A Bayesian Semiparametric Multiplicative Error Model with an Application to Realized Volatility

Reza Solgi

(reza.solgi@usi.ch)

Antonietta Mira

(antonietta.mira@usi.ch)

Swiss Finance Institute, University of Lugano, Switzerland.

November 6, 2012

Abstract

A semiparametric multiplicative error model (MEM) is proposed. In traditional MEM, the innovations are typically assumed to be Gamma distributed (with one free parameter that ensures unit mean of the innovations and thus identifiability of the model), however empirical investigations unveils the inappropriateness of this choice. In the proposed approach, the conditional mean of the time series is modeled parametrically, while we model its conditional distribution nonparametrically by Dirichlet process mixture of Gamma distributions. Bayesian inference is performed using Markov chain Monte Carlo simulation. This model is applied to the time series of daily realized volatility of some indices, and is compared to similar parametric models available in the literature. Our simulations and empirical studies show better predictive performance, flexibility and robustness to mis-specification of our Bayesian semiparametric approach.

Keywords: Dirichlet process mixture model, multiplicative error model, slice sampler, realized volatility, parameter expansion.

1 Introduction

In the context of financial time series analysis, one of the interesting problems is to model nonnegative persistence time series. For instance, durations between the transactions in the financial markets, number or volume of transactions within a fixed time interval, absolute returns, the low/high range of price of assets within an interval, bid/ask spread, estimators of integrated volatility such as realized volatility, etc. A common feature of these times series, besides their persistence, is that they are, by definition, nonnegative and might touch the zero lower bound or assume, with positive probability, values arbitrarily close to zero. As discussed in Engle, 2002, there are two naïve approaches to this problem: The first is to employ an additive model (a zero mean noise added to the conditional mean of the process μ_t),

and the second is to model their logarithm. As pointed out in Engle, 2002, in order to let the model assign positive probabilities to values arbitrarily close to zero, the support (and the higher moments) of the distribution of the additive noise should change through time, and must be discontinue at $-\mu_t$. This results in difficult estimation of these models. The second approach, in turn, has its own pitfalls: The presence of zero observations makes the log-transformation infeasible (and therefore the zeros should be substituted by arbitrary small values), and the presence of very small observations (very large negative values in the log-transformation) influences the estimators dramatically. As a consequence of the above considerations, an appropriate treatment of these processes should consider their nonnegativity in the model construction, and accommodate the presence of zero values in small samples with positive probability.

The dominant structure of the models that have been proposed for this class of financial time series has a multiplicative form. For studying the durations between the transaction in financial markets, Engle and Russell, 1998, introduced the Autoregressive Conditional Duration (ACD) model. More precisely for the nonnegative process x_t we have,

$$\mathbb{E}(x_t|\mathcal{F}_{t-1}) \quad = \quad \mu_t(\mathcal{F}_{t-1};\theta) =: \mu_t > 0$$

where $\mathcal{F}_{t-1}$ is the information set available at time $t-1$, θ is the vector of parameters, and $x_t = \mu_t\varepsilon_t$, where ε_t's are i.i.d. from a random variable indexed by the parameters ϕ, and with support $[0, +\infty)$. In order to ensure the identifiability of the model, we should restrict ε_t to have unit mean. As a consequence, for all t, and for any arbitrarily small $\delta > 0$, we have $\mathbb{P}(x_t < \delta|\mathbf{x}^{t-1}) = F_\varepsilon(\delta/\mu_t) > 0$ (where $F_\varepsilon(\cdot)$ is the cdf of ε). The ACD model originally proposed for the duration, was generalized by Engle 2002 for a wider set of applications and renamed Multiplicative

Error Model (MEM). Chou, 2005, applied MEM for modeling the low/high range of prices in daily and weekly scales, and used a Weibull distribution for the innovations. Since then, several extensions of MEM have appeared in the literature. Extensions include generalization of the univariate model to multivariate settings and flexible structure for the conditional mean equation, more precisely modeling the conditional distribution of the process through generalizations of the innovation's distribution to more flexible families.

In this paper we propose a Bayesian semiparametric MEM in which the distribution of the innovations is modeled nonparametrically. In particular we use a countable infinite mixture of Gamma distributions with two free parameters. Using this rich family allows us to approximate any continuous distribution on the positive axis to any precision level. The advantage of the Bayesian approach is that using a mixture distribution provides a very flexible model for the innovations, without the need to fix the number of components of the mixture *a priori*. Inference is performed by Markov chain Monte Carlo (MCMC) simulation using a sampling scheme based on slice sampler for mixture models (Kalli et al 2011). A final contribution of this paper is to present a new modification of the sampling scheme that improves the mixing of the MCMC dramatically.

The rest of this paper is organized as follows: In Section 2 the MEM is introduced and our semiparametric extension is presented. In the following section the sampling scheme of the MCMC for conducting Bayesian inference on the proposed model is presented. The model and the sampling scheme are evaluated, via a simulation study, in Section 4, and in the following section the model is fitted to real financial time series. The last section is devoted to the conclusions.

2 The Model

In this section the MEM formulation is explained. In the first part we focus on the parametric MEM that has been developed in the literature in the recent decade. Our semiparametric extension is then proposed.

2.1 The Multiplicative Error Model

In MEM, as its name suggests, the stochastic process x_t is constructed by multiplying the innovation term ϵ_t, by the conditional mean of the process $\mu_t := \mathbb{E}(x_t|\mathcal{F}_{t-1})$, where $\mathcal{F}_{t-1}$ is the information set available at time $t-1$. In other words, for the discrete time stochastic process, $\{x_t\}_{t=1}^{+\infty}$, we have:

$$x_t = \mu_t(\theta)\varepsilon_t$$

where, for any t, $\mu_t(\theta)$ is a nonnegative process, measurable with respect to the sigma algebra $\mathcal{F}_{t-1}$ (where θ is the vector of parameters to be estimated), and ε_t is restricted to have conditionally unit mean:

$$\mathbb{E}(\varepsilon_t|\mathcal{F}_{t-1}) = 1$$

The unit mean constrain is necessary to ensure identifiability of the model. In most applications ε_ts are i.i.d. draws from a unit mean distribution, i.e. x_ts, conditional on $\mathcal{F}_{t-1}$, are draws from a scale-family of distributions, in which the scale parameter evolves in time according to μ_t, and the *shape* of the distribution remains unchanged. But in principle, as long as the conditional unit mean constraint is honored, the *shape* of this distribution can be seen as a function of the elements of the information set $\mathcal{F}_{t-1}$, and may change through time. For instance Drost and Werker 2004, argue that the i.i.d. assumption for the innovations, is too strong, and they let this distri-

bution depend on $\mathcal{H}_{t-1} \subset \mathcal{F}_{t-1}$.

In the base MEM, μ_t is formulated as a linear combination of the first p and q lagged x_ts and μ_ts, respectively:

$$\mu_t = \omega + \sum_{j=1}^{p} \alpha_j x_{t-j} + \sum_{i=1}^{q} \beta_i \mu_{t-i}$$

With this structure, the persistence property of x_t can be modeled parsimoniously. It is well known that this model for μ_t is equivalent to an ARMA$(\max(p,q),q)$ model for x_t, in which $\eta_t = x_t - \mu_t$ (that is a Martingale difference sequence) plays the role of the innovations. With $p = q = 1$ we obtain the base MEM(1, 1) (from now on MEM for brevity):

$$\mu_t = \omega + \alpha x_{t-1} + \beta \mu_{t-1} \tag{1}$$

that is usually sufficient in empirical studies.

Several generalization of the base MEM have been proposed. In the asymmetric MEM, the conditional mean reacts asymmetrically to the sign of some elements of the information set. For instance, it has been shown that the conditional mean of realized volatility of an equity, reacts asymmetrically in response to positive and negative returns. It is also possible to include other $\mathcal{F}_{t-1}$-measurable variables $z_{t-1}^{(k)}$, in the conditional mean equation:

$$\mu_t = \omega + \sum_{i=1}^{p} \alpha_i x_{t-i} + \sum_{j=1}^{q} \beta_j \mu_{t-j} + \sum_{k=1}^{r} \gamma_k z_{t-1}^{(k)}$$

In our empirical analysis, we consider the following asymmetric MEM (AMEM from now on):

$$\mu_t = \omega + \alpha x_{t-1} + \beta \mu_{t-1} + \gamma |r_{t-1}| \mathbb{I}(r_{t-1} < 0) \tag{2}$$

where $\mathbb{I}(\cdot)$ is the indicator function, and r_t is the daily return.

To model the intra daily volume forecasting, Brownlees, Cipollini and Gallo, 2009, developed the component MEM (CMEM). In this model the conditional mean is broken down into three multiplicative components which model the daily dynamic, intra daily periodic, and intra daily dynamic of the conditional mean. Obviously the base MEM family is not rich enough to reproduce such a dynamic in the conditional mean. In another extension, the conditional mean has a composite structure, since it is modeled as the summation of a time-varying level and an additive stationary noise (for example see Brownlees, Cipollini and Gallo 2012 for its univariate, and Cipollini and Gallo 2012 for its multivariate extension).

In parametric MEMs, the common choices for the distribution of innovations have been Weibull (Engle and Russell 1998, Chou 2005), Gamma (Engle and Gallo 2006), log-Normal and inverse Gamma. In order to model the non trivial fraction of zeros in high-frequency cumulated trading volumes, Hautsch, Malec, Schienle 2010, proposed a mixture of a point probability mass at $\varepsilon = 0$ and a continuous distribution on the positive real line. A univariate mixture MEM is proposed in Lanne 2006, later generalized to a bivariate version by Ahoniemi and Lanne 2009, and applied to the put and call implied volatilities. Although these papers show that using a mixture of two components improves the forecast performance of the model, fixing a priori the number of components of the mixture seems restrictive. Ahoniemi and Lanne 2011, generalize the mixture MEMs allowing the mixing probabilities to change through time. In the next subsection we illustrate the proposed framework where the distribution of the innovations in modeled nonparametrically by a Dirichlet process mixture, with Gamma as the kernel distribution.

2.2 The Semiparametric Multiplicative Error Model

A common technique for modeling a complex distribution is to consider it as a mixture of simpler distributions. From a Bayesian perspective, a finite mixture model with K components can be formulated as follows:

$$\begin{aligned} \varepsilon_t \mid d_t, \boldsymbol{\phi} &\sim F(\phi_{d_t}) \\ d_t \mid \mathbf{p} &\sim \text{Discrete}(p_1, ..., p_K) \end{aligned}$$

where $\boldsymbol{\phi} = (\phi_1, ..., \phi_K)$, $\mathbf{p} = (p_1, ..., p_K)$ and d_t are categorical variables that determine to which mixture component the observation ε_t belongs. In order to fully specify this model in a Bayesian setting, we should assign priors to ϕ_d and $\mathbf{p}$:

$$\begin{aligned} \phi_d &\sim G_0 \\ \mathbf{p} &\sim \text{Dirichlet}(\alpha/K, ..., \alpha/K) \end{aligned}$$

where G_0 is a distribution on the parameter space of F, and Dirichlet$(\alpha/K, ..., \alpha/K)$ is the Dirichlet distribution on the K-dimensional simplex. There are two important problems with the finite component mixtures: It is usually difficult to determine the number of components *a priori*, and they lack the degree of flexibility that is needed in many applications. Dirichlet process mixture (DPM) models, first introduced by Antoniak 1974, can be seen as the limit of the finite mixture model specified above, when K approaches infinity.

