Università degli Studi Milano-Bicocca

Dipartimento di Economia, Metodi quantitativi e Strategie d'impresa

Ph.D. course in Statistics

Ph.D. Thesis

Bayesian semiparametric analysis of vector Multiplicative Error Models

Nicola Donelli

Supervisor:
Prof.ssa Antonietta Mira

Internal Supervisor:
Prof.ssa Fulvia Mecatti

Ph.D. Cycle XXVIII

To Manuela Pivetta, who taught me the love for Mathematics

Contents

Introduction

In many fields it is interesting to study the interactions between time series of different measures of the same unobservable characteristic or between time series of the same characteristic measured on different subjects. To study the case in which these time series are non-negative, Engle and Gallo (2006) introduced a multivariate version of the Multiplicative Error Model proposed by Engle (2002). They called it "vector Multiplicative Error Model" (vMEM). In this model the vector of observations is represented as the element-by-element product of a conditional mean vector and a random innovation vector. The main limitation of the early approaches to the analysis of these models was the necessity to use some restrictive hypotheses on the multidimensional distribution of the random term. The main purpose of our research line is to overcome this issue using a Bayesian Semiparametric approach: we model the multidimensional distribution of the innovation vector as an infinite location-scale mixture of multidimensional kernels with support on the positive orthant obtaining a very flexible specification that outperforms the classical ones both in terms of fitting and in predictive power. Furthermore we introduce a very general specification for the conditional mean vector term so that the model could take into account several different features of the observed time series and could be used to model different kind of data. We will in particular use this model on real financial datasets to study the dynamic interactions among different volatility proxies measured on the same market index. Thanks to the general specification of the conditional mean term and to the flexibility of the distribution of the random term, we strongly believe that our model could be a useful improvement for applications in which different vMEMs are already used and maybe even allow for new ones.

Chapter 1

Random measures and Dirichlet processes

1.1 Introduction

In this chapter we will give a brief introduction to the fundamentals of the branch of statistics known as Bayesian nonparametrics. The results we are going to show are nowadays classical results whose study started with the papers by Ferguson (1973) , Antoniak (1974), Cifarelli and Regazzini (1979) and many other authors after them.

1.2 Random Measures

1.2.1 The space of all the probability measures

Let $(\mathscr{X}, \mathscr{A})$ be a Polish space with its Borel σ-field and let $\left([0,1]^{\mathscr{A}}, \mathscr{W}\right)$ be the space of probability measures on $(\mathscr{X}, \mathscr{A})$ [1] equipped with the Borel σ-field generated by the weak topology. Before introducing the concept of random probability measure we give, without proof, a result that should explain why, although other measurability structures are possible, we choose exactly the $\mathscr{W}$ σ-field.

Proposition 1.1. *Let $\mathscr{X}$ be Polish and $\mathscr{W}$ the Borel σ-field generated by the class of all open subsets of $[0,1]^{\mathscr{A}}$ defined using the weak convergence. Then $\mathscr{W}$ is also:*

1. *The smallest σ-field on $[0,1]^{\mathscr{A}}$ making all the maps $P \mapsto P(A)$ measurable, $\forall A \in \mathscr{A}$.*
2. *The smallest σ-field on $[0,1]^{\mathscr{A}}$ making all the maps $P \mapsto P(A)$ measurable, $\forall A$ in a generator $\mathscr{A}_0$ of $\mathscr{A}$.*
3. *The smallest σ-field on $[0,1]^{\mathscr{A}}$ making all the maps $P \mapsto \int \psi dP$ measurable, $\forall \psi \in C_b^0(\mathscr{X})$.*

[1] Technically this is an abuse of notation, since $[0,1]^{\mathscr{A}}$ usually refers to the space of all the functions from $\mathscr{A}$ to $[0,1]$ and instead here we are using that name for its subspace composed only by the σ-additive functions (i.e. the measures).

Consequently, a finite measure on $\left([0,1]^{\mathscr{A}}, \mathscr{W}\right)$ *is completely determined by the family of distributions induced by the maps* $P \mapsto (P(A_1), \ldots, P(A_k))$, *for* $A_1, \ldots, A_k \in \mathscr{A}_0$ *and* $k \in \mathbb{N}$ *and also by the maps* $P \mapsto \int \psi dP$ *for* $\psi \in C_b^0(\mathscr{X})$.

Furthermore, the space of probability measures on $(\mathscr{X}, \mathscr{A})$, *equipped with the weak convergence metric, is a Polish space.*

1.2.2 Definitions of random probability measure

Having equipped the space $[0,1]^{\mathscr{A}}$ of probability measures on $(\mathscr{X}, \mathscr{A})$ with the described σ-field $\mathscr{W}$ we can give the following definition.

Definition 1.1. Let $(\mathscr{X}, \mathscr{A})$ and $\left([0,1]^{\mathscr{A}}, \mathscr{W}\right)$ be as defined before. A random probability measure (r.p.m.) on $(\mathscr{X}, \mathscr{A})$ is a random variable P from some probability space $(\Omega, \mathscr{F}, \mathbb{P})$ to the space of all the probability measures on $(\mathscr{X}, \mathscr{A})$, namely:

$$P : (\Omega, \mathscr{F}, \mathbb{P}) \to \left([0,1]^{\mathscr{A}}, \mathscr{W}\right).$$

The first point of Proposition 1.1 implies that a map $P : (\Omega, \mathscr{F}, \mathbb{P}) \to \left([0,1]^{\mathscr{A}}, \mathscr{W}\right)$ is measurable if and only if the induced "random probability" $P(A)$ of every set $A \in \mathscr{X}$, viewed as a function from $\left([0,1]^{\mathscr{A}}, \mathscr{W}\right)$ to $([0,1], \mathscr{B}([0,1]))$, is measurable (i.e. it is a random variable). Thus a random probability measure can be identified with a random variable $\{P(A)\}_{A \in \mathscr{A}}$ in the product space $[0,1]^{\mathscr{A}}$. Hence can be given also the subsequent definition:

Definition 1.2. A random probability measure (r.p.m.) P on $(\mathscr{X}, \mathscr{A})$ is a stochastic process indexed by the elements of the σ-field $\mathscr{A}$ s.t.

$$P : \Omega \times \mathscr{A} \to ([0,1], \mathscr{B}([0,1])) \text{ s.t.}$$
1) $\omega \mapsto P(\omega, A)$ is $\mathscr{W}$-measurable $\forall A \in \mathscr{A}$
2) $A \mapsto P(\omega, A)$ is a probability measure on $(\mathscr{X}, \mathscr{A})$ $\forall \omega \in \Omega$.

From this definition follows that a general way to construct a r.p.m. is to build a stochastic process indexed by the elements of the σ-field $\mathscr{A}$ using Kolmogorov's extension theorem and next show that this process can be realized within $[0,1]^{\mathscr{A}}$, viewed as a subset of $\mathbb{R}^{\mathscr{A}}$. Before starting the description of the construction, let us report, without proof, the Kolmogorov's extension theorem.

Theorem 1.1. *For every finite subset* S *of an arbitrary set* T *let* P_S *be a probability distribution on* $\mathbb{R}^S$. *Then there exists a probability space* $(\Omega, \mathscr{F}, \mathbb{P})$ *and measurable maps* $X_t : \Omega \to \mathbb{R}$ *such that* $(X_t)_{t \in S} \sim P_S$ *for every finite set* S *if and only if for every pair* $S' \subset S$ *of finite subsets of* T, $P_{S'}$ *is the marginal distribution of* P_S *on* $\mathbb{R}^{S'}$.

From the theorem stems that to build our target stochastic process we can start defining the joint distributions of the random variables $(P(A_1), \ldots, P(A_k))$ for every k and every sequence $A_1, \ldots, A_k$ of measurable sets (these distributions are also sometimes called "marginals" or "finite dimensional laws" of the stochastic process). If the given distributions respect the consistency condition required by Kolmogorov's theorem, it ensures that there exists a probability measure $\mathscr{P}$

on $\left([0,1]^{\mathscr{A}}, \sigma\left(\mathscr{C}^{\mathscr{A}}\right)\right)$ that has them as finite-dimensional laws. To ensure that the process we built is a r.p.m. it is further necessary (but not sufficient) that

$$\begin{aligned} &i) P(\emptyset) = 1,\ P(\mathscr{X}) = 1\ \mathbb{P}-a.s. \\ &ii) P(A_1 \cap A_2) = P(A_1) + P(A_2)\ \mathbb{P}-a.s.,\ \forall A_1, A_2 \in \mathscr{A} \text{ s.t. } A_1 \cap A_2 = \emptyset \end{aligned} \tag{1.2.1}$$

These requirements are not sufficient to guarantee that our process is a r.p.m., since the null set in which $ii)$ does not hold depends in general on the pair of measurable, disjoint sets $\{A_1, A_2\}$ considered and $\mathscr{A}$ has countably many disjoint elements (so a direct extension of $ii)$ to σ-additivity is not possible).

Anyway we can overcome the problem proving the existence of a so called "mean measure", i.e. a map $\mu : \mathscr{A} \to \mathbb{R}$ s.t.

$$\mu(A) = E[P(A)].$$

If the defined process P is a r.p.m. this clearly defines a regular measure on $\mathscr{A}$. But, as shown in the theorem below, the existence of a mean measure is also sufficient for the existence of a version of $\{P(A)\}_{A\in\mathscr{A}}$ that is a r.p.m. on $(\mathscr{X}, \mathscr{A})$.

Theorem 1.2. *Suppose that $\{P(A)\}_{A\in\mathscr{A}}$ is a stochastic process that satisfies $i)$ and $ii)$ and whose mean function $A \mapsto E[P(A)]$ is a (regular) Borel measure on $(\mathscr{X}, \mathscr{A})$. Then there exists a measurable map $\tilde{P} : (\Omega, \mathscr{F}, \mathbb{P}) \to \left([0,1]^{\mathscr{A}}, \mathscr{W}\right)$ such that $\tilde{P}(A) = P(A)\ \mathbb{P}-a.s.$ for every $A \in \mathscr{A}$.*

Proof. Let $\mathscr{A}_0 = \{A_0, A_1, \ldots\}$ be a countable field that generates $\mathscr{A}$. Since the mean measure μ is regular, for every $i, m \in \mathbb{N}$ there exists a compact set $K_{i,m} \subset A_i$ with $\mu(A_i - K_{i,m}) \leq 2^{-2i-2m}$. By Markov's inequality we have that

$$\mathbb{P}\left\{P(A_i - K_{i,m}) > 2^{-i-m}\right\} \leq 2^{i+m} E[P(A_i - K_{i,m})] \leq 2^{-i-m}$$

hence, defined $\Omega_m = \bigcap_{i=1}^{\infty} \left\{P(A_i - K_{i,m}) \leq 2^{-i-m}\right\}$, we have that $\mathbb{P}\{\Omega_m\} \geq 1 - 2^{-m}$ and so $\mathbb{P}\left\{\liminf_{m\to\infty} \Omega_m\right\} = 1$ by the Borel-Cantelli lemma.

Since $\mathscr{A}_0$ is countable, the nulls set involved in $i)$ and $ii)$ with $\{A_i\}_{i\geq 1} \in \mathscr{A}_0$ can be aggregated into a single null set N. For every $\omega \notin N$ the process P is a finitely additive measure on $\mathscr{A}_0$ with the resulting usual properties of monotonicity and sub-additivity. We can also ensure that it is sub-additive on all finite unions of sets $A_i - K_{i,m}$ by increasing N if necessary.

Let $A_{i_1} \supset A_{i_2} \supset \ldots$ be an arbitrary decreasing sequence of sets in $\mathscr{A}_0$ with empty intersection. Then, for every m, the corresponding compacts $K_{i_j,m}$ posses empty intersection too: for every m there exists a finite J_m such that $\bigcap_{j\leq J_m} K_{i_j,m} = \emptyset$. This implies that

$$A_{i_{J_m}} = \bigcap_{j=1}^{J_m} A_{i_j} - \bigcap_{j=1}^{J_m} K_{i_j,m} \subset \bigcup_{j=1}^{J_m} \left(A_{i_j} - K_{i_j,m}\right)$$

and consequently, on the event $\Omega_m - N$,

$$\limsup_{j\to\infty} P\left(A_{i_j}\right) \leq P\left(A_{i_{J_m}}\right) \leq \sum_{j=1}^{J_m} \left(A_{i_j} - K_{i_j,m}\right) \leq 2^{-m}.$$

Thus on the event $\Omega_0 = \liminf_{m\to\infty} \Omega_m - N$ the limit is zero. We can conclude that for every $\omega \in \Omega_0$ the restriction of $A \mapsto P(A)$ to $\mathscr{A}_0$ is countably additive and, by Carathéodory's theorem it extends to a measure $\tilde{P}$ on $\mathscr{A}$.

By construction $\tilde{P}(A) = P(A)$ $\mathbb{P}$-a.s. for every $A \in \mathscr{A}_0$ and, in particular,

$$E\left[\tilde{P}(A)\right] = E[P(A)] = \mu(A) \ \forall A \in \mathscr{A}_0.$$

Hence, by uniqueness of extension, the mean measure of $\tilde{P}$ coincides with μ on $\mathscr{A}$.

Since $\mathscr{A} = \sigma(\mathscr{A}_0)$, for every $A \in \mathscr{A}$ there exists a sequence $\{A_m\} \subset \mathscr{A}_0$ s.t. $\mu(A \Delta A_m) \underset{m\to\infty}{\longrightarrow} 0$. Then both $P(A\Delta A_m)$ and $\tilde{P}(A\Delta A_m)$ tend to zero in mean. Further, finite additivity of P gives that $|P(A) - P(A_m)| \leq P(A\Delta A_m)$ and the same holds for $\tilde{P}$ thanks to its σ-additivity.

Hence $P(A) = \tilde{P}(A)$ $\mathbb{P} - a.s.$ for every $A \in \mathscr{A}$. This also proves that $\tilde{P}(A)$ is a random variable for every $A \in \mathscr{A}$ and so $\tilde{P}$ is a measurable map on $\left([0,1]^{\mathscr{A}}, \mathscr{W}\right)$. □

Rather than starting with a process $\{P(A)\}_{A\in\mathscr{A}}$ indexed on all the Borel sets, we may wish to start from a smaller set $\{P(A)\}_{A\in\mathscr{A}_0}$ of variables, for some $\mathscr{A}_0 \subset \mathscr{A}$. As shown by the previous proof a countable collection $\mathscr{A}_0$ suffices with just an additional specification regarding compact sets.

Theorem 1.3. *Suppose that $\{P(A)\}_{A\in\mathscr{A}_0}$ is a stochastic process that satisfies i) and ii) for a countable field $\mathscr{A}_0$ that generates $\mathscr{A}$ and is such that for every $A \in \mathscr{A}_0$ and $\varepsilon > 0$ there exists a compact $K_\varepsilon \in \mathscr{X}$ and $A_\varepsilon \in \mathscr{A}_0$ such that $A_\varepsilon \subset K_\varepsilon \subset A$ and $\mu(A - A_\varepsilon) < \varepsilon$, where μ is the mean measure $\mu(A) = E[P(A)]$. Then there exists a random measure that extends P to $\mathscr{A}$.*

1.2.3 Discrete random probability measures

Since discrete random probability measures (i.e. random probability measures which sample discrete distributions almost surely) are of particular interest in itself and, above all, for the rest of our discussion, we will now describe a general way to build them.

Given a random vector $(N, w_{1,N}, w_{2,N}, \ldots, w_{N,N}, \theta_{1,N}, \theta_{2,N}, \ldots \theta_{N,N})$ where $N \in \mathbb{N} \cup \{\infty\}$, the "weights" $(w_{1,N}, w_{2,N}, \ldots, w_{N,N})$ are non-negative random variables such that $\sum_{i=1}^{N} w_{i,N} = 1$ and the "locations" $(\theta_{1,N}, \theta_{2,N}, \ldots \theta_{N,N})$ are random variables taking values in $(\mathscr{X}, \mathscr{A})$, Definition 1.1 implies that we can define a random probability measure by

$$P = \sum_{i=1}^{N} w_{i,N} \delta_{\theta_{i,N}}, \tag{1.2.2}$$

where δ_θ is the measure that gives mass 1 to the point θ.

This random probability measure has support included in the subset of $[0,1]^{\mathscr{A}}$ made of the discrete probability distributions on $(\mathscr{X}, \mathscr{A})$ with finitely or countably many support points.

Proposition 1.2. *If the support of N is unbounded and, given N, the weights and the locations are mutually independent and if, given N=n, the weights have full support on the n-dimensional simplex, $\mathbb{S}_n$, and the locations have full support $\mathscr{X}^n$, for every n, then P has full support.*

Proof. Since the distribution with finite support are (weakly) dense in the space of the discrete distributions on $(\mathscr{X}, \mathscr{A})$ and this latter is (weakly) dense in the space of all the probability distributions on $(\mathscr{X}, \mathscr{A})$, it suffices to show that P gives positive probability to any (weak) neighbourhood of a distribution with finite support, $P^\star = \sum_{i=1}^{k} w_i^\star \delta_{\theta_i^\star}$.

All distributions $P' = \sum_{i=1}^{k} w_i \delta_{\theta_i}$ with $(w_1, \ldots, w_k)$ and $(\theta_1, \ldots, \theta_k)$ sufficiently close to $(w_1^\star, \ldots, w_k^\star)$ and $(\theta_1^\star, \ldots, \theta_k^\star)$ are in such a neighbourhood and so are the measures $P' = \sum_{i=1}^{\infty} w_i \delta_{\theta_i}$ with $\sum_{i=k+1}^{\infty} w_i$ small enough and $(w_1, \ldots, w_k)$ and $(\theta_1, \ldots, \theta_k)$ sufficiently close to $(w_1^\star, \ldots, w_k^\star)$ and $(\theta_1^\star, \ldots, \theta_k^\star)$.

If N is not identically infinite, then the assertion follows from the assumed positive probability of the events $\left\{ N = k',\ \max_{i \leq k'} (|w_{i,k'} - w_i^\star| \vee |\theta_{i,k'} - \theta_i^\star|) < \varepsilon \right\}$ for every $\varepsilon > 0$ and a given $k' > k$ (where we define $w_i^\star = 0$ and $\theta_i^\star$ arbitrarily for $k < i < k'$).

If N is infinite with probability one then the assertion follows considering the events

$$\left\{ \sum_{i>k} w_{i,\infty} < \varepsilon,\ \max_{i \leq k} (|w_{i,k} - w_i^\star| \vee |\theta_{i,k} - \theta_i^\star|) < \varepsilon \right\}.$$

These events have positive probability since they refer to an open subset of $\mathbb{S}_\infty$ and the weights have full support on $\mathbb{S}_\infty$ and so the assertions follows as before. □

A practically very important way to construct the sequence of weights $(w_1, w_2, \ldots)$ is the so called "stick-breaking" algorithm.

Given a sequence of random variables $v_1, v_2, \ldots$ with values in $[0, 1]$, we consider the interval $[0, 1]$ as a stick of unit length and we break it at the point v_1. We then consider the remaining part of the original stick, namely the interval $[v_1, 1]$, as a stick of unit length and we break it at the point v_2. We then consider the remaining part of the original stick, namely $[v_1 + v_2 (1 - v_1), 1]$, as a stick of unit length and we break it at the point v_3, etc... If we identify the sequence of weights $(w_1, w_2, \ldots)$ with the sequence of breaking points of the original stick we have that

$$w_j = v_j \prod_{i=1}^{j-1} (1 - v_i) \tag{1.2.3}$$

and, under mild conditions on the sequence $(v_1, v_2, \ldots)$, we have that $\sum_{j=1}^{\infty} w_j = 1$.

Proposition 1.3. *Given a sequence of random variables $(v_1, v_2, \ldots)$ with values in $[0, 1]$, the stick-breaking sequence $(w_1, w_2, \ldots)$ defined as in equation (1.2.3) sums one almost surely if and only if* $E\left[\prod_{j=1}^{n} (1 - v_j)\right] \underset{n \to \infty}{\longrightarrow} 0$.

If the variables $(v_1, v_2, \ldots)$ are independent the previous condition is equivalent to $\sum_{j=1}^{\infty} \log (E[1 - v_j]) = -\infty$. *In particular if $(v_1, v_2, \ldots)$ are i.i.d. it suffices that $Pr(v_1 > 0) > 0$.*

If for every $k \in \mathbb{N}$ the support of $(v_1, v_2, \ldots, v_k)$ is $[0, 1]^k$, then the support of $(w_1, w_2, \ldots)$ is the whole infinite-dimensional simplex, $\mathbb{S}_\infty$.

Proof. By construction the leftover mass at stage n of the stick-breaking algorithm is equal to $1 - \sum_{i=1}^{n} w_j = \prod_{i=1}^{n} (1 - v_i)$. Hence

$$(w_1, w_2, \ldots) \in \mathbb{S}_\infty \text{ a.s.} \iff \prod_{j=1}^{n} (1 - v_j) \underset{n\to\infty}{\longrightarrow} 0.$$

Since the sequence of leftover mass is decreasing, non-negative and bounded by 1, the convergence a.s. is equivalent to the convergence in mean, and thus is proved the first part of the proposition.

If $v_1, v_2, \ldots$ are independent variables the condition $E\left[\prod_{j=1}^{n} (1 - v_j)\right] \underset{n\to\infty}{\longrightarrow} 0$ becomes $\prod_{j=1}^{n} E[1 - v_j] \underset{n\to\infty}{\longrightarrow} 0$ which is equivalent to the condition $\sum_{j=1}^{\infty} \log(E[1 - v_j]) = -\infty$ and, if the variables are also identically distributed, this last condition is implied by $\Pr(v_1 > 0) > 0$.

Since, for every $k \in \mathbb{N}$, the random vector $(w_1, w_2, \ldots, w_k)$ is a continuous function of the random vector$(v_1, v_2, \ldots, v_k)$ and this latter has support $[0, 1]^k$, then the former has support $\mathbb{S}_k$ and the sequence $(w_1, w_2, \ldots)$ has support $\mathbb{S}_\infty$. □

An important special case of random probability measure built as in equation (1.2.2) is obtained when $N \equiv \infty$, the sequence of locations $\theta_1, \theta_2, \ldots$ is i.i.d. on $\mathscr{X}$ and the sequence of weights $w_1, w_2, \ldots$ is stick-breaking. In this case, if the common distribution of the $\theta_i s$ has support the full space $\mathscr{X}$ and the sequence of weights has support $\mathbb{S}_\infty$, then the r.p.m. defined as in (1.2.2) has full support.

1.3 The Dirichlet Process

We will now introduce the Dirichlet process, which is "the normal distribution of Bayesian nonparametrics". It is the default prior on spaces of probability measures and will be the building block of the models we will present in the next chapters. But first let us introduce the finite-dimensional analogue of the Dirichlet process, the Dirichlet distribution.

1.3.1 Dirichlet distribution

The Dirichlet distribution makes its appearance in problems involving order statistics and is known to Bayesians as the conjugate prior for the parameters of a multinomial distribution.

Definition 1.3. Let $G_1, \ldots, G_k \overset{ind}{\sim} \text{Gamma}(\alpha_j, 1)$ with $\alpha_j \geq 0 \ \forall j = 1, \ldots, k$ and $\alpha_j > 0$ for some $j \in \{1, \ldots, k\}$ (if $\alpha_j = 0$, then $G_j \sim \delta_0$). The Dirichlet distribution with parameter$(\alpha_1, \ldots, \alpha_k)$ is defined as the joint distribution of $\left(Y_1 = \frac{G_1}{\sum_{j=1}^{k} G_j}, \ldots, Y_k = \frac{G_k}{\sum_{j=1}^{k} G_j}\right)$. In symbols

$$\left(\frac{G_1}{\sum_{j=1}^{k} G_j}, \ldots, \frac{G_k}{\sum_{j=1}^{k} G_j}\right) \sim \text{Dir}(\alpha_1, \ldots, \alpha_k).$$

This distribution is always singular with respect to the Lebesgue measure on $\mathbb{R}^k$, since $Y_1 + \ldots + Y_k = 1$. In addition, if $\alpha_j > 0 \; \forall j = 1, \ldots, k$, the distribution of $(Y_1, \ldots, Y_{k-1})$ on $\mathbb{R}^{k-1}$ is absolutely continuous (with respect to the Lebesgue measure on $\mathbb{R}^{k-1}$) with density

$$f(y_1, \ldots, y_{k-1} | \alpha_1, \ldots, \alpha_k) = \frac{\Gamma\left(\sum_{j=1}^{k} \alpha_j\right)}{\prod_{j=1}^{k} \Gamma(\alpha_j)} \left(\prod_{j=1}^{k-1} y_j^{\alpha_j - 1}\right) \left(1 - \sum_{j=1}^{k-1} y_j\right)^{\alpha_k - 1} \mathbb{I}_{\mathbb{D}^k}(y_1, \ldots, y_{k-1}),$$

where $\mathbb{D}^k = \left\{ (y_1, \ldots, y_{k-1}) \in \mathbb{R}^{k-1} \middle| y_j \geq 0 \, \forall j = 1, \ldots, k-1, \; \sum_{j=1}^{k-1} y_j \leq 1 \right\}$ is a subspace of the k-dimensional simplex $\mathbb{S}_k$.

The main properties of the Dirichlet distribution are:

Proposition 1.4. *If $(Y_1, \ldots, Y_k) \sim Dir(\alpha_1, \ldots, \alpha_k)$ and is given a partition $I_1, \ldots, I_n$ of $\{1, \ldots, k\}$ then*

$$\left(\sum_{j \in I_1} Y_j, \ldots, \sum_{j \in I_n} Y_j\right) \sim Dir\left(\sum_{j \in I_1} \alpha_j, \ldots, \sum_{j \in I_n} \alpha_j\right).$$

Proof. It follows directly from the additive property of the Gamma distribution.

Since $W_i = \sum_{j \in I_i} G_j \overset{ind}{\sim} \text{Gamma}\left(\sum_{j \in I_i} \alpha_j, 1\right)$ for $i = 1, \ldots, n$, $\sum_{i=1}^{n} W_i = \sum_{j=1}^{k} G_j$ and $\sum_{j \in I_i} Y_j = \frac{W_i}{\sum_{j=1}^{n} W_j}$ the result follows directly. □

Proposition 1.5. *If $(Y_1, \ldots, Y_k) \sim Dir(\alpha_1, \ldots, \alpha_k)$ and we define $|\alpha| = \sum_{j=1}^{k} \alpha_j$, then , marginally, $Y_j \sim Beta(\alpha_j, |\alpha| - \alpha_j)$, for $j = 1, \ldots, k$ and hence*

$$E[Y_j] = \frac{\alpha_j}{|\alpha|}$$
$$E\left[Y_j^2\right] = \frac{\alpha_j(\alpha_j + 1)}{|\alpha|(|\alpha| + 1)}$$
$$E[Y_j Y_h] = \frac{\alpha_j \alpha_h}{|\alpha|(|\alpha| + 1)} \text{ if } j \neq h.$$

Proof. From Proposition 1.4 with $n = 2$, $I_1 = \{i\}$ and $I_2 = \{1, \ldots, k\} - \{i\}$, we have that

$$\left(Y_i, \sum_{j \neq i} Y_j\right) \sim \text{Dir}\left(\alpha_i, \sum_{j \neq i} \alpha_j\right) = \text{Beta}(\alpha_i, |\alpha| - \alpha_i) \; \forall i = 1, \ldots, k$$

Hence the expression for the first two central moments follow from the moments of the Beta distribution.

For the last assertion, using again Proposition 1.4 with $n = 2$, $I_1 = \{j, h\}$ and $I_2 = \{1, \ldots, k\} - \{j, h\}$ we have that

$$\left(Y_j + Y_h, \sum_{i \neq j,h} Y_i\right) \sim \text{Dir}\left(\alpha_j + \alpha_h, \sum_{i \neq j,h} \alpha_j\right) = \text{Beta}(\alpha_j + \alpha_h, |\alpha| - \alpha_j - \alpha_h)$$

and this gives $\mathrm{Var}(Y_h + Y_j) = \frac{(\alpha_j+\alpha_h)(|\alpha|-\alpha_j-\alpha_h)}{|\alpha|^2(|\alpha|+1)}$. Then, remembering that

$$2\mathrm{Cov}(Y_j, Y_h) = \mathrm{Var}(Y_j + Y_h) - \mathrm{Var}(Y_j) - \mathrm{Var}(Y_h)$$

we have that

$$\begin{aligned}
E[Y_j Y_h] &= \frac{1}{2}\left(\mathrm{Var}(Y_j + Y_h) - \mathrm{Var}(Y_j) - \mathrm{Var}(Y_h)\right) + E(Y_j)E(Y_h) = \\
&= \frac{1}{2}\left(\frac{(\alpha_j+\alpha_h)(|\alpha|-\alpha_j-\alpha_h) - \alpha_j(|\alpha|-\alpha_j) - \alpha_h(|\alpha|-\alpha_h)}{|\alpha|^2(|\alpha|+1)}\right) + \frac{\alpha_j}{|\alpha|}\frac{\alpha_h}{|\alpha|} = \\
&= \frac{-\alpha_j\alpha_h}{|\alpha|^2(|\alpha|+1)} + \frac{\alpha_j\alpha_h}{|\alpha|^2} = \\
&= \frac{\alpha_j\alpha_h}{|\alpha|(|\alpha|+1)}
\end{aligned}$$

□

Finally we give, without proof, the very well known conjugacy property mentioned at the start of this subsection.

Proposition 1.6. *If* $Pr\{X = j|Y_1, \ldots, Y_k\} = Y_j$ *a.s.* $\forall j = 1, \ldots, k$ *and* $(Y_1, \ldots, Y_k) \sim Dir(\alpha_1, \ldots, \alpha_k)$, *then*

$$(Y_1, \ldots, Y_k) \,|\, \{X = j\} \sim Dir(\alpha_1, \ldots, \alpha_{j-1}, \alpha_j + 1, \alpha_{j+1}, \ldots, \alpha_k).$$

It is useful to enlighten that from the last two propositions it follows that the equality

$$\Pr\{X = j, Y_1 \le y_1, \ldots, Y_k \le y_k\} = \Pr\{X = j\}\Pr\{Y_1 \le y_1, \ldots, Y_k \le y_k | X = j\}$$

may be expressed, in terms of the distribution function $D(y_1, \ldots, y_k|\alpha_1, \ldots, \alpha_k)$ of the Dirichlet distribution, as

$$\int_0^{y_1} \ldots \int_0^{y_k} z_j dD(z_1, \ldots, z_k|\alpha_1, \ldots, \alpha_k) = \frac{\alpha_j}{|\alpha|} D(y_1, \ldots, y_k|\alpha_1, \ldots, \alpha_{j-1}, \alpha_j + 1, \alpha_{j+1}, \ldots, \alpha_k).$$

1.3.1.1 An Urn Scheme for the Dirichlet Distribution

Consider an urn containing α balls of k different colours: α_1 balls are of colour 1, ... , α_k balls are of colour k. Thus the initial proportions for each colour are

$$\left(f_1^{(0)}, \ldots, f_k^{(0)}\right) = \left(\frac{\alpha_1}{\alpha}, \ldots, \frac{\alpha_k}{\alpha}\right).$$

Then at the ith step we draw a ball randomly from this urn, record its colour X_i, and then put it back in the urn with another ball of the same colour. After the first step hence we have

$$\left(f_1^{(1)}, \ldots, f_k^{(1)}\right) = \left(\frac{\alpha_1 + \delta_1(X_1)}{\alpha+1}, \ldots, \frac{\alpha_k + \delta_k(X_1)}{\alpha+1}\right).$$

If we continue with these sampling scheme, the asymptotic fraction of the balls of different colours in the urn will be a random draw from $\mathrm{Dir}(\alpha_1, \ldots, \alpha_k)$, i.e.

$$\left(f_1^{(n)}, \ldots, f_k^{(n)}\right) = \left(\frac{\alpha_1 + \sum_{i=1}^n \delta_1(X_i)}{\alpha+n}, \ldots, \frac{\alpha_k + \sum_{i=1}^n \delta_k(X_i)}{\alpha+n}\right) \underset{n\to\infty}{\Longrightarrow} \mathrm{Dir}(\alpha_1, \ldots, \alpha_k).$$

1.3.1.2 A Stick-Breaking construction

Proposition 1.7. *Given* $(\alpha_1, \dots, \alpha_k)$ *with* $\alpha_j \geq 0 \ \forall j \in \{1, \dots, k\}$ *and* $\alpha_j > 0$ *for some* $j \in \{1, \dots, k\}$*, let us define the random variables* $v_j \overset{ind}{\sim} Beta\left(\alpha_j, \sum_{i=j+1}^{k} \alpha_i\right) \ \forall j = 1, \dots, k-1$ *and* $v_k \sim \delta_1$*. Then, if we define*

$$p_j = \begin{cases} v_1 & \text{if } j = 1 \\ v_j \prod_{i=1}^{j-1} (1 - v_i) & \text{if } j = 2, \dots, k \end{cases}$$

we have that

$$(p_1, \dots, p_k) \sim Dir(\alpha_1, \dots, \alpha_k)$$

Proof. From the definition of $v_1, \dots, v_k$ we have that $\sum_{j=1}^{k} p_k = 1$. Furthermore if we define the function $g(v_1, \dots, v_{k-1}) = \left(p_1 = v_1, p_2 = v_2(1 - v_1), \dots, p_{k-1} = v_{k-1} \prod_{j=1}^{k-2} (1 - v_j)\right)$ its inverse is $g^{-1}(p_1, \dots, p_{k-1}) = \left(v_1 = p_1, v_2 = \frac{p_2}{1-p_1}, \dots, v_{k-1} = \frac{p_{k-1}}{1 - \sum_{j=1}^{k-2} p_j}\right)$ and the Jacobian of the inverse transformation is $\left|\frac{\partial g^{-1}(p_1, \dots p_{k-1})}{\partial (p_1, \dots p_{k-1})}\right| = \frac{1}{1-p_1} \frac{1}{1-p_1-p_2} \dots \frac{1}{1 - \sum_{j=1}^{k-2} p_j}$.

Hence the joint density of $(p_1, \dots, p_{k-1})$ is

$$\begin{aligned} f_{p_1, \dots, p_{k-1}}(p_1, \dots, p_{k-1}) &= f_{v_1, \dots, v_{k-1}}\left(p_1, \frac{p_2}{1-p_1}, \dots, \frac{p_{k-1}}{1 - \sum_{j=1}^{k-2} p_j}\right) \frac{1}{1-p_1} \frac{1}{1-p_1-p_2} \dots \frac{1}{1 - \sum_{j=1}^{k-2} p_j} = \\ &= f_{v_1}(p_1) \dots f_{v_{k-1}}\left(\frac{p_{k-1}}{1 - \sum_{j=1}^{k-2} p_j}\right) \frac{1}{1-p_1} \frac{1}{1-p_1-p_2} \dots \frac{1}{1 - \sum_{j=1}^{k-2} p_j} \propto \\ &\propto p_1^{\alpha_1 - 1} (1 - p_1)^{\sum_{j=2}^{k} \alpha_j - 1} \left(\frac{p_2}{1-p_1}\right)^{\alpha_2 - 1} \left(\frac{1-p_1-p_2}{1-p_1}\right)^{\sum_{j=3}^{k} \alpha_j - 1} \times \dots \times \\ &\times \left(\frac{p_{k-1}}{1 - \sum_{j=1}^{k-2} p_j}\right)^{\alpha_{k-1} - 1} \left(\frac{1 - \sum_{j=1}^{k-1} p_j}{1 - \sum_{j=1}^{k-2} p_j}\right)^{\alpha_k - 1} \times \frac{1}{1-p_1} \frac{1}{1-p_1-p_2} \dots \frac{1}{1 - \sum_{j=1}^{k-2} p_j} \propto \\ &\propto p_1^{\alpha_1 - 1} p_2^{\alpha_2 - 1} \dots p_{k-1}^{\alpha_{k-1} - 1} \left(1 - \sum_{j=1}^{k-1} p_j\right)^{\alpha_k - 1} \end{aligned}$$

□

1.3.2 Dirichlet process

Definition 1.4. Let α be a non-null finite regular Borel measure (non-negative and σ-additive[2]) on $(\mathscr{X}, \mathscr{A})$. We say that a random probability measure P on $(\mathscr{X}, \mathscr{A})$ possess a Dirichlet process distribution with base measure α (shortly, $P \sim DP(\alpha)$) if we have

$$(P(B_1), \ldots, P(B_k)) \sim \text{Dir}(\alpha(B_1), \ldots, \alpha(B_k)),$$

$\forall k \in \mathbb{N}$ and $\forall \{B_1, \ldots, B_k\}$ measurable partition of $\mathscr{X}$. We will use the notation $|\alpha| = \alpha(\mathscr{X})$ for the total mass of the base measure (also called "concentration parameter"), $\bar{\alpha} = \frac{\alpha}{|\alpha|}$ for the probability measure obtained normalizing the base measure (also called "centred base measure") and we will write $P \sim DP(\alpha)$ or $P \sim DP(|\alpha|, \bar{\alpha})$ to indicate that the random probability measure P has a Dirichlet process distribution with given base measure.

The existence of the Dirichlet process is not obvious, so we will now prove that this is a good definition .

Definition 1.4 specifies the marginal distribution for any measurable partition of the sample space. Let us first prove that the given specification of distributions can be extended to any vector of the type $(P(A_1), \ldots, P(A_k))$, for any class of measurable sets (not only partitions) respecting the consistency conditions required by Kolmogorov's extension theorem.

An arbitrary class of measurable sets $(A_1, \ldots, A_k)$ defines a collection of 2^k atoms of the form $A_1^\star \cap \ldots \cap A_k^\star$, where $A^\star$ stands for A or A^c. These atoms $\{B_j\}_{j=1}^{2^k}$ form a partition of the sample space and hence the joint distribution of $\left(P(B_j) : j = 1, \ldots, 2^k\right)$ is defined by Definition 1.4. Furthermore every A_i can be written as a union of atoms and $P(A_i)$ can be defined accordingly as the sum of the respective $P(B_j)\,s$, so we have a definition for the law of $(P(A_1), \ldots, P(A_k))$.

To prove the consistency conditions required by Kolmogorov's theorem let us consider the distribution of the vector $(P(A_1), \ldots, P(A_{k-1}))$. This distribution has been defined using the coarser partitioning in the 2^{k-1} sets of the form $A_1^\star \cap \ldots \cap A_{k-1}^\star$ and every set in this coarser partition is necessarily a union of two stets in the finer partition used before to define the distribution of $(P(A_1), \ldots, P(A_k))$. Therefore, to prove the required consistency conditions for every class of measurable sets, it is enough to prove the same conditions for every couple of partition, one finer than the other. Let $\{B_1, \ldots, B_k\}$ be measurable partition of $\mathscr{X}$ and $\{B_{i1}, B_{i2}\}$ be a further measurable partition of B_i for every $i = 1, \ldots, k$. Then Definition 1.4 specifies that

$$\begin{gathered}(P(B_{11}), P(B_{12}), P(B_{21}), \ldots, P(B_{k1}), P(B_{k2})) \sim \\ \text{Dir}(\alpha(B_{11}), \alpha(B_{12}), \alpha(B_{21}), \ldots, \alpha(B_{k1}), \alpha(B_{k2}))\end{gathered}$$

and, this implies that,

$$\begin{gathered}(P(B_1) = P(B_{11}) + P(B_{12}), \ldots, P(B_k) = P(B_{k1}) + P(B_{k2})) \sim \\ \text{Dir}(\alpha(B_1) = \alpha(B_{11}) + \alpha(B_{12}), \ldots, \alpha(B_k) = \alpha(B_{k1}) + \alpha(B_{k2}))\end{gathered}$$

because of Proposition 1.4 and finite additivity of the base measure. Hence the consistency condition is respected for any partition and so it is proved that there exists a stochastic process with finite dimensional laws as the ones specified in Definition 1.4.

[2] In the original definition in Ferguson (1973), was required only finite additivity.

Furthermore directly from Definition 1.4 we also have that $(P(A), P(A^c)) \sim \text{Dir}(\alpha(A), \alpha(A^c))$ and that $P(A) \sim \text{Beta}(\alpha(A), |\alpha| - \alpha(A))$ for every measurable set. Hence $P(\emptyset) = 0$ $\mathbb{P}$-a.s., $P(\mathscr{X}) = 1$ $\mathbb{P}$-a.s. and the mean measure is $A \mapsto E[P(A)] = \bar{\alpha}(A)$, which is a regular Borel measure by assumption. Thus Theorem 1.2 implies the existence of the Dirichlet process distribution with base measure α.

We will now present some propositions that shows the relationship between the properties of the random probability measure P and the properties of the parameter measure α.

