Statistica Sinica **23** (2013), 1573-1594
doi:http://dx.doi.org/10.5705/ss.2012.224s

STOCHASTIC CHANGE-POINT ARX-GARCH MODELS AND THEIR APPLICATIONS TO ECONOMETRIC TIME SERIES

Tze Leung Lai and Haipeng Xing

Stanford University and SUNY at Stony Brook

Abstract: This paper shows that the commonly encountered volatility persistence in fitting GARCH models to financial time series can arise if the possibility of structural changes is not incorporated in the time series model. To avoid spurious long memory in modeling volatilities of econometric time series, we consider two time-scales and use the "short" time-scale to define GARCH dynamics and the "long" time-scale to incorporate parameter jumps. This leads to a Bayesian change-point ARX-GARCH model, whose unknown parameters can undergo occasional changes at unspecified times and can be estimated by explicit recursive formulas when the hyperparameters of the Bayesian model are specified. Efficient estimators of the hyperparameters of the Bayesian model are developed, yielding empirical Bayes estimates of the piecewise constant parameters in the stochastic change-point model. The empirical Bayes approach is applied to the frequentist problem of partitioning the time series into segments under sparsity assumptions on the change-points. Simulation and empirical studies of its performance are also given.

Key words and phrases: ARX-GARCH models, empirical Bayes, long memory, multiple change-points, recursive adaptive filters, segmentation.

1. Introduction

Volatility modeling is a cornerstone of empirical finance, as portfolio theory, asset pricing, and hedging all involve volatilities. Since the seminal works of Engle (1982) and Bollerslev (1986), generalized autoregressive conditionally heteroskedastic (GARCH) models have been widely used to model and forecast volatilities of financial time series. In many empirical studies of stock returns and exchange rates, estimation of the parameters ν, a, and b in the GARCH(1,1) model

$$y_t = \sigma_t \epsilon_t, \quad \sigma_t^2 = (1-a-b)\nu^2 + a y_{t-1}^2 + b\sigma_{t-1}^2 \tag{1.1}$$

reveals high volatility persistence, with the maximum likelihood estimate of $a+b$ close to 1. To model such persistence, Engle and Bollerslev (1986) considered the "integrated" GARCH (IGARCH) models, and Baillie, Bollerslev, and Mikkelsen (1996) introduced fractional integration in their FIGARCH models, with a slow

hyperbolic rate of decay for the influence of the past innovations, to quantify the long memory of exchange rate volatilities. However, it has been pointed out that if the model parameters undergo occasional changes, then the fitted models that assume time-invariant parameters tend to exhibit long memory; see Diebold (1986), Perron (1989), Lamoureux and Lastrapes (1990), Mikosch and Starica (2004), and Hillebrand (2005).

The frequentist approach to incorporating change-points in regression models without GARCH-type dynamics for the error variances assumes that the change-points are unknown parameters to be estimated from the data; see Quandt (1958, 1960), Andrews, Lee, and Ploberger (1996), Bai, Lumsdaine, and Stock (1998), Bai and Perron (1998), Davis, Lee, and Rodriguez-Yam (2006), Qu and Perron (2007), Spokoiny (2009), Galeano and Tsay (2010), Aue and Lee (2011), and the references therein. An alternative approach is Bayesian and assumes that the change-points and the associated regimes are generated by some stochastic process so that the unknown regression parameters can be estimated from their posterior distribution via Bayes' theorem; see Goldfeld and Quandt (1973) and Hamilton (1989). Albert and Chib (1993) consider ARX models (autoregressive models with exogenous inputs) whose autoregressive parameters and error variances are subject to regime changes determined by a two-state Markov chain with unknown transition probabilities. McCulloch and Tsay (1993) consider AR models with random shifts in mean and error variance. Wang and Zivot (2000) assume the number of change-points to be known for random changes in level, trend, and variance of a time series. Hamilton and Susmel (1994) consider regime switching in ARCH models, and So, Lam, and Li (1998) consider regime-switching stochastic volatility (SV) models. These are hidden Markov-models (HMMs). Markov chain Monte Carlo (MCMC) methods, in particular Gibbs samplers, are used to approximate the optimal smoothers in these HMMs.

In this paper we introduce a new class of change-point ARX-GARCH models for which there are explicit recursive filters and smoothers, thereby obviating the reliance on MCMC methods whose convergence properties and performance in change-point time series models have not been systematically studied because of their computational complexity. This decoupling of long-run and short-term volatilities in the GARCH model, so that jumps can be included in the long-run volatility, is what makes the proposed HMM much easier to handle than previous ones. The optimal Bayes estimates in our change-point ARX-GARCH model involve unspecified hyperparameters, which can in principle be estimated by the EM algorithm. For regime-switching ARCH models, Cai (1994) noted the "tremendous complication" of the normal equations of the EM algorithm, making it "extremely difficult" to implement for sample sizes exceeding 50. By using recursive representations of the summands of the log-likelihood function,

we have a relatively simple algorithm to evaluate the log-likelihood function and estimate the hyperparameters of the change-point model. In Section 2 we introduce the change-point ARX-GARCH model and describe the associated filters for estimating the piecewise constant ARX parameter $\boldsymbol{\beta}_n$ and long-run volatility ν_n based on the observations $y_1, \ldots, y_n$, assuming known hyperparameters that include the GARCH parameters associated with short-term volatility changes. We also develop bounded complexity mixture (BCMIX) approximations for efficient computation of these filters and of the likelihood function to estimate the hyperparameters. We then use these closed-form recursions to develop BCMIX approximations to the Bayes estimates (smoothers) of $\boldsymbol{\beta}_t$ and ν_t for $1 \leq t \leq n$. In Section 3 we prove volatility persistence of maximum likelihood estimates that ignore change-points in the stochastic change-point ARX-GARCH model. Previous works in this direction, which are reviewed in Section 3, have not considered maximum likelihood estimates that are commonly used to fit GARCH models.

In contrast to our empirical Bayes (EB) approach that assumes a relatively simple stochastic model for change-points, the frequentist approach, often called "segmentation", assumes the change-points and the pre- and post-change regression coefficients in regression models to be unknown parameters, and uses maximum likelihood to estimate them and a model selection criterion to determine the number of change-ponts. In Section 4, following our previous work in Lai and Xing (2011) on multiple parameter changes in multiparameter exponential families, we show how the relative simplicity of the EB smoothers can be used to resolve difficulties in the frequentist segmentation problem. Asymptotic theory and simulation studies of the proposed segmentation procedure are presented. In Section 5 we apply the change-point AR(1)-GARCH(1,1) model to an empirical analysis of the weekly returns of the SP500 index from 1950 to 2009. We also evaluate BCMIX predictors and use them for forecasting 1-week ahead returns to calculate VaR (value at risk). Section 6 gives further discussion and concluding remarks.

2. A Stochastic Change-point ARX-GARCH Model

To incorporate structural changes in the regression coefficients and the unconditional variance of the random disturbances in an autoregressive model with exogenous inputs (ARX), while allowing the conditional variances to follow a GARCH model, we consider the change-point ARX-GARCH model

$$y_t = \boldsymbol{\beta}_t^T \mathbf{x}_t + \nu_t \sqrt{h_t} \epsilon_t, \tag{2.1}$$

in which the parameter vector $\boldsymbol{\beta}_t$ and the unconditional variance ν_t^2 are piecewise constant, with jumps at times of structural change, the vector $\mathbf{x}_t$ consists of exogenous variables and the past observations $y_{t-1}, y_{t-2}, \ldots, y_{t-\kappa}$, and we use h_t

to represent short-term proportional fluctuations in variance generated by the GARCH model

$$h_t = (1 - \sum_{i=1}^{k} a_i - \sum_{l=1}^{k'} b_l) + \sum_{i=1}^{k} a_i w_{t-i}^2 + \sum_{l=1}^{k'} b_l h_{t-l}, \quad \text{with } w_s = \sqrt{h_s}\epsilon_s. \quad (2.2)$$

The ϵ_t are assumed to be i.i.d. standard normal random variables such that ϵ_t are independent of $\mathbf{x}_t$, and the time-invariant GARCH parameters $a_1, \ldots, a_k$, $b_1, \ldots, b_{k'}$ are assumed to satisfy $a_i \geq 0$, $b_l \geq 0$ and $\sum_{i=1}^{k} a_i + \sum_{l=1}^{k'} b_l \leq 1$. Letting $\tau_t = 1/(2\nu_t^2)$, we assume $\boldsymbol{\theta}_t = (\boldsymbol{\beta}_t^T, \tau_t)^T$ to be piecewise constant and satisfy the following conditions.

(A1) For $t > t_0 = \max(k, k')$, the change-times of $\boldsymbol{\theta}_t$ form a renewal process with i.i.d. inter-arrival times that are geometrically distributed with parameter p or, equivalently,

$$I_t := 1_{\{\boldsymbol{\theta}_t \neq \boldsymbol{\theta}_{t-1}\}} \text{ are i.i.d. Bernoulli}(p) \text{ with } P(I_t = 1) = p,$$

$I_{t_0} = 1$, and there is no change-point prior to t_0.

(A2) $\boldsymbol{\theta}_t = (1 - I_t)\boldsymbol{\theta}_{t-1} + I_t(\mathbf{z}_t^T, \gamma_t)^T$, where $(\mathbf{z}_1^T, \gamma_1)^T, (\mathbf{z}_2^T, \gamma_2)^T, \ldots$ are i.i.d. random vectors such that $\mathbf{z}_t|\gamma_t \sim \text{Normal}(\mathbf{z}, \mathbf{V}/(2\gamma_t))$, $\gamma_t \sim \chi_d^2/\rho$, with χ_d^2 the chi-square distribution with d degrees of freedom.