The main building block of DPM, is the Dirichlet process (DP, Ferguson 1973), which, in the Bayesian nonparametric framework, provides a prior on the space of distributions. Assume G_0 is a probability distribution on the measurable space Θ, and let α be a positive real number. A random probability distribution G is a draw from the DP with parameters G_0 and α, DP(G_0, α), if, for any finite measurable

partition $(\Theta_1, ..., \Theta_n)$ of Θ, we have:

$$(G(\Theta_1), ..., G(\Theta_n)) \quad \sim \quad \text{Dirichlet}(\alpha G_0(\Theta_1), ..., \alpha G_0(\Theta_1))$$

The probability distribution G_0, and the scalar α are called the centering (or base) measure, and the concentration parameter, respectively. From the definition it follows that, for any measurable set A, $G(A)$ has a Beta distribution:

$$G(A) \quad \sim \quad \text{Beta}(\alpha G_0(A), \alpha G_0(A^c))$$

As a consequence $\mathbb{E}[G(A)] = G_0(A)$ and $\mathbb{V}[G(A)] = G_0(A)[1 - G_0(A)]/(\alpha + 1)$. In other words, the random probability distribution G is centered at the base measure G_0, and its dispersion is inversely proportional to the concentration parameter α. Moreover, using the normalized gamma process representation of DP, it can be shown that G is, almost surely, a discrete distribution.

The definition of DP given above is non-constructive, while the stick-breaking representation of DP (Sethuraman 1994), provides a constructive definition. Assume $v_j \overset{iid}{\sim} \text{Beta}(1, \alpha)$, and define the stick-breaking weights $w_j = v_j \prod_{k<j}(1 - v_k)$. Moreover assume $\phi_j \overset{iid}{\sim} G_0$ (and independent from v_j), and define the random probability distribution G:

$$G(\cdot) \quad = \quad \sum_{i=1}^{+\infty} w_j \delta_{\phi_j}(\cdot)$$

where δ_{ϕ_j} is the point mass distribution with unit mass at ϕ_j. The weights w_j could be thought as the result of breaking a stick of unit size in countably infinite parts. In the first step, the stick is broken in two parts with length v_1 and $1 - v_1$, and the weight of the first component in the infinite mixture is set to be equal to the first part of the stick: $w_1 := v_1$. Then, the remaining part (having length $1 - v_1$)

is broken in two further fragments with proportions v_2 and $1 - v_2$, so that the two fragments have length $(1-v_1)v_2$ and $(1-v_1)(1-v_2)$ respectively. We set the weight of the second component in the infinite mixture equal to the length of the first one (of these two additional fragments), $w_2 := (1 - v_1)v_2$. This procedure is continued infinitely many times in order to determine all the weights. It can be shown that G, constructed as detailed above, is distributed according to a DP with centering measure G_0 and concentration parameter α (Sethuraman 1994). In order to represent the weight process $\mathbf{w} = (w_1, w_2, ...)$, we use the notation $\mathbf{w} \sim \text{GEM}(\alpha)$, where GEM stands for Griffiths, Engen and McCloskey (Piman 2002).

As mentioned earlier, samples from a DP are almost surely discrete, which is not a desirable fact when modeling nonparametrically continuous distributions. To handle this problem a hierarchical model can be employed. We assume the observations ε_t are conditionally independent from a parametric family of distributions, parameterized by the vector, $\theta_t \sim G$: $\varepsilon_t|\theta_t \sim k(\cdot|\theta_t)$. If we put a DP prior on G, $G \sim \text{DP}(\alpha, G_0)$, the result is called a DPM (Antoniak 1974 and Lo 1984). In other words, a DPM is the result of nonparametric DP mixing of a parametric family of distributions:

$$\begin{aligned} f_\varepsilon(\cdot) &= \int k(\cdot|\theta) dG(\theta) \\ G &\sim \text{DP}(\alpha, G_0) \end{aligned}$$

The stick breaking representation of DP implies that:

$$\begin{aligned} f_\varepsilon(\cdot) &= \sum_{j=1}^{+\infty} w_j \; k(\cdot|\theta_j) \\ \mathbf{w} &\sim \text{GEM}(\alpha) \\ \theta_j &\overset{iid}{\sim} G_0 \end{aligned}$$

A computationally convenient choice for k and G_0 could be a member of exponential family and its conjugate prior distribution, respectively. It can be shown that DPM is the limiting model of a K-component mixture model, when $K \to +\infty$ (Neal 2000). Therefore this nonparametric model bypasses the issue of choosing *the correct number* of components in a finite mixture model.

In the proposed semiparametric MEM we suggest to model the innovations by a DPM. Since in this model the innovations are nonnegative, we should use a suitable kernel distribution. Our suggestion is to use the Gamma distribution. As mentioned earlier, in parametric MEMs, the distribution of the innovations is usually restricted to have unit mean. At first glance it seems thus natural to use the Gamma distribution with unit mean as the kernel distribution of our DPM:

$$\begin{aligned} f_\varepsilon(\varepsilon) &= \int k(\varepsilon|\phi) dG(\phi) \\ G &\sim \mathrm{DP}(\alpha, G_0) \\ k(\varepsilon|\phi) &= \frac{\phi^\phi}{\Gamma(\phi)} \varepsilon^{\phi-1} \exp(-\phi\varepsilon) \end{aligned}$$

Since the kernel $k(\varepsilon|\phi)$, by construction, has unit mean for any $\phi > 0$, also the DPM $f_\varepsilon(\varepsilon)$ has unit mean. However extending the unit mean constraint to all the mixture components may be too bounding for some applications. In fact it can be easily shown that the random distribution $f_\varepsilon(\varepsilon)$ does not range over all distributions on the positive axis. Since this kernel has only one free parameter, introducing components with thicker tails in the mixture (components with smaller ϕ), will increase, at the same time, the probability of the neighborhood around zero; Hence in presence of fat tail innovations in the data, while this DPM attempts to assign higher weights to the components with smaller ϕ, it will, at the same time, increase the likelihood of the innovations close to zero. As a consequence, this model, *de facto*, is not nonparametric, because it does not range over all potentially *true* distributions

on the positive axis. This is the first nonparametric model for the distribution of innovations that will be studied in this paper (since this is a DPM of Gammas with one free parameter, the resulting models will be called DPMG1-MEM and DPMG1-AMEM, when it is combined with Eq. 1 and Eq. 2, respectively). Our empirical investigations demonstrate that DPMG1-MEM and DPMG1-AMEM (although they improve the forecast performance in comparison with their parametric counterparts) are not flexible enough to model the empirical financial data properly, however they might be sufficient for other applications.

In a more flexible view, we replace the kernel with a Gamma distribution with two free parameters:

$$k(\varepsilon|\phi,m) \quad = \quad \frac{\phi^{\phi}}{m^{\phi}\Gamma(\phi)}\varepsilon^{\phi-1}\exp(-\frac{\phi}{m}\varepsilon)$$

where ϕ is the shape parameter and m is the expected value of the kernel. Wu and Goshal, 2008, have demonstrated the Kullback-Leibler property (positivity of the prior probability in a Kullback-Leibler neighborhood of the true density) of the mixture of Gamma distribution, assuming very mild conditions for the true distribution.

The stick breaking representation of the model is:

$$f_{\varepsilon}(\varepsilon) \quad = \quad \sum_{j=1}^{+\infty} w_j k(\varepsilon|\phi_j,m_j)$$

where $\mathbf{w} \sim \text{GEM}(\alpha)$. By this definition, clearly $f_{\varepsilon}(\varepsilon)$ does not have unit mean:

$$\bar{m} := \mathbb{E}_{f_{\varepsilon}}(\varepsilon) = \sum_{j=1}^{+\infty} w_j m_j \neq 1$$

A potential solution to this unidentifiability issue, is to restrict the constant term of the conditional mean to unity, in other words for the base MEM(1, 1) we would

have:

$$\mu_t^* = 1 + \alpha^* x_{t-1} + \beta \mu_{t-1}^* \tag{3}$$

and similarly for the Asymmetric-MEM(1, 1) we should use:

$$\mu_t^* = 1 + \alpha^* x_{t-1} + \beta \mu_{t-1}^* + \mathbb{I}(r_{t-1} < 0)\gamma^* |r_{t-1}|$$

(where in this new parametrization, ω is the mean of the innovations, $\mu_t^* = \mu_t/\omega$, $\alpha^* = \alpha/\omega$ and $\gamma^* = \gamma/\omega$.) In theory this will solve the problem, and a Metropolis-within-Gibbs can be designed for iteratively sampling the parameters of μ_t and f_ε. However this parametrization will cause slow mixing of the MCMC simulation as explained below. In a parametric MEM, ω is strongly correlated with α and β and in the proposed semiparametric model, this correlation translates into a strong correlation among $(w_1, m_1, w_2, m_2, ...)$ and (α^*, β) which are sampled in different stages of the Gibbs sampler, thus resulting in slow mixing of the Markov chain.

As an alternative the support of the random distribution of the innovations may be restricted to have unit mean by modifying the priors. More specifically, the distribution of the innovations could be specified as the follows:

$$g_\varepsilon(\varepsilon) = \sum_{j=1}^{+\infty} w_j k(\varepsilon | \phi_j, m_j/\bar{m})$$

which, by construction, ensures $\mathbb{E}_{g_\varepsilon}(\varepsilon) = 1$. Combining this model for the distribution of innovations with Eq. 1 and Eq. 2, results in two models that will be called DPMG2-MEM and DPMG2-AMEM, respectively. Unfortunately direct sampling of these models by the sampling schemes available in the literature is not possible, since the kernel of each of the components of the mixture depends on all w_j and m_j, and this makes the sampling scheme very complicated (if not impossible).