Proposition 1.8. *Let P be a DP(α) on $(\mathscr{X}, \mathscr{A})$ and let $A \in \mathscr{A}$. If $\alpha(A) = 0$, then $P(A) = 0$ $\mathbb{P}$-a.s. and if $\alpha(A) > 0$ then $P(A) > 0$ $\mathbb{P}$-a.s..*

Proof. Let us consider the measurable partition $\{A, A^c\}$. Then $P(A) \sim \text{Beta}(\alpha(A), \alpha(A^c))$. So, if $\alpha(A) = 0$ then $P(A) \sim \delta_0$ (i.e. $\mathbb{P}\{P(A) = 0\} = 1$) and if $\alpha(A) > 0$ then $P(A) \nsim \delta_0$ (and, since the Beta distribution with positive parameters is a continuous distribution, $\mathbb{P}\{P(A) = 0\} = 0$). □

It is important to enlighten that the $\mathbb{P}$-null set outside of which the conclusion of Proposition 1.8 holds may depend on A. Hence Proposition 1.8 does not imply that α and the random probabilities P drawn from a $DP(\alpha)$ are mutually absolutely continuous.

Proposition 1.9. *Let P be a DP(α) on $(\mathscr{X}, \mathscr{A})$. If α is σ-additive then so is P, in the sense that for a decreasing sequence of measurable sets $A_m \searrow \emptyset$ we have $P(A_m) \underset{m \to +\infty}{\longrightarrow} 0$ $\mathbb{P}$-a.s..*

Proof. Since $A_m \searrow \emptyset$ and α is σ-additive we have that $\alpha(A_m) \underset{m \to \infty}{\longrightarrow} 0$. Hence there exists a subsequence $\{m_j\}$ s.t. $\sum_{j=1}^{\infty} \alpha(A_{m_j}) < \infty$. Using the Chebichev inequality we have, for fixed $\varepsilon > 0$,

$$\sum_{j=1}^{\infty} \mathbb{P}\{P(A_{m_j}) > \varepsilon\} \leq \frac{1}{\varepsilon} \sum_{j=1}^{\infty} E[P(A_{m_j})] = \frac{1}{\varepsilon} \sum_{j=1}^{\infty} \frac{\alpha(A_{m_j})}{|\alpha|} < \infty.$$

Hence, from the Borel-Cantelli lemma, we have that $\mathbb{P}\left\{\limsup_{j \to \infty} P(A_{m_j}) > \varepsilon\right\} = 0$ that implies that $\mathbb{P}\left\{\lim_{j \to \infty} P(A_{m_j}) = 0\right\} = 1$. Furthermore, since $A_m \searrow \emptyset$, we have that $P(A_1) > P(A_2) > \ldots$ $\mathbb{P}$-a.s. and so we can conclude that $\mathbb{P}\left\{\lim_{m \to \infty} P(A_m) = 0\right\} = 1$. □

Proposition 1.10. *Let P be a DP(α) on $(\mathscr{X}, \mathscr{A})$ and let Q be a fixed probability measure on $(\mathscr{X}, \mathscr{A})$ with $Q \ll \alpha$. Then $\forall m \in \mathbb{N}$, $\forall A_1, \ldots, A_m \in \mathscr{A}$ and $\forall \varepsilon > 0$ we have*

$$\mathbb{P}\{|P(A_i) - Q(A_i)| < \varepsilon \text{ for } i = 1, \ldots, m\} > 0.$$

Proof. Let $B_{\nu_1 \ldots \nu_m} = \bigcap_{j=1}^{m} A_j^{\nu_j}$ with $v_j = 0, 1$, $A_j^1 = A_j$ and $A_j^0 = A_j^c$. Then $P(A_i) = \sum_{\{\nu_1 \ldots \nu_m\} \ni \nu_i = 1} P(B_{\nu_1 \ldots \nu_m})$ and hence

$$\mathbb{P}\{|P(A_i) - Q(A_i)| < \varepsilon \text{ for } i = 1, \ldots, m\} \geq \mathbb{P}\left\{\sum_{\{\nu_1 \ldots \nu_m\} \ni \nu_i = 1} |P(B_{\nu_1 \ldots \nu_m}) - Q(B_{\nu_1 \ldots \nu_m})| < \varepsilon \text{ for } i = 1, \ldots, m\right\}.$$

Therefore it is sufficient to show that

$$\mathbb{P}\left\{\left|P\left(B_{\nu_1\ldots\nu_m}\right)-Q\left(B_{\nu_1\ldots\nu_m}\right)\right|<\frac{\varepsilon}{2^m}\text{ for all }\left(\nu_1\ldots\nu_m\right)\right\}>0.$$

If $\alpha\left(B_{\nu_1\ldots\nu_m}\right)=0$ then $Q\left(B_{\nu_1\ldots\nu_m}\right)=0$ (since $Q\ll\alpha$) and $P\left(B_{\nu_1\ldots\nu_m}\right)=0$ $\mathbb{P}$-a.s. (because of Proposition 1.8), so that $\left|P\left(B_{\nu_1\ldots\nu_m}\right)-Q\left(B_{\nu_1\ldots\nu_m}\right)\right|=0$ $\mathbb{P}$-a.s.. For those $\left(\nu_1\ldots\nu_m\right)$ for which $\alpha\left(B_{\nu_1\ldots\nu_m}\right)>0$, because of Proposition 1.8, the distribution of the corresponding $P\left(B_{\nu_1\ldots\nu_m}\right)$ gives positive weight to all open subsets of

$$\sum_{\{\nu_1\ldots\nu_m\}\text{s.t.}\alpha\left(B_{\nu_1\ldots\nu_m}\right)>0}P\left(B_{\nu_1\ldots\nu_m}\right)=1$$

and this completes the proof. □

Remark. This proposition states that the support of the DP(α) is "large". The discussion of the support of a random probability measure depends on the topology considered on the space of probability measures on $(\mathscr{X},\mathscr{A})$. If the topology chosen would have been the one of pointwise convergence (i.e. $Q_n\to Q$ iff $Q_n\left(A\right)\to Q\left(A\right)$ $\forall A\in\mathscr{A}$), by Proposition 1.10 the support of the DP(α) on $(\mathscr{X},\mathscr{A})$ would have contained only the set of all probability measures that are absolutely continuous with respect to α (and, conversely, any probability measure that is not absolutely continuous with respect to α would not have been included).

With the topology of the weak convergence (that is commonly adopted and that we described in the previous section) it may be shown that, if α is σ-additive, the support of P is the set of all σ-additive probability measures whose support is contained in the support of α. So the adoption of the σ-field $\mathscr{W}$ of subsets of $[0,1]^{\mathscr{A}}$ induced by the topology of the weak convergence is once again practically convenient.

Definition 1.5. Let $P\sim DP\left(\alpha\right)$ on $(\mathscr{X},\mathscr{A})$. We say that $X_1,\ldots,X_n$ is an independent and identically distributed sample of size n from the Dirichlet process if, for any $m=1,2,\ldots$ and measurable sets $A_1,\ldots,A_m,C_1,\ldots,C_n$

$$\Pr\left\{X_1\in C_1,\ldots,X_n\in C_n\left|P\left(A_1\right),\ldots,P\left(A_m\right),P\left(C_1\right),\ldots,P\left(C_n\right)\right.\right\}=\prod_{j=1}^{n}P\left(C_j\right)\ \mathbb{P}-a.s.\quad(1.3.1)$$

Roughly speaking, the definition given above means that $X_1,\ldots,X_n$ is an independent sample of size n sampled from a distribution P which was drawn from a Dirichlet process distribution on $\left([0,1]^{\mathscr{A}},\mathscr{W}\right)$. We will adopt the widely used notation:

$$\begin{aligned}&P\sim DP\left(\alpha\right)\\&X_1,\ldots,X_n\left|P\right.\overset{iid}{\sim}P.\end{aligned}$$

This definition determines the joint distribution of $X_1,\ldots,X_n,P\left(A_1\right),\ldots,P\left(A_m\right)$, since

$$\begin{aligned}&\Pr\left\{P\left(A_1\right)\leq y_1,\ldots,P\left(A_m\right)\leq y_m,\,X_1\in C_1,\ldots,X_n\in C_n\right\}=\\&=\int_0^{y_1}\cdots\int_0^{y_m}\int_{[0,1]^n}\prod_{j=1}^{n}P\left(C_j\right)dD_{P(A_1),\ldots,P(A_m),P(C_1),\ldots,P(C_n)}\end{aligned}\quad(1.3.2)$$

can be found integrating (1.3.1) with respect to the joint distribution of $P(A_1), \ldots, P(A_m), P(C_1), \ldots, P(C_n)$ over $[0, y_1] \times \ldots \times [0, y_m] \times [0,1]^n$. The Kolmogorov consistency conditions may easily be checked to show that (1.3.2) determines a probability measure $\mathscr{P}$ on the product space $\left(\mathscr{X}^n \times [0,1]^{\mathscr{A}}, \mathscr{A}^n \otimes \mathscr{W}\right)$.

From this consideration we obtain the subsequent properties for a DP.

Proposition 1.11. *Let $P \sim DP(\alpha)$ on $(\mathscr{X}, \mathscr{A})$ and let X be a sample of size 1 from P. Then*

$$\mathscr{P}\{X \in A\} = \frac{\alpha(A)}{|\alpha|} = \bar{\alpha}(A).$$

Proof. Since $\mathscr{P}\{X \in A | P(A)\} = P(A)$ $\mathbb{P}$-a.s.,

$$\mathscr{P}\{X \in A\} = E[\mathscr{P}\{X \in A | P(A)\}] = E[P(A)] = \frac{\alpha(A)}{|\alpha|}.$$

□

Proposition 1.12. *Let $P \sim DP(\alpha)$ on $(\mathscr{X}, \mathscr{A})$ and let X be a sample of size 1 from P. Let $\{B_1, \ldots, B_k\}$ be a measurable partition of $\mathscr{X}$ and $A \in \mathscr{A}$. Then*

$$\begin{aligned} &\mathscr{P}\{X \in A, P(B_1) \leq y_1, \ldots, P(B_k) \leq y_k\} = \\ &= \sum_{j=1}^{k} \frac{\alpha(B_j \cap A)}{|\alpha|} D(y_1, \ldots, y_k | \alpha(B_1), \ldots, \alpha(B_{j-1}), \alpha(B_j)+1, \alpha(B_{j+1}), \ldots, \alpha(B_k)), \end{aligned} \tag{1.3.3}$$

where $D(y_1, \ldots, y_k | \alpha(B_1), \ldots, \alpha(B_{j-1}), \alpha(B_j)+1, \alpha(B_{j+1}), \ldots, \alpha(B_k))$ is the distribution function of the Dirichlet distribution with parameters $(\alpha(B_1), \ldots, \alpha(B_{j-1}), \alpha(B_j)+1, \alpha(B_{j+1}), \ldots, \alpha(B_k))$ evaluated in $(y_1, \ldots, y_k)$.

Proof. Define $B_{j,1} = B_j \cap A$ and $B_{j,0} = B_j \cap A^c$ for all $j = 1, \ldots, k$. Let $Y_{j,\nu} = P(B_{j,\nu})$ for all $j = 1, \ldots, k$ and $\nu = 0, 1$. Then, from (1.3.1),

$$\mathscr{P}\left\{X \in A \middle| \{Y_{j,\nu}\}_{j=1,\ldots,k}^{\nu=0,1}\right\} = \sum_{j=1}^{k} Y_{j,1} \ \mathbb{P} - \text{a.s.} \tag{1.3.4}$$

Hence, for arbitrary $y_{j,\nu} \in [0,1]$, for $j = 1, \ldots, k$ and $\nu = 0, 1$,

$$\mathscr{P}\{X \in A, Y_{j,\nu} \leq y_{j,\nu}, \text{ for } j = 1, \ldots, k \text{ and } \nu = 0, 1\}$$

can be found by integrating (1.3.4) w.r.t. the Dirichlet distribution of the $Y_{j,\nu}s$ over the set $\{Y_{j,\nu} \leq y_{j,\nu},\ j = 1, \ldots, k \text{ and } \nu = 0, 1\}$:

$$\begin{aligned} &\sum_{j=1}^{k} \int_0^{y_{1,0}} \cdots \int_0^{y_{k,0}} \int_0^{y_{1,1}} \cdots \int_0^{y_{k,1}} Y_{j,1} dD(Y_{1,0}, \ldots, Y_{k,0}, Y_{1,1}, \ldots, Y_{k,1} | \alpha(B_{1,0}), \ldots, \alpha(B_{k,0}), \alpha(B_{1,1}), \ldots, \alpha(B_{k,1})) = \\ &= \sum_{j=1}^{k} \frac{\alpha(B_{j,1})}{|\alpha|} D(y_{1,0}, \ldots, y_{k,0}, y_{1,1}, \ldots, y_{k,1} | \alpha(B_{1,0}), \ldots, \alpha(B_{j,0})+1, \ldots, \alpha(B_{k,0}), \alpha(B_{1,1}), \ldots, \alpha(B_{j,1})+1, \ldots, \alpha(B_{k,1})). \end{aligned}$$

Since $P(B_j) = Y_{j,1} + Y_{j,0}$ $\mathbb{P}$-a.s., the conclusion of the proof follows using Proposition 1.4 because the process of finding marginal distributions (i.e. the multi-integral) is linear. □

One of the most remarkable properties of the Dirichlet process prior is that the posterior distribution is again Dirichlet process.

Theorem 1.4. *Let $P \sim DP(\alpha)$ on $(\mathscr{X}, \mathscr{A})$ and let $X_1, \ldots, X_n | P \sim P$. Then*

$$P | X_1, \ldots, X_n \sim DP\left(\alpha + \sum_{j=1}^{n} \delta_{X_j}\right).$$

Proof. It is sufficient to prove the theorem for $n = 1$, since the general case then follows by induction upon repeated application of the case $n = 1$.

Let $\{B_1, \ldots, B_k\}$ be a measurable partition of $\mathscr{X}$ and let $A \in \mathscr{A}$. The marginal distributions of the conditional distribution of a process are identical to the conditional distributions of the marginals (given a stochastic process $P = \{\xi_t\}_{t \in T}$ on a Polish space, there exists a version of the random variable $P|X$ that can be represented as the family $\{\xi_t | X\}_{t \in T}$). Hence we must show that the conditional distribution of $(P(B_1), \ldots, P(B_k))$ given X has distribution function

$$D(y_1, \ldots, y_k | \alpha(B_1) + \delta_X(B_1), \ldots, \alpha(B_k) + \delta_X(B_k)). \tag{1.3.5}$$

This may be done showing that the integral of (1.3.5) over A with respect to the marginal distribution of X is equal to the probability (1.3.3), since if, after the integration, we obtain the joint distribution of $X, P(B_1), \ldots, P(B_k)$ this means that the integrand was the conditional distribution of $P(B_1), \ldots, P(B_k)$ given X. Using the marginal probability of X as found in Proposition 1.11, we compute

$$\begin{aligned}
&\int_A D(y_1, \ldots, y_k | \alpha(B_1) + \delta_x(B_1), \ldots, \alpha(B_k) + \delta_x(B_k)) \, d\frac{\alpha(x)}{|\alpha|} = \\
&= \sum_{j=1}^{k} \int_{B_j \cap A} D(y_1, \ldots, y_k | \alpha(B_1), \ldots, \alpha(B_{j-1}), \alpha(B_j) + 1, \alpha(B_{j+1}), \ldots, \alpha(B_k)) \, d\frac{\alpha(x)}{|\alpha|} \\
&= \sum_{j=1}^{k} \frac{\alpha(B_j \cap A)}{|\alpha|} D(y_1, \ldots, y_k | \alpha(B_1), \ldots, \alpha(B_{j-1}), \alpha(B_j) + 1, \alpha(B_{j+1}), \ldots, \alpha(B_k)),
\end{aligned}$$

which completes the proof. □

This theorem essentially gives an update rule for the base measure of the Dirichlet process distribution when passing from the prior to the posterior. In terms of the parametrization $(|\alpha|, \alpha)$ it is

$$\begin{aligned}
|\alpha| &\mapsto |\alpha| + n \\
\bar{\alpha} &\mapsto \frac{|\alpha|}{|\alpha| + n} \bar{\alpha} + \frac{n}{|\alpha| + n} \mathbb{F}_n
\end{aligned}$$

where $\mathbb{F}_n = \frac{1}{n} \sum_{j=1}^{n} \delta_{X_j}$ is the empirical distribution of the sample $X_1, \ldots, X_n$. Since the mean measure of a Dirichlet process is the normalized base measure we obtain

$$E[P(A) | X_1, \ldots, X_n] = \frac{|\alpha|}{|\alpha| + n} \bar{\alpha}(A) + \frac{n}{|\alpha| + n} \mathbb{F}_n(A), \tag{1.3.6}$$

that means that the posterior mean (i.e. the "Bayes estimator" of P) is a convex combination of the prior mean $\bar{\alpha}$ and the empirical distribution of the data. For a given sample size n, the posterior mean is close to the prior mean if the concentration parameter $|\alpha|$ is large and is close to the empirical distribution if the concentration parameter is small[3]. Thus $|\alpha|$ determines the extent to which the prior controls the posterior mean.

Furthermore equation (1.3.6) implies that, for a fixed $|\alpha|$, the posterior mean behaves asymptotically as $n \to \infty$, $\mathbb{P}-a.s.$, like the empirical distribution of the sample to the order $O\left(n^{-1}\right)$. Hence it possesses the same asymptotic properties as the empirical distribution: in particular, if $X_1, X_2, \ldots$ are sampled from a "true distribution" P_0 then the posterior mean will tend to P_0 $\mathbb{P}-a.s.$.

Combining Theorem 1.4 with the formula for the variance of a Dirichlet variable we obtain that

$$\operatorname{Var}\left(P\left(A\right)|X_1,\ldots,X_n\right) = \frac{E\left[P\left(A\right)|X_1,\ldots,X_n\right]E\left[P\left(A^c\right)|X_1,\ldots,X_n\right]}{1+|\alpha|+n} \leq \frac{1}{4\left(1+|\alpha|+n\right)}. \tag{1.3.7}$$

Consequently, the posterior distribution contracts to its mean as $n \to \infty$: this implies that, if the data are sampled from a "true distribution" P_0, then the posterior distribution of P converges weakly to the measure degenerate at P_0.

1.3.3 Stick-breaking representation

The stick-breaking representation of the Dirichlet process due to Sethuraman (1994), allows to represent the Dirichlet process as a discrete random measure of the type expressed in equation (1.2.2) with stick-breaking weights.

Before approaching the Sethuraman's construction it is necessary to prove a preliminary Lemma.

Lemma 1.1. *Let us consider a finite, non-negative, regular Borel measure α on $(\mathscr{X}, \mathscr{A})$. For given independent random variables $\theta \sim \bar{\alpha}$ and $Y \sim Beta\left(1, |\alpha|\right)$, the Dirichlet process $DP\left(\alpha\right)$ is the unique solution of the distributional equation*[4]

$$P =_d Y\delta_\theta + \left(1-Y\right)P. \tag{1.3.8}$$

Proof. For a given measurable partition $\{B_1, \ldots, B_k\}$ the equation requires that

$$Q \coloneqq \left(P\left(B_1\right), \ldots, P\left(B_k\right)\right)$$

has the same distribution as the vector $YN + (1-Y)\,Q$ for $N \sim \text{Multinomial}\left(1; \bar{\alpha}\left(A_1\right), \ldots, \bar{\alpha}\left(A_k\right)\right)$ and (Y, N) independent of Q.

We first show that the solution is unique in distribution. Let (Y_n, N_n) be a sequence of i.i.d. copies of (Y, N) and, for two solutions Q and Q' of the equation that are independent of this sequence and suitably defined on the same probability space, set $Q_0 = Q$, $Q'_0 = Q'$ and recursively define

$$\begin{aligned} Q_n &= Y_n N_n + \left(1-Y_n\right)Q_{n-1} \\ Q'_n &= Y_n N_n + \left(1-Y_n\right)Q'_{n-1} \end{aligned} \quad \forall n \in \mathbb{N}.$$

[3] Clearly the terms "large" and "small" are to be intended with respect to the sample size n

[4] We say that a random probability measure that is independent of (Y, θ) is a solution of the distributional equation (1.3.8) if for every measurable partition of the sample space the random vectors obtained by evaluating the random probability measures on the right and left side of the equation have the same distribution in $\mathbb{R}^k$.

Then, since Q and Q' are solutions of equation (1.3.8), every Q_n is distributed as Q and every Q'_n is distributed as Q'_n. Also

$$\|Q_n - Q'_n\| = |1 - Y_n| \left\|Q_{n-1} - Q'_{n-1}\right\| = \prod_{i=1}^{n} |1 - Y_i| \|Q - Q'\| \underset{n\to\infty}{\longrightarrow} 0 \ a.s.$$

since the Y_i are i.i.d. and are in $(0,1)$ with probability 1. This proves the uniqueness of the solution for the distributional equation (1.3.8).

To prove that the Dirichlet process is the solution let $G_0, G_1, \ldots, G_k \overset{ind}{\sim} \text{Gamma}(\alpha_i, 1)$ for $i = 0, \ldots, k$ and $\alpha_0 = 1$. Then the vector $\left(G_0, G = \sum_{j=1}^{k} G_j\right)$ is independent of the vector

$$Q \coloneqq \left(\frac{G_1}{G}, \ldots, \frac{G_k}{G}\right) \sim \text{Dir}(\alpha_1, \ldots, \alpha_k).$$

Furthermore $Y \coloneqq \frac{G_0}{G_0+G} \sim \text{Beta}(1, |\alpha|)$ and $(Y, (1-Y)Q) \sim \text{Dir}(1, \alpha_1, \ldots, \alpha_k)$. Thus, for any $i = 1, \ldots, k$, merging the 0th cell with the ith, we obtain from Proposition 1.4 that, with e_i the ith unit vector,

$$Ye_i + (1-Y)Q \sim \text{Dir}(\alpha_1, \ldots, \alpha_{i-1}, \alpha_i + 1, \alpha_{i+1}, \ldots, \alpha_k)$$

and this gives the conditional distribution of the vector $YN + (1-Y)Q$ given $N = e_i$. It follows that $YN + (1-Y)Q$ given N is distributed as $\text{Dir}(\alpha_1 + N_1, \ldots, \alpha_k + N_k)$, just as $(Y_1, \ldots, Y_k)$ given X in Proposition 1.6. Since also the marginal distributions of N are the same as the ones of X in Proposition 1.6 so must be the marginal distributions of $YN + (1-Y)Q$ and $(Y_1, \ldots, Y_k)$, i.e. $YN + (1-Y)Q \sim \text{Dir}(\alpha_1, \ldots, \alpha_k)$. □

We are now ready to give the subsequent representation theorem.

Theorem 1.5. *Let $\theta_1, \theta_2, \ldots \overset{iid}{\sim} \bar{\alpha}$ and $v_1, v_2, \ldots \overset{iid}{\sim} Beta(1, M)$ are independent random variables and $w_j = v_j \prod_{i=1}^{j-1} (1 - v_i)$, then $P = \sum_{j=1}^{\infty} w_j \delta_{\theta_j} \sim DP(M\bar{\alpha})$.*

Proof. Since $E\left[\prod_{i=1}^{j} (1 - v_i)\right] = \left(\frac{M}{1+M}\right)^j \underset{j\to\infty}{\longrightarrow} 0$, it follows from Proposition 1.3 that the vector of stick-breaking weights sums 1 a.s. and hence P is a probability measure a.s.. For $j \geq 2$ define $w'_j = v_j \prod_{i=2}^{j-1} (1 - v_i)$ and $\theta'_j = \theta_{j+1}$. Then $w_j = (1 - v_1) w'_{j-1}$ for every $j \geq 1$ and hence

$$P = w_1 \delta_{\theta_1} + \sum_{j=2}^{\infty} w_j \delta_{\theta_j} = v_1 \delta_{\theta_1} + (1 - v_1) \sum_{j=1}^{\infty} w'_j \delta_{\theta'_j}.$$

But the random probability measure $\sum_{j=1}^{\infty} w'_j \delta_{\theta'_j}$ has exactly the same distribution as P (since it has the same structure) and is independent of (v_1, θ_1). So we conclude that P satisfies the distributional equation (1.3.8) and hence we have from the previous Lemma that $P \sim DP(M\bar{\alpha})$. □

From this fundamental result follow to important corollaries.

Corollary 1.1. *The realizations from the Dirichlet process distribution are discrete probability measures* $\mathbb{P}-a.s.$.

Proof. This result stems directly from the atomic representation on the Dirichlet process distribution given in Theorem 1.5. □

Corollary 1.2. *The Dirichlet process distribution with base measure* α *has full support on* $\left([0,1]^{\mathscr{A}}, \mathscr{W}\right)$.

Proof. This result stems directly from Propositions 1.2 and 1.3. □

In order to represent the weight process $\mathbf{w} = \{w_1, w_2, ...\}$ constructed as described above we will use the notation $\mathbf{w} \sim GEM(\alpha)$ where GEM stands for Griffiths, Engen, McCloskey, as used in Pitman 2002 and Johnson, Kotz, and Balakrishnan 1997.

1.3.4 Properties of samples from Dirichlet Processes

We now enlighten some properties of samples from Dirichlet processes (as defined by Definition 1.5) that comes directly from the propositions stated above.

The joint distribution of a sequence $X_1, X_2, \ldots$ generated from $P \sim DP(\alpha)$ on $(\mathscr{X}, \mathscr{A})$ has a complicated structure but can be conveniently described by the sequence of its predictive distributions (i.e. the laws of X_1, $X_1 | X_2$, $X_3 | X_1 X_2$, ...).

From Proposition 1.11 we have that, marginally, each X_i is distributed as $\bar{\alpha}$. If X_1, X_2 were an i.i.d. sample from $\bar{\alpha}$ we would expect that $\mathscr{P}\{X_1 = X_2\} = 0$ but that is not the case. Although α is non-atomic the conditional distribution of P given X is a Dirichlet process with parameter $\alpha + \delta_X$ and that is an atomic measure with an atom of mass 1 at X. Hence the probability that X_2 is equal to X_1, given X_1 is $\frac{1}{|\alpha|+1}$. In other words, since $X_2 | P, X_1 \sim P$ and $P | X_1 \sim DP(\alpha + \delta_{X_1})$ then, by the same reasoning applied in the proof of Proposition 1.11, we have that

$$\begin{aligned}\mathscr{P}\{X_2 \in A | X_1\} &= E[\mathscr{P}\{X_2 \in A | P\} | X_1] = E[P(A) | X_1] = \\ &= \frac{\alpha(A) + \delta_{X_1}(A)}{|\alpha| + 1} = \frac{|\alpha|}{|\alpha| + 1}\bar{\alpha}(A) + \frac{1}{|\alpha| + 1}\delta_{X_1}(A)\end{aligned}$$

By the same logic, since $P | X_1, \ldots, X_{n-1} \sim DP\left(\alpha + \sum_{j=1}^{n-1} \delta_{X_j}\right)$ by Theorem 1.4, we obtain that

$$X_n | X_1, \ldots, X_{n-1} \sim \frac{|\alpha|}{|\alpha| + n - 1}\bar{\alpha} + \sum_{j=1}^{n-1} \frac{1}{|\alpha| + n - 1}\delta_{X_j} \tag{1.3.9}$$

and the probability of obtaining i equal samples is then calculated as

$$\mathscr{P}\{X_1 = \ldots = X_n\} = \frac{(n-1)!\,|\alpha|!}{(|\alpha| + n - 1)!} \ \forall n = 1, 2, \ldots$$

It is clear from equation (1.3.9) that the joint distribution of $(X_1, \ldots, X_n)$, being equal to the product of the predictive distribution, is not equal to the product of the single marginal distributions (i.e. $X_1, \ldots, X_i$ are not independent). But from (1.3.9) if follows also that the joint distribution of $(X_1, \ldots, X_n)$ is exactly the same as the distribution of $\left(X_{s(1)}, \ldots, X_{s(n)}\right)$ for every permutation s

of the elements of the set $\{1,\dots,n\}$ and this holds for every $n \in \mathbb{N}$. Thus any sequence $X_1, X_2, \dots$ generated from $P \sim DP(\alpha)$ on $(\mathscr{X}, \mathscr{A})$ is exchangeable.

Equation (1.3.9) can also be viewed as an iterative rule for building an urn initially composed by a infinitely continuous number of "balls" distributed according to $\bar{\alpha}$ and is thus also called "generalized Polya urn scheme". This property, studied first by Blackwell and MacQueen (1973), can be used to characterize the Dirichlet process.

Let us now consider the probability that X_n is a new value, distinct from any previous observation $X_1, \dots, X_{n-1}$, in the case that the base measure α is non-atomic. Under this hypothesis on α equation (1.3.9) implies that the nth value is different from all the previous ones if it is drawn from $\bar{\alpha}$. If we define the random variables

$$W_i = \begin{cases} 1 & \text{if } X_i \notin \{1, \dots, X_{i-1}\} \\ 0 & \text{otherwise,} \end{cases}$$

then W_i is independent from W_j for all $i \neq j$ and, by equation (1.3.9), we have that

$$\mathscr{P}\{W_i = 1\} = \frac{|\alpha|}{|\alpha| + i - 1} \ \forall i = 1, 2, \dots.$$

We can further define the number of distinct values obtained in the first n observations as $Z_n = \sum_{i=1}^{n} W_i$ and obtain the subsequent results.

Proposition 1.13. *Given a sample $X_1, X_2, \dots, X_n$ generated from $P \sim DP(\alpha)$ on $(\mathscr{X}, \mathscr{A})$ with α non-atomic and fixed mass $|\alpha|$, we have that*

1. $E[Z_n] \approx |\alpha| \log n \approx Var(Z_n)$
2. $\frac{Z_n}{\log n} \underset{n\to\infty}{\longrightarrow} |\alpha|$
3. $\frac{Z_n - E[Z_n]}{\sqrt{Var(Z_n)}} \Longrightarrow N(0,1)$

Proof. The first point follows from easy calculations:

$$E[Z_n] = \sum_{i=1}^{n} E[W_i] = \sum_{i=1}^{n} \mathscr{P}\{W_i = 1\} = \sum_{i=1}^{n} \frac{|\alpha|}{|\alpha| + i - 1} \approx |\alpha| \log(n)$$
$$\mathrm{Var}(Z_n) = \sum_{i=1}^{n} \mathrm{Var}[W_i] = \sum_{i=1}^{n} \frac{|\alpha|(i-1)}{(|\alpha| + i - 1)^2} \approx |\alpha| \log(n)$$

The second point follows from the strong law of large numbers for independent variables, since $E\left[\frac{Z_n}{\log(n)}\right] \to |\alpha|$ and

$$\sum_{i=1}^{\infty} \mathrm{Var}\left(\frac{Z_i}{\log i}\right) = \sum_{i=1}^{\infty} \frac{\mathrm{Var}(W_i)}{(\log i)^2} = \sum_{i=1}^{\infty} \frac{|\alpha|(i-1)}{(|\alpha| + i - 1)(\log i)^2} < \infty.$$

Finally the last point is a direct consequence of the Lindeberg central limit theorem. □

So, although $\mathscr{P}\{W_n = 1\} = \frac{|\alpha|}{|\alpha|+n-1} \searrow 0$ and thus new distinct values are increasingly rare, we are assured nonetheless of a steadily increasing number of distinct values[5]. Moreover, since the distribution of the distinct values is simply $\bar{\alpha}$, this property can be used to obtain information about the shape of $\alpha(\cdot)$ if it is unknown.

On the other hand, the rate at which new distinct values appear when increasing the sample depends only on the magnitude of $|\alpha|$, and not on the shape of $\alpha(\cdot)$. In terms of the stick breaking representation given in equation (1.3.8), this means that only the magnitude of $|\alpha|$ determines the differences in magnitude of the weights w_i: a small value of $|\alpha|$ implies that w_1 is big compared to w_2, this latter is big compared to w_3 (and so on...) while a large value of $|\alpha|$ means that there are many small weights that tail off to zero slowly. Since $|\alpha|$ characterizes the rate at which new values appear it is justified the name "concentration parameter" and is also justified the use of the notation $P \sim DP(|\alpha|, \bar{\alpha})$ in the cases in which it is of some interest the rate at which new values appear.

1.3.5 Mixtures of Dirichlet processes

The application of the Dirichlet process prior requires a choice of a base measure α. It is often reasonable to choose the centre measure $\bar{\alpha}$ from a specific family of statistical distributions depending on some parameters. It is then natural to give a further prior distribution to the parameters of the centre measure. Similarly, to incorporate prior beliefs and uncertainty on the stick-breaking weights, one can also put a prior on the precision parameter $|\alpha|$. For a base measure $\alpha(\xi)$ that depends on a parameter ξ the model then consists of the hierarchy

$$\begin{aligned} X_1, \ldots, X_n \,|P, \xi &\overset{iid}{\sim} P \\ P\,|\xi &\sim DP(\alpha(\xi)) \\ \xi &\sim \pi \end{aligned}$$

and the induced marginal prior on P, called "mixture Dirichlet prior", is usually denoted by $MDP(\alpha(\xi), \pi)$. Many properties of this mixture Dirichlet prior follow immediately from the those of Dirichlet process (e.g. any probability measure P drawn from a mixture Dirichlet process prior is almost surely discrete) and, given ξ, we can use theorem 1.4 to obtain

$$P\,|\xi, X_1, \ldots, X_n \sim DP\left(\alpha + \sum_{j=1}^{n} \delta_{X_j}\right).$$

To obtain the posterior distribution of P given the data we thus need to integrate this with respect to the posterior distribution of ξ given $X_1, \ldots, X_n$, which is proportional to $\pi(\xi)\, p(X_1, \ldots, X_n\,|\xi)$, obtaining that $P\,|X_1, \ldots, X_n \sim MDP\left(\alpha + \sum_{j=1}^{n} \delta_{X_j},\, \pi(\xi)\, p(X_1, \ldots, X_n\,|\xi)\right)$. The joint marginal density $p(X_1, \ldots, X_n\,|\xi)$ is described by the conditional distributions given in equation (1.3.9) with $\alpha(\xi)$ instead of α: in general this has a complicated structure due to the random ties between the observations. However, as far as the posterior estimation is concerned, we condition on the observed

[5] More precisely, the number of distinct values in a large sample from a distribution drawn from a fixed Dirichlet process prior is logarithmic in the sample size and the fluctuations of this number around its mean are of the order $\sqrt{\log n}$.

data $X_1, \ldots, X_n$ and know their ties. Given this information, the density $p(X_1, \ldots, X_n | \xi)$ takes a simpler form. If, for instance, the observations are distinct (which happens with probability 1 when the observations actually follow a continuous distribution), then equation (1.3.9) implies that $p(X_1, \ldots, X_n | \xi) = \prod_{i=1}^{n} \frac{\alpha_\xi(X_i)}{|\alpha(\xi)|+i-1}$, where $\alpha_\xi(\cdot)$ is a density of $\alpha(\xi)$.

1.3.6 Dirichlet process mixtures

Since the Dirichlet process prior is discrete, it is useless when we wish to estimate a continuous distribution but this can be remedied by convolving it with a continuous kernel density.

Definition 1.6. For each θ in a probability space $(\Theta, \mathscr{B}(\Theta), P)$ let $x \mapsto \varphi(x, \theta)$ be a probability density function measurable in its two arguments. If $P \sim DP(\alpha)$ then we say that the mixture density $p_P(x) = \int \varphi(x, \theta) dP(\theta)$ is a Dirichlet process mixture.

If the data are distributed as a Dirichlet process mixture, then the resulting statistical model, called "Dirichlet process mixture" (DPM) model, can be written as follow:

$$\begin{aligned} P &\sim DP(\alpha) \\ \theta_i | P &\overset{iid}{\sim} P \\ X_i | \theta_i, P &\overset{ind}{\sim} \varphi(\cdot, \theta_i) \end{aligned} \tag{1.3.10}$$

The following result, due to Antoniak (1974), shows that if we obtain an unobservable sample from a mixture Dirichlet process and the observed sample is distorted by a random error (i.e. the distribution drawn from the mixture Dirichlet process prior is used as mixing distribution for the mixture distribution of observed the data) then the posterior distribution of the process given the observed sample is again a mixture Dirichlet process.

Theorem 1.6. *Let $P \sim MDP(\alpha(\xi), \pi)$ on a Polish space $(\Theta, \mathscr{B}(\Theta))$ where π is a probability measure on a Polish space $(U, \mathscr{B}(U))$ and α is a transition measure on $U \times \mathscr{B}(\Theta)$. Let $(\mathscr{X}, \mathscr{B}(\mathscr{X}))$ be a Polish space and F a transition probability from $\Theta \times \mathscr{B}(\mathscr{X})$ to $[0, 1]$. If*

$$\begin{aligned} \xi &\sim H \\ P | \xi &\sim DP(\alpha(\xi)) \\ \theta | P, \xi &\sim P \\ X | P, \xi, \theta &\sim F(\theta, \cdot) \end{aligned}$$

then $P | X \sim MDP(\alpha(\xi) + \delta_\theta, p(\theta, \xi | X))$, where $p(\theta, \xi | X)$ is the conditional distribution of (θ, ξ) given X.

Proof. From Theorem 1.4 it follows that $P | \theta, \xi, X \sim DP(\alpha(\xi) + \delta_\theta)$, independently of X. From Proposition 1.11 we have that marginally (w.r.t. P) $\theta | \xi \sim \frac{\alpha(\xi, \cdot)}{\alpha(\xi, \Theta)}$. By hypothesis $X | P, \xi, \theta \sim F(\theta, \cdot)$ is independent of ξ and $\xi \sim H$. Hence the joint distribution of (θ, ξ, X) is defined by the (conditional) distributions of $X | P, \xi, \theta$, $\theta | \xi$ and ξ, and thus the conditional distribution of $(\theta, \xi) | X$ is well defined. Finally the conditional distribution of P given X is obtained integrating the conditional distribution of $P | \theta, \xi, X$ with respect to the conditional distribution of $(\theta, \xi) | X$: since $P | \theta, \xi, X \sim DP(\alpha(\xi) + \delta_\theta)$, we recognize that the conditional distribution $P | X$ is a mixture Dirichlet process with the conditional distribution of $(\theta, \xi) | X$ as mixing measure. □

From this result follows directly the subsequent Corollary which states that, if the observed data are distributed according to a Dirichlet process mixture model, the posterior distribution of the Dirichlet process given the observed data is a mixture Dirichlet process with mixing distribution the joint conditional distribution of the unobserved sample from the Dirichlet process (i.e. the parameters of the Dirichlet process mixture model) given the observed data.

Corollary 1.3. *Let $P \sim DP(\alpha)$ on a Polish space $(\Theta, \mathscr{B}(\Theta))$. Let $(\mathscr{X}, \mathscr{B}(\mathscr{X}))$ be a Polish space and $\varphi : \Theta \times \mathscr{X} \to \mathbb{R}$ such that $\theta \mapsto \varphi(x, \theta)$ is measurable for all $x \in \mathscr{X}$ and $x \mapsto \varphi(x, \theta)$ is a probability density function for all $\theta \in \Theta$. If*

$$\begin{aligned} P &\sim DP(\alpha) \\ \theta_i \,|P &\overset{iid}{\sim} P \qquad \text{for } i = 1, \dots, n \\ X_i \,|\theta_i, P &\overset{ind}{\sim} \varphi(\cdot, \theta_i) \end{aligned}$$

then $P\,|X_1, \dots, X_n \sim MDP\left(\alpha + \sum_{i=1}^{n} \delta_{\theta_i},\, p(\theta_1, \dots, \theta_n \,|X_1, \dots, X_n)\right)$, where $p(\theta_1, \dots, \theta_n \,|X_1, \dots, X_n)$ is the conditional distribution of $\theta_1, \dots, \theta_n$ given $X_1, \dots, X_n$.

Proof. From Theorem 1.4 it follows that $P\,|\theta_1, \dots, \theta_n, X_1, \dots, X_n \sim DP\left(\alpha + \sum_{i=1}^{n} \delta_{\theta_i}\right)$, independently of $X_1, \dots, X_n$. From Proposition 1.11 we have that marginally $\theta_i \overset{iid}{\sim} \bar{\alpha}$ and hence their joint distribution is simply the product of these marginals. By hypothesis $(X_1, \dots X_n)\,|\theta_1, \dots, \theta_n, P \sim \prod_{i=1}^{n} \varphi(\cdot, \theta_i)$. Hence the joint distribution of $(\theta_1, \dots, \theta_n, X_1, \dots, X_n)$ is well defined by the distributions of $(X_1, \dots X_n)\,|\theta_1, \dots, \theta_n, P$ and $\theta_1, \dots, \theta_n$, and thus the conditional distribution of $\theta_1, \dots, \theta_n \,|X_1, \dots, X_n$ is well defined too. Finally the conditional distribution of P given X is obtained integrating the conditional distribution of $P\,|\theta_1, \dots, \theta_n, X_1, \dots, X_n$ with respect to the conditional distribution of $\theta_1, \dots, \theta_n \,|X_1, \dots, X_n$:
since $P\,|\theta_1, \dots, \theta_n, X_1, \dots, X_n \sim DP\left(\alpha + \sum_{i=1}^{n} \delta_{\theta_i}\right)$, we recognize that the conditional distribution $P\,|X_1, \dots, X_n$ is a mixture Dirichlet process with the conditional distribution of $\theta_1, \dots, \theta_n \,|X_1, \dots, X_n$ as mixing measure. □

From this results it follows that the posterior mean density of a Dirichlet process mixture density $p_P(x) = \int \varphi(x, \theta)\, dP(\theta)$ given $\theta_1, \dots, \theta_n, X_1, \dots, X_n$ is

$$E\left[p_P(x)\,|\theta_1, \dots, \theta_n, X_1, \dots, X_n\right] = \frac{1}{|\alpha| + n}\left(\int \varphi(x, \theta)\, d\alpha(\theta) + \sum_{i=1}^{n} \varphi(x, \theta_i)\right) \tag{1.3.11}$$

which is a combination of a part attributable to the prior and a part due to observations (has it happens in equation (1.3.6) for the posterior mean of a Dirichlet process). It is also interesting to enlighten the role of the concentration parameter $|\alpha|$ in the context of a Dirichlet process mixture model. As we have already said at the end of subsection 1.3.4 the concentration parameter rules the appearance of new values (distinct from all the ones previously obtained) in samples from the Dirichlet process: in the context of Dirichlet process mixture models this means a large value of $|\alpha|$ is associated with a little number of mixture components with weights much heavier than the ones of the remaining components, while a small value of the concentration parameter is associated with a more even distribution of the weights between the components.