(A3) The processes $\{I_t\}$, $\{(\mathbf{z}_t^T, \gamma_t)\}$, and $\{(\mathbf{x}_t, \epsilon_t)\}$ are independent.

2.1. Closed-form recursive filters

Conditions (A1)−(A3) specify a Markov chain with unobserved states $(I_t, \boldsymbol{\theta}_t)$. The observations $(\mathbf{x}_t, y_t)$ are such that $(y_t - \boldsymbol{\beta}_t^T \mathbf{x}_t)/\nu_t$ forms a GARCH process. This hidden Markov model (HMM) has hyperparameters $p, \mathbf{z}, \mathbf{V}, \rho, d, a_1, \ldots, a_k$, $b_1, \ldots, b_{k'}$. To estimate $\boldsymbol{\theta}_t$ assuming known hyperparameters, we extend the method of Lai, Liu, and Xing (2005) for the special case $\mathbf{x}_t = (y_{t-1}, \ldots, y_{t-k})^T$ and $h_t \equiv 1$, in which the error variances σ_n^2 have jumps but do not undergo GARCH dynamics. Let $J_n = \max\{t \leq n : I_t = 1\}$ and note that $n - J_n \geq k$ by (A1). Define $\mathcal{Y}_n = (\mathbf{x}_1, y_1, \ldots, \mathbf{x}_n, y_n)$ and $\mathcal{Y}_{j,n} = (\mathbf{x}_j, y_j, \ldots, \mathbf{x}_n, y_n)$. The estimates $\widehat{\boldsymbol{\beta}}_n$ and $\widehat{\nu}_n^2$ based on $\mathcal{Y}_n$ are weighted averages of $\widehat{\boldsymbol{\beta}}_{n,j}$ and $\widehat{\nu}_{n,j}^2$ based on $\mathcal{Y}_{j,n}$, with the weights $p_{n,j}$ to be specified. The $\widehat{\boldsymbol{\beta}}_{n,j}$ and $\widehat{\nu}_{n,j}^2$ can be computed recursively (with increasing n and fixed j). Initializing at $n = j-1$ with $\widehat{\boldsymbol{\beta}}_{n,j} = \mathbf{z}$, $\widehat{\mathbf{V}}_{n,j} = \mathbf{V}$, and $\widehat{\nu}_{n,j}^2 = \rho/(2d)$, define for $n \geq j$,

$$\widehat{h}_{n,j} = (1 - \sum_{i=1}^{k} a_i - \sum_{l=1}^{k'} b_l) + \sum_{l=1}^{k'} b_l \widehat{h}_{n-l,j} + \sum_{i=1}^{k} a_i \frac{(y_{n-i} - \widehat{\boldsymbol{\beta}}_{n-i,j}^T \mathbf{x}_{n-i})^2}{\widehat{\nu}_{n-i,j}^2}. \quad (2.3a)$$

$$\mathbf{V}_{n,j} = \mathbf{V}_{n-1,j} - \Big\{ \frac{\mathbf{V}_{n-1,j}\mathbf{x}_n\mathbf{x}_n^T\mathbf{V}_{n-1,j}}{\widehat{h}_{n,j} + \mathbf{x}_n^T\mathbf{V}_{n-1,j}\mathbf{x}_n} \Big\}, \tag{2.3b}$$

$$\widehat{\boldsymbol{\beta}}_{n,j} = \widehat{\boldsymbol{\beta}}_{n-1,j} + \Big\{ \frac{\mathbf{V}_{n-1,j}\mathbf{x}_n\big(y_n - \widehat{\boldsymbol{\beta}}_{n-1,j}^T\mathbf{x}_n\big)}{\widehat{h}_{n,j} + \mathbf{x}_n^T\mathbf{V}_{n-1,j}\mathbf{x}_n} \Big\}, \tag{2.3c}$$

$$\widehat{\nu}_{n,j}^2 = \frac{d+n-j-2}{d+n-j-1}\widehat{\nu}_{n-1,j}^2 + \frac{1}{d+n-j-1} \cdot \frac{\big(y_n - \widehat{\boldsymbol{\beta}}_{n-1,j}^T\mathbf{x}_n\big)^2}{\widehat{h}_{n,j} + \mathbf{x}_n^T\mathbf{V}_{n-1,j}\mathbf{x}_n}, \tag{2.3d}$$

The weights $p_{n,j}$ are the posterior probabilities $P(J_n = j|\mathcal{Y}_n)$ and are given recursively by

$$p_{n,j} \propto p_{n,j}^* := \begin{cases} \frac{pf_{nn}}{f_{00}} & \text{if } j = n, \\ \frac{(1-p)p_{n-1,j}f_{nj}}{f_{n-1,j}} & \text{if } j \le n-1, \end{cases} \tag{2.4}$$

where, letting $\mathbf{z}_{n,j} = \mathbf{V}_{n,j}\big(\mathbf{V}^{-1}\mathbf{z} + \sum_{t=j}^n \mathbf{x}_t y_t/\widehat{h}_{t,j}\big)$ and $\rho_{n,j} = \rho/2 + \mathbf{z}^T\mathbf{V}^{-1}\mathbf{z} - \mathbf{z}_{n,j}^T\mathbf{V}_{n,j}^{-1}\mathbf{z}_{n,j} + \sum_{t=j}^n y_t^2/\widehat{h}_{t,j}$, $p_{n,j} = p_{n,j}^*/\sum_{j'=1}^n p_{n,j'}^*$ and the f_{nj} are given explicitly by $f_{nj} = |\mathbf{V}_{n,j}|^{1/2}\Gamma((d+n-j+1)/2)\rho_{n,j}^{-(d+n-j+1)/2}$ and $f_{00} = |\mathbf{V}|^{1/2}\Gamma(d/2)(\rho/2)^{-d/2}$.

Note that the preceding extension from the case $h_t \equiv 1$ in Lai, Liu, and Xing (2005) to more general known h_t basically amounts to extending OLS (given the most recent change-point J_n) to GLS. Although (2.1) has decomposed the coefficient of ϵ_t into the factor ν_t that undergoes occasional changes and a GARCH part $\sqrt{h_t}$ that involves the time-invariant parameters a_i and b_l, the potential change-point j enters not only in (2.3d) for $\widehat{\nu}_{n,j}^2$, but also in (2.3a) for $\widehat{h}_{n,j}$ which needs $\widehat{\nu}_{n-i,j}^2$ and $\widehat{\boldsymbol{\beta}}_{n-i,j}$ to estimate w_{n-i}^2. Note that given $J_n = j$, (2.3a) corresponds to replacing h_n and w_{n-i}^2 in (2.2) by $\widehat{h}_{n,j}$ and $(y_{n-i} - \widehat{\boldsymbol{\beta}}_{n-i,j}^T\mathbf{x}_{n-i})^2/\widehat{\nu}_{n-i,j}^2$, respectively. The recursion (2.3a) for $\widehat{h}_{n,j}$ is a major advance over Section 2 of Lai, Liu, and Xing (2005) in which GARCH dynamics for short-term volatility is absent.

2.2. Estimation of hyperparameters

Another advance of the present paper over Lai, Liu, and Xing (2005) is related to estimation of hyperparameters. The above Bayesian filter involves $\mathbf{z}, \mathbf{V}, \rho, d, p$, and $\boldsymbol{\eta} = (a_1, \ldots, a_k, b_1, \ldots, b_{k'})$. In particular, the parameter vector $\boldsymbol{\eta}$ is pivotal to GARCH dynamics for short-term volatility and has to be well estimated. Note that $\mathbf{z}$ is the prior mean of $\boldsymbol{\beta}_t$ and $\rho/(d-2)$ is the prior mean of $2\nu^2$ at time t when parameter changes occur. As noted in Lai and Xing (2008, p.107), it is more convenient to represent the χ_d^2/ρ prior distribution for $(2\nu_t^2)^{-1}$ as a Gamma$(d/2, \rho/2)$ distribution so that d does not need to be

an integer. The recursions (2.3b) and (2.3c) are basically recursions for ridge regression which shrinks the GLS estimate (generalized least squares using the weights $\widehat{h}_{t,j}$) towards $\mathbf{z}$, with $\mathbf{V}^{-1}$ and $\sum_{t=j}^{n}\mathbf{x}_t\mathbf{x}_t^T$ being the matrix weights for the shrinkage target and the GLS estimator, respectively. The shrinkage target $\mathbf{z}$ and its associated weight matrix $\mathbf{V}^{-1}$ are relevant when $n-j$ is small, but become increasingly negligible with increasing $n-j$. We can estimate $\mathbf{z}$, $\mathbf{V}$, ρ, and d by applying the method of moments to the stationary distribution of the Markov chain $(I_t, \boldsymbol{\theta}_t, \epsilon_t)$ that is partially observed via $(\mathbf{x}_t, y_t)$. Details are given below. With $\mathbf{z}$, $\mathbf{V}$, ρ, and d replaced by these estimates, we then estimate $\boldsymbol{\eta}$ and p by maximum likelihood. The log-likelihood function ℓ_n based on $y_1, \ldots, y_n$ has the representation

$$\log \ell_n(\boldsymbol{\eta}, p) = \sum_{t=1}^{n} \log f(y_t|\mathcal{Y}_{t-1}, \mathbf{x}_t) = \sum_{t=1}^{n} \log\big[\sum_{j=1}^{t} p^*_{t,j}(\boldsymbol{\eta}, p)\big], \tag{2.5}$$

where $f(\cdot|\cdot)$ denotes conditional density and $p^*_{t,j}$ is given by (2.4); see Section 2.3 of Lai and Xing (2011). Further simplifications for implementation are given in the next section. The hyperparameter estimates in Section 5 of Lai, Liu, and Xing (2005) use a coarser accumulated prediction error criterion or sequential Monte Carlo, which becomes much harder to implement when the GARCH parameter $\boldsymbol{\eta}$ is also present and has to be estimated well.