In this paper, we propose to use the unconstrained DPM for the distribution of the innovations and unconstrained conditional mean equation:

$$\begin{aligned} f_\varepsilon(\varepsilon) &= \int k(\varepsilon|\theta) dG(\theta) \\ G &\sim \mathrm{DP}(\alpha, G_0) \\ k(\varepsilon|\phi, m) &= \frac{\phi^\phi}{m^\phi \Gamma(\phi)} \varepsilon^{\phi-1} \exp(-\frac{\phi}{m}\varepsilon) \end{aligned}$$

where $\theta = (\phi, m)$. That is a parameter expanded (PX, in the sense of Liu and Wu 1999, van Dyk and Meng 2000 and Liu, Rubin and Wu 1998) version of the model with unit-mean DPM. In a similar manner this model may be considered as the parameter expanded version of the model with constrained mean for the distribution of innovations. Combining this model for the distribution of innovations with Eq. 1 and Eq. 2, results in two models that will be called PX-DPMG2-MEM and PX-DPMG2-AMEM, respectively. The use of proper priors results in proper posteriors for these models (even if the likelihood is improper). Note that a prior on the parameter space of the PX-models induces a prior on the parameters of the original models. We call these priors the *induced or implied priors*. An MCMC simulation can be set up to target the PX-model, at the end of which the sample obtained are post-processed by transforming each sampled model to an *equivalent* model in the family of identifiable models (for instance the family of DPMG2-MEM and DPMG2-AMEM as defined above). Note that, after post-processing the models, the obtained sample is a sample from the posterior of DPMG2-MEM and DPMG2-AMEM (whose prior is the prior induced by the prior on the PX-model). For instance for the PX-DPMG2-MEM, either of these reduction functions that map the sampled model to a model in the family of models with unit-mean distribution of innovations or family

of models with unit mean innovations, may be used:

$$(\omega, \alpha, \beta, \mathbf{w}, \boldsymbol{\phi}, \mathbf{m}) \rightarrow (\bar{m}\omega, \bar{m}\alpha, \beta, \mathbf{w}, \boldsymbol{\phi}, \mathbf{m}/\bar{m}) \quad (4)$$

$$(\omega, \alpha, \beta, \mathbf{w}, \boldsymbol{\phi}, \mathbf{m}) \rightarrow (1, \alpha/\omega, \beta, \mathbf{w}, \boldsymbol{\phi}, \omega\mathbf{m}) \quad (5)$$

Equivalently for PX-DPMG2-AMEM, the following mappings may be used:

$$(\omega, \alpha, \beta, \gamma, \mathbf{w}, \boldsymbol{\phi}, \mathbf{m}) \rightarrow (\bar{m}\omega, \bar{m}\alpha, \beta, \bar{m}\gamma, \mathbf{w}, \boldsymbol{\phi}, \mathbf{m}/\bar{m}) \quad (6)$$

$$(\omega, \alpha, \beta, \gamma, \mathbf{w}, \boldsymbol{\phi}, \mathbf{m}) \rightarrow (1, \alpha/\omega, \beta, \gamma/\omega, \mathbf{w}, \boldsymbol{\phi}, \omega\mathbf{m}) \quad (7)$$

Note that, in order to use the reduction functions (4) and (6), we need also to sample $\bar{m}$, the mean of the DPM, that is an infinite sum. The distribution of the mean of DP and DPM has been the subject of several studies (for instance see Lijoi, Regazzini, 2004, among others), however sampling directly from this distribution is not trivial. In fact even evaluation of the distribution of the mean of a DP (in very simple examples) might be subject to computation of some numerical integrals. Here we propose to approximate the infinite sum of $\bar{m}$ by a finite sum in such a way that the truncated sum of weights is close to 1. In practice, at the tolerance ε, in order to obtain a sample from the mean of the DPM, we need to truncate the DPM at K_ε, where:

$$K_\varepsilon = \inf\{j \in \mathbb{N}; 1 - \sum_{k=1}^{j} w_j < \varepsilon\}$$

(In our simulations we have set $\varepsilon = 10^{-10}$.) In Muliere and Tardella, 1998, it has been shown that,

$$K_\varepsilon - 1 \sim \text{Poisson}(-\alpha \log \varepsilon)$$

Therefore the expected value of the truncation level is proportional to $-\log(\varepsilon)$, so

that, with small and moderate values of precision parameter, extremely accurate results may be obtained in a reasonable computational time.

This strategy (sample from the posterior of the PX-model using an MCMC simulation, and then post-process the path of the Markov chain in order to obtain a sample from the posterior of the original model) has been used also in Gelman 2006. In this paper, in order to facilitate the sampling in a hierarchical model, a PX-version of the original model is considered. Post-processing of a sample drawn from the posterior of the PX-model (via a Gibbs sampler), results in a sample from the posterior of the original hierarchical model.

Although there is a relatively rich literature on Bayesian nonparametric modelling with constraint on the median (see Hanson and Johnson 2002, Burr and Doss 2005, among others), the nonparametric modelling of distributions with moments constraints have been considered only recently. Yang, Dunson and Baird, 2010, estimate a semiparametric latent factor model that involves nonparametric estimation of the distribution of latent factors with constrained moments. Our sampling algorithm resembles theirs, however they approximate the DPM by a finite mixture obtained by truncating the stick breaking representation of DP (Ishwaran and James 2001). This approximation enables them to implement a Gibbs sampler (Ishwaran and Zarepour 2000).

Here we give a straightforward justification of the validity of the applied sampling strategy. Assume the parameters of the original model is $\theta \in \Theta$, and let $l(\theta)$ be the likelihood of this model. Moreover let $\vartheta = (\theta, \alpha)$ be the parameters of the PX-model (where $\alpha \in A$), with likelihood $l^*(\vartheta) = l(R(\theta, \alpha))$, in which $R : \Theta \times A \to \Theta$ is the reduction function (that maps an unidentifiable model to an equivalent model in the family of identifiable models.) Define $A(\theta) = \{(\theta^*, \alpha^*); R(\theta^*, \alpha^*) = \theta\}$, and

let $h^*(\vartheta) = h^*(\theta, \alpha)$ be the prior on the parameter space of the PX-model. We sample from the posterior of the PX-model $\pi^*(\theta, \alpha) = h^*(\theta, \alpha) l(R(\theta, \alpha))$, and them post-process the sample using the reduction function $R(\theta, \alpha)$. The distribution of the obtained sample is

$$\begin{aligned} \pi(\theta) &= \int_{A(\theta)} \pi^*(\theta^*, \alpha^*) d\theta^* d\alpha^* \\ &= \int_{A(\theta)} h^*(\theta^*, \alpha^*) l(R(\theta^*, \alpha^*)) d\theta^* d\alpha^* = l(\theta) \int_{A(\theta)} h^*(\theta^*, \alpha^*) d\theta^* d\alpha^* \end{aligned}$$

Therefore $\pi(\theta)$ is the posterior of the original model, with prior $h(\theta) = \int_{A(\theta)} h^*(\theta^*, \alpha^*) d\theta^* d\alpha^*$, that is the prior induced on the parameter space of the original model by $h^*(\theta, \alpha)$.

3 Bayesian Inference

The analytical intractability of DPM had restricted its application until the mid 90s. By popularization of Markov chain Monte Carlo methods in Bayesian inference, and their particular application to DPM (Escobar 1994, and Escobar and West 1995), these models found several applications in a variety of fields. In Escobar 1994, and Escobar and West 1995, a Pólya urn representation of DPM is employed to design the posterior sampling scheme. The main pitfall of this algorithm is its poor mixing. Improved versions of the algorithm have been presented in Bush and MacEachern, 1996 (the partially collapsed Gibbs sampler), Neal, 1991 and MacEachern,1994 (the fully collapsed Gibbs sampler). These algorithms rely on the conjugacy properties of the model (i.e. the kernel distribution and the centering distribution of the DP need to be conjugate). For dealing with non-conjugate models, West, Müller and Escobar 1994, proposed to use numerical integration to evaluate the integral of the kernel distribution with respect to the centering distribution. Alternative approaches to deal with the non-conjugate models, are the "no-gaps" algorithm of MacEachern and

Müller, 1998, the Metropolis-Hastings updates of the indicator variables, introducing the temporarily auxiliary variables of Neal 2000, and the "split-merge" algorithm of Jain and Neal 2004 and 2007. The above mentioned algorithms are examples of *marginal methods*, because, in principle, they are based on integrating out the random distribution that is an infinite dimensional object.

A parallel family of methods for inference in DPM are the so called *conditional methods* that rely on the stick-breaking representation of the DP. Ishwaran and James 2000 truncated the infinite sum of the stick-breaking representation and proposed to use a blocked Gibbs sampler. Since the weights of the infinite mixture in the stick-breaking representation decays exponentially in expectation, the truncated mixture could provide an acceptable approximation to the exact model. A part from this approximate inference, other exact MCMC sampling schemes based on the stick-breaking representation have been proposed in the recent years. Among them it is worth mentioning the retrospective sampler of Papaspiliopoulos and Roberts 2005, the slice sampler of Walker 2007, its improved version by Papaspiliopoulos 2008 and the independent slice-efficient sampler of Kalli, Griffin and Walker, 2011. Kalli, Walker and Damien 2011 use the latter for inference in a semiparametric GARCH model. Here we briefly explain how the slice-efficient sampler can be adapted to conduct Bayesian inference also in our setting. Following Walker 2007, by augmenting the model with the latent variable u, the joint density of (ε, u) is:

$$f_{\varepsilon,u}(\varepsilon, u) \quad = \quad \sum_{j=1}^{+\infty} I(w_j > u) k(\varepsilon|\theta_j)$$

where, in DPMG1-MEM and DPMG1-AMEM, $\theta_j = \phi_j$ and $k(\cdot|\phi)$ is the unit mean Gamma probability density function with shape parameter ϕ, and, in PX-DPMG2-MEM and PX-DPMG2-AMEM, $\theta_j = (\phi_j, m_j)$ and $k(\cdot|\phi, m)$ is the Gamma probability density function with shape parameter ϕ and mean m. Therefore, given u, the infinite mixture will reduce to a finite mixture (since only a finite number of

weights w_j could be bigger than the given positive real number u). Moreover, by introducing the latent allocation variable d (indicating to which component of the mixture ε belongs), the joint density of (ε, u, d) will be:

$$f_{\varepsilon,u,d}(\varepsilon, u, d) \quad = \quad I(w_d > u)k(\varepsilon|\theta_d)$$

Clearly it is not possible to sample the infinite set of parameters θ_j, however it can be show that, by augmenting the model with the latent variable u, we only need to sample a finite set of these parameters, still guaranteeing that the Markov chain retains the correct target as its stationary distribution.

Based on this result, Walker 2007, presents a Gibbs sampler for inference in stick-breaking mixture models. By augmenting the model with the latent variables u_t and d_t for $1 \leq t \leq T$, the posterior will be:

$$\text{Priors} \ \times \ \prod_{t=1}^{T} \mathbf{1}(u_t < w_{d_t}) \frac{1}{\mu_t} k\left(\frac{x_t}{\mu_t}; \theta_{d_t}\right)$$

where μ_t is the conditional mean in the MEM formulation. In order to improve the efficiency of the slice sampler of Walker 2007, the slice-efficient sampler (Kalli, Griffin and Walker, 2011) was proposed. The Authors suggest to rewrite the joint density of (ε, u, d) as

$$f_{\varepsilon,u,d}(\varepsilon, u, d) \quad = \quad I(\xi_d > u)\frac{w_d}{\xi_d}k(\varepsilon|\theta_d)$$

where ξ_d is an infinite series decreasing in d. Introducing the ξ_d series enhances the sampling efficiency dramatically, mainly because, in the original slice sampler of Walker 2007, u and w are strongly correlated. In principle the ξ_d series could be any decreasing series, however it controls the efficiency and the computational time of the algorithm. Kalli, Griffin, and Walker 2011 found that the mixing depends on the rate of increase of $\mathbb{E}(w_j)/\xi_j$; A higher rate of increase implies a better mixing and,

at the same time, a longer simulation time. In their examples they find that with $\mathbb{E}(w_j)/\xi_j \propto 1.5^j$ an acceptable balance is achieved. Here we set $\xi_j = g(j)$, where $g(j)$ is a deterministic decreasing function of d and $g(d) \propto \mathbb{E}(w_j)/1.5^j$.