Chapter 2

Multiplicative error models

2.1 Introduction

Modelling non-negative time series is quite an important task in the econometric literature. For example one could think to the volume of shares traded over a 10-minute period or the difference between the highest price and the lowest price in a given time period or the ask price minus the bid price or again sales volumes, demand levels, insurance claims, weather derivatives, etc...

In particular, being able to accurately forecast the volatility of asset returns, which is a positive quantity, is vital in many fields of finance such as derivative pricing and risk management. Different variants of the ARCH model introduced by Engle (1982) and the stochastic volatility model Taylor (1987) have been widely employed to model daily data since the introduction of the original models. These are effectively models for the squared returns, where the conditional variance is a latent factor, and they produce relatively noisy forecasts of the volatility. With the recent wide availability of intra-daily data, it has become possible to measure directly the daily volatility more accurately than before. Andersen and Bollerslev (1998) showed that the realized variance, computed as the sum of squared intra-daily returns, provides a measure of the volatility with desirable properties. In particular Andersen, Bollerslev, Diebold, and Labys (2001) and Andersen, Bollerslev, Diebold, and Ebens (2001) demonstrated that the square root of realized variance (also called realized volatility) of stock and exchange rates can be described as unconditionally lognormally distributed and exhibiting long memory.

As remarked in Engle (2002), in the classical literature there are two conventional approaches to the analysis of these non-negative-valued processes: the first is to ignore the non-negativity, the second is to take the logs and, in both cases, then it is used a linear regression model to represent the observations.

Let us consider a univariate time series $\{x_t\}_{t=1}^T$ where $x_t \geq 0 \ \forall t = 1, \ldots, T$ and such that

$$\Pr\{x_t < \delta \,|x_{t-1}, \ldots, x_1\} > 0 \ \forall \delta > 0, \, t = 1, \ldots, T. \tag{2.1.1}$$

Let us indicate as $\mu_t = E[x_t \,|x_{t-1}, \ldots, x_1]$ and $\sigma_t^2 = \mathrm{Var}(x_t \,|x_{t-1}, \ldots, x_1)$ the conditional mean and the conditional variance of the series at time t given its past. Then, ignoring the non-negativity, we can specify a linear model as

$$x_t = \mu_t + \varepsilon_t, \ \varepsilon_t \,\big|\mathscr{F}_{t-1} \sim \mathscr{D}\left(0, \sigma_t^2\right)$$

where $\mathscr{F}_{t-1}$ is the information available at time $t-1$ and $\mathscr{D}$ is a continuous probability distribution with mean 0 and variance σ_t^2. It is clear that the definition of the distribution $\mathscr{D}$ is quite delicate since by the fact that the mean μ_t is positive and x_t is non negative implies that the disturbances cannot be more negative than the mean. Thus the support of the distribution $\mathscr{D}$ will be different for every observation. Furthermore, since the probability of a near zero is given by

$$\Pr\{x_t < \delta \,|x_{t-1}, \ldots, x_1\} = \Pr\{\varepsilon_t < \delta - \mu_t \,|x_{t-1}, \ldots, x_1\},$$

then the error distribution must be discontinuous at $-\mu_t$ to satisfy equation (2.1.1). These facts makes the estimation process quite cumbersome (both in the classical and in the Bayesian context).

If instead we consider the idea of taking the logs, the model becomes

$$\log(x_t) = m_t + u_t$$

and the conditional mean and variance become $\mu_t = e^{m_t} E[e^{u_t}]$ and $\sigma_t^2 = e^{2m_t} \mathrm{Var}(e^{u_t})$. But this solution will not work if there are exact zeros in the time series $\{x_t\}$ and the solution sometimes adopted to add a small constant to eliminate exact zeros is more of a theoretical solution than a practical one, since it has been proven that the finite sample estimates are typically heavily influenced by the size of this constant.

2.2 The univariate case

Since both the classical approaches reported above have been proven to be flawed Engle and Russell (1998) introduced the Autoregressive Conditional Duration (ACD) model to model the duration between transactions in financial markets. This model had then been generalized for a wider set of applications by Engle (2002) who also renamed it Multiplicative Error Model, also referred to with the acronym MEM. This is a model in which the data, at every time t, are specified as the product of a random term representing the disturbances in the data and a deterministic term representing the conditional mean of the data (given the past information at that time). The model is then completed specifying a distribution for the data and an equation that describes the dynamics of their conditional mean. The MEM approach has the advantage that it both reduces the difficulties in the specification of the distribution of the data and avoids the problems due to the log transformation of zero data, allowing an easy and parsimonious analysis of non-negative-valued time series.

2.2.1 Model formulation

Given a non-negative time series $\{x_t\}_{t=1}^T$, the MEM model is specified as

$$x_t = \mu_t \varepsilon_t \tag{2.2.1}$$

where

- $\mu_t = \mu(\boldsymbol{\eta}, \mathscr{F}_{t-1})$ is a positive, deterministic, term that evolves according to the vector of parameters $\boldsymbol{\eta}$.
- $\varepsilon_t \,|\mathscr{F}_{t-1} \sim \mathscr{D}(1, \sigma^2)$ where $\mathscr{D}(1, \sigma^2)$ is a continuous probability distribution on $[0, \infty)$ with unit mean and unknown constant variance σ^2.

Hence we have that

$$\begin{aligned} E\left[x_t \,|\mathscr{F}_{t-1}\right] &= \mu_t, \\ \mathrm{Var}\left[x_t \,|\mathscr{F}_{t-1}\right] &= \mu_t^2 \sigma^2. \end{aligned} \tag{2.2.2}$$

2.2.2 Specification of the innovations

In principle the conditional distribution of the innovations ε_t can be specified by means of any probability density with support on $[0, \infty)$ with unit mean. Examples include the Gamma, Log-Normal, Weibull, Inverted-Gamma and mixtures of them. Engle (2002) proposed an exponential distribution with unit mean, which allows for easy and consistent quasi-maximum likelihood estimates but lacked flexibility. Chou (2005), who applied a MEM to model the high/low range of prices in daily and weekly scales, used a Weibull distribution. Lanne (2006) adopted a mixture of unit-mean Gamma densities to model realized volatilities. Engle and Gallo (2006) choose Gamma distributions with equal parameters (so that it has unit mean and it is more flexible than the exponential) to model, simultaneously, squared returns, realized variances and high-low ranges.

An alternative strategy is to leave the distribution of the disturbances unspecified except for the two conditional moments in equation (2.2.2), to obtain a classical semiparametric specification of the model to be estimated using the generalized method of moments as done by Brownlees, Cipollini, and Gallo (2012).

Finally it is also possible to adopt a Bayesian semiparametric point of view to specify the distribution of the disturbances as done by Solgi and Mira (2013), who proposed to adopt a Dirichlet process mixture of Gamma distributions with two free parameter as distribution of the disturbances. The use of this rich family of distributions allowed them to approximate any continuous distribution on $[0, \infty)$ to any precision level, thus obtaining great flexibility and very good fit of the distribution of the data.

2.2.3 Specification of the conditional mean

Any practical specification of μ_t must depend on the available information $\mathscr{F}_{t-1}$ and must ensure that μ_t is always positive for every t. We will now show some examples from the literature for illustrative purposes but the list will not be exhaustive of the many different specifications of the conditional mean adopted for MEMs in the past decade.

Engle (2002), supposing the existence of a vector of weakly exogenous variables $\boldsymbol{z}_t$, proposed to specify μ_t as

$$\mu_t = \omega + \sum_{i=1}^{p} \alpha_i x_{t-i} + \sum_{i=1}^{q} \beta_i \mu_{t-i} + \boldsymbol{\gamma}' \boldsymbol{z}_t.$$

To ensure the positiveness for all possible realizations, if $\boldsymbol{z}$ are positive variables, it is sufficient to require that all the parameters are positive[1]. To ensure covariance stationarity for x_t it is sufficient that $\boldsymbol{z}_t$ is covariance stationary and $\sum_{i=1}^{p} \alpha_i + \sum_{i=1}^{q} \beta_i < 1$. With this structure, the persistence property[2] of x_t can be modelled parsimoniously and it is a well known fact that this model for μ_t is equivalent to an ARMA $(\max(p,q), q)$ model for x_t, with heteroskedastic errors specified as $\nu_t = x_t - \mu_t$.

[1] Note however that this is not a necessary condition.

[2] A time series process is called "persistent" if the effect of infinitesimally small shock will be influencing the future predictions of the time series for a very long time.

Removing the exogenous variables and posing $p = q = 1$ we obtain the simplest specification, also called "baseline" MEM

$$\mu_t = \omega + \alpha_1 x_{t-1} + \beta_1 \mu_{t-1} \tag{2.2.3}$$

which often in the literature took the role of benchmark model and is usually sufficient in empirical studies.

It is a known fact that some financial time series, like the realized volatility of an equity, reacts asymmetrically in response to positive and negative realizations of some exogenous variable, like the returns of the same asset. It had then been introduced the so called asymmetric MEM, in which the conditional mean of the non-negative process $\{x_t\}$ is specified as

$$\mu_t = \omega + \sum_{i=1}^{p} \alpha_i x_{t-i} + \sum_{i=1}^{q} \beta_i \mu_{t-i} + \sum_{i=1}^{s} \gamma_i x_{t-i} \mathbb{I}_{\{r_{t-i}<0\}}, \tag{2.2.4}$$

where r_{t-i} denotes a lagged exogenous variable (usually the lagged returns) such that $E\left[x_{t-i}\mathbb{I}_{\{r_{t-i}<0\}} \,|\mathscr{F}_{t-1}\right] = \frac{\mu_t}{2}$ and $\mathbb{I}_{\{.\}}$is the indicator function. Also in this case the specification most used in practice is the one with $p = q = s = 1$. To ensure mean stationarity for the time series x_t is sufficient that $\sum_{i=1}^{p} \alpha_i + \sum_{i=1}^{q} \beta_i + \sum_{i=1}^{s} \frac{\gamma_i}{2} < 1$ and again the positiveness of the conditional mean is trivially guaranteed if all the parameters are positive (but, also in this case, this is not a necessary condition).

Beyond these two basic specifications for the conditional mean are possible many extensions in many directions, mainly depending on the characteristics of the time series under analysis. Since it is far beyond the aim of this work to describe all the possible specifications of the conditional mean in a MEM model, we limit ourselves to cite the Component MEM by Brownlees, Cipollini, and Gallo (2011) built to model intra-daily volumes of transactions and the specification introduced by Manganelli (2005) to model the durations between subsequent trades.

2.2.4 Estimation methods

Since the introduction of the multiplicative error models many methods had been proposed to estimate both the parameters of the conditional mean and the parameter of the distribution of the disturbances. Engle (2002), under the hypothesis of unit-mean exponential disturbances, proved that, under some mild conditions on the process $\frac{x_t}{\mu_t}$, the quasi-maximum likelihood estimator of the parameters of the conditional mean is consistent and asymptotically normal. Brownlees, Cipollini, and Gallo (2012), considering disturbances with Gamma distribution with equal shape and rate, showed that the maximum likelihood estimator of the parameters of the parameters of the conditional mean does not depend on the "dispersion parameter" of the disturbances. This last can be estimated again by maximum likelihood, based on the estimates of the parameters of μ_t. On the other hand, in the same paper, the authors propose also to resort to Generalized Method of Moments to estimate the parameters of μ_t (thus avoiding distributional assumptions on the disturbances) and give a formulation for the instruments that give the GMM estimator with smaller asymptotic variance.

As far as we know, the only Bayesian formulation of a MEM model is the semiparametric one given by Solgi and Mira (2013). In this case the estimation of the parameters of interest is carried out using a conditional method, known as "slice sampler". We will detail this algorithm in the next chapter.

2.3 The multivariate case

As described in the Introduction, to study multivariate time series defined on the non-negative orthant Cipollini, Engle, and Gallo (2006) proposed a multivariate version of the multiplicative error model, called vector MEM (vMEM). In this case different mono-dimensional time series are represented together as a single time series of vectors. The vectors of data are modelled as an element by element product of a random vector representing the disturbances and a deterministic vector representing the conditional mean of the data (given the past information). The model is then completed specifying a multidimensional unit-mean distribution for the data vector and a vector equation that describes the dynamics of the mean-vector of the data.

From now on we will indicate the non-negative orthant of $\mathbb{R}^n$ as $(\mathbb{R}^+)^n$ and we will use the notation $\boldsymbol{\iota}_n$ to refer to the n-dimensional vector with every component equal to 1.

Furthermore bold capital letters will in general refer to matrices, bold Greek letters will refers to vectors and normal Greek letters will refer to elements of the corresponding matrix/vector (e.g. α_{ij} is an element of the matrix $\mathbf{A}$ and ω_i is an element of the vector $\boldsymbol{\omega}$).

2.3.1 Model formulation

Given a d-dimensional time series $\{\mathbf{x}_t\}_t$ contained into $(\mathbb{R}^+)^d$, the vMEM model is specified as:

$$\mathbf{x}_t = \boldsymbol{\mu}_t \odot \boldsymbol{\varepsilon}_t \tag{2.3.1}$$

where

- $\boldsymbol{\mu}_t = \boldsymbol{\mu}(\boldsymbol{\eta}, \mathscr{F}_{t-1})$ is a deterministic vector with non-negative components that evolves according to the vector of parameters $\boldsymbol{\eta}$.
- $\boldsymbol{\varepsilon}_t | \mathscr{F}_{t-1} \sim \mathscr{D}(\boldsymbol{\iota}_d, \boldsymbol{\Sigma})$ where $\mathscr{D}(\boldsymbol{\iota}_d, \boldsymbol{\Sigma})$ is a continuous probability distribution on $(\mathbb{R}^+)^d$ with unit-vector mean $\boldsymbol{\iota}_d$ and unknown constant covariance matrix Σ.

Thus we have that

$$\begin{aligned} E[\mathbf{x}_t | \mathscr{F}_{t-1}] &= \boldsymbol{\mu}_t, \\ \mathrm{Var}[\mathbf{x}_t | \mathscr{F}_{t-1}] &= \boldsymbol{\mu}_t \boldsymbol{\mu}_t^{'} \odot \boldsymbol{\Sigma}. \end{aligned} \tag{2.3.2}$$

and hence $\mathrm{Var}[\mathbf{x}_t | \mathscr{F}_{t-1}]$ is guaranteed to be a positive definite matrix.

2.3.2 Specification of the innovations

As in the univariate case, in principle the conditional distribution of the innovations $\boldsymbol{\varepsilon}_t$ can be specified by means of any probability density with support on $\mathbb{R}^{+^d}$ with unit mean. The natural extension of the univariate case is to limit ourselves to the assumption that at each time t, given the past information, all the components, $\varepsilon_i^{(t)}$ of the vector $\boldsymbol{\varepsilon}_t$ are independently distributed with Gamma densities with equal shape and rate, ϕ_i. We could also use a multivariate log-normal distribution (defined as the distribution of the exponential transformation of a multivariate normal random variable).

Another possibility is to use one of the multivariate generalizations of the Gamma distribution described in Johnson, Kotz, and Balakrishnan (2000), but most of them are quite unhandy since they are defined via the joint characteristic function (and so require numerical inversion formulas

to find the corresponding probability density function). Because of this Cipollini, Engle, and Gallo (2006) chose to adopt the multivariate version of the Gamma distribution due to Cheryan and Ramabhadran with unit mean,

$$f_{\boldsymbol{\varepsilon}_t|\mathscr{F}_{t-1}}(\boldsymbol{x};\phi_0,\boldsymbol{\phi},\boldsymbol{\phi}) = \frac{1}{\Gamma(\phi_0)}\prod_{i=1}^{d}\frac{\phi_i}{\Gamma(\phi_i-\phi_0)}e^{-\sum_{i=1}^{d}\phi_i x_i}\int_0^{\min_{i=1,\dots,d}(\phi_i x_i)} y^{\phi_0-1}e^{(d-1)y}\prod_{i=1}^{d}(\phi_i x_i - y)^{\phi_i-\phi_0-1}\,dy$$

where $0 < \phi_0 < \min(\phi_1,\dots,\phi_d)$. In this case each $\varepsilon_i^{(t)}$ has univariate gamma marginal distribution with shape and rate equal to ϕ_i and $\text{Cov}\left(\varepsilon_i^{(t)},\varepsilon_j^{(t)}\right) = \frac{\phi_0}{\phi_i\phi_j}$ (so no negative correlations are allowed under this specification).

A different way to define the distribution of $\boldsymbol{\varepsilon}_t|\mathscr{F}_{t-1}$ is to start from the assumption that all the univariate marginal probability density functions are $Gamma(\phi_i,\phi_i)$ and then use copula functions. Many copula functions have been proposed in the literature: Cipollini, Engle, and Gallo (2006) choose to use a Normal copula because of its analytic tractability and the easy of simulation. The advantage of using copulas over a multivariate Gamma distribution are mainly that copulas allows more flexible covariance structures (allowing also negative correlations), the correlations do not depend anymore on the variances of the marginals (but are completely specified by the copula function adopted) and there are no complicated constraints on the parameters of the marginal distributions.

Anyway, as we could see from the previous example and had been enlightened also by Cipollini, Engle, and Gallo (2013), there are some problems in the complete specification of the conditional distribution:

- the distributions defined on the positive orthant are often not enough flexible or they are defined via the characteristic function.
- the distributions obtained via copulas often present not realistic symmetric behaviours due to the fact that the choice of the copula function is driven by computation reasons. Furthermore, even when combined with correct formulations of the marginals, copulas are not always able to model adequately the association among components of the error term. Finally, as far as the marginals are concerned, in principle each one of them could follow a different distribution. Hence a copula approach could require an expensive tuning.
- the exact specification of the distribution of the disturbance term may be not interesting[3].

Because all of these motivations, as we will describe in the next chapters, we will consider a Bayesian semiparametric framework and we will propose as conditional distribution for the disturbances a DPM of multivariate log-normal distributions.

2.3.3 Specification of the conditional mean

Just like in the univariate case, any practical definition of $\boldsymbol{\mu}_t$ must depend on $\mathscr{F}_{t-1}$ and must ensure that $\boldsymbol{\mu}_t \in (\mathbb{R}^+)^d \ \forall t \geq 0$. Cipollini, Engle, and Gallo (2006) proposed the so called base vMEM, in which the specification of $\boldsymbol{\mu}_t$ is the multivariate analogous of the base MEM represented in equation (2.2.3):

$$\boldsymbol{\mu}_t = \boldsymbol{\omega} + \mathbf{B}\boldsymbol{\mu}_{t-1} + \mathbf{A}\mathbf{x}_{t-1} \tag{2.3.3}$$

[3]This happens in particular when the analysis is mainly focused on the dynamics of the conditional mean.

where $\boldsymbol{\omega}$ is a $d \times 1$ vector and $\mathbf{B}$ and $\mathbf{A}$ are $d \times d$ matrices. Sufficient (but not necessary) conditions so that $\boldsymbol{\mu}_t \in \mathbb{R}^{+^d}\ \forall t \geq 0$ are again that all the parameters $\omega_i,\ \beta_{ij},\ \alpha_{ij}$ are positive for every $i, j = 1, \ldots, d$[4], while sufficient condition for the stationarity of $\boldsymbol{\mu}_t$ is that all the characteristic roots of $\mathbf{B}+\mathbf{A}$ lie inside the unit circle[5].

A direct generalization of the specification (2.3.3), used when we are modelling jointly different measures of the volatility of one asset and analogous to the univariate asymmetric specification expressed in equation (2.2.4), is the so-called asymmetric vMEM (AvMEM):

$$\boldsymbol{\mu}_t = \boldsymbol{\omega} + \mathbf{B}\boldsymbol{\mu}_{t-1} + \mathbf{A}\mathbf{x}_{t-1} + \mathbf{A}^{(-)}\mathbf{x}_{t-1}\mathbb{I}_{\{r_{t-1}\leq 0\}} \tag{2.3.4}$$

where $\mathbf{A}^{(-)}$ is another $d \times d$ matrix and r_t is the return at time t of the asset whose volatilities are measured. Also in this case for $\boldsymbol{\mu}_t$ to belong to the positive orthant it is sufficient that all the element of all the matrices are non-negative, while sufficient condition for the stationarity of $\boldsymbol{\mu}_t$ is that all the characteristic roots of $\mathbf{B}+\mathbf{A}+\frac{\mathbf{A}^{(-)}}{2}$ lie inside the unit circle. Some generalizations of the AvMEM presented in equation (2.3.4), have been used by Cipollini, Engle, and Gallo (2013) to study inter-dependencies across volatility measures.

Furthermore Engle, Gallo, and Velucchi (2009) and Engle, Gallo, and Velucchi (2012), used another augmented version of the base vMEM to describe the volatility spillover effect in the East Asian financial markets in the period before, during and after the Asian currency crisis of 1997-1998. In these papers they use this specification for the conditional mean:

$$\boldsymbol{\mu}_t = \boldsymbol{\omega} + \operatorname{diag}(\beta_1, \ldots, \beta_d)\,\boldsymbol{\mu}_{t-1} + \mathbf{A}^{\star}\mathbf{x}_{t-1} + \operatorname{diag}(\psi_1, \ldots, \psi_d)\,\mathbf{x}_{t-2} + \boldsymbol{\delta} DC_{t-1} + \boldsymbol{\lambda} PC_{t-1} + \mathbf{G}^{\star}\mathbf{x}_{t-1} DC_{t-1} \tag{2.3.5}$$

where

$$\mathbf{A}^{\star} = \begin{bmatrix} \alpha_{1,1}^{(+)}\mathbb{I}_{\left\{r_1^{(t-1)}\geq 0\right\}} + \alpha_{1,1}^{(-)}\mathbb{I}_{\left\{r_1^{(t-1)}<0\right\}} & \alpha_{1,2} & \cdots & \alpha_{1,d} \\ \alpha_{2,1} & \ddots & & \vdots \\ \vdots & & \ddots & \alpha_{d-1,d} \\ \alpha_{d,1} & \cdots & \alpha_{d,d-1} & \alpha_{d,d}^{(+)}\mathbb{I}_{\left\{r_d^{(t-1)}\geq 0\right\}} + \alpha_{d,d}^{(-)}\mathbb{I}_{\left\{r_d^{(t-1)}<0\right\}} \end{bmatrix},$$

$$\mathbf{G}^{\star} = \begin{bmatrix} \gamma_{1,1}^{(+)}\mathbb{I}_{\left\{r_1^{(t-1)}\geq 0\right\}} + \gamma_{1,1}^{(-)}\mathbb{I}_{\left\{r_1^{(t-1)}<0\right\}} & \gamma_{1,2} & \cdots & \gamma_{1,d} \\ \gamma_{2,1} & \ddots & & \vdots \\ \vdots & & \ddots & \gamma_{d-1,d} \\ \gamma_{d,1} & \cdots & \gamma_{d,d-1} & \gamma_{d,d}^{(+)}\mathbb{I}_{\left\{r_d^{(t-1)}\geq 0\right\}} + \gamma_{d,d}^{(-)}\mathbb{I}_{\left\{r_d^{(t-1)}<0\right\}} \end{bmatrix},$$

$\boldsymbol{\delta}$ and $\boldsymbol{\lambda}$ are $d \times 1$ vectors of parameters, $\mathbf{r}_t = \left(r_1^{(t)}, \ldots, r_d^{(t)}\right)$ is the vector of returns, at time t, of the d market indexes whose volatility is measures and DC_{t-1}, PC_{t-1} are dummy variables that assume value 1 if the $(t-1)$th observation belongs to the "during crisis" period or the "post crisis" period, respectively. Alternatively an easier specification for the conditional mean proposed by Gallo and

[4]Note however that this is not a necessary condition and, in multivariate context, it is also quite restrictive. It is therefore often omitted in applications in favour of a simple non-negativity check of the values of the conditional mean obtained with the estimates of its parameters.

[5]This stationarity condition follows directly from the classical theory of vector autoregressive models.

Velucchi (2009) and Giovannetti and Velucchi (2011) always to study the volatility spillover effect, is:

$$\boldsymbol{\mu}_t = \boldsymbol{\omega} + \operatorname{diag}(\beta_1, \ldots, \beta_p)\, \boldsymbol{\mu}_{t-1} + \mathbf{A}^{(+)}\mathbf{x}_{t-1} \odot \mathbb{I}_{\{\mathbf{r}_{t-1} \geq \mathbf{0}\}} + \mathbf{A}^{(-)}\mathbf{x}_{t-1} \odot \mathbb{I}_{\{\mathbf{r}_{t-1} < \mathbf{0}\}} \tag{2.3.6}$$

where

$$\mathbf{A}^{(+)} = \begin{bmatrix} \alpha_{1,1}^{(+)} & \ldots & \alpha_{1,d}^{(+)} \\ \vdots & \ddots & \vdots \\ \alpha_{d,1}^{(+)} & \ldots & \alpha_{d,d}^{(+)} \end{bmatrix}, \ \mathbf{A}^{(-)} = \begin{bmatrix} \alpha_{1,1}^{(-)} & \ldots & \alpha_{1,d}^{(-)} \\ \vdots & \ddots & \vdots \\ \alpha_{d,1}^{(-)} & \ldots & \alpha_{d,d}^{(-)} \end{bmatrix} \text{ and } \mathbb{I}_{\{\mathbf{r}_{t-1} \lesseqgtr \mathbf{0}\}} = \begin{bmatrix} \mathbb{I}_{\left\{r_1^{(t-1)} \lesseqgtr 0\right\}} \\ \vdots \\ \mathbb{I}_{\left\{r_d^{(t-1)} \lesseqgtr 0\right\}} \end{bmatrix}.$$

2.3.4 Estimation methods

To estimate the parameters of vMEMs in the literature had been proposed different methods. In their seminal paper Cipollini, Engle, and Gallo (2006), considering a Normal copula of Gamma marginals as distribution of the innovations, showed that the neat maximum likelihood approach doesn't work in this context because the likelihood equations for the correlation matrix of the Normal copula cannot be solved. So they proposed a compromise to be able to solve the likelihood equation and then obtain an estimator (that formally cannot be interpreted as a maximum likelihood estimator) for that matrix. In the same paper the authors, to avoid functional specifications for the distribution of the innovations, proposed also to use the Estimating Function approach in which the advantage of not formulating assumptions about the shape of the conditional distribution of the data is balanced by a loss of efficiency with respect to the maximum likelihood under correct specification of the complete model.

To avoid parametric assumptions on the distribution of the innovations, Cipollini, Engle, and Gallo (2013) proposed to estimate the parameters of the conditional mean using an efficient version of Generalized Method of Moments that allowed the estimator obtained to have the smallest asymptotic covariance matrix. In this context, the covariance matrix of the innovation is treated as a nuisance parameter and the choice of the instruments necessary for the estimation of the parameters of the conditional mean ensured that the resulting estimator is "nuisance parameter insensitive" (in the meaning of Jørgensen and Knudsen (2004)), so that the goodness of the estimation of $\boldsymbol{\Sigma}$ did not affect the goodness of the estimation of the parameters of the conditional mean.

Finally, to study the volatility spillover effect, Engle, Gallo, and Velucchi (2012) and Giovannetti and Velucchi (2011), used independent gamma distributions as marginals for each unidimensional time series and estimated the parameters of their model for the conditional mean using equation-by-equation.

Chapter 3

A Bayesian semiparametric vector multiplicative error model

3.1 Introduction

In this chapter we are gonna describe a new vMEM in which the specification of the conditional mean is general enough to contain all the specifications described in the previous chapter as particular cases and the distribution of the innovation is more flexible than the ones described in the literature.

3.2 Specification of the conditional mean

As described in Chapter 2, we model the d-dimensional non-negative stochastic process $\mathbf{x}_t$ as an element-by-element product of the innovation term $\boldsymbol{\varepsilon}_t$ times the conditional mean of the process $\boldsymbol{\mu}_t = \mathbb{E}\left[\mathbf{x}_t|\mathscr{F}_{t-1}\right]$, where $\mathscr{F}_{t-1}$ is the information set available at time $t-1$. In other words we will have

$$\mathbf{x}_t = \boldsymbol{\mu}_t\left(\boldsymbol{\eta}\right) \odot \boldsymbol{\varepsilon}_t$$

where, for any t, $\boldsymbol{\mu}_t\left(\boldsymbol{\eta}\right)$ is a non-negative process, measurable with respect to the σ-algebra $\mathscr{F}_{t-1}$, and $\boldsymbol{\eta}$ is a vector of parameters, to be estimated, that characterize the specification of the conditional mean. As described in the previous chapter, in the literature the innovation term is often modelled as an i.i.d. process with conditional unit-vector mean, i.e.

$$\mathbb{E}\left[\boldsymbol{\varepsilon}_t|\mathscr{F}_{t-1}\right] = \boldsymbol{\iota}_d.$$

This unit mean assumption is necessary to guarantee the identifiability of the model. The i.i.d. assumption instead is not necessary: it implies that the $\mathbf{x}_t$s, conditional on $\mathscr{F}_{t-1}$, are draws from a scale-family of distributions in which the scale parameter evolves in time according to $\boldsymbol{\mu}_t$ and the shape of the distribution remains unchanged, but, in principle, as long as the conditional unit mean constraint holds, the shape of the distribution may change through time as a function of the elements of the information set $\mathscr{F}_{t-1}$.

Assuming that we have a partition of the set of times $\{1,..,T\}$ in ℓ subsets, a specification of the conditional mean that could nest the specifications (2.3.3), (2.3.4), (2.3.5) and (2.3.6) presented above is:

$$\begin{cases} \mathbf{x}_t = \boldsymbol{\mu}_t \odot \boldsymbol{\varepsilon}_t \\ \boldsymbol{\mu}_t = \boldsymbol{\omega} + \mathbf{B}\boldsymbol{\mu}_{t-1} + \mathbf{A}\mathbf{x}_{t-1} + \mathbf{A}^{(-)}\mathbf{x}_{t-1} \odot \mathbb{I}_{\{\mathbf{r}_{t-1} < \mathbf{0}_d\}} + \mathbf{F}\mathbf{x}_{t-2} + \\ \qquad + \sum_{j=1}^{\ell-1} \left(\boldsymbol{\kappa}^{(j)} + \mathbf{G}^{(j)}\mathbf{x}_{t-1} + \mathbf{G}^{(j-)}\mathbf{x}_{t-1} \odot \mathbb{I}_{\{\mathbf{r}_{t-1} < \mathbf{0}_d\}} \right) \delta_j^{(t-1)} \end{cases} \tag{3.2.1}$$

where

$$\mathbb{I}_{\{\mathbf{r}_{t-1} < \mathbf{0}_d\}} = \begin{bmatrix} \mathbb{I}_{\left\{r_1^{(t-1)} < 0\right\}} \\ \vdots \\ \mathbb{I}_{\left\{r_d^{(t-1)} < 0\right\}} \end{bmatrix}_{d \times 1}$$

$$\boldsymbol{\omega} = \begin{bmatrix} \omega_1 \\ \vdots \\ \omega_d \end{bmatrix}_{d \times 1}$$

$$\mathbf{B} = \begin{bmatrix} \beta_{1,1} & \cdots & \beta_{1,d} \\ \vdots & \ddots & \vdots \\ \beta_{d,1} & \cdots & \beta_{d,d} \end{bmatrix}_{d \times d}$$

$$\mathbf{A} = \begin{bmatrix} \alpha_{1,1} & \cdots & \alpha_{1,d} \\ \vdots & \ddots & \vdots \\ \alpha_{d,1} & \cdots & \alpha_{d,d} \end{bmatrix}_{d \times d}$$

$$\mathbf{A}^{(-)} = \begin{bmatrix} \alpha_{1,1}^{(-)} & \cdots & \alpha_{1,d}^{(-)} \\ \vdots & \ddots & \vdots \\ \alpha_{d,1}^{(-)} & \cdots & \alpha_{d,d}^{(-)} \end{bmatrix}_{d \times d}$$

$$\mathbf{F} = \begin{bmatrix} \varphi_{1,1} & \cdots & \varphi_{1,d} \\ \vdots & \ddots & \vdots \\ \varphi_{d,1} & \cdots & \varphi_{d,d} \end{bmatrix}_{d \times d}$$

$$\boldsymbol{\kappa}^{(j)} = \begin{bmatrix} \kappa_1^{(j)} \\ \vdots \\ \kappa_d^{(j)} \end{bmatrix}_{d \times 1} \qquad \forall j = 1, \ldots, \ell - 1$$

$$\mathbf{G}^{(j)} = \begin{bmatrix} \gamma_{1,1}^{(j)} & \cdots & \gamma_{1,d}^{(j)} \\ \vdots & \ddots & \vdots \\ \gamma_{d,1}^{(j)} & \cdots & \gamma_{d,d}^{(j)} \end{bmatrix}_{d \times d} \qquad \forall j = 1, \ldots, \ell - 1$$

$$\mathbf{G}^{(j-)} = \begin{bmatrix} \gamma_{1,1}^{(j-)} & \cdots & \gamma_{1,d}^{(j-)} \\ \vdots & \ddots & \vdots \\ \gamma_{d,1}^{(j-)} & \cdots & \gamma_{d,d}^{(j-)} \end{bmatrix}_{d \times d} \qquad \forall j = 1, \ldots, \ell - 1$$

$\delta_j^{(t-1)}$ is a time-dummy that assumes the value 1 if the $(t-1)$th observation lies in the $(j+1)$th time interval and the operation $\odot\mathbb{I}_{\{\dots\}}$ is meant to be an Hadamard product if $\mathbb{I}_{\{\dots\}}$ is a vector (as in the case of "volatility spillovers analysis") or a simple product if $\mathbb{I}_{\{\dots\}}$ is just a scalar value (as in the context of volatility proxies comparison).

The parts containing the time dummies do have an empirical motivation mainly in the "volatility spillover analysis" context, but can have some sense also in other contexts if the data analysed show apparent changes of dynamics in different time periods. Anyway, that part of the specification of the conditional mean might be generally excluded or included simply looking at the volatility dynamics.

The asymmetric reaction to negative returns can be taken in account also introducing a term dependent on the lagged returns (as done in *Engle and Gallo 2006*). Since it is a known stylized fact for volatility proxies that they depend asymmetrically from their past values, this cannot be a cheap substitute to the introduction of the signed lagged term , but it can be a possible additional term.

3.3 Specification of the innovations

We model the innovation term using a mixture of simple multivariate distributions. A finite mixture with K components can be formulated as

$$\begin{aligned} \boldsymbol{\varepsilon}_t | d_t, \boldsymbol{\phi} &\sim F\left(\boldsymbol{\phi}_{d_t}\right) \\ d_t | \mathbf{p} &\sim \text{Discrete}\left(p_1, \dots, p_K\right) \end{aligned}$$

where $\boldsymbol{\phi}_{d_t} = \left(\phi_1^{(d_t)}, \dots, \phi_K^{(d_t)}\right)$, $\mathbf{p} = (p_1, \dots, p_K)$ and d_t are categorical variables (also called "latent labels") that determine to which mixture component the observation $\boldsymbol{\varepsilon}_t$ belongs. In order to fully specify this model in a Bayesian setting, we should assign priors to $\boldsymbol{\phi}_d$ and $\mathbf{p}$:

$$\begin{aligned} \boldsymbol{\phi}_d &\sim G_0 \\ \mathbf{p} &\sim \text{Dir}\left(\frac{\alpha}{K}, \dots, \frac{\alpha}{K}\right) \end{aligned}$$

where G_0 is a distribution on the parameter space of F and $\text{Dir}\left(\frac{\alpha}{K}, \dots, \frac{\alpha}{K}\right)$ is the Dirichlet distribution on the K-dimensional simplex described in Chapter 1.

There are two important problems with finite component mixtures:

- it is usually difficult to determine a priori the required number of components,
- they lack the degree of flexibility that is needed in many applications.

To solve this problems we will use the Dirichlet Process Mixture (DPM) model described in Chapter 1, that can be seen as the limit of the finite mixture model specified above for $K \longrightarrow \infty$.

As seen in subsection (1.3.6), a DPM is the result of nonparametric DP mixing of a parametric family of distribution:

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

The stick breaking construction of the DP, described in subsection (1.3.3), implies that

$$\begin{aligned} f_{\boldsymbol{\varepsilon}}(\cdot) &= \sum_{j=1}^{\infty} w_j k(\cdot|\boldsymbol{\theta}_j) \\ \mathbf{w} &\sim GEM(\alpha) \\ \theta_j &\overset{i.i.d.}{\sim} G_0 \end{aligned}$$

therefore this model bypasses the problem of choosing the correct number of components, typical of the finite mixture models.

In our vMEM framework we propose to model the innovations by a DPM. Since in our model the data vector belongs to the positive orthant, we choose to use multivariate log-normal densities as a convenient choice for the kernels of the infinite mixture. As mentioned earlier, in parametric vMEM the distribution of innovations is restricted to have unit-vector mean. Hence, at a first glance, it seems natural to use multivariate log-normal densities with unit-vector mean and a positive-definite scale matrices, $\boldsymbol{\Sigma}_j$, obtaining [1]:

$$\begin{aligned} f_{\boldsymbol{\varepsilon}}(\cdot) &= \sum_{j=1}^{\infty} w_j logN_d(\cdot|\mathbf{m}, \boldsymbol{\Sigma}_j) \\ logN_d(\boldsymbol{\varepsilon}|\boldsymbol{\Sigma}) &= \prod_{i=1}^{d} \frac{1}{\varepsilon_i} (2\pi)^{-\frac{d}{2}} |\boldsymbol{\Sigma}|^{-\frac{1}{2}} \exp\left\{ -\frac{1}{2} (\log \boldsymbol{\varepsilon} - \mathbf{m})' \boldsymbol{\Sigma}^{-1} (\log \boldsymbol{\varepsilon} - \mathbf{m}) \right\} \mathbb{I}_{(\mathbb{R}^+)^d}(\boldsymbol{\varepsilon}) \\ m_i &= -\frac{1}{2}\Sigma_{i,i} \ \forall i = 1, \ldots, d. \\ \mathbf{w} &\sim GEM(\alpha) \\ \boldsymbol{\Sigma}_j &\overset{i.i.d.}{\sim} G_0. \end{aligned} \tag{3.3.1}$$

Constricting the vector-means of all the components to depend only from the diagonal elements of the scale matrix of the same component, restricts in some way the ability of the model to cover all the possible distributions on the positive orthant. In fact, in the univariate log-normal case (just like in the univariate Gamma case with only one parameter analysed by Solgi and Mira (2013)) introducing components with thicker tails in the mixture (i.e. with bigger variance) not modifying the mean will increase at the same time the probability of the neighbourhood around zero. Hence, in presence of fat tailed innovations in the data, while this univariate DPM attempts to assign higher weights to the components with smaller precision, it will, at the same time, increase the likelihood of the innovations close to zero. In the multivariate case, this same of reasoning can be directly applied to the marginals and can also be extended, mutati mutandis, to the joint distribution. As

[1] If we have a d-dimensional multivariate log-normal random variable $\boldsymbol{\varepsilon}$ with log-scale $\boldsymbol{m}$ and shape matrix $\boldsymbol{\Sigma}$ (i.e. we have a multidimensional random variable $\boldsymbol{\varepsilon}$ such that $\mathbf{y} = \log \boldsymbol{\varepsilon} \sim N_d(\boldsymbol{m}, \boldsymbol{\Sigma})$) we have that the components of the mean vector are

$$E[\varepsilon_i] = e^{m_i + \frac{1}{2}\Sigma_{i,i}} \ \forall i = 1, \ldots, d,$$

and hence we obtain a unit mean vector if and only if

$$m_i = -\frac{1}{2}\Sigma_{i,i} \ \forall i = 1, \ldots, d.$$

a consequence, this model does not effectively range over all the possibly true distributions on the positive orthant.

