/m), net equity expansion (ntis), inflation rate (inf) and consumption-wealth ratio (cay).

Our sample period is from 1927 to 2012, and both the quarterly and monthly horizons are examined. We plot the time series of monthly absolute returns in Figure 18. Using the BP procedure, one structural break of absolute return is identified in the mid-1940s.

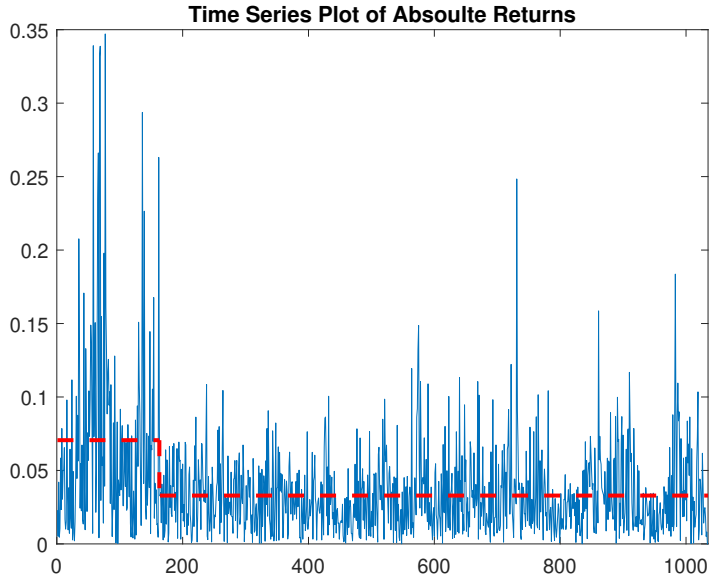


Figure 19: Absolute return of S&P 500 Index with one Level Shift

Note that this is consistent with the finding in Kim et al. (2005). Using Bayesian marginal likelihood analysis, those authors also find a permanent reduction in the general level of stock market volatility in the 1940s. The break date and the average level before and after the break are shown in Figure 18 by the dashed red line.

We first examine the predictability of these potential predictors using the IVX-Wald test. The results for quarterly and monthly data are reported in the first and third columns of Table 12, respectively. In the monthly case, all variables we consider except for earnings-price ratio are highly significant predictors. As for the quarterly horizon, all but long-term yield and earning-price ratio exhibit great in-sample predictive power. Given the existence of a level shift in absolute return, it is natural to ask whether these results are consequences of spurious regression effect. To this end, we carry out tests based on sample-splitting at the estimated break date. The results are reported in the second and fourth columns of Table 8, which is for quarterly and monthly data, respectively. It can be seen that, for several variables, the remarkable predictability disappears after the break is taken into consideration. For instance, the T-bill rate is much less predictive for both monthly and quarterly horizons than before. All these examples indicate that, when the predictability of stock return volatility is investigated, it is necessary to treat the level shift

Table 12: Predictability of Absolute Return of Stock Index

predictors	quarterly		monthly	
	IVX-Wald	W_b	IVX-Wald	W_b
Dividend payout ratio	42.82***	11.90***	99.10***	27.55***
Long-term yield	2.61	1.77	6.10***	2.30
Dividend yield	14.63***	12.90***	35.40***	37.80***
Dividend-price ratio	25.64***	31.12***	42.13***	36.03***
T-bill rate	5.88**	0.67	13.74***	5.07*
Earnings-price ratio	0.0502	12.28***	1.76	4.58
Book-to-market value ratio	37.43***	30.68***	62.85***	37.32***
Default yield spread	137.67***	40.76***	252.61***	63.51***
Net equity expansion	8.77***	10.06***	16.55***	11.33
Term spread	5.59**	2.65	13.20***	8.25**
Inflation rate	28.29***	11.23***	11.01***	4.39
Consumption-wealth ratio	11.351***	16.457**	-	-

explicitly before any test is carried out.

4.5.2 Predicting Growth Rate of Housing Price Index

The second empirical application we consider is the predictability of the growth rate of the quarterly housing price index (HPI), which was recently studied in Yang et al. (2020). In that paper, the authors use the IVX-based test as well, but they emphasize another potential failure of this approach. Specifically, it is argued that when $u_y t$ has serial correlation, the IVX-Wald test will suffer severe size distortion. To solve this problem, a modification based on Cochrane–Orcutt procedure is proposed and shown to work well. After conducting adjustment, they find that most macroeconomic variables they consider have no predictive power, which dramatically contradicts the conclusion led to by a direct application of the IVX-Wald test.

We re-examine here this exercise with a focus on the level shifts. The predicted variable, HPI, is collected from the Federal Housing Finance Agency (FHFA), and the time span is from the first quarter of 1975 to the second quarter of 2018. Using this time series, we can easily obtain the growth rate of HPI. Both the level and growth rate of HPI are plotted in Figure 19. BP procedure suggests that the growth rate of HPI experiences four level shifts. The timings of these breaks are approximately 1980, 2000, 2005, and 2012. All the break dates and the average level of HPI growth rate in each sub-period are also indicated in Figure 19 by the red dashed line. In the presence of these shifts, spurious

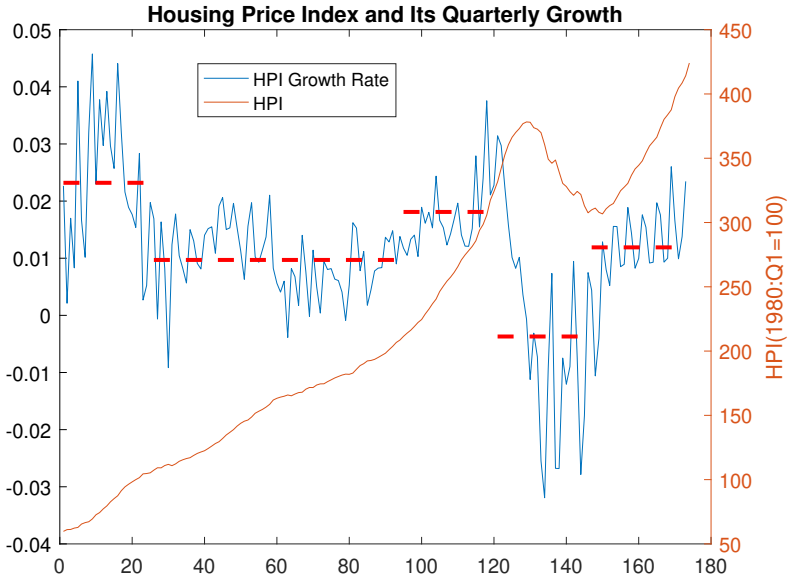


Figure 20: Housing Price Index (HPI) with four level shifts

predictability may well be detected by the IVX-Wald test.

Following Yang et al. (2020), we also consider following ten macroeconomic variables, which are available from Federal Reserve Economic Data (FRED):

- 1. CPI: Consumer price index with all items less shelter for all urban consumers (Index 1982-1984 = 100).
- 2. DEF: Implicit price deflator of GDP (Index 2012=100).
- 3. GDP: Percent change of GDP from last period.
- 4. INC: Percent change of real disposable personal income from the same quarter in last year.
- 5. IND: Industrial production index (Index 2012 = 100).
- 6. INT: Effective federal funds rate.
- 7. INV: The shares of the residential fixed investment in GDP.
- 8. MOG: 30-year mortgage rate.

Table 13: Predictability of Housing Price Index (univariate regressions)

predictors	1975:Q1-2018:Q2		2000:Q1-2018:Q2	
	IVX-Wald	W_b	IVX-Wald	W_b
CPI	35.542***	20.206***	11.3039***	1.7742
DEF	24.115***	31.1201***	3.4594*	3.7734
GDP	21.841***	9.024	6.8410***	7.1247*
INC	4.924**	11.314**	5.9425**	0.2463
IND	5.229**	17.158***	1.3217	1.1243
INT	6.915***	12.804**	0.8632	5.6860
INV	94.097***	61.795***	38.9980***	7.9640**
MOG	4.528**	11.181**	0.8115	1.9936
RES	8.671***	6.032***	3.1489*	4.6965
UNE	22.718***	13.030**	17.1028***	1.9788

- 9. RES: Total reserve balances maintained with the Federal Reserve banks.
- 10. UNE: Civilian unemployment rate.

We first consider univariate predictive regressions. The results are reported in Table 13. As in Yang et al. (2020), in addition to the full sample, we also consider the sub-sample starting from the first quarter of 2000. The reason for the choice of this sub-period is that HPI is highly volatile during this time, and it consists of both decline and surge of the housing market. To make a comparison, we report both the results with and without adjustment for level shifts. As shown by the first two columns of Table 13, taking breaks into account will not make a big difference for the full-sample test, with only the GDP becoming an insignificant predictor. More interestingly, however, when we turn to the sub-period in the 21st century, the conclusion changes remarkably and is in line with the one obtained in Yang et al. (2020). Indeed, most macroeconomic variables that are deemed as significantly predictive before adjusting for level shifts show little or no predictive power after we use the sample-splitting procedure. Similar to Yang et al. (2020), we also find that INV, the share of the residential fixed investment in GDP, remains significant after taking breaks into account and thereby is a pretty robust predictor of the growth rate of HPI.

Now let us move on to the multivariate regressions. Following Yang et al. (2020), we consider five combinations

- 1. INV+UNE;

Table 14: Predictability of Housing Price Index (multivariate regressions)

predictors	1975:Q1-2018:Q2		2000:Q1-2018:Q2	
	IVX-Wald	W_b	IVX-Wald	W_b
Combination 1	94.43***	80.51***	38.51***	10.6
Combination 2	106.60***	115.12***	46.92***	30.38**
Combination 3	22.89***	92.48***	14.14***	22.38**
Combination 4	22.99***	35.49***	17.11***	7.70
Combination 5	216.60***	244.00***	232.18***	111.86***

- 2. INV+UNE+IND+GDP+INC;
- 3. CPI+DEF+INT+RES;
- 4. CPI+INT+MOG;
- 5. A "kitchen sink" which includes all ten variables.

See Section 4.4 of that paper for a discussion of these choices. The results for multivariate regressions are reported in Table 14. In general, for the full-sample case, adjusting for level shifts does not make any difference from the IVX-Wald test. This is, however, not the case for sub-sample starting from the year 2000. For this specific period, taking breaks into account will make combination 2 and 3 less significant, while combination 1 and 4 completely lose predictive power. Again, it seems that the predictability in this period is partly induced by the spurious regression effect.

4.6 Conclusions

In this chapter, we study the potential consequences of predicted variable's level shifts on the detection of predictability. We show by both theoretical derivation and simulation that this type of structural breaks will result in the spurious regression phenomenon and thereby seriously distort the test based on OLS and IVX estimators. We discussed two cases. In the first case, the break size is large and the t-statistic tends to infinity in the limit, which is similar to conventional spurious regression. The second case is concerned with moderate shifts, which leaves the t-statistic stochastically bounded in large sample. However, the limiting distribution in that case is not the same as the one used to compute the critical values, and thus the size distortion will still be present.

To deal with the large breaks, We propose a testing procedure based on sample-splitting and show that it enjoys reasonable size control when breaks can be precisely detected by popular methods like Bai and Perron (1998). Two empirical applications are provided to illustrate the empirical relevance of our results. As for moderate shifts, we carefully study the single break case and argue that size control can be partially achieved by pre-specifying the number of breaks. Simulation suggests that this strategy works well in various cases. We leave a thorough understanding of moderate break case as future work.

5 Multivariate Stochastic Volatility Model with Flexible Dynamic Correlations

5.1 Introduction

Characterizing the dynamic behavior of volatility of individual asset returns is of great importance to both portfolio allocation and risk management. Starting from the seminal paper by Engel (1982), a wide range of univariate models have been considered in the literature, most of which can be categorized as either GARCH-based models or stochastic volatility (SV) models. Another notable development during the past few decades is the focus on multivariate financial data analysis. It is increasingly recognized that analyzing the asset return individually is not adequate, and the dependence structure among different assets must be taken into account. To this end, a plethora of multivariate extensions of GARCH and SV have emerged and been applied in the empirical researches as well as industry practice. See Bauwens et al. (2006) for an extensive review of multivariate GARCH (MGARCH) and Asai et al. (2006) for multivariate SV (MSV).

The first MSV model is proposed in Harvey et al. (1994) and is an analogy of the constant conditional correlation (CCC) model in the MGARCH literature. In this basic setup, the volatility of each individual asset is assumed to follow a univariate SV process, while the correlation matrix among all assets is constant over time. This is obviously a rather restrictive and perhaps unrealistic assumption, as it implies that the correlation structure keeps unchanged over the whole time span. Great efforts have been dedicated to the relaxation of the constant correlation assumption ever since. For instance, Yu and Meyer (2006) proposed a model that mirrors the dynamic conditional correlation (DCC) model of Engle (2002) in MGARCH. Another parametrization that is also based on DCC can be found in Asai and McAleer (2009).

All the models mentioned above are built upon a variance-correlation decomposition of the covariance matrix. It is also possible to characterize the dynamic conditional correlation by modeling the covariance matrix per se. Similar to the case in the MGARCH literature, the major challenge here is to make sure that the model can produce a posi-

tive definite covariance matrix. One possibility is to model the matrix logarithm of the original covariance matrix, which must exist and is free of algebraic constraint by design. This idea, as an extension of exponential GARCH in Kawakatsu (2006), is explored in Ishihara et al. (2016). A significant drawback of this approach is that the volatilities and correlations are interwoven with each other, and thus the estimated model is very hard to interpret. Another existing approach, considered in Lopes et al. (2002), takes advantage of the well-known Cholesky decomposition of a symmetric positive definite matrix. The useful property of this method is that there is a clear and meaningful connection between elements of the original covariance matrix and that of the component matrices, while the shortcoming is that the driving forces underlying volatilities and correlations are not completely separated. Last but not least, the Wishart autoregressive multivariate process provides a flexible modeling tool for MSV as well; see Philipov and Glickman (2004) and Gourieroux et al. (2016) for details.

In this paper, we propose to apply a recently developed new parameterization of the correlation matrix to develop a new MSV model. Such a parameterization, originally proposed in Archakov and Hansen (2018), has been successfully implemented in the MGARCH context, see Archakov et al. (2019). They show that this approach can be deemed as a generalization of the well-known Fisher's z-transformation to the higher dimensional case. We seek to extend their work to the MSV setup. Under this new modeling design, the underlying latent states that determine the correlation among assets are allowed to have an unrestricted domain, and no algebraic constraint of any kind is necessary. This is due to the fact that this new parameterization by construction automatically guarantees that the estimated correlation matrix is indeed valid. Meanwhile, the underlying shocks to volatilities and correlations are fully separated in our model. This is an appealing feature, as in practice, these two quantities may be determined by completely distinct factors. Last but not least, our model is invariant to the reordering of assets, and thus no ex-ante ordering is necessary. All these features indicate that our model is very flexible in terms of capturing dynamic patterns of volatilities and correlations, imposing a minimum level of ex-ante restrictions.

As is the case for most MSV models, estimation of our model is challenging due to a

large number of parameters as well as a high dimension of latent variables. To accommodate the former, we follow the existing work and apply Bayesian Markov Chain Monte Carlo (MCMC) method, as the maximum-likelihood-based approach will be unstable or even fail to reach a result under this circumstance. Departing from much of the literature, where carefully designed single-move sampler or multi-move Gibbs sampler are used to tackle the latent variables issue, we opt to work with the recently proposed particle-filter based MCMC algorithm. Ever since the seminal paper by Andrieu et al. (2010), the research on the theoretical foundation of PMCMC and its empirical applications to many different fields has exploded quickly. Though theoretically applicable under a very general setup, the practical performance of this approach for a particular model depends on many factors and requires careful examination. To strike a balance between satisfactory estimation accuracy and acceptable computational cost, we choose a method called Particle Gibbs Ancestor Sampling (PGAS). It is a modified version of Particle Gibbs Sampling (PG) considered in the Andrieu et al. (2010), which dramatically improves the mixing property under a small number of particles. Extensive simulation results are presented to justify the choice and also provide some guidance for empirical application.

The structure of this chapter is as follows. Section 5.2 briefly reviews the parameterization and estimation method used in the current literature. Section 5.3 introduces the new parametrization proposed in Archakov and Hansen (2018) and discuss some of its properties. The detailed specification of our new multivariate stochastic volatility model is presented in Section 5.4. Section 5.5 contains how we make inferences for the proposed model using PGAS and present some simulation results. Then, in Section 5, the new model is implemented to characterize the dynamics of two sets of assets. The first one is the return of foreign exchange rates, and the second is equity returns. Section 5.6 concludes the paper and points out some potential improvements to our model.

5.2 A Selective Literature Review

We review some literature on modeling multivariate stochastic volatility in this section. Yu and Meyer (2006) and Asai et al. (2006) summarized this area of research up to that time, discussing both estimation techniques and model comparison. A similar ac-

count can be found in Chib et al. (2009). For early literature, we refer the reader to those review papers. Here, we mainly consider the models proposed in the most recent decade, paying particular attention to how they achieve the validity of resulting covariance matrices and the choice of inference method.[16]

5.2.1 Model Setup

The basic structure of the MSV model is

$$r_t|C_t \sim N(0, C_t), \tag{5.1}$$

where r_t is a vector of asset returns, and we aim at characterizing the dynamics of its variance-covariance matrix C_t. Clearly, C_t must be symmetric and positive definite for all t. Different models usually rely on different designs to guarantee this property. Broadly speaking, we can categorize MSV models into two groups. For the first group, one directly builds a model for C_t per se, while, for the latter, we first carry out a volatility-correlation decomposition and then model each part separately.

Within the first group of models, three methods have been considered. The first one is based on the matrix exponential. Ishihara et al. (2016) assume that

$$C_t = \exp(H_t/2)$$

and propose to model $vech(H_t)$ as a vector autoregression process. Due to the definition of the matrix exponential, C_t is guaranteed to be a valid variance-covariance matrix. The major drawback of this model is that the relationship between those modeled latent factors and original volatilies/correlations is highly nonlinear and thus very hard to interpret.

The second one utilizes the well-known Cholesky decomposition. For instance, Lopes et al. (2012) propose to decompose C_t as

$$C_t = A_t H_t A_t',$$

[16] Note that we do not consider the models based on factor structure in this review, as our new model is based on direct modeling of covariance matrix

where H_t is a diagonal matrix and A_t is a lower triangular matrix, and then model all the nonzero elements in A_t and H_t as the autoregression process. Similarly, Shirota et al. (2017) also use this decomposition to set up their MSV model. Though under this parameterization we can easily link the latent variables to volatilities and correlations, other problems arise with the Cholesky decomposition. For instance, the resulting matrix depends on the ordering of assets, which is undesirable. Meanwhile, the underlying dynamics of volatilities and correlations are not fully separated.

The third set of models take advantage of the Wishart distribution, whose support includes only positive definite matrices. This is considered in Gourieroux et al. (2009), where a Wishart autoregression process is used. Specifically, they assume that

$$\begin{aligned} C_t &= \sum_{i=1}^{m} x_{it} x_{it}', \\ x_{it} &= A x_{i,t-1} + \epsilon_{it} \text{ and } \epsilon_{it} \sim N(0, \Sigma), \end{aligned}$$

where (v, d, A) are unknown parameters. Alternatively, one can also model C_t uses the inverse Wishart as in Philipov and Glickman (2006). In this case, we have

$$C_t^{-1} | v, C_{t-1}^{-1} \sim Wishart\left(v, \frac{1}{v}(A^{1/2})(C_{t-1}^{-1})^d (A^{1/2})'\right),$$

where (m, A, Σ) are unknown parameters. A similar model along this line is presented in Jin and Maheu (2013).

The models in the second group treat volatility of each asset and the correlation matrix separately, based on the following decomposition

$$C_t = V_t^{1/2} R_t V_t^{1/2},$$

where V_t is a diagonal matrix collecting all the variances and R_t correlation matrix. For our purpose, the major difference of designs among this group lies in how they parameterize the R_t. Note that the critical issue in this setup is again how to make sure of the validity of the model-generated correlation matrix. The first and simplest model in this

fashion is the constant correlation MSV in Harvey et al. (1994), where

$$R_t = R, \text{ for all } t.$$

A similar assumption is made in Chan et al. (2006) and Asai and McAleer (2006). It can be easily seen that in these models, the dynamic movement of correlations is not allowed. Although the inference under this assumption is quite simple, it is clearly too restrictive for modeling most financial data.

To incorporate a time-evolving correlation structure, Asai and McAleer (2009) considered two models motivated by the Dynamic Conditional Correlation (DCC) model of Engle (2002) in the MGARCH literature. The idea is to write the correlation matrix in the following form

$$R_t = \tilde{Q}_t^{-1} Q_t \tilde{Q}_t^{-1},$$

where $\tilde{Q}_t = (diag(vecd(Q_t))^{1/2}$. By this construction, the diagonal elements of R_t must be 1 and symmetric positive definiteness can be achieved if Q_t is symmetric positive definite. Two Wishart-driven models for Q_t they propose are

$$Q_{t+1} = (1-\phi)\bar{Q} + \phi Q_t + \Xi_t, \text{ where } \Xi_t \sim Wishart(k, \Lambda)$$

and

$$Q_{t+1}^{-1} | k, Q_t^{-1} \sim Wishart\left(k, \frac{1}{k} Q_t^{-\phi/2} \Lambda Q_t^{-\phi/2}\right),$$

where unknown parameters are (k, ϕ, Λ). Between these two parameterizations, the authors argued that the second one is preferred.

Inspired by the Dynamic Equicorrelation (DECO) model in the MGARCH literature proposed in Engel and Kelly (2012), Kurose and Omori (2016) proposed to model R_t in the following way:

$$R_t = (1-\rho_t)I + \rho_t J,$$

where I is an identity matrix, and J is a square matrix with all elements equal to 1. To ensure that ρ_t is within $(-1, 1)$, they model the Fisher transformation of ρ_t as an autore-

gressive process. Note that in this model, R_t is positive definite only if ρ_t is large than some lower bound depending on the number of assets. When more assets are considered, this lower bound will approach 0. Kurose and Omori (2018) further extended this model to the multiple-block case and included some other features.

More recently, Yamauchi and Omori (2020) propose to model pairwise correlations by the Fisher transformation. Their parameterization is

$$R_t = \{\rho_{ij,t}\}, \text{ where } \rho_{ij,t} = \frac{\exp(g_{ij,t}) - 1}{\exp(g_{ij,t}) + 1},$$

and $g_{ij,t}$ is assumed to follow a random walk. Since this element-wise operation does not guarantee the validity of R_t as a correlation matrix, they further derive algebraic bounds for $\rho_{ij,t}$ that guarantee the positive definiteness of R_t. Note that the bounds for one particular $\rho_{ij,t}$ are conditional on all other elements in R_t. Therefore, it is well suited for the single-move Gibbs sampling technique, but hard to be extended to other estimation method.

5.2.2 Estimation Method

Unlike observation-driven models like univariate and multivariate GARCH, which can be estimated straightforwardly by the maximum likelihood method, stochastic volatility models are particularly challenging in terms of estimation and inference. This is due to the high-dimensional latent variables involved in the models, as well as a large number of parameters. To deal with these complications, historically, dozens of methods were proposed and implemented to make inferences for SV models. For a detailed survey on this topic, see Broto and Ruiz (2004).

The earliest and perhaps most straightforward method for estimating SV models is the method of moments, which is considered in Taylor (1986). Other moment-based procedures include the generalized method of moments as in Melino and Turnbull (1990) and Andersen and Sorensen (1996), and the simulated method of moments proposed in Duffie and Singleton (1993). Though very easy to implement, these methods have less satisfactory finite sample properties. Moreover, they do not provide estimates of underlying

latent processes. Later, a quasi-maximum-likelihood estimator (QMLE) is proposed in Harvey et al. (1994), which relies on a log-linearization of the original model. This transformation results in a linear state space form with log chi-squared errors, and the Kalman filter is subsequently used to compute the likelihood. However, QMLE does not rely on the exact likelihood of data, and the approximation of log chi-squared errors by Gaussian density could be somewhat inappropriate. A better approximation to the exact likelihood can be achieved either by using a mixture of normals or through numerical simulation. The former is proposed in Kim et al. (1998). Monte Carlo Likelihood approach proposed in Sandmann and Koopman (1998) provides an example of the latter.

A majority of recent papers on stochastic volatility work under the Bayesian framework and base their inference on the Markov Chain Monte Carlo technique. This is particularly the case for multivariate models, as they usually involve a large number of parameters, and optimizing over a high-dimensional parameter space is well known to be complicated. For a Bayesian method, this problem becomes computing high-dimensional integral, which can be efficiently done by MCMC sampling. The critical step of this procedure is to draw a sample of latent processes given a particular set of static parameters. The first method proposed in the literature is the single-move sampler in Jacquier et al. (1994), where each latent variable is drawn one at a time, given all the other ones. This approach is well known to be inefficient, as it will generate seriously autocorrelated samples from the Markov chain, suggesting a vast amount of random draws is required to achieve a satisfactory accuracy of estimation. This is particularly true for empirically relevant cases of strongly persistent volatility. In light of this inefficiency, Kim et al. (1998) provide a highly efficient alternative MCMC method. Their method also starts with a log-linearization of the model, after which they introduce extra auxiliary variables to reach a linear state-space form. Simulation smoother of de Jong and Shephard (1994) is then applied to draw the whole latent process simultaneously. Though very popular in the univariate SV literature, this method is not easy to be extended to MSV models with complicated correlation structures. Another significant improvement of the single-move sampler is proposed in Shephard and Pitt (1997). These authors suggest blocking to improve the speed of convergence for simulators of non-Gaussian state-space models.