By introducing ξ_d, the posterior of our models becomes:

$$\text{Priors} \times \prod_{t=1}^{T} \mathbf{1}(u_t < \xi_{d_t}) \frac{w_{d_t}}{\xi_{dt}} \frac{1}{\mu_t} k\left(\frac{x_t}{\mu_t}; \theta_{d_t}\right)$$

In our MCMC simulations we sample (v_j, θ_j) for $j = 1, 2, ...,$ (d_t, u_t) for $t = 1, ..., T$, and the parameters of the conditional mean's equation. In the case of the PX-DPMG2-MEM and PX-DPMG2-AMEM, we post-process the obtained sample by the transformation (4) and (6), respectively, in order to obtain a sample from the posterior of DPMG2-MEM and DPMG2-AMEM. In the next subsections we show how the parameters of the all four models maybe sampled using a Metropolis-within-Gibbs strategy.

a. Sampling u_t

In all four models, the full conditional of u_t is: $p(u_t|\cdot) \propto \mathbf{1}(u_t < \xi_{d_t})$, therefore, conditionally on the remaining parameters, u_t are uniformly distributed on $(0, \xi_{d_t})$.

b. Sampling v_j

In all four models, the full conditional of v_j is:

$$\begin{aligned} p(v_j|\cdot) &\propto \pi(v_j) \prod_{t;d_t \geq v_j} w_{d_t} \\ &\propto v_j(1-v_j)^M \prod_{t;d_t=j} v_{d_t} \prod_{t;d_t>v_j} (1-v_{d_t}) \\ &= v_j^{1+n_j}(1-v_j)^{1+n'_j} \end{aligned}$$

where $n_j = \sum_{t=1}^{T} \mathbf{1}(d_t = j)$ and $n'_j = \sum_{t=1}^{T} \mathbf{1}(d_t > j)$. Therefore, conditioned on the

rest of the parameters, the v_j's follow a Beta distribution:

$$v_j|\cdot \;\sim\; \text{Beta}(1+n_j, M+n'_j)$$

Note that, $n_j = 0$ for any $j > \bar{d}$, and $n'_j = 0$ for any $j \geq \bar{d}$ where $\bar{d} = \max\{d_t\}$ is a finite integer. This means that the distribution of v_j will be updated if and only if there exists at least one innovation associated with the component k of the mixture where $k \geq j$. Otherwise the full conditional of v_j is equal to the prior distribution. Therefore at this step of the Gibbs sampler we only need to sample a finite number of v_js, namely v_1, ..., $v_{\bar{d}}$.

c. Sampling ϕ_j

In the DPMG1-MEM and DPMG1-AMEM, the full conditional of ϕ is:

$$\begin{aligned} p(\phi_j|\cdot) &\propto p(\phi_j) \;\times \prod_{t;d_t=j} \frac{\phi_j^{\phi_j}}{\Gamma(\phi_j)} \frac{x_t^{\phi_j-1}}{\mu_t^{\phi_j}} \exp\left(-\phi_j \frac{x_t}{\mu_t}\right) \\ &\propto \frac{\phi_j^{n_j\phi_j+a_0-1}}{[\Gamma(\phi_j)]^{n_j}} P_j^{\phi_j-1} \exp\left(-S_j\phi_j - \frac{a_0}{b_0}\phi_j\right) \end{aligned}$$

while the full conditional of ϕ_j in PX-DPMG2-MEM and PX-DPMG2-AMEM is:

$$\begin{aligned} p(\phi_j|\cdot) &\propto p(\phi_j) \;\times \prod_{t;d_t=j} \frac{\phi_j^{\phi_j}}{\Gamma(\phi_j)} \frac{x_t^{\phi_j-1}}{(\mu_t m_j)^{\phi_j}} \exp\left(-\frac{\phi_j}{m_j}\frac{x_t}{\mu_t}\right) \\ &\propto \frac{\phi_j^{n_j\phi_j+a_0-1}}{[\Gamma(\phi_j)]^{n_j} m_j^{n_j\phi_j}} P_j^{\phi_j-1} \exp\left(-\frac{S_j}{m_j}\phi_j - \frac{a_0}{b_0}\phi_j\right) \end{aligned}$$

where $P_j = \prod_{t;d_t=j} \frac{x_t}{\mu_t}$ and $S_j = \sum_{t;d_t=j} \frac{x_t}{\mu_t}$, and a Gamma prior with shape parameter a_0, and mean b_0 is used. We should note that, for the empty components of the infinite mixture (i.e. components with no innovation assigned to them), we have $n_j = 0$, $P_j = 1$, and $S_j = 0$, and this will reduce the full conditional given above to the prior. As a consequence, again we need to sample at most the first $\bar{d}$ elements of

$(\phi_1, \phi_2, ...)$. It can be shown that these two full conditionals, despite not being standard distributions, are log-concave and thus, to sample them, the adaptive rejection sampling of Gilks and Wild 1992, can be used. We prefer instead to implement a Metropolis-Hastings sampler with Gamma proposal distribution whose parameters depend on the current state of the Markov chain as detailed below. The mode, $\tilde{\phi}_j$, of the log-concave distribution of ϕ_j can be found easily and extremely fast. We thus use a Gamma proposal with mode equal to $\tilde{\phi}_j$ and set the decay rate of its logarithm approximately equal to the decay rate of the full conditional of ϕ_j at points $\frac{\tilde{\phi}_j}{\lambda}$ and $\lambda\tilde{\phi}_j$ (for some constant $\lambda > 1$). In particular, the shape and scale parameters of the Gamma approximation to the full conditional ϕ_j are:

$$\begin{aligned} \alpha_q &= \frac{1}{2}\left(\frac{h_1'}{\frac{\lambda}{\tilde{\phi}_j} - \frac{1}{\tilde{\phi}_j}} + \frac{h_2'}{\frac{1}{\lambda\tilde{\phi}_j} - \frac{1}{\tilde{\phi}_j}}\right) + 1 \\ \beta_q &= \frac{\tilde{\phi}_j}{\alpha_q - 1} \end{aligned}$$

where h_1' and h_2' are the first derivative of the logarithm of full conditional of ϕ_j at the points $\frac{\tilde{\phi}_j}{\lambda}$ and $\lambda\tilde{\phi}_j$ respectively (Here we have dropped the j's subscripts for ease of notation). Using this approximation (with $\lambda = 3$) an almost perfect fit of the full conditional is obtained and using this approximation as the proposal distribution of the MH step, an average acceptance rate above 99% is achieved.

d. Sampling m_j

The full conditional of m_j in the PX-DPMG2-MEM and PX-DPMG2-AMEM is:

$$\begin{aligned} p(m_j|\cdot) \propto p(m_j) \times \prod_{t;d_t=j} \frac{\phi_j^{\phi_j}}{(\mu_t m_j)^{\phi_j}\Gamma(\phi_j)} x_t^{\phi_j - 1} \exp(-\frac{\phi_j}{\mu_t m_j} x_t) \\ \propto p(m_j) \frac{1}{m_j^{n_j\phi_j}} \exp(-\frac{\phi_j S_j}{m_j}) \end{aligned}$$

Therefore the conjugate prior of m_j is the Inverse-Gamma distribution:

$$p(m_j) \propto \frac{1}{m_j^{c+1}} \exp(-\frac{d}{m_j})$$

where c and d are the shape and scale parameters, and this implies:

$$m_j|\cdot \sim \text{InvGamma}(n_j\phi_j + c, \phi_j S_j + d)$$

Here as well, only a finite number m_js will be sampled in each sweep of the Gibbs sampler, since the full conditional of the rest of them is equal to their prior.

e. Sampling d_t

In all four models, the full conditional of d_t is:

$$p(d_t = i|\cdot) \propto \mathbf{1}(u_t < \xi_i)\frac{w_i}{\xi_i}k(x_t; \phi_i, \mu_t m_i)$$

where i is a positive integer. Since $\xi_i = g(i)$ is a decreasing series in i, for every $i \geq g^{-1}(u_t)$, we have $\xi_i \leq u_t$, that implies $p(d_t = i|\cdot) = 0$. Consequently, conditioned on all other parameters, d_t takes values on the finite set $\{1, ..., \lceil g^{-1}(u_t)\rceil - 1\}$, and therefore its sampling is trivial.

f. Sampling conditional mean parameters

The full conditional of the parameters of the conditional mean, $\eta = (\omega, \alpha, \beta)$ for MEM and $\eta = (\omega, \alpha, \beta, \gamma)$ for AMEM, is:

$$p(\eta|\cdot) \propto p(\eta) \times \prod_{t=1}^{T} \frac{1}{\mu_t(\eta)} k\left(\frac{x_t}{\mu_t}; \theta_{d_t}\right)$$

which is not a standard distribution. For the prior of η we use an independent

truncated Normal distribution with very large variances:

$$p(\eta) \;=\; 2^d \; \varphi_d(\eta; \mathbf{0}_d, s\mathbf{I}_d) \; \mathbb{I}(\eta \in \mathbb{R}_+^d)$$

where $d = 3$ for MEM and $d = 4$ for AMEM, and $\mathbf{0}_d$, $\mathbf{I}_d$ and φ_d are the vector of zeros, the identity matrix and the pdf of a d-dimensional multivariate Normal distribution, respectively. Moreover s is a large positive constant (in our simulation we set $s = 100$).

To sample η an adaptive version of the Metropolis-Adjusted Langevin algorithm (MALA, Roberts and Tweedie 1996, and Roberts and Rosenthal 1998) is used. Our proposal distribution is:

$$Q(\eta_{k+1}|\cdot) \;=\; \mathcal{N}_d\left(\eta_k + \frac{\lambda^2}{2}\mathbf{\Lambda}_k \nabla \log p(\eta_k|\cdot), \lambda^2 \mathbf{\Lambda}_k\right)$$

In the DPMG1-MEM and DPMG1-AMEM, $\mathbf{\Lambda}_k = \Sigma_k$, where Σ_k is the empirical covariance matrix of $(\eta_1 ..., \eta_k)$. The choice of $\mathbf{\Lambda}_k$ is more delicate in the PX-models. In this case we propose $\mathbf{\Lambda}_k = \mathbf{C}(\bar{m}_k) \circ \Sigma_k$, where for PX-DPMG2-MEM

$$C(\bar{m}_k) \;=\; \begin{bmatrix} \bar{m}_k^{-2} & \bar{m}_k^{-2} & \bar{m}_k^{-1} \\ \bar{m}_k^{-2} & \bar{m}_k^{-2} & \bar{m}_k^{-1} \\ \bar{m}_k^{-1} & \bar{m}_k^{-1} & 1 \end{bmatrix},$$

and for PX-DPMG2-AMEM

$$C(\bar{m}_k) \;=\; \begin{bmatrix} \bar{m}_k^{-2} & \bar{m}_k^{-2} & \bar{m}_k^{-1} & \bar{m}_k^{-2} \\ \bar{m}_k^{-2} & \bar{m}_k^{-2} & \bar{m}_k^{-1} & \bar{m}_k^{-2} \\ \bar{m}_k^{-1} & \bar{m}_k^{-1} & 1 & \bar{m}_k^{-1} \\ \bar{m}_k^{-2} & \bar{m}_k^{-2} & \bar{m}_k^{-1} & \bar{m}_k^{-2} \end{bmatrix},$$

"$\circ$" is the Hadamard operator, and Σ_k is the empirical covariance matrix of $(\tilde{\eta_1}, ..., \tilde{\eta_k})$,

where $\tilde{\eta}_j = (\bar{m}_j\omega_j, \bar{m}_j\alpha_j, \beta_j)$ in PX-DPMG2-MEM, and $\tilde{\eta}_j = (\bar{m}_j\omega_j, \bar{m}_j\alpha_j, \beta_j, \bar{m}_j\gamma_j)$ in PX-DPMG2-AMEM. In our proposal λ is a constant parameter that should be tuned; after few pilot runs we set $\lambda = 1$.