Then, in a more flexible view, we can replace the previous kernels with log-normal densities with location vectors $\mathbf{m}_j$, obtaining

$$\begin{aligned} f_{\boldsymbol{\varepsilon}}(\cdot) &= \sum_{j=1}^{\infty} w_j logN_d(\cdot|\mathbf{m}_j, \boldsymbol{\Sigma}_j) \\ logN_d(\boldsymbol{\varepsilon}|\mathbf{m}, \boldsymbol{\Sigma}) &= \prod_{i=1}^{d} \frac{1}{\varepsilon_i} (2\pi)^{-\frac{d}{2}} |\boldsymbol{\Sigma}|^{-\frac{1}{2}} \exp\left\{-\frac{1}{2} (\log \boldsymbol{\varepsilon} - \mathbf{m})' \boldsymbol{\Sigma}^{-1} (\log \boldsymbol{\varepsilon} - \mathbf{m})\right\} \mathbb{I}_{(\mathbb{R}^+)^d}(\boldsymbol{\varepsilon}) \\ \mathbf{w} &\sim GEM(\alpha) \\ (\mathbf{m}, \boldsymbol{\Sigma})_j &\overset{i.i.d.}{\sim} G_0 \end{aligned} \tag{3.3.2}$$

By this definition clearly $f_{\boldsymbol{\varepsilon}}(\cdot)$ does not have unit mean. In fact, if we call $\boldsymbol{\sigma}_j = \left(\sigma_{1,1}^{(j)}, \ldots, \sigma_{d,d}^{(j)}\right)$ the vector of the diagonal elements of the matrix $\boldsymbol{\Sigma}_j$, we have

$$\bar{\mathbf{m}} = \mathbb{E}_f[\boldsymbol{\varepsilon}] = \sum_{j=1}^{\infty} w_j \exp\left\{\mathbf{m}_j + \frac{1}{2}\boldsymbol{\sigma}_j\right\} \neq \boldsymbol{\iota}_d.$$

To solve the arising identification issue we could modify the support of the random mixing distribution so that the infinite mixture has unit-vector mean. This could be done simply modifying the mixture kernels so that the density function of the innovations results specified as

$$g_{\boldsymbol{\varepsilon}}(\cdot) = \sum_{j=1}^{\infty} w_j logN_d(\cdot|\mathbf{m}_j - \log \bar{\mathbf{m}}, \boldsymbol{\Sigma}_j).$$

This specification ensures that

$$\begin{aligned} \mathbb{E}_g[\boldsymbol{\varepsilon}] &= \sum_{j=1}^{\infty} w_j \exp\left\{\mathbf{m}_j - \log \bar{\mathbf{m}} + \frac{1}{2}\boldsymbol{\sigma}_j\right\} = \\ &= \sum_{j=1}^{\infty} w_j \exp\left\{\mathbf{m}_j + \frac{1}{2}\boldsymbol{\sigma}_j\right\} \oslash \exp\{\log \bar{\mathbf{m}}\} = \\ &= \left[\sum_{j=1}^{\infty} w_j \exp\left\{\mathbf{m}_j + \frac{1}{2}\boldsymbol{\sigma}_j\right\}\right] \oslash \bar{\mathbf{m}} = \boldsymbol{\iota}_d. \end{aligned}$$

Combining this model for the distribution of innovations with (3.2.1) results in a model that we will call DPMLN2-vMEM. Unfortunately direct sampling of this model 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 the infinitely many w_js and $\mathbf{m}_j$s.

Thus, here we propose to consider the unconstrained DPM for the distribution of the innovations

obtaining:

$$\begin{aligned}
\mathbf{x}_t =& \boldsymbol{\mu}_t \odot \boldsymbol{\varepsilon}_t \\
\boldsymbol{\mu}_t =& \boldsymbol{\omega} + \mathbf{B}\boldsymbol{\mu}_{t-1} + \mathbf{A}\mathbf{x}_{t-1} + \mathbf{A}^{(-)}\mathbf{x}_{t-1} \odot \mathbb{I}_{\{\mathbf{r}_{t-1}<0\}} + \mathbf{F}\mathbf{x}_{t-2} + \\
&+ \sum_{j=1}^{\ell-1} \left(\boldsymbol{\kappa}^{(j)} + \mathbf{G}^{(j)}\mathbf{x}_{t-1} + \mathbf{G}^{(j-)}\mathbf{x}_{t-1} \odot \mathbb{I}_{\{\mathbf{r}_{t-1}<0\}} \right) \delta_j^{(t-1)} \\
f_{\boldsymbol{\varepsilon}}(\cdot) =& \sum_{j=1}^{\infty} w_j logN_d(\cdot|\mathbf{m}_j, \boldsymbol{\Sigma}_j) \qquad (3.3.3) \\
logN_d(\boldsymbol{\varepsilon}|\mathbf{m}, \boldsymbol{\Sigma}) =& \prod_{i=1}^{d} \frac{1}{\varepsilon_i} (2\pi)^{-\frac{d}{2}} |\boldsymbol{\Sigma}|^{-\frac{1}{2}} \exp\left\{ -\frac{1}{2} (\log\boldsymbol{\varepsilon} - \mathbf{m})' \boldsymbol{\Sigma}^{-1} (\log\boldsymbol{\varepsilon} - \mathbf{m}) \right\} \mathbb{I}_{(\mathbb{R}^+)^d}(\boldsymbol{\varepsilon}) \\
\mathbf{w} \sim& GEM(\alpha) \\
(\mathbf{m}, \boldsymbol{\Sigma})_j \overset{i.i.d.}{\sim}& G_0
\end{aligned}$$

which is a parameter expanded [2] (PX) version of the DPMLN2-vMEM proposed in (3.3.1) and will be called PX-DPMLN2-vMEM. It is important to enlighten that a prior on the parameters of the PX model induces a prior on the parameters of the original model and that the use of proper priors results in proper posteriors for this model (even if the likelihood is improper).

Hence we will set up a Markov Chain Monte Carlo (MCMC) simulation to target the PX-DPMLN2-vMEM and, at the end of this simulation, we will post-process the sample obtained from the that parameter-expanded model to an equivalent sample from the DPMLN2-vMEM. To map the sample from the PX-DPMLN2-VMEM to one from the DPMLN2-vMEM we will use this

[2] in the sense of Liu 1999, Van Dyk 2001 and Liu 1998

transformation:

$$\begin{pmatrix} \boldsymbol{\omega} \\ \mathbf{B} \\ \mathbf{A} \\ \mathbf{A}^{(-)} \\ \boldsymbol{\varphi} \\ \mathbf{K} \\ \mathbf{G}^{(1)} \\ \vdots \\ \mathbf{G}^{(\ell-1)} \\ \mathbf{G}^{(1-)} \\ \vdots \\ \mathbf{G}^{(\ell-1-)} \\ w_1 \\ w_2 \\ \vdots \\ \mathbf{m}_1 \\ \mathbf{m}_2 \\ \vdots \\ \boldsymbol{\Sigma}_1 \\ \boldsymbol{\Sigma}_2 \\ \vdots \end{pmatrix} \rightarrow \begin{pmatrix} \boldsymbol{\omega} \odot \bar{\mathbf{m}} \\ \mathbf{B} \\ \mathbf{A} \odot \left[\bar{\mathbf{m}} \boldsymbol{\iota}_d^{'}\right] \\ \mathbf{A}^{(-)} \odot \left[\bar{\mathbf{m}} \boldsymbol{\iota}_d^{'}\right] \\ \boldsymbol{\varphi} \odot \bar{\mathbf{m}} \\ \mathbf{K} \odot \left[\bar{\mathbf{m}} \boldsymbol{\iota}^{'}\right] \\ \mathbf{G}^{(1)} \odot \left[\bar{\mathbf{m}} \boldsymbol{\iota}_d^{'}\right] \\ \vdots \\ \mathbf{G}^{(\ell-1)} \odot \left[\bar{\mathbf{m}} \boldsymbol{\iota}_d^{'}\right] \\ \mathbf{G}^{(1-)} \odot \left[\bar{\mathbf{m}} \boldsymbol{\iota}_d^{'}\right] \\ \vdots \\ \mathbf{G}^{(\ell-1-)} \odot \left[\bar{\mathbf{m}} \boldsymbol{\iota}_d^{'}\right] \\ w_1 \\ w_2 \\ \vdots \\ \mathbf{m}_1 - \log\left(\bar{\mathbf{m}}\right) \\ \mathbf{m}_2 - \log\left(\bar{\mathbf{m}}\right) \\ \vdots \\ \boldsymbol{\Sigma}_1 \\ \boldsymbol{\Sigma}_2 \\ \vdots \end{pmatrix} \tag{3.3.4}$$

Note that, in order to use this post-processing function, we need to sample $\bar{\mathbf{m}}$, the mean of the DPM, that is an infinite sum. Although the distribution of the mean of the DP and DPMs has been the subject of several studies [3], we are not aware of a simple way to sample from these distributions since their evaluation is generally subject to computation of some numerical integrals.

To solve this problem here we propose to approximate the infinite sum constituting $\bar{\mathbf{m}}$ by a finite one so that the truncated sum of weights is close enough to 1. In practice, in order to obtain a sample from the mean of the DPM, we need to truncate the it at

$$K_{\varepsilon_{\bar{m}}} = \inf \left\{ k \in \mathbb{N} \left| 1 - \sum_{j=1}^{k} w_j < \varepsilon_{\bar{m}} \right. \right\} \tag{3.3.5}$$

where$\varepsilon_{\bar{m}}$ is a fixed tolerance level. In Muliere and Tardella (1998) it has been shown that

$$K_{\varepsilon_{\bar{m}}} - 1 \sim \text{Poisson}\left(-\alpha \log \varepsilon_{\bar{m}}\right),$$

therefore the expected value of the truncation level $K_{\varepsilon_{\bar{m}}}$ is proportional to $-\log \varepsilon_{\bar{m}}$ so that, with a small value of the concentration parameter α, extremely accurate results may be obtained in a reasonable computational time.

[3] For further insights see Lijoi 2004, Cifarelli and Regazzini 1990 and Regazzini, Lijoi, and Prünster 2003

3.4 Bayesian Inference

The Bayesian Inference on DPM models has the big issue that DPMs are infinite dimensional objects. There are substantially two main families of methods to deal with this problem: the "marginal methods", that are based on integrating out the random distribution and the "conditional methods" that explicitly instantiate the DP and rely on its stick-breaking representation. One of the most used conditional methods is the so-called "slice sampler", introduced by Walker (2007). Here we will describe, adapt to our goals and finally use its efficient version due to Kalli, Griffin, and Walker (2011).

Following Walker (2007), we augment the model with the latent variable u such that the joint density of $(\boldsymbol{\varepsilon}, u)$ is

$$f_{\boldsymbol{\varepsilon},u}(\boldsymbol{\varepsilon}, u) = \sum_{j=1}^{\infty} \mathbb{I}(w_j > u)\, logN_d(\boldsymbol{\varepsilon} \,|\mathbf{m}_j, \boldsymbol{\Sigma}_j), \tag{3.4.1}$$

Therefore, given u, the infinite mixture reduces to a finite mixture: for every fixed value of u in $[0,1]$, only a finite number of $w_j s$ can be greater than u, since $\sum_{j=1}^{\infty} w_j = 1$. Moreover, introducing the latent label l that indicates to which component of the mixture $\boldsymbol{\varepsilon}$ belongs, the joint density of $(\boldsymbol{\varepsilon}, u, l)$ is

$$f_{\boldsymbol{\varepsilon},u,l}(\boldsymbol{\varepsilon}, u, l) = \mathbb{I}(w_l > u)\, logN_d(\boldsymbol{\varepsilon} \,|\mathbf{m}_l, \boldsymbol{\Sigma}_l). \tag{3.4.2}$$

Obviously it is not possible to sample the infinite set of parameters $(\mathbf{m}_j, \boldsymbol{\Sigma}_j)_{j>1}$ but it had been shown by Walker (2007) that, by augmenting the model with the latent variable u, we only need to sample a finite set of these parameters to obtain a sample from the target "DPM distribution" (i.e. distribution that is a trajectory of a DPM).

In order to improve the efficiency of the slice sampler, Kalli, Griffin, and Walker (2011) proposed to sample in a block u and w and to rewrite the joint density (3.4.2) as

$$f_{\boldsymbol{\varepsilon},u,l}(\boldsymbol{\varepsilon}, u, l) = \mathbb{I}(\xi_l > u)\, \frac{w_l}{\xi_l} logN_d(\boldsymbol{\varepsilon} \,|\mathbf{m}_l, \boldsymbol{\Sigma}_l),$$

where $\{\xi_l\}$ is an infinite sequence decreasing in l. The block sampling increases the efficiency with respect to the original algorithm since u and w are strongly correlated, while the introduction of ξ_l reduces the sampling of useless $w_j s$ [4]. In what follows, we will use a deterministic, decreasing sequence $\{\xi_j\}_{j\in\mathbb{N}}$ but, in general, a random sequence could also be considered. Kalli, Griffin, and Walker (2011) found that the mixing of the resulting Markov chain depends on the rate of increase of $\frac{E[w_j]}{\xi_j}$: higher rates of increase are associated with better mixing but longer running times, since the average size of the sets $\{j \,|w_j > u\}$ increases. They suggest increasing the rate of increase of $\frac{E[w_j]}{\xi_j}$ until the gains in mixing are counter-balanced by the longer running time. In their examples, Kalli, Griffin, and Walker (2011) find that $\frac{E[w_j]}{\xi_j} \propto \left(\frac{3}{2}\right)^j$ strikes a good balance. Thus here we set $\xi_j \propto \frac{E[w_j]}{(1.5)^j}$.

[4] Since the sampling of u can cause changes in the set $\bigcup_{i=1}^{n} \{j \,|w_j > u_i\}$, this can lead to the simulation of superfluous $w_j s$: a scaling factor can reduce the number the average dimension of the sets $\{j \,|w_j > u_i\}$ and mitigate the problem.

At this stage we have that $\mathbf{x}_t \oslash \boldsymbol{\mu}_t = \boldsymbol{\varepsilon}_t$ and $f_{\boldsymbol{\varepsilon}_t}(\cdot) = \sum_{j=1}^{\infty} w_j logN_d(\cdot\,|\mathbf{m}_j, \boldsymbol{\Sigma}_j)$ so

$$
\begin{aligned}
f_{\mathbf{x}_t|\ldots}(\boldsymbol{x}) &= f_{\boldsymbol{\varepsilon}_t}(\boldsymbol{x} \oslash \boldsymbol{\mu}) \left|\frac{\partial \boldsymbol{\varepsilon}}{\partial \boldsymbol{x}}\right| = \\
&= \sum_{j=1}^{\infty} w_j logN_d(\boldsymbol{x} \oslash \boldsymbol{\mu}\,|\mathbf{m}_j, \boldsymbol{\Sigma}_j) \prod_{i=1}^{d} \frac{1}{\mu_i} = \\
&= \sum_{j=1}^{\infty} w_j \prod_{i=1}^{d} \frac{1}{\mu_i} \prod_{i=1}^{d} \frac{\mu_i}{x_i} (2\pi)^{-\frac{d}{2}} |\boldsymbol{\Sigma}_j|^{-\frac{1}{2}} \exp\left\{-\frac{1}{2}(\log(\boldsymbol{x} \oslash \boldsymbol{\mu}) - \mathbf{m}_j)' \boldsymbol{\Sigma}^{-1} (\log(\boldsymbol{x} \oslash \boldsymbol{\mu}) - \mathbf{m}_j)\right\} = \\
&= \sum_{j=1}^{\infty} w_j \prod_{i=1}^{d} \frac{1}{x_i} (2\pi)^{-\frac{d}{2}} |\boldsymbol{\Sigma}_j|^{-\frac{1}{2}} \exp\left\{-\frac{1}{2}(\log \boldsymbol{x} - \log \boldsymbol{\mu} - \mathbf{m}_j)' \boldsymbol{\Sigma}_j^{-1} (\log \boldsymbol{x} - \log \boldsymbol{\mu} - \mathbf{m}_j)\right\} = \\
&= \sum_{j=1}^{\infty} w_j logN_d(\boldsymbol{x}\,|\mathbf{m}_j + \log \boldsymbol{\mu}, \boldsymbol{\Sigma}_j)
\end{aligned}
$$

Hence the posterior of our PX-DPMLN2-vMEM model is:

$$
\begin{aligned}
&p(\boldsymbol{\eta}, \mathbf{m}_1, \mathbf{m}_2, \ldots, \boldsymbol{\Sigma}_1, \boldsymbol{\Sigma}_2, \ldots, \mathbf{w}, \mathbf{d}, \mathbf{u}\,|\mathbf{x}_1, \ldots, \mathbf{x}_t) = \\
&= \text{Priors} \times \prod_{t=1}^{T} \mathbb{I}_{(\xi_{d_t} > u_t)} \frac{w_{l_t}}{\xi_{l_t}} logN_d(\mathbf{x}_t \oslash \boldsymbol{\mu}_t\,|\mathbf{m}_{l_t}, \boldsymbol{\Sigma}_{l_t}) \prod_{i=1}^{d} \frac{1}{\mu_i^{(t)}} = \\
&= \text{Priors} \times \prod_{t=1}^{T} \mathbb{I}_{(\xi_{l_t} > u_t)} \frac{w_{l_t}}{\xi_{l_t}} \prod_{i=1}^{d} \frac{1}{x_i^{(t)}} (2\pi)^{-\frac{d}{2}} |\boldsymbol{\Sigma}_{l_t}|^{-\frac{1}{2}} \exp\left\{-\frac{1}{2}(\log \mathbf{x}_t - \log \boldsymbol{\mu}_t - \mathbf{m}_{l_t})' \boldsymbol{\Sigma}_{l_t}^{-1} (\log \mathbf{x}_t - \log \boldsymbol{\mu}_t - \mathbf{m}_{l_t})\right\},
\end{aligned}
$$

where$\boldsymbol{\eta}$ is vector of all the parameters from which depends the conditional mean, $\mathbf{w} = \{w_1, w_2, ...\}$ is the weight process, $\mathbf{l} = (l_1, \ldots, l_T)$ is the vector of latent labels and $\mathbf{u} = (u_1, \ldots, u_T)$ is the vector of latent variables such that (3.4.1) holds.

In our MCMC simulations we sample l_t,u_t $\forall t = 1, \ldots, T$, $v_j, \mathbf{m}_j, \boldsymbol{\Sigma}_j$ for all the required js and $\boldsymbol{\eta}$. Then we post process the sample obtained using the map (3.3.4) in order to obtain a sample from the posterior of DPMLN2-vMEM.

We will now detail the steps of the slice sampler.

3.4.1 Sampling u_t

The full conditional probability density function of u_t is

$$p(u_t\,|\ldots) \propto \mathbb{I}(\xi_{l_t} > u_t).$$

Therefore, we can simply sample u_t from the uniform distribution $U(0, \xi_{l_t})$.

3.4.2 Sampling v_j

As described in the first chapter, $v_j \overset{i.i.d.}{\sim} \text{Beta}(1, \alpha)$. Thus the full conditional probability density function of v_j is

$$\begin{aligned} p(v_j|\dots) &\propto \pi(v_j) \prod_{t:\, l_t \geq j} w_{l_t} \propto \\ &\propto v_j^0 (1-v_j)^{\alpha-1} \prod_{t:\, l_t \geq j} v_{l_t} \prod_{k=1}^{l_t-1} (1-v_k) \propto \\ &\propto (1-v_j)^{\alpha-1} \prod_{t:\, l_t = j} v_{l_t} \prod_{k=1}^{l_t-1} (1-v_k) \prod_{t:\, l_t > j} v_{l_t} \prod_{k=1}^{l_t-1} (1-v_k) \propto \\ &\propto (1-v_j)^{\alpha-1} \prod_{t:\, l_t = j} v_{l_t} \prod_{t:\, l_t > j} (1-v_j) = \\ &= v_j^{n_j} (1-v_j)^{\alpha-1+g_j}. \end{aligned}$$

Therefore, the full conditional distribution of v_j is $\text{Beta}(1+n_j, \alpha+g_j)$, where $n_j = \sum_{t=1}^{T} \mathbb{I}_{(l_t=j)}$ and $g_j = \sum_{t=1}^{T} \mathbb{I}_{(l_t>j)}$.

Note that $\sum_{t=1}^{T} \mathbb{I}_{(l_t=j)} = \sum_{t=1}^{T} \mathbb{I}_{(l_t>j)} = 0 \; \forall j \geq \bar{d} = \max\{l_1, \dots, l_T\}$: this means that the distribution of v_j will be updated if and only if there exists at least one innovation coming from a component with index greater than (or equal to) j. Otherwise the full conditional of v_j is equal to the prior distribution. Therefore at this step of the sampling we only need to update a finite number, N, of v_js: the others will not be updated and, if we will ever need them in other steps of the sampler, we will sample them from their prior.

3.4.3 Sampling $\left(\mathbf{m}_j, \Sigma_j^{-1}\right)$

In our PX-DPMLN2-vMEM model we put a d-dimensional Normal-Wishart prior on $\left(\mathbf{m}_j, \mathbf{\Sigma}_j^{-1}\right)$. This prior is the conjugate prior for a Bayesian model with normal data, so we consider a transformation of the data:

$$\begin{cases} \boldsymbol{\varepsilon}_t = \mathbf{x}_t \oslash \boldsymbol{\mu}_t \; \forall t = 1, \dots, T \\ f_{\boldsymbol{\varepsilon}}(\cdot) = \sum_{j=1}^{\infty} w_j \log N_d(\boldsymbol{\varepsilon}|\mathbf{m}_j, \mathbf{\Sigma}_j) \\ f_{\mathbf{x}}(\boldsymbol{x}) = f_{\boldsymbol{\varepsilon}}(\boldsymbol{x} \oslash \boldsymbol{\mu}) \left| \dfrac{\partial \boldsymbol{\varepsilon}}{\partial \boldsymbol{x}} \right| \end{cases} \Longrightarrow \begin{cases} \log \boldsymbol{\varepsilon}_t = \log(\mathbf{x}_t \oslash \boldsymbol{\mu}_t) = \mathbf{y}_t \; \forall t = 1, \dots, T \\ f_{\log \boldsymbol{\varepsilon}}(\cdot) = \sum_{j=1}^{\infty} w_j N_d(\log \boldsymbol{\varepsilon}|\mathbf{m}_j, \mathbf{\Sigma}_j) \\ f_{\mathbf{y}}(\boldsymbol{y}) = f_{\log \boldsymbol{\varepsilon}}(\log(\mathbf{x} \oslash \boldsymbol{\mu})) \left| \dfrac{\partial \log \boldsymbol{\varepsilon}}{\partial \log(\mathbf{x} \oslash \boldsymbol{\mu})} \right| \end{cases} \tag{3.4.3}$$

So for every $j = 1, 2, \dots$, we put

$$\begin{aligned} \mathbf{\Sigma}_j^{-1} &\sim \text{Wishart}_d(a, \mathbf{W}) \\ \mathbf{m}_j | \mathbf{\Sigma}_j^{-1} &\sim N_d\left(\boldsymbol{\nu}, n_0 \mathbf{\Sigma}_j^{-1}\right) \end{aligned} \tag{3.4.4}$$

where $a \geq d$, $n_0 > 0$, $\mathbf{W}$ is a positive definite, symmetric $d \times d$ matrix and $n_0 \boldsymbol{\Sigma}^{-1}$ is the precision matrix. Thus we obtain that

$$\boldsymbol{\Sigma}_j^{-1} | \mathbf{y}_1, \ldots, \mathbf{y}_T \sim$$
$$\sim \text{Wishart}_d \left(a + n_j, \left[\mathbf{W}^{-1} + \sum_{t:\, l_t = j} (\mathbf{y}_t - \bar{\mathbf{y}}_j)(\mathbf{y}_t - \bar{\mathbf{y}}_j)' + \frac{n_0 n_j}{n_j + n_0} (\bar{\mathbf{y}}_j - \boldsymbol{\nu})(\bar{\mathbf{y}}_j - \boldsymbol{\nu})' \right]^{-1} \right)$$
$$\mathbf{m}_j \left| \boldsymbol{\Sigma}_j^{-1}, \mathbf{y}_1, \ldots, \mathbf{y}_T \right. \sim N_d \left(\frac{n_0 \boldsymbol{\nu} + n_j \bar{\mathbf{y}}_j}{n_0 + n_j}, (n_0 + n_j) \boldsymbol{\Sigma}_j^{-1-1} \right)$$

with $\bar{\mathbf{y}}_j = \frac{1}{n_j} \sum_{t:\, l_t = j} \mathbf{y}_t$ and where $(n_0 + n_j) \boldsymbol{\Sigma}_j^{-1}$ is the precision matrix. Note that, although $j = 1, 2, \ldots$, only a finite number of $\left(\mathbf{m}_j, \boldsymbol{\Sigma}_j^{-1}\right)$ s will be updated at each step of the Gibbs sampler, since the full conditional of all the couples for which $n_j = 0$ is equal to their prior.

3.4.4 Sampling l_t

The full conditional distribution of l_t is given by the probabilities:

$$\Pr\{l_t = k \,|\ldots\} \propto$$
$$\propto \mathbb{I}_{(\xi_k > u_t)} \frac{w_k}{\xi_k} \prod_{i=1}^{d} \frac{1}{x_i^{(t)}} (2\pi)^{-\frac{d}{2}} |\boldsymbol{\Sigma}_k|^{-\frac{1}{2}} \exp\left\{ -\frac{1}{2} (\log \mathbf{x}_t - \log \boldsymbol{\mu}_t - \mathbf{m}_k)' \boldsymbol{\Sigma}_k^{-1} (\log \mathbf{x}_t - \log \boldsymbol{\mu}_t - \mathbf{m}_k) \right\}.$$

Since $\xi_k \propto \left(\frac{2}{3}\right)^k E[w_k] = \left(\frac{2}{3}\right)^k \frac{1}{1+\alpha} \left(\frac{\alpha}{1+\alpha}\right)^{k-1} = \frac{1}{\alpha} \left(\frac{2\alpha}{3+3\alpha}\right)^k$ is a decreasing function of k, for every $k \geq \log_{\frac{2\alpha}{3+3\alpha}} (\alpha u_t)$ we have that $\xi_k \leq u_t$ and hence $\Pr\{l_t = k \,|\ldots\} = 0$. Consequently, given all the other parameters, l_t takes values in the finite set $\left\{1, \ldots, \left\lfloor \log_{\frac{2\alpha}{3+3\alpha}} (\alpha u_t) \right\rfloor \right\}$,where $\lfloor a \rfloor$ stands for the integer part of the real number a.

3.4.5 Sampling η

The full conditional probability density function of the vector of parameters of the conditional mean, $\boldsymbol{\eta}_{1\times m} = \text{vec}\left(\left[\boldsymbol{\omega}, \mathbf{B}, \mathbf{A}, \mathbf{A}^{(-)}, \mathbf{F}, \mathbf{K}, \mathbf{G}^{(1)}, \ldots, \mathbf{G}^{(\ell-1)}, \mathbf{G}^{(1-)}, \ldots, \mathbf{G}^{(\ell-1-)}\right]\right)_{1\times m}$, is

$$p(\boldsymbol{\eta} \,|\ldots) \propto p(\boldsymbol{\eta}) \times \prod_{t=1}^{T} logN_d (\mathbf{x}_t \oslash \boldsymbol{\mu}_t(\boldsymbol{\eta}) \,|\mathbf{m}_{l_t}, \boldsymbol{\Sigma}_{l_t}) \prod_{i=1}^{d} \frac{1}{\mu_i^{(t)}(\boldsymbol{\eta})} \propto$$
$$\propto p(\boldsymbol{\eta}) \times \prod_{t=1}^{T} \exp\left\{ -\frac{1}{2} (\log \mathbf{x}_t - \log \boldsymbol{\mu}_t(\boldsymbol{\eta}) - \mathbf{m}_{l_t})' \boldsymbol{\Sigma}_{l_t}^{-1} (\log \mathbf{x}_t - \log \boldsymbol{\mu}_t(\boldsymbol{\eta}) - \mathbf{m}_{l_t}) \right\}, \tag{3.4.5}$$

which is not a standard distribution. For the prior of $\boldsymbol{\eta}$ we use an independent Normal distribution with large variances:

$$p(\boldsymbol{\eta}) = N_m(\boldsymbol{\eta}; \mathbf{0}_m, 20\mathbf{I}_m),$$

where $m = d + 4d^2 + d(\ell - 1) + 2(\ell - 1)d^2$ and $N_m(\cdot; \mathbf{0}_m, 20\mathbf{I}_m)$ is the density function of the m-dimensional Normal distribution with parameters $(\mathbf{0}_m, 20\mathbf{I}_m)$.

To sample $\boldsymbol{\eta}$ we will use an adaptive version of the random-walk Metropolis-Hastings algorithm with proposal density

$$q\left(\boldsymbol{\eta}_n, \boldsymbol{\eta}_{n+1}\right) = p \cdot N_m\left(\boldsymbol{\eta}_{n+1}; \boldsymbol{\eta}_n, \frac{\boldsymbol{\Lambda}_n}{m}\sigma_1^2\right) + (1-p) \cdot N_m\left(\boldsymbol{\eta}_{n+1}; \boldsymbol{\eta}_n, \frac{\boldsymbol{\Lambda}_n}{m}\sigma_2^2\right).$$

The $\boldsymbol{\Lambda}_n$ component of the proposal covariance matrix is adapted as

$$\boldsymbol{\Lambda}_n = \hat{\boldsymbol{\Sigma}}_n \oslash \mathbf{C} + 10^{-6}\mathbf{I}_m$$

where $\hat{\boldsymbol{\Sigma}}_n$ is the empirical covariance matrix of the vectors obtained from $\boldsymbol{\eta}_1, \ldots, \boldsymbol{\eta}_n$ using transformation (3.3.4), and

$$\mathbf{C} = \begin{bmatrix} \bar{\mathbf{m}} \\ \boldsymbol{\iota}_{d^2} \\ \bar{\mathbf{m}} \\ \vdots \\ \bar{\mathbf{m}} \end{bmatrix} \begin{bmatrix} \bar{\mathbf{m}}' & \boldsymbol{\iota}_{d^2}' & \bar{\mathbf{m}}' & \ldots & \bar{\mathbf{m}}' \end{bmatrix}$$

is the transformation matrix to be applied to $\hat{\boldsymbol{\Sigma}}_n$ to recover from it the empirical covariance matrix of $[\boldsymbol{\eta}_1, \ldots, \boldsymbol{\eta}_k]$.

For what it takes the scale parameters, σ_1 and σ_2, and the mixture weight, p, we consider them as constants that should be tuned.

Finally it is important to enlighten that at the k-th iteration $\hat{\boldsymbol{\Sigma}}_n$ changes only by $O\left(\frac{1}{n}\right)$ by definition. Therefore this adaptation mechanism satisfies the "diminishing adaptation" condition of Roberts and Rosenthal (2007) and thus the correct target distribution is preserved.

Chapter 4

Simulations

To study the performance of the different specifications of our model and of the proposed sampling scheme we run several simulation studies. We will now present some of most significant ones using different specifications of the conditional mean.

4.1 Bivariate simulations

For all the bivariate simulations we will use this common hyperparameters:

- the concentration parameter of the DP is $\alpha = 1$.
- the truncation level defined in equation (3.3.5) is $\varepsilon_{\tilde{\mathbf{m}}} = 10^{-6}$
- the parameters of the base Normal-Wishart measure defined in equation (3.4.4) are

$$a = 10 + d,\ \mathbf{W}_{ij} = \begin{cases} \frac{1}{T-1} \sum_{t=1}^{T} \left(\log \boldsymbol{\epsilon}_{t,i} - \overline{\log \boldsymbol{\epsilon}}_i \right)^2 & if\ i = j \\ 0 & else \end{cases},\ \boldsymbol{\nu} = \overline{\log \boldsymbol{\epsilon}},\ n_0 = 1$$

  where $\boldsymbol{\epsilon}_t = \mathbf{x}_t \oslash \boldsymbol{\mu}_t(\boldsymbol{\eta}_0)$, $d = 2$ is the length of the vector $\mathbf{x}_t$ and the over-line indicates sample mean over t.

- the initial value of the vector of parameters of the conditional mean, $\boldsymbol{\eta}_0$, is the maximum likelihood estimate found assuming a parametric model with log-Normal distributed innovations.

The samples from the posterior of the DPMLN2-vMEM are derived by post-processing the samples obtained from the posterior of the PX-DPMLN2-vMEM using the direct adaptation of transformation (3.3.4) to the case under analysis.

For every simulation we sample 3000 bivariate observations from the specified model. The location vectors of both the components of the mixture distribution of the true innovations are always chosen such that both of them (and thus also the whole mixture) have unit-vector mean. To compute the effective sample sizes will use the "coda" R package. Furthermore, we will compare the estimates of the density of the innovations and of the predictive density obtained with our DPMLN2-vMEM with the ones obtained using Maximum Likelihood from a vMEM with parametric

Log-Normal innovations with vector mean obtained as a transformation of the diagonal elements of the scale matrix (so that the unit-vector-mean constraint is satisfied). From now on we will call this last parametric model LN1-vMEM.

4.1.1 Base diagonal specification

We start our simulations with the simplest specification for the conditional mean, that we will call the "base diagonal specification":

$$\begin{cases} \mathbf{x}_t = & \boldsymbol{\mu}_t \odot \boldsymbol{\varepsilon}_t \\ \boldsymbol{\mu}_t = & \boldsymbol{\omega} + \boldsymbol{\beta} \odot \boldsymbol{\mu}_{t-1} + \mathbf{A}\mathbf{x}_{t-1} \end{cases}$$

with

$$\boldsymbol{\omega} = \begin{bmatrix} 0.40 \\ 0.25 \end{bmatrix}, \ \boldsymbol{\beta} = \begin{bmatrix} 0.35 \\ 0.30 \end{bmatrix}, \ \mathbf{A} = \begin{bmatrix} 0.27 & 0.19 \\ 0.18 & 0.31 \end{bmatrix}$$

and

$$\boldsymbol{\varepsilon}_t \overset{i.i.d.}{\sim} 0.7 * logN_2\left(\begin{bmatrix} -0.10 \\ -0.05 \end{bmatrix}, \begin{bmatrix} 0.20 & 0.09 \\ 0.09 & 0.10 \end{bmatrix}\right) + 0.3 * logN_2\left(\begin{bmatrix} -0.15 \\ -0.10 \end{bmatrix}, \begin{bmatrix} 0.30 & 0.10 \\ 0.10 & 0.20 \end{bmatrix}\right).$$

We run the algorithm for $N = 51,000$ iterations and then we discard the first 6,000 as burn in. To sample the parameters of the conditional mean as described in subsection 3.4.5 we set $p = 1$ as the weight for the proposal density and $\sigma_1 = 1.5$ as its scale factor. The simulation time on a server running at 2.60GHz and with 128GB RAM is about 2 hours. In Table 4.1.1 we report the posterior means and 95% credible intervals for the parameters of the conditional mean, along with their true values. As it can be seen all the true values lie inside the 95% credible intervals ensuring thus the goodness of the estimates. All the estimates are based on effective sample size greater than 1100. In Figures 4.1.1, 4.1.2, 4.1.3 we reported the traces, the posterior histograms and the autocorrelation functions parameters of the conditional mean of the DPMLN2-vMEM model. As it can be seen from the traces and is also confirmed by the Geweke and Heidelberg-Welch tests of convergence, all the Markov chains of the parameters of the conditional mean seem to have reached convergence and the autocorrelation functions remain significant for less than 80 lags. In Figure 4.1.4 are reported the traces and the running averages of the maximum number of components and of the number of "active components[1]" at each step along with the traces of the mixture weights. As it can be seen both the average number of components and the average number of active ones converge quite fast. Since we choose a fixed value for the concentration parameter for computational efficiency reasons, the average number of active components converge to 7, which is quite bigger than the true number of components. But, from the rightmost graph of that same figure we can notice that, at each step, only two of the active components have significant weight. So the choice of a fixed concentration parameter in our simulations do not really affect that much the ability of the model to recover the true number of components of the distribution of the disturbances. The contour plot on the left of Figure 4.1.5 reports the density of the innovations estimated with DPMLN2-vMEM (black line), the density of the innovations obtained with LN1-vMEM (green lines) and the true density of the simulated innovations. As we can see the approximation of the density of the innovations obtained

[1] With the expression "active components" we will refer to those components to which, at a given iteration of the MCMC simulation, is assigned at least one observation.

with our specification is very good and slightly better than the one obtained with the parametric model (which, with this simple model, is the same quite good). Finally the contour plot on the right of Figure 4.1.5 shows the one-step-ahead predictive density obtained with DPMLN2-vMEM (black line), the one-step-ahead predictive density obtained with LN1-vMEM (green line) and the true density of x_{T+1} given $x_1, \ldots, x_T$. The approximation obtained with both models is quite good although the one obtained with our model is slightly better than the one obtained with the parametric model.

Table 4.1.1: Posterior mean and 95% credible intervals for the parameters of the conditional mean.

	ω_1	ω_2	β_1	β_2
True	0.40	0.25	0.35	0.30
Est.	0.4764	0.2419	0.3295	0.3230
(95% C.I.)	(0.3880, 0.5740)	(0.2031, 0.2823)	(0.2494, 0.4069)	(0.2782, 0.3673)
	α_{11}	α_{21}	α_{12}	α_{22}
True	0.27	0.18	0.19	0.31
Est.	0.2768	0.1741	0.1600	0.2970
(95% C.I.)	(0.2363, 0.3191)	(0.1498, 0.1973)	(0.1005, 0.2265)	(0.2581, 0.3376)

Figure 4.1.1: MCMC traces, posterior densities and ACF of the components of the post-processed vector $\boldsymbol{\omega}$. The green lines in the histogram represent the 95% C.I. while the red one is the true value.

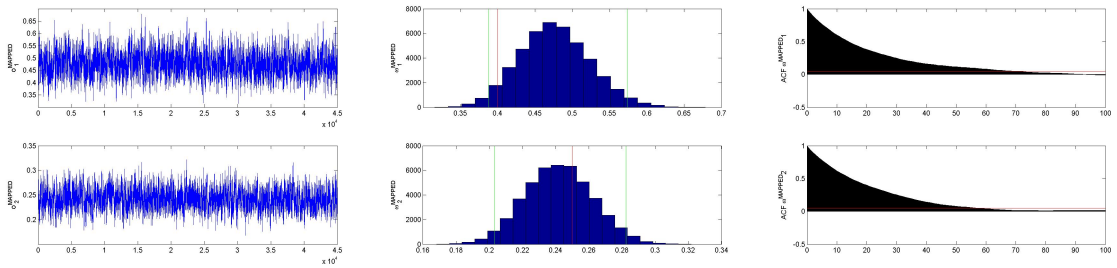


Figure 4.1.2: MCMC traces, posterior densities and ACF of the components of the post-processed vector $\boldsymbol{\beta}$. The green lines in the histogram represent the 95% C.I. while the red one is the true value.

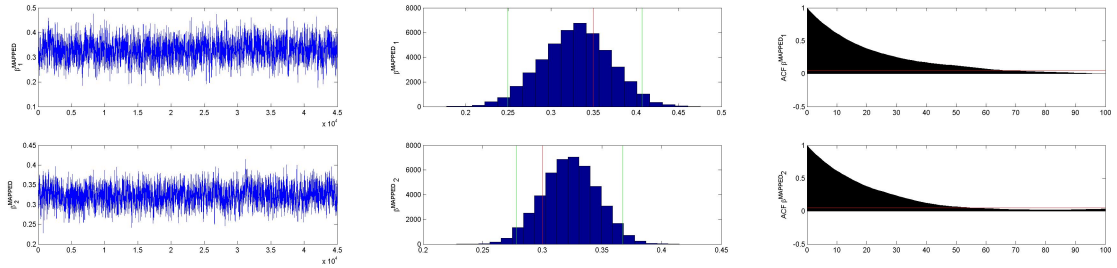


Figure 4.1.3: MCMC traces, posterior densities and ACF of the components of the post-processed matrix $\mathbf{A}$. The green lines in the histogram represent the 95% C.I. while the red one is the true value.

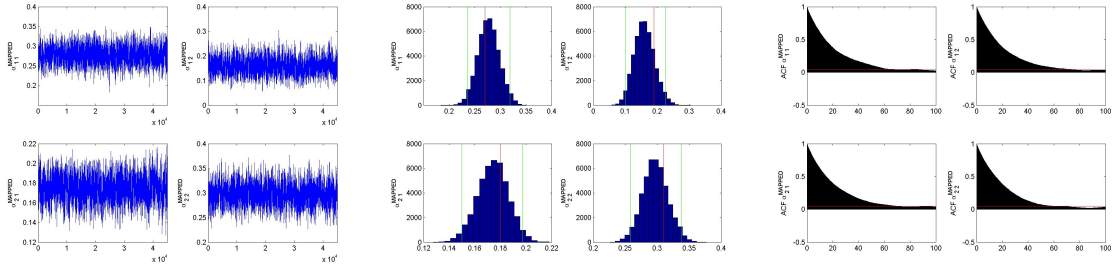


Figure 4.1.4: The upper left plot shows the traces of total number of components and of the number of active components at each step. The lower left plot shows the corresponding running averages. The plot on the right shows the traces of the mixture weights.