We now describe the method-of-moments estimates of $\mathbf{z}$, $\mathbf{V}$, ρ, and d based on $(\mathbf{x}_t, y_t)$, $1 \le t \le n$. From (A2) and (A3), it follows that $E(\boldsymbol{\beta}_t) = \mathbf{z}$, $\mathrm{Cov}(\boldsymbol{\beta}_t) = (E\nu_t^2)\mathbf{V}$, and $E(\mathbf{x}_t y_t) = E(\mathbf{x}_t\mathbf{x}_t^T)\mathbf{z}$. From $n-L$ moving windows $\{(\mathbf{x}_t, y_t) : s \le t \le s+L\}$ of these data, compute the least squares estimate and denote it as $\widehat{\boldsymbol{\beta}}^{(s)}$. Each $\widehat{\boldsymbol{\beta}}^{(s)}$ is a method-of-moments estimate of $\mathbf{z}$, and so is $\overline{\boldsymbol{\beta}} = (n-L)^{-1}\sum_{s=1}^{n-L}\widehat{\boldsymbol{\beta}}^{(s)}$. If an oracle would reveal the change-times up to time n, then one would segment the time series accordingly and use the least squares estimate for each segment to estimate the regression parameter for that segment. The average of these least squares estimates over the segment would provide a method-of-moments estimate of $\mathbf{z}$. Similarly, the average squared residual in each segment is a method-of-moments estimate of $E(\nu_t^2) = \rho/[2(d-2)]$ and so is the average of these values over the segments; see Engle and Mezrich (1996). In ignorance of the change-points, we replace the segments by moving windows of length $L+1$ in $\widehat{\boldsymbol{\beta}}^{(s)}$ and estimate $\mathbf{z}$ by the average $\overline{\boldsymbol{\beta}}$ of the $\widehat{\boldsymbol{\beta}}^{(s)}$. Likewise we estimate ρ and d by equating the mean and variance of the inverted gamma distribution for ν_t^2 to their sample counterparts for the average squared residuals. And similarly we can estimate $\mathbf{V}$. By the Ergodic Theorem, the averages over s in $\overline{\boldsymbol{\beta}}$ and these sample counterparts converge to their means under the stationary distribution of the Markov chain $(I_t, \boldsymbol{\theta}_t, \epsilon_t, \mathbf{x}_t)$, with probability 1 as $n \to \infty$,

assuming that $\mathbf{x}_t$ is also an ergodic Markov chain. Moreover, if L is chosen much smaller than $1/p$, then most of the moving windows do not have change-points, and therefore the method-of-moments estimates differ from the preceding oracle estimates by $o_p(1)$ as $pL \to 0$.

2.3. Bounded complexity mixture approximations

Although (2.4) provides closed-form recursions for updating the weights $p_{t,i}$, $1 \le i \le t$, the number of weights increases with t, resulting in rapidly increasing computational complexity and memory requirements for estimating θ_n as n increases. A natural idea to reduce the complexity and to facilitate the use of parallel algorithms for the recursions is to keep only a fixed number M of weights at every stage n (which is tantamount to setting the other weights to be 0). Following Lai, Liu, and Xing (2005), we keep the most recent m weights $p_{n,i}$ (with $n - m < i \le n$) and the largest $M - m$ of the remaining weights, where $1 \le m < M$. Specifically, let $\mathcal{K}_{n-1}$ denote the set of indices i for which $p_{n-1,i}$ is kept at stage $n-1$; thus $\mathcal{K}_{n-1} \supset \{n-1, , \cdots, n-m\}$. At stage n, define $p^*_{n,i}$ by (2.4) for $i \in \{n\} \cup \mathcal{K}_{n-1}$, and let i_n be the index not belonging to $\{n, n-1, \cdots, n-m+1\}$ such that $p^*_{n,i_n} = \min\{p^*_{n,i} : j \in \mathcal{K}_{n-1}$ and $j \le n-m\}$, choosing i_n to be the one farthest from n if the minimizing set in p^*_{n,i_n} has more than one element. Define $\mathcal{K}_n = \{n\} \cup (\mathcal{K}_{n-1} - \{i_n\})$ and let $p_{n,i} = p^*_{n,i} \Big/ \sum_{j\in\mathcal{K}_n} p^*_{n,j}$.

We use these BCMIX not only to approximate the filters $(\boldsymbol{\beta}_t, \nu_t)|\mathcal{Y}_t$ but also to approximate the likelihood function (2.5), in which we replace $\sum_{j=1}^t$ by $\sum_{j\in\mathcal{K}_t}$. Letting $\boldsymbol{\lambda} = (p, \boldsymbol{\eta})$, the maximizer $\boldsymbol{\lambda}$ of $\ell_{n-1}(\boldsymbol{\lambda})$ is used to replace $\boldsymbol{\lambda}$ in the recursions (2.3) and (2.4). For reasons explained in Lai and Xing (2011, p.548), we use a grid of the form $\{2^j/n : j_0 \le j \le j_1\}$, where $j_0 < 0 < j_1$ are integers, to search for the maximum of $\ell_n(p; \widehat{\boldsymbol{\eta}}_n)$. The update $\widehat{\boldsymbol{\eta}}_n$ of the GARCH parameters after observing $(\mathbf{x}_n, y_n)$ uses a single iteration of the Newton-Raphson iteration procedure to maximize $\ell_n(\widehat{p}_{n-1}, \boldsymbol{\eta})$ when $n \ge n_0$, and uses more iterations until convergence for small n. Therefore relatively fast updates of the hyperparameters estimates can be used to implement the adaptive BCMIX filters.

2.4. BCMIX smoothers

We begin by deriving the Bayes estimate (smoother) of $\boldsymbol{\theta}_t = (\boldsymbol{\beta}_t^T, \tau_t)^T$ given $\mathcal{Y}_n$ for $1 \le t \le n$ in the "oracle" setting in which the h_t are specified exactly (by the oracle) so that there are explicit recursive representations of the posterior mean of $\boldsymbol{\theta}_t$ given $\mathcal{Y}_n$ for $1 \le t \le n$. To obtain the optimal smoother $E(\boldsymbol{\theta}_t|\mathcal{Y}_n)$ for $1 \le t \le n$, we use Bayes' theorem to combine the forward filter $\boldsymbol{\theta}_t|\mathcal{Y}_t$ with the backward filter $\boldsymbol{\theta}_t|\mathcal{Y}_{t+1,n}$; see Section 2.2 of Lai and Xing (2011). Because the

h_t are assumed known in $(y_t - \boldsymbol{\beta}_t^T \mathbf{x}_t)/\sqrt{h_t} = \nu_t \epsilon_t$ and the ϵ_t are i.i.d. standard normal, the backward filter has the same form as the forward filter. Note that assumptions (A1)−(A3) define a reversible Markov chain of jump times and jump magnitudes, assuming $I_{n-t_0+1} = 1$ and no change-points afterwards. Let π denote the density function of the stationary distribution. Letting $\widetilde{J}_{t+1} = \min\{s \geq t+1 : I_s = 1\}$ and $q_{t+1,j} = P(\widetilde{J}_{t+1} = j | \mathcal{Y}_{t+1,n})$ for $j \geq t+1$, we can reverse time and obtain a backward filter that is similar to the forward filter:

$$f(\boldsymbol{\theta}_t | \mathcal{Y}_{t+1,n}) = p\pi(\boldsymbol{\theta}_t) + (1-p) \sum_{j=t+1}^{n} q_{t+1,j} f(\boldsymbol{\theta}_{t+1} | \mathcal{Y}_{t+1,n}, \widetilde{J}_{t+1} = j),$$

in which

$$q_{t+1,j} \propto q^*_{t+1,j} = \begin{cases} \frac{p f_{jj}}{f_{00}} & \text{if } j = t+1, \\ \frac{(1-p) q_{t+2,j} f_{t+1,j}}{f_{t+2,j}} & \text{if } j \geq t+2. \end{cases}$$

Application of Bayes' theorem then yields

$$f(\boldsymbol{\theta}_t | \mathcal{Y}_n) = \sum_{1 \leq i \leq t \leq j \leq n} \alpha_{ijt} f(\boldsymbol{\theta}_t | \mathcal{Y}_n, C_{ij}), \tag{2.6}$$

where $C_{ij} = \{I_i = 1, I_{i+1} = \cdots = I_j = 0, I_{j+1} = 1\}$ and the α_{ijt} are determined recursively by

$$\begin{aligned} \alpha_{ijt} &= \frac{\alpha^*_{ijt}}{A_t}, \qquad A_t = \sum_{1 \leq i \leq t \leq j \leq n} \alpha^*_{ijt}, \\ \alpha^*_{ijt} &= \begin{cases} p p_{i,t}, & i \leq t, j = t, \\ \frac{a p_{i,t} q_{t+1,j} f_{00} f_{ij}}{f_{it} f_{t+1,j}}, & i \leq t, j > t. \end{cases} \end{aligned} \tag{2.7}$$