At the k-th iteration, Σ_n changes only by $O(1/k)$, therefore this adaptation mechanism satisfies the diminishing adaptation condition (Roberts and Rosenthal 2007), and thus the correct target distribution is preserved.

In our simulations (for all models) we use $N_0 = 2 \times 10^3$ iterations as the burn-in, and then run the MCMC for $N = 10^4$ iterations. Samples from the posterior of DPMG2-MEM and DPMG2-AMEM are derived by post-processing the samples obtained from the posterior of the PX-DPMG2-MEM and PX-DPMG2-AMEM using the transformations 4 and 6, respectively. The simulation time on a desktop computer with the CPU running at 2.70GHz and 8GB RAM is around 4-5 minutes.

4 Simulation Study

To study the performance of our models and the proposed sampling scheme, the following simulation study has been conducted. For the sake of brevity only the results for the MEM models are reported (similar results are obtained for AMEM). We have generated a sample of 3000 observation from a symmetric MEM with parameters $\omega = 0.4$, $\alpha = 0.3$ and $\beta = 0.65$:

$$\begin{aligned} \mu_t &= \omega + \alpha x_{t-1} + \beta\mu_{t-1} \\ x_t &= \mu_t\varepsilon_t \end{aligned}$$

Table 1: Estimation of the MEM in a simulation study with $\omega = 0.4$, $\alpha = 0.3$ and $\beta = 0.65$. Standard errors are reported in parentheses.

	$\hat{\omega}$	$\hat{\alpha}$	$\hat{\beta}$
Gamma-MEM (MLE Estimates)	0.442 (0.059)	0.286 (0.016)	0.658 (0.019)
DPMG1-MEM	0.437 (0.057)	0.285 (0.014)	0.657 (0.017)
DPMG2-MEM	0.429 (0.058)	0.288 (0.014)	0.658 (0.016)

where the innovations ε_t are i.i.d. draws from a mixture of Gamma and LogNormal distributions:

$$\varepsilon_t \overset{iid}{\sim} p\ \text{Gamma}\left(\phi, 1\right) + (1-p)\ \text{LN}\left(-\sigma^2/2, \sigma^2\right)$$

with $p = 0.7$, $\phi = 15$ and $\sigma = 0.45$, where the logarithm of $\text{LN}\left(a, b^2\right)$ is Normal with mean a and variance b^2. Both components of the mixture are unit mean distributions, so the distribution of the innovations has unit mean by construction. We have found the MLE by assuming a parametric model with Gamma distributed innovations. The MLE along with their standard errors are reported in Table 1 (first row). Figure 1 shows the simulated time series and the QQ-Plot of the empirical quantiles of the estimated innovations (using the parametric MEM with Gamma distribution).

On the same simulated data we also estimate DPMG1-MEM and DPMG2-MEM. In Table 1 the estimates of the parameters of the conditional mean equation are reported (second and third row). The MCMC traces, the posterior distribution of the parameters and the autocorrelation function (ACF) along the chains are presented in Figure 2 and 3. Moreover in Figure 4 we show the posterior distribution of the innovations (1000 samples) along with the true distribution (the thick black curve), and the QQ-Plot of the empirical quantiles of the estimated innovations. As it can be seen in the plot of the posterior distribution of f_ε, the mixture of unit mean Gamma distributions is not able to appropriately fit the left tail of the distribution:

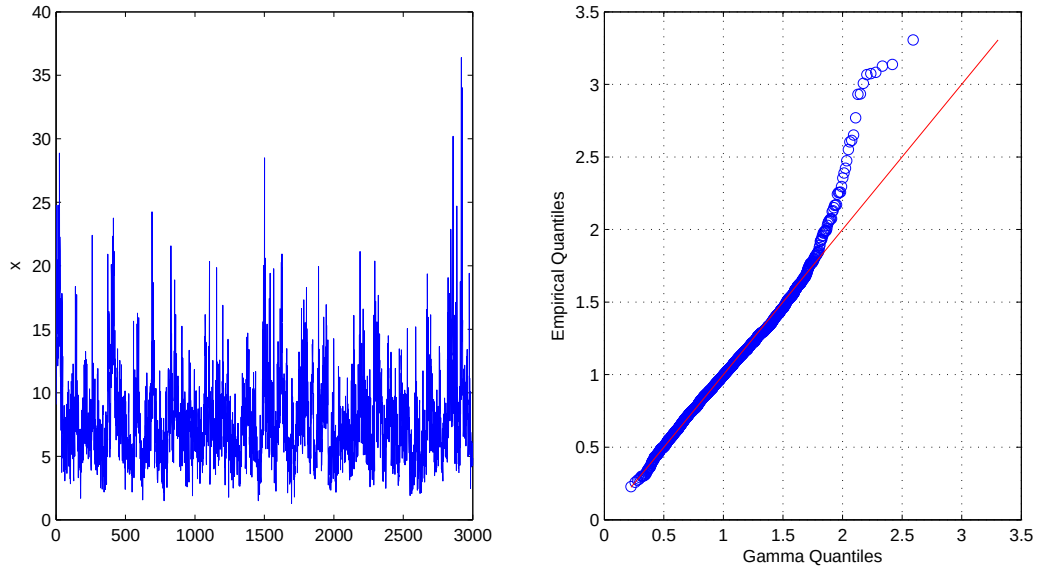

Figure 1: The simulated time series (left) and the QQ-Plot of the estimated innovations of the parametric MEM with Gamma distributed innovations.

The DPMG1-MEM has improved the fitting in the left tail of the distribution, but this comes at the cost of a worse fit in the zero neighborhood (Because the kernel of the mixture has only one free parameter ϕ. By decreasing ϕ the left tail of the kernel becomes thicker, while, at the same time, the probability of the zero neighborhood increases.) In contrast to this, the prior of DPMG2-MEM assigns positive probability in a Kullback-Leibler neighborhood of the true distribution of innovation (Kullback-Leibler property), and therefore the model can consistently estimate this distribution. As we can see in Figure 4 this model has almost perfectly recovered the true distribution of the innovations.

As pointed out in Section 2, an alternative to the parameter expanded model, is to restrict the constant term of the conditional mean to unity and model the distribution of the innovations by a DPM of Gamma distributions:

$$\begin{aligned} \mu_t &= 1 + \alpha^* x_{t-1} + \beta \mu_{t-1} \\ x_t &= \mu_t \varepsilon_t \end{aligned}$$

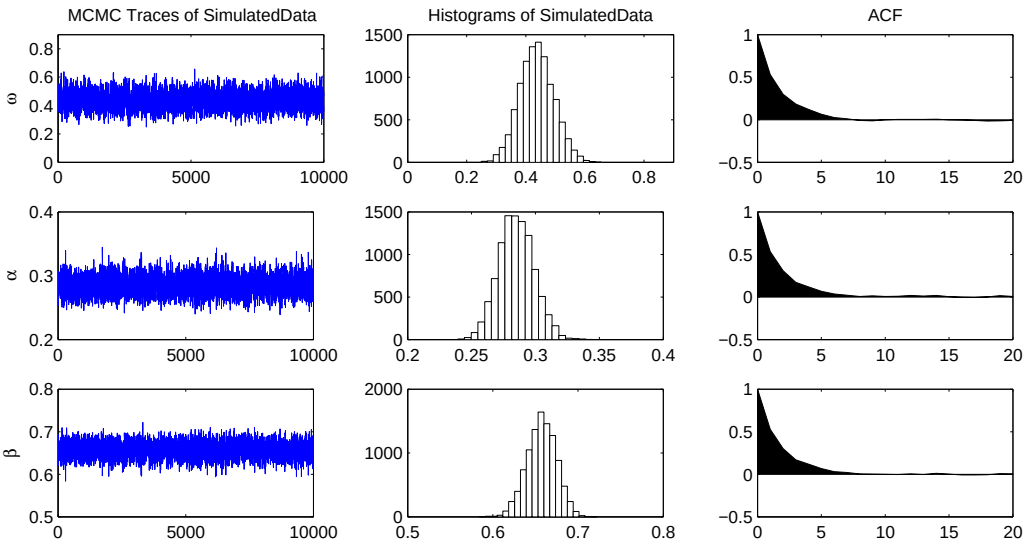

Figure 2: Simulation study: MCMC traces, posterior distributions, and ACF of DPMG1-MEM

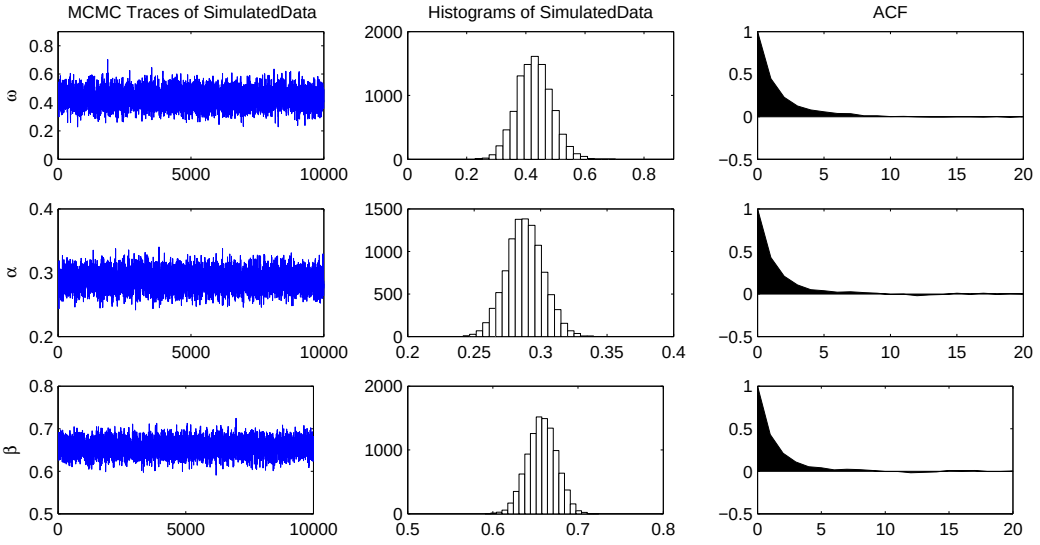

Figure 3: Simulation study: MCMC traces, posterior distributions, and ACF of DPMG2-MEM

We can adapt our sampling scheme also to this setting. However since the parameters $(\alpha^*,\ \beta)$ are highly anti-correlated to the mean parameters of the components of the DPM, the resulting Markov chain mixes extremely slowly (In this formulation, for the true parameters used in the simulations, the correlation between the MLE estimates ω, and α and β are -0.93 and -0.63, respectively). Figure 5 shows the 10^6 MCMC traces of ω, α^* and β, along with their auto-correlation function (ACF).