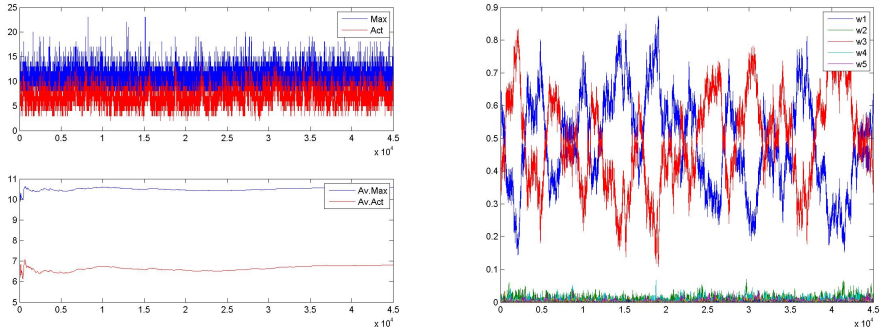
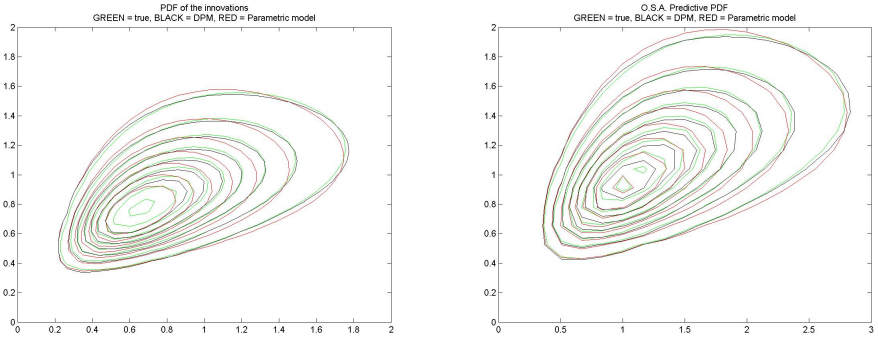


Figure 4.1.5: On the left, the true and estimated contour plots of the densities of the innovations. On the right, the contour plots of the true and estimated one step ahead predictive densities.



4.1.2 Base full specification

We now extend the previous base diagonal specification of the conditional mean to include also off-diagonal elements of the $\mathbf{B}$ matrix:

$$\begin{cases} \mathbf{x}_t = & \boldsymbol{\mu}_t \odot \boldsymbol{\varepsilon}_t \\ \boldsymbol{\mu}_t = & \boldsymbol{\omega} + \mathbf{B}\boldsymbol{\mu}_{t-1} + \mathbf{A}\mathbf{x}_{t-1} \end{cases}$$

with

$$\boldsymbol{\omega} = \begin{bmatrix} 0.40 \\ 0.25 \end{bmatrix}, \ \mathbf{B} = \begin{bmatrix} 0.35 & 0.10 \\ 0.07 & 0.30 \end{bmatrix}, \ \mathbf{A} = \begin{bmatrix} 0.27 & 0.19 \\ 0.18 & 0.31 \end{bmatrix}$$

and

$$\boldsymbol{\varepsilon}_t \overset{i.i.d.}{\sim} 0.7 * logN_2\left(\begin{bmatrix} -0.10 \\ -0.05 \end{bmatrix}, \begin{bmatrix} 0.20 & 0.09 \\ 0.09 & 0.10 \end{bmatrix}\right) + 0.3 * logN_2\left(\begin{bmatrix} -0.15 \\ -0.10 \end{bmatrix}, \begin{bmatrix} 0.30 & 0.10 \\ 0.10 & 0.20 \end{bmatrix}\right).$$

We run the algorithm for $N = 200,000$ iterations and then we discard the first 5,000 as burn-in. To sample the parameters of the conditional mean we set $p = 0.9$ as the mixture weight for the proposal density and $\sigma_1 = 0.5$, $\sigma_2 = \sqrt{10}$ as scale factors. The simulation time on server running at 3.5GHz and with 128GB RAM is about 7 and a half hours. In Table 4.1.2 we report the posterior means and 95% credible intervals for the parameters of the conditional mean, along with their true values. As it can be seen all the true values lie inside the 95% credible intervals. All the estimates are based on effective sample sizes greater than 328. In Figures 4.1.6, 4.1.7, 4.1.8 we reported the traces, the posterior densities and the autocorrelation functions of the post-processed parameters of the conditional mean. The traces seem to mix a little worse with respect to the corresponding ones presented in Figures 4.1.1, 4.1.2, 4.1.3 but still seem to have reached convergence and the autocorrelation functions remain significant for at most 2500 lags. These are clues of the fact that the Markov chain struggles to move with this target density more than with the one of the base diagonal specification. This is likely due to the newly introduced off-diagonal elements of the $\mathbf{B}$ matrix, that are highly correlated with other parameters. Figure 4.1.9 shows the traces and the running averages of the maximum number of components and of the number of active

components at each step along with the traces of the mixture weights. As in the case of the model considered in subsection 4.1.1, there are on average 7 active components but only two of them has really significant weights. The contour plot on the left of Figure 4.1.10 reports the density of the innovations estimated with DPMLN2-vMEM (black line), the density of the innovations obtained with LN1-vMEM (green lines) and the true density of the simulated innovations. As we can see the approximation of the density of the innovations obtained with our specification is very good and slightly better than the one obtained with the parametric model. Finally the contour plot on the right of Figure 4.1.10 shows the one-step-ahead predictive density obtained with DPMLN2-vMEM (black line), the one-step-ahead predictive density obtained with LN1-vMEM (green line) and the true density of x_{T+1} given $x_1, \ldots, x_T$. As we can see, the approximation obtained with our semiparametric model is quite good and much better than the one obtained with its parametric counterpart.

Table 4.1.2: Posterior mean and 95% credible intervals for the parameters of the conditional mean.

	ω_1	ω_2	β_{11}	β_{21}	β_{12}	β_{22}
True	0.40	0.25	0.35	0.07	0.10	0.30
Est.	0.499751	0.247714	0.137292	0.103098	0.333920	0.252589
(95% C.I.)	(0.3703, 0.6725)	(0.1378, 0.3405)	(−0.2711, 0.4915)	(−0.1511, 0.4219)	(−0.0555, 0.7963)	(−0.0678, 0.5170)
	α_{11}	α_{21}	α_{12}	α_{22}		
True	0.27	0.18	0.19	0.31		
Est.	0.253322	0.200371	0.167585	0.292701		
(95% C.I.)	(0.2132, 0.2962)	(0.1758, 0.2262)	(0.1005, 0.2336)	(0.2526, 0.3317)		

Figure 4.1.6: MCMC traces, posterior densities and ACF of the components of the post-processed vector $\boldsymbol{\omega}$. The green lines in the histogram represent the 95% C.I. while the red one is the true value.

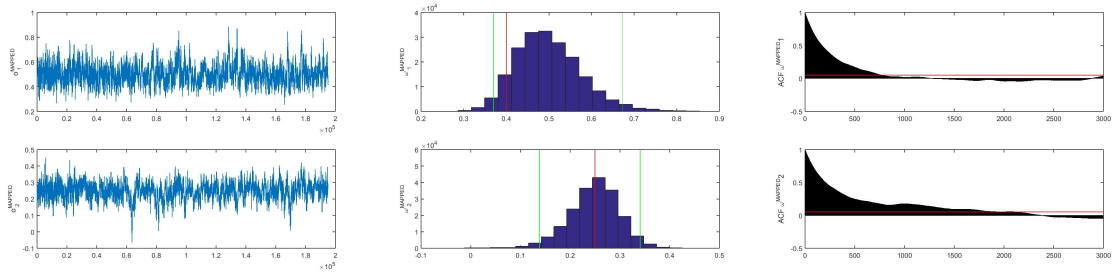


Figure 4.1.7: MCMC traces, posterior densities and ACF of the components of the post-processed matrix **B**. The green lines in the histogram represent the 95% C.I. while the red one is the true value.

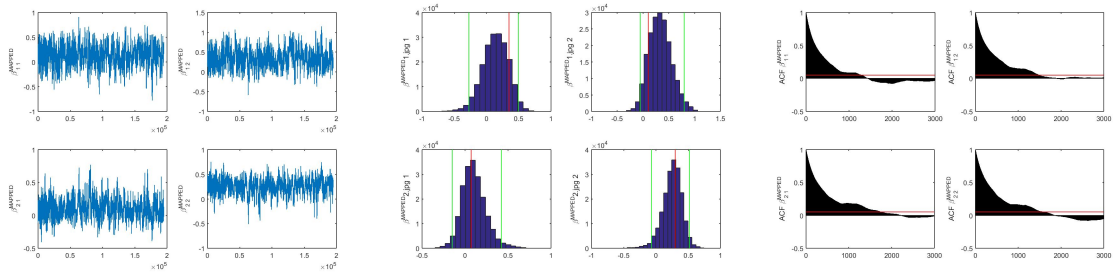


Figure 4.1.8: MCMC traces, posterior densities and ACF of the components of the post-processed matrix **A**. The green lines in the histogram represent the 95% C.I. while the red one is the true value.

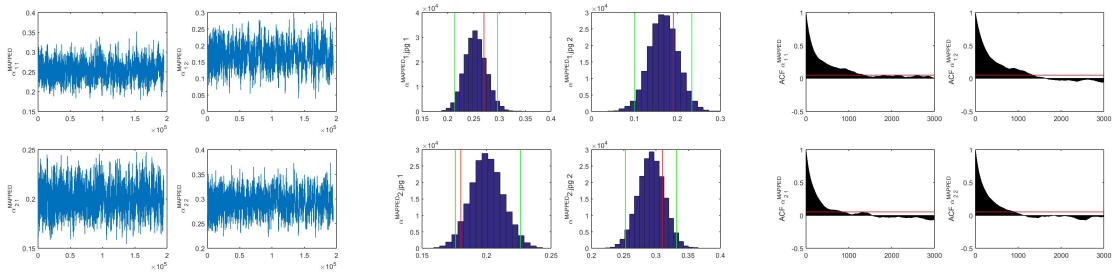


Figure 4.1.9: The upper left plot shows the traces of total number of components and of the number of active components at each step. The lower left plot shows the corresponding running averages. The plot on the right shows the traces of the mixture weights.

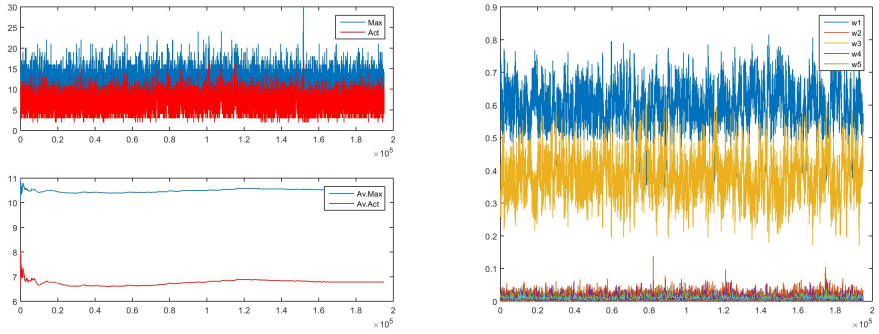
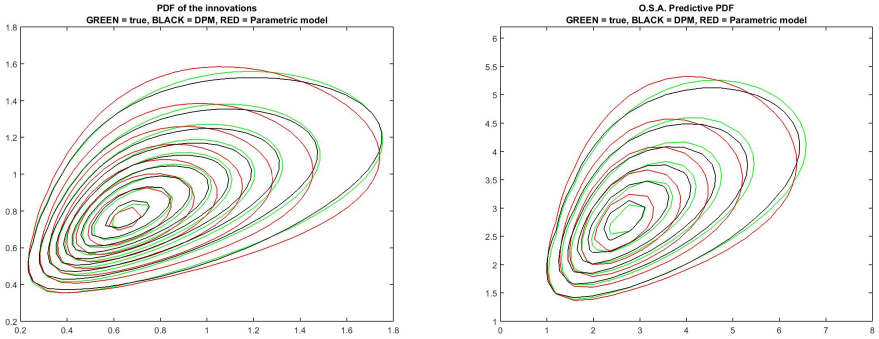


Figure 4.1.10: On the left, the true and estimated contour plots of the densities of the innovations. On the right, the contour plots of the true and estimated one step ahead predictive densities.



4.1.3 Complete full specification

Finally we simulate our sample from the full model described in equation (3.2.1):

$$\begin{cases} \mathbf{x}_t = & \boldsymbol{\mu}_t \odot \boldsymbol{\varepsilon}_t \\ \boldsymbol{\mu}_t = & \boldsymbol{\omega} + \mathbf{B}\boldsymbol{\mu}_{t-1} + \mathbf{A}\mathbf{x}_{t-1} + \mathbf{A}^{(-)}\mathbf{x}_{t-1}\mathbb{I}_{\{r_{t-1}<0\}} + \mathbf{F}\mathbf{x}_{t-2} + \\ & + \sum_{j=1}^{\ell-1} \left(\boldsymbol{\kappa}^{(j)} + \mathbf{G}^{(j)}\mathbf{x}_{t-1} + \mathbf{G}^{(j-)}\mathbf{x}_{t-1}\mathbb{I}_{\{r_{t-1}<0\}}\right) \delta_j^{(t-1)} \end{cases}$$

with

$$\boldsymbol{\varepsilon}_t \overset{i.i.d.}{\sim} 0.7 * logN_2\left(\begin{bmatrix} -0.200 \\ -0.175 \end{bmatrix}, \begin{bmatrix} 0.40 & 0.30 \\ 0.30 & 0.35 \end{bmatrix}\right) + 0.3 * logN_2\left(\begin{bmatrix} -0.185 \\ -0.195 \end{bmatrix}, \begin{bmatrix} 0.37 & 0.15 \\ 0.15 & 0.39 \end{bmatrix}\right).$$

We set $r_t \sim N_2(0, x_{t,1})$ and divide the time period into three intervals: $(0, 1000]$,$(1000, 2100]$, $(2100, 3000]$.

We run the algorithm for $N = 200,000$ iterations and we discard the first 10,000 as burn-in. To sample the parameters of the conditional mean we set $p = 0.9$ as the weight for the proposal distribution and $\sigma_1 = 1$, $\sigma_2 = \sqrt{38}$. The simulation time on a server running at 3.5GHz and with 128GB RAM is about 18 hours. In Table 4.1.3 we report the posterior means and 95% credible intervals for the parameters of the conditional mean, along with their true values. As we can see from the table, all the true values of the parameters lie inside the 95% posterior credible intervals. All the estimates are based on effective sample sizes greater than 547. Since also in this case the number of parameters is quite high and the graphs for all the variables are similar, we will report in Figure 4.1.11 only a representative example of the traces, the histograms and the autocorrelation functions of the sampled values[2]. As we can see the traces seem to have reached convergence and the autocorrelation functions of most of the variables becomes non-significant after about 3000 lags. Figure 4.1.12 shows the traces and the running averages of the maximum number of components and the active components on the left and the traces of the mixture weights on the rightmost figure: as we can see, there are on average 5 active components every step but, correctly, only two of them

[2]The remaining traces, histograms and autocorrelation functions are reported in Appendix A

have significant weights. The contour plot on the left of Figure 4.1.13 reports the density of the innovations estimated with DPMLN2-vMEM (black line), the density of the innovations obtained with LN1-vMEM (green lines) and the true density of the simulated innovations. Also in this case the approximation of the density of the innovations obtained with our semiparametric model is good in general and much better than the one obtained with the parametric model. Finally the contour plot on the right of Figure 4.1.13 shows the one-step-ahead predictive density obtained with DPMLN2-vMEM (black line), the one-step-ahead predictive density obtained with LN1-vMEM (green line) and the true density of x_{T+1} given $x_1, \ldots, x_T$. With this complex specification of the conditional mean the approximation obtained with our semiparametric model is not as good as the one obtained with simpler specifications but is still significantly better than the one obtained with the parametric model.

Table 4.1.3: Posterior mean and 95% credible intervals for the parameters of the conditional mean.

	True	Est.	(95% C.I.)		True	Est.	(95% C.I.)
ω_1	0.27	0.2426	$(0.1360, 0.3589)$	κ_2^1	0.11	0.1305	$(0.0003, 0.2646)$
ω_2	0.57	0.5040	$(0.3984, 0.6263)$	κ_1^2	0.12	0.1633	$(0.0232, 0.3175)$
β_{11}	0.25	0.0844	$(-0.1123, 0.2831)$	κ_2^2	0.05	0.1065	$(-0.0454, 0.2655)$
β_{21}	0.10	0.0601	$(-0.1448, 0.2745)$	γ_{11}^1	0.08	0.0984	$(-0.0013, 0.2078)$
β_{12}	0.08	0.2174	$(-0.0251, 0.4420)$	γ_{21}^1	0.03	0.0077	$(-0.0946, 0.1171)$
β_{22}	0.23	0.2290	$(-0.0204, 0.4613)$	γ_{12}^1	0.05	0.0568	$(-0.0451, 0.1527)$
α_{11}	0.15	0.1392	$(0.0558, 0.2158)$	γ_{22}^1	0.06	0.0105	$(-0.0949, 0.1085)$
α_{21}	0.10	0.1086	$(0.0115, 0.1952)$	γ_{11}^2	0.07	0.0790	$(-0.0362, 0.2051)$
α_{12}	0.11	0.1294	$(0.0635, 0.2062)$	γ_{21}^2	0.04	0.0638	$(-0.0511, 0.1905)$
α_{22}	0.17	0.1988	$(0.1220, 0.2899)$	γ_{12}^2	0.05	0.0268	$(-0.0906, 0.1322)$
α_{11}^-	0.13	0.2126	$(0.0854, 0.3517)$	γ_{22}^2	0.06	-0.0116	$(-0.1267, 0.0955)$
α_{21}^-	0.04	0.099413	$(-0.0333, 0.2456)$	γ_{11}^{1-}	0.08	0.0527	$(-0.1377, 0.2288)$
α_{12}^-	0.05	-0.0133	$(-0.1232, 0.0851)$	γ_{21}^{1-}	0.03	-0.0504	$(-0.2268, 0.1228)$
α_{22}^-	0.20	0.1806	$(0.0591, 0.3019)$	γ_{12}^{1-}	0.04	0.0038	$(-0.1484, 0.1656)$
φ_{11}^-	0.10	0.1183	$(0.0599, 0.1793)$	γ_{22}^{1-}	0.12	0.0724	$(-0.0960, 0.2335)$
φ_{21}^-	0.05	0.0469	$(-0.0132, 0.1099)$	γ_{11}^{2-}	0.10	-0.0436	$(-0.2437, 0.1327)$
φ_{12}^-	0.07	0.0400	$(-0.0151, 0.0951)$	γ_{21}^{2-}	0.06	-0.0901	$(-0.2815, 0.0939)$
φ_{22}^-	0.09	0.0542	$(0.0005, 0.1126)$	γ_{12}^{2-}	0.02	0.0530	$(-0.1059, 0.2309)$
κ_1^1	0.10	0.0680	$(-0.0627, 0.2034)$	γ_{22}^{2-}	0.09	0.1342	$(-0.0328, 0.3125)$

Figure 4.1.11: MCMC traces, posterior densities and ACF of the components of the post-processed matrix **A**. The green lines in the histogram represent the 95% C.I. while the red one is the true value.

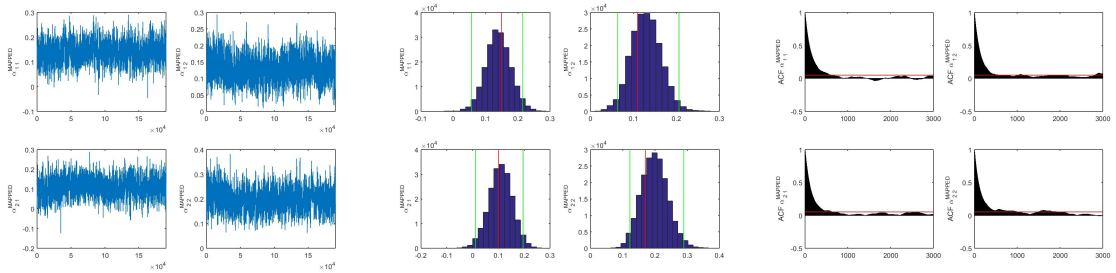


Figure 4.1.12: The upper left plot shows the traces of total number of components and of the number of active components at each step. The lower left plot shows the corresponding running averages. The plot on the right shows the traces of the mixture weights.

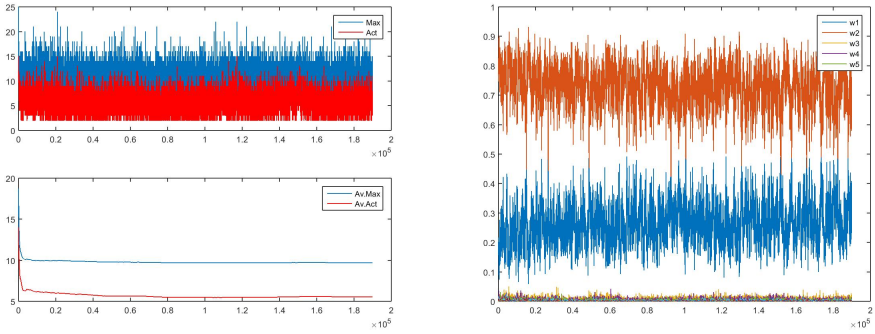
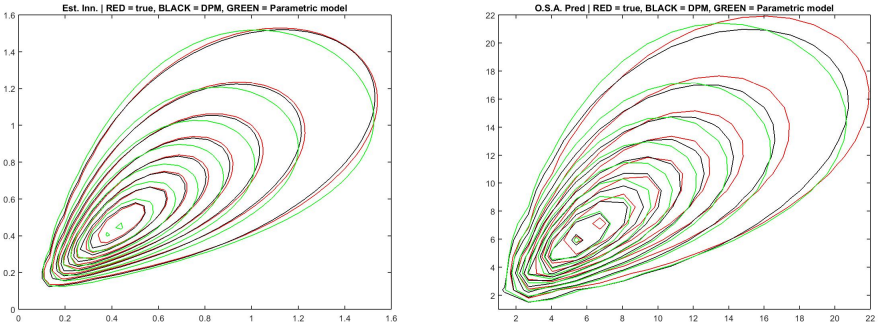


Figure 4.1.13: On the left, the true and estimated contour plots of the densities of the innovations. On the right, the contour plots of the true and estimated one step ahead predictive densities.



4.2 Trivariate simulations

We will report now some trivariate simulations. Since the number of parameter rapidly increase with the complexity of the specification of the conditional mean, we will report only simulations with two simple specifications of the conditional mean.

For these trivariate simulations we will use the same hyperparameters used for the bivariate simulations, described at the beginning of the previous section. Also in this case the samples from the posterior of the DPMLN2-vMEM are derived by post-processing the samples obtained from the posterior of the PX-DPMLN2-vMEM using the direct adaptation of transformation (3.3.4) to the case under analysis.

For every simulation we sample 3000 trivariate observations by the specified model. The location vectors of both the components of the mixture distribution of the innovations are again chosen such that both of them (and thus also the whole mixture) have unit-vector mean.

4.2.1 Base diagonal specification

We start again simulating the base diagonal specification

$$\begin{cases} \mathbf{x}_t = & \boldsymbol{\mu}_t \odot \boldsymbol{\varepsilon}_t \\ \boldsymbol{\mu}_t = & \boldsymbol{\omega} + \boldsymbol{\beta} \odot \boldsymbol{\mu}_{t-1} + \mathbf{A}\mathbf{x}_{t-1} \end{cases}$$

with

$$\boldsymbol{\omega} = \begin{bmatrix} 0.27 \\ 0.57 \\ 0.37 \end{bmatrix}, \ \boldsymbol{\beta} = \begin{bmatrix} 0.35 \\ 0.23 \\ 0.31 \end{bmatrix}, \ \mathbf{A} = \begin{bmatrix} 0.21 & 0.12 & 0.06 \\ 0.11 & 0.27 & 0.10 \\ 0.08 & 0.09 & 0.28 \end{bmatrix}$$

and

$$\boldsymbol{\varepsilon}_t \overset{i.i.d.}{\sim} 0.7 * logN_2\left(\begin{bmatrix} -0.20 \\ -0.175 \\ -0.15 \end{bmatrix}, \begin{bmatrix} 0.40 & 0.30 & 0.20 \\ 0.30 & 0.35 & 0.25 \\ 0.20 & 0.25 & 0.30 \end{bmatrix}\right) + 0.3 * logN_2\left(\begin{bmatrix} -0.185 \\ -0.195 \\ -0.125 \end{bmatrix}, \begin{bmatrix} 0.37 & 0.15 & 0.24 \\ 0.15 & 0.39 & 0.18 \\ 0.24 & 0.18 & 0.25 \end{bmatrix}\right).$$

We run the algorithm for $N = 50,000$ iterations and we discard the first 5000 as burn in. To sample the parameters of the conditional mean we set $p = 1$ as the weight for the proposal mixture density and $\sigma_1 = 1.5$ as its scale factor. The simulation time on a server running at 2.60GHz and with 128GB RAM is about 2 and a half hours. In Table 4.2.1 we report the posterior means and 95% credible intervals for the parameters of the conditional mean, along with their true values. As it can be seen, all the true values lie inside the 95% credible intervals. All the estimates are based on effective sample sizes greater than 582. In Figures 4.2.1, 4.2.2, 4.2.3 we reported the traces, the posterior densities and the autocorrelation functions of the post-processed parameters of the conditional mean. As it can be seen, all the traces have reached convergence and all the autocorrelation functions become non-significant in at most 500 lags. In Figure 4.2.4 are reported the traces and the running averages of the maximum number of components and of the number of active components at each step along with the traces of the mixture weights. Clearly both the average maximum number of components and the average number of active components converge quite fast, resulting in an average number of 6 active components. But, from the rightmost graph of that same figure we can also notice that also in this case, at each step, only two of the active components have

significant weights (which are, on average, remarkably close to the true ones). Finally in Figure 4.2.5 we report the true marginal densities of the innovations together with the estimated marginal densities obtained with DMPLN2-vMEM and the LN1-vMEM. The approximation obtained with the DPMLN2-vMEM is quite good and slightly better than the one obtained with the LN1-vMEM.

Table 4.2.1: Posterior mean and 95% credible intervals for the parameters of the conditional mean.

	True	Est.	(95% C.I.)		True	Est.	(95% C.I.)
ω_1	0.27	0.2738	$(0.2104, 0.3363)$	α_{31}	0.08	0.0889	$(0.0323, 0.1464)$
ω_2	0.57	0.5463	$(0.4411, 0.6525)$	α_{12}	0.12	0.1282	$(0.0766, 0.1786)$
ω_3	0.37	0.3485	$(0.2779, 0.4228)$	α_{22}	0.27	0.2592	$(0.1891, 0.3273)$
β_1	0.35	0.3143	$(0.2437, 0.3838)$	α_{32}	0.09	0.1002	$(0.0513, 0.1472)$
β_2	0.23	0.2311	$(0.1558, 0.3072)$	α_{13}	0.06	0.0830	$(0.0294, 0.1417)$
β_3	0.31	0.2980	$(0.2283, 0.3668)$	α_{23}	0.10	0.1397	$(0.0663, 0.2158)$
α_{11}	0.21	0.2080	$(0.1436, 0.2725)$	α_{33}	0.28	0.2877	$(0.2296, 0.3507)$
α_{21}	0.11	0.1029	$(0.0257, 0.1845)$				

Figure 4.2.1: MCMC traces, posterior densities and ACF of the components of the post-processed vector $\boldsymbol{\omega}$. The green lines in the histogram represent the 95% C.I. while the red one is the true value.

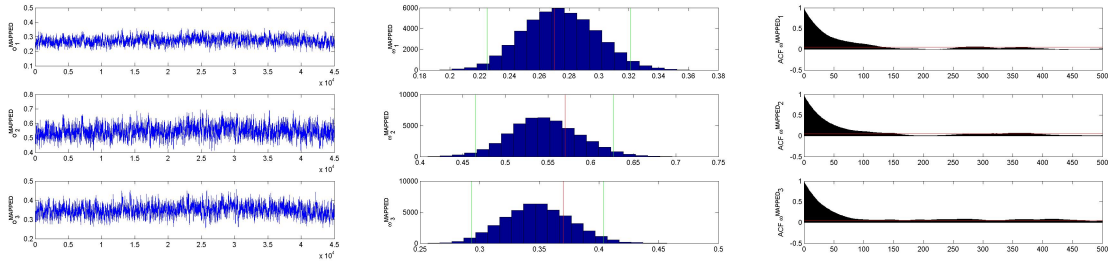


Figure 4.2.2: MCMC traces, posterior densities and ACF of the components of the post-processed vector $\boldsymbol{\beta}$. The green lines in the histogram represent the 95% C.I. while the red one is the true value.

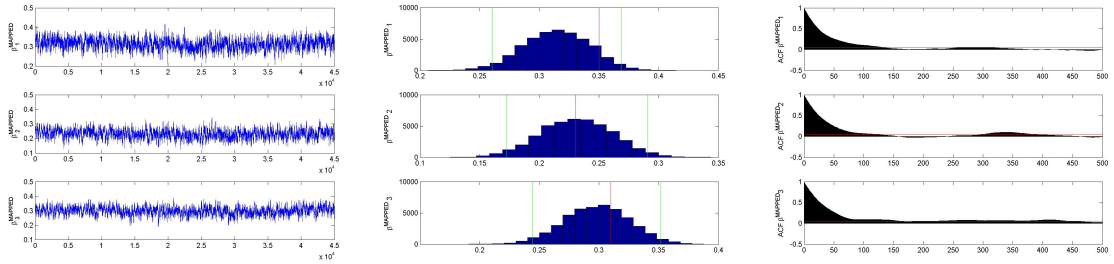


Figure 4.2.3: MCMC traces, posterior densities and ACF of the components of the post-processed matrix **A**. The green lines in the histogram represent the 95% C.I. while the red one is the true value.

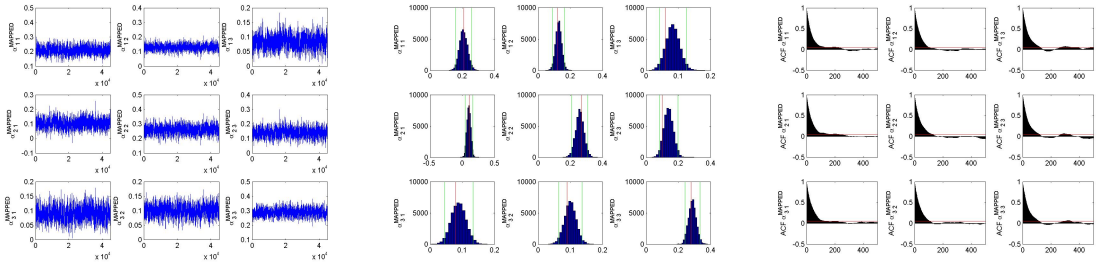


Figure 4.2.4: The upper left plot shows the traces of total number of components and of the number of active components at each step. The lower left plot shows the corresponding running averages. The plot on the right shows the traces of the mixture weights.

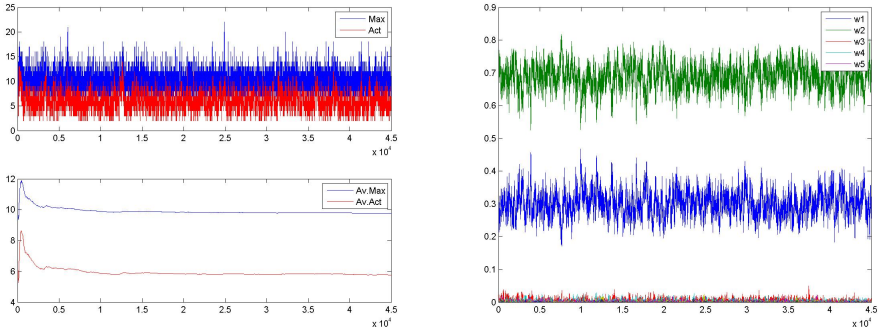
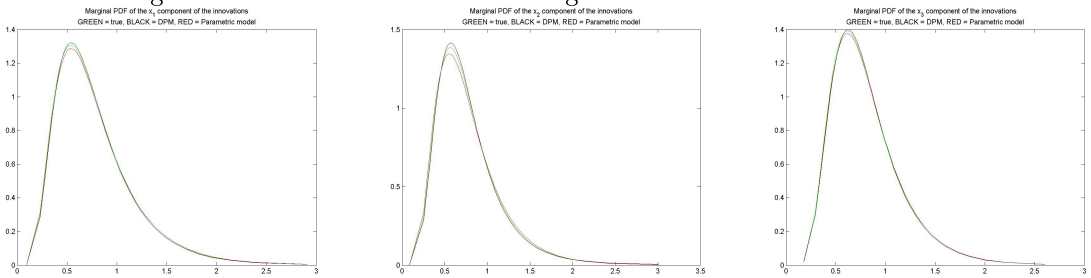


Figure 4.2.5: True and estimated marginal densities of the innovations.



4.2.2 Base full specification

We now simulate our sample from the base full specification

$$\begin{cases} \mathbf{x}_t = & \boldsymbol{\mu}_t \odot \boldsymbol{\varepsilon}_t \\ \boldsymbol{\mu}_t = & \boldsymbol{\omega} + \mathbf{B}\boldsymbol{\mu}_{t-1} + \mathbf{A}\mathbf{x}_{t-1} \end{cases}$$

with

$$\boldsymbol{\omega} = \begin{bmatrix} 0.35 \\ 0.59 \\ 0.43 \end{bmatrix}, \ \mathbf{B} = \begin{bmatrix} 0.36 & 0.07 & 0.18 \\ 0.20 & 0.24 & 0.14 \\ 0.01 & 0.10 & 0.41 \end{bmatrix}, \ \mathbf{A} = \begin{bmatrix} 0.21 & 0.14 & 0.04 \\ 0.13 & 0.28 & 0.09 \\ 0.07 & 0.08 & 0.30 \end{bmatrix}$$

and

$$\boldsymbol{\varepsilon}_t \overset{i.i.d.}{\sim} 0.7 * logN_2 \left(\begin{bmatrix} -0.200 \\ -0.175 \\ -0.150 \end{bmatrix}, \begin{bmatrix} 0.40 & 0.30 & 0.20 \\ 0.30 & 0.35 & 0.25 \\ 0.20 & 0.25 & 0.30 \end{bmatrix} \right) + 0.3 * logN_2 \left(\begin{bmatrix} -0.185 \\ -0.195 \\ -0.125 \end{bmatrix}, \begin{bmatrix} 0.37 & 0.15 & 0.24 \\ 0.15 & 0.39 & 0.18 \\ 0.24 & 0.18 & 0.25 \end{bmatrix} \right).$$

We run the algorithm for $N = 200,000$ iterations and we discard the first 20,000 as burn-in. To sample the parameters of the conditional mean we set $p = 0.9$ as the weight for the proposal mixture density and $\sigma_1 = 1$, $\sigma_2 = \sqrt{21}$ as its scale factors. The simulation time on a server running at 2.60GHz and with 128GB RAM is about 13 hours. In Table 4.2.2 we report the posterior means and 95% credible intervals for the parameters of the conditional mean, along with their true values. As it can be seen, again, all the true values of the parameters lie inside the 95% posterior credible intervals. All the estimates are based on effective sample sizes greater than 327. In Figures 4.2.6, 4.2.7, 4.2.8 we reported the traces, the posterior densities and the autocorrelation functions of the post-processed parameters of the conditional mean. These figures show that all the traces have reached convergence and all the autocorrelation functions become non-significant in less than 3000 lags (most of them in less than 2000). In Figure 4.2.9 are reported the traces and the running averages of the maximum number of components and of the number of active components, at each step, along with the traces of the mixture weights. We can see that there are on average 7 active components but, correctly, only two of them have really significant weights. Finally in Figure 4.2.10 we report the true marginal densities of the innovations together with the estimated marginal densities obtained with DMPLN2-vMEM and the LN1-vMEM. Also in this case, the approximation obtained with the DPMLN2-vMEM is better than the one obtained with the LN1-vMEM.

Figure 4.2.6: MCMC traces, posterior densities and ACF of the components of the post-processed vector $\boldsymbol{\omega}$. The green lines in the histogram represent the 95% C.I. while the red one is the true value.

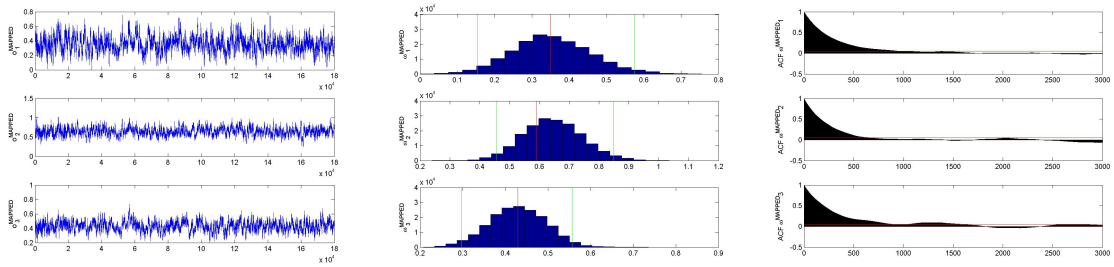


Table 4.2.2: Posterior mean and 95% credible intervals for the parameters of the conditional mean.

	True	Est.	(95% C.I.)		True	Est.	(95% C.I.)
ω_1	0.35	0.3532	$(0.1531, 0.5743)$	α_{11}	0.21	0.2083	$(0.1658, 0.2521)$
ω_2	0.59	0.6455	$(0.4572, 0.8484)$	α_{21}	0.13	0.1359	$(0.0946, 0.1770)$
ω_3	0.43	0.4290	$(0.2965, 0.5566)$	α_{31}	0.07	0.0792	$(0.0473, 0.1095)$
β_{11}	0.36	0.1417	$(-0.1222, 0.3800)$	α_{12}	0.14	0.1255	$(0.0744, 0.1765)$
β_{21}	0.10	0.0864	$(-0.2156, 0.3987)$	α_{22}	0.28	0.2469	$(0.1966, 0.2977)$
β_{31}	0.01	-0.0741	$(-0.3631, 0.2235)$	α_{32}	0.08	0.0737	$(0.0348, 0.1102)$
β_{12}	0.07	0.1346	$(-0.0989, 0.4058)$	α_{13}	0.04	0.0534	$(0.0014, 0.1074)$
β_{22}	0.24	0.2126	$(-0.0449, 0.4671)$	α_{23}	0.09	0.1046	$(0.0512, 0.1572)$
β_{32}	0.10	0.1290	$(-0.1039, 0.3946)$	α_{33}	0.30	0.2739	$(0.2308, 0.3190)$
β_{13}	0.18	0.2390	$(0.0909, 0.4126)$				
β_{23}	0.14	0.1884	$(0.0206, 0.3706)$				
β_{33}	0.41	0.4788	$(0.3109, 0.6359)$				

Figure 4.2.7: MCMC traces, posterior densities and ACF of the components of the post-processed vector $\boldsymbol{\beta}$. The green lines in the histogram represent the 95% C.I. while the red one is the true value.

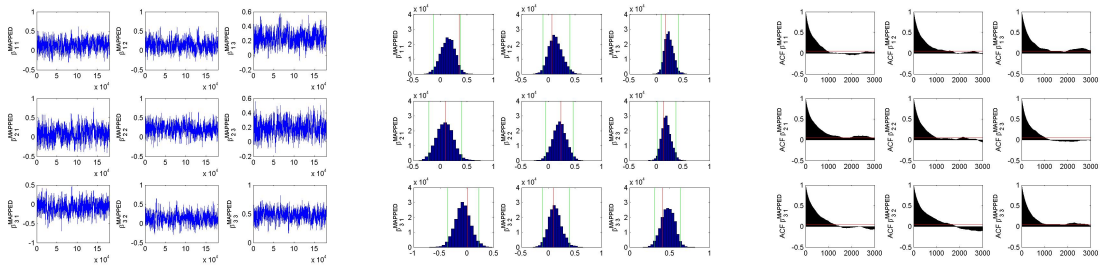


Figure 4.2.8: MCMC traces, posterior densities and ACF of the components of the post-processed matrix $\mathbf{A}$. The green lines in the histogram represent the 95% C.I. while the red one is the true value.