Within each block, a second-order Gaussian approximation is used to obtain a good proposal density. Since sampling the whole sequence of latent variables in one block leads to a high rejection rate, they also discuss how to choose a balanced block size. This method is later modified by, e.g., Omori and Watanabe (2004, 2007), and attracts a lot of applications.

5.3 Generalized Fisher's z-Transformation and GFT-MSV Model

In this section, we introduce the new parameterization of the correlation matrix proposed in Archakov and Hansen (2018), which can be considered as a high-dimensional generalization of widely used Fisher's z-transformation in bivariate correlation modeling. Based on this new tool, a flexible multivariate stochastic volatility model is proposed.

5.3.1 Parametrization of Correlation Matrix

When the correlation coefficient between two random variables, say ρ, is to be modeled, an essential constraint is that its value must be within the interval $(-1, 1)$. To avoid complexity brought by this constraint in inference, one can instead model the Fisher's z-transformation of ρ, defined as $F(\rho) = \frac{1}{2}\log\frac{1+\rho}{1-\rho}$. We know that $F(\cdot)$ is a one-to-one mapping and for any $F(\rho) \in \mathscr{R}$, there exists a unique corresponding $\rho \in (-1, 1)$. Therefore, one can impose any structure on $F(\rho)$ and transform it back to obtain ρ without worrying about the validity of the resulting correlation coefficient. This idea was first introduced to multivariate stochastic volatility literature in Yu and Meyer (2006). Unfortunately, it is acknowledged by the authors that this approach "is not easy to be generalized into higher dimension situations". In particular, a pair-wise transformation applied to each entry in a high-dimensional correlation matrix, though seems to be natural, is not a valid choice as it fails to guarantee the positive definiteness of the resulting correlation matrix.

Fisher's z-transformation has many known nice properties, and it is, therefore, desirable to obtain a valid high-dimensional version of this transformation. This is achieved in a recent paper by Archakov and Hansen (2018). Their proposal is to apply the matrix logarithm to the original correlation matrix and then model all the lower off-diagonal elements in that new matrix. To fix the idea, suppose we have a p-dimensional corre-

lation matrix C and let $G = \log C = \sum_{k=1}^{\infty} \frac{(-1)^k (C-I)^k}{k}$. Note that the convergence of the infinite summation and thus the existence of G is ensured by the fact that C, as a correlation matrix, is symmetric and positive definite. Further, denote $q = vecl(G)$ as the $\frac{p(p-1)}{2}$-dimensional vector containing all lower off-diagonal entries of G. In summary, the new parametrization of the original correlation matrix is now defined by the mapping $q = vecl(\log C)$. The key theoretical contribution of Archakov and Hansen (2018) is that, this mapping from any vector in $\mathscr{R}^{p \times p}$ to valid correlation matrix in $\mathscr{R}^{\frac{p(p-1)}{2} \times 1}$, is shown to be one-to-one. That is to say, given any $\frac{p(p-1)}{2}$-dimensional vector q, there exists a unique valid p-dimensional correlation matrix C. Although the inverse mapping from q to C does not admit a closed-form analytical solution, C can easily be reconstructed from q using an iterative algorithm.[17]

It is straightforward to show that when $p = 2$, the above-defined transformation will reduce to Fisher's z-transformation. As a generalization, the new parametrization inherits a few advantages from the z-transformation and enjoy some additional desirable properties. First and foremost, it is very flexible in the sense that when modeling q, we do not need to impose any algebraic constraint. This suggests that we can consider any reasonable dynamics for q without worrying about the positive-definiteness of the resulting correlation matrix, which is attractive in the stochastic volatility model. Second, compared with original elements in correlation matrix C, the sampling distribution of elements in transformed vector q usually appears to be closer to Gaussian distribution. Hence it is reasonable to model q through a Gaussian autoregressive process. Third, this transformation is invariant to the reordering of variables, in contrast to those approaches based on Cholesky decomposition. Fourth, although elements of q depend on C on a nonlinear way, there are some interesting properties that do carry over to matrix $G = \log(C)$, notably equicorrelation and block-equicorrelation structure, see Archakov et al. (2019).

Note: For the sake of notational simplicity, in the rest of the paper, we will refer to the mapping $vecl(\log(\cdot))$ by $F(\cdot)$ and the corresponding inverse mapping by $F^{-1}(\cdot)$.

[17] This algorithm has been implemented in Matlab, Python, Ox, and the corresponding computer code can be found in the online appendix of Archakov and Hansen (2018). We thank the authors for making it public.

5.3.2 Multivariate Stochastic Volatility Model with Generalized Fisher Transformation

We specify our new MSV model in this section. Let $r_t = (r_{1t}, ..., r_{pt})'$ denote the $p \times 1$ vector of asset returns and $h_t = (h_{1t}, ..., h_{pt})'$ the corresponding vector of latent log-volatilities at time t. We denote the vector of latent factors at time t that underlie the movement of correlations by $q_t = (q_{1t}, ..., q_{dt})'$, where $d = p \times (p-1)/2$ by construction . This vector q_t is connected to correlation matrix R_t through the transformation detailed in Section 5.3. Then, the basic model, which we refer to as GFT-MSV, is given by

$$r_t = V_t^{1/2}\epsilon_t, \text{ where } \epsilon_t \sim N(0, R_t), t = 1, ...T, \tag{5.2}$$

$$V_t = \exp(H_t), \text{ and } h_t = diag(H_t), t = 1, ...T, \tag{5.3}$$

$$q_t = F(R_t), t = 1, ...T, \tag{5.4}$$

$$h_{t+1} = \mu_h + \Phi_h \cdot (h_t - \mu_h) + \eta_{ht}, \text{ where } \eta_{ht} \sim N(0, \Sigma_h), t = 1, ...T-1, \tag{5.5}$$

$$q_{t+1} = \mu_q + \Phi_q \cdot (q_t - \mu_q) + \eta_{qt}, \text{ where } \eta_{qt} \sim N(0, \Sigma_q), t = 1, ...T-1, \tag{5.6}$$

$$h_1 \sim N(\mu_h, (I_p - \Phi_h^2)^{-1}\Sigma_h), \text{ and } q_1 \sim N(\mu_q, (I_d - \Phi_q^2)^{-1}\Sigma_q), \tag{5.7}$$

where $\epsilon_t = (\epsilon_{1t}, ...\epsilon_{pt})'$, $\eta_{ht} = (\eta_{h1t}, ...\eta_{hpt})'$, $\eta_{qt} = (\eta_{q1t}, ...\eta_{qdt})'$, $\mu_h = (\mu_{h1}, ...\mu_{hp})'$ and $\mu_q = (\mu_{q1}, ...\mu_{qp})'$. We assume that Φ_h and Φ_q are both diagonal matrix, whose diagonal elements are $\phi_h = (\phi_{h1}, ..., \phi_{hp})$ and $\phi_q = (\phi_{q1}, ..., \phi_{qd})$ respectively. Σ_h and Σ_q are also supposed to be diagonal matrices, with diagonal entries being $\sigma_h^2 = (\sigma_{h1}^2, ...\sigma_{hp}^2)$ and $\sigma_q^2 = (\sigma_{q1}^2, ..., \sigma_{qd}^2)$. Hence, all the latent variables are generated by completely independent autoregressive processes. As usual, h's represent the log-variance of each individual asset return. On the other hand, q's capture the dynamic correlation structure among all assets and are related to the observations through a nonlinear link transformation $F(\cdot)$.

Note that in the GFT-MSV model, persistence is allowed to vary across all the factors driving the dynamics of the correlation matrix. This is in sharp contrast to models based on Engle's DCC framework or the Wishart autoregression, where the persistence of all the correlations is usually governed by one scalar parameter. This flexibility could be

consequential in practice, as in most cases, such 'equi-persistence' can not be justified ex ante. To empirically examine the relevance of this relaxation, we also consider a restricted model in this paper, which is same as the general model except for the additional assumption that the three driving factors of dynamic correlations share the same persistence and variance, i.e., $\phi_{q1} = ... = \phi_{qd}$ and $\sigma^2_{q1} = ... = \sigma^2_{qd}$. As can be easily seen, this structure is similar to the equicorrelation model proposed by Engle and Kelly (2012), with a minor difference that we allow the variation in the mean of each correlation.

5.4 Particle Filter and Markov Chain Monte Carlo Estimation

Due to the difficulty of evaluating likelihood, the most recent literature on MSV models carries out inference under a Bayesian framework. The seminal paper by Jacquire et al. (1994) introduced the single-move Gibbs sampler, in which all the latent variables are sampled using full conditional distribution one by one. Kim et al. (1998) showed that this approach could be quite inefficient and produce highly autocorrelated MCMC samples. To improve efficiency, the multi-move sampler proposed in Sheppard and Pitt (1997) is employed by many papers. See also the discussion in Watanabe and Omori (2004) and Omori and Watanabe (2008). In this paper, we take advantage of a recently proposed technique known as PMCMC, which builds an efficient, high-dimensional MCMC kernel. Specifically, we use an improved Particle Gibbs sampler that enjoys the good mixing property even with a small number of particles. Details of our estimation procedure are discussed in this section.

5.4.1 Review of Particle Filter

Before moving to PMCMC, we first briefly introduce the concept of particle filter. Consider a general non-linear state space model given by

$$y_t|x_t = x \sim f_\theta(\cdot|x), \tag{5.8}$$

$$x_{t+1}|x_t = x \sim g_\theta(\cdot|x), \text{ and } x_1 \sim \mu_\theta(\cdot), \tag{5.9}$$

where y_t is the observable variable, x_t is latent variable and θ contains all the parameters of interest. In this subsection, we assume for the moment that θ is given and our target is to infer the latent variables $(x_1, ...x_t)$ using observations $(y_1, ...y_t)$. Specifically, we wish to obtain an estimator of the conditional distribution $p_\theta(x_1, ...x_t|y_1, ..., y_t)$. A closed-form analytical solution for this problem is usually unavailable, except for the very special cases like Gaussian linear model, where Kalman filter is applicable. Approximation must be relied upon in general models. Particle filter, also known as sequential Monte Carlo (SMC) in the literature, is exactly the tool to apply in such a case. The only requirements for the validity of particle filter are that (i). the measurement density $f_\theta(\cdot|x)$ can be numerically evaluated and (ii). one can simulate from the transition density $g_\theta(\cdot|x)$.

The methodology of particle filter combines importance sampling and Monte Carlo simulations to approximate the target distribution. The key idea is to represent the distribution by a set of random samples with corresponding weights and calculate the quantity of interest based on these samples and weights. To fix the idea, let $\{x_{1:t}^{(i)}, w_t^{(i)}\}_{i=1}^N$ be a random measure, where $\{x_{0:t}^{(i)}, i = 1, ..., N\}$ is a set of support points and $\{w_t^{(i)}, i = 1, ..., N\}$ are associated weights. Here, we use $x_{1:t} = \{x_j, j = 1, ...t\}$ to denote the set of all states up to time t. Each point is called a particle, and N is the number of particles used. The approximated distribution can then be written as

$$\hat{p}_\theta(dx_{1:t}|y_{1:t}) = \sum_{i=1}^{N} w_t^{(i)} \delta_{x_{1:t}^{(i)}}(dx_{1:t}), \tag{5.10}$$

where $y_{1:t}$ is similarly defined and $\delta(\cdot)$ is Dirac function. $\hat{p}_\theta$ is a discrete weighted approximation to the target distribution p_θ. Apparently, the accuracy of the approximation can be improved as an increasing number of particles are included. Doing so, however, will also dramatically raise the computational burden.

To obtain the weights, we resort to the importance sampling. That is to say, we sample N times from a candidate distribution, say $q_\theta(x_{1:t}|y_{1:t})$, and assign the weight

$$w_t^{(i)} \propto p_\theta(x_{1:t}^{(i)}|y_{1:t}) / q_\theta(x_{1:t}^{(i)}|y_{1:t})$$

to each sample drawn. In practice, it is hard, if not impossible, to pick up a proper importance density for joint distribution of $x_{1:t}$ conditional on the data when sample size becomes large. Hence, this approach usually proceeds in a sequential fashion. Specifically, importance density is chosen to admit the factorization such that

$$q_\theta(x_{1:t}|y_{1:t}) = q_\theta(x_t|x_{t-1}, y_t)q_\theta(x_{1:t-1}|y_{1:t-1}).$$

For any existing weighted sample $\{x^{(i)}_{1:t-1}, w^{(i)}_{1:t-1}\}$ that follows from $p_\theta(x_{1:t-1}|y_{1:t-1})$, we augment it with the new state $x^{(i)}_t$ randomly drawn from $q_\theta(x_t|x_{t-1}, y_t)$. The joint sample, $(x^{(i)}_{t-1}, x^{(i)}_t)$ is then a realization from the targeted joint importance density. The corresponding weight for i^{th} sample can easily be updated through

$$\tilde{w}^{(i)}_t \propto w^{(i)}_{t-1} \frac{f_\theta(y_t|x^{(i)}_t)g_\theta(x^{(i)}_t|x^{(i)}_{t-1})}{q_\theta(x^{(i)}_t|x^{(i)}_{t-1}, y_t)},$$

and normalized to be $w^{(i)}_t = \frac{1}{N}\sum_{i=1}^N \tilde{w}^{(i)}_t$. An unavoidable problem of this procedure, known as degeneracy, is that after a few iterations, only one particle will still have non-negligible weight, which means a large computational cost is spent on particles with almost no contribution. To alleviate this problem, usually a resampling step is necessary.

An important by-product of this filtering strategy is an approximation to the conditional marginal likelihood $p_\theta(y_{1:t}|y_{1:t-1})$, which has a simple formula

$$\hat{p}_\theta(y_{1:t}|y_{1:t-1}) = \frac{1}{N}\sum_{i=1}^N w^{(i)}_t.$$

The joint likelihood can then be easily obtained as

$$\hat{p}_\theta(y_{1:T}) = \prod_{t=2}^T \hat{p}_\theta(y_{1:t}|y_{1:t-1}) \cdot \hat{p}_\theta(y_1).$$

Do note that, unlike the case for Kalman filter, where the exact likelihood is available through analytic formula, the likelihood computed using particle filter is an estimation of the true likelihood, and thus subject to approximation error from random sampling.

Despite its general applicability, when implementing particle filter for a particular model, many subtle points must be considered. These include how to choose a proper importance density $q_\theta(x_t|x_{t-1}, y_t)$, how many particles to use and whether a resampling step should be added. For a thorough discussion, see Arulampalam et al. (2002) and Doucet and Johansen (2008).

5.5 Particle Gibbs Sampler and Ancestor Sampling

In practice, model parameters are unknown, and in a Bayesian framework with MCMC sampling, they must be drawn together with the latent variables. To sample from joint density $p(\theta, x_{1:T}|y_{1:T})$, typically we proceed by running a Gibbs sampler, which means drawing alternately from two conditional density, namely $p(\theta|x_{1:T}, y_{1:T})$ and $p(x_{1:T}|y_{1:T}, \theta)$. Usually, the former is easy to sample, either by imposing conjugate priors or through Metropolis-Hasting algorithm. The latter, on the other hand, can be handled by the particle filter approach introduced above. This is the basic idea of particle Gibbs (PG) sampler proposed in Andrieu et al. (2010).

One subtlety of this algorithm, as suggested in that seminal paper, is that to ensure the targeted joint distribution is indeed the invariant distribution of Markov chain, we have to modify the sequential Monte Carlo step a little bit. Specifically, one particle trajectory must be specified a priori, and it serves as a reference trajectory. Therefore, after a complete pass of the particle filter is run, a trajectory $x_{1:T}^{(i)}$ is randomly picked with probability proportional to the corresponding weight $w_T^{(i)}$. For the next MCMC iteration, when running particle filter, we only draw $N-1$ particles, and the N^{th} particle is fixed at the chosen one from the last iteration. The intuition of such modification is that this path can guide the simulated particles to move within a relevant region of state space. For a formal justification, see Theorem 5 of Andrieu et al. (2010).

A particle-filter-based MCMC procedure has a few desirable properties, which make it preferred compared with traditional methods. The most significant improvement is in terms of efficiency. Traditional single-move sampler is well known to be quite inefficient, as it will usually produce highly autocorrelated samples across MCMC iterations. Such strong dependency implies that one has to draw a vast number of samples to achieve

satisfactory accuracy. As suggested by the simulation studies below, the PG method we use significantly reduces the sample autocorrelation and thereby is a much more efficient sampler. Another alternative that could also alleviate the inefficiency in the single-move sampler is various types of multi-move approaches. Those methods, however, in most cases, require the derivation of a Gaussian second-order approximation, which could be tedious and difficult for multivariate non-linear models, like the one we consider in the current article. Designing a PG sampler, on the contrary, requires a minimal modification across different models, as long as they could be cast into a state-space form.

As for model comparison, one can also gain a lot from adopting sequential Monte Carlo. In the Bayesian paradigm, models are usually compared using Bayes factor based on the posterior model probability, which in turn, is calculated using the marginal likelihood of observed data $p(y_{1:T})$. To obtain this likelihood for a model with latent processes, theoretically, we should integrate out all those latent variables and parameters. In practice, however, if either single-move or multi-move sampler is used, a one-time particle filter will be conducted after plugging in the posterior mean of model parameters $\hat{\theta} = E(\theta|y_{1:T})$. The likelihood, which is actually a conditional one, will then be computed as $p(y_{1:T}|\hat{x}_{1:T}, \hat{\theta})$, based on the filtered states $\hat{x}_{1:T} = E(x_{1:T}|y_{1:T}, \hat{\theta})$. In the PG, on the other hand, this extra particle filter after MCMC sampling is unnecessary, for we obtain an estimated likelihood as a by-product during each SMC we run to sample the latent processes. Therefore, we can just calculate the marginal likelihood of data by averaging those likelihoods.

Though theoretically correct, PG has been shown to perform quite poorly when the underlying SMC sampler suffers from path degeneracy. As observed in Lindsten and Schön (2013) and Chopin and Singh (2014), the mixing of the Markov kernel induced by PG is rather slow under those circumstances. What makes things worse is that for the high-dimensional problem, like the one we consider in the current paper, path degeneracy is inevitable. To overcome this severe drawback, Lindsten et al. (2014) proposed a new method, which includes an additional step called ancestor sampling. While this is a small modification, the new particle Gibbs with ancestor sampling (PGAS) enjoys fast mixing of the Markov kernel even when only a seemingly small number of particles are used in

the underlying SMC. Informally, in the original PG, when degeneracy occurs, the particle system will collapse toward the chosen reference trajectory, while in the PGAS, it will degenerate toward something entirely different. As a consequence, the update rates of latent variables will be much higher with the additional ancestor sampling step, and thereby the mixing will be much faster. They also provide a proof for the fact that for a state-space model, PGAS is probabilistically equivalent to particle Gibbs sampler with a backward smoothing step under certain conditions.

For our purpose, a fast mixing under a small number of particles is highly desirable, as our likelihood function contains a component that has no closed-form solution and thus must be computed numerically. Although the cost for one-time computation is relatively low, it will soon become infeasible when a vast number of particles are included in the system. Indeed, for MCMC with S iterations, if the sample size is T and N particles are used, $F^{-1}(\cdot)$ must be evaluated $S \times T \times N$ times. As S and T are usually quite large in the empirical application, we can gain a lot in terms of computational efficiency by choosing the PGAS approach. Due to the same consideration, the particle Metropolis-Hasting approach is not chosen either as it requires an accurate calculation of marginal likelihood and thereby a large number of particles. In summary, we think that PGAS could be a suitable estimation tool given our model setup. Its applicability will be further examined below through extensive simulation studies.

5.5.1 Model Estimation

After introducing the critical component of our estimation strategy, we can now present our particle filter-based MCMC sampling algorithm. First of all, we need to specify the prior distribution of all the parameters $\theta \equiv (\mu_h, \mu_q, \phi_h, \phi_q, \sigma_h^2, \sigma_q^2)$. In this regard, our specification follows previous papers, especially the classical Kim et al. (1998). For the prior distribution of μ_h and μ_q, we assume independent multivariate normal distributions. The persistence parameters ϕ_h and ϕ_q are assumed to have a Beta priors. The prior distribution of σ_h and σ_q are chosen to be inverse gamma. In summary, we impose the following prior distribution:

- $\mu_{hi} \sim N(m_{\mu 0}, s_{\mu 0}^2)$ and $\mu_{qj} \sim N(m_{\mu 0}, s_{\mu 0}^2)$;

- $\frac{\phi_{hi}+1}{2} \sim Beta(a,b)$ and $\frac{\phi_{qj}+1}{2} \sim Beta(a,b)$;
- $\sigma_{hi}^2 \sim IG(\frac{n_{m0}}{2}, \frac{d_{m0}}{2})$ and $\sigma_{hi}^2 \sim IG(\frac{n_{m0}}{2}, \frac{d_{m0}}{2})$,

for $i = 1, ..., p$ and $j = 1, ..., d$ and $m_{\mu 0}, s_{\mu 0}^2, a, b, n_{m0}, d_{m0}$ are hyperparameters.

Let $r = (r1', ..., r_T')$, $h = (h_1', ...h_T')$ and $q = (q_1', ...q_T')$. To carry out the inference, we implement a Gibbs sampler with four blocks. In the following, we use $\theta_{/\alpha}$ to denote the parameters θ excluding α. Then, the algorithm proceeds as:

1. Initialize h, q and θ.
2. Draw $h, q|r, \theta$.
3. Draw $\mu_h, \mu_q|r, \theta_{/(\mu_h,\mu_q)}$.
4. Draw $\phi_h, \phi_q|r, \theta_{/(\phi_h,\phi_q)}$.
5. Draw $\sigma_h^2, \sigma_q^2|r, \theta_{/(\sigma_h^2,\sigma_q^2)}$.