The next step is to approximate α_{ijt} by $\widehat{\alpha}_{ijt}$ that replaces the h_t, which is actually unknown, by the estimates $\widehat{h}_{j,i}$ defined recursively for $j \geq i$ by (2.3a). As in Section 2.1, we assume known hyperparameters p and $\boldsymbol{\eta}$ for the time being. Using the BCMIX approximation to the forward and backward filters, we approximate the sum in (2.6) and that defining A_t in (2.7) by

$$\boldsymbol{\beta}_t | \boldsymbol{\tau}_n, \mathcal{Y}_n \sim \sum_{i \in \mathcal{K}_t, j \in \{t\} \cup \widetilde{\mathcal{K}}_{t+1}} \alpha_{ijt} N(\mathbf{z}_{j,i}, \frac{\mathbf{V}_{j,i}}{(2\tau_{j,i})}), \quad \tau_{j,i} | \mathcal{Y}_n \sim \frac{\chi^2_{d+j-i+1}}{\rho_{j,i}}, \tag{2.8}$$

where $\mathcal{K}_t$ is the same as that in Section 2.3 for the forward filter and $\widetilde{\mathcal{K}}_{t+1}$ is the corresponding set for the backward filter; see Section 4.2 of Lai and Xing (2011). Assuming known $\boldsymbol{\eta}$ and p, the BCMIX estimates for $\boldsymbol{\beta}_t$ and ν_t given $\mathcal{Y}_n$ are

$$\widehat{\boldsymbol{\beta}}_{t|n} = \sum_{i \in \mathcal{K}_t, j \in \{t\} \cup \widetilde{\mathcal{K}}_{t+1}} \alpha_{ijt} \mathbf{z}_{j,i}, \quad \widehat{\nu}^2_{t|n} = \sum_{i \in \mathcal{K}_t, j \in \{t\} \cup \widetilde{\mathcal{K}}_{t+1}} \alpha_{ijt} \frac{\rho_{j,i}}{d+j-i-1}. \tag{2.9}$$

Moreover, the conditional probability of a change point at time t given $\mathcal{Y}_n$ is estimated by

$$\widehat{P}(I_{t+1} = 1|\mathcal{Y}_n) = \sum_{1\le i\le t} \widehat{P}(C_{it}|\mathcal{Y}_n) = \frac{p}{A_t}. \tag{2.10}$$

Without assuming p and $\boldsymbol{\eta}$ to be known, we can use the BCMIX approximation in the log-likelihood function (2.5) based on $\mathcal{Y}_n$, and evaluate its maximizer $(\widehat{p}, \widehat{\boldsymbol{\eta}})$. Replacing $(p, \boldsymbol{\eta})$ by $(\widehat{p}, \widehat{\boldsymbol{\eta}})$ in (2.8) yields the BCMIX empirical Bayes smoother.

3. Long-range Dependence in GARCH Models Ignoring Change-points

Mikosch and Starica (2004) pointed out that although the "integrated GARCH effect" had been related to structural changes in the financial econometrics literature dating back to Diebold (1986), the studies to date made use of "either simulations or indirect approaches to substantiate their claims." They showed that the sample autocorrelation function of the squared log-returns of a piecewise stationary GARCH sequence approaches a positive constant as the lag approaches ∞ and that the spectral density estimates at arbitrarily small frequencies can become arbitrarily large. They also showed that the maximizers of the Whittle likelihood of a fitted GARCH(1,1) model converges to that of a nonrandom function $-\Delta$ as the sample size approaches ∞, and noted that although "it is not possible to obtain an explicit form of the minimizer" of Δ, one can minimize Δ "numerically" to show that the sum $\widehat{a} + \widehat{b}$ of the estimated parameters converges to 1 when the underlying GARCH model is piecewise stationary. Hillebrand (2005) considered the ideal case in which ϵ_t and its coefficient $\sigma_t = \sqrt{h_t}\nu_t$ are completely observable so that method-of-moments estimation of an assumed GARCH(1,1) model can be based on (σ_t, ϵ_t). Assuming that "the influence of a single realization" of (σ_t, ϵ_t) on the estimator of the GARCH parameters "vanishes with growing sample size," he showed that $\widehat{a} + \widehat{b}$ converges to 1 when the underlying model is piecewise stationary. He called this "spurious almost-integration" and conjectured that "if spurious almost-integration occurs when σ_t and ϵ_t are observable, it will also occur when there is less perfect information about the volatility condition of the market."

In this section we prove that this spurious almost-integration property holds for maximum likelihood estimates in the proposed stochastic change-point ARX-GARCH model satisfying (A1)−(A3), and such that $\mathbf{x}_t$ is an ergodic Markov chain with stationary transition probabilities. We begin by considering the GARCH(1,1) model without the ARX component, which can be written as

$$(1 - \lambda^* B)y_t^2 = \omega_t + \zeta_t - b^*\zeta_{t-1}, \tag{3.1}$$

where $\lambda^* = a^* + b^*$, a^* and b^* denote the true values of a and b, $\zeta_t = y_t^2 - \sigma_t^2$, $\sigma_t^2 = h_t\nu_t$, $\omega_t = (1-\lambda^*)\nu_t^2$, and B denotes the back-shift operator. Note that $(y_t, \sigma_t^2, \omega_t)$ is an ergodic Markov chain with stationary distribution π and that y_t is a component of this chain. As shown below, fitting the GARCH model (1.1) by maximum likelihood is asymptotically equivalent to choosing a, b, and $\omega = (1-a-b)\nu^2$ in (1.1) to yield the smallest Kullback-Leibler divergence (or relative entropy) of the fitted Markov chain $(y_t, \sigma_t^2(\omega, a, b))$ to the actual chain. The relative entropy, which is of basic importance in information theory, is a measure of the discrepancy of an approximating distribution Q from the actual distribution P. If P is the distribution of a sequence of random variables, one would like Q to capture the time series properties of P, and the better Q is able to do so, the less is the relative entropy.

In the actual Markov chain (3.1), ω_t has occasional jumps at random times generated by a geometric renewal process. We can express ω_t as $\omega^* + \delta_t$, where $\omega^* = (1-\lambda^*)E_\pi(y_t^2)$, $E_\pi(\delta_t) = 0$, δ_t is piecewise constant with expected duration $1/p$ between consecutive jumps, and E_π denotes expectation when the initial distribution of the Markov chain is the stationary distribution π. The fitted GARCH(1,1) model ignores the component δ_t in

$$(1-\lambda^* B)y_t^2 = \omega^* + \delta_t + \zeta_t - b^*\zeta_{t-1}, \tag{3.2}$$

and considers only the moving average $\zeta_t - b^*\zeta_{t-1}$ of uncorrelated innovations ζ_i, which are in fact martingale differences. Choosing $\lambda = a+b$ not sufficiently close to 1 in (1.1) would miss the time-scale of $1/p$ for the average jump time of δ_t, and time series of length n generated from the fitted model would have too short memory in comparison with the observed $y_1, \ldots, y_n$. The smallest Kullback-Leibler divergence of the fitted GARCH(1,1) model is attained by choosing $\lambda = \lambda(p)$, with $\lambda(p) \to 1$ as $p \to 0$ to capture the long time-scale.

Mikosch and Starica (2004) have shown that the sample autocorrelation function of y_t^2 approaches a positive constant as the lag approaches ∞, suggesting long-range dependence, when the underlying GARCH(1,1) model is piecewise stationary. Their piecewise stationarity means that the GARCH(1,1) model has r change-points for its parameters (ω, a, b) so that the parameters between successive change-points are time-invariant. They assume that r is fixed and that the number of observations between two successive change-points becomes infinite as $n \to \infty$. In this general framework they "cannot provide a (theoretical) result for the asymptotic behavior" of the maximum likelihood estimator. By using the proposed stochastic change-point model to embed the piecewise constant stationary GARCH(1,1) model into an ergodic Markov chain with stationary transition probabilities, we can use the stationary distribution to derive the Kullback-Leibler divergence of a fitted GARCH model from the true one

that has change-points. Technical details are given below and rely on a representation of the likelihood function provided by Berkes, Horváth, and Kokoszka (2003), abbreviated BHK. Letting $n \to \infty$ and then $p \to 0$ establishes the spurious almost-integration property in fitting a GARCH(1,1) model by maximum likelihood to the stochastic change-point model (3.2).