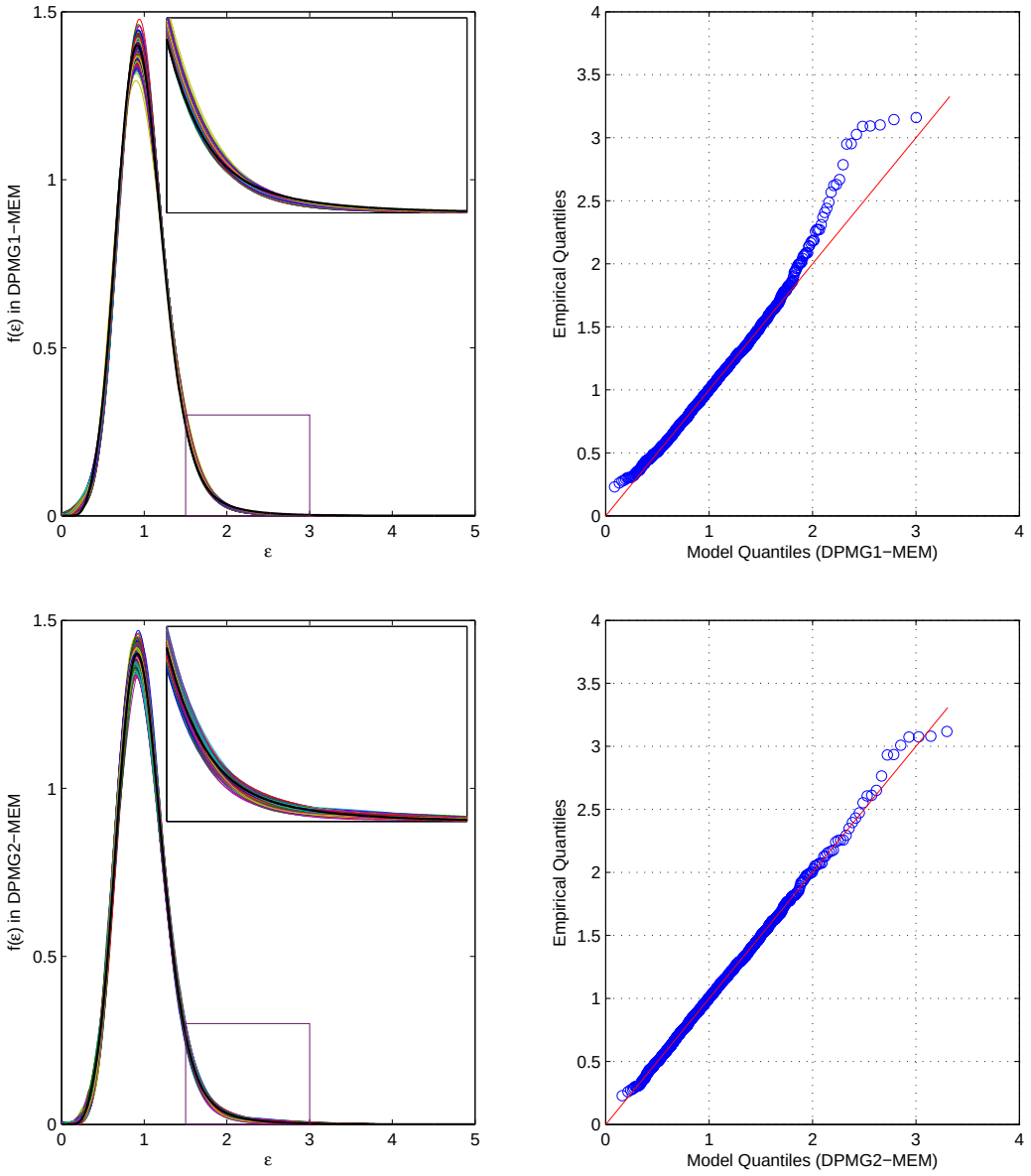

Figure 4: Simulation study: Posterior distribution along with the true distribution (the black thick line) of the innovations and the corresponding QQ-Plot of the DPG1-MEM (top) and DPG2-MEM (bottom)

Note that the ACF remain significant even up to 2×10^4 lags. Comparing this to Figure 3 demonstrates that the algorithm proposed for inference in DPMG2 MEM is significantly more efficient than the sampling schemes implemented on the naïvely reparameterized model.

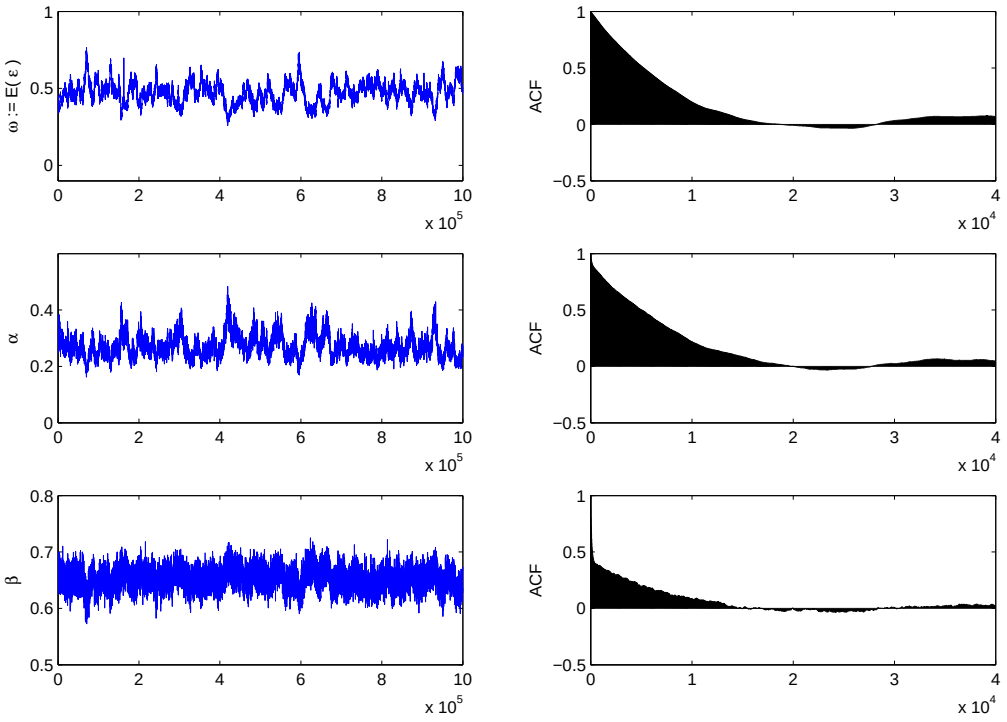

Figure 5: Estimation of the semiparametric model with the formulation $\mu_t = 1 + \alpha^* x_{t-1} + \beta\mu_{t-1}$, and innovations modeled by a DPM with unrestricted mean: MCMC traces and their ACF

5 Empirical Analysis

The proposed semiparametric models is now fitted to the daily realized volatility of Standard & Poor 500 (S&P 500), Dow Jones Industrial Average (DJIA) and FTSE 100. The data is obtained from the Oxford-Man Institute's "realised library" publicly available[1]. In particular we have used the realized kernel (Barndorff-Nielsen, Hansen, Lunde, and Shephard 2008) that is proved to be robust to market microstructure noises. The data covers the period from Jan. 1996 to Feb. 2009 for the S&P 500 and DJIA (3261 observations) and from Nov. 1997 to Feb. 2009 for the FTSE (2844 observations). In our empirical analysis, the Realized Kernel time series, RK_t, is transformed to annualized realized volatility in percentages:

$$x_t := \sqrt{252\ RK_t} \times 100$$

[1] http://realized.oxford-man.ox.ac.uk/

The proposed semiparameteric models is fitted to these time series and inference is conducted by MCMC simulation using the algorithm detailed in the previous sections. Figures 6 and 7 present the traces of the MCMC simulations for estimation of DPMG1-MEM and DPMG2-MEM for the realized volatility of S&P 500, along with the posterior distributions of the parameters, and the ACF of the Markov chain (for the sake of brevity, very similar figures for DJIA and FTSE, and for DPMG1-AMEM and DPMG2-AMEM are not reported). The estimated parameters $\hat{\omega}$, $\hat{\alpha}$ and $\hat{\beta}$ are reported in Table 2 and 3.

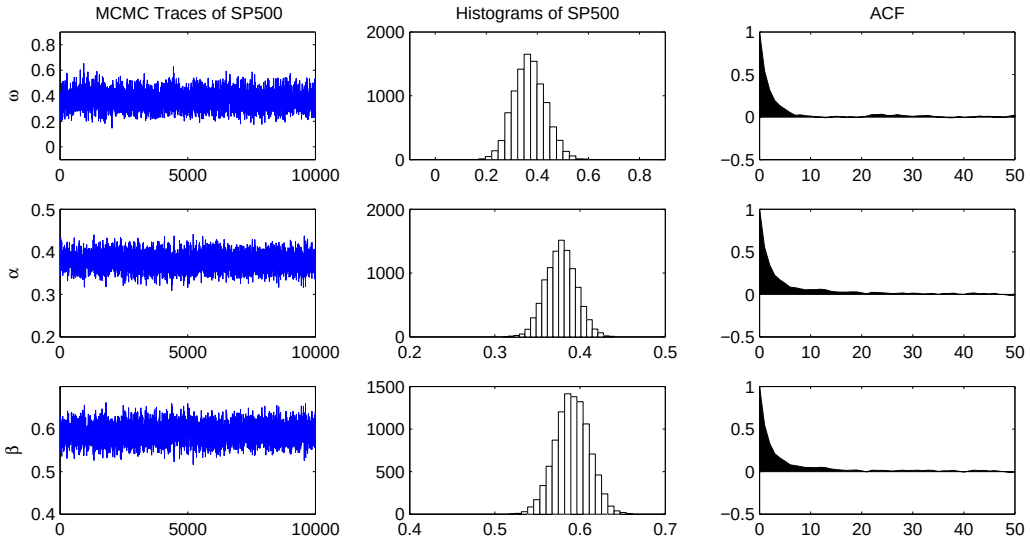

Figure 6: S&P 500: MCMC traces, posterior distributions, and ACF of DPMG1-MEM

Table 2: Estimation of DPMG1-MEM and DPMG2-MEM

	DPMG1-MEM			DPMG2-MEM		
	$\hat{\omega}$	$\hat{\alpha}$	$\hat{\beta}$	$\hat{\omega}$	$\hat{\alpha}$	$\hat{\beta}$
S&P 500	0.369	0.378	0.591	0.346	0.363	0.611
DJIA	0.354	0.386	0.583	0.358	0.377	0.596
FTSE 100	0.144	0.281	0.704	0.153	0.27	0.719