Iterating over steps (2)-(4) consists of a complete sweep of MCMC sampler. The joint

posterior function of our model can be written as

$$
\begin{aligned}
p(\theta, h, q|r) &\propto p(r|\theta, h, q)p(\theta, h, q) \\
&= f(r|h, q)g_\theta(h)g_\theta(q)\pi(\theta) \\
&= f(r_1|h_1, q_1)g_\theta(h_1)g_\theta(q_1)\prod_{t=2}^{T}\left[f(r_t|h_t, q_t)g_\theta(h_t|h_{t-1})g_\theta(q_t|q_{t-1})\right]\pi(\theta) \\
&= \prod_{t=1}^{T}\left[\left(\sum_{i=1}^{p} h_{it}\right)|R_t|^{-1/2}\exp\left[-\frac{1}{2}r_t'(V_t^{1/2}R_tV_t^{1/2})^{-1}r_t\right]\right] \\
&\times \prod_{t=2}^{T}\prod_{i=1}^{p}\left[(\sigma_{hi}^2)^{-1/2}\exp\left(-\frac{1}{2\sigma_{hi}^2}\left(h_{it+1}-\mu_{hi}-\phi_{hi}(h_{it}-\mu_{hi})\right)^2\right)\right] \\
&\times \prod_{t=2}^{T}\prod_{j=1}^{d}\left[(\sigma_{qj}^2)^{-1/2}\exp\left(-\frac{1}{2\sigma_{qj}^2}\left(q_{jt+1}-\mu_{qj}-\phi_{qj}(q_{jt}-\mu_{qj})\right)^2\right)\right] \\
&\times \prod_{i=1}^{p}\left(\frac{\sigma_{hi}^2}{1-\phi_{hi}^2}\right)^{-1/2}\exp\left(-\frac{(h_{i1}-\mu_{h1})^2}{2\sigma_{hi}^2/(1-\phi_{hi}^2)}\right) \\
&\times \prod_{j=1}^{d}\left(\frac{\sigma_{qj}^2}{1-\phi_{qj}^2}\right)^{-1/2}\exp\left(-\frac{(q_{j1}-\mu_{q1})^2}{2\sigma_{qj}^2/(1-\phi_{qj}^2)}\right) \\
&\times \pi(\theta).
\end{aligned}
\tag{5.11}
$$

We apply the PGAS introduced in the last subsection to sample from the latent variables h and q given all the observations y and one particular set of parameter values. The detailed description of the algorithm is presented in Appendix D.1. On the other hand, from the joint posterior density, it is straightforward to sample from the marginal posterior of each parameter in θ given one particular realization of latent processes h and q. Specifically, we can do the following:

1. As the prior and likelihood are both normal, we can directly sample full conditional posterior of μ_{hi} and μ_{qi} from a normal distribution. For $i = 1, ..., p$ and $j = 1, ..., d$,

$$\mu_{hi}|y, h, q, \theta_{/\mu_{hi}} \sim N(\tilde{m}_{h\mu}, \tilde{s}_{h\mu}^2) \text{ and } \mu_{qj}|y, h, q, \theta_{/\mu_{qj}} \sim N(\tilde{m}_{q\mu}, \tilde{s}_{q\mu}^2),$$

   where

$$\tilde{m}_{h\mu} = \tilde{s}_{h\mu}^2\left\{\frac{1-\phi_{hi}^2}{\sigma_{hi}^2}h_{i1} + \frac{1-\phi_{hi}}{\sigma_{hi}^2}\sum_{t=1}^{T-1}(h_{it+1}-\phi_{hi}h_{it})\right\},$$

$$\tilde{m}_{q\mu} = \tilde{s}^2_{q\mu} \left\{ \frac{1-\phi^2_{qj}}{\sigma^2_{qj}} q_{j1} + \frac{1-\phi_{qj}}{\sigma^2_{qj}} \sum_{t=1}^{T-1} (q_{jt+1} - \phi_{qj} q_{jt}) \right\},$$

and

$$\tilde{s}^2_{h\mu} = \sigma^2_{hi} \left[(T-1)(1-\phi_{hi})^2 + (1-\phi_{hi})^2 \right]^{-1},$$

$$\tilde{s}^2_{q\mu} = \sigma^2_{qj} \left[(T-1)(1-\phi_{qj})^2 + (1-\phi_{qj})^2 \right]^{-1}.$$

2. To draw random samples from the full conditional posterior density of ϕ_{hi} and ϕ_{qi}, one can resort to Metropolis-Hasting sampler. Since

$$\begin{aligned} \log p(\phi_{hi}|y,h,q,\theta_{/\phi_{hi}}) &\propto \log p(h_i|\phi_{hi},\theta_{/\phi_{hi}}) + \log \pi(\phi_{hi}) \\ &= \log \pi(\phi_{hi}) - \frac{(h_{i1}-\mu_{hi})^2(1-\phi^2_{hi})}{2\sigma^2_{hi}} + \frac{1}{2}\log(1+\phi^2_{hi}) \\ &\quad - \frac{\sum_{t=1}^{T-1}\left[(h_{it+1}-\mu_{hi}) - \phi_{hi}(h_{it}-\mu_{hi})\right]^2}{2\sigma^2_{hi}}, \end{aligned}$$

we draw ϕ^*_{hi} from the proposal normal density $N(\hat{\phi}_{hi}, V_{\phi_{hi}})$, where

$$\hat{\phi}_{hi} = \left[\sum_{t=1}^{T-1} (h_{it+1}-\mu_{hi})(h_{it}-\mu_{hi}) \right] \Big/ \left[\sum_{t=1}^{T-1} (h_{it}-\mu_{hi})^2 \right],$$

is the ordinary least square estimator of ϕ_{hi} given h_i and

$$V_{\phi_{hi}} = \sigma^2_{hi} \left[\sum_{t=1}^{T-1} (h_{it}-\mu_{hi})^2 \right]^{-1}.$$

Then, the sample drawn is accepted with probability $\min\left[1, \exp\left\{g(\phi^*_{hi})/g(\phi^{(i-1)}_{hi})\right\}\right]$, where $\phi^{(i-1)}_{hi}$ is the sample from last MCMC iteration and

$$g(\phi_{hi}) = \log \pi(\phi_{hi}) - \frac{(h_{i1}-\mu_{hi})^2(1-\phi^2_{hi})}{2\sigma^2_{hi}} + \frac{1}{2}\log(1+\phi^2_{hi}).$$

ϕ_{qi} can be treated in a completely same fashion.

3. 3. Similar to the case for μ, due to the conjugacy, full conditional posterior of σ^2_{hi} can be immediately drawn from a inverse gamma distribution. For $i = 1, ..., p$ and

$j = 1, ..., d,$

$$\sigma_{hi}^2|y, h, q, \theta_{/\sigma_{hi}^2} \sim IG\left(\frac{\tilde{n}_m}{2}, \frac{\tilde{d}_{hm}}{2}\right) \text{ and } \sigma_{qj}^2|y, h, q, \theta_{/\sigma_{qj}^2} \sim IG\left(\frac{\tilde{n}_m}{2}, \frac{\tilde{d}_{qm}}{2}\right),$$

where $\tilde{n}_m = n_{m0} + T$ and

$$\tilde{d}_{hm} = d_{m0} + (h_{i1} - \mu_{hi})^2(1 - \phi_{hi}^2) + \sum_{t=1}^{T-1}\left[(h_{it+1} - \mu_{hi}) - \phi_{hi}(h_{it} - \mu_{hi})\right]^2,$$

$$\tilde{d}_{qm} = d_{m0} + (q_{j1} - \mu_{qj})^2(1 - \phi_{qj}^2) + \sum_{t=1}^{T-1}\left[(q_{jt+1} - \mu_{qj}) - \phi_{qj}(q_{jt} - \mu_{qj})\right]^2.$$

5.5.2 Simulation Studies

To investigate the performance of our chosen estimation procedure, we conduct some simulation exercises in this section. Our simulation is frequentist in essence as we generate data from the same data generating process for 100 times and use the posterior mean as a point estimator for all the parameters in the model. Since we know the true value of those parameters, we are thus able to examine the estimation bias as well as standard deviation[18].

To appreciate the sampling efficiency, following Kim et al. (1998), we also calculate the average inefficiency factor (IF), which is defined as the variance of the sample mean from MCMC sampling divided by that from a hypothetical sampler which draws independent samples. The variance of the MCMC sample mean is the square of the numerical standard error estimated by

$$NSE = 1 + \frac{2B_M}{B_M - 1}\sum_{i=1}^{B_M} K\left(\frac{i}{B_M}\right)\hat{\rho}(i),$$

where $\hat{\rho}(i)$ is estimated autocorrelation at lag i, B_M is the bandwidth and $K(\cdot)$ is the Parzen kernel. We choose the bandwidth B_M to be 1000. Apparently, a smaller IF in general indicates a faster mixing of Markov chain and thereby a better sampling efficiency.

[18]Note that here, standard deviation refers to the variation across replications, rather than the numerical standard error of MCMC sampler introduced below.

Table 15: Posterior statistics of μ with simulated data

			μ_{h1}	μ_{h2}	μ_{h3}	μ_{q1}	μ_{q2}	μ_{q3}
	N	True Value	0.3	0.3	0.3	0.7	0.7	0.7
T=500	50	Mean	0.311	0.317	0.305	0.632	0.639	0.631
		Std	0.094	0.083	0.080	0.063	0.054	0.059
		IF	11.5	9.8	9.5	17.3	16.0	14.3
	100	Mean	0.305	0.313	0.301	0.646	0.654	0.645
		Std	0.093	0.083	0.082	0.064	0.055	0.061
		IF	8.5	7.6	7.8	16.4	15.1	12.5
	200	Mean	0.304	0.311	0.299	0.657	0.665	0.656
		Std	0.092	0.083	0.082	0.064	0.055	0.061
		IF	6.4	5.8	5.5	12.2	11.6	12.0
T=1000	50	Mean	0.316	0.315	0.322	0.646	0.651	0.648
		Std	0.055	0.051	0.048	0.037	0.038	0.034
		IF	11.5	10.8	10.3	19.6	19.1	19.5
	100	Mean	0.310	0.310	0.317	0.661	0.666	0.664
		Std	0.053	0.052	0.049	0.036	0.038	0.034
		IF	9.0	7.6	6.7	17.2	16.9	14.7
	200	Mean	0.307	0.308	0.315	0.671	0.676	0.674
		Std	0.053	0.051	0.049	0.037	0.038	0.035
		IF	6.4	6.7	5.3	14.4	13.5	15.4
T=2000	50	Mean	0.316	0.316	0.316	0.651	0.657	0.655
		Std	0.036	0.033	0.034	0.028	0.025	0.024
		IF	10.1	9.3	8.6	21.2	25.1	20.8
	100	Mean	0.312	0.311	0.312	0.666	0.672	0.671
		Std	0.036	0.033	0.035	0.028	0.025	0.024
		IF	6.8	6.4	5.9	16.9	19.6	20.2
	200	Mean	0.308	0.308	0.309	0.675	0.682	0.681
		Std	0.037	0.032	0.035	0.028	0.026	0.024
		IF	6.3	4.7	5.4	15.7	15.4	15.5

1. T is the number of time series observations for each asset and N is the number of particles used in PGAS.
2. Mean is the average posterior mean across replications.
3. Std is the standard error of posterior mean across replications.
4. IF is the average inefficiency factor across replications calculated as suggested in Kim et al. (1998).

Table 16: Posterior statistics of ϕ with simulated data

			ϕ_{h1}	ϕ_{h2}	ϕ_{h3}	ϕ_{q1}	ϕ_{q2}	ϕ_{q3}
	N	True Value	0.9	0.9	0.9	0.8	0.8	0.8
T=500	50	Mean	0.894	0.885	0.887	0.784	0.791	0.790
		Std	0.033	0.043	0.046	0.051	0.058	0.06
		IF	56.4	54.6	55.6	56.9	59.5	58.7
	100	Mean	0.898	0.891	0.892	0.794	0.801	0.800
		Std	0.034	0.041	0.045	0.055	0.060	0.063
		IF	69.1	61.4	68.6	61.3	65.9	64.4
	200	Mean	0.902	0.893	0.893	0.798	0.807	0.803
		Std	0.033	0.043	0.047	0.059	0.061	0.062
		IF	61.3	69.3	71.1	60.7	65.2	72.6
T=1000	50	Mean	0.899	0.889	0.895	0.785	0.787	0.781
		Std	0.032	0.036	0.029	0.063	0.051	0.055
		IF	91.5	92.6	90.9	94.5	93.3	91.8
	100	Mean	0.900	0.891	0.898	0.793	0.795	0.791
		Std	0.032	0.032	0.028	0.060	0.051	0.057
		IF	103.86	87.1	95.1	101.0	92.5	91.7
	200	Mean	0.900	0.894	0.899	0.799	0.800	0.798
		Std	0.033	0.032	0.029	0.059	0.056	0.056
		IF	102.9	102.0	92.7	103.2	107.3	110.2
T=2000	50	Mean	0.900	0.897	0.899	0.793	0.785	0.787
		Std	0.022	0.022	0.023	0.047	0.040	0.044
		IF	112.4	105.8	107.0	128.2	125.0	120.9
	100	Mean	0.901	0.898	0.900	0.798	0.792	0.792
		Std	0.023	0.024	0.022	0.051	0.042	0.042
		IF	114.0	106.1	100.9	135.7	144.4	136.3
	200	Mean	0.900	0.898	0.900	0.801	0.795	0.796
		Std	0.023	0.023	0.024	0.049	0.041	0.044
		IF	103.6	99.2	102.0	134.9	131.5	141.8

1. T is the number of time series observations for each asset and N is the number of particles used in PGAS.
2. Mean is the average posterior mean across replications.
3. Std is the standard error of posterior mean across replications.
4. IF is the average inefficiency factor across replications calculated as suggested in Kim et al. (1998).

Table 17: Posterior statistics of σ^2 with simulated data

			σ^2_{h1}	σ^2_{h2}	σ^2_{h3}	σ^2_{q1}	σ^2_{q2}	σ^2_{q3}
	N	True Value	0.05	0.05	0.05	0.05	0.05	0.05
T=500	50	Mean	0.026	0.030	0.028	0.027	0.029	0.029
		Std	0.013	0.013	0.013	0.012	0.012	0.012
		IF	173.6	150.7	148.0	124.2	124.1	122.8
	100	Mean	0.032	0.036	0.034	0.032	0.034	0.035
		Std	0.017	0.016	0.016	0.015	0.015	0.018
		IF	170.7	134.4	143.3	124.3	133.0	125.5
	200	Mean	0.036	0.041	0.038	0.037	0.039	0.040
		Std	0.019	0.020	0.019	0.018	0.017	0.020
		IF	138.0	136.8	134.8	121.8	126.2	138.2
T=1000	50	Mean	0.030	0.030	0.029	0.026	0.028	0.026
		Std	0.015	0.011	0.010	0.009	0.010	0.010
		IF	198.0	195.1	198.5	185.6	184.5	188.3
	100	Mean	0.036	0.037	0.035	0.032	0.034	0.032
		Std	0.016	0.012	0.012	0.011	0.012	0.013
		IF	182.3	157.5	170.8	180.43	169.5	163.0
	200	Mean	0.041	0.040	0.039	0.037	0.038	0.037
		Std	0.018	0.014	0.012	0.013	0.014	0.015
		IF	170.1	164.7	153.1	169.5	175.9	188.5
T=2000	50	Mean	0.030	0.031	0.031	0.025	0.026	0.024
		Std	0.010	0.009	0.009	0.008	0.007	0.006
		IF	198.9	196.7	193.0	240.4	239.4	235.0
	100	Mean	0.036	0.038	0.037	0.032	0.033	0.031
		Std	0.011	0.011	0.010	0.010	0.010	0.008
		IF	183.2	173.3	167.5	216.5	232.4	230.0
	200	Mean	0.041	0.042	0.041	0.037	0.039	0.037
		Std	0.012	0.011	0.011	0.012	0.012	0.010
		IF	161.8	157.3	157.8	205.7	200.4	219.6

1. T is the number of time series observations for each asset and N is the number of particles used in PGAS.
2. Mean is the average posterior mean across replications.
3. Std is the standard error of posterior mean across replications.
4. IF is the average inefficiency factor across replications calculated as suggested in Kim et al. (1998).

The number of assets considered for simulation is three and the true value of in total 18 parameters are given by:

1. $\mu_{h1} = \mu_{h2} = \mu_{h3} = 0.3$ and $\mu_{q1} = \mu_{q2} = \mu_{q3} = 0.7$.
2. $\phi_{h1} = \phi_{h2} = \phi_{h3} = 0.9$ and $\phi_{q1} = \phi_{q2} = \phi_{q3} = 0.8$.
3. $\sigma^2_{h1} = \sigma^2_{h2} = \sigma^2_{h3} = 0.05$ and $\sigma^2_{q1} = \sigma^2_{q2} = \sigma^2_{q3} = 0.05$.

All the simulation results reported in this section is based on 5000 MCMC iterations, among which the first 1000 samples are discarded as burn-in period[19]. We consider three different sample sizes, namely $T = 500$, $T = 1000$ and $T = 2000$, as well as three number of particles, namely $N = 50$, $N = 100$ and $N = 200$. It is worthwhile to mention that, the simulated data used across different particle numbers for given sample size are the same, while it changes when the sample size increases.

The estimation of the mean parameters μ's is reported in Table 15. It can be seen that even for a short sample span like 500 and a relatively small number of particles like 50, the estimated mean for both h and q are close to the true values, in spite of an upward bias for μ_h and downward bias for μ_q. Given a particular number of particles used, as sample size increases, the bias in the estimator for μ_q improves, but that is not the case for μ_h. Nevertheless, as expected, the standard deviation of both estimators shrinks toward zero with more observations. On the other hand, an increasing number of particles seems to alleviate the bias substantially. For example, when the sample size is 2000, the bias in the estimator for the mean of h diminish from 0.016 to 0.008 if 200 particles are included instead of 50. A similar improvement applies to μ_q. In terms of standard deviation, an increasing number of particles has no effect.

Table 16 presents the results related to the persistence parameter ϕ's. The estimation is already very accurate, with 500 observations and 50 particles, with very small bias and a low standard deviation. When 200 particles are used, the bias almost vanishes.

The performance of estimated variance parameters σ^2's is shown in Table 17. A substantial downward bias can be observed for the variance estimator of both volatilities and

[19]Plot of autocorrelation function suggests that the MCMC sampling has converged after at most 1000 iterations.

correlations when 50 particles are considered. This bias is insensitive to the increasing of the number of observations. Fortunately, it can be improved by using more particles. Indeed, when N=200, the bias becomes much smaller, although it seems that a larger number of particles are necessary to fully cancel this bias.

As for average IF, in general, it does not vary much as we either change the sample size or the number of particles. Consistent with previous researches, the IF is lowest for mean parameter μ and highest for σ^2. Compared with traditional single-move or multi-move Gibbs sampler, our new PGAS sampler enjoys a much better mixing than the former, and in most cases, is comparable with the latter. In summary, the simulation results confirm that our chosen approach works well for the particular model we consider in this paper, and in the light of excellent performance, 200 particles will be used for all of our empirical applications.

At last, we report the filtered latent volatility and correlation factors, together with the 95% confidence interval and the true processes, in Figure 21 and 22. These figures show that our sampling of latent variables based on particle filter is reliable.

5.6 Empirical Analysis

In this section, we consider two empirical implementation of our models. The first one studies weekly exchange rate returns of three European currencies, while the second one looks at the daily returns of three major Asian equity index.

As a comparison, we consider the following four models:

1. GFT-MSV model: The general model that we propose in Section 3.
2. GFT-MSV model with equi-persistence: Same as general GFT-MSV model except for the additional assumption of 'equi-persistence' on correlation factors.
3. CC-MSV: Model with constant correlation matrix over time, similar to the one considered in Harvey et al. (1994)
4. DCC-MSV: Model proposed in Asai and McAleer (2009), where a DCC structure with a Wishart transition dynamics is used to characterize the movement of correlation matrix.

Figure 21: True and filtered latent volatily factors

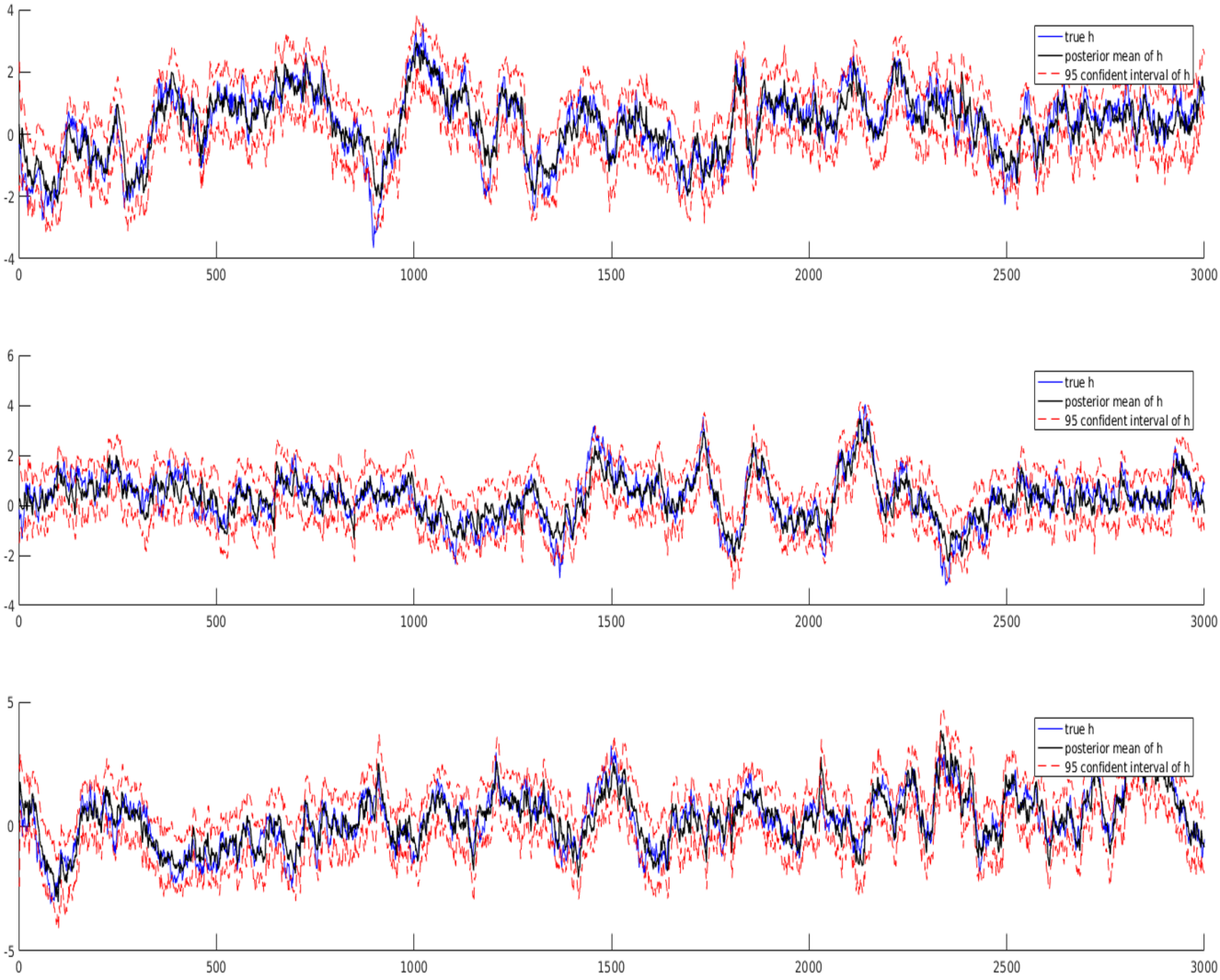
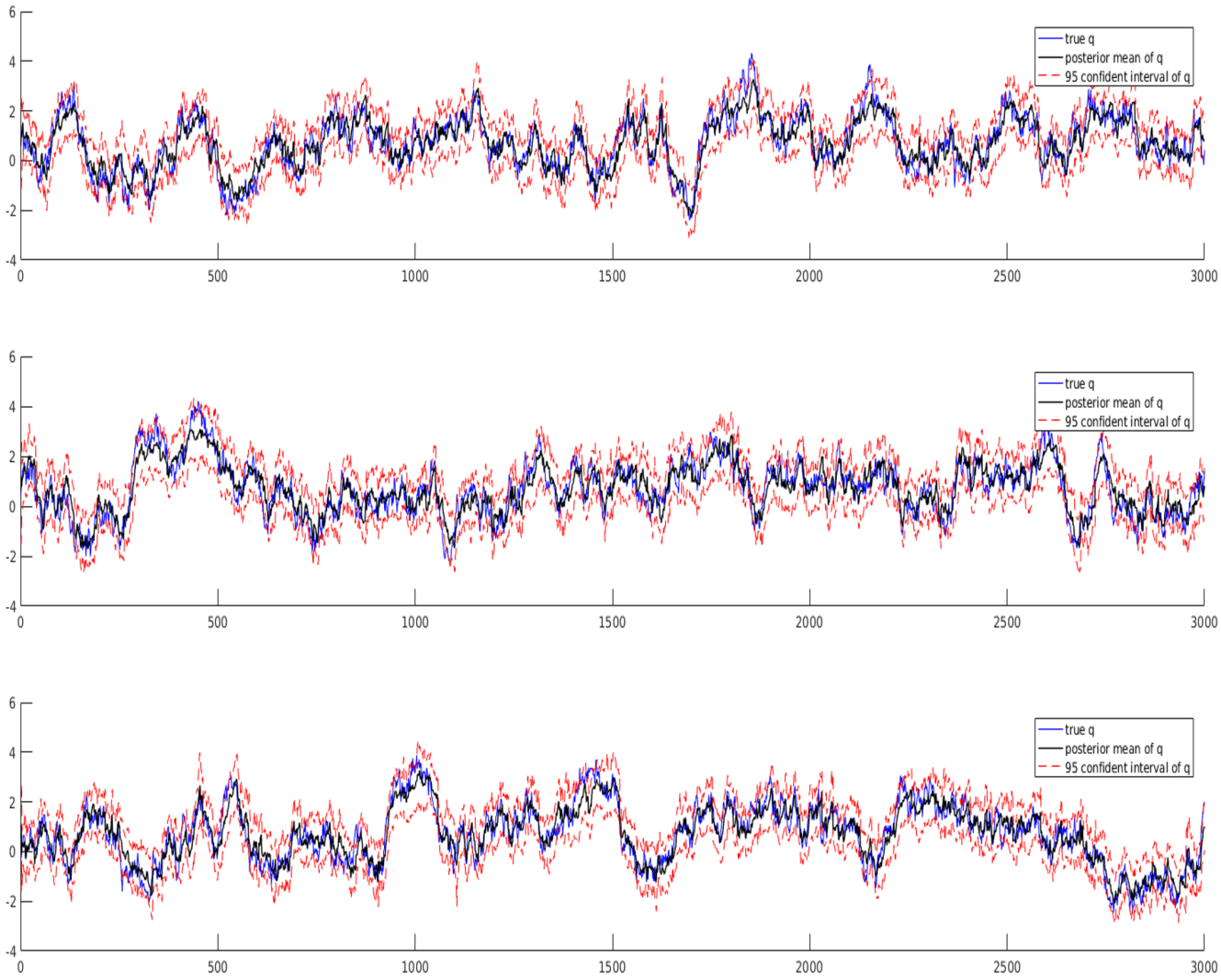


Figure 22: True and filtered latent correlation factors



Note that in all these models, the parameterization of volatility movement is the same, and we focus on the different assumptions imposed upon the dynamics of correlations. To estimate and compare these models in a unified framework, we treat all of them as a nonlinear state-space model and then exploit the PGAS algorithm introduced above. The details of the estimation procedure and the computation of likelihood are presented in the appendix.

5.6.1 Weekly Foreign Exchange Rates

In the first application, the data used are 1406 weekly mean-corrected log-returns of Euro, Pound sterling and Swiss franc, all against the US dollar, from January 1993 to December 2019. Ex ante, these three series are expected to be significantly correlated, as the underlying economies are very closely connected. The three time series are plotted in Figure 23. Filtered log-variance of each pair of exchange rate and filtered correlations between them are plotted in Figure 25 and Figure 26 respectively.