The preceding argument can clearly be extended to stochastic change-point GARCH(k, k') models. In fact, BHK considers general k and k' and writes the log-likelihood function, based on $y_1, \ldots, y_n$, of an assumed GARCH(k, k') model (without change-points), as

$$L_n(\boldsymbol{\theta}) = -\frac{1}{2}\sum_{t=1}^{n}\Big\{\log u_t(\boldsymbol{\theta}) + \frac{y_t^2}{u_t(\boldsymbol{\theta})}\Big\}, \tag{3.3}$$

where $\boldsymbol{\theta} = (\omega, a_1, \ldots, a_k, b_1, \ldots, b_{k'})$, $u_t(\boldsymbol{\theta}) = c_0(\boldsymbol{\theta}) + \sum_{i=1}^{t-1} c_i(\boldsymbol{\theta}) y_{t-i}^2$, and $c_i(\boldsymbol{\theta})$ are defined by explicit recursions; see p.216 and Section 3 of BHK. The sequence $c_1(\boldsymbol{\theta}), c_2(\boldsymbol{\theta}), \ldots$ "decays exponentially fast," uniformly in $\Theta_{\epsilon,\epsilon',\epsilon''} := \big\{\boldsymbol{\theta} : b_1 + \cdots + b_{k'} \le \epsilon, \min(\boldsymbol{\theta}) \ge \epsilon', \max(\boldsymbol{\theta}) \ge \epsilon''\big\}$ for any $0 < \epsilon < 1$, $0 < \epsilon' < \epsilon/k'$ and $\epsilon'' > \epsilon'$, where $\min(\boldsymbol{\theta})$ and $\max(\boldsymbol{\theta})$ denote the minimum and maximum of the components of $\boldsymbol{\theta}$, respectively (BHK, pp.205 and 211). Since the Markov chain $(y_t, \sigma_t^2, \omega_t)$ is ergodic, it follows from the Ergodic Theorem that $n^{-1}L_n(\boldsymbol{\theta})$ converges a.s. to

$$\ell(\boldsymbol{\theta}) := -\frac{1}{2}E_\pi\Big\{\log\big[c_0(\boldsymbol{\theta}) + \sum_{i=1}^{\infty} c_i(\boldsymbol{\theta}) y_{-i}^2\big] + \frac{y_0^2}{\big[c_0(\boldsymbol{\theta}) + \sum_{i=1}^{\infty} c_i(\boldsymbol{\theta}) y_{-i}^2\big]}\Big\} \tag{3.4}$$

uniformly over $\Theta_{\epsilon,\epsilon',\epsilon''}$, where we extend the stationary sequence $\{y_t, t \ge 1\}$ (under π) to $\{y_t, -\infty < t < \infty\}$; see p.214 of BHK and the second paragraph of Section 2.2. The log-likelihood, based on $y_1, \ldots, y_n$, under the true model is $\ell(\boldsymbol{\eta}^*, p^*)$ given by (2.5), with $\boldsymbol{\eta}^*$ and p^* denoting the actual hyperparameter values in the change-point GARCH(k, k') model and with the other hyperparameters also replaced by their actual values. A similar argument using the Ergodic Theorem shows that $n^{-1}\ell_n$ converges a.s. to

$$\ell^* := \lim_{t\to\infty} E_\pi\big\{\log f(y_t|\mathcal{Y}_{t-1})\big\}.$$

The Kullback-Leibler divergence of the GARCH model with parameter vector $\boldsymbol{\theta}$ from the actual stochastic change-point model is $\ell^* - \ell(\boldsymbol{\theta})$. Therefore choosing $\boldsymbol{\theta}$ to maximize the log-likelihood (3.3) is asymptotically equivalent to finding the $\boldsymbol{\theta}$ that minimizes the Kullback-Leibler divergence $\ell^* - \ell(\boldsymbol{\theta})$. The best approximating GARCH(k, k') model to the stochastic change-point model in terms of Kullback-Leibler divergence would choose the GARCH parameters to capture the time-scale of $1/p$ for the average jump time of ω_t, explaining the volatility persistence

of the fitted model when p is small. This argument can be readily extended to the ARX-GARCH model with contemporaneous jumps in the ARX and GARCH parameters if $\mathbf{x}_t$ in (2.1) is also assumed to be an ergodic Markov chain.

The piecewise stationary GARCH model of Mikosch and Starica (2004) corresponds to the case $n \to \infty$ and $p \to 0$ such that $np \to r > 0$ if we embed it in the stochastic change-point GARCH(1,1) model. Although this does not fall in the asymptotic regime $n \to \infty$ and then $p \to 0$ of the preceding argument, we can use a more direct argument to show that $n^{-1}L_n(\boldsymbol{\theta})$ still converges in this case; in fact, the limit is a convex combination of the limits over different stationary pieces. The maximizing $\boldsymbol{\theta}$ again exhibits spurious almost-integration because of the time-scale of $1/p$ for the average jump time of ω_t.

As noted by BHK (p.202), the GARCH(k, k') model can be defined recursively by

$$\mathbf{Y}_{t+1} = \mathbf{A}_t \mathbf{Y}_t + (\omega, 0, \ldots, 0)^T, \tag{3.5}$$

where $\mathbf{Y}_t = (\sigma_t^2, \ldots, \sigma_{t-k'+1}^2, y_{t-1}^2, \ldots, y_{n-k+1}^2)^T \in \mathbb{R}^{k+k'-1}$ and $\mathbf{A}_t$ is a $(k + k' - 1) \times (k + k' - 1)$ matrix whose entries involve $a_1, \ldots, a_k, b_1, \ldots, b_{k'}$, and ϵ_n^2. Our change-point GARCH(k, k') model also has this representation with ω replaced by ω_{t+1} in (3.5). Siegmund (2001) has considered such recursions in the univariate case, $Y_{t+1} = A_t Y_t + W_t$, and derived an asymptotic recursion in the univariate case for the tail behavior of the stationary distribution of Y_t. He says that "one motive for studying (the recursion) is to obtain information about the ARCH(1) process," which is a special case of the recursion.

4. Application to the Segmentation Problem

In principle, the frequentist approach to multiple change-point problems for regression models reviewed in Section 1 can be extended to ARX-GARCH models by maximizing the log-likelihood over the locations of the change-points and the piecewise constant parameters when it is assumed that there are k change-points. This optimization problem, however, is much more difficult than that for regression models and only constitutes an inner loop of an algorithm whose outer loop is another minimization, over k, of a suitably chosen model selection criterion to determine k. One such selection criterion is Siegmund's (2004) modified BIC for non-smooth (such as change-point) models. In this section, we use the relative simplicity of the BCMIX smoothers in the empirical Bayes approach in Section 2 to circumvent the computational complexity of the segmentation problem.

For computational and analytic tractability, the frequentist approach typically assumes that k is small relative to n and that adjacent change-points are sufficiently far apart so that the segments are identifiable except for relatively small neighborhoods of change-points; see e.g., Bai and Perron (1998), Mikosch and Starica (2004), Galeano and Tsay (2010). Lai and Xing (2011) formulate these assumptions for the piecewise constant parameter vectors $\boldsymbol{\theta}_t$ as follows.

(B1) The true change-points occur at $t_1^{(n)} < \cdots < t_k^{(n)}$ such that $\liminf_{n\to\infty} n^{-1}(t_i^{(n)} - t_{i-1}^{(n)}) > 0$ for $1 \le i \le k+1$, with $t_0^{(n)} = 0$ and $t_{k+1}^{(n)} = n$.

(B2) There exists $\delta > 0$, which does not depend on n, such that $\min_{1\le i\le k} ||\boldsymbol{\theta}_{t_i^{(n)}} - \boldsymbol{\theta}_{t_{i-1}^{(n)}}|| \ge \delta$ for all large n.

In addition, we also assume that the stochastic regressors satisfy a stability condition.

(B3) $\max_{1\le t\le n} ||\mathbf{x}_t||^2/n \xrightarrow{P} 0$ and $\sum_{t=1}^n \mathbf{x}_t\mathbf{x}_t^T/n$ converge a.s. to a positive definite non- random matrix.

In the case of an AR(κ) model for which $\mathbf{x}_t = (y_{t-1}, \ldots, y_{t-\kappa})^T$, the stationary assumption that $1 - \beta_{t,1}z - \cdots - \beta_{t,\kappa}z^\kappa$ has roots inside and uniformly bounded away from the unit circle for $t \in \{1, t_1^{(n)}, \ldots, t_k^{(n)}\}$, as in Lai, Liu, and Xing (2005), ensures (B3). For an ARX model, besides this stationarity assumption on the coefficients of y_{t-j}, we also require that the subvector consisting of the other components of $\mathbf{x}_t$ (representing the exogenous inputs) satisfies (B3). Making use of (B1)−(B3) and an argument similar to that used in the proof of Theorem 2 in Lai and Xing (2011, pp.563-567), we can prove the following.

Theorem 1. *Assume* (B1)−(B3) *and that* $m \sim |\log n|^{1+\epsilon}$ *and* $M - m = O(1)$ *as* $n \to \infty$, *for some* $\epsilon > 0$. *Then the BCMIX smoother* $\widehat{\boldsymbol{\theta}}_{t|n}$ *satisfies*

$$\max_{1\le t\le n:\min_{1\le i\le k}|t-t_i^{(n)}|\ge m} ||\widehat{\boldsymbol{\theta}}_{t|n} - \boldsymbol{\theta}_t|| \to 0, \quad \text{as } n \to \infty$$

uniformly in $a_1/n \le p \le a_2/n$.