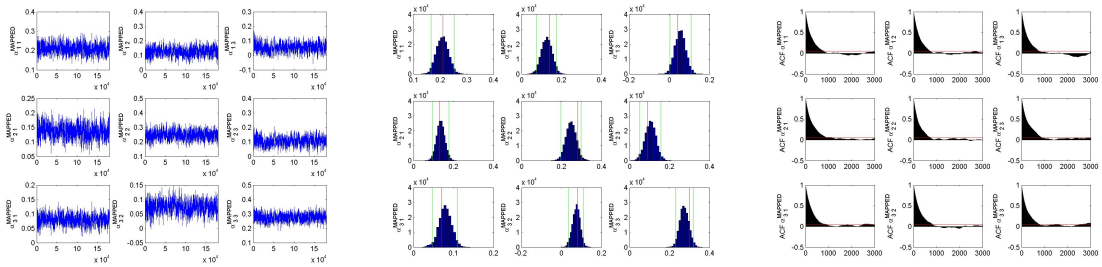


Figure 4.2.9: The upper left plot shows the traces of total number of components and of the number of active components at each step. The lower left plot shows the corresponding running averages. The plot on the right shows the traces of the mixture weights.

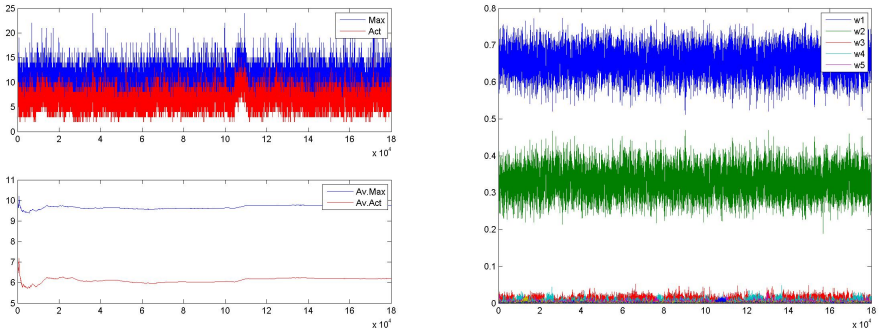
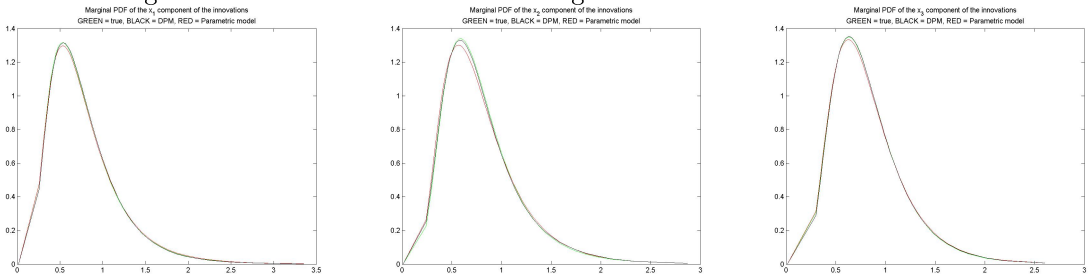


Figure 4.2.10: True and estimated marginal densities of the innovations.



Chapter 5

Empirical Analyses: Interdependence across volatility measures.

5.1 Introduction

Volatility measurement using intra-daily data was born with the simplest daily range proposed by Parkinson (1980) and evolved since then passing through the plain vanilla realized volatility proposed by Andersen and Bollerslev (1998) arriving to the more recent realized kernels proposed by Barndorff-Nielsen, Hansen, Lunde, and Shephard (2008). In parallel to the evolution of these measures has been done a natural complementary effort to build adequate models for volatility prediction and Multiplicative Error Models have been used for this purpose (see, for example, Cipollini, Engle, and Gallo (2013)). The time series which are most commonly used in this respect are the squared close-to-close returns r_t^2, the realized variances rv_t^2 (in any of their flavours), the absolute returns $|r_t|$, the realized volatilities rv_t, and the daily ranges hl_t. All these series posses some very well known general features like persistence and alternance of periods of high and low volatility. From an empirical point of view, we will now illustrate the characteristics of our DPMLN2-vMEM in modelling the interaction among several volatility measures for the purpose of forecasting. We will consider four different specifications of the conditional mean for our semiparametric model:

Specification 1: $\boldsymbol{\mu}_t = \boldsymbol{\omega} + \boldsymbol{\beta} \odot \boldsymbol{\mu}_{t-1} + \mathbf{A}\mathbf{x}_{t-1}$

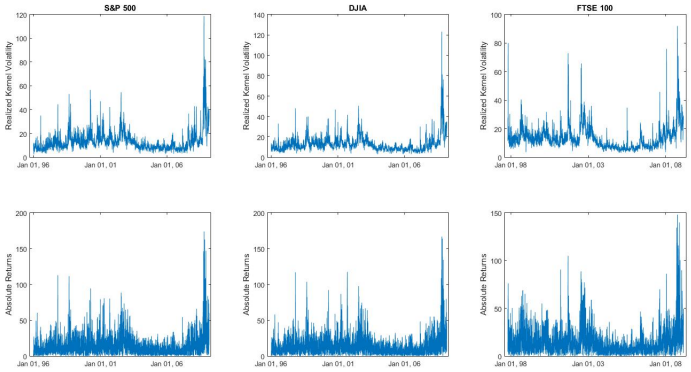
Specification 2: $\boldsymbol{\mu}_t = \boldsymbol{\omega} + \boldsymbol{\beta} \odot \boldsymbol{\mu}_{t-1} + \mathbf{A}\mathbf{x}_{t-1} + \boldsymbol{\alpha}^{(-)} \odot \mathbf{x}_{t-1}\mathbb{I}_{\{r_{t-1}<0\}} + \boldsymbol{\varphi} \odot \mathbf{x}_{t-2}$

Specification 3: $\boldsymbol{\mu}_t = \boldsymbol{\omega} + \boldsymbol{\beta} \odot \boldsymbol{\mu}_{t-1} + \mathbf{A}\mathbf{x}_{t-1} + \boldsymbol{\alpha}^{(-)} \odot \mathbf{x}_{t-1}\mathbb{I}_{\{r_{t-1}<0\}} + \mathbf{F}\mathbf{x}_{t-2}$

Specification 4: $\boldsymbol{\mu}_t = \boldsymbol{\omega} + \boldsymbol{\beta} \odot \boldsymbol{\mu}_{t-1} + \mathbf{A}\mathbf{x}_{t-1} + \mathbf{A}^{(-)}\mathbf{x}_{t-1}\mathbb{I}_{\{r_{t-1}<0\}} + \mathbf{F}\mathbf{x}_{t-2}$

We will make a comparison between the DPMLN2-vMEM with each specification of the conditional mean and the analogue LN1-vMEM (estimated using MLE) in terms of their (in the sample) predictive performance. To do this we will use the Log-Predictive Score (LPS) proposed by Kim,

Figure 5.1.1: Time series of realized volatilities and absolute returns of S&P 500, DJIA, FTSE 100



Shephard, and Chib (1998), defined as:

$$LPS = -\frac{1}{T}\sum_{t=1}^{T} \log \hat{f}_{\mathbf{x}_t}\left(\mathbf{x}_t\right) = -\frac{1}{T}\sum_{t=1}^{T} \log\left(\prod_{h=1}^{d} \frac{1}{\hat{\mu}_h^{(t)}} \hat{f}_{\boldsymbol{\epsilon}_t}\left(\mathbf{x}_t \oslash \hat{\boldsymbol{\mu}}_t\right)\right)$$

where the probability density function of the innovations has been estimated as:

$$\hat{f}_{\boldsymbol{\epsilon}_t}\left(\cdot\right) = \frac{1}{N}\sum_{n=1}^{N}\sum_{j=1}^{N_{Trunc}} w_j^{(n)} logN_d\left(\cdot; \mathbf{m}_j^{(n)}, \boldsymbol{\Sigma}_j^{(n)}\right)$$

and N_{trunc} is the minimum number of components that appear in all the MCMC steps and is such that $\sum_{j=1}^{N_{Trunc}} w_j^{(n)} > 0.99$. By this definition, a lower LPS is an indication of a better predictive performance.

For our analysis we make use of a bivariate series composed by daily absolute returns and realized kernel volatilities, $(|r_t|, rv_t)$. We take the data from the Oxford Man Institute "Realized Library" (Shephard and Sheppard (2010)) and we express them in annualized percentage terms through the transformation:

$$x_t^{AP} = 100\sqrt{252}x_t.$$

We run our analysis on three stock indices: Standard & Poor 500 (S&P 500), Dow Jones Industrial Average (DJIA), Financial Times Stock Exchange 100 (FTSE 100). The covered period is the one between January 1996 and February 2009 for a total of 3261 observations for the S&P 500 series, 3260 observations for the DJIA series and 2840 observations of the FTSE 100 series. From the time series plotted in Figure 5.1.1, we can see that both the measures of all the three indices share some common features like alternance of periods of high and low volatility and persistence.

In the next sections we will compare the estimates obtained for the different models and the different time series using our DPMLN2-vMEM and the corresponding ones obtained using a LN1-vMEM estimated with Maximum Likelihood.

For the first specification of the conditional mean, for all the time series, we run 150,000 iterations of the algorithm described in Section 3.4 and then discard the first 30,000 of them as burn-in. For

all the other specifications of the conditional mean, for all the time series, we run 200,000 iterations of the same algorithm and then discard the first 50,000 of them as burn-in. In Tables 5.2.1, 5.3.1, 5.4.1 and 5.5.1 we report, for each specification of the conditional mean, the estimates of its parameters and the 95% credible intervals obtained with our model for the three time series, together with the maximum likelihood estimates of the same parameters and the corresponding standard errors obtained from the LN1-vMEM, for all the time series. In all the analyses, all the effective sample sizes of the variables obtained from the MCMC simulations are bigger than 500. In Tables 5.2.2, 5.3.2, 5.4.2 and 5.5.2 we report, for each specification of the conditional mean, the Log-Predictive scores of the DPMLN2-vMEM and the LN1-vMEM, for all the three time series. Figures 5.2.1, 5.3.1, 5.4.1 and 5.5.1 show a significant examples of the traces, the posterior histograms and the autocorrelation functions of the parameters of each specification of the conditional mean. Finally Figures 5.2.2, 5.3.2, 5.4.2 and 5.5.2 show the estimated joint and marginal densities of the innovations over the estimated innovations obtained with all the four specifications[1].

5.2 Specification 1

We first fit to the time series described above our semiparametric model with Specification 1 of the conditional mean.

Table 5.2.1: Posterior mean, 95% credible intervals, ML estimates and corresponding standard errors, for the parameters of the conditional mean.

	S&P 500		DJIA		FTSE 100	
	MCMC Est. (95% C.I.)	MLE Est. (S.d.)	MCMC Est. (95% C.I.)	MLE Est. (S.d.)	MCMC Est. (95% C.I.)	MLE Est. (S.d.)
ω_1	0.00835 $(-0.1266, 0.3211)$	-0.0361 0.2473	0.0157 $(-0.2783, 0.3423)$	-0.1157 0.2527	0.0748 $(-0.1332, 0.3047)$	-0.0486 0.2155
ω_2	0.4343 $(0.3102, 0.5623)$	0.4685 0.0603	0.4446 $(0.3229, 0.5698)$	0.4519 0.0571	0.2244 $(0.1415, 0.3140)$	0.2089 0.0401
β_1	0.7111 $(0.6357, 0.7754)$	0.6379 0.0523	0.6115 $(0.5189, 0.6967)$	0.6386 0.0525	0.6455 $(0.5692, 0.7149)$	0.6629 0.0624
β_2	0.5609 $(0.5221, 0.5980)$	0.5603 0.0157	0.5486 $(0.5091, 0.5862)$	0.5621 0.0154	0.6421 $(0.5978, 0.6852)$	0.6735 0.012506
α_{11}	-0.0428 $(-0.0667, -0.0183)$	-0.1131 0.0230	-0.0451 $(-0.0750, -0.0151)$	-0.0925 0.0251	-0.0245 $(-0.0552, 0.0065)$	-0.0574 0.0282
α_{21}	0.0405 $(0.0308, 0.0505)$	0.0393 0.0052	0.0370 $(0.0280, 0.0461)$	0.0368 0.0047	0.0305 $(0.0210, 0.0403)$	0.0325 0.0046
α_{12}	0.3452 $(0.2723, 0.4276)$	0.5915 0.0760	0.4514 $(0.3537, 0.5558)$	0.5610 0.0763	0.4042 $(0.3167, 0.5000)$	0.5137 0.0973
α_{22}	0.3636 $(0.3307, 0.3973)$	0.3625 0.0150	0.3782 $(0.3447, 0.4134)$	0.3641 0.0144	0.3086 $(0.2701, 0.3486)$	0.2758 0.0125

[1]For the sake of brevity, we will plot hereafter only the traces, histograms and autocorrelation functions of the elements of the $\boldsymbol{\omega}$ vector and the estimations of the joint density and its marginals obtained analysing the time series of the S&P 500 index with DPMLN2-vMEM and LN1-vMEM. Very similar figures for other parameters and for the other two time series are reported in Appendix B.

Table 5.2.2: Log-Predictive Scores for parametric and semiparametric models, for all the three series.

	S&P 500	DJIA	FTSE 100
LN1-vMEM	6.1795	6.0238	6.1318
DPMLN2-vMEM	6.0176	5.8676	5.9124

Figure 5.2.1: MCMC traces, posterior histograms and ACFs of the components of the post-processed vector $\boldsymbol{\omega}$. The green lines in the histogram represent the 95% C.I.

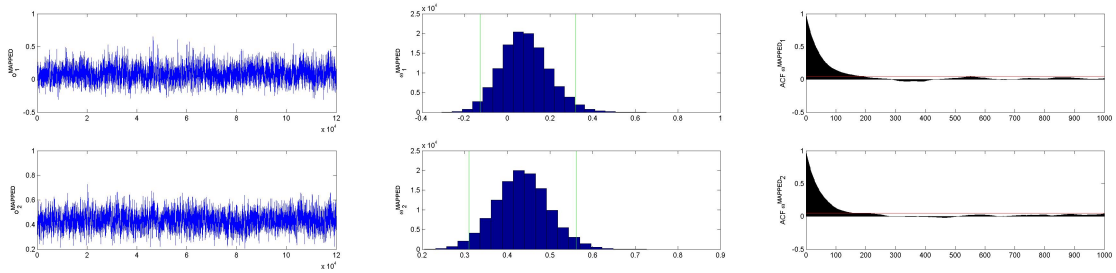
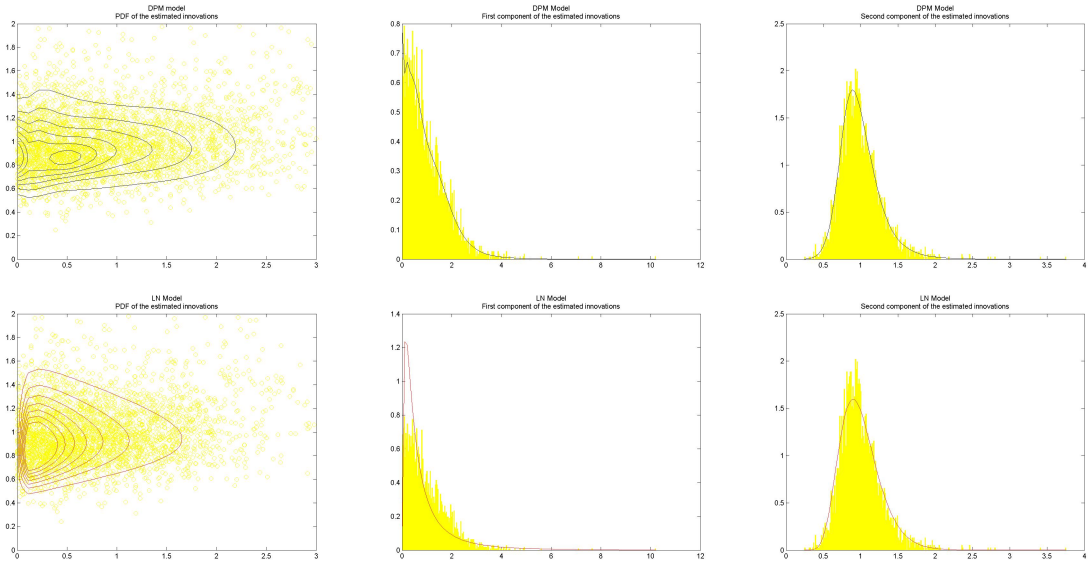


Figure 5.2.2: Estimated joint and marginal densities of the innovations over the estimated innovations. On the upper row there are the results obtained with the DPMLN2-vMEM, on the lower row the ones obtained with LN1-vMEM



5.3 Specification 2

We will now fit our semiparametric model with Specification 2 of the conditional mean.

Table 5.3.1: Posterior mean, 95% credible intervals, ML estimates and corresponding standard errors, for the parameters of the conditional mean.

	S&P 500		DJIA		FTSE 100	
	MCMC Est. (95% C.I.)	MLE Est. (S.d.)	MCMC Est. (95% C.I.)	MLE Est. (S.d.)	MCMC Est. (95% C.I.)	MLE Est. (S.d.)
ω_1	0.1543 $(-0.0025, 0.3211)$	0.1030 0.1490	0.0635 $(-0.1573, 0.3039)$	0.0368 0.1245	0.1147 $(-0.0512, 0.2931)$	0.1423 0.1084
ω_2	0.3810 $(0.2812, 0.4913)$	0.3360 0.0452	0.3368 $(0.2496, 0.4306)$	0.2812 0.0378	0.1453 $(0.0936, 0.2039)$	0.1196 0.0215
β_1	0.7896 $(0.7285, 0.8393)$	0.7914 0.0359	0.7092 $(0.6184, 0.7858)$	0.8358 0.0312	0.7179 $(0.6224, 0.8049)$	0.8314 0.0348
β_2	0.6488 $(0.5952, 0.6960)$	0.7111 0.0195	0.6790 $(0.6299, 0.7237)$	0.7517 0.0189	0.7720 $(0.7300, 0.8116)$	0.8376 0.0132
α_{11}	-0.1077 $(-0.1416, -0.0731)$	-0.2068 0.0267	-0.0959 $(-0.1287, -0.0614)$	-0.1690 0.0298	-0.0831 $(-0.1210, -0.0436)$	-0.1776 0.0324
α_{21}	0.0352 $(0.0261, 0.0442)$	0.0310 0.0046	0.0319 $(0.0240, 0.0400)$	0.0290 0.0040	0.0209 $(0.0130, 0.0292)$	0.0187 0.0033
α_{12}	0.2262 $(0.1665, 0.2991)$	0.2926 0.0563	0.3156 $(0.2263, 0.4188)$	0.2161 0.0483	0.2888 $(0.1835, 0.4038)$	0.2129 0.0548
α_{22}	0.3143 $(0.2757, 0.3518)$	0.3411 0.0193	0.3316 $(0.2951, 0.3684)$	0.3531 0.0181	0.2934 $(0.2534, 0.3344)$	0.3256 0.0171
$\alpha_1^{(-)}$	0.0896 $(0.0608, 0.1194)$	0.1291 0.0249	0.0850 $(0.0520, 0.1167)$	0.1245 0.0221	0.0566 $(0.0161, 0.0949)$	0.1604 0.0236
$\alpha_2^{(-)}$	0.0612 $(0.0465, 0.0759)$	0.0568 0.0071	0.0549 $(0.0432, 0.0668)$	0.0514 0.0059	0.0353 $(0.0237, 0.0469)$	0.0389 0.0055
φ_1	0.0492 $(0.0181, 0.0815)$	0.1137 0.0280	0.0380 $(0.0069, 0.0690)$	0.0961 0.0280	0.0615 $(0.0272, 0.0973)$	0.0971 0.0315
φ_2	-0.0604 $(-0.1112, -0.0067)$	-0.1398 0.0238	-0.0979 $(-0.1480, -0.0468)$	-0.1831 0.0241	-0.1168 $(-0.1661, -0.0692)$	-0.2127 0.0197

Table 5.3.2: Log-Predictive Scores for parametric and semiparametric models, for all the three series.

	S&P 500	DJIA	FTSE 100
LN1-vMEM	6.1606	6.0020	6.1069
DPMLN2-vMEM	5.9986	5.8469	5.8908

Figure 5.3.1: MCMC traces, posterior histograms and ACFs of the components of the post-processed vector $\boldsymbol{\omega}$. The green lines in the histogram represent the 95% C.I.

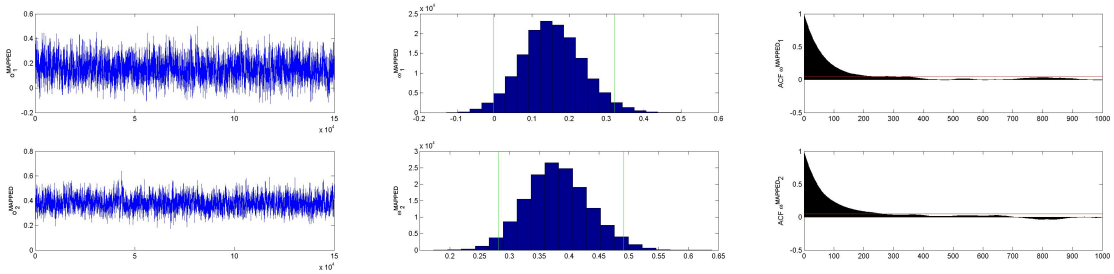
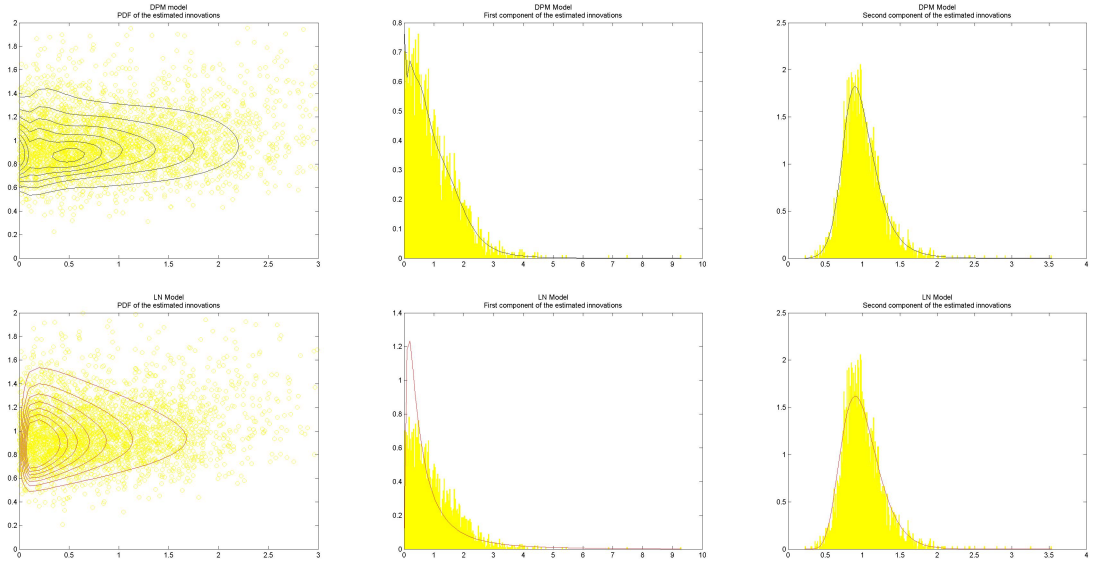


Figure 5.3.2: Estimated joint and marginal densities of the innovations over the estimated innovations. On the upper row there are the results obtained with the DPMLN2-vMEM, on the lower row the ones obtained with LN1-vMEM



5.4 Specification 3

Now we fit to the time series our DPMLN2-vMEM, with Specification 3 of the conditional mean.

Table 5.4.1: Posterior mean, 95% credible intervals, ML estimates and corresponding standard errors, for the parameters of the conditional mean.

	S&P 500		DJIA		FTSE 100	
	MCMC Est. (95% C.I.)	MLE Est. (S.e.)	MCMC Est. (95% C.I.)	MLE Est. (S.e.)	MCMC Est. (95% C.I.)	MLE Est. (S.e.)
ω_1	0.1404 $(-0.0165, 0.3051)$	0.0623 0.1127	0.0418 $(-0.1682, 0.2602)$	0.0136 0.0862	0.1022 $(-0.0760, 0.2880)$	0.1408 0.0939
ω_2	0.3701 $(0.2715, 0.4832)$	0.3087 0.0434	0.3407 $(0.2500, 0.4457)$	0.2655 0.0371	0.1425 $(0.0901, 0.2015)$	0.1049 0.0196
β_1	0.7910 $(0.7097, 0.8532)$	0.8486 0.0384	0.6409 $(0.4790, 0.7843)$	0.8931 0.0289	0.6832 $(0.5311, 0.8194)$	0.8559 0.0353
β_2	0.6541 $(0.5990, 0.7042)$	0.7280 0.0193	0.6741 $(0.6224, 0.7220)$	0.7619 0.0190	0.7708 $(0.7243, 0.8143)$	0.8530 0.0123
α_{11}	-0.1062 $(-0.1376, -0.0730)$	-0.2025 0.0271	-0.0891 $(-0.1212, -0.0549)$	-0.1636 0.0300	-0.0733 $(-0.1114, -0.0346)$	-0.1640 0.0332
α_{21}	0.0370 $(0.0264, 0.0479)$	0.0359 0.0058	0.0333 $(0.0233, 0.0431)$	0.0335 0.0053	0.0335 $(0.0226, 0.0446)$	0.0382 0.0056
α_{12}	0.2501 $(0.1472, 0.3561)$	0.4694 0.1095	0.3593 $(0.2454, 0.4811)$	0.4251 0.1082	0.3122 $(0.1867, 0.4454)$	0.2546 0.1177
α_{22}	0.3162 $(0.2768, 0.3543)$	0.3475 0.0192	0.3355 $(0.2981, 0.3741)$	0.3586 0.0181	0.2914 $(0.2511, 0.3325)$	0.3176 0.0179
$\alpha_1^{(-)}$	0.0866 $(0.0572, 0.1167)$	0.1042 0.0246	0.0743 $(0.0406, 0.1071)$	0.0963 0.0211	0.0522 $(0.0082, 0.0930)$	0.1572 0.0243
$\alpha_2^{(-)}$	0.0601 $(0.0460, 0.0745)$	0.0534 0.0070	0.0546 $(0.0422, 0.0671)$	0.0494 0.0059	0.0346 $(0.0232, 0.0460)$	0.0376 0.0052
φ_{11}	0.0470 $(0.0140, 0.0788)$	0.1281 0.0288	0.0273 $(-0.0072, 0.0609)$	0.1078 0.0284	0.0463 $(0.0087, 0.0815)$	0.0845 0.0330
φ_{21}	-0.0035 $(-0.0143, 0.0076)$	-0.0095 0.0060	-0.0012 $(-0.0114, 0.0085)$	-0.0073 0.0056	-0.0165 $(-0.0277, -0.0054)$	-0.0251 0.0059
φ_{12}	-0.0218 $(-0.1497, 0.1114)$	-0.2543 0.1279	0.0398 $(-0.1354, 0.2199)$	-0.282525 0.120276	0.0237 $(-0.1476, 0.1936)$	-0.0736 0.1277
φ_{22}	-0.0642 $(-0.1159, -0.0105)$	-0.1543 0.0235	-0.0971 $(-0.1491, -0.0455)$	-0.193756 0.023992	-0.1089 $(-0.1598, -0.0600)$	-0.2123 0.0201

Table 5.4.2: Log-Predictive Scores for parametric and semiparametric models, for all the three series.

	S&P 500	DJIA	FTSE 100
LN1-vMEM	6.1595	6.0009	6.1033
DPMLN2-vMEM	6.0007	5.8437	5.8926

Figure 5.4.1: MCMC traces, posterior histograms and ACFs of the components of the post-processed vector $\boldsymbol{\omega}$. The green lines in the histogram represent the 95% C.I.

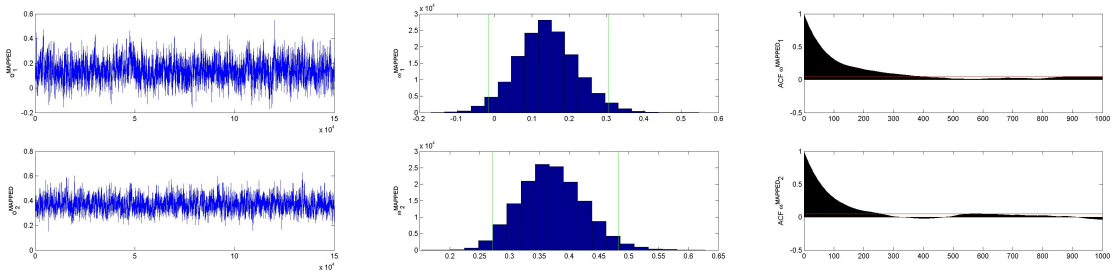
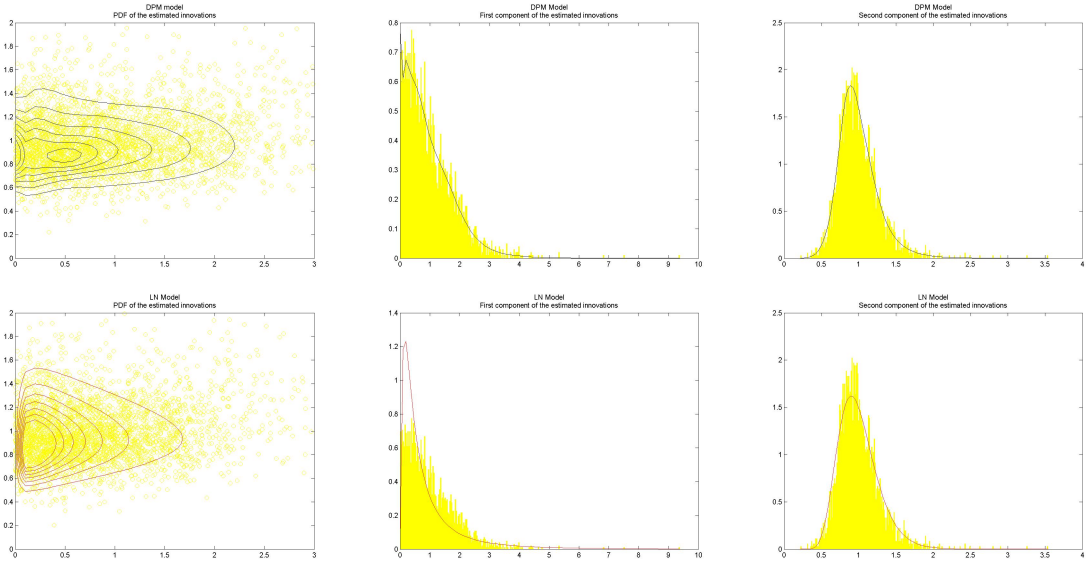


Figure 5.4.2: Estimated joint and marginal densities of the innovations over the estimated innovations. On the upper row there are the results obtained with the DPMLN2-vMEM, on the lower row the ones obtained with LN1-vMEM



5.5 Specification 4

Finally we fit to the time series our DPMLN2-vMEM, with Specification 4 of the conditional mean.

Table 5.5.1: Posterior mean, 95% credible intervals, ML estimates and corresponding standard errors, for the parameters of the conditional mean.

	S&P 500		DJIA		FTSE 100	
	MCMC Est. (95% C.I.)	MLE Est. (S.e.)	MCMC Est. (95% C.I.)	MLE Est. (S.e.)	MCMC Est. (95% C.I.)	MLE Est. (S.e.)
ω_1	0.1682 (0.0305, 0.3112)	0.1225 0.1133	0.0757 (−0.1403, 0.3074)	0.0425 0.0919	0.1274 (0.0127, 0.2535)	0.1633 0.0917
ω_2	0.3833 (0.2993, 0.4815)	0.3816 0.0413	0.3603 (0.2739, 0.4552)	0.3061 0.0364	0.1271 (0.0847, 0.1758)	0.1175 0.0187
β_1	0.8270 (0.7673, 0.8760)	0.8475 0.0354	0.6805 (0.5275, 0.8035)	0.8856 0.0297	0.8305 (0.7315, 0.8954)	0.8608 0.0319
β_2	0.7123 (0.6666, 0.7539)	0.7307 0.0180	0.6906 (0.6443, 0.7334)	0.7596 0.0184	0.8217 (0.7840, 0.8543)	0.8580 0.0117
α_{11}	-0.1170 (−0.1504, −0.0819)	-0.2026 0.0274	-0.0963 (−0.1332, −0.0586)	-0.1690 0.0307	-0.0885 (−0.1257, −0.0484)	-0.1621 0.0345
α_{21}	-0.0039 (−0.0162, 0.0081)	-0.0045 0.0069	0.0032 (−0.0074, 0.0142)	0.0046 0.0063	0.013283 (0.0010, 0.0259)	0.0166 0.0066
α_{12}	0.2221 (0.1175, 0.3268)	0.4052 0.1094	0.3152 (0.1954, 0.4402)	0.3910 0.1097	0.2381 (0.1260, 0.3551)	0.2155 0.1193
α_{22}	0.2884 (0.2519, 0.3257)	0.3106 0.0190	0.3121 (0.2760, 0.3496)	0.3336 0.0181	0.2684 (0.2305, 0.3074)	0.3048 0.0180
$\alpha_{11}^{(-)}$	0.1054 (0.0687, 0.1442)	0.1027 0.0321	0.0851 (0.0379, 0.1339)	0.1084 0.0301	0.0870 (0.0405, 0.1315)	0.1478 0.0378
$\alpha_{21}^{(-)}$	0.0934 (0.0780, 0.1089)	0.0902 0.0080	0.0692 (0.0552, 0.0836)	0.0643 0.0072	0.0450 (0.0317, 0.0594)	0.0433 0.0069
$\alpha_{12}^{(-)}$	0.0022 (−0.0438, 0.0462)	0.0504 0.0387	0.0277 (−0.0350, 0.0883)	0.0083 0.0353	0.0046 (−0.0466, 0.0549)	0.0386 0.0479
$\alpha_{22}^{(-)}$	-0.0206 (−0.0385, −0.0025)	-0.0216 0.0097	-0.0032 (−0.0198, 0.0127)	-0.0060 0.0086	-0.0037 (−0.0200, 0.0121)	-0.0016 0.0087
φ_{11}	0.0551 (0.0234, 0.0864)	0.1295 0.0286	0.0343 (−0.0019, 0.0692)	0.1085 0.0285	0.0554 (0.0210, 0.0900)	0.0854 0.0327
φ_{21}	-0.0005 (−0.0114, 0.0099)	-0.0025 0.0059	0.0013 (−0.0078, 0.0107)	-0.0046 0.0056	-0.0184 (−0.0294, −0.0072)	-0.0234 0.0059
φ_{12}	-0.0420 (−0.1657, 0.0830)	-0.2194 0.1251	0.0210 (−0.1468, 0.1899)	-0.2465 0.1215	-0.0784 (−0.2184, 0.0683)	-0.0602 0.1251
φ_{22}	-0.0629 (−0.1119, −0.0146)	-0.0990 0.0227	-0.0689 (−0.1165, −0.0209)	-0.1468 0.0238	-0.1164 (−0.1626, −0.0711)	-0.1872 0.0202

Table 5.5.2: Log-Predictive Scores for parametric and semiparametric models, for all the three series.

	S&P 500	DJIA	FTSE 100
LN1-vMEM	6.1398	5.9885	6.0961
DPMLN2-vMEM	5.9783	5.8305	5.8848

Figure 5.5.1: MCMC traces, posterior histograms and ACFs of the components of the post-processed vector $\boldsymbol{\omega}$. The green lines in the histogram represent the 95% C.I.

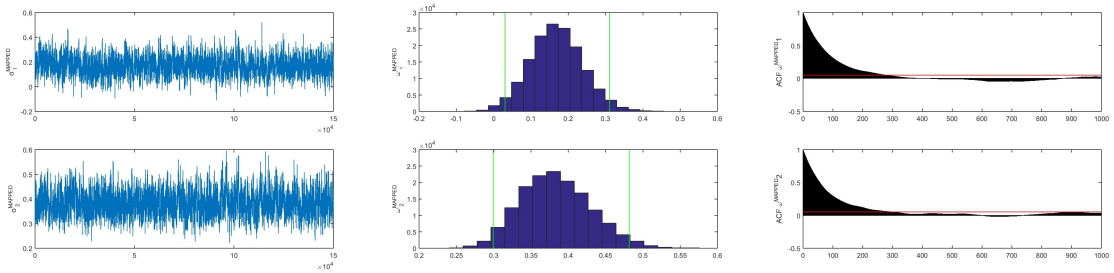
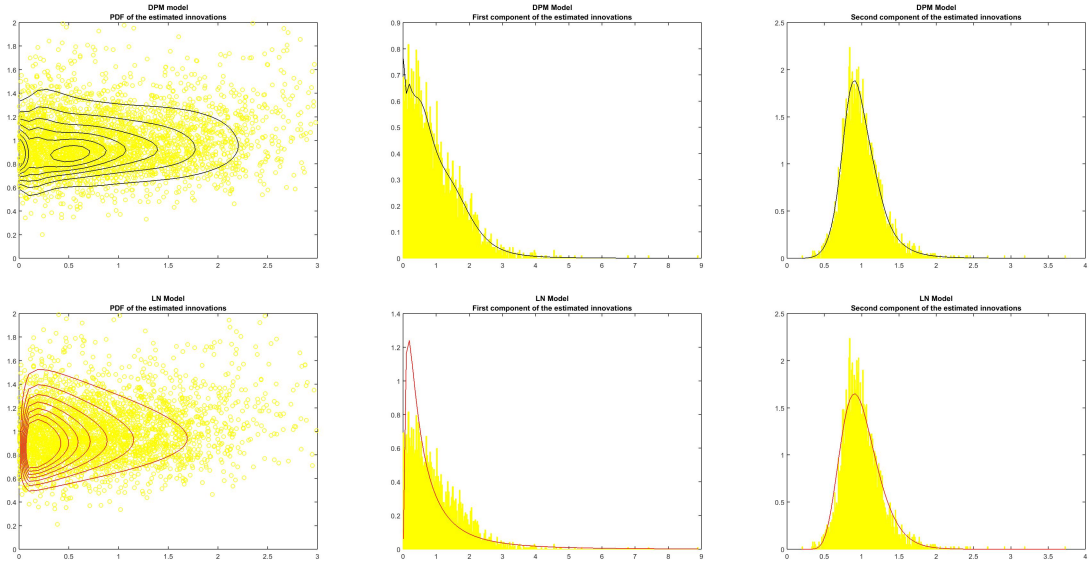


Figure 5.5.2: Estimated joint and marginal densities of the innovations over the estimated innovations. On the upper row there are the results obtained with the DPMLN2-vMEM, on the lower row the ones obtained with LN1-vMEM



5.6 Discussion

First of all, from Figures 5.2.1, 5.3.1, 5.4.1 and 5.5.1 we can see that, although there is some autocorrelation, for every specification of the conditional mean, the traces of the MCMC simulations have all reached convergence and the posterior histograms look informative.

For all the specifications of the conditional mean, for all the three time series and both the models considered, we obtain some common qualitative features of the point estimates. First of all the βs are always the biggest coefficients, meaning that the factor that influences the most the evolution of the conditional mean is always its lagged realization. Second, the estimates of the coefficients of the second column of matrix $\mathbf{A}$ are always bigger, in absolute value, than the ones in the first column of the same matrix. This suggests that the lagged realizations of the realized kernel volatility influence the evolution of both the components of the conditional mean vector more than the lagged realizations of the absolute returns. This fact, that could look strange at first sight, simply means that the lagged observation of the realized volatility contains more information on the present realization of the conditional mean of the absolute returns than the lagged absolute returns and this can be viewed as a further proof of the fact that the realized volatilities are more informative about the latent volatility than the absolute returns. Furthermore the α_{11} coefficient is always negative meaning that the conditional mean of the absolute returns depends inversely from the lagged realizations of the absolute returns. For what it takes the parameters in matrices $\mathbf{A}^{(-)}$ and $\mathbf{F}$, when present, they are always smaller than the coefficients of $\boldsymbol{\beta}$ and of the second column of the matrix $\mathbf{A}$ meaning that, although their introduction contribute to the increment of the predictive power of the model (as proven by the Log-Predictive Scores), their influence on the evolution of the conditional mean is of secondary importance.

In all the empirical analyses we tried, for all the specifications adopted for the conditional mean, there are always about the same average number of active and total components of the DPM, for all the time series (as shown by the graphs in Appendix B). Furthermore for all the time series and all the specifications of the conditional mean there are always at least eight or nine components of the DPM for the whole MCMC run. The sum of the average of the weights of these components is always bigger than 0.99, meaning that, even in the MCMC steps in which there are more components, the first eight or nine are dominant.

Regarding the approximation of the distribution of the data obtained with the DPMLN2-vMEM and with the LN1-vMEM with all the four specifications of the conditional mean, Figures 5.2.2, 5.3.2, 5.4.2 and 5.5.2 show that our semiparametric model outperforms its parametric counterpart. This can be perceived from the graphs of the joint distribution, but it becomes even more clear looking at the graphs of the marginals: for all the specifications of the conditional mean, the approximation (in particular of the first marginals) obtained with DPMLN2-vMEM is much better than the one obtained with LN1-vMEM.