Table 18 reports the posterior statistics of parameters related to volatility dynamics. Three measures are provided namely posterior mean, posterior standard deviation, and 95% confidence interval. It can be seen that the result is very similar to the existing literature. All log-variances have a very high level of persistence, with autoregressive root close to 1. This reaffirms that our new estimation strategy works well for the current setup.

Table 19 reports the posterior statistics of parameters related to underlying correlation factors. For the second model, since persistence levels and standard deviations are restricted to be the same among all q's, we have only one ϕ_q and one σ_q^2. For the third model, no ϕ_q and one σ_q^2 is present as all correlations are assumed to be constant over time. An important finding from this table is that, if these three correlation factors are allowed to have separate dynamics, q_1 will have a very low level of persistence and seems to be highly stationary. Specifically, the posterior mean of ϕ_q in this case is $(0.212, 0.911, 0.812)'$. This is in sharp contrast to the model with only one parameter governing all the factors. In the latter case, a ϕ_q as high as 0.957 will be concluded. To see whether this additional flexibility brings any improvement in terms of model performance, we further compare the

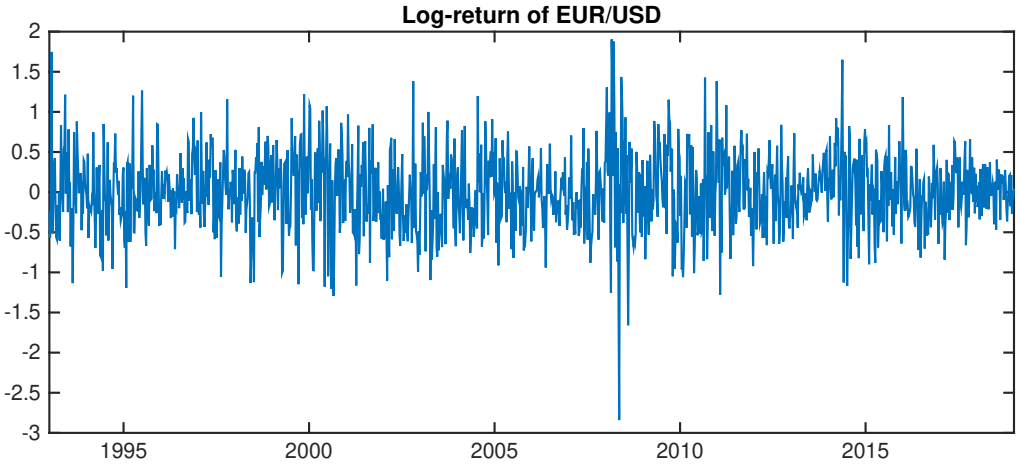
Table 18: Posterior statistics of volatilities with exchange rate data

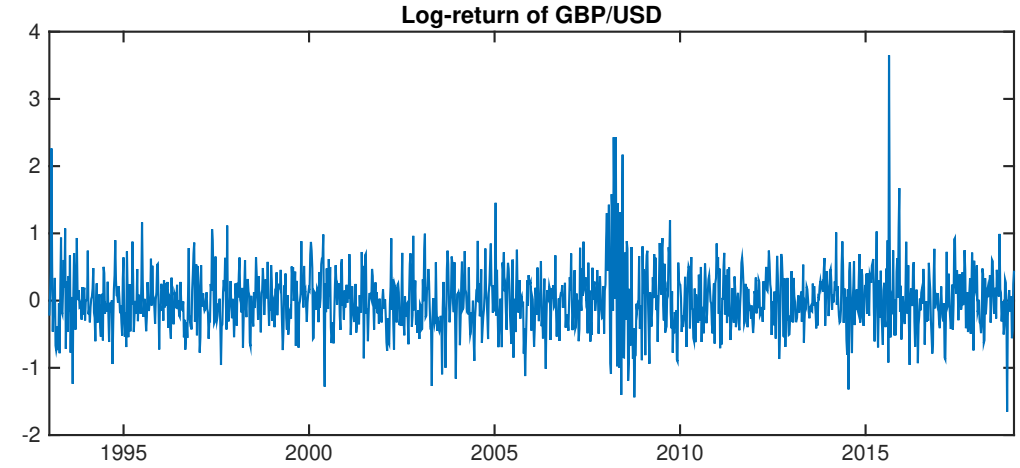
		GFT-MSV	GFT-MSV (Equipersistence)	CC-MSV	DCC-MSV
μ_{h1}	Mean	-1.701	-1.68	-1.872	-1.764
	SD	0.120	0.108	0.182	0.098
	95%CI	[-1.963,-1.470]	[-1.904,-1.468]	[-2.231,-1.507]	[-1.964,-1.575]
μ_{h2}	Mean	-1.804	-1.774	-1.864	-1.893
	SD	0.101	0.099	0.142	0.097
	95%CI	[-2.005,-1.599]	[-1.973,-1.582]	[-2.132,-1.565]	[-2.034,-1.655]
μ_{h3}	Mean	-1.504	-1.481	-1.688	-1.578
	SD	0.106	0.093	0.131	0.083
	95%CI	[-1.733,-1.317]	[-1.672,-1.311]	[-1.939,-1.406]	[-1.752,-1.422]
ϕ_{h1}	Mean	0.980	0.977	0.98	0.972
	SD	0.016	0.016	0.01	0.015
	95%CI	[0.958,0.994]	[0.952,0.993]	[0.962,0.992]	[0.939,0.990]
ϕ_{h2}	Mean	0.962	0.960	0.966	0.970
	SD	0.019	0.031	0.015	0.036
	95%CI	[0.921,0.987]	[0.849,0.989]	[0.929,0.987]	[0.843,0.990]
ϕ_{h3}	Mean	0.970	0.957	0.960	0.962
	SD	0.027	0.037	0.014	0.036
	95%CI	[0.927,0.991]	[0.888,0.988]	[0.928,0.982]	[0.847,0.987]
σ^2_{h1}	Mean	0.005	0.004	0.01	0.005
	SD	0.005	0.004	0.006	0.005
	95%CI	[0.002,0.008]	[0.002,0.008]	[0.005,0.020]	[0.002,0.012]
σ^2_{h2}	Mean	0.013	0.013	0.017	0.006
	SD	0.008	0.016	0.009	0.016
	95%CI	[0.004,0.028]	[0.004,0.080]	[0.008,0.037]	[0.002,0.066]
σ^2_{h3}	Mean	0.006	0.008	0.022	0.005
	SD	0.007	0.008	0.009	0.012
	95%CI	[0.002,0.021]	[0.003,0.017]	[0.008,0.037]	[0.002,0.032]

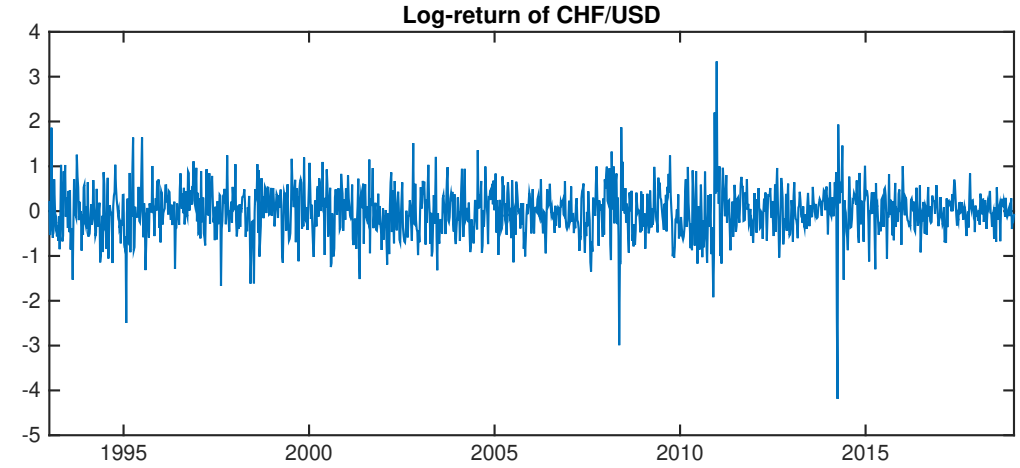
1. Mean is the posterior mean based on 4000 MCMC samples after a 1000 burn-in period.
2. SD is the numerical standard errors of the posterior means.
3. 95% CI is constructed using the 2.5th and 97.5th percentiles of the MCMC draws.

integrated log-likelihood of each model, and the results are reported in the last row of Table 16. Apparently, the fully flexible model has a much better likelihood and dominates all other restricted versions. We can then conclude that adopting the generalized Fisher transformation is indeed beneficial for modeling multivariate asset returns.

Figure 23: Time series of exchange rate returns







5.6.2 Daily Stock Market Indices

Our second application uses 2237 daily mean-corrected log-returns of three stock market indices, namely Hong Kong Heng Seng Index, Nikkei 225 Index of Tokyo Stock Exchange, and SSE composite Index of Shanghai Stock Exchange. The time span of the data is from the beginning of 2005 to the end of 2014. Note that we intentionally pick a data span that includes the Great Financial Crisis period. This exercise is interesting, as it can provide some insights on the co-movement of these three arguably most important market indices in Asia. The three series are plotted in Figure 26. Filtered log-variance of each stock index and filtered correlations between them are plotted in Figure 27 and Figure 28 respectively.

The general conclusion from this dataset is similar to the case in exchange rate data. For example, as shown in Table 20, all the log-variances are strongly persistent with autoregressive root approximately equal to one. As for correlation factors, presented in Table 21, the flexible model results in a ϕ_q equal to $(0.519, 0.654, 0.664)'$ while the equi-persistence model suggests a higher persistence level equal to 0.859. In this dataset, however, the improvement of integrated log-likelihood is not as impressive as in the previous example. The reason is that the autoregressive roots of three correlation factors are not as dispersed as in exchange rate examples, and thus equi-persistence is not as restrictive as in that example. Anyway, even for such a scenario, using the new flexible parameterization lead to a better result. Indeed, an increase of log-likelihood by four could have significant empirical relevance.

5.7 Conclusion

We propose a novel multivariate stochastic volatility model in this paper, using a generalized version of Fisher's z-transformation to characterize the dynamics of correlation structure in a highly flexible manner. The leading features are that our model will automatically generate a positive definite correlation matrix, and the driving forces underlying volatilities and correlations are fully separated. Different from much of the existing literature, when making inference for our model, we apply the state-of-the-art particle-

Table 19: Posterior statistics of correlations and log-likelihood with exchange rate data

		GFT-MSV	GFT-MSV (Equipersistence)	CC-MSV		DCC-MSV
μ_{q1}	Mean	0.705	0.615	0.697	k	9.690
	SD	0.029	0.126	0.127		0.145
	95%CI	[0.647,0.762]	[0.327,0.812]	[0.438,0.946]		[9.430,10.064]
μ_{q2}	Mean	1.511	1.392	1.103	d	0.005
	SD	0.085	0.135	0.123		0.014
	95%CI	[1.337,1.674]	[1.078,1.600]	[0.871,1.354]		[-0.015,0.026]
μ_{q3}	Mean	0.444	0.379	0.359	a_{11}	5.639
	SD	0.047	0.126	0.121		0.487
	95%CI	[0.351,0.533]	[0.114,0.576]	[0.128,0.589]		[4.900,6.506]
ϕ_{q1}	Mean	0.212	0.957		a_{21}	-1.232
	SD	0.107	0.019			0.137
	95%CI	[0.014,0.417]	[0.916,0.988]			[-1.484,-1.001]
ϕ_{q2}	Mean	0.911			a_{22}	-4.312
	SD	0.022				0.425
	95%CI	[0.862,0.948]				[-5.053,-3.644]
ϕ_{q3}	Mean	0.812			a_{31}	1.751
	SD	0.055				0.085
	95%CI	[0.685,0.898]				[1.612,1.912]
σ^2_{q1}	Mean	0.116	0.018		a_{32}	0.098
	SD	0.018	0.007			0.097
	95%CI	[0.083,0.155]	[0.010,0.032]			[-0.086,0.298]
σ^2_{q2}	Mean	0.061			a_{33}	4.779
	SD	0.014				0.397
	95%CI	[0.039,0.091]				[4.135,5.459]
σ^2_{q3}	Mean	0.055				
	SD	0.016				
	95%CI	[0.031,0.095]				
Integrated log-likelihood		-1271.4	-1449.4	-1424.2		-1441.2

1. Mean is the posterior mean based on 4000 MCMC samples after a 1000 burn-in period.
2. SD is the numerical standard errors of the posterior means.
3. 95% CI is constructed using the 2.5th and 97.5th percentiles of the MCMC draws.
4. Integrated likelihood is computed by averaging the marginal likelihood produced by particle filter at each MCMC iteration.
5. For DCC-MSV model, the notation of parameters is consistent with the usage in Asai and McAleer (2009).

Figure 24: Filtered log-variance of each exchange rate pair

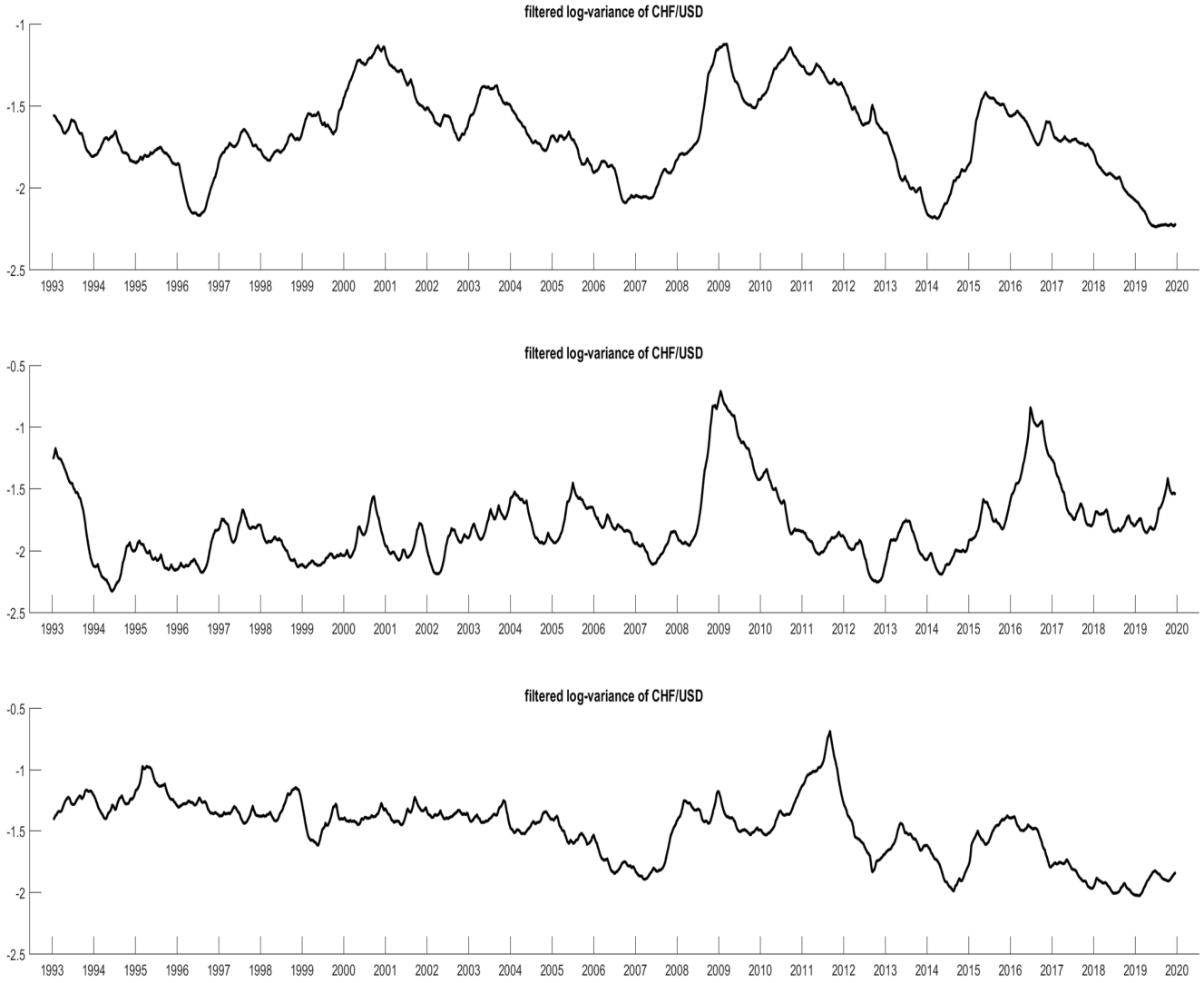


Figure 25: Filtered correlation between each exchange rate pair

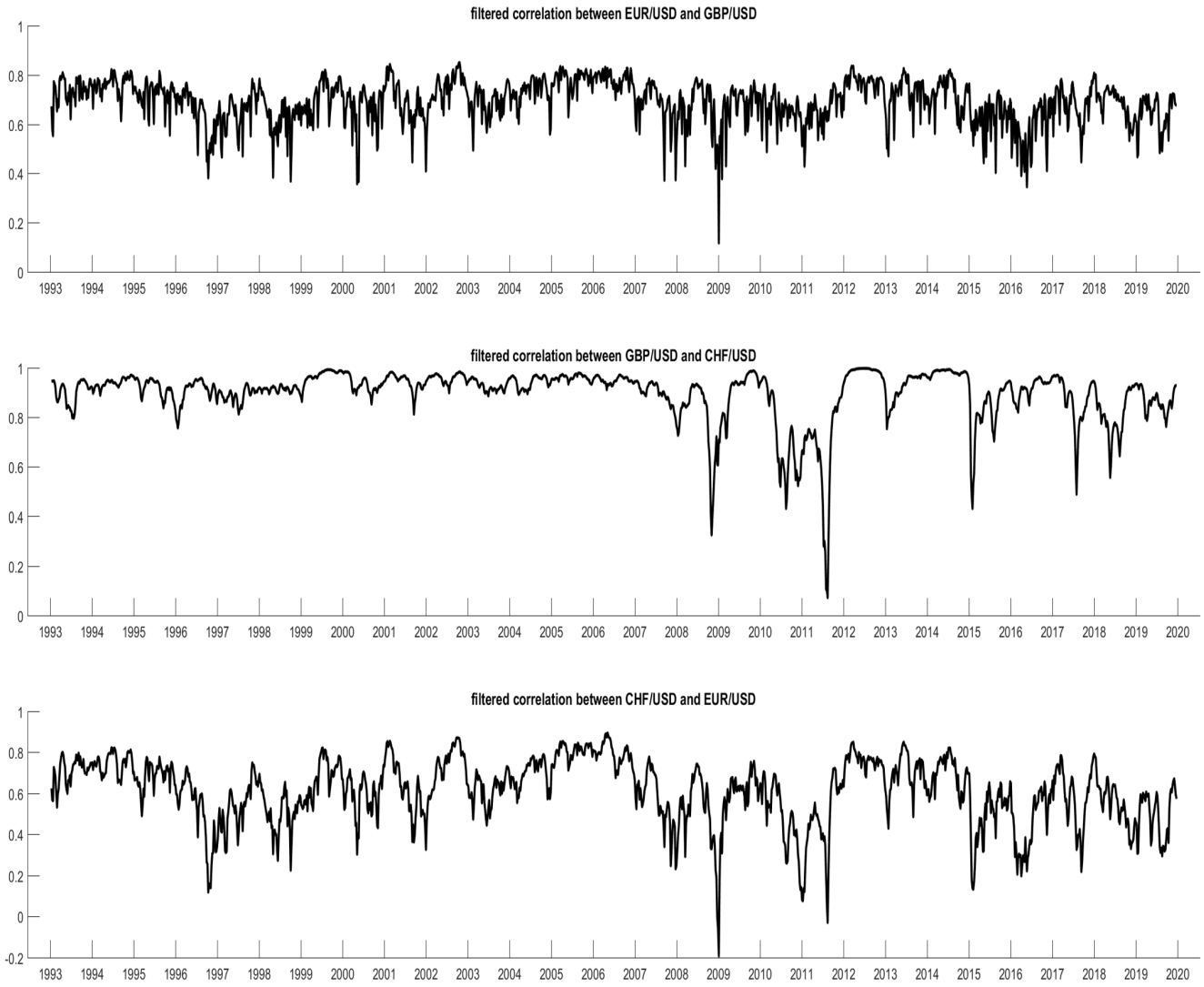


Figure 26: Time series of stock index returns

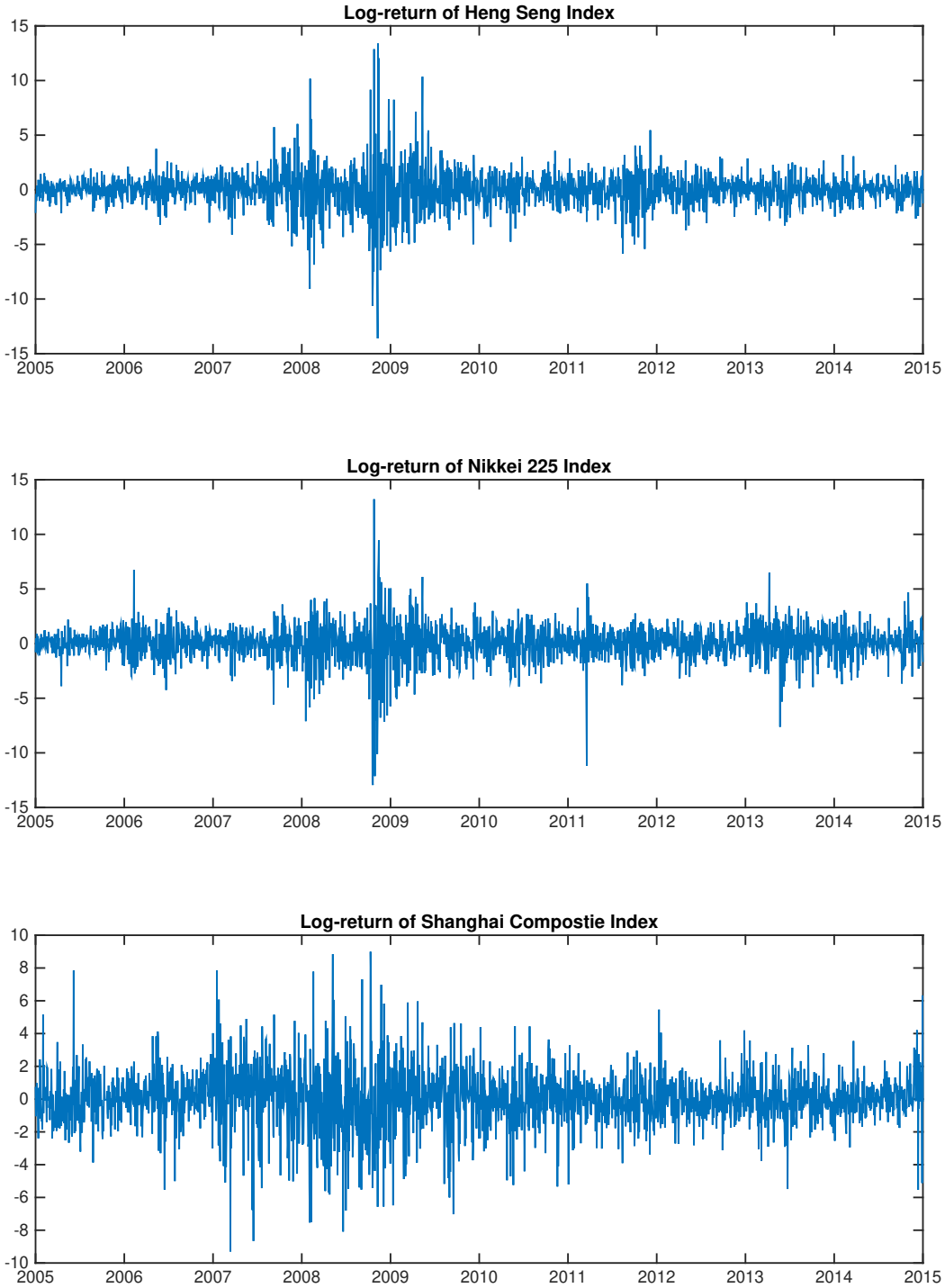


Table 20: Posterior statistics of volatilities with daily stock index data

		GFT-MSV	GFT-MSV (Equipersistence)	CC-MSV	DCC-MSV
μ_{h1}	Mean	0.370	0.343	0.397	0.375
	SD	0.246	0.257	0.247	0.232
	95%CI	[-0.144,0.851]	[-0.180,0.845]	[-0.096,0.889]	[-0.085,0.849]
μ_{h2}	Mean	0.525	0.502	0.539	0.525
	SD	0.175	0.175	0.171	0.161
	95%CI	[0.186,0.872]	[0.147,0.836]	[0.195,0.887]	[0.203,0.839]
μ_{h3}	Mean	0.758	0.738	0.767	0.757
	SD	0.203	0.216	0.215	0.195
	95%CI	[0.357,1.156]	[0.287,1.155]	[0.335,1.194]	[0.347,1.144]
ϕ_{h1}	Mean	0.991	0.990	0.987	0.989
	SD	0.004	0.005	0.005	0.005
	95%CI	[0.982,0.995]	[0.981,0.95]	[0.977,0.994]	[0.979,0.995]
ϕ_{h2}	Mean	0.981	0.980	0.978	0.979
	SD	0.007	0.006	0.007	0.008
	95%CI	[0.968,0.991]	[0.965,0.990]	[0.962,0.989]	[0.961,0.991]
ϕ_{h3}	Mean	0.983	0.984	0.983	0.983
	SD	0.007	0.006	0.006	0.007
	95%CI	[0.968,0.993]	[0.970,0.994]	[0.968,0.993]	[0.967,0.994]
σ^2_{h1}	Mean	0.011	0.014	0.016	0.012
	SD	0.006	0.006	0.008	0.007
	95%CI	[0.006,0.021]	[0.008,0.023]	[0.009,0.027]	[0.006,0.026]
σ^2_{h2}	Mean	0.020	0.023	0.024	0.020
	SD	0.006	0.006	0.007	0.008
	95%CI	[0.013,0.030]	[0.013,0.035]	[0.016,0.040]	[0.010,0.036]
σ^2_{h3}	Mean	0.022	0.020	0.022	0.019
	SD	0.009	0.007	0.008	0.009
	95%CI	[0.011,0.035]	[0.011,0.034]	[0.011,0.039]	[0.008,0.034]

1. Mean is the posterior mean based on 4000 MCMC samples after a 1000 burn-in period.
2. SD is the numerical standard errors of the posterior means.
3. 95% CI is constructed using the 2.5th and 97.5th percentiles of the MCMC draws.