Figure 8 exhibits the QQ-Plots of the estimated innovations of these time series obtained by the parametric Gamma-MEM, DPMG1-MEM and DPMG2-MEM. We observe that a parametric MEM is not able to fit the conditional distribution of

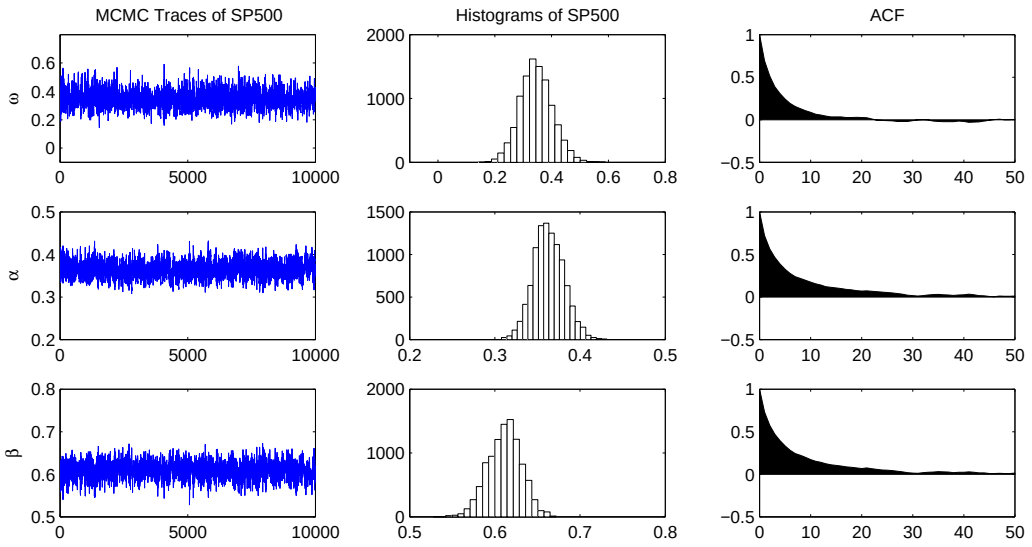

Figure 7: S&P 500: MCMC traces, posterior distributions, and ACF of DPMG2-MEM

Table 3: Estimation of DPMG1-AMEM and DPMG2-AMEM

	DPMG1-AMEM				DPMG2-AMEM			
	$\hat{\omega}$	$\hat{\alpha}$	$\hat{\beta}$	$\hat{\gamma}$	$\hat{\omega}$	$\hat{\alpha}$	$\hat{\beta}$	$\hat{\gamma}$
S&P 500	0.458	0.242	0.673	0.091	0.447	0.244	0.677	0.089
DJIA	0.401	0.258	0.668	0.077	0.405	0.262	0.668	0.076
FTSE 100	0.180	0.185	0.770	0.055	0.181	0.181	0.777	0.052

the realized volatility. In comparison, the DPMG1-MEM has a better performance: QQ-Plots show a better fit for the right tail of the distribution, however on the other hand, the fit has worsen in the neighborhood of zero. Finally an almost perfect fit can be achieved using the DPMG2-MEM. (The very similar graphs obtained for AMEM models have been dropped for the sake of brevity.)

We have also compared the models in terms of their (in the sample) predictive performance. In particular we have used the log-predictive score and log-predictive

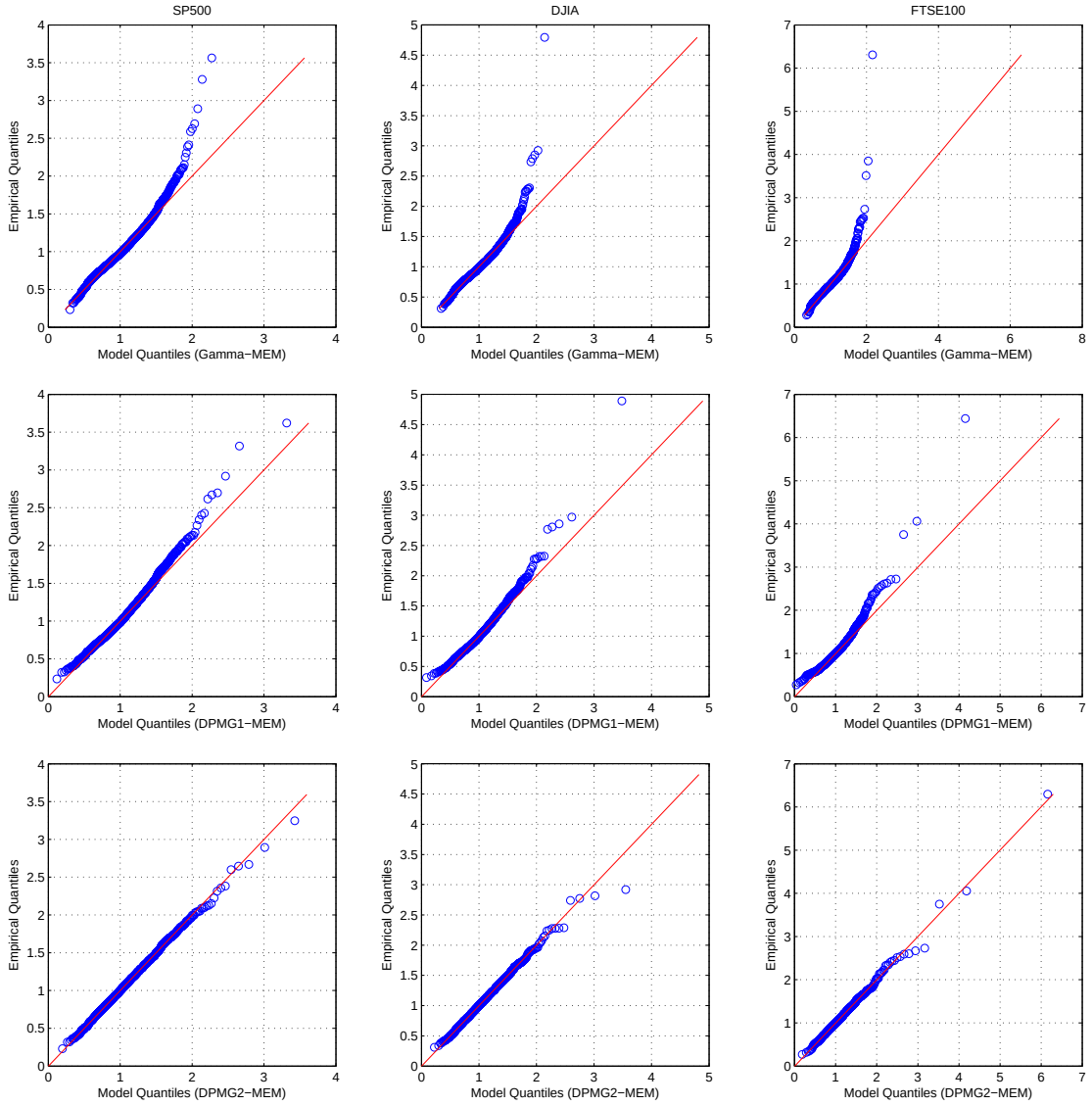

Figure 8: The QQ-Plots of the estimated innovations of the parametric MEM with Gamma distributed innovations (first row) and the semiparametric MEM models DPMG1-MEM (second row) and DPMG1-MEM (third row).

tail score (LPS and LPTS) of Delatola and Griffin 2010:

$$\begin{aligned}
\text{LPS} &:= -\frac{1}{n}\sum_{t=1}^{n}\log \hat{f}_{x_t}(x_t) = -\frac{1}{n}\sum_{t=1}^{N}\log\left(\frac{1}{\hat{\mu}_t}\hat{f}_{\varepsilon}(x_t/\hat{\mu}_t)\right) \\
\text{LPTS}_q &:= -\frac{1}{\sum_{t=1}^{n}\mathbf{1}(x_t > q_\alpha)}\sum_{t=1}^{n}\mathbf{1}(x_t > q_\alpha)\log \hat{f}_{x_t}(x_t) \\
&= -\frac{1}{\sum_{t=1}^{n}\mathbf{1}(x_t > q_\alpha)}\sum_{t=1}^{n}\mathbf{1}(x_t > q_\alpha)\log\left(\frac{1}{\hat{\mu}_t}\hat{f}_{\varepsilon}(x_t/\hat{\mu}_t)\right)
\end{aligned}$$

where q_α is the quantile of x_t (In our comparisons we have used have used $\alpha = 0.95$ and $\alpha = 0.99$). The probability density function of the innovations in the DPMG1-MEM and DPMG1-MEM have been estimated by

$$\hat{f}_{\varepsilon}(\varepsilon) = \frac{1}{N}\sum_{i=1}^{N}\sum_{j=1}^{+\infty} w_j^{(i)}\text{Gam}(\varepsilon;\phi_j^{(i)},1)$$

Similarly for DPMG2-MEM and DPMG2-AMEM we have used the following estimator:

$$\hat{f}_{\varepsilon}(\varepsilon) = \frac{1}{N}\sum_{i=1}^{N}\sum_{j=1}^{+\infty} w_j^{(i)}\text{Gam}(\varepsilon;\phi_j^{(i)},\mu_j^{(i)})$$

(where $N = 10^4$ is the number of MCMC sweeps.) The inner summation has been truncated in such way that $\sum_{j=1}^{N_i^{(Trunc)}} w_j > 0.999$. By this definition, a lower LPS or LPTS is an indication of a better (in the sample) predictive performance. We have reported the LPS and LPTS of the parametric and the semiparametric models in Table 4 and Table 5. The results demonstrate that the DPMG1-MEM and DPMG1-AMEM perform better than thier parametric counterparts. Moreover DPMG2-MEM and DPMG2-AMEM outperform DPMG1-MEM and DPMG1-AMEM. The above comparisons are true for all three financial time series considered.