For what it takes the predictive performances in the sample, Tables 5.2.2, 5.3.2, 5.4.2 and 5.5.2 clearly show that the DPMLN2-vMEM performs always better the LN1-vMEM. In Table 5.6.1 we report all together the Log-Predictive Scores that we already presented singularly in the previous sections, to compare more easily the predictive powers of the two models when the specification of the conditional mean changes. As we can see looking at the columns of this table, for all the time series considered and with both the DPMLN2 and LN1 models we have almost the same trend of the Log-Predictive Scores for increasing model complexity. In particular, passing from the first to the second specification causes a small increase of the predictive power. Passing from the second to the third specification generally does not cause a significant decrease of the Log-Predictive Scores

Table 5.6.1: Log-Predictive Scores

	DPMLN2			LN1		
	SP500	FTSE100	DJIA	SP500	FTSE100	DJIA
Spec. 1	6.0176	5.9124	5.8676	6.1795	6.1318	6.0238
Spec. 2	5.9986	5.8908	5.8469	6.1606	6.1069	6.0020
Spec. 3	6.0007	5.8926	5.8437	6.1595	6.1033	6.0009
Spec. 4	5.9783	5.8848	5.8305	6.1398	6.0961	5.9885

(and also, in the case of the DPMLN2-vMEM for the S&P 500 and FTSE 100 time series, there is a slight increase), meaning that the introduction of doubly-lagged observations of the other volatility proxy do not increase significantly the predictive power of the model. Finally passing from the third to the fourth specification there is a general decrease of the Log-Predictive Scores, meaning that introducing the asymmetric observations of the other volatility proxy generally increases the predictive power of the model.

Conclusions

With this work we gave a novel contribution to the modelling and computational aspects of vector Multiplicative Error Models. We proposed a semiparametric specification for the innovation term of a vector MEM using a location-scale Dirichlet Process Mixture of multivariate log-normal distribution and obtained a framework that is more flexible and efficient in comparison with the standard parametric one. We also introduced a very general specification for the conditional mean that nests most of the specifications used in the literature allowing the model to represent many features of financial time series. The parameter expansion technique enabled us to simplify some aspects of our adaptation of the slice sampler obtaining an efficient and relatively fast algorithm. The proposed semiparametric model showed better performances than the naive parametric counterpart in the fitting of both the simulations and the empirical data. Possible development of this research line include (but are not limited to) a further refinement of the sampling technique used to sample the parameters of the conditional mean from their full posterior to make this step even more efficient (this is a crucial point especially when considering time series with many dimensions), the use of a more complex specification of the conditional mean to compare volatility proxies and the use of the proposed model for other applications (e.g. for the analysis of spillover effect between different market indices).

Acknowledgements

I want to express my most sincere gratitude for the help received by Dr. Reza Solgi in our frequent calls and emails: his many useful insights and suggestions have been fundamental in the development of this research. I also want to thank my supervisor, Prof. Antonietta Mira, for the support she always demonstrated to me, for the advices on the research and writing processes and for the chance she gave me to stay a whole year at Università della Svizzera italiana. Finally I want to thank Prof. Conception Ausìn, all the participants to the 10th Conference on Bayesian Nonparametrics (Raleigh, NC, June 2015) and in particular Dr. Fabrizio Leisen, Dr. Antonio Canale, Dr. Bernardo Nipoti, Prof. Jim Griffin and Prof. Sonia Petrone for the many useful comments on my intermediate results.

Appendix A

Remaining graphs of the bivariate simulations

We will now include the graphs of the parameters of the model analysed in subsection 4.1.3 that we have not already included there.

Figure A.0.1: MCMC traces, posterior densities and ACF of the components of the post-processed vector $\boldsymbol{\omega}$. The green lines in the histogram represent the 95% C.I. while the red one is the true value.

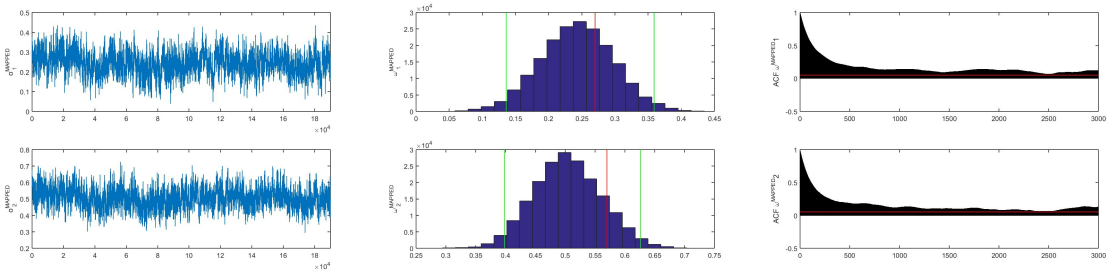


Figure A.0.2: MCMC traces, posterior densities and ACF of the components of the post-processed matrix **B**. The green lines in the histogram represent the 95% C.I. while the red one is the true value.

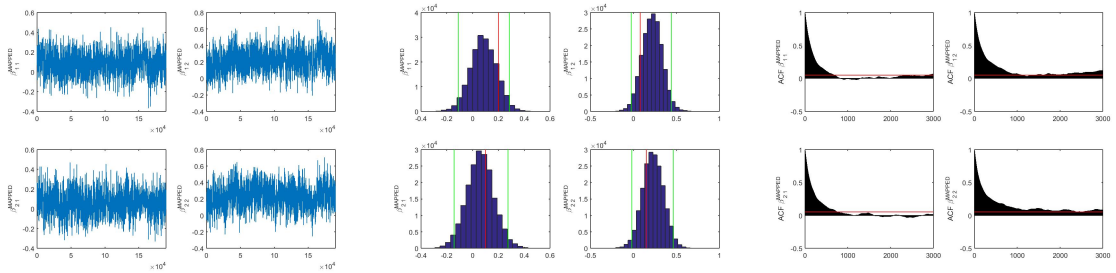


Figure A.0.3: MCMC traces, posterior densities and ACF of the components of the post-processed matrix $\mathbf{A}^{(-)}$. The green lines in the histogram represent the 95% C.I. while the red one is the true value.

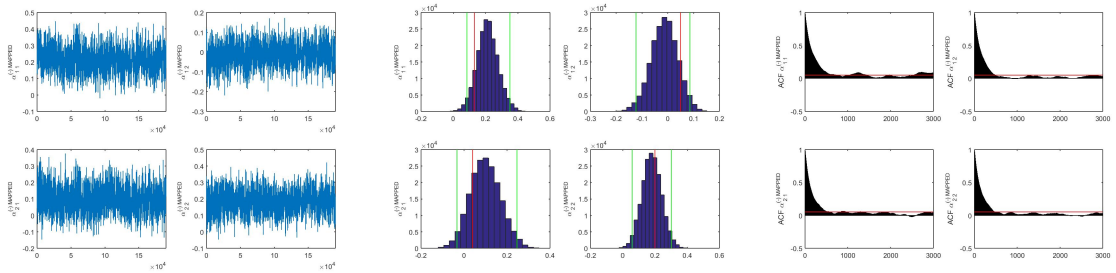


Figure A.0.4: MCMC traces, posterior densities and ACF of the components of the post-processed matrix **F**. The green lines in the histogram represent the 95% C.I. while the red one is the true value.

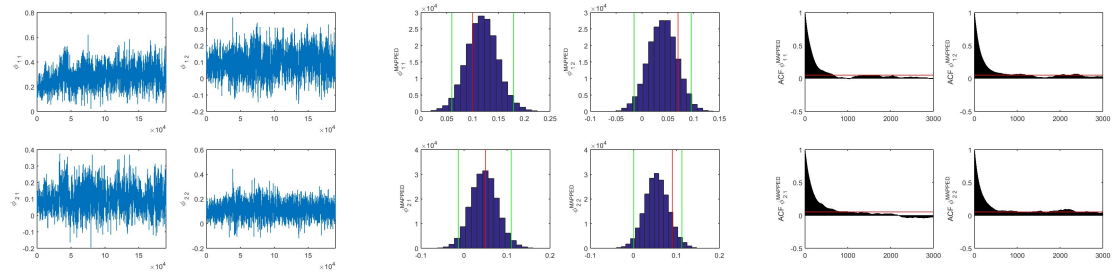


Figure A.0.5: MCMC traces, posterior densities and ACF of the components of the post-processed matrix $\mathbf{K}$. The green lines in the histogram represent the 95% C.I. while the red one is the true value.

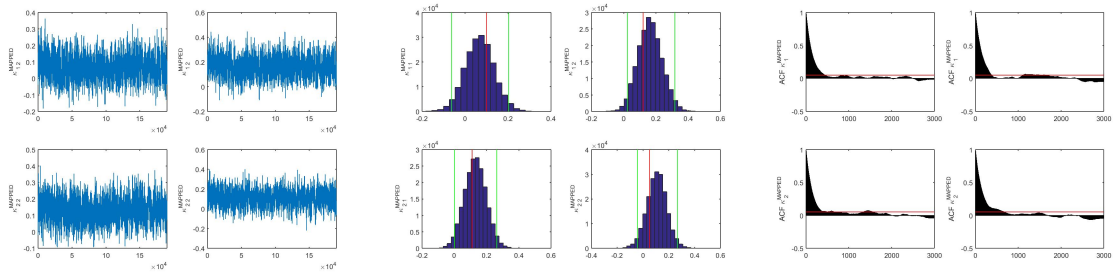


Figure A.0.6: MCMC traces, posterior densities and ACF of the components of the post-processed matrix $\mathbf{G}^{(1)}$. The green lines in the histogram represent the 95% C.I. while the red one is the true value.

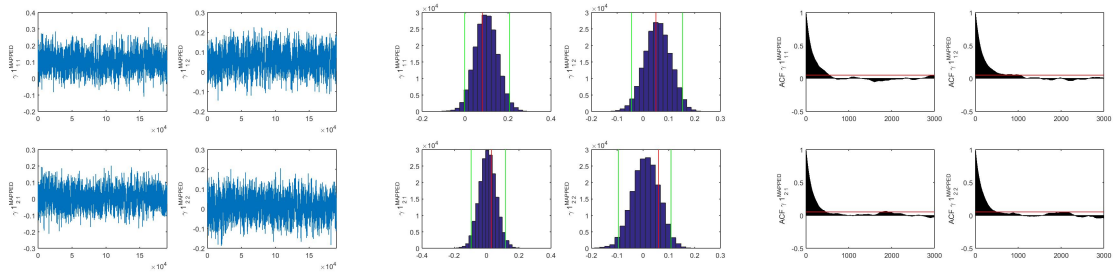


Figure A.0.7: MCMC traces, posterior densities and ACF of the components of the post-processed matrix $\mathbf{G}^{(2)}$. The green lines in the histogram represent the 95% C.I. while the red one is the true value.

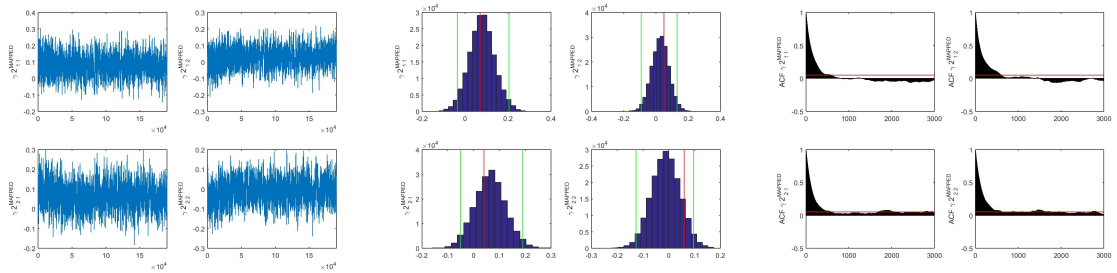


Figure A.0.8: MCMC traces, posterior densities and ACF of the components of the post-processed matrix $\mathbf{G}^{(1-)}$. The green lines in the histogram represent the 95% C.I. while the red one is the true value.

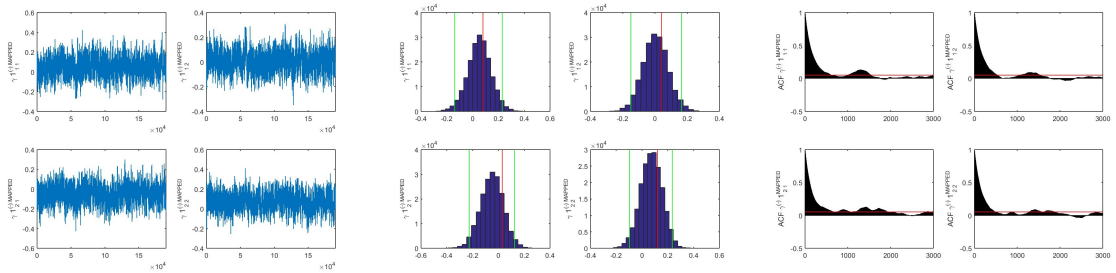
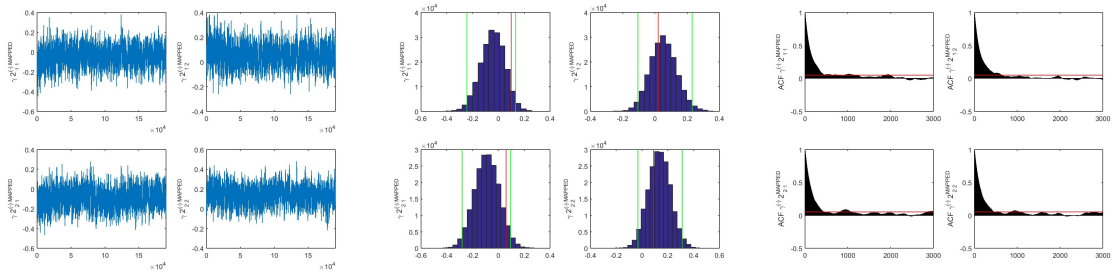


Figure A.0.9: MCMC traces, posterior densities and ACF of the components of the post-processed matrix $\mathbf{G}^{(2-)}$. The green lines in the histogram represent the 95% C.I. while the red one is the true value.



Appendix B

Remaining graphs of the empirical analyses

B.1 Specification 1:

B.1.1 S&P 500

Figure B.1.1: MCMC traces, posterior histograms, ACFs and running averages of the components of the post-processed vector $\boldsymbol{\omega}$. The green lines in the histogram represent the 95% C.I.

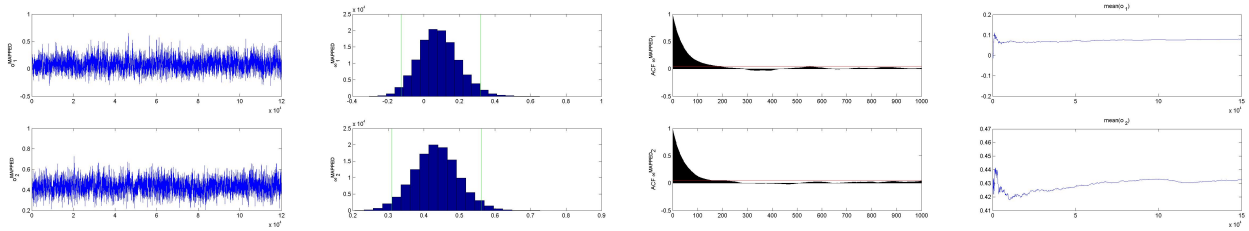


Figure B.1.2: MCMC traces, posterior histogram, ACFs and running averages of the components of the post-processed vector $\boldsymbol{\beta}$. The green lines in the histogram represent the 95% C.I.

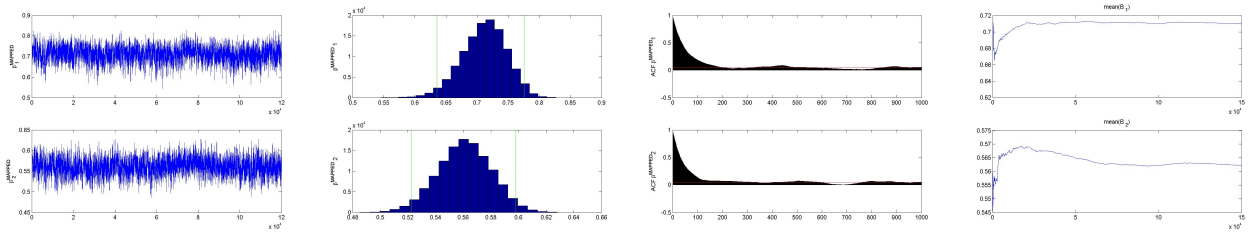


Figure B.1.3: MCMC traces, posterior histograms, ACFs and running averages of the components of the post-processed matrix **A**. The green lines in the histogram represent the 95% C.I.

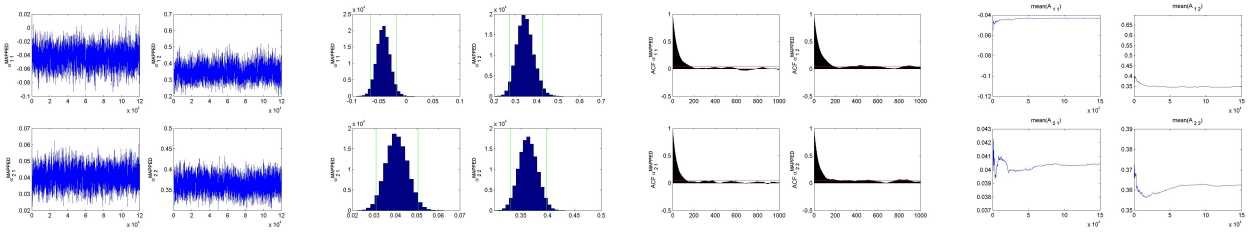


Figure B.1.4: The upper plot shows the traces of total number of components and of the number of active components at each step. The lower plot shows the corresponding running averages.

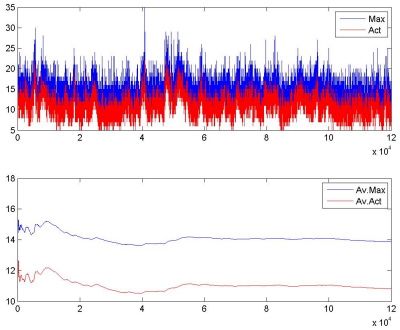
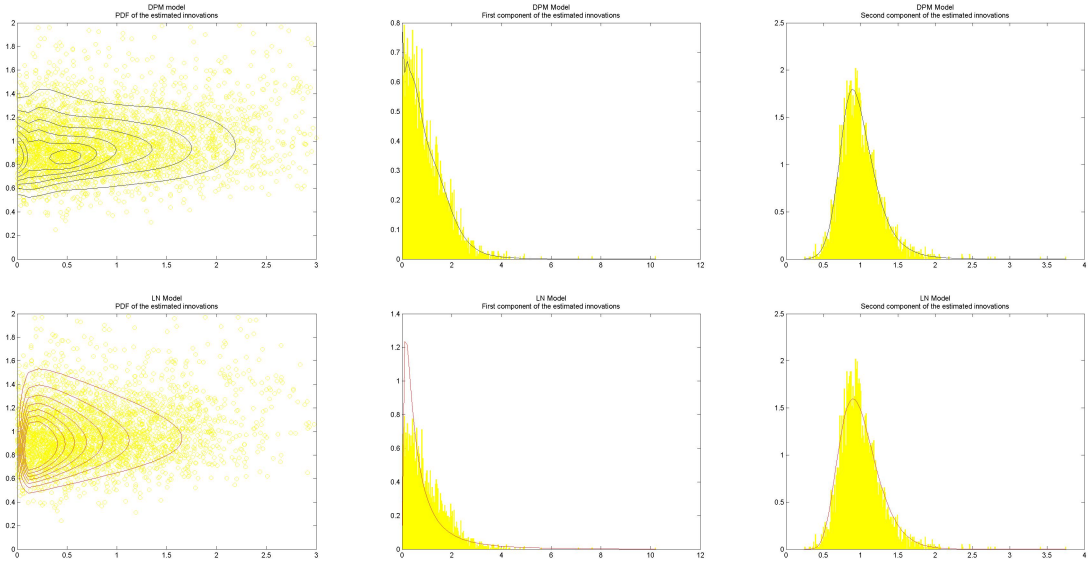


Figure B.1.5: Estimated joint and marginal densities of the innovations over the estimated innovations. On the upper row there are the results obtained with the DPMLN2-vMEM, on the lower row the ones obtained with LN1-vMEM



B.1.2 DJIA

Figure B.1.6: MCMC traces, posterior histograms, ACFs and running averages of the components of the post-processed vector $\boldsymbol{\omega}$. The green lines in the histogram represent the 95% C.I.

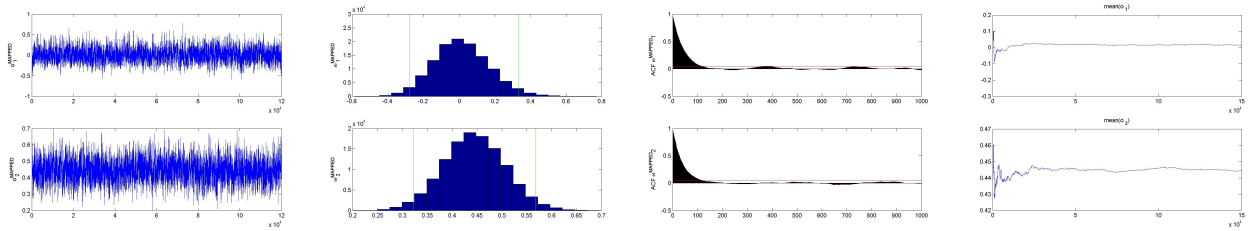


Figure B.1.7: MCMC traces, posterior histogram, ACFs and running averages of the components of the post-processed vector $\boldsymbol{\beta}$. The green lines in the histogram represent the 95% C.I.

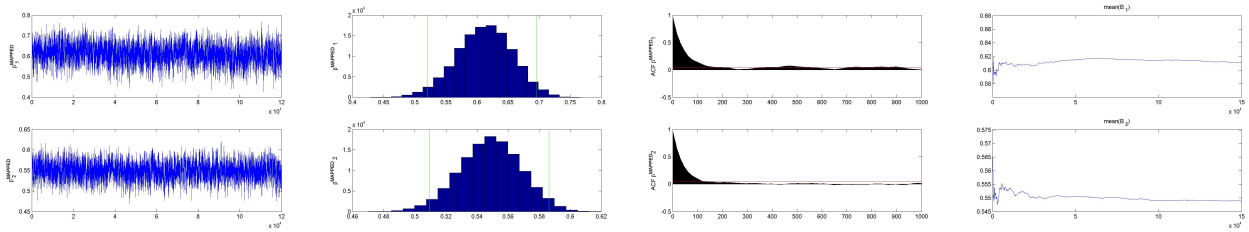


Figure B.1.8: MCMC traces, posterior histograms, ACFs and running averages of the components of the post-processed matrix $\mathbf{A}$. The green lines in the histogram represent the 95% C.I.

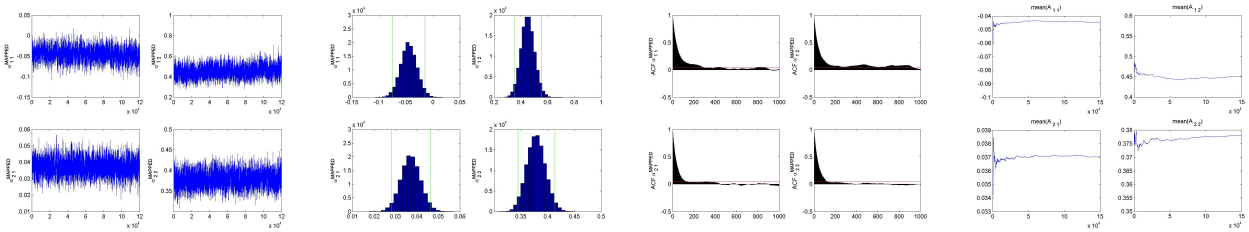


Figure B.1.9: The upper plot shows the traces of total number of components and of the number of active components at each step. The lower plot shows the corresponding running averages.

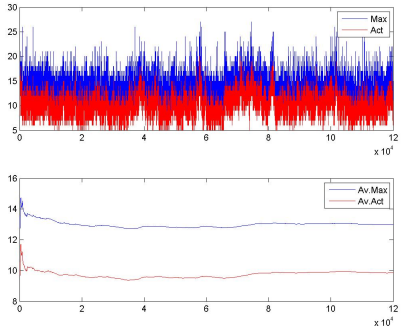
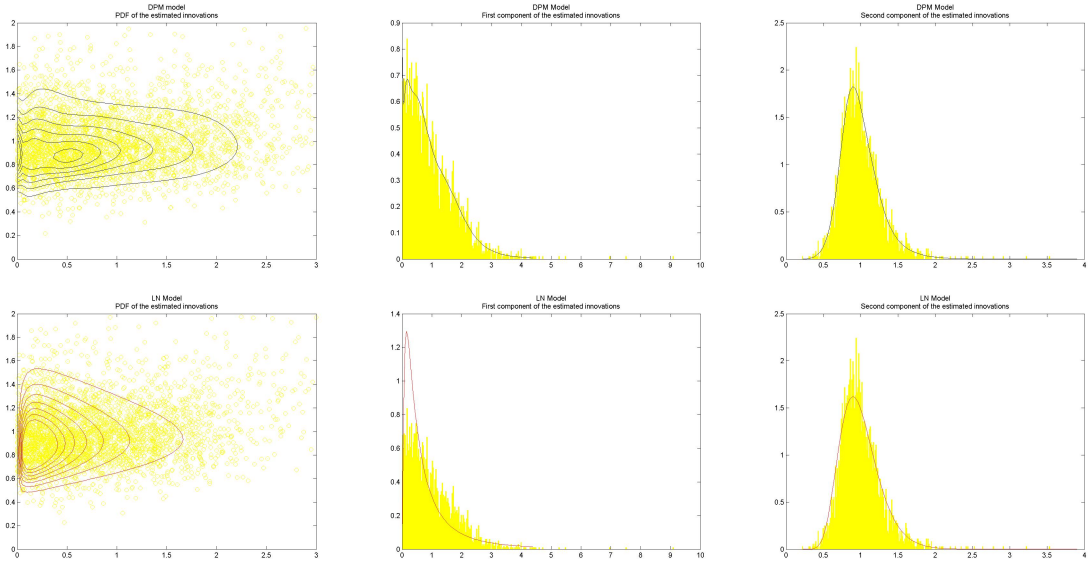


Figure B.1.10: Estimated joint and marginal densities of the innovations over the estimated innovations. On the upper row there are the results obtained with the DPMLN2-vMEM, on the lower row the ones obtained with LN1-vMEM



B.1.3 FTSE 100

Figure B.1.11: MCMC traces, posterior histograms, ACFs and running averages of the components of the post-processed vector $\boldsymbol{\omega}$. The green lines in the histogram represent the 95% C.I.

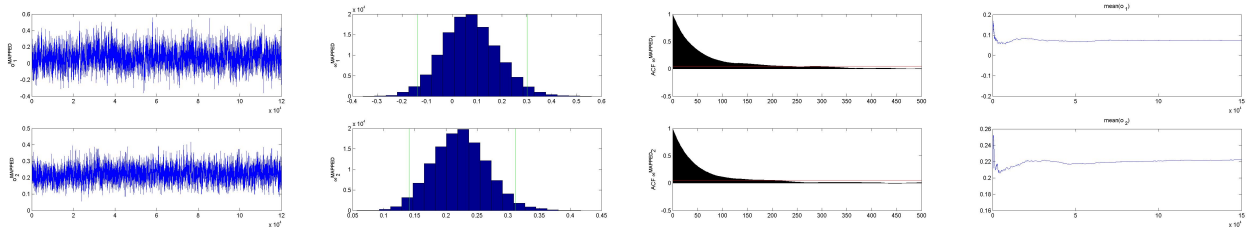


Figure B.1.12: MCMC traces, posterior histogram, ACFs and running averages of the components of the post-processed vector $\boldsymbol{\beta}$. The green lines in the histogram represent the 95% C.I.

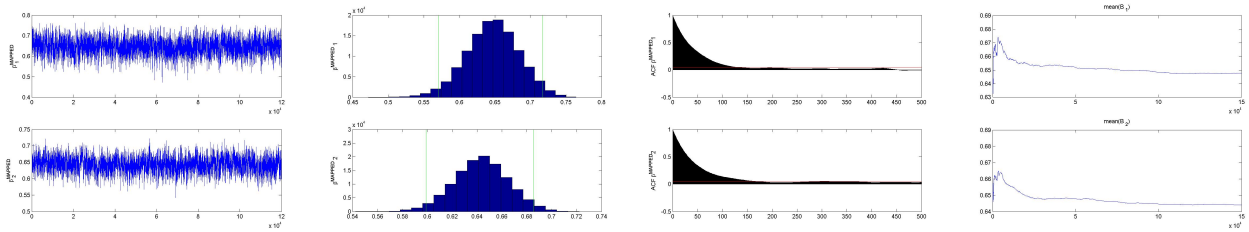


Figure B.1.13: MCMC traces, posterior histograms, ACFs and running averages of the components of the post-processed matrix $\mathbf{A}$. The green lines in the histogram represent the 95% C.I.

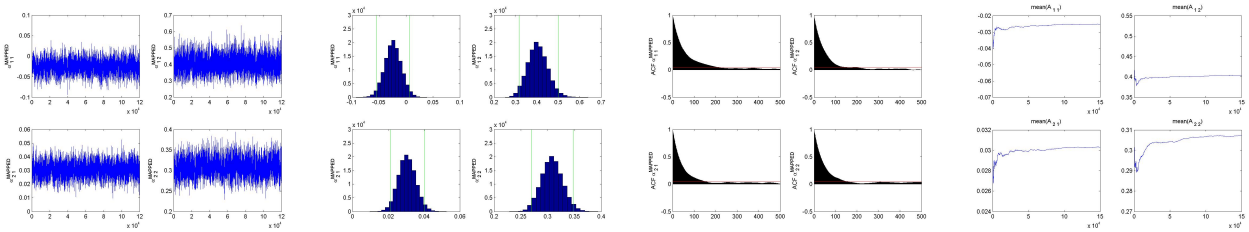


Figure B.1.14: The upper plot shows the traces of total number of components and of the number of active components at each step. The lower plot shows the corresponding running averages.

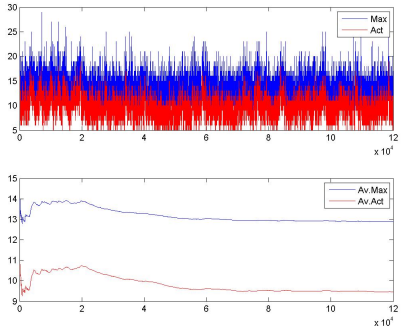
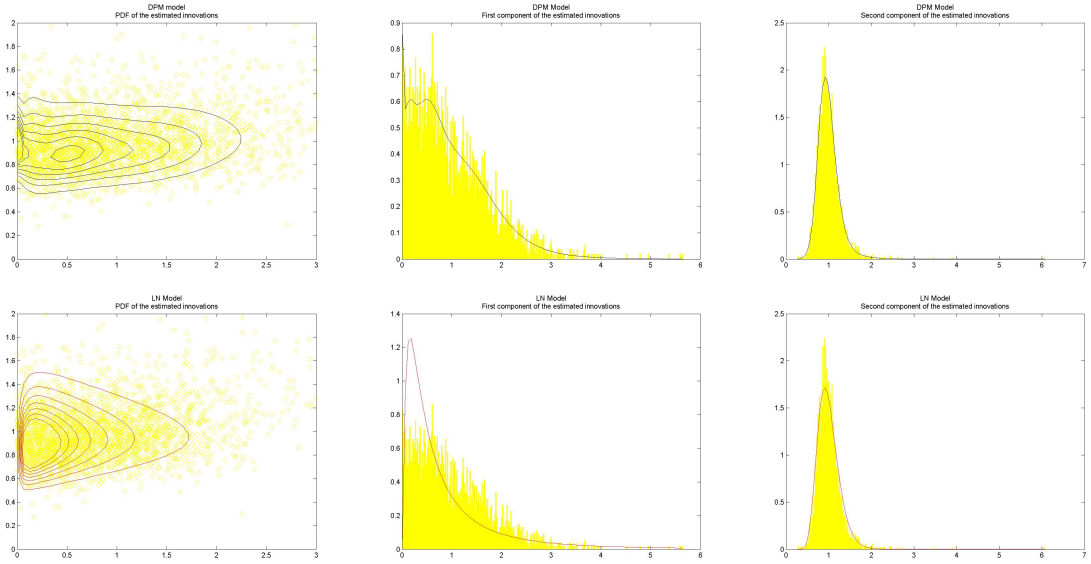


Figure B.1.15: Estimated joint and marginal densities of the innovations over the estimated innovations. On the upper row there are the results obtained with the DPMLN2-vMEM, on the lower row the ones obtained with LN1-vMEM



B.2 Specification 2:

B.2.1 S&P 500

Figure B.2.1: MCMC traces, posterior histograms, ACFs and running averages of the components of the post-processed vector $\boldsymbol{\omega}$. The green lines in the histogram represent the 95% C.I.

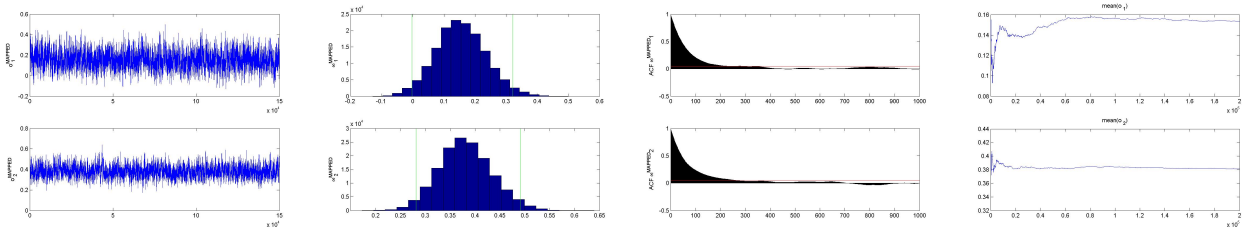


Figure B.2.2: MCMC traces, posterior histogram, ACFs and running averages of the components of the post-processed vector $\boldsymbol{\beta}$. The green lines in the histogram represent the 95% C.I.

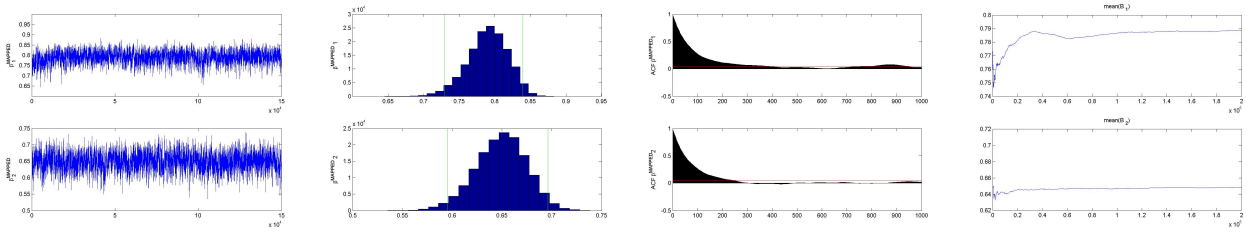


Figure B.2.3: MCMC traces, posterior histograms, ACFs and running averages of the components of the post-processed matrix $\mathbf{A}$. The green lines in the histogram represent the 95% C.I.

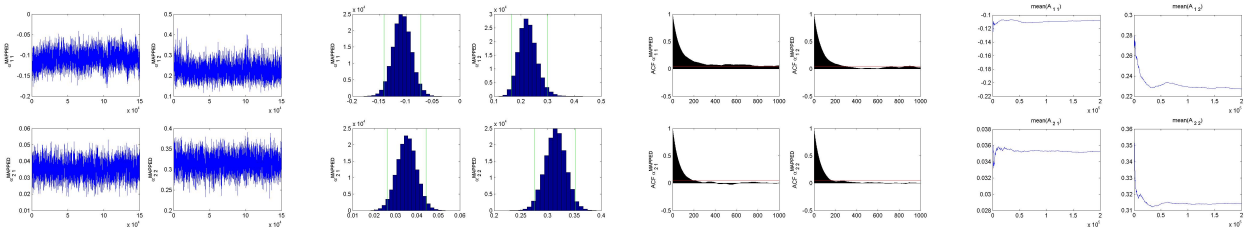


Figure B.2.4: MCMC traces, posterior histograms, ACFs and running averages of the components of the post-processed matrix $\mathbf{A}^{(-)}$. The green lines in the histogram represent the 95% C.I.

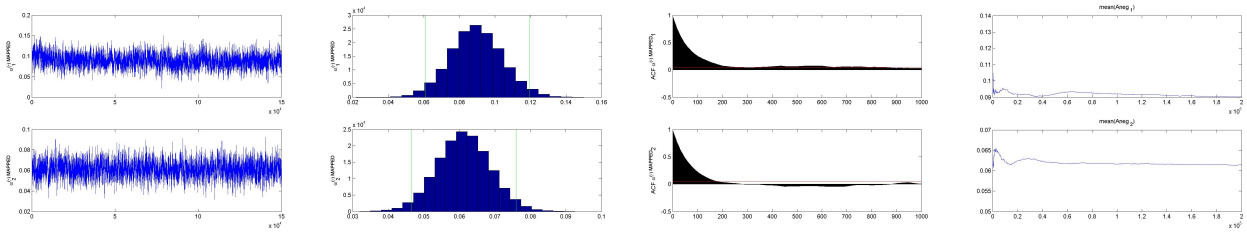


Figure B.2.5: MCMC traces, posterior histograms, ACFs and running averages of the components of the post-processed matrix $\mathbf{F}$. The green lines in the histogram represent the 95% C.I.

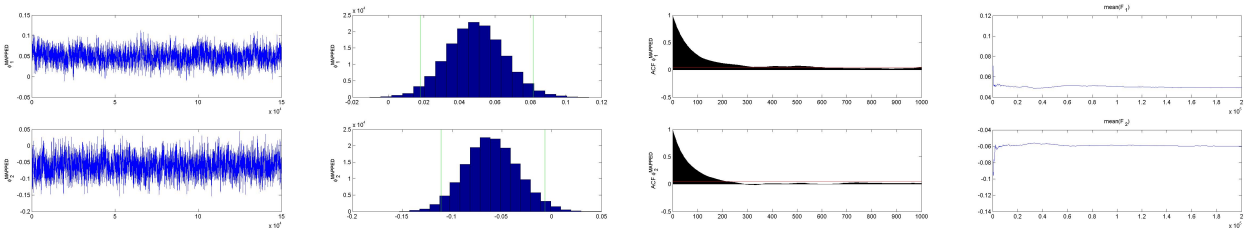


Figure B.2.6: The upper plot shows the traces of total number of components and of the number of active components at each step. The lower plot shows the corresponding running averages.

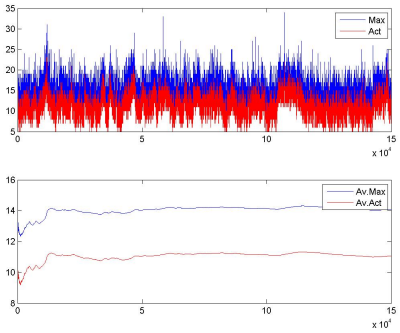
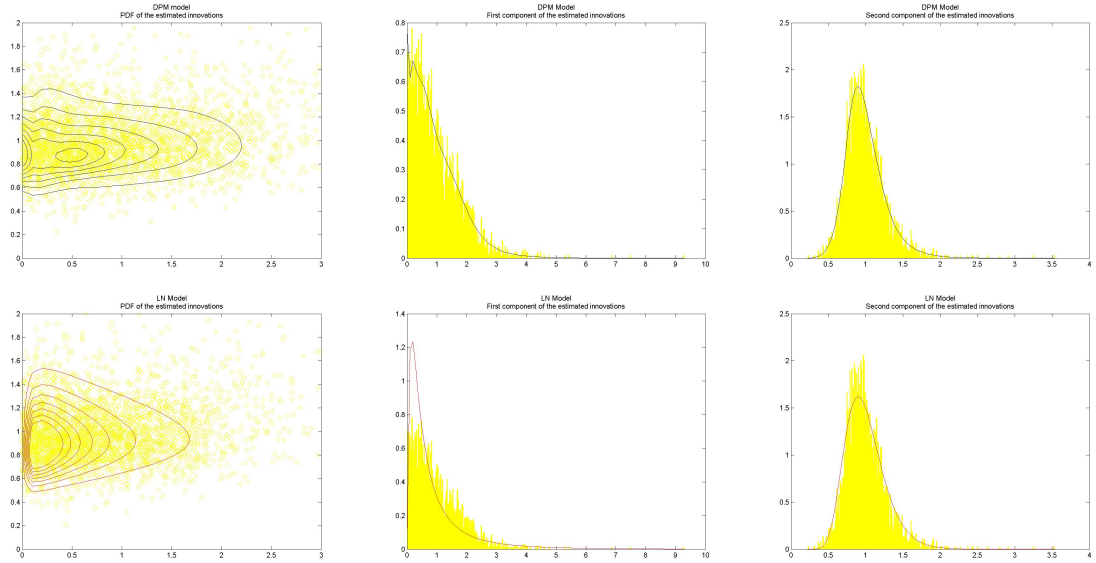


Figure B.2.7: Estimated joint and marginal densities of the innovations over the estimated innovations. On the upper row there are the results obtained with the DPMLN2-vMEM, on the lower row the ones obtained with LN1-vMEM



B.2.2 DJIA

Figure B.2.8: MCMC traces, posterior histograms, ACFs and running averages of the components of the post-processed vector $\boldsymbol{\omega}$. The green lines in the histogram represent the 95% C.I.