Table 21: Posterior statistics of correlations and log-likelihood with daily stock index data

		GFT-MSV	GFT-MSV (Equipersistence)	CC-MSV		DCC-MSV
μ_{q1}	Mean	0.715	0.631	0.674	k	39.634
	SD	0.028	0.065	0.071		0.174
	95%CI	[0.663,0.771]	[0.476,0.736]	[0.540,0.827]		[39.261,39.968]
μ_{q2}	Mean	0.542	0.5	0.535	d	0.000
	SD	0.027	0.052	0.067		0.015
	95%CI	[0.489,0.595]	[0.368,0.578]	[0.408,0.658]		[-0.017,0.015]
μ_{q3}	Mean	0.123	0.109	0.123	a_{11}	1.915
	SD	0.024	0.047	0.065		0.115
	95%CI	[0.077,0.171]	[-0.003,0.179]	[-0.009,0.248]		[1.745,2.063]
ϕ_{q1}	Mean	0.519	0.859		a_{21}	-0.962
	SD	0.115	0.141			0.098
	95%CI	[0.277,0.733]	[0.524,0.994]			[-1.077,-0.829]
ϕ_{q2}	Mean	0.654			a_{22}	-0.685
	SD	0.104				0.075
	95%CI	[0.429,0.836]				[-0.793,-0.583]
ϕ_{q3}	Mean	0.664			a_{31}	1.561
	SD	0.090				0.070
	95%CI	[0.479,0.820]				[1.451,1.654]
σ^2_{q1}	Mean	0.069	0.007		a_{32}	0.058
	SD	0.020	0.010			0.039
	95%CI	[0.037,0.114]	[0.001,0.038]			[-0.012,0.133]
σ^2_{q2}	Mean	0.029			a_{33}	1.323
	SD	0.010				0.044
	95%CI	[0.015,0.050]				[1.256,1.399]
σ^2_{q3}	Mean	0.017				
	SD	0.008				
	95%CI	[0.008,0.038]				
Integrated log-likelihood		-11035	-11039	-11056		-11152

1. Mean is the posterior mean based on 4000 MCMC samples after a 1000 burn-in period.
2. SD is the numerical standard errors of the posterior means.
3. 95% CI is constructed using the 2.5th and 97.5th percentiles of the MCMC draws.
4. Integrated likelihood is computed by averaging the marginal likelihood produced by particle filter at each MCMC iteration.

Figure 27: Filtered log-variance of each stock index

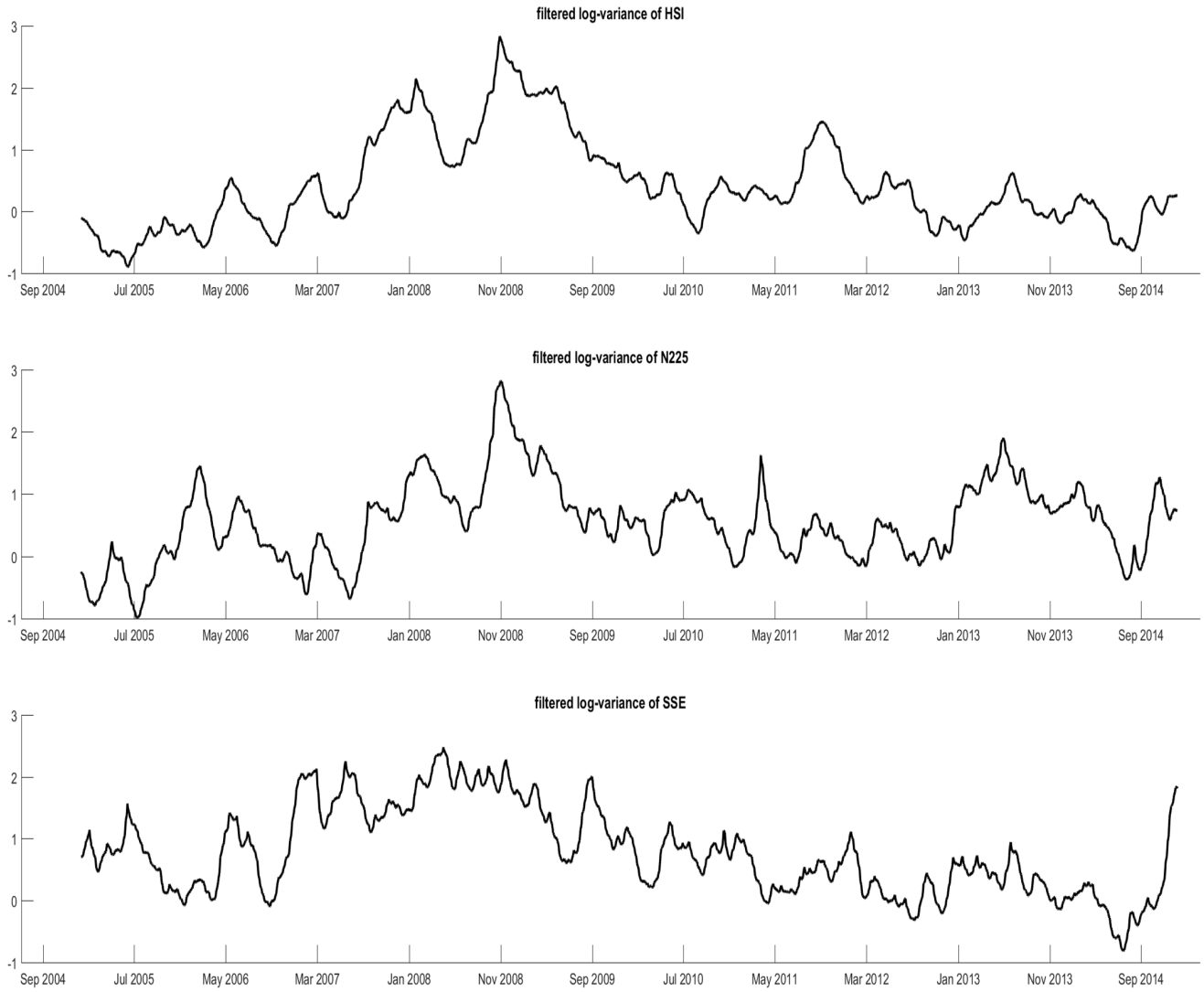
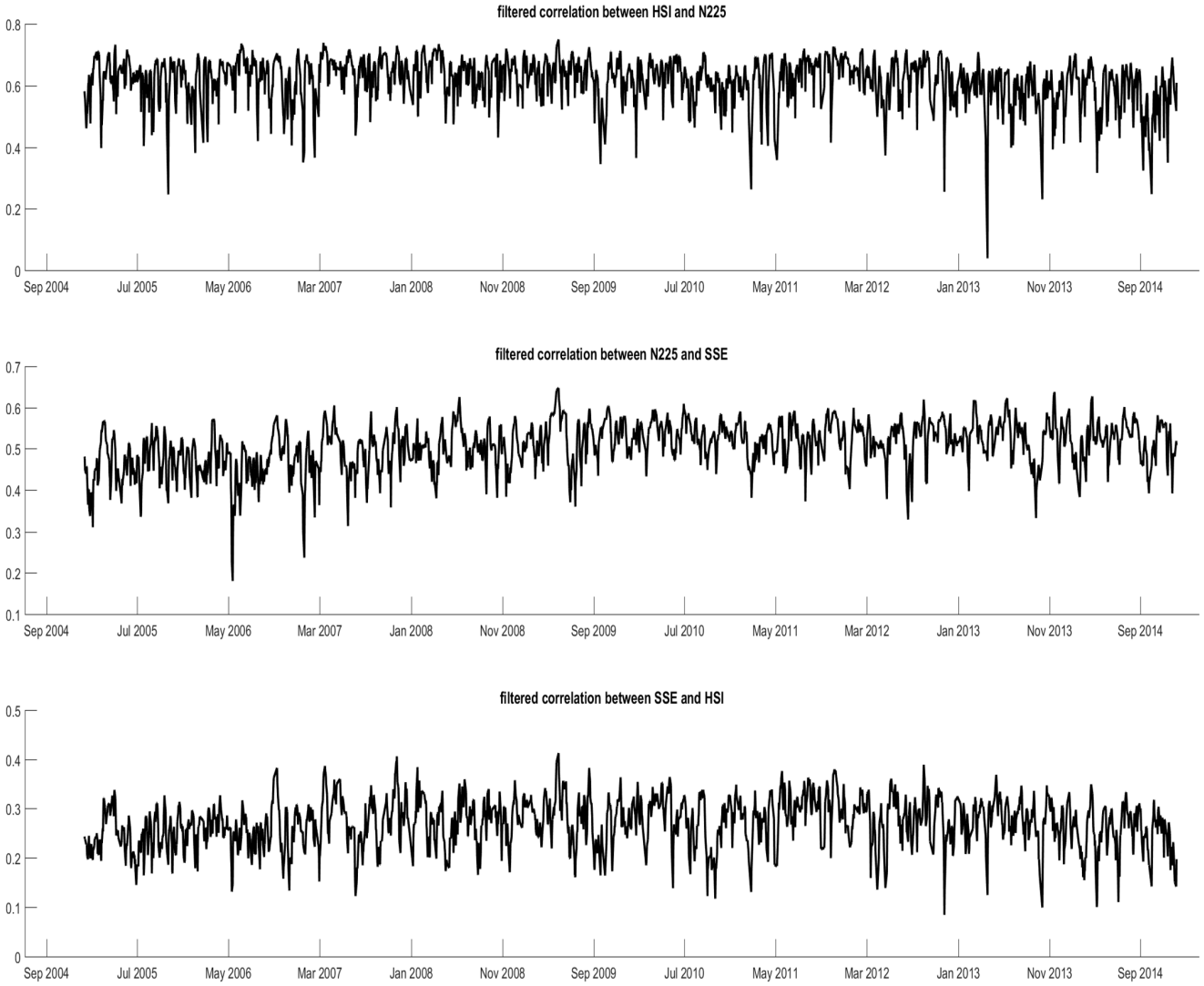


Figure 28: Filtered correlation between each stock index



filter-based MCMC technique. We show through simulation that the technique we choose works well for our particular model.

Though our model enjoys excellent properties, there are still many potential improvements. First of all, we do not address the leverage effect in our model. To accommodate this salient empirical fact, one can extend the current model by adding correlations between innovations to returns and to volatilities. Another fundamental task is to find a parsimonious yet plausible way to tackle the vast parameter space when a large pool of assets are considered simultaneously in our framework. In this regard, particular assumptions like (block) equicorrelation structures or latent factor models must be imposed. Last but not least, a comparison of our model with existing competitors based on out-of-sample performance is also of great interest. These issues are left for our future work.

References

ANDREWS, D. W. (1993): "Tests for parameter instability and structural change with unknown change point," Econometrica, 821–856.

ANDREWS, D. W. AND W. PLOBERGER (1994): "Optimal tests when a nuisance parameter is present only under the alternative," Econometrica, 1383–1414.

ANDRIEU, C., A. DOUCET, AND R. HOLENSTEIN (2010): "Particle markov chain monte carlo methods," Journal of the Royal Statistical Society: Series B (Statistical Methodology), 72, 269–342.

ANG, A. AND G. BEKAERT (2007): "Stock return predictability: Is it there?" The Review of Financial Studies, 20, 651–707.

ARCHAKOV, I. AND P. R. HANSEN (2018a): "A new parametrization of correlation matrices," Working paper, 1–32.

——— (2018b): "Web Appendix for "A new parametrization of correlation matrices"," Working paper, 1–44.

ARULAMPALAM, M. S., S. MASKELL, N. GORDON, AND T. CLAPP (2002): "A tutorial on particle filters for online nonlinear/non-Gaussian Bayesian tracking," IEEE Transactions on signal processing, 50, 174–188.

ASAI, M. AND M. MCALEER (2006): "Asymmetric multivariate stochastic volatility," Econometric Reviews, 25, 453–473.

——— (2009): "The structure of dynamic correlations in multivariate stochastic volatility models," Journal of Econometrics, 150, 182–192.

ASAI, M., M. MCALEER, AND J. YU (2006): "Multivariate stochastic volatility: a review," Econometric Reviews, 25, 145–175.

BAI, J. (1997): "Estimation of a change point in multiple regression models," Review of Economics and Statistics, 79, 551–563.

BAI, J. AND P. PERRON (1998): "Estimating and testing linear models with multiple structural changes," Econometrica, 47–78.

BAUWENS, L., S. LAURENT, AND J. V. ROMBOUTS (2006): "Multivariate GARCH models: a survey," Journal of applied econometrics, 21, 79–109.

BROTO, C. AND E. RUIZ (2004): "Estimation methods for stochastic volatility models: a survey," Journal of Economic Surveys, 18, 613–649.

CAI, Z., Y. WANG, AND Y. WANG (2015): "Testing instability in a predictive regression model with nonstationary regressors," Econometric Theory, 31, 953–980.

CAMPBELL, J. Y. AND T. VUOLTEENAHO (2004): "Bad beta, good beta," American Economic Review, 94, 1249–1275.

CAMPBELL, J. Y. AND M. YOGO (2006): "Efficient tests of stock return predictability," Journal of Financial Economics, 81, 27–60.

CAVANAGH, C. L., G. ELLIOTT, AND J. H. STOCK (1995): "Inference in models with nearly integrated regressors," Econometric Theory, 11, 1131–1147.

CHAN, D., R. KOHN, AND C. KIRBY (2006): "Multivariate stochastic volatility models with correlated errors," Econometric Reviews, 25, 245–274.

CHIB, S., Y. OMORI, AND M. ASAI (2009): "Multivariate stochastic volatility," in Handbook of Financial Time Series, Springer, 365–400.

CHOPIN, N., S. S. SINGH, ET AL. (2015): "On particle Gibbs sampling," Bernoulli, 21, 1855–1883.

CIZEAU, P., Y. LIU, M. MEYER, C.-K. PENG, AND H. STANLEY (1997): "Volatility distribution in the S&P500 stock index," Physica A: Statistical Mechanics and its Applications, 245, 441–445.

DEMETRESCU, M., I. GEORGIEV, P. M. RODRIGUES, AND A. R. TAYLOR (2020): "Testing for episodic predictability in stock returns," Journal of Econometrics.

DUFFIE, D. AND K. J. SINGLETON (1990): "Simulated moments estimation of Markov models of asset prices," .

ELLIOTT, G. (2011): "A control function approach for testing the usefulness of trending variables in forecast models and linear regression," Journal of Econometrics, 164, 79–91.

ELLIOTT, G. AND U. K. MÜLLER (2006): "Efficient tests for general persistent time variation in regression coefficients," The Review of Economic Studies, 73, 907–940.

——— (2007): "Confidence sets for the date of a single break in linear time series regressions," Journal of Econometrics, 141, 1196–1218.

ENGLE, R. AND B. KELLY (2012): "Dynamic equicorrelation," Journal of Business & Economic Statistics, 30, 212–228.

FERSON, W. E. AND R. W. SCHADT (1996): "Measuring fund strategy and performance in changing economic conditions," The Journal of Finance, 51, 425–461.

FORSBERG, L. AND E. GHYSELS (2007): "Why do absolute returns predict volatility so well?" Journal of Financial Econometrics, 5, 31–67.

GEORGIEV, I., D. I. HARVEY, S. J. LEYBOURNE, AND A. R. TAYLOR (2018): "Testing for parameter instability in predictive regression models," Journal of Econometrics, 204, 101–118.

GIRAITIS, L. AND P. C. B. PHILLIPS (2006): "Uniform limit theory for stationary autoregression," Journal of Time Series Analysis, 27, 51–60.

GONZALO, J. AND J.-Y. PITARAKIS (2012): "Regime-specific predictability in predictive regressions," Journal of Business & Economic Statistics, 30, 229–241.

——— (2017): "Inferring the predictability induced by a persistent regressor in a predictive threshold model," Journal of Business & Economic Statistics, 35, 202–217.

GOURIÉROUX, C., J. JASIAK, AND R. SUFANA (2009): "The Wishart autoregressive process of multivariate stochastic volatility," Journal of Econometrics, 150, 167–181.

GRANGER, C. W. AND N. HYUNG (2004): "Occasional structural breaks and long memory with an application to the S&P 500 absolute stock returns," Journal of Empirical Finance, 11, 399–421.

HAMILTON, J. D. (1994): Time series analysis, vol. 2, Princeton university press Princeton.

HANSEN, B. E. (2000): "Testing for structural change in conditional models," Journal of Econometrics, 97, 93–115.

HARVEY, A., E. RUIZ, AND N. SHEPHARD (1994): "Multivariate stochastic variance models," The Review of Economic Studies, 61, 247–264.

ISHIHARA, T., Y. OMORI, AND M. ASAI (2016): "Matrix exponential stochastic volatility with cross leverage," Computational Statistics & Data Analysis, 100, 331–350.

JACQUIER, E., N. G. POLSON, AND P. E. ROSSI (2002): "Bayesian analysis of stochastic volatility models," Journal of Business & Economic Statistics, 20, 69–87.

JANSSON, M. AND M. J. MOREIRA (2006): "Optimal inference in regression models with nearly integrated regressors," Econometrica, 74, 681–714.

JIANG, L., X. WANG, AND J. YU (2018): "New distribution theory for the estimation of structural break point in mean," Journal of Econometrics, 205, 156–176.

KAWAKATSU, H. (2006): "Matrix exponential GARCH," Journal of Econometrics, 134, 95–128.

KIM, C.-J., J. C. MORLEY, AND C. R. NELSON (2005): "The structural break in the equity premium," Journal of Business & Economic Statistics, 23, 181–191.

KIM, S., N. SHEPHARD, AND S. CHIB (1998): "Stochastic volatility: likelihood inference and comparison with ARCH models," The Review of Economic Studies, 65, 361–393.

KOSTAKIS, A., T. MAGDALINOS, AND M. P. STAMATOGIANNIS (2014): "Robust econometric inference for stock return predictability," The Review of Financial Studies, 28, 1506–1553.

KOTHARI, S. P. AND J. SHANKEN (1997): "Book-to-market, dividend yield, and expected market returns: A time-series analysis," Journal of Financial Economics, 44, 169–203.

KUROSE, Y. AND Y. OMORI (2016): "Dynamic equicorrelation stochastic volatility," Computational Statistics & Data Analysis, 100, 795–813.

——— (2018): "Multiple-block dynamic equicorrelations with realized measures, leverage and endogeneity," Econometrics and Statistics.

LAMONT, O. (1998): "Earnings and expected returns," Journal of Finance, 53, 1563–1587.

LETTAU, M. AND S. LUDVIGSON (2001): "Consumption, aggregate wealth, and expected stock returns," Journal of Finance, 56, 815–849.

LINDSTEN, F., M. I. JORDAN, AND T. B. SCHÖN (2014): "Particle Gibbs with ancestor sampling," The Journal of Machine Learning Research, 15, 2145–2184.

LINDSTEN, F. AND T. B. SCHÖN (2013): "Backward simulation methods for Monte Carlo statistical inference," Foundations and Trends® in Machine Learning, 6, 1–143.

LIU, X. AND L. PENG (2019): "Asymptotic theory and unified confidence region for an autoregressive model," Journal of Time Series Analysis, 40, 43–65.

MAGDALINOS, T. AND P. C. B. PHILLIPS (2009): "Limit theory for cointegrated systems with moderately integrated and moderately explosive regressors," Econometric Theory, 25, 482–526.

MELINO, A. AND S. M. TURNBULL (1990): "Pricing foreign currency options with stochastic volatility," Journal of Econometrics, 45, 239–265.

NORIEGA, A. E. AND D. VENTOSA-SANTAULÀRIA (2007): "Spurious regression and trending variables," Oxford Bulletin of Economics and Statistics, 69, 439–444.

OMORI, Y. AND T. WATANABE (2008): "Block sampler and posterior mode estimation for asymmetric stochastic volatility models," Computational Statistics & Data Analysis, 52, 2892–2910.

PAYE, B. S. (2012): "'Déjà vol': Predictive regressions for aggregate stock market volatility using macroeconomic variables," Journal of Financial Economics, 106, 527–546.

PAYE, B. S. AND A. TIMMERMANN (2006): "Instability of return prediction models," Journal of Empirical Finance, 13, 274–315.

PHILIPOV, A. AND M. E. GLICKMAN (2006): "Multivariate stochastic volatility via Wishart processes," Journal of Business & Economic Statistics, 24, 313–328.

PHILLIPS, P. C. B. AND J. H. LEE (2013): "Predictive regression under various degrees of persistence and robust long-horizon regression," Journal of Econometrics, 177, 250–264.

PHILLIPS, P. C. B. AND T. MAGDALINOS (2007): "Limit theory for moderate deviations from a unit root," Journal of Econometrics, 136, 115–130.

——— (2009): "Econometric inference in the vicinity of unity," Singapore Management University, CoFie Working Paper, 7.

PHILLIPS, P. C. B., S. SHI, AND J. YU (2015a): "Testing for multiple bubbles: Historical episodes of exuberance and collapse in the S&P 500," International Economic Review, 56, 1043–1078.

——— (2015b): "Testing For Multiple Bubbles: Limit Theory Of Real-Time Detectors," International Economic Review, 56, 1079–1134.

PHILLIPS, P. C. B., Y. WU, AND J. YU (2011): "Explosive behavior in the 1990s Nasdaq: When did exuberance escalate asset values?" International Economic Review, 52, 201–226.

PITARAKIS, J.-Y. (2017): "A simple approach for diagnosing instabilities in predictive regressions," Oxford Bulletin of Economics and Statistics, 79, 851–874.

RAPACH, D. E. AND M. E. WOHAR (2006): "Structural breaks and predictive regression models of aggregate US stock returns," Journal of Financial Econometrics, 4, 238–274.

ROSSI, B. (2005): "Optimal tests for nested model selection with underlying parameter instability," Econometric Theory, 21, 962–990.

SANDMANN, G. AND S. J. KOOPMAN (1998): "Estimation of stochastic volatility models via Monte Carlo maximum likelihood," Journal of Econometrics, 87, 271–301.

SHEPHARD, N. AND M. K. PITT (1997): "Likelihood analysis of non-Gaussian measurement time series," Biometrika, 84, 653–667.

TAYLOR, S. J. (1986): Modelling financial time series, World Scientific.

WANG, X. AND J. YU (2015): "Limit theory for an explosive autoregressive process," Economics Letters, 126, 176–180.

——— (2016): "Double asymptotics for explosive continuous time models," Journal of Econometrics, 193, 35–53.

WATANABE, T. AND Y. OMORI (2004): "A multi-move sampler for estimating non-Gaussian time series models: Comments on Shephard & Pitt (1997)," Biometrika, 246–248.

WELCH, I. AND A. GOYAL (2008): "A comprehensive look at the empirical performance of equity premium prediction," The Review of Financial Studies, 21, 1455–1508.

YAMAUCHI, Y. AND Y. OMORI (2019): "Multivariate stochastic volatility model with realized volatilities and pairwise realized correlations," Journal of Business & Economic Statistics, 1–17.

YANG, B., W. LONG, L. PENG, AND Z. CAI (2020): "Testing the predictability of US housing price index returns based on an IVX-AR model," Journal of the American Statistical Association, 1–22.

YU, J. AND R. MEYER (2006): "Multivariate stochastic volatility models: Bayesian estimation and model comparison," Econometric Reviews, 25, 361–384.

A Appendix to Chapter 2

This appendix includes the technical details and proof of the results reported in Chapter 2.

Proof of Theorem 2.1. (a). Note first that $k_n = o(n)$ by assumption and $y_n^0 = o_p(\sqrt{n})$ by the part (a) of Lemma 3.1 in PM. Also note that $\rho_n^n = o(1)$ when $c < 0$. Then, using expression (2.4), we have $y_n = \frac{d}{c}k_n(\rho_n^n - 1) + y_n^0 = \frac{d}{c} \cdot o(n) \cdot o(1) + o_p(\sqrt{n}) = o_p(n)$.