An out of sample test has also been performed. Both the parametric and semi-

Table 4: LPS and LPTS of the parametric MEM, DPMG1-MEM and DPMG2-MEM (in the sample, using the whole sample)

		Gamma-MEM	DPMG1-MEM	DPMG2-MEM
S&P 500	LPS	2.5753	2.5600	2.5482
	LPTS 5%	4.4552	4.4171	4.3341
	LPTS 1%	5.2798	5.1340	5.0178
Dow Jones Industrial	LPS	2.4683	2.4421	2.4306
	LPTS 5%	4.5489	4.2928	4.2052
	LPTS 1%	5.6303	5.2286	5.0814
FTSE 100	LPS	2.5158	2.4668	2.4528
	LPTS 5%	5.0485	4.5209	4.3950
	LPTS 1%	7.3766	6.1834	5.8474

Table 5: LPS and LPTS of the parametric AMEM, DPMG1-AMEM and DPMG2-AMEM (in the sample, using the whole sample)

		Gamma-AMEM	DPMG1-AMEM	DPMG2-AMEM
S&P 500	LPS	2.5316	2.5113	2.5032
	LPTS 5%	4.2516	4.1883	4.1396
	LPTS 1%	5.087	4.8136	4.7518
Dow Jones Industrial	LPS	2.4292	2.3987	2.3918
	LPTS 5%	4.3621	4.1095	4.0485
	LPTS 1%	5.2931	4.9619	4.8668
FTSE 100	LPS	2.4867	2.4375	2.4269
	LPTS 5%	4.9357	4.3588	4.267
	LPTS 1%	7.0836	6.026	5.7249

Table 6: LPS and LPTS of the parametric MEM, DPMG1-MEM and DPMG2-MEM (out of sample)

		Gamma-MEM	DPMG1-MEM	DPMG2-MEM
S&P 500	LPS	2.4684	2.4483	2.4395
	LPTS 5%	4.3523	4.3391	4.2858
	LPTS 1%	5.7612	5.6096	5.4449
Dow Jones Industrial	LPS	2.3804	2.3439	2.3365
	LPTS 5%	4.7351	4.4303	4.3348
	LPTS 1%	6.3302	5.8249	5.5693
FTSE 100	LPS	2.3922	2.3757	2.3647
	LPTS 5%	5.0034	4.5131	4.3961
	LPTS 1%	6.7100	6.2676	5.9387

parametric models have been estimated using the training dataset (first half of the samples). Then the estimated models have been used to predict the realized volatility on the test dataset (second half of the samples). Using the estimated innovations' distribution and the estimated parameters $\hat{\omega}$, $\hat{\alpha}$ and $\hat{\beta}$ (all estimated using the training dataset), we have computed the LPS and LPTSs for the test dataset. The results of the out of sample test are reported in Table 6 and Table 7 and confirm the same ordering of the models observed in the sample test.

6 Conclusions

This paper offers a novel contribution both on the modeling and on the computational aspects of Bayesian MEM. We propose semiparametric MEMs where the distribution of the innovations is modeled by a DPM resulting in a more flexible and efficient framework in comparison with the standard parametric setting. Both symmetric (MEM) and asymmetric (AMEM) models have been considered. Bayesian inference is conducted via MCMC simulations. A sampling algorithm is proposed that is based on a parameter expanded model and results in an efficient and fast simulation algorithm. Our empirical studies show that the proposed semiparametric

Table 7: LPS and LPTS of the parametric AMEM, DPMG1-AMEM and DPMG2-AMEM (out of sample)

		Gamma-AMEM	DPMG1-AMEM	DPMG2-AMEM
S&P 500	LPS	2.4235	2.3969	2.3902
	LPTS 5%	4.2284	4.1756	4.1491
	LPTS 1%	5.4846	5.2963	5.2103
Dow Jones Industrial	LPS	2.3424	2.2987	2.2934
	LPTS 5%	4.6186	4.2984	4.2257
	LPTS 1%	6.1527	5.6346	5.4481
FTSE 100	LPS	2.4032	2.3727	2.3613
	LPTS 5%	5.0000	4.4246	4.3242
	LPTS 1%	6.9320	6.2166	5.9100

models are able to fit the financial time series better than their parametric counterparts. Also in terms of orediction (in and out of sample), the proposed models show a better performance for all three financial time series considered (S&P 500, Dow Jones Industrial and FTSE 100).

7 Acknowledgments

The authors gratefully thank the two referees for their helpful comments and suggestions that greatly improved this manuscript, and Antonio Lijoi and Sonia Petrone for fruitful discussions on the Bayesian nonparametric aspects related to this research line.

8 References

[1] K. Ahoniemi, M. Lanne, Joint Modeling of Call and Put Implied Volatility. International Journal of Forecasting, 25, pp. 239-258, 2009.

[2] K. Ahoniemi and M. Lanne, Time-Varying Mixture Multiplicative Error Models

for Implied Volatility, 2011.

[3] C. E. Antoniak, Mixtures of Dirichlet processes with applications to Bayesian nonparametric problems, Annals of Statistics, 2(6), pp. 1152–1174, 1974.

[4] O. E. Barndorff-Nielsen, P. R. Hansen, A. Lunde, and N. Shephard, Designing realised kernels to measure the ex-post variation of equity prices in the presence of noise, Econometrica, 76, pp. 1481-1536, 2008.

[5] C. T. Brownlees, F. Cipollini, and G. M. Gallo, Intra-daily Volume Modeling and Prediction for Algorithmic Trading, Journal of Financial Econometrics, 9, pp. 489-518, 2011.

[6] C. T. Brownlees, F. Cipollini, and G. M. Gallo, Multiplicative Error Models, in "Handbook in Financial Engineering and Econometrics: Volatility Models and Their Applications", L. Bauwens, C. Hafner and S. Laurent, Wiley, 2012.

[7] D. Burr, H. Doss, A Bayesian semiparametric model for random-effects meta-analysis. Journal of the American Statistical Association, 100, pp. 242–251. 2005.

[8] R. Y. Chou, Forecasting financial volatilities with extreme values: The conditional autoregressive range (CARR) model, Journal of Money, Credit and Banking, 37(3), pp. 561-582, 2005.

[9] E. Delatola, J. Griffin, Bayesian Nonparametric Modelling of the Return Distribution with Stochastic Volatility, Technical Report, University of Kent, 2010.

[10] F. C. Drost, B. J. M. Werker, Semiparametric duration models, Journal of Business and Economic Statistics, 22(1), pp. 40-50, 2004.

[11] R. F. Engle, New Frontiers For ARCH Models, Journal Of Applied Econometrics, 17, pp. 425–446, 2002.

[12] Engle, R. F., G. M. Gallo 2006: A multiple indicators model for volatility using intra-daily data, Journal of Econometrics, 131, 3-27.

[13] R. F. Engle and J. R. Russell, Autoregressive Conditional Duration: A New Model for Irregularly Spaced Transaction Data, Econometrica, 66(5), pp. 1127–1162, 1998.

[14] M. D. Escobar, Estimating normal means with a Dirichlet process prior, J. Am. Stat. Assoc., 89, pp. 268-277, 1994.

[15] m. D. Escobar, M. West, Bayesian density estimation and inference using mixtures, J. Am. Stat. Assoc., 90, pp. 577-588, 1995.

[16] T. S. Ferguson, A Bayesian analysis of some nonparametric problems, Annals of Statistics, 1(2), pp. 209–230, 1973.

[17] A. Gelman, Prior distributions for variance parameters in hierarchical models, Bayesian Analysis, 1(3), pp. 515–533, 2006.

[18] W. R. Gilks, P. Wild, Adaptive Rejection Sampling for Gibbs Sampling, Applied Statistics, 41, pp. 337-348, 1992.

[19] T. Hanson, W. O. Johnson, Modeling regression error with a mixture of Pólya trees. Journal of the American Statistical Association, 97, pp. 1020–1033, 2002.

[20] N. Hautsch, P. Malec, M. Schienle, Capturing the Zero: A New Class of Zero-Augmented Distributions and Multiplicative Error Processes, Discussion Paper 2010-055, CRC 649, Berlin, 2010.

[21] Haario H., Saksman E., Tamminen J., An adaptive Metropolis algorithm, Bernoulli, 7, 223-242, 2001.

[22] H. Ishwaran, M. Zarepour, Markov Chain Monte Carlo in Approximate Dirichlet and Beta Two-Parameter Process Hierarchical Models, Biometrika, 87(2), 371-390, 2000.

[23] S. Jain, R. M. Neal, A split-merge Markov chain Monte Carlo procedure for the Dirichlet process mixture model, Journal of Computational and Graphical Statistics, 13, p. 158–182, 2004.

[24] S. Jain, R. M. Neal, Splitting and merging components of a nonconjugate Dirichlet process mixture model, Technical Report 0507, Department of Statistics, University of Toronto, 2005.

[25] M. Kalli, J. Griffin, S. Walker, Slice Sampling Mixture Models, Statistics and Computing, 21, pp. 93-105, 2011.

[26] M. Kalli, S. Walker, P. Damien Modelling the conditional distribution of daily stock index returns: an alternative Bayesian semiparametric model, Technical report, University of Kent, Center for Health Services Studies, 2011.

[27] A. Y. Lo, On a class of Bayesian nonparametric estimates: I. Density estimates, Annals of Statistics, 12(1), pp. 351-357, 1984.

[28] M. Lanne, A Mixture Multiplicative Error Model for Realized Volatility, Journal of Financial Econometrics, 4, pp. 594616, 2006.

[29] A. Lijoi, E. Regazzini, Means of a Dirichlet process and multiple hypergeometric functions, The Annals of Probability, 32, pp. 14691495, 2004.

[30] J. S. Liu, Y. N. Wu, Parameter Expansion for Data Augmentation, J. Amer. Statist. Assoc., 94, pp. 1264–74, 1999.

[31] C. Liu, D. B. Rubin, Y. N. Wu, Parameter expansion to accelerate EM: The PX-EM algorithm, Biometrika, 85(4), pp. 755-770 1998.

[32] P. Muliere, L. Tardella, Approximating Distributions of Random Functionals of Ferguson-Dirichlet Priors, The Canadian Journal of Statistics, 26(2), pp. 283–297, 1998.

[32] R. M. Neal, Markov chain sampling methods for Dirichlet process mixture models, Journal of Computational and Graphical Statistics, 9(2), pp. 249–265, 2000.

[33] O. Papaspiliopoulos, A note on posterior sampling from Dirichlet mixture models, Preprint, 2008.

[34] O. Papaspiliopoulos, G. O. Roberts, Retrospective Markov chain Monte Carlo methods for Dirichlet process hierarchical models, Biometrika 95, pp. 169-186, 2008.

[35] J. Pitman, Combinatorial stochastic processes, Technical Report 621, U. C. Berkeley Department of Statistics, August 2002.

[36] G. O. Roberts, J. S. Rosenthal, Coupling and ergodicity of adaptive Markov chain Monte Carlo algorithms, Journal of Applied Probability, 44(2), pp. 458-475, 2007.

[37] J. Sethuraman, A constructive definition of Dirichlet priors, Statistica Sinica,

4, pp. 639-50, 1994.

[38] D. A. van Dyk, X. L. Meng, The Art of Data Augmentation (with discussion). Journal of Computational and Graphical Statistics, 10, pp. 1–111, 2001.

[39] M. Yang, D. B. Dunson, D. Baird, Semiparametric Bayes hierarchical models with mean and variance constraints, Computational Statistics and Data Analysis, 54 (9), pp. 2172-2186, 2010.

[40] Y. Wu, S. Ghosal, Kullback Leibler property of kernel mixture priors in Bayesian density estimation. Electronic Journal of Statistics, 2, pp. 298-331, 2008.