The proof of Theorem 1 proceeds by first assuming the GARCH parameter vector $\boldsymbol{\eta}$ to be known and showing that the BCMIX smoother $\widehat{\boldsymbol{\theta}}_{t|n}(p)$ has the desired convergence property uniformly in $a_1/n \le p \le a_2/n$. It then proves consistency of $\widehat{\boldsymbol{\eta}}$ and thereby establishes the desired result for $\widehat{\boldsymbol{\theta}}_{t|n}$. Following Lai and Xing (2011), we can apply Theorem 1 to estimate the change-times $t_1^{(n)}, \ldots, t_k^{(n)}$ in (B1). Let

$$\Delta_t = ||\widehat{\boldsymbol{\theta}}_{t+m} - \widehat{\boldsymbol{\theta}}_{t-m}||^2, \tag{4.1}$$

and let $\widehat{\tau}_1$ be the maximizer of Δ_t over $m < t < n - m$. After $\widehat{\tau}_1, \ldots, \widehat{\tau}_{j-1}$ have been defined, take

$$\widehat{\tau}_j = \arg \max_{t:m<t<n-m,\min_{1\le i\le j-1}|t-\widehat{\tau}_i|\ge m} \Delta_t. \tag{4.2}$$

Note that the estimates $\widehat{\tau}_j$ of the change-times are unordered and do not depend on the number k of change-points. Assuming that there are k change-points, we can order $\widehat{\tau}_1, \ldots, \widehat{\tau}_k$ as $\widehat{t}_{(1),k} < \cdots < \widehat{t}_{(k),k}$ to provide estimates of $t_1^{(n)} < \cdots < t_k^{(n)}$. Let $\boldsymbol{\theta}^{(j)}$ be the common value of $\boldsymbol{\theta}_t$ in the interval $t_{j-1}^{(n)} \le t < t_j^{(n)}$. Assuming $t_1^{(n)}, \ldots, t_k^{(n)}$ to be known, we can estimate the parameter vectors $\boldsymbol{\theta}^{(1)}, \boldsymbol{\theta}^{(2)}, \ldots, \boldsymbol{\theta}^{(k+1)}$, which are the parameter values prior to $t_1^{(n)}$, between $t_1^{(n)}$ and $t_2^{(n)}$, ..., and after $t_k^{(n)}$, respectively, by maximum likelihood. Replacing the $t_1^{(n)}, \ldots, t_k^{(n)}$ by the estimates $\widehat{t}_{(1),k} < \cdots < \widehat{t}_{(k),k}$ in these MLE's leads to the quasi-likelihood estimators $\widetilde{\boldsymbol{\theta}}^{(1)}, \ldots, \widetilde{\boldsymbol{\theta}}^{(k+1)}$, $\widetilde{\boldsymbol{\eta}}$ and the quasi-likelihood $\Lambda_n(k) = \sum_{j=1}^{k+1} \sum_{t=\widehat{t}_{(j-1),k}}^{\widehat{t}_{(j),k}-1} \log f(y_t; \boldsymbol{\theta}^{(j)}, \boldsymbol{\eta})$, in which $f(\cdot; \boldsymbol{\theta}, \boldsymbol{\eta})$ is the density function of y_t given the piecewise constant parameter values, noting that $(y_t - \boldsymbol{\beta}_t^T \mathbf{x}_t)/(\nu_t \sqrt{h_t})$ is standard normal. Assuming a known upper bound K on the number k of change-points in (B1), Lai and Xing (2011) propose to estimate k by $\widehat{k}_n = \arg\max_{1 \le k \le K} \{\Lambda_n(k) - (k+1)C_n\}$, where C_n is a penalty term that satisfies

$$C_n \to \infty \text{ and } C_n/n \to 0 \text{ as } n \to \infty. \tag{4.3}$$

Examples of (4.3) are the BIC and Siegmund's (2004) modified BIC. Making use of Theorem 1, we prove in the Appendix the following result on the consistency of $\widehat{k}_n$ and $\widetilde{\boldsymbol{\eta}}$ and $\widetilde{\boldsymbol{\theta}}^{(j)}$.

Theorem 2. *Under the assumptions of Theorem* 1, $\widehat{k}_n \xrightarrow{P} k$, *and* $\widetilde{\boldsymbol{\eta}} \xrightarrow{P} \boldsymbol{\eta}$ *and* $\widetilde{\boldsymbol{\theta}}^{(j)} \xrightarrow{P} \boldsymbol{\theta}^{(j)}$ *for* $1 \le j \le k+1$.

To compare the true parameter process $\{\boldsymbol{\theta}_t\}$ and an estimated process $\{\widehat{\boldsymbol{\theta}}_t\}$, we consider their relative entropy (Kullback-Leibler information) $\mathrm{KL} = n^{-1} \sum_{t=1}^n \mathrm{KL}_t$, where in our change-point stochastic regression model with standard normal ϵ_t, the relative entropy KL_t is given by

$$2\mathrm{KL}_t = \frac{\left[\mathbf{x}_t^T(\boldsymbol{\beta}_t - \widehat{\boldsymbol{\beta}}_t)\right]^2}{\widehat{v}_t^2 \widehat{h}_t} + \frac{v_t^2 h_t}{\widehat{v}_t^2 \widehat{h}_t} - \log \frac{v_t^2 h_t}{\widehat{v}_t^2 \widehat{h}_t} - 1.$$

A simulation study of the performance of the preceding segmentation procedure considered the mean-shift GARCH(1,1) model (which is a special case of (2.1) with $\mathbf{x}_t \equiv 1$):

$$y_t = \mu_t + \nu_t \sqrt{h_t} \epsilon_t, \qquad 1 \le t \le n = 1{,}000,$$

in which $h_t = 1 - a - b + a w_{t-1}^2 + b h_{t-1}$, with seven choices of the GARCH parameters: $(a, b) = (0.1, 0.3)$, $(0.1, 0.5)$, $(0.1, 0.7)$, $(0.3, 0.3)$, $(0.3, 0.5)$, $(0.5, 0.3)$, $(0.5, 0.4)$, and five scenarios of the number of change-points and the piecewise constant (μ_t, ν_t):

Table 1. Kullback-Leibler information KL (with standard error in parentheses) and relative frequency of $\widehat{k}$.

	a	0.1	0.1	0.1	0.3	0.3	0.5	0.5
	b	0.3	0.5	0.7	0.3	0.5	0.3	0.4
(a)	KL	0.0015	0.0021	0.0047	0.0065	0.0181	0.0233	0.0352
		(4.5e-5)	(5.6e-5)	(9.4e-5)	(1.1e-4)	(2.0e-4)	(2.3e-4)	(2.7e-4)
	$\widehat{k}=0$	1.000	1.000	1.000	1.000	1.000	1.000	1.000
(b)	KL	0.0033	0.0038	0.0060	0.0074	0.0256	0.0234	0.091
		(7.3e-5)	(8.2e-5)	(1.1e-4)	(1.5e-4)	(7.0e-4)	(6.4e-4)	(2.4e-3)
	$\widehat{k}=1$	0.977	0.968	0.923	0.912	0.735	0.759	0.510
	$\widehat{k}=2$	0.023	0.031	0.073	0.078	0.192	0.180	0.286
	$\widehat{k}\geq 3$	0	0.001	0.004	0.010	0.073	0.061	0.204
(c)	KL	0.0060	0.0067	0.0098	0.0111	0.0238	0.0246	0.0499
		(9.9e-5)	(1.1e-4)	(1.6e-4)	(1.7e-4)	(4.6e-4)	(4.5e-4)	(1.1e-3)
	$\widehat{k}=2$	0.960	0.943	0.878	0.882	0.724	0.726	0.596
	$\widehat{k}=3$	0.038	0.056	0.109	0.107	0.200	0.203	0.244
	$\widehat{k}\geq 4$	0.002	0.002	0.013	0.011	0.076	0.071	0.160
(d)	KL	0.0080	0.0105	0.0213	0.0199	0.0477	0.0418	0.1152
		(1.2e-4)	(1.7e-4)	(3.2e-4)	(2.9e-4)	(7.5e-4)	(7.2e-4)	(2.1e-3)
	$\widehat{k}=2$	0	0	0.016	0.002	0.002	0	0
	$\widehat{k}=3$	0.929	0.878	0.637	0.689	0.430	0.507	0.221
	$\widehat{k}=4$	0.069	0.116	0.237	0.207	0.318	0.270	0.239
	$\widehat{k}\geq 5$	0.002	0.006	0.110	0.102	0.250	0.223	0.540
(e)	KL	0.0206	0.0207	0.0256	0.0333	0.1649	0.1813	0.1927
		(5.5e-4)	(5.6e-4)	(8.4e-4)	(1.2e-3)	(3.7e-3)	(4.2e-3)	(4.3e-4)
	$\widehat{k}-k$	0.182	0.196	0.240	0.310	1.286	1.291	3.451
	$\mathrm{se}(\widehat{k}-k)$	0.016	0.015	0.017	0.022	0.042	0.042	0.113

(a) no change-point and $(\mu_t, \nu_t) \equiv (0, 1)$;

(b) one change-point and $(\mu_t, \nu_t) = (-0.5, 0.5)1_{\{t\leq n/2\}} + (0.5, 0.75)1_{\{t>n/2\}}$;

(c) two change-points and $(\mu_t, \nu_t) = (-0.5, 0.5)1_{\{t\leq n/4\}} + (0.5, 0.75)1_{\{n/4<t\leq 3n/4\}} + (0, 0.6)1_{\{t>3n/4\}}$;

(d) three change-points and $(\mu_t, \nu_t) = (-0.5, 0.5)1_{\{t\leq n/4\}} + (0.5, 0.75)1_{\{n/4<t\leq n/2\}} + (0, 0.6)1_{\{n/2<t\leq 3n/4\}} + (-0.5, 0.8)1_{\{t>3n/4\}}$;

(e) the number of change-points and (μ_t, ν_t) follow (A1)−(A3) with $p = 0.01$, $z = 0$, $V = 1$, $d = 5$ and $\rho = 1$.

The results on KL and the distribution of $\widehat{k}$, based on 1,000 simulations, are summarized in Table 1; it shows that the proposed segmentation procedure performs well in the frequentist and Bayesian scenarios.