(b). This comes immediately from Lindeberg-Feller central limit theorem.

(c). From model (2.1), it is easy to get $y_t - y_{t-1} = d + \frac{c}{k_n}y_{t-1} + u_t$. Then, taking summation on both sides, $\frac{c}{k_n}\sum y_{t-1} = y_n - y_0 - dn - \sum u_t$. Based on the limiting distribution derived in (a) and (b), we have $\frac{c}{nk_n}\sum y_{t-1} = -d + o_p(1)$ and thus $\frac{\sum y_{t-1}}{nk_n} \Rightarrow -\frac{d}{c}$.

(d). Expression (2.4) leads to $y_{t-1} = -\frac{d}{c}k_n(1-\rho_n^{t-1}) + y_{t-1}^0$. Hence, $n^{-1/2}k_n^{-1}\sum y_{t-1}u_t = \frac{d}{c}\frac{1}{\sqrt{n}}\sum \rho_n^{t-1}u_t - \frac{d}{c}\frac{1}{\sqrt{n}}\sum u_t + n^{-1/2}k_n^{-1}\sum y_{t-1}^0 u_t$. The second term will converge to a normal distribution by (b). The third term is $o_p(1)$ because PM shows that $\sum y_{t-1}^0 u_t$ is $O_p(\sqrt{nk_n})$[20]. Therefore, we have $n^{-1/2}k_n^{-1}\sum y_{t-1}u_t = -\frac{d}{c}\frac{1}{\sqrt{n}}\sum u_t + o_p(1) \Rightarrow -\frac{d}{c}Z$.

(e). Squaring both sides of Model (2.1), we get $y_t^2 = d^2 + \rho_n^2 y_{t-1}^2 + u_t^2 + 2d\rho_n y_{t-1} + 2du_t + 2\rho_n y_{t-1}u_t$, which leads to $(\rho_n^2 - 1)\sum y_{t-1}^2 = y_n^2 - y_0^2 - d^2 n - \sum u_t^2 - 2d\rho_n \sum y_{t-1} - 2d\sum u_t - 2\rho_n \sum y_{t-1}u_t$. Based on the limiting results above and the assumption about y_0, it is straightforward to derive that $(\rho_n^2 - 1)/(nk_n)\sum y_{t-1}^2 = -2d\frac{\sum y_{t-1}}{nk_n} + o_p(1) \Rightarrow 2d^2/c$. Using the fact that $\rho_n^2 - 1 = (2c/k_n)[1 + o(1)]$, we can write this result as $\frac{\sum y_{t-1}^2}{nk_n^2} \Rightarrow \frac{d^2}{c^2}$.

■

Proof of Theorem 2.3. (a). We first derive the order of y_n^0. Since $y_n^0 = \rho_n^n y_0 + \sum \rho_n^{n-t}u_t$, we have $\frac{y_n^0}{\rho_n^n\sqrt{k_n}} = \frac{y_0}{\sqrt{k_n}} + \frac{1}{\sqrt{k_n}}\sum \rho_n^{-t}u_t = o_p(1) + Y_n = O_p(1)$, due to the assumption on initialization and the definition of Y_n. Note that when $c > 0$, $\rho_n^{-n} = o(1)$. Hence, we

[20] See part (b) of Theorem 3.2 in that paper.

conclude that $\frac{y_n}{\rho_n^n k_n} = \frac{d}{c}(1-\rho_n^{-n}) + \frac{1}{\sqrt{k_n}} \cdot O_p(1) \Rightarrow \frac{d}{c}$.

(b). Identical to the proof of Theorem 2.1(c), we use the relationship that $\frac{c}{k_n}\sum y_{t-1} = y_n - y_0 - dn - \sum u_t$. Based on the convergence in (a), we easily obtain that $\frac{c}{\rho_n^n k_n^2}\sum y_{t-1} = \frac{d}{c} + o_p(1) \Rightarrow \frac{d}{c}$, which is exactly the result we need to prove.

(c). Since $\sum y_{t-1}u_t = \sum[\frac{d}{c}k_n(\rho_n^{t-1}-1) + y_{t-1}^0]u_t =$
$\frac{d}{c}k_n(\sum \rho_n^{t-1}u_t - \sum u_t) + \sum y_{t-1}^0 u_t$, we have that
$\rho_n^{-n}k_n^{-3/2}\sum y_{t-1}u_t = \rho_n^{-n}k_n^{-3/2}\sum y_{t-1}^0 u_t + \frac{d}{c}\frac{1}{\sqrt{k_n}}\sum \rho_n^{-(n-t)-1}u_t - \frac{d}{c}\frac{1}{\sqrt{k_n}}\rho_n^{-n}\sum u_t$
$= O_p(1)/\sqrt{k_n} + \frac{d}{c}X_n - \frac{d}{c}(\rho_n^{-n}\sqrt{\frac{n}{k_n}})\frac{1}{\sqrt{n}}\sum u_t = \frac{d}{c}X_n + o_p(1) \Rightarrow \frac{d}{c}X$.

(d). Again, parallel to the proof of Theorem 2.1(e), we have $(\rho_n^2-1)\sum y_{t-1}^2 = y_n^2 - y_0^2 - d^2 n - \sum u_t^2 - 2d\rho_n\sum y_{t-1} - 2d\sum u_t - 2\rho_n\sum y_{t-1}u_t$. Using the limiting behavior justified above, one can compare the order of terms on the right-hand side. It's not hard to find that y_n^2 goes to infinity faster than all other terms. Thus, we have $\frac{(\rho_n^2-1)}{\rho_n^{2n}k_n^2}\sum y_{t-1}^2 = \frac{d^2}{c^2} + o_p(1) \Rightarrow \frac{d^2}{c^2}$.

■

Proof of Corollary 2.1. (a). It is obvious and hence the proof is omitted.

(b). Starting from the definition of t_{ρ_n}, it can be obtained that

$$
\begin{aligned}
t_{\rho_n} &= \frac{(\hat{\rho}_n - \rho_n)[n\sum y_{t-1}^2 - (\sum y_{t-1})^2]^{1/2}}{[n\cdot\hat{\sigma}^2]^{1/2}} \\
&= \left[\rho_n^n(\rho_n-1)^{-3/2}\frac{\hat{\rho_n}-\rho_n}{\hat{\sigma}}\right]\cdot\frac{[n\sum y_{t-1}^2 - (\sum y_{t-1})^2]^{1/2}}{\rho_n^n(\rho_n-1)^{-3/2}} \\
&= \left[\rho_n^n(\rho_n-1)^{-3/2}\frac{\hat{\rho_n}-\rho_n}{\hat{\sigma}}\right]\cdot c^{\frac{3}{2}}\left[\frac{\sum y_{t-1}^2}{k_n^3\rho_n^{2n}} - \left(\frac{\sum y_{t-1}}{\sqrt{n}k_n^{3/2}\rho_n^{2n}}\right)^2\right]^{1/2} \\
&= \left[\rho_n^n(\rho_n-1)^{-3/2}\frac{\hat{\rho_n}-\rho_n}{\hat{\sigma}}\right]\cdot c^{\frac{3}{2}}\left(\frac{d^2}{2c^3}+o_p(1)\right)^{1/2} \\
&\Rightarrow c^{3/2}\left(\frac{d^2}{2c^3}\right)^{1/2} N\left(0,\frac{2}{d^2}\right) = N(0,1).
\end{aligned}
$$

(c). It suffices to show that Z and X are independent of each other. To prove that, let

$$\tilde{W}_n = \frac{1}{\sqrt{n}} \sum_{t=1}^{\lfloor n \rfloor} u_t, \quad W^* = \frac{1}{\sqrt{n}} \sum_{t=\lfloor \sqrt{n} \rfloor + 1}^{n} u_t,$$

and

$$\tilde{X}_n = \frac{1}{\sqrt{n}} \sum_{t=\lfloor \sqrt{n} \rfloor + 1}^{n} \rho_n^{-(n-t)-1} u_t, \quad X^* = \frac{1}{\sqrt{n}} \sum_{t=1}^{\lfloor n \rfloor} \rho_n^{-(n-t)-1} u_t.$$

Then X_n^* and W_n^* are independently distributed because they involve disjoint sets of u's. As $n \to \infty$, we have

$$E(W_n - W_n^*)^2 = E(\tilde{W}_n)^2 = \frac{\lfloor \sqrt{n} \rfloor}{n} \sigma^2 \to 0$$

and

$$\begin{aligned} E(X_n - X_n^*)^2 = E(\tilde{X}_n)^2 &= \frac{\sigma^2}{n \rho_n^{2(n-1)}} \sum_{t=\lfloor \sqrt{n} \rfloor + 1}^{n} \rho_n^{2t} \\ &= \frac{\rho_n^{2(\lfloor \sqrt{n} \rfloor + 1)} (1 - \rho_n^{2(n - \lfloor \sqrt{n} \rfloor)})}{(1 - \rho_n^2) n \rho_n^{2(n-1)}} \sigma^2 \\ &= \frac{(1 - \rho_n^{2(\lfloor \sqrt{n} \rfloor - n)}) \rho_n^4}{2c(n/k_n)} \sigma^2 \to 0 \end{aligned}$$

Hence, $W_n - W_n^*$ and $X_n - X_n^*$ converge with probability 1 to 0. Therefore, the asymptotic independence between X and Z follows. The independence between t_d and t_{ρ_n} as n goes to infinity then comes immediately. ■

B Appendix to Chapter 3

This appendix includes the technical details and proof of the results reported in the Chapter 3. Before proceeding to the proof, it's convenience to summarize in the following lemma the limiting theory from KMS and Phillips and Magdalinos (2009) that we will use.

Lemma B.1 *Let* $V_{xz} = \int_0^\infty e^{rC}(\int_0^\infty e^{sC}\Omega_{xx}e^{sC}ds)e^{rC_z}dr$*, where* Ω_{xx} *is the long-run variance of* u_{xt}*. Then, absent of break in both* α *and* β*, we have under Assumption 1:*
(i). the sample covariance satisfies that

$$\frac{1}{T^{(1+\min(\delta_z,\gamma_x))/2}}\sum_{t=1}^{[\tau T]}\tilde{z}_{t-1}(u_{yt}-\bar{u}_T) \Rightarrow U(\tau),$$

where $U(\cdot)$ *is a Brownian motion with variance* $\sigma_y^2\tilde{V}$ *and*

$$\tilde{V} = \begin{cases} \int_0^\infty e^{rC_z}\Omega_{xx}e^{rC_z}dr & ,\text{if } \gamma_x > \delta_z \\ \int_0^\infty e^{rC_z}(CV_{xz}C_z + C_zV_{xz}'C)e^{rC_z}dr & ,\text{if } \gamma_x = \delta_z \\ \int_0^\infty e^{rC}\Omega_{xx}e^{rC}dr & ,\text{if } 0 < \gamma_x < \delta_z \\ E(x_{0,1}x_{0,1}') & ,\text{if } \gamma_x = 0 \end{cases}$$

where $x_{0,t} = \sum_{j=0}^\infty (I_k + C)^j u_{x,t-j}$ *is stationary version of* x_t *when* $\gamma_x = 0$*.*

(ii). the sample second moment satisfies that

$$\frac{1}{T^{1+\min(\delta_z,\gamma_x)}}\sum_{t=1}^{[\tau T]}\tilde{z}_{t-1}(x_{t-1}-\bar{x}_{n-1})' \Rightarrow \Psi(\tau)$$

$$\Psi(\tau) = \begin{cases} -C_z^{-1}\Omega_{xx}(\tau + \int_0^\tau \underline{J}_C dJ_C') & ,\text{if } \gamma_x > 1 \\ -C_z^{-1}\Omega_{xx}(\tau + \int_0^\tau \underline{B} dB') & ,\text{if } \gamma_x = 1 \\ -\tau C_z^{-1}(\Omega_{xx} + \int_0^\infty e^{rC}\Omega_{xx}e^{rC}drC) & ,\text{if } \delta_z < \gamma_x < 1 \\ -\tau CV_{xz} & ,\text{if } \gamma_x = \delta_z \\ \tau\int_0^\infty e^{rC}\Omega_{xx}e^{rC}dr & ,\text{if } 0 < \gamma_x < \delta_z \\ \tau E(x_{0,1}x_{0,1}') & ,\text{if } \gamma_x = 0 \end{cases}$$

where $B(\cdot)$ is a k-dimensional standard Brownian motion, $J_C(\tau) = \int_0^\tau e^{C(\tau-s)}dB(s)$ is an Ornstein-Uhlenbeck process and $\underline{B}(\tau) = B(\tau) - \int_0^1 B(s)ds$ and $\underline{J}_C(\tau) = J_C(\tau) - \int_0^1 J_C(s)ds$ are corresponding demeaned process.

(iii). The joint convergence applies and the limiting objects in (i) and (ii) are independent.

Proof of Theorem 3.1. Let Y denote the vector stacking all demeaned y_t and X be the matrix collecting all demeaned x_{t-1}, i.e.

$$Y = (y_1 - \bar{y}_T, y_2 - \bar{y}_T, ..., y_n - \bar{y}_T)' \text{ and } X = (x_0 - \bar{x}_{T-1}, x_1 - \bar{x}_{T-1}, ...x_{T-1-}\bar{x}_{T-1})'.$$

Also, use u_y to stack all the demeaned u_{yt} in a vector. For any $1 \leqslant t \leqslant T$, we define X_t to be a $T \times k$ matrix, whose first t rows are the same as X while the rest are all zeros. In a similar fashion, let $Z = (z_0, z_1, ...z_{T-1})'$ collect all the IVX, and $Z_t = (z_0, ..., z_t, 0, ...0)'$ be the corresponding time-t truncated matrix. Given these notations, we can rephrase the original model as

$$Y = X\beta_2 + X_t\theta + u_y, \tag{B.1}$$

where $\theta = \beta_2 - \beta_1$ measures the magnitude of structural break. In the following, we will use θ_t to denote the break size associated with partition at time t. Therefore, given any particular t, testing the structural break in β is equivalent to testing the null hypothesis $\theta_t = 0$. Define $M_{xz} = I_k - X(Z'X)^{-1}Z'$, which is idempotent and orthogonal to both X and Z. Multiplying M_{xz} on both sides of equation (B.1), we deduce that $M_{xz}Y = M_{xz}X_t + M_{xz}u_y$. Using $M_{xz}Z_t$ as the instrumental variables for $M_{xz}X_t$, we obtain an estimator for parameter of interest θ

$$\tilde{\theta}_t = (Z_t'M_{xz}X_t)^{-1}Z_t'M_{xz}Y. \tag{B.2}$$

It's straightforward to show that $\tilde{\theta}_t = \tilde{\beta}_1(t) - \tilde{\beta}_2(t)$. To obtain its limiting characterization,

first notice that

$$\begin{aligned}
\tilde{\theta}_t - \theta_t &= (Z_t' M_{xz} X_t)^{-1} Z_t' M_{xz} u_y \\
&= (Z_t' X_t - Z_t' X (Z'X)^{-1} Z' X_t)^{-1} (Z_t' u_y - Z_t' X (Z'X)^{-1} Z' u_y) \\
&= (Z_t' X_t - Z_t' X_t (Z'X)^{-1} Z_t' X_t)^{-1} (Z_t' u_y - Z_t' X_t (Z'X)^{-1} Z' u_y) \\
&= \left[\sum_{t=1}^{[\tau T]} \tilde{z}_t (x_t - \bar{x}_{T-1})' - \sum_{t=1}^{[\tau T]} \tilde{z}_t (x_t - \bar{x}_{T-1})' \left(\sum_{t=1}^{T} \tilde{z}_t (x_t - \bar{x}_{T-1})' \right)^{-1} \sum_{t=1}^{[\tau T]} \tilde{z}_t (x_t - \bar{x}_{T-1})' \right]^{-1} \\
&\quad \left[\sum_{t=1}^{[\tau T]} \tilde{z}_t (u_{yt} - \bar{u}_{T-1})' - \sum_{t=1}^{[\tau T]} \tilde{z}_t (x_t - \bar{x}_{T-1})' \left(\sum_{t=1}^{T} \tilde{z}_t (x_t - \bar{x}_{T-1})' \right)^{-1} \sum_{t=1}^{T} \tilde{z}_t (u_{yt} - \bar{u}_y) \right]
\end{aligned}$$

Applying the results in Lemma B.1, we can derive that

$$T^{\frac{1+\min[\gamma_x, \delta_z]}{2}} (\tilde{\theta}_t - \theta_t) \Rightarrow \left[\Psi(\tau) - \Psi(\tau)\Psi(1)^{-1}\Psi(\tau)' \right]^{-1} \left(U(\tau) - \Psi(\tau)\Psi(1)^{-1} U(1) \right). \tag{B.3}$$

Now, we turn to analyze the variance estimator $\tilde{Q}_1(t)$ and $\tilde{Q}_2(t)$. Ignoring the second-order bias correction, we have

$$\begin{aligned}
\tilde{Q}_1(t) &= (Z_t' X_t)^{-1} Z_t' Z_t (X_t' Z_t)^{-1} \\
&= \left(\sum_{t=1}^{[\tau n]} \tilde{z}_t (x_t - \bar{x}_{n-1})' \right)^{-1} \left(\sum_{t=1}^{[\tau n]} \tilde{z}_t \tilde{z}_t' \right) \left(\sum_{t=1}^{[\tau n]} (x_t - \bar{x}_{n-1}) \tilde{z}_t' \right)^{-1}
\end{aligned}$$

and thus

$$T^{1+\min[\gamma_x, \delta_z]} \tilde{Q}_1(t) \Rightarrow \Psi(\tau)^{-1} \left(\tau \sigma_y^2 \tilde{V} \right) \Psi(\tau)^{-1'} \tag{B.4}$$

Similarly, we can obtain that

$$T^{1+\min[\gamma_x, \delta_z]} \tilde{Q}_2(t) \Rightarrow (\Psi(1) - \Psi(\tau))^{-1} \left((1-\tau) \sigma_y^2 \tilde{V} \right) (\Psi(1) - \Psi(\tau))^{-1'} \tag{B.5}$$

Combining all these together leads to the result that

$$
\begin{aligned}
W_{bt} &= \left(\tilde{\beta}_2(t) - \tilde{\beta}_1(t)\right)' \left[\tilde{Q}_1(t) + \tilde{Q}_2(t)\right]^{-1} \left(\tilde{\beta}_2(t) - \tilde{\beta}_1(t)\right) \\
&= \left[T^{\frac{1+\min[\delta_z,\gamma_x]}{2}} \left(\tilde{\beta}_2(t) - \tilde{\beta}_1(t)\right)\right]' \left[T^{1+\min[\delta_z,\gamma_x]} \left(\tilde{Q}_1(t) + \tilde{Q}_2(t)\right)\right]^{-1} \left[T^{\frac{1+\min[\delta_z,\gamma_x]}{2}} \left(\tilde{\beta}_2(t) - \tilde{\beta}_1(t)\right)\right] \\
&\Rightarrow \left(U(\tau) - \Psi(\tau)\Psi(1)^{-1}U(1)\right)' \left[\Psi(\tau) - \Psi(\tau)\Psi(1)^{-1}\Psi(\tau)\right]^{-1'} \\
&\quad \left[\Psi(\tau)^{-1}\left(\tau\sigma_y^2\tilde{V}\right)\Psi(\tau)^{-1'} + (\Psi(1) - \Psi(\tau))^{-1}\left((1-\tau)\sigma_y^2\tilde{V}\right)(\Psi(1) - \Psi(\tau))^{-1'}\right]^{-1} \\
&\quad \left[\Psi(\tau) - \Psi(\tau)\Psi(1)^{-1}\Psi(\tau)\right]^{-1} \left(U(\tau) - \Psi(\tau)\Psi(1)^{-1}U(1)\right)
\end{aligned}
$$

Let $A = \Psi(\tau)$, $C = \Psi(1)$ and $\Sigma = \sigma_y^2\tilde{V}$, then the above object can be written as

$$
\begin{aligned}
W_{bt} \Rightarrow & \left(U(\tau) - AC^{-1}U(1)\right)' \left(A - AC^{-1}A\right)^{-1'} \left[\tau A^{-1}\Sigma A^{-1'}\right. \\
& \left. + (1-\tau)(C-A)^{-1}\Sigma(C-A)^{-1'}\right]^{-1} \left(A - AC^{-1}A\right)^{-1} \left(U(\tau) - AC^{-1}U(1)\right) \\
= & \left(U(\tau) - AC^{-1}U(1)\right)' \left\{\left(A - AC^{-1}A\right)\left[\tau A^{-1}\Sigma A^{-1'}\right.\right. \\
& \left.\left. + (1-\tau)(C-A)^{-1}\Sigma(C-A)^{-1'}\right]\left(A - AC^{-1}A\right)'\right\}^{-1} \left(U(\tau) - AC^{-1}U(1)\right) \\
= & \left(U(\tau) - AC^{-1}U(1)\right)' \left\{\tau\left[(A - AC^{-1}A)A^{-1}\Sigma A^{-1'}(A - AC^{-1}A)'\right]\right. \\
& \left. + (1-\tau)\left[(A - AC^{-1}A)(C-A)^{-1}\Sigma(C-A)^{-1'}(A - AC^{-1}A)'\right]\right\}^{-1} \left(U(\tau) - AC^{-1}U(1)\right) \\
= & \left(U(\tau) - AC^{-1}U(1)\right)' \left\{\tau(I - AC^{-1})\Sigma(I - AC^{-1})' + (1-\tau)(AC^{-1})\Sigma(AC^{-1})'\right\}^{-1} \\
& \left(U(\tau) - AC^{-1}U(1)\right)
\end{aligned}
$$

Since $U(\cdot)$ is known to be a Brownian motion with variance Σ, this can be further simpli-

fied to be

$$
\begin{aligned}
W_{bt} \Rightarrow & \Big(B(\tau) - AC^{-1}B(1)\Big)' \Big\{\tau(I - AC^{-1})(I - AC^{-1})' \\
& + (1-\tau)(AC^{-1})(AC^{-1})'\Big\}^{-1} \Big(B(\tau) - AC^{-1}B(1)\Big) \\
= & \Big(B(\tau) - \Psi(\tau)\Psi(1)^{-1}B(1)\Big)' \Big\{\tau(I - \Psi(\tau)\Psi(1)^{-1})(I - \Psi(\tau)\Psi(1)^{-1})' \\
& + (1-\tau)(\Psi(\tau)\Psi(1)^{-1})(\Psi(\tau)\Psi(1)^{-1})'\Big\}^{-1} \Big(B(\tau) - \Psi(\tau)\Psi(1)^{-1}B(1)\Big),
\end{aligned}
\tag{B.6}
$$

where $B(\cdot)$ is a standard Brownian motion. From Lemma 1, it's easy to observe that

$$
\Psi(\tau)\Psi(1)^{-1} = \begin{cases} (\tau I_k + \int_0^\tau \underline{B} dB')(I_k + \int_0^1 \underline{B} dB')^{-1} & \text{, if } \gamma_x > 1 \\ (\tau I_k + \int_0^\tau \underline{J}_C dJ_C')(I_k + \int_0^1 \underline{J} dJ_C')^{-1} & \text{, if } \gamma_x = 1 \\ \tau I_k & \text{, otherwise} \end{cases}.
$$

This completes the proof of Theorem 3.1. ■

Proof of Corollary 3.1. When $\gamma_x < 1$, we have $R(\tau) = \tau I_k$ and thus $M(\tau) = \tau(I_k - \tau I_k)(I_k - \tau I_k)' + (1-\tau)\tau I_k(\tau I_k)' = [\tau(1-\tau)^2 + (1-\tau)\tau^2] I_k = \tau(1-\tau) I_k$. Hence, in this case, the limiting object in the Theorem 3.1 will reduce to $(B(\tau) - \tau B(1))'(B(\tau) - \tau B(1))/[\tau(1-\tau)]$ and justify the claim. ■

Proof of Corollary 3.2. (i). Using the above notations, it's straightforward to obtain that under the null hypothesis $\beta_1 = \beta_2 = 0$, we have

$$
\tilde{\beta}' \tilde{Q}^{-1} \tilde{\beta} \Rightarrow U(1)' [\Psi(1)^{-1'} \Sigma \Psi(1)^{-1}]^{-1} U(1).
$$

Combining this result with Theorem 3.1, we can deduce that

$$
\begin{aligned}
&W_{\beta=0} + W_{bt} \Rightarrow \\
&\begin{pmatrix} \Psi(1)^{-1}U(1) \\ U(\tau) - \Psi(\tau)\Psi(1)^{-1}U(1) \end{pmatrix}' \begin{pmatrix} \Psi(1)^{-1'}\Sigma\Psi(1)^{-1} & 0_{k\times k} \\ 0_{k\times k} & \tau(I-\Psi(\tau)\Psi(1)^{-1})\Sigma(I-\Psi(\tau)\Psi(1)^{-1})' \\ & +(1-\tau)(\Psi(\tau)\Psi(1)^{-1})\Sigma(\Psi(\tau)\Psi(1)^{-1})' \end{pmatrix}^{-1} \\
&\quad \begin{pmatrix} \Psi(1)^{-1}U(1) \\ U(\tau) - \Psi(\tau)\Psi(1)^{-1}U(1) \end{pmatrix} \\
&= \begin{pmatrix} \Psi(1)^{-1}B(1) \\ B(\tau) - \Psi(\tau)\Psi(1)^{-1}B(1) \end{pmatrix}' \begin{pmatrix} \Psi(1)^{-1'}\Psi(1)^{-1} & 0_{k\times k} \\ 0_{k\times k} & \tau(I-\Psi(\tau)\Psi(1)^{-1})(I-\Psi(\tau)\Psi(1)^{-1})' \\ & +(1-\tau)(\Psi(\tau)\Psi(1)^{-1})(\Psi(\tau)\Psi(1)^{-1})' \end{pmatrix}^{-1} \\
&\quad \begin{pmatrix} \Psi(1)^{-1}B(1) \\ B(\tau) - \Psi(\tau)\Psi(1)^{-1}B(1) \end{pmatrix} \\
&= \begin{pmatrix} B(1) \\ B(\tau) - R(\tau)B(1) \end{pmatrix}' \begin{pmatrix} I_k & 0_{k\times k} \\ 0_{k\times k} & M(\tau) \end{pmatrix}^{-1} \begin{pmatrix} B(1) \\ B(\tau) - R(\tau)B(1) \end{pmatrix} \\
&= B(1)'B(1) + H(\tau)'M(\tau)^{-1}H(\tau)
\end{aligned}
$$

Since the first component is independent of τ, then by Continuous Mapping Theorem, we conclude that $W_\beta = \sup_{\tau\in[\tau_L,\tau_U]}(W_{\beta=0}+W_{bt}) \Rightarrow B(1)'B(1)+\sup_{\tau\in[\tau_L,\tau_U]} H(\tau)'M(\tau)^{-1}H(\tau)$.