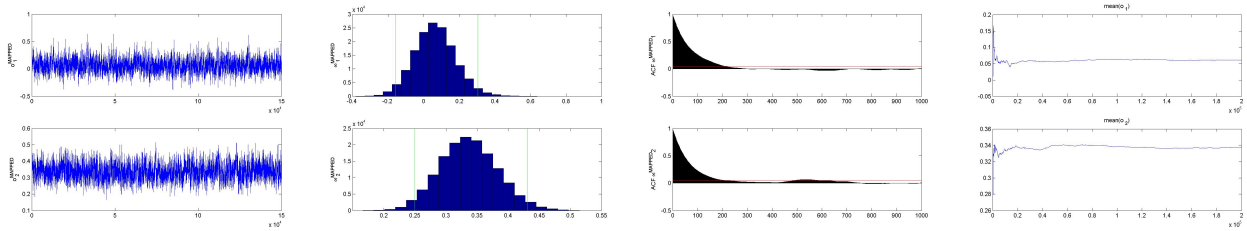


Figure B.2.9: MCMC traces, posterior histogram, ACFs and running averages of the components of the post-processed vector $\boldsymbol{\beta}$. The green lines in the histogram represent the 95% C.I.

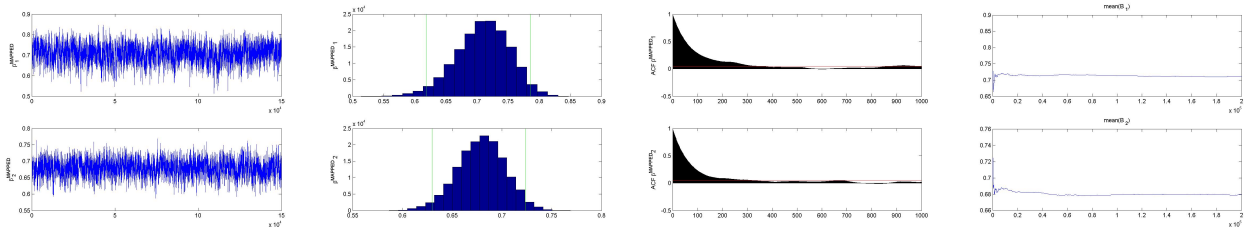


Figure B.2.10: MCMC traces, posterior histograms, ACFs and running averages of the components of the post-processed matrix $\mathbf{A}$. The green lines in the histogram represent the 95% C.I.

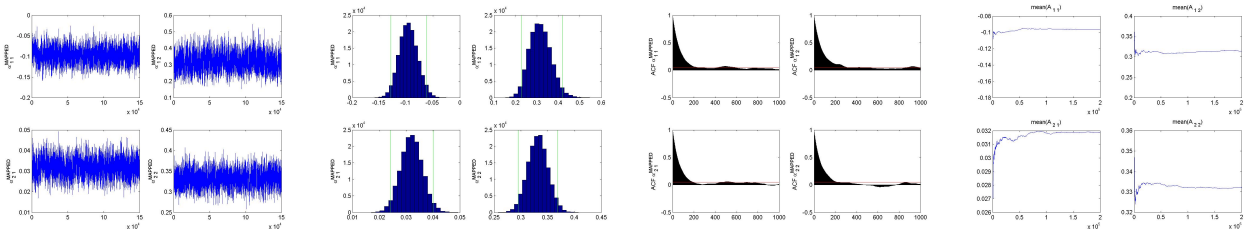


Figure B.2.11: MCMC traces, posterior histograms, ACFs and running averages of the components of the post-processed matrix $\mathbf{A}^{(-)}$. The green lines in the histogram represent the 95% C.I.

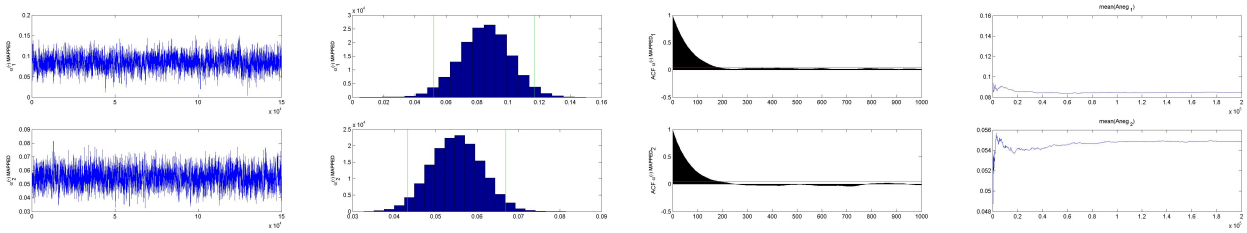


Figure B.2.12: MCMC traces, posterior histograms, ACFs and running averages of the components of the post-processed matrix $\mathbf{F}$. The green lines in the histogram represent the 95% C.I.

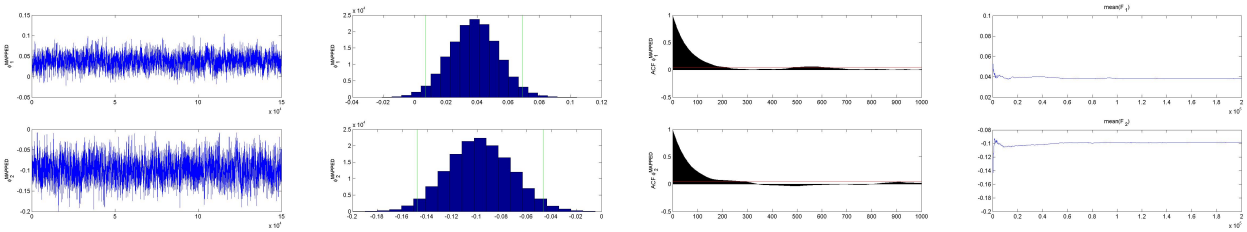


Figure B.2.13: The upper plot shows the traces of total number of components and of the number of active components at each step. The lower plot shows the corresponding running averages.

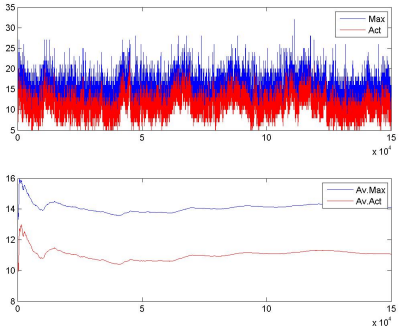
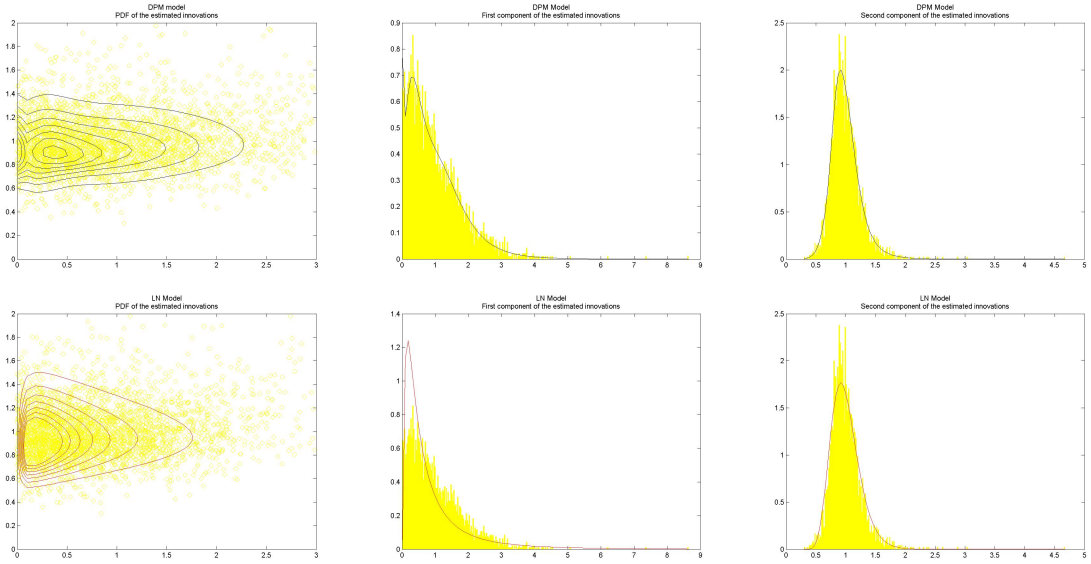


Figure B.2.14: Estimated joint and marginal densities of the innovations over the estimated innovations. On the upper row there are the results obtained with the DPMLN2-vMEM, on the lower row the ones obtained with LN1-vMEM



B.2.3 FTSE 100

Figure B.2.15: MCMC traces, posterior histograms, ACFs and running averages of the components of the post-processed vector $\boldsymbol{\omega}$. The green lines in the histogram represent the 95% C.I.

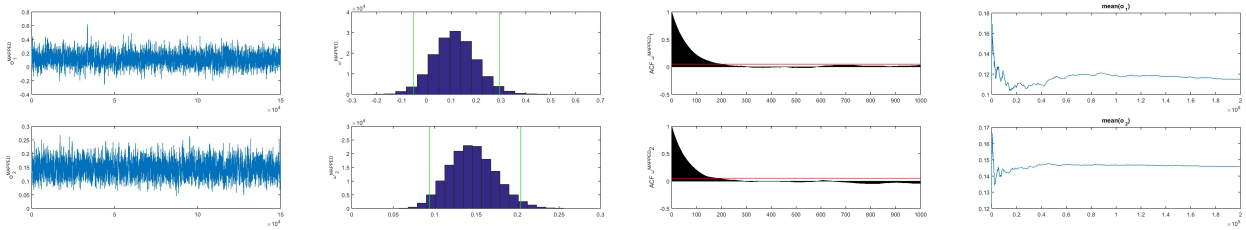


Figure B.2.16: MCMC traces, posterior histogram, ACFs and running averages of the components of the post-processed vector $\boldsymbol{\beta}$. The green lines in the histogram represent the 95% C.I.

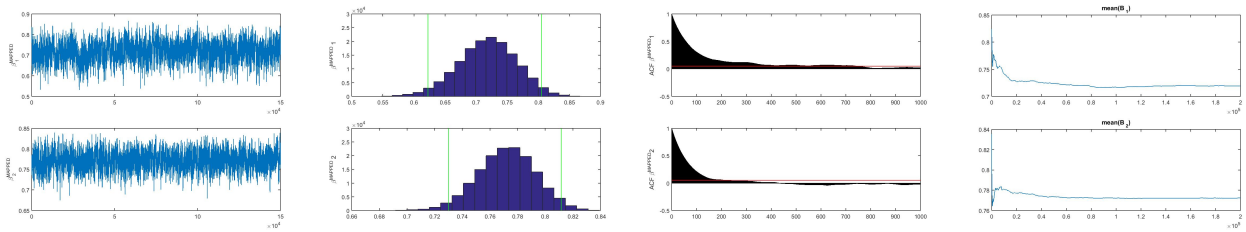


Figure B.2.17: MCMC traces, posterior histograms, ACFs and running averages of the components of the post-processed matrix $\mathbf{A}$. The green lines in the histogram represent the 95% C.I.

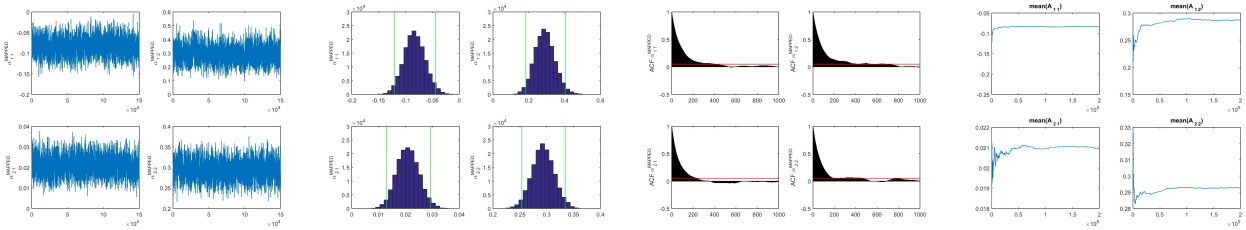


Figure B.2.18: MCMC traces, posterior histograms, ACFs and running averages of the components of the post-processed matrix $\mathbf{A}^{(-)}$. The green lines in the histogram represent the 95% C.I.

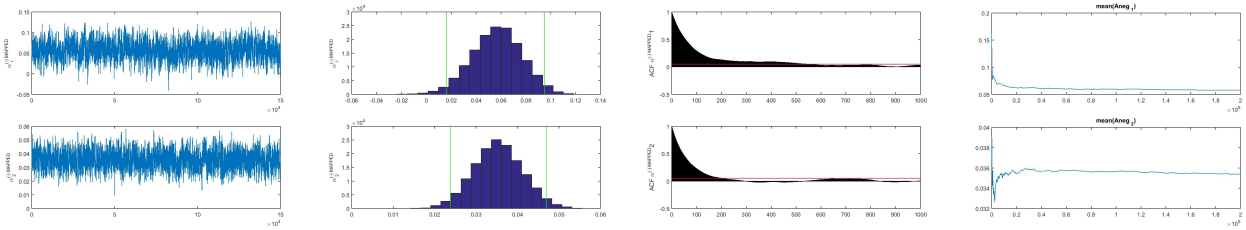


Figure B.2.19: MCMC traces, posterior histograms, ACFs and running averages of the components of the post-processed matrix $\mathbf{F}$. The green lines in the histogram represent the 95% C.I.

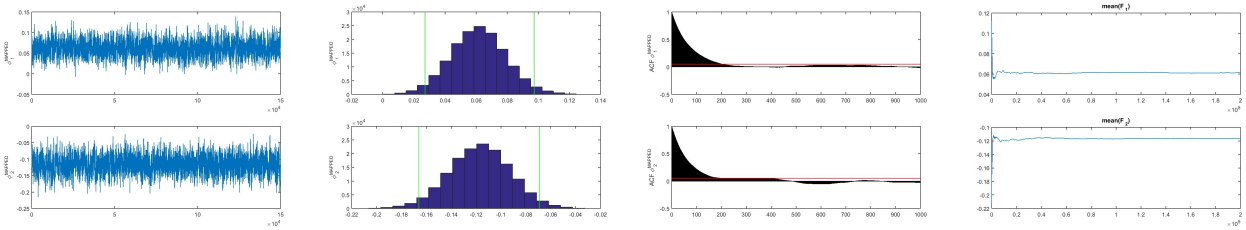


Figure B.2.20: The upper plot shows the traces of total number of components and of the number of active components at each step. The lower plot shows the corresponding running averages.

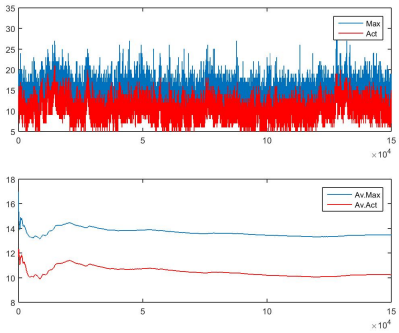
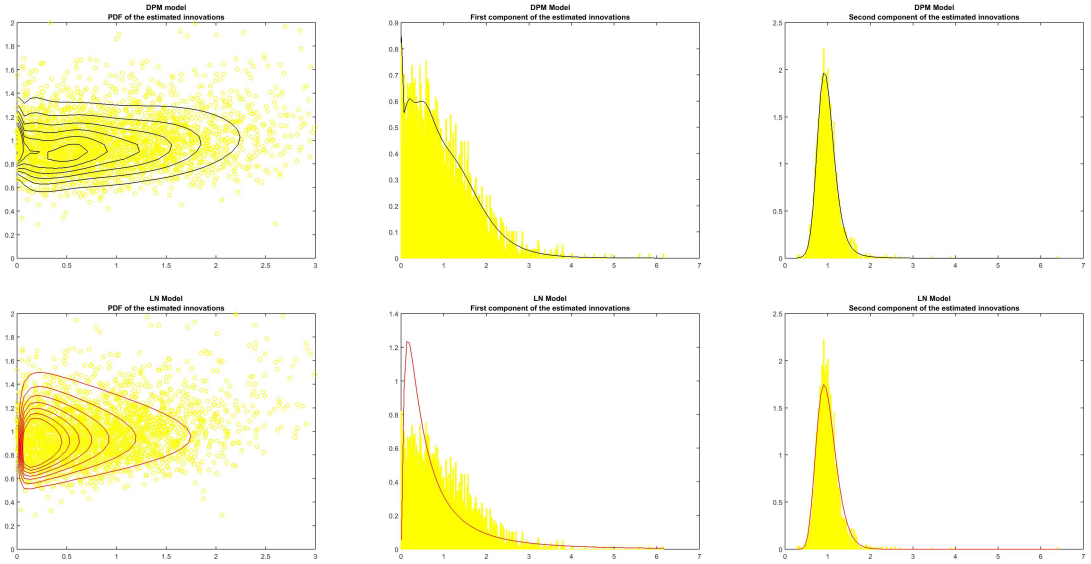


Figure B.2.21: Estimated joint and marginal densities of the innovations over the estimated innovations. On the upper row there are the results obtained with the DPMLN2-vMEM, on the lower row the ones obtained with LN1-vMEM



B.3 Specification 3:

B.3.1 S&P 500

Figure B.3.1: MCMC traces, posterior histograms, ACFs and running averages of the components of the post-processed vector $\boldsymbol{\omega}$. The green lines in the histogram represent the 95% C.I.

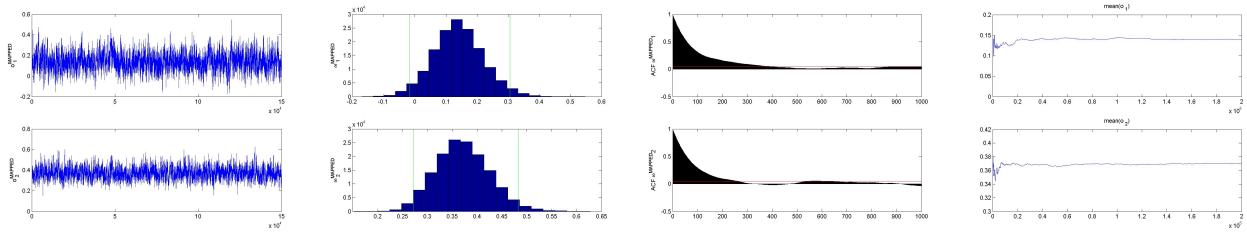


Figure B.3.2: MCMC traces, posterior histogram, ACFs and running averages of the components of the post-processed vector $\boldsymbol{\beta}$. The green lines in the histogram represent the 95% C.I.

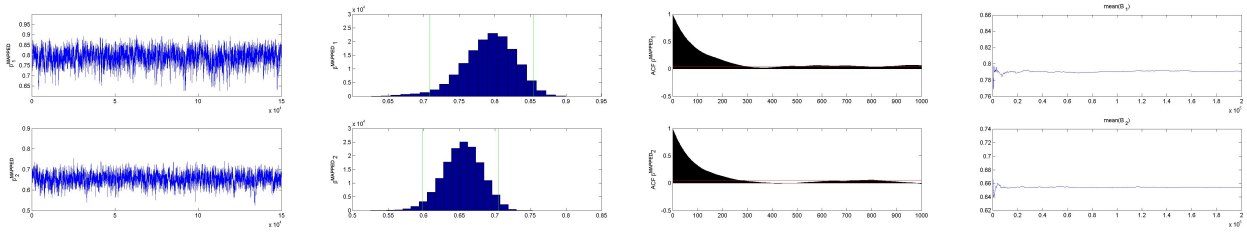


Figure B.3.3: MCMC traces, posterior histograms, ACFs and running averages of the components of the post-processed matrix $\mathbf{A}$. The green lines in the histogram represent the 95% C.I.

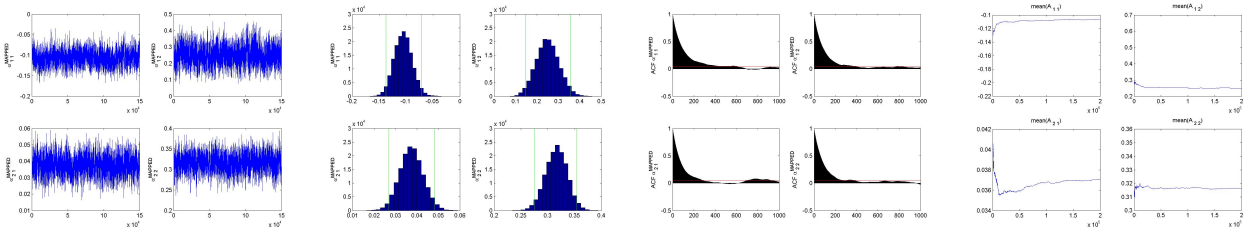


Figure B.3.4: MCMC traces, posterior histograms, ACFs and running averages of the components of the post-processed matrix $\mathbf{A}^{(-)}$. The green lines in the histogram represent the 95% C.I.

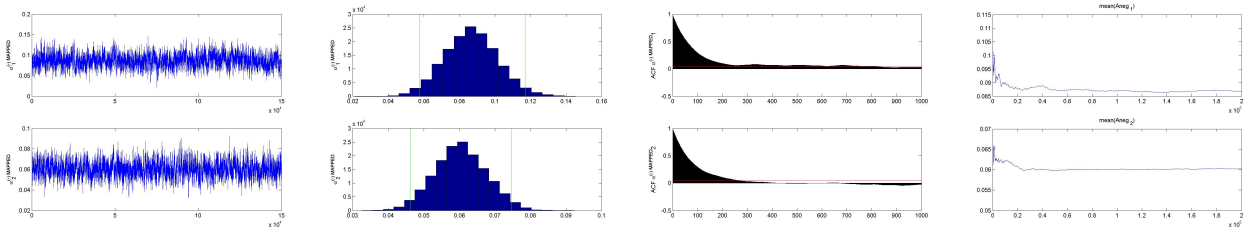


Figure B.3.5: MCMC traces, posterior histograms, ACFs and running averages of the components of the post-processed matrix $\mathbf{F}$. The green lines in the histogram represent the 95% C.I.

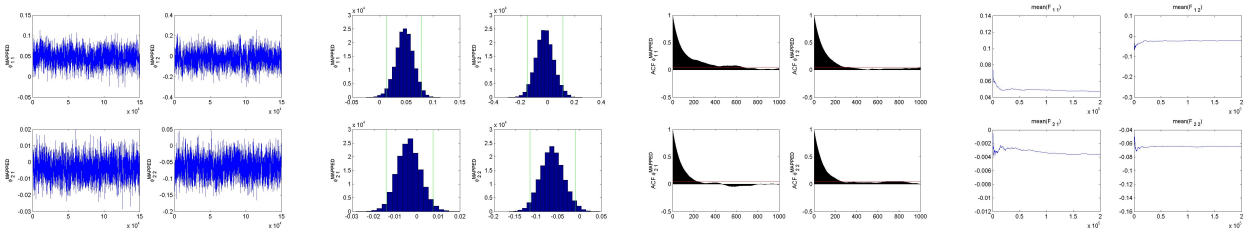


Figure B.3.6: The upper plot shows the traces of total number of components and of the number of active components at each step. The lower plot shows the corresponding running averages.

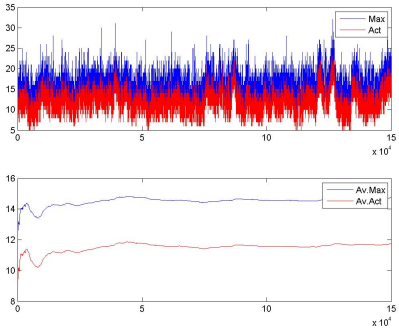
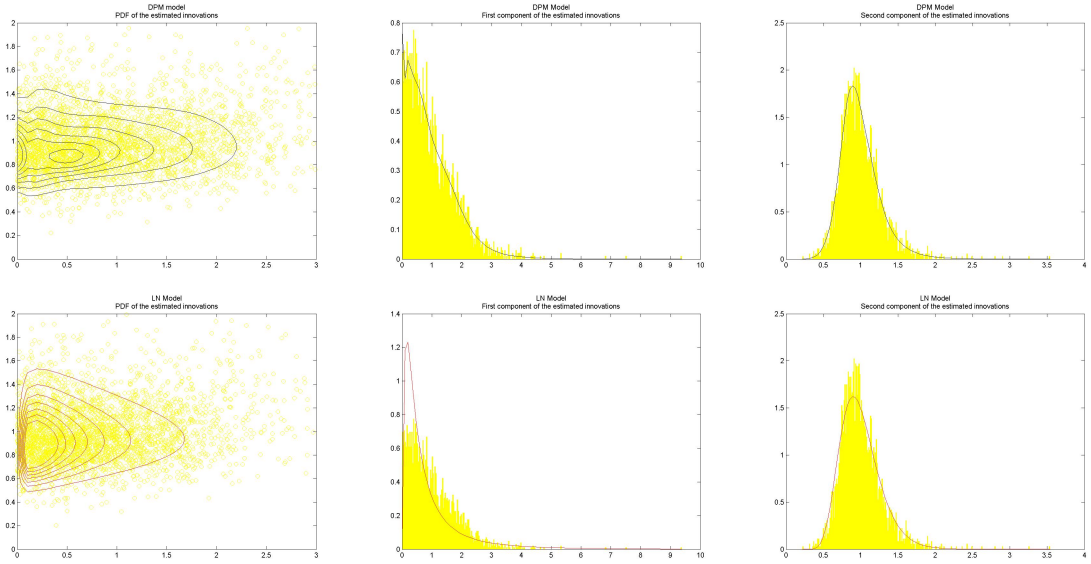


Figure B.3.7: Estimated joint and marginal densities of the innovations over the estimated innovations. On the upper row there are the results obtained with the DPMLN2-vMEM, on the lower row the ones obtained with LN1-vMEM



B.3.2 DJIA

Figure B.3.8: MCMC traces, posterior histograms, ACFs and running averages of the components of the post-processed vector $\boldsymbol{\omega}$. The green lines in the histogram represent the 95% C.I.

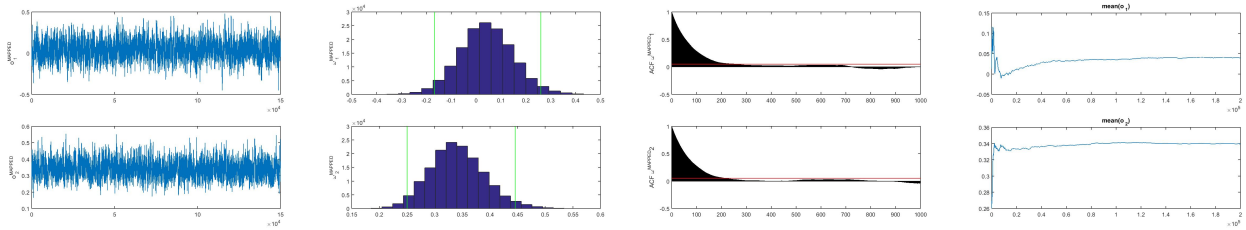


Figure B.3.9: MCMC traces, posterior histogram, ACFs and running averages of the components of the post-processed vector $\boldsymbol{\beta}$. The green lines in the histogram represent the 95% C.I.

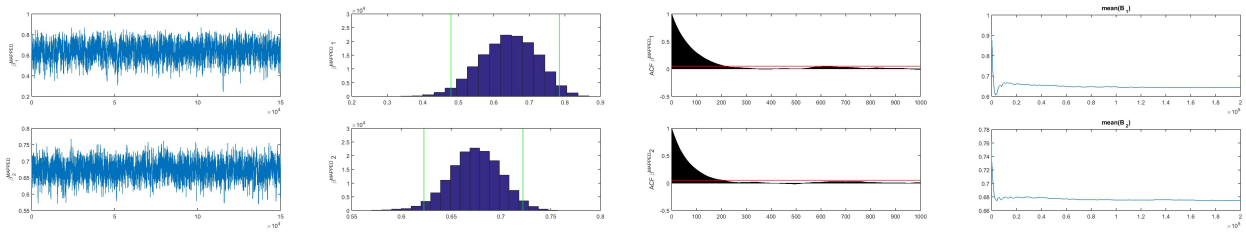


Figure B.3.10: MCMC traces, posterior histograms, ACFs and running averages of the components of the post-processed matrix $\mathbf{A}$. The green lines in the histogram represent the 95% C.I.

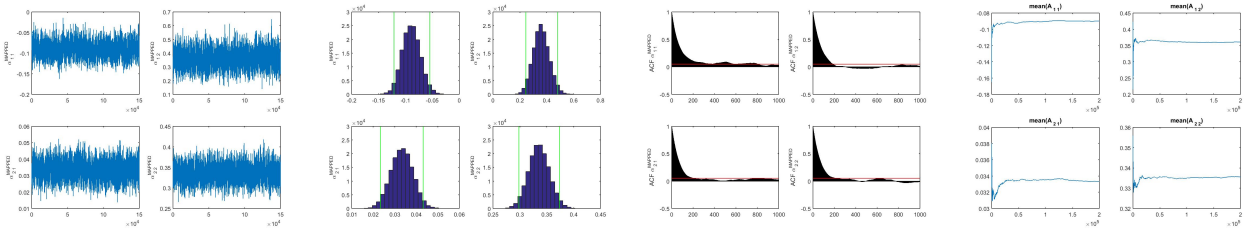


Figure B.3.11: MCMC traces, posterior histograms, ACFs and running averages of the components of the post-processed matrix $\mathbf{A}^{(-)}$. The green lines in the histogram represent the 95% C.I.

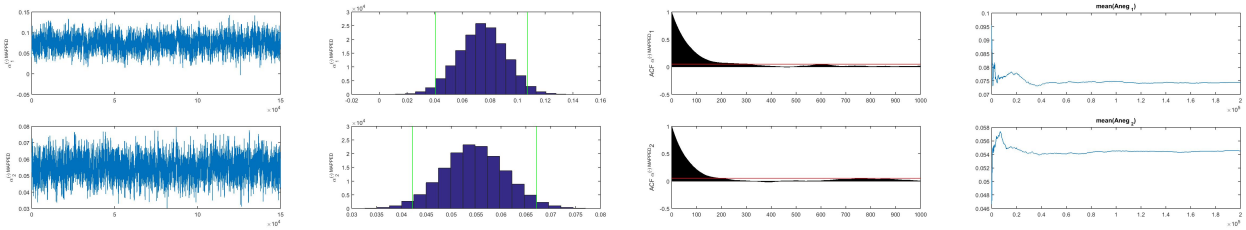


Figure B.3.12: MCMC traces, posterior histograms, ACFs and running averages of the components of the post-processed matrix $\mathbf{F}$. The green lines in the histogram represent the 95% C.I.

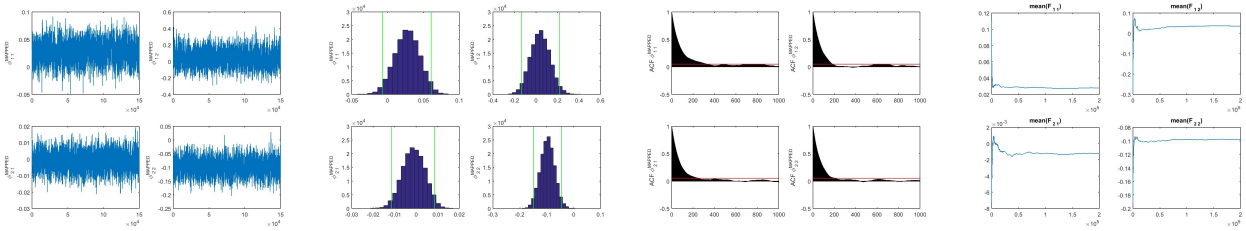


Figure B.3.13: The upper plot shows the traces of total number of components and of the number of active components at each step. The lower plot shows the corresponding running averages.

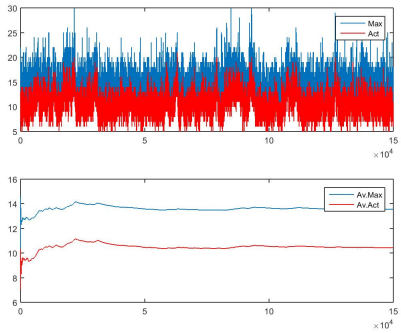
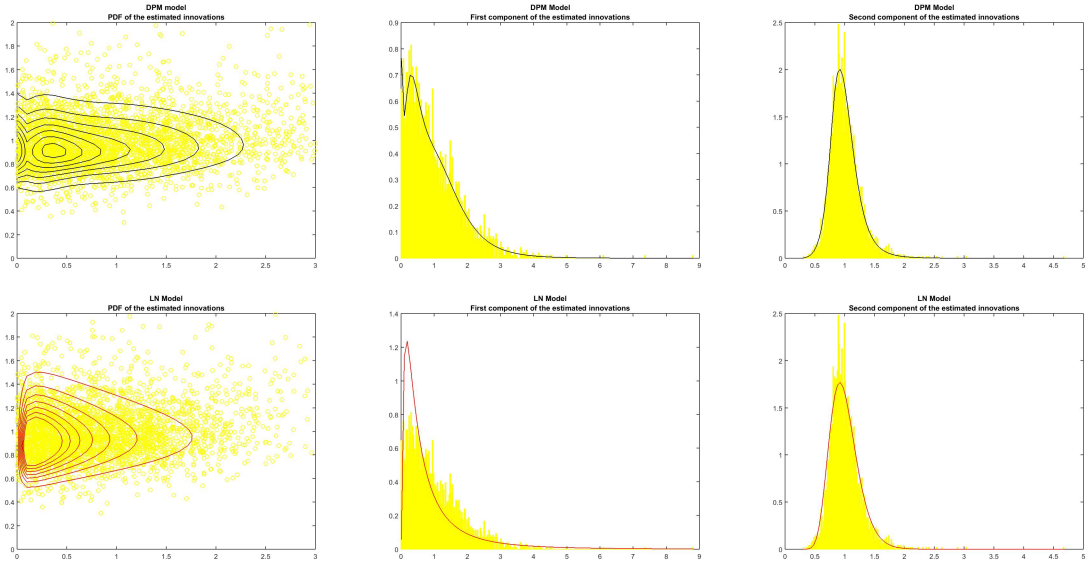


Figure B.3.14: Estimated joint and marginal densities of the innovations over the estimated innovations. On the upper row there are the results obtained with the DPMLN2-vMEM, on the lower row the ones obtained with LN1-vMEM



B.3.3 FTSE 100

Figure B.3.15: MCMC traces, posterior histograms, ACFs and running averages of the components of the post-processed vector $\boldsymbol{\omega}$. The green lines in the histogram represent the 95% C.I.

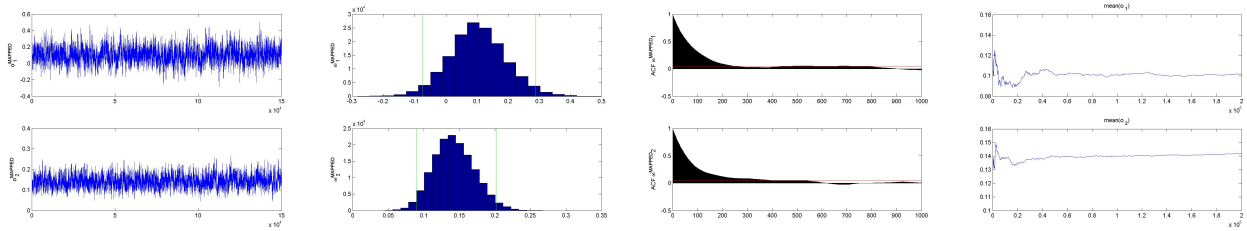


Figure B.3.16: MCMC traces, posterior histogram, ACFs and running averages of the components of the post-processed vector $\boldsymbol{\beta}$. The green lines in the histogram represent the 95% C.I.

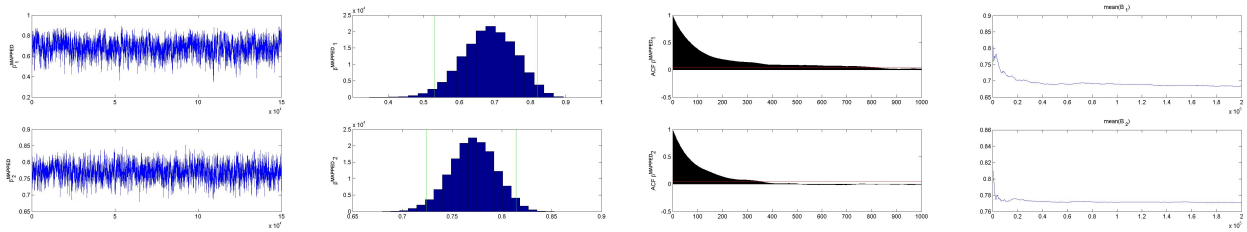


Figure B.3.17: MCMC traces, posterior histograms, ACFs and running averages of the components of the post-processed matrix $\mathbf{A}$. The green lines in the histogram represent the 95% C.I.

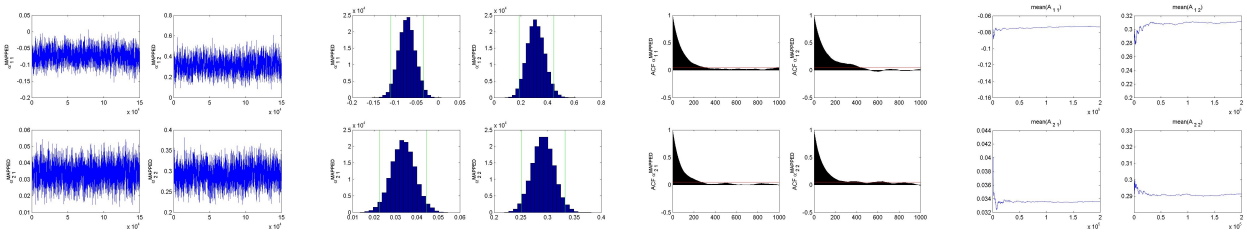


Figure B.3.18: MCMC traces, posterior histograms, ACFs and running averages of the components of the post-processed matrix $\mathbf{A}^{(-)}$. The green lines in the histogram represent the 95% C.I.

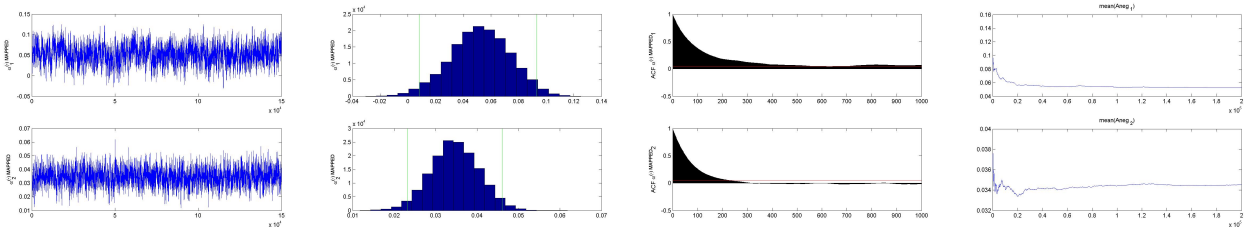


Figure B.3.19: MCMC traces, posterior histograms, ACFs and running averages of the components of the post-processed matrix $\mathbf{F}$. The green lines in the histogram represent the 95% C.I.

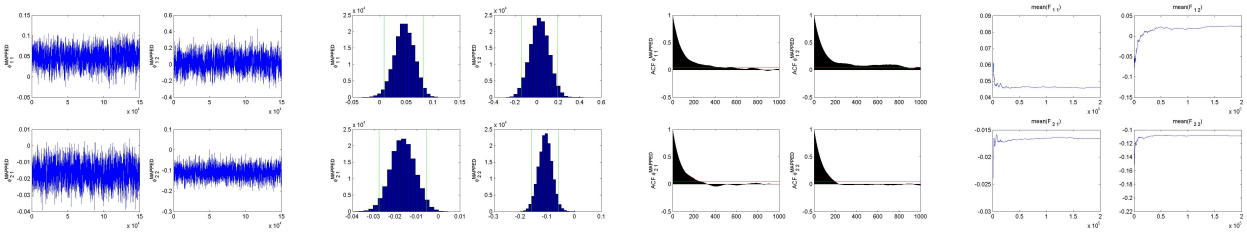


Figure B.3.20: The upper plot shows the traces of total number of components and of the number of active components at each step. The lower plot shows the corresponding running averages.