5. An Empirical Study

Figure 1, top panel, plots the weekly returns r_t of the SP500 index, from the week starting on January 2, 1990 to the week starting on August 24, 2009. The dataset consists of $n = 1024$ closing prices P_t on the last day of the week from which the returns $y_t = P_t/P_{t-1} - 1$ are computed. The mean, variance, skewness, and kurtosis of the return series are 1.327×10^{-3}, 5.570×10^{-4}, -0.477, and 8.992, respectively. We fit the change-point AR(1)-GARCH(1,1) model

$$y_t = \mu_t + \alpha_t y_{t-1} + \nu_t w_t, \qquad w_t = \sqrt{h_t}\epsilon_t, \qquad h_t = 1 - a - b + a w_{t-1}^2 + b h_{t-1} \quad (5.1)$$

to these data, assuming (A1)−(A3) for the stochastic dynamics of (μ_t, α_t, ν_t). The bottom panel of Figure 1 plots the posterior probability $P(I_t = 1|\mathcal{Y}_n)$ of a change-point at time t given by the fitted model via (2.10). For comparison, we also use `garchfit` in MATLAB to fit the AR(1)-GARCH(1,1) model

$$y_t = \mu + \alpha y_{t-1} + w_t, \qquad w_t = \sigma_t \epsilon_t, \qquad \sigma_t^2 = \omega + \psi w_{t-1}^2 + \phi \sigma_{t-1}^2, \qquad (5.2)$$

with time-invariant parameters, to these data. In both (5.1) and (5.2), ϵ_t are assumed to be i.i.d. standard normal. The maximum likelihood estimates of the parameters of (5.2), based on the entire time series y_t, $1 \le t \le n$, are

$$\widehat{\mu} = 2.046 \times 10^{-3}, \quad \widehat{\alpha} = -0.110, \quad \widehat{\omega} = 1.1340 \times 10^{-5}, \quad \widehat{\psi} = 0.841, \quad \widehat{\phi} = 0.147. \qquad (5.3)$$

Note that $\widehat{\psi} + \widehat{\phi} = 0.988$ is very near 1, suggesting high volatility persistence. In comparison, for the change-point AR(1)-GARCH(1,1) model (5.1), $\widehat{a} + \widehat{b} = 0.02$. We use the method in Sections 2.2 and 2.3 with $L = 30$ and $m = 10$, $M = 20$ to estimate (μ_t, α_t, ν_t) in model (5.1) based on observations up to time t. Figure 2 plots the time-varying estimates of $\widehat{\mu}_t$ (top) $\widehat{\alpha}_t$ (middle), and unconditional volatilities $\widehat{\nu}_t$ (bottom), respectively.

We use the segmentation procedure in Section 3 to locate structural breaks in a frequentist framework of change-points. The procedure yields six change-points and segments the time series into seven segments given in the first column of Table 2. As noted in Section 1, the high persistence in GARCH models might be spurious in the presence of structural changes. The second column in Table 2 shows that the fitted value of $\phi + \psi$ for model (5.2) in each segmented period is markedly less than 1. Table 2 also compares the average of the estimated unconditional volatilities $\widehat{\nu}_t$ over t in model (5.1) with $\widetilde{\nu} = \{\widetilde{\omega}/(1 - \widetilde{\psi} - \widetilde{\phi})\}^{1/2}$ for model (5.2) in different segments; we use $(\widetilde{\omega}, \widetilde{\psi}, \widetilde{\phi}, \widetilde{\alpha}, \widetilde{\mu})$ to denote the MLE for a segment to distinguish it from the estimate (5.3) for the entire time series. The estimated average volatility in (5.1) is quite similar to its counterpart $\widetilde{\nu}$ in (5.2) for each segment, and has the same pattern as that shown in Figure

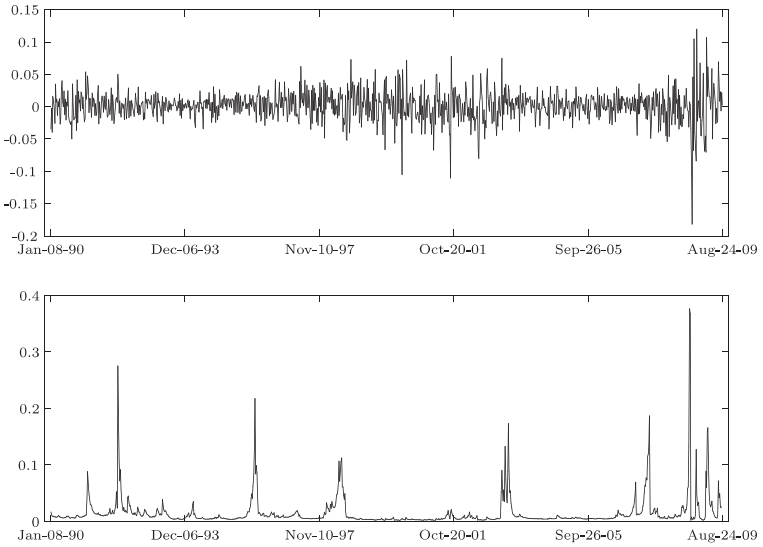

Figure 1. Weekly returns of SP500 index (top) and the posterior probability (bottom).

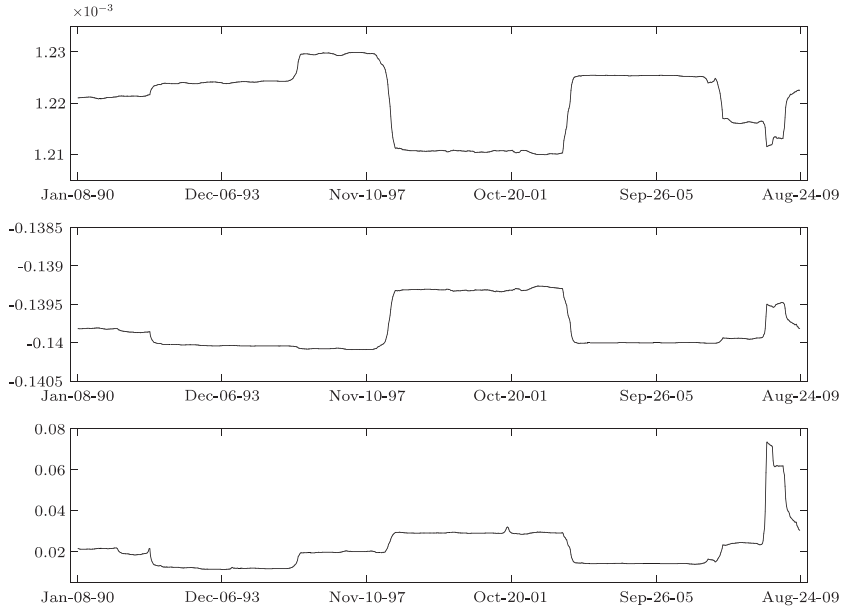

Figure 2. Estimates of time-varying intercepts μ_t (top), AR coefficients α_t (middle) and unconditional volatilities ν_t (bottom).

3. In particular, the periods Jan 1991 – Jan 1996 and June 2003 – June 2007 have very low weekly volatilities ($\leq 1.4\%$), while the weekly volatilities of the

Table 2. Unconditional volatilities in segmented periods.

Period	$\widehat{\phi}+\widehat{\psi}$	Ave. $\widehat{\nu}_t$ in (4.1)	$\widetilde{\nu}$ in (4.2)
Aug/04/90–Dec/30/91	0.5992	0.0204	0.0208
Jan/06/91–Jan/02/96	0.5611	0.0115	0.0115
Jan/08/96–Jul/06/98	0.8242	0.0191	0.0190
Jul/13/98–Jun/09/03	0.1434	0.0287	0.0289
Jun/16/03–Jun/18/07	0.7112	0.0140	0.0140
Jun/25/07–Sep/29/08	0.6445	0.0291	0.0344
Oct/06/08–Aug/24/09	0.8211	0.0454	0.0537

other periods are substantially higher, ranging between 2% and 5%. The high-volatility segments are associated with tumultuous economic events in the U.S. economy. For example, during the period July 1998 – June 2003, the Russian default occurred in the summer of 1998, Brazil currency depreciation occurred in late 1998, the internet bubble burst occurred in 2001, and the 9/11 terrorist attacks and the March 2003 Iraq war caused great anxiety in the U.S. stock market. The period June 2007 – Sept 2008 witnessed the subprime mortgage meltdown and the collapse of Bear Stearns and Lehman-Brothers, and Oct 2008 – Aug 2009 witnessed the Great Recession.

6. Discussion

The idea of representing the GARCH model by $\nu_t\sqrt{h_t}\epsilon_t$, in which ν_t^2 is the unconditional variance and h_t follows the GARCH dynamics in (2.2), has also been used by Engle and Rangel (2008) in their spline-GARCH model that uses a deterministic function of time and exogenous variables to model ν_t by $\log\nu_t = \boldsymbol{\beta}^T\mathbf{x}_t + \phi_0 t + \sum_{i=1}^I \phi_i(t-t_i)_+^2$. We use a piecewise constant function to model ν_t instead, and relate the exogenous variables $\mathbf{x}_t$ to y_t via the regression model (2.1), allowing contemporaneous jumps in the regression coefficients and the unconditional variances. As in Lai, Liu, and Xing (2005) who consider the special case $h_t \equiv 1$ and $\mathbf{x}_t = (y_{t-1},\ldots,y_{t-\kappa})^T$, our stochastic model for jumps in $(\boldsymbol{\beta}_t,\nu_t)$ involves linear Bayes methods and conjugate priors, yielding BCMIX approximations to the Bayes estimate of $(\boldsymbol{\beta}_t,\nu_t^2)$. The BCMIX approximations also provide estimates of the hyperparameters in the Bayesian model with relatively low computational complexity, yielding EB estimates of the piecewise constant parameters that are efficient from both computational and statistical viewpoints.