(ii). In case that $\gamma_x < 1$, by Corollary 3.2, the second component in the limiting process would be a function of Brownian bridge $B(\tau) - \tau B(1)$. Since both $B(1)$ and $B(\tau) - \tau B(1)$ are Gaussian and we have $Cov(B(1), B(\tau) - \tau B(1)) = \tau - \tau = 0$, these two objects are independent of each other. Hence, we prove the statement. ■

Proof of Theorem 3.2. As discussed in the main text, under the null that there's no break in intercept, $\tilde{\alpha} - \alpha$ can be decomposed into following three parts, viz. $\bar{\epsilon}_T$, $(\tilde{\beta}^{IVX} - \beta)\bar{z}_{T-1}$ and $\beta(\bar{x}_{T-1} - \bar{z}_{T-1})$. We first show that the second term is asymptotically dominated by the first term. Note that we have

$$
\sqrt{T}(\tilde{\beta}^{IVX} - \beta)\bar{z}_{T-1} = \left[T^{\frac{1+\min(\gamma_x,\delta_z)}{2}}(\tilde{\beta}^{IVX} - \beta)\right] \cdot \left[T^{-(1+\frac{\min(\gamma_x,\delta_z)}{2})} \sum_{t=1}^{T} \tilde{z}_{t-1}\right].
$$

The term in the first bracket is $O_p(1)$ due to convergence property of IVX estimator established in Phillips and Magdalinos (2009). The order of the term in the second bracket is determined by the convergence rate of $\sum_{t=1}^{T} \tilde{z}_{t-1}$. Part (i) of Lemma A1 in the Online Appendix of KMS indicates that $\sum_{t=1}^{T} \tilde{z}_{t-1} = \begin{cases} O_p(T^{\frac{\min(\gamma_x,1)}{2}+\delta_z}) & \text{, if } \delta_z < \gamma_x \\ O_p(T^{\gamma_x+\frac{\delta_z}{2}}) & \text{, if } 0 < \gamma_x \leqslant \delta_z < 1 \end{cases}$ Hence, the term in the second bracket will be $O_p(T^{\frac{\min(\gamma_x,1)+\delta_z}{2}-1})$. As $\min(\gamma_x, 1)+\delta_z < 2$, this also suggests that it is $o_p(1)$. Therefore, we prove that the second component of $\tilde{\alpha}-\alpha$ is asymptotically dominated by the first part.

Now, we turn to show that the third component is also dominated by first term if β is small enough. First we rewrite it as $\sqrt{T}\beta(\bar{x}_{T-1}-\bar{z}_{T-1}) = \beta\frac{1}{\sqrt{n}}\sum_{t=1}^{T}(x_{t-1}-\tilde{z}_{t-1})$. Using the representation formula in equation (23) of Phillips and Magdalinos (2009), this term can be represented as $-\beta C_z T^{(\frac{1}{2}+\delta_z)}\sum_{t=1}^{n}\psi_{n,t-1}$, where $\psi_{nt} = \sum_{j=1}^{t} R_T^{t-j} x_{j-1}$. Hence, if we set $C_z = I_k$ for simplicity, we have

$$
\begin{aligned}
||\sqrt{T}\beta(\bar{x}_{T-1}-\bar{z}_{T-1})|| \leqslant & ||\beta|| T^{-(\frac{1}{2}+\delta_z)} \sum_{t=1}^{T} ||\psi_{T,t-1}|| \\
\leqslant & ||\beta|| T^{-(\frac{1}{2}+\delta_z)} n \sup_{2\leqslant t\leqslant t} ||\psi_{T,t-1}|| \\
= & ||\beta|| T^{\frac{1}{2}-\delta_z} \sup_{2\leqslant t\leqslant n} ||\psi_{T,t-1}|| \\
\leqslant & ||\beta|| T^{\frac{1}{2}-\delta_z} O_p(T^{\frac{\gamma_x}{2}+\delta_z}) \\
= & ||\beta|| O_p(T^{\frac{1+\gamma_x}{2}}),
\end{aligned}
$$

where the forth line is justified by the uniform bound of $||\psi_{T,t-1}||$, which is shown to be $O_p(T^{\frac{\gamma_x}{2}+\delta_z})$ in the Phillips and Magdalinos (2009). Thus, if $\beta = o_p(T^{-\frac{1+\gamma_x}{2}})$, then $||\sqrt{T}\beta(\bar{x}_{T-1}-\bar{z}_{T-1})|| = o_p(1)$ and the third component of $\tilde{\alpha}-\alpha$ will also be asymptotically dominated. Also note that if β is a nonzero constant, this third term will dominate.

Combining the above two results together, we obtain that $\sqrt{T}(\tilde{\alpha}-\alpha) = \frac{1}{\sqrt{T}}\sum_{t=1}^{T} u_{yt} + o_p(1)$. The required result then follows straightforwardly from the use of standard algebra and continuous mapping theorem. ■

Proof of Theorem 3.3. Given Theorem 3.1 and Theorem 3.2, the only step left is to show that W_{at} and W_{bt} are asymptotically independent of each other. Note that W_{at} is driven by $\sum_{t=1}^{[Ts]} u_{yt}$, while W_{bt} is determined by $\sum_{t=1}^{[Ts]} \tilde{z}_{t-1} u_{yt}$. These two partial sums will converge jointly to two independent Brownian motions, as shown in the Proposition A1 in Phillips and Magdalinos (2009). Hence the asymptotic independence is guaranteed. Convergence of $W_{\alpha\beta}$ then follows from continuous mapping theorem. ■

C Appendix to Chapter 4

This appendix includes the technical details and proof of the results reported in the Chapter 4. Before proceeding to the proof, it's convenience to summarize some notation we use in the following. Let α_t be the true value of α at time t and

$$\bar{\alpha} = \frac{1}{T}\sum_{i=1}^{m+1}(t_i - t_{i-1})\alpha_i = \sum_{i=1}^{m+1}(\tau_i - \tau_{i-1})\alpha_i$$

be the weighted average value of α. We will use Ω_x to denote the long-run variance of the predictor.

Proof of Theorem 4.1. The least square estimator for slope coefficient β can be expressed as

$$\begin{aligned}
\hat{\beta}^{ols} &= \left[\sum_{t=1}^{T}(x_{t-1}-\bar{x})^2\right]^{-1}\sum_{t=1}^{T}(x_{t-1}-\bar{x})(y_t-\bar{y}) \\
&= \left[\sum_{t=1}^{T}(x_{t-1}-\bar{x})^2\right]^{-1}\sum_{t=1}^{T}(x_{t-1}-\bar{x})\Big((\alpha_t+u_{yt})-(\bar{\alpha}+\bar{u})\Big) \\
&= \left[\sum_{t=1}^{T}(x_{t-1}-\bar{x})^2\right]^{-1}\sum_{t=1}^{T}(x_{t-1}-\bar{x})\,u_{yt} + \left[\sum_{t=1}^{T}(x_{t-1}-\bar{x})^2\right]^{-1}\sum_{t=1}^{T}(x_{t-1}-\bar{x})\,(\alpha_t-\bar{\alpha})
\end{aligned} \tag{C.1}$$

Since $\alpha_t = \sum_{i=1}^{m+1}\alpha_i I(t_{i-1} < t \leqslant t_i)$, the second component can be written as

$$\left[\sum_{t=1}^{T}(x_{t-1}-\bar{x})^2\right]^{-1}\sum_{i=1}^{m+1}\left[(\alpha_i-\bar{\alpha})\sum_{t=t_{i-1}}^{t_i}(x_{t-1}-\bar{x})\right]$$

and thus we have

$$\begin{aligned}
\sqrt{T}\hat{\beta}^{ols} &= \left[\frac{1}{T}\sum_{t=1}^{T}(x_{t-1}-\bar{x})^2\right]^{-1}\frac{1}{\sqrt{T}}\sum_{t=1}^{T}(x_{t-1}-\bar{x})\,u_{yt} \\
&\quad + \left[\frac{1}{T}\sum_{t=1}^{T}(x_{t-1}-\bar{x})^2\right]^{-1}\sum_{i=1}^{m+1}\left[(\alpha_i-\bar{\alpha})\frac{1}{\sqrt{T}}\sum_{t=t_{i-1}}^{t_i}(x_{t-1}-\bar{x})\right] \\
&\to^d N\left(0, \frac{\sigma_y^2 + \sum_{i=1}^{m+1}(\alpha_i-\bar{\alpha})^2(\tau_i-\tau_{i-1})}{E(x_{t-1}^2)}\right)
\end{aligned} \tag{C.2}$$

As for estimated variance of u_{yt}, we have following decompostion

$$\begin{aligned}
\hat{\sigma}^2 &= \frac{1}{T}\sum_{t=1}^{T}(y_t - \hat{\alpha} - x_{t-1}\hat{\beta})^2 \\
&= \frac{1}{T}\sum_{t=1}^{T}(\alpha_t + u_{yt} - \bar{y} + \bar{x}\hat{\beta} - x_{t-1}\hat{\beta})^2 \\
&= \frac{1}{T}\sum_{t=1}^{T}(\alpha_t - \bar{\alpha} + u_{yt} - \bar{u}_y + \bar{x}\hat{\beta} - x_{t-1}\hat{\beta})^2 \\
&= \frac{1}{T}\sum_{t=1}^{T}(\alpha_t - \bar{\alpha})^2 + \frac{1}{T}\sum_{t=1}^{T}(u_{yt} - \bar{u}_y)^2 + \frac{1}{T}\sum_{t=1}^{T}(x_{t-1} - \bar{x})^2\hat{\beta}^2 \\
&\quad \frac{2}{T}\sum_{t=1}^{T}(\alpha_t - \bar{\alpha})(u_{yt} - \bar{u}_y) + \frac{2}{T}\sum_{t=1}^{T}(\alpha_t - \bar{\alpha})(x_{t-1} - \bar{x})\hat{\beta} + \frac{2}{T}\sum_{t=1}^{T}(u_{yt} - \bar{u}_y)(x_{t-1} - \bar{x})\hat{\beta}
\end{aligned} \tag{C.3}$$

It's easy to see by simple application of law of large number that

$$\frac{1}{T}\sum_{t=1}^{T}(\alpha_t - \bar{\alpha})^2 = \sum_{i=1}^{m+1}(\alpha_i - \bar{\alpha})^2(\tau_i - \tau_{i-1})$$

$$\frac{1}{T}\sum_{t=1}^{T}(u_{yt} - \bar{u}_y)^2 \to^p \sigma_y^2$$

$$\frac{1}{T}\sum_{t=1}^{T}(x_{t-1} - \bar{x})^2\hat{\beta}^2 = \frac{1}{T}(\sqrt{T}\hat{\beta})^2\left[\frac{1}{T}\sum_{t=1}^{T}(x_{t-1} - \bar{x})^2\right] = o_p(1)$$

and all cross-prodcut terms are of smaller order. Therefore, we can conclude that

$$\hat{\sigma}^2 \to^p \sigma_y^2 + \sum_{i=1}^{m+1}(\alpha_i - \bar{\alpha})^2(\tau_i - \tau_{i-1}). \tag{C.4}$$

Combining (C.2) and (C.4), it's obvious that t-statistics converges to a standard normal random variable. ■

Proof of Theorem 4.2. It's well known that when the regressors is nonstationary, least square estimator has a faster rate of convergence. Similar to (C.2), we have following

expression with a different normalizing rate

$$
\begin{aligned}
T\hat{\beta}^{ols} = & \left[\frac{1}{T^2}\sum_{t=1}^{T}\left(x_{t-1}-\bar{x}\right)^2\right]^{-1}\frac{1}{T}\sum_{t=1}^{T}\left(x_{t-1}-\bar{x}\right)u_{yt} \\
& + \left[\frac{1}{T^2}\sum_{t=1}^{T}\left(x_{t-1}-\bar{x}\right)^2\right]^{-1}\sum_{i=1}^{m+1}\left[\sqrt{T}(\alpha_i-\bar{\alpha})\frac{1}{T\sqrt{T}}\sum_{t=t_{i-1}}^{t_i}\left(x_{t-1}-\bar{x}\right)\right]
\end{aligned}
\tag{C.5}
$$

Using the standard results in Phillips (1987), we have

$$
\frac{1}{T^2}\sum_{t=1}^{T}\left(x_{t-1}-\bar{x}\right)^2 \Rightarrow \int_0^1 \underline{J}_c(r)^2 dr
$$

$$
\frac{1}{T}\sum_{t=1}^{T}\left(x_{t-1}-\bar{x}\right)u_{yt} \Rightarrow \int_0^1 \underline{J}_c(r) dW_y(r)
$$

$$
\frac{1}{T\sqrt{T}}\sum_{t=t_{i-1}}^{t_i}\left(x_{t-1}-\bar{x}\right) \Rightarrow \int_{\tau_{i-1}}^{\tau_i} \underline{J}_c(r) dr
$$

where W_y is Brownian motion with standard deviation equal to σ_y. Therefore, by comparing the terms in (C.5), we find that the limiting distribution is determined by the asymptotic behavior of $\sqrt{T}(\alpha_i-\bar{\alpha})$. Note that, by simple algebraic manipulation, we can deduce

$$
\begin{aligned}
\alpha_i-\bar{\alpha} &= \alpha_i - \sum_{j=1}^{m+1}\alpha_j(\tau_j-\tau_{j-1}) \\
&= \alpha_i - [\alpha_1\tau_1 + \alpha_2(\tau_2-\tau_1) + ... + \alpha_{m+1}(1-\tau_m)] \\
&= \alpha_i - [(\alpha_1-\alpha_2)\tau_1 + (\alpha_2-\alpha_3)\tau_2 + ... + (\alpha_{m-1}-\alpha_m)\tau_m + \alpha_{m+1}] \\
&= (\alpha_2-\alpha_1)\tau_1 + (\alpha_3-\alpha_2)\tau_2 + ... + (\alpha_{m+1}-\alpha_m)\tau_m - (\alpha_{m+1}-\alpha_i)
\end{aligned}
\tag{C.6}
$$

(i). When all breaks are small, for $i = 1, 2, ..., m+1$, $\sqrt{T}(\alpha_i-\bar{\alpha})$ are $o(1)$. Therefore, the second compnent in (C.5) will be dominated. It's also obvious in this case that $\hat{\sigma}^2 \to^p \sigma_y^2$. This implies that the breaks are asympotically neglible and thus the limiting distribution

of t-statistics is the same as in the no-break case, which is

$$t_{\beta}^{ols} \Rightarrow \frac{\int_0^1 \underline{J}_c(r)dW(r)}{\sqrt{\int_0^1 \underline{J}_c(r)^2 dr}}$$

(ii). For moderate breaks case, we have $\sqrt{T}(\alpha_i - \bar{\alpha}) = \sum_{i=1}^{m} \tau_i d_i - \sum_{j=i}^{m} d_j$. Combining all these limiting components together, we have that

$$\begin{aligned} T\hat{\beta}^{ols} &\Rightarrow \frac{\int_0^1 \underline{J}_c(r)dW_y(r)}{\int_0^1 \underline{J}_c(r)^2 dr} + \frac{\sum_{i=1}^{m+1}\left[\sum_{i=1}^{m}\tau_i d_i - \sum_{j=i}^{m} d_j\right]\int_{\tau_{i-1}}^{\tau_i}\underline{J}_c(r)dr}{\int_0^1 \underline{J}_c(r)^2 dr} \\ &= \frac{\int_0^1 \underline{J}_c(r)dW_y(r)}{\int_0^1 \underline{J}_c(r)^2 dr} + \frac{\left(\sum_{i=1}^{m}\tau_i d_i\right)\int_0^1 J_c(r)dr - \sum_{i=1}^{m+1}\left[(\sum_{j=i}^{m} d_j)\int_{\tau_{i-1}}^{\tau_i}\underline{J}_c(r)dr\right]}{\int_0^1 \underline{J}_c(r)^2 dr} \end{aligned} \tag{C.7}$$

Now, we derive the limiting behavior of $\hat{\sigma}^2$. To this end, we can still rely on (C.3). The only difference is that now we have

$$\frac{1}{T}\sum_{t=1}^{T}(\alpha_t - \bar{\alpha})^2 = \frac{1}{T}\sum_{i=1}^{m+1}\left[\sqrt{T}(\alpha_i - \bar{\alpha})\right]^2(\tau_i - \tau_{i-1}) = o(1)$$

and thus $\hat{\sigma}^2 \to^p \sigma_y^2$. Therefore, we have that

$$\begin{aligned} t_{\beta}^{ols} &= \frac{T\hat{\beta}^{ols}}{\sqrt{\hat{\sigma}^2\left[\frac{1}{T^2}\sum_{t=1}^{T}(x_{t-1} - \bar{x})^2\right]^{-1}}} \\ &\Rightarrow \frac{\int_0^1 \underline{J}_c(r)dW(r)}{\sqrt{\int_0^1 \underline{J}_c(r)^2 dr}} + \frac{\left(\sum_{i=1}^{m}\tau_i d_i\right)\int_0^1 \underline{J}_c(r)dr - \sum_{i=1}^{m+1}\left[(\sum_{j=i}^{m} d_j)\int_{\tau_{i-1}}^{\tau_i}\underline{J}_c(r)dr\right]}{\sigma_y\sqrt{\int_0^1 \underline{J}_c(r)^2 dr}} \end{aligned} \tag{C.8}$$

(iii). If any of the breaks is large and $\sqrt{T}(\alpha_i - \bar{\alpha}) \to \infty$ as sample size increases, then the second term in (C.5) will dominate the limting behavior. More importantly, the convergence rate of $\hat{\beta}^{ols}$ will be slower. At the meantime, the convergence rate of the terms in denominator will remain the same. Therefore, t-statistics will explode to infinity in this

case. ■

Proof of Theorem 4.3. Similar to the proof of Theorme 4.2, we re-normalize the least square estimator with a different rate to accommadate the limiting behavior of mildly stationary process.

$$T^{\frac{1+\gamma_x}{2}}\hat{\beta}^{ols} = \left[\frac{1}{T^{1+\gamma_x}}\sum_{t=1}^{T}(x_{t-1}-\bar{x})^2\right]^{-1}\frac{1}{T^{\frac{1+\gamma_x}{2}}}\sum_{t=1}^{T}(x_{t-1}-\bar{x})\,u_{yt} + \left[\frac{1}{T^{1+\gamma_x}}\sum_{t=1}^{T}(x_{t-1}-\bar{x})^2\right]^{-1}\sum_{i=1}^{m+1}\left[T^{\frac{\gamma_x}{2}}(\alpha_i-\bar{\alpha})\frac{1}{T^{\frac{1}{2}+\gamma_x}}\sum_{t=t_{i-1}}^{t_i}(x_{t-1}-\bar{x})\right]. \tag{C.9}$$

As shown in Phillips and Magdalinos (2005) and Giraitis and Phillips (2006), when x_{t-1} is mildly stationary, we have the following asymptotic results

$$\frac{1}{T^{1+\gamma_x}}\sum_{t=1}^{T}(x_{t-1}-\bar{x})^2 \to^p \Omega_x\int_0^\infty e^{2r}dr = \frac{1}{2}\Omega_x,$$

$$\frac{1}{T^{\frac{1+\gamma_x}{2}}}\sum_{t=1}^{T}(x_{t-1}-\bar{x})\,u_{yt} \Rightarrow U(1),$$

$$\frac{1}{T^{\frac{1}{2}+\gamma_x}}\sum_{t=t_{i-1}}^{t_i}(x_{t-1}-\bar{x}) \Rightarrow B(\tau_i)-B(\tau_{i-1}),$$

where $U(\cdot)$ and $B(\cdot)$ are independent Brownian motions with variance equal to $\frac{1}{2}\sigma_y^2\Omega_x$ and Ω_x, respectively. Therefore, the limiting behavior of t-statistics is captured by $T^{\frac{\gamma_x}{2}}(\alpha_i-\bar{\alpha})$. It's also easy to show that we still have $\hat{\sigma}_y^2 \to^p \sigma_y^2$. Hence, when all the breaks are

moderate, we can deduce that

$$
\begin{aligned}
t_{\beta}^{ols} =& \frac{\hat{\beta}^{ols}}{\hat{\sigma}_y \sqrt{\left[\sum_{t=1}^{T}(x_{t-1}-\bar{x})^2\right]^{-1}}} \\
=& \frac{T^{\frac{1+\gamma_x}{2}}\hat{\beta}^{ols}}{\hat{\sigma}_y \sqrt{\left[\frac{1}{T^{1+\gamma_x}}\sum_{t=1}^{T}(x_{t-1}-\bar{x})^2\right]^{-1}}} \\
\Rightarrow& \frac{(\frac{1}{2}\Omega_x)^{-1}U(1)}{\sigma_y(\frac{1}{2}\Omega_x)^{-1/2}} + \frac{(\frac{1}{2}\Omega_x)^{-1}\sum_{i=1}^{m+1}(\sum_{i=1}^{m}\tau_i d_i - \sum_{j=i}^{m} d_j)(B(\tau_i)-B(\tau_{i-1}))}{\sigma_y(\frac{1}{2}\Omega_x)^{-1/2}} \\
=& N\left(0, 1+\frac{1}{\sigma_y^2}\sum_{i=1}^{m+1}\left(\sum_{i=1}^{m}\tau_i d_i - \sum_{j=i}^{m} d_j\right)^2(\tau_i-\tau_{i-1})\right)
\end{aligned}
\tag{C.10}
$$

When all the breaks are small, the second component of (C.9) is dominated and thus we have $t_{\beta}^{ols} \Rightarrow N(0,1)$. Oppositely, if any break is large, the second component will dominate, which results in a slower convergence rate of $\hat{\beta}^{ols}$ and consequently an explosive t-statistics. ■

Proof of Theorem 4.4. We focus on the proof of part (ii), i.e. moderate break case, as other two cases can seen as trivial implication of it. In the following, we assume that the predictor is local to unit root. The proof under unit-root regressor is similar and thus omitted. The IVX estimator of slope coefficient β can be written as

$$
\begin{aligned}
\tilde{\beta}^{ivx} &= \left[\sum_{t=1}^{T}(x_{t-1}-\bar{x})\tilde{z}_{t-1}\right]^{-1}\sum_{t=1}^{T}\tilde{z}_{t-1}(y_t-\bar{y}) \\
&= \left[\sum_{t=1}^{T}(x_{t-1}-\bar{x})\tilde{z}_{t-1}\right]^{-1}\sum_{t=1}^{T}\tilde{z}_{t-1}\left[(\alpha_t-\bar{\alpha})+(u_{yt}-\bar{u}_y)\right] \\
&= \left[\sum_{t=1}^{T}(x_{t-1}-\bar{x})\tilde{z}_{t-1}\right]^{-1}\left[\sum_{t=1}^{T}\tilde{z}_{t-1}(u_{yt}-\bar{u}_y) + \sum_{i=1}^{m+1}(\alpha_i-\bar{\alpha})\sum_{t=t_{i-1}}^{t_i}\tilde{z}_{t-1}\right]
\end{aligned}
\tag{C.11}
$$

which can be further normalized to obtain that

$$
\begin{aligned}
T^{\frac{1+\delta_z}{2}}\tilde{\beta}^{ivx} = & \left[\frac{1}{T^{1+\delta_z}}\sum_{t=1}^{T}(x_{t-1}-\bar{x})\tilde{z}_{t-1}\right]^{-1}\left[\frac{1}{T^{\frac{1+\delta_z}{2}}}\sum_{t=1}^{T}\tilde{z}_{t-1}(u_{yt}-\bar{u}_y)\right] \\
& + \left[\frac{1}{T^{1+\delta_z}}\sum_{t=1}^{T}(x_{t-1}-\bar{x})\tilde{z}_{t-1}\right]^{-1} \\
& \left[\sum_{i=1}^{m+1}\left(T^{\frac{\min[1,\gamma_x]}{2}+\delta_z-\frac{1+\delta_z}{2}}(\alpha_i-\bar{\alpha})\right)\left(\frac{1}{T^{\frac{\min[1,\gamma_x]}{2}+\delta_z}}\sum_{t=t_{i-1}}^{t_i}\tilde{z}_{t-1}\right)\right]
\end{aligned}
\tag{C.12}
$$

Since the standard deviation of IVX estimator is estimated by

$$
\hat{\sigma}_y\left(\sum_{t=1}^{T}\tilde{z}_{t-1}^2\right)^{1/2}\left(\sum_{t=1}^{T}\tilde{z}_{t-1}(x_{t-1}-\bar{x})\right)^{-1},
$$

the IVX t-statistics can be expressed as

$$
\begin{aligned}
t_\beta^{ivx} &= \frac{T^{\frac{1+\delta_z}{2}}\tilde{\beta}^{ivx}}{\hat{\sigma}_y\left(\frac{1}{T^{1+\delta_z}}\sum_{t=1}^{T}\tilde{z}_{t-1}^2\right)^{1/2}\left(\frac{1}{T^{1+\delta_z}}\sum_{t=1}^{T}\tilde{z}_{t-1}(x_{t-1}-\bar{x})\right)^{-1}} \\
&= \frac{\frac{1}{T^{\frac{1+\delta_z}{2}}}\sum_{t=1}^{T}\tilde{z}_{t-1}(u_{yt}-\bar{u}_y)}{\hat{\sigma}_y\left(\frac{1}{T^{1+\delta_z}}\sum_{t=1}^{T}\tilde{z}_{t-1}^2\right)^{1/2}} + \frac{\sum_{i=1}^{m+1}\left(T^{\frac{\delta_z+\min[0,\gamma_x-1]}{2}}(\alpha_i-\bar{\alpha})\right)\left(\frac{1}{T^{\frac{\min[1,\gamma_x]}{2}+\delta_z}}\sum_{t=t_{i-1}}^{t_i}\tilde{z}_{t-1}\right)}{\hat{\sigma}_y\left(\frac{1}{T^{1+\delta_z}}\sum_{t=1}^{T}\tilde{z}_{t-1}^2\right)^{1/2}}
\end{aligned}
\tag{C.13}
$$

As proved in Phillips and Magdalinos (2009) and Kostakis et al. (2015), following limiting theory applies

$$
\frac{1}{T^{1+\delta_z}}\sum_{t=1}^{T}\tilde{z}_{t-1}^2 \to^p \Omega_x\int_0^\infty e^{2r}dr = \frac{1}{2}\Omega_x
$$

$$
\frac{1}{T^{\frac{1+\delta_z}{2}}}\sum_{t=1}^{T}\tilde{z}_{t-1}(u_{yt}-\bar{u}_y) \Rightarrow N(0,\frac{1}{2}\sigma_y^2\Omega_x)
$$

$$
\frac{1}{T^{\frac{\min[1,\gamma_x]}{2}+\delta_z}}\sum_{t=t_{i-1}}^{t_i}\tilde{z}_{t-1} = \frac{1}{\sqrt{T}}x_T + o_p(1) \Rightarrow \Omega_x^{1/2}\Big(J_c(\tau_i)-J_c(\tau_{i-1})\Big)
$$

Combining these results with (C.11), we can deduce that

$$t_\beta^{ivx} \Rightarrow N(0,1)+ \\ \frac{\left(\sum_{i=1}^{m} \tau_i d_i\right) J_c(1) - \sum_{i=1}^{m+1} \left(\sum_{j=i}^{m} d_j\right) \left(J_c(\tau_i) - J_c(\tau_{i-1})\right)}{\sigma_y/\sqrt{2}} \tag{C.14}$$

This is exactly the result that we want to prove. ■

Proof of Theorem 4.5. The proof is very similar as in the proof of Theorem 4.4. The only difference to note is that in this case, we have $\sum_{t=1}^{T} \tilde{z}_{t-1} = O_p(T^{\gamma_x + \frac{\delta_z}{2}})$, as is proved in Online Appendix of Kostakis et al. (2015). Meanwhile, we have

$$\frac{1}{T^{1+\gamma_x}} \sum_{t=1}^{T} \tilde{z}_{t-1}^2 \to^p \Omega_x \int_0^\infty e^{2r} dr = \frac{1}{2}\Omega_x$$

$$\frac{1}{T^{\frac{1+\gamma_x}{2}}} \sum_{t=1}^{T} \tilde{z}_{t-1}(u_{yt} - \bar{u}_y) \Rightarrow N(0, \frac{1}{2}\sigma_y^2 \Omega_x)$$

Hence, by normalizing the $\tilde{\beta}^{ivx}$ using $T^{\frac{1+\gamma_x}{2}}$, it is straightforward to obtain the results. ■

Proof of Theorem 4.7. Under our assumption, the model under null hythothesis is a special case of the model considered in Bai and Perron (1998). Thus, we have that $\hat{m}$ is consistent and $\hat{\tau}_i$ converges to τ_i with rate T. Meanwhile, we know that the convergence rate of $\tilde{\beta}^{ivx}$ is $T^{\frac{\min[\delta_z, \gamma_x]+1}{2}}$, which must be slower than T. Therefore, by similar argument as in Bai (1997) and Bai and Perron (1998), we can easily see that the estimation error induced by break detection is of smaller order and thereby asymptotically negligible. Consequently, the IVX-Wald statistics in each regime converges to a $\chi^2(k)$ random variable, which implies that their summation goes to $\chi^2(k(m+1))$ ■

D Appendix to Chapter 5

D.1 Details of PGAS algorithm

Recall that a state space model is in the form

$$y_t|x_t = x \sim f_\theta(\cdot|x), \tag{D.1}$$

$$x_{t+1}|x_t = x \sim g_\theta(\cdot|x), \text{ and } x_1 \sim \mu_\theta(\cdot), \tag{D.2}$$

The output of a PGAS algorithm is a random draw from the joint smoothing distribution $p_\theta(x_{1:T}|y_{1:T})$, conditional on one particular set of parameter values. In the following, we will omit parameters in all densities with an understanding that they depend on a same θ. The input of this algorithm, except for θ, is a reference trajectory of $x_{1:T}$, which is a sample from last MCMC iteration. Let's denote that reference trajectory by $x'_{1:T}$. Then, the algorithm proceeds as following:

- 1. Draw $x_1^{(i)}$ from $q_1(x_1|y_1)$, for $i = 1, 2, ..., N-1$.
- 2. Set $x_1^{(N)} = x'_1$.
- 3. Set $w_1^{(i)} = f(y_1|x_1^{(i)})/q_1(x_1^{(i)}|y_1)$, for $i = 1, 2, ..., N$.
- 4. For $t = 2$ to T, do the following:
  - (a). Generate $\{\tilde{x}_{1:t-1}^{(i)}\}_{i=1}^{N-1}$ by sampling with replacement $N-1$ times from $\{x_{1:t-1}^{(i)}\}_{i=1}^{N}$ with probabilities proportional to the importance weights $\{w_{t-1}^{(i)}\}_{i=1}^{N}$.
  - (b). Draw J from $\{1, 2, ..., N\}$ with probabilities proportional to $w_{t-1}^{(i)} g(x'_t|x_{t-1}^{(i)})$ and then set $\tilde{x}_{1:t-1}^{(N)} = x_{1:t-1}^{(J)}$.
  - (c). Simulate $x_t^{(i)}$ from $q_t(x_t|\tilde{x}_{t-1}^{(i)}, y_t)$, for $i = 1, 2, ..., N-1$.
  - (d). Set $x_t^{(N)} = x'_t$.
  - (e). Set $x_{i:t}^{(i)} = (\tilde{x}_{1:t-1}^{(i)}, x_t^{(i)})$
  - (f). Set weight to be $w_t^{(i)} = f(y_t|x_t^{(i)}) g(x_t^{(i)}|\tilde{x}_{t-1}^{(i)})/q_t(x_t^{(i)}|\tilde{x}_{t-1}^{(i)}, y_t)$, for $i = 1, 2, ..., N$.

- 5. Draw k from $\{1, 2, ..., N\}$ with probabilities proportional to $w_T^{(i)}$ and return $x_{1:T}^* = x_{1:T}^{(k)}$.

Note that this procedure is very similar to the original conditional sequential Monte Carlo-based particle Gibbs sampler. The major modification is in the Step 4(b), where a new index is drawn and thus the N^{th} trajectory may not be the reference one from last iteration. In conditional PG, on the contrary, we fix the last particle to follow the input trejectory $x'_{1:T}$.

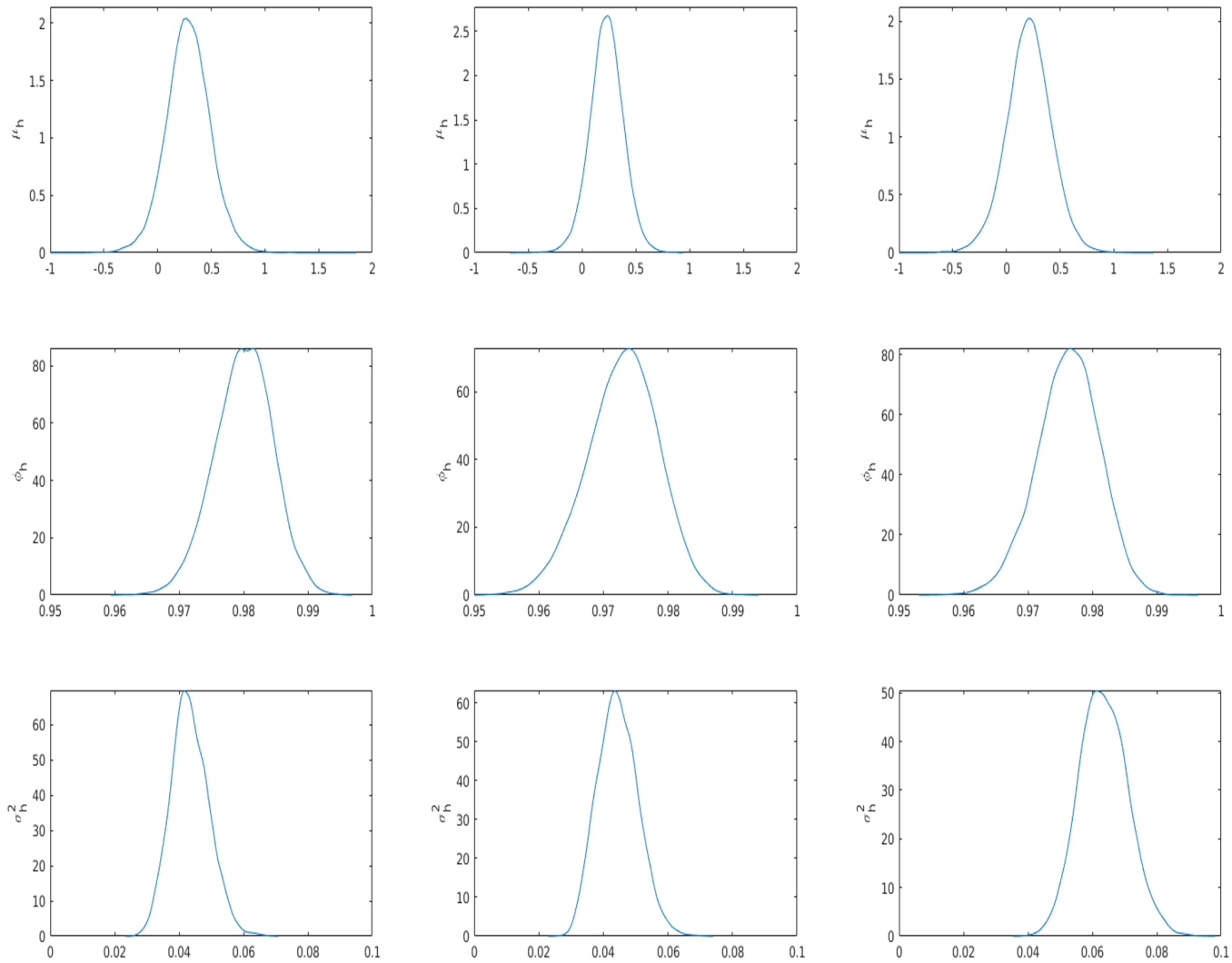
It is worth mentioning that the probability of drawing J depends on $g(x'_t|x_{t-1}^{(i)})$ and x'_t is drawn in the last iteration conditional on all observations $y_{1:T}$. Therefore, this step makes the algorithm more like a smoothing instead of filtering.

D.2 Additional Figures

This section includes some additional figures that are related to the empirical application with weekly exchange rate data we consider in the main text. Corresponding figures for stock return data are similar and thus omitted for saving the space. Specifically, we present the following figures:

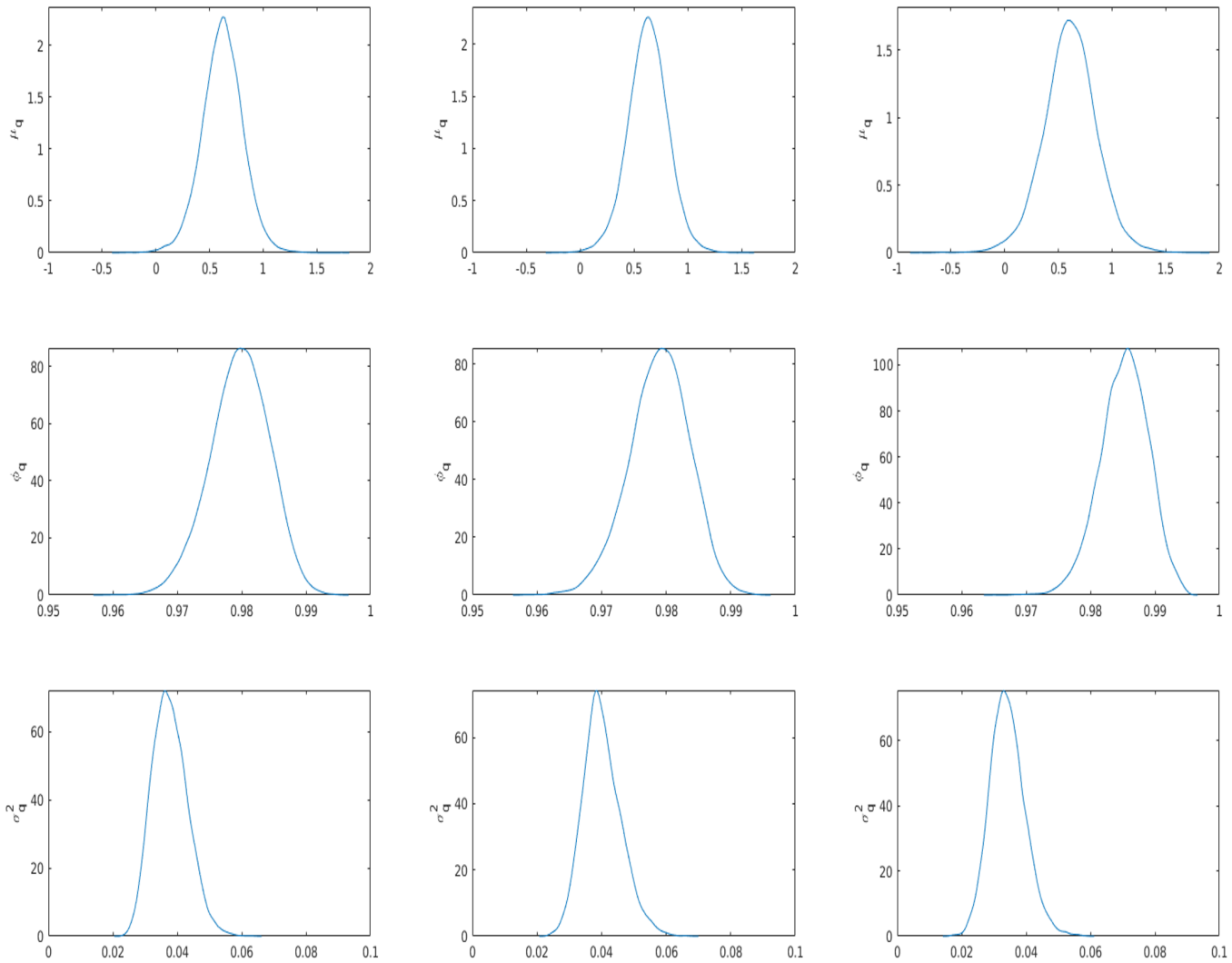
- Figure 29: Kernel-smoothed posterior distributions of volatility-related parameters.
- Figure 30: Kernel-smoothed posterior distributions of correlation-related parameters.
- Figure 31: Mixing of volatility-related parameters represented by the decay of autocorrelation.
- Figure 32: Mixing of correlation-related parameters represented by the decay of autocorrelation.

Figure 29: Posterior distributions of volatility-related parameters for weekly exchange rate data



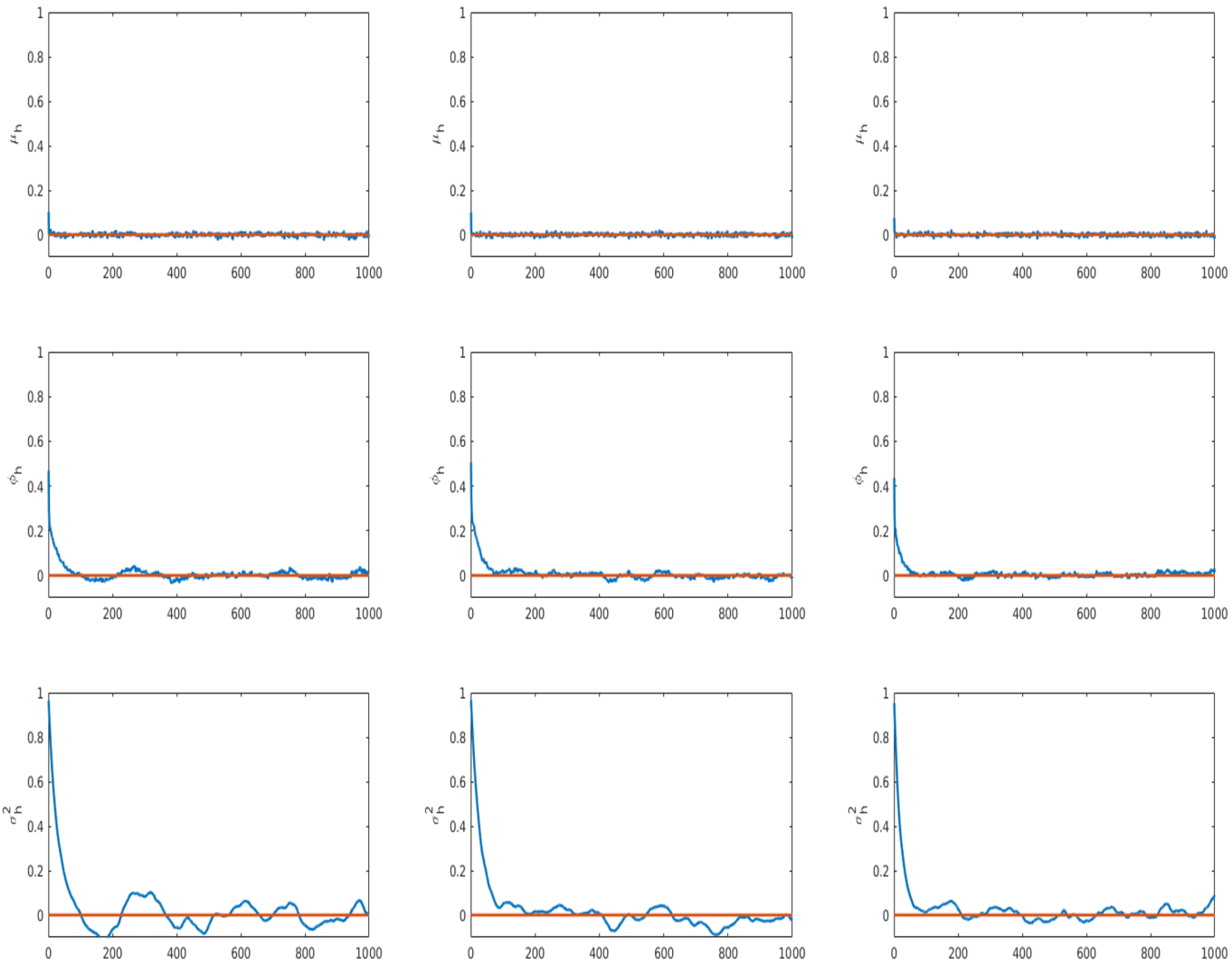
Note: The first row includes μ's, the second row includes ϕ's and the third row includes σ^2's. All densities are kernel-smoothed.

Figure 30: Posterior distributions of correlation-related parameters for weekly exchange rate data



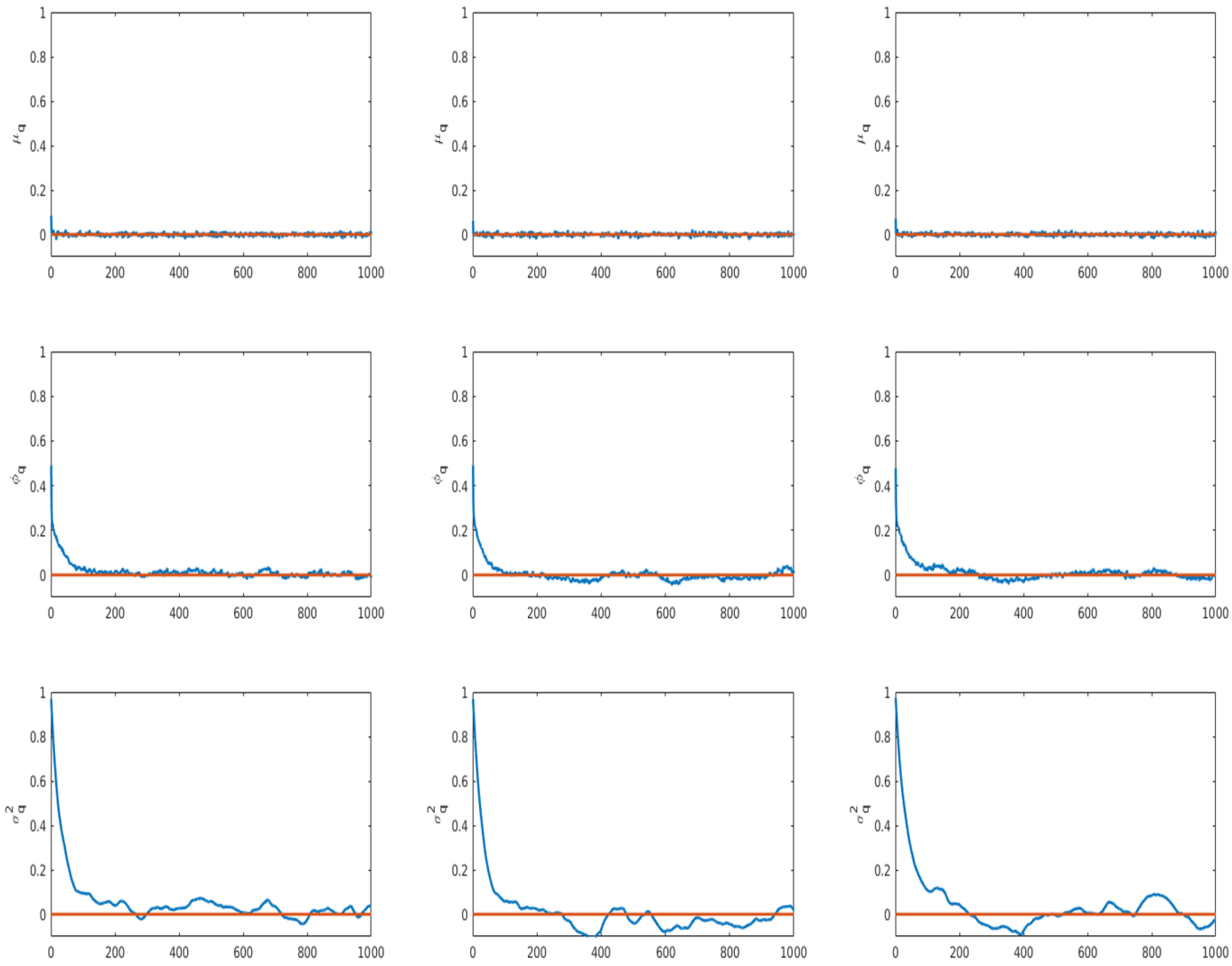
Note: The first row includes μ's, the second row includes ϕ's and the third row includes σ^2's. All densities are kernel-smoothed.

Figure 31: Mixing of volatility-related parameters for weekly exchange rate data



Note: The first row includes μ's, the second row includes ϕ's and the third row includes σ^2's. Blue curves show the decay of sample autocorrelations of MCMC sampler.

Figure 32: Mixing of correlation-related parameters for weekly exchange rate data



Note: The first row includes μ's, the second row includes ϕ's and the third row includes σ^2's. Blue curves show the decay of sample autocorrelations of MCMC sampler.