We have shown in Section 4 how the computationally attractive EB estimates can be used to address the challenging frequentist problem of segmentation. The empirical study in Section 4 shows that segmenting the data can remove the spurious long memory in volatility exhibited by fitting the AR-GARCH model to the entire time series without incorporating possible parameter changes during a

long period that undergoes several structural changes. The apparent long memory arises from the (long) time-scale for parameter changes. The segments are more general than the "regimes" in regime-switching volatility models (which are HMMs) reviewed in Section 1, in which difficulties in estimating the hyperparameters are noted. To address these difficulties, Gray (1996, pp.35-36) modifies the usual regime-switching GARCH model by aggregating the conditional variances from different regimes at each time step. In our segmentation approach, the GARCH parameters are separately estimated for different segments, as in Table 2. However, to determine the segments using the EB estimates, the Bayesian model assumes changes only in the unconditional variance ν_t^2 but not in the GARCH parameters $a_1, \ldots, a_k, b_1, \ldots, b_{k'}$. Not only does this model circumvent the computational difficulties of regime-switching GARCH (or even ARCH) models noted by Cai (1994) and Gray (1996), but it also captures the short-run dynamics of the conditional variance and the structural changes of the long-run volatility. Although not allowing the GARCH parameters to change over time may appear too restrictive, we can in fact estimate them and the other hyperparameter p from moving windows of current and past data, instead of from the entire past history as in (2.5), thereby implicitly allowing these hyperparameters to change slowly over time.

To estimate the magnitude and assess the significance of volatility dynamics and jump risk premia in option pricing, contemporaneous jumps in prices and in volatility have been incorporated into dynamic models of asset prices in the finance literature; see Broadie, Chernov, and Johannes (2007) for a review and discussion in support of contemporaneous jumps in both price and volatility. In particular, Duffie, Pan, and Singleton (2000) introduced a continuous-time stochastic volatility (SV) model that incorporates contemporaneous jumps (CJ) in returns and volatility, and developed analytic methods for pricing under this SVCJ model, which has since become very popular in the finance literature. However, parameter estimation and empirical analysis of the SVCJ model has been a challenging problem. Eraker, Johannes, and Polson (2003) developed a simulation-based Bayes estimator, using MCMC methods to estimate both the hidden states and the model parameters after discretizing the continuous-time bivariate process of returns and their volatilities into an HMM. In contrast, the stochastic change-point AR-GARCH model proposed herein is much simpler to implement, and offers a promising alternative to the SVCJ model. Moreover, it can easily incorporate exogenous covariates, as we have shown in the more general ARX setting.

Acknowledgement

This research is supported by the National Science Foundation under grants DMS-1106535 at Stanford University and DMS-0906593 and DMS-1206321 at SUNY at Stony Brook. We are grateful for remarks from Robert Engle, Peter Phillips, Junshan Bai, Donald Andrews and Edward Vytlacil.

References

Albert, J. H. and Chib, S. (1993). Bayes inference via Gibbs sampling of autoregressive time series subject to Markov mean and variance shifts. *J. Bus. Econ. Statist.* **11**, 1-15.

Andrews, D. W. K., Lee, I. and Ploberger, W. (1996). Optimal changepoint tests for normal linear regression. *J. Econometrics* **70**, 9-36.

Aue, A. and Lee, T. C. M. (2011). On image segmentation using information theoretic criteria. *Ann. Statist.* **39**, 2912-2935.

Bai, J., Lumsdaine, R. and Stock, J. H. (1998). Testing for and dating common breaks in multivariate time series. *Rev. Econ. Stud.* **65**, 395-432.

Bai, J. and Perron, P. (1998). Testing for and estimation of multiple structural changes. *Econometrica* **66**, 817-858.

Baillie, R. T., Bollerslev, T. and Mikkelsen, H. O. (1996). Fractionally integrated generalized autoregressive conditional heteroskedasticity. *J. Econometrics* **74**, 3-30.

Berkes, I., Horváth, L. and Kokoszka, P. S. (2003). GARCH processes: Structure and estimation. *Bernoulli* **9**, 201-227.

Bollerslev, T. (1986). Generalized autoregressive conditional heteroskedasticity. *J. Econometrics* **31**, 307-327.

Broadie, M., Chernov, M. and Johannes, M. (2007). Model specification and risk premia: Evidence from futures options. *J. Finance* **62**, 1453-1490.

Cai, J. (1994). A Markov model of switching-regime ARCH. *J. Bus. Econ. Statist.* **12**, 309-316.

Davis, R., Lee, T. C. M., and Rodriguez-Yam, G. A. (2006). Structural break estimation for nonstationary time series models. *J. Amer. Statist. Assoc.* **101**, 223-239.

Diebold, F. X. (1986). Comment on modeling the persistence of conditional variance. *Econometric Rev.* **5**, 51-56.

Duffie, D., Pan, J., and Singleton, K. (2000). Transform analysis and asset pricing for affine jump-diffusions. *Econometrica* **68**, 1343-1376.

Engle, R. F. (1982). Autoregressive conditional heteroskedasticity with estimates of variance of United Kingdom inflation. *Econometrica* **50**, 987-1008.

Engle, R. F. and Bollerslev, T. (1986). Modeling the persistence of conditional variances. *Econometric Rev.* **5**, 1-50.

Engle, R. F. and Mezrich, J. (1996). GARCH for groups. *Risk*, 36-40.

Engle, R. F. and Rangel, J. G. (2008). The spline-GARCH models for low-frequency volatility and its global macroeconomic causes. *Rev. Finan. Stud.* **21**, 1187-1222.

Eraker, B., Johannes, M. S., and Polson, N. G. (2003). The impact of jumps in returns and volatility. *J. Finance* **53**, 1269-1300.

Galeano, P. and Tsay, R. S. (2010). Shifts in individual parameters of a GARCH model. *J. Financ. Econometrics* **8**, 122-153.

Goldfeld, S. M. and Quandt, R. E. (1973). A Markov model for switching regressions. *J. Econometrics* **1**, 3-16.

Gray, S. F. (1996). Modeling the conditional distribution of interest rates as a regime-switching process. *J Finan. Econ.* **42**, 27-62.

Hamilton, J. D. (1989). A new approach to the economic analysis of nonstationary time series and the business cycle. *Econometrica* **57**, 357-384.

Hamilton, J. D. and Susmel, R. (1994). A conditional heteroskedasticity and change in regime. *J. Econometrics* **64**, 307-333.

Hillebrand, E. (2005). Neglecting parameter changes in GARCH models. *J. Econometrics* **129**, 121-138.

Lamoureux, C. G. and Lastrapes, W. D. (1990). Persistence in variance, structural change and the GARCH model. *J. Bus. Econ. Statist.* **8**, 225-234.

Lai, T. L., Liu, H., and Xing, H. (2005). Autoregressive models with piecewise constant volatility and regression parameters. *Statist. Sinica* **15**, 279-301.

Lai, T. L. and Xing, H. (2008). *Statistical Models and Methods for Financial Markets*. Springer, New York.

Lai, T. L. and Xing, H. (2011). A simple Bayesian approach to multiple change-points. *Statist. Sinica* **21**, 539-569.

McCulloch, R. E. and Tsay, R. S. (1993). Bayesian inference and prediction for mean and variance shifts in autoregressive time series. *J. Amer. Statist. Assoc.* **88**, 968-978.

Mikosch, T. and Starica, C. (2004). Nonstationarities in financial time series, the long-range dependence, and the IGARCH effects. *Rev. Econ. Stat.* **86**, 378-390.

Perron, P. (1989). The great crash, the oil price shock and the unit root hypothesis. *Econometrica* **57**, 1361-1401.

Qu, Z. and Perron, P. (2007). Estimating and testing structural changes in multivariate regressions. *Econometrica* **75**, 459-502.

Quandt, R. E. (1958). The estimation of the parameters of a linear regression system obeying two separate regimes. *J. Amer. Statist. Assoc.* **53**, 873-880.

Quandt, R. E. (1960). Tests of the hypothesis that a linear regression system obeys two separate regimes. *J. Amer. Statist. Assoc.* **55**, 324-330.

Siegmund, D. (2001). Note on a stochastic recursion. In *State of the Art in Probability and Statistics: A Festschrift for Willem R. van Zwet*, 547-554. Institute of Mathematical Statistics Lecture Notes.

Siegmund, D. (2004). Model selection in irregular problems: Applications to mapping quantitative trait loci. *Biometrika* **91**, 785-800.

So, M. K. P., Lam, K. and Li, W. K. (1998). A stochastic volatility model with Markov switching. *J. Bus. Econ. Statist.* **16**, 244-253.

Spokoiny, V. (2009). Multiscale local change point detection with applications to Value-at-Risk. *Ann. Statist.* **37**, 1405-1436.

Wang, J. and Zivot, E. (2000). A Bayesian time series model of multiple structural changes in level, trend, and variance. *J. Bus. Econ. Statist.* **18**, 374-386.

Department of Statistics, Stanford University, Stanford, CA 94305-4065, U.S.A.

E-mail: lait@stanford.edu

Department of Applied Mathematics and Statistics, State University of New York at Stony Brook, Stony Brook, NY 11794, U.S.A.

E-mail: xing@ams.sunysb.edu

(Received July 2012; accepted January 2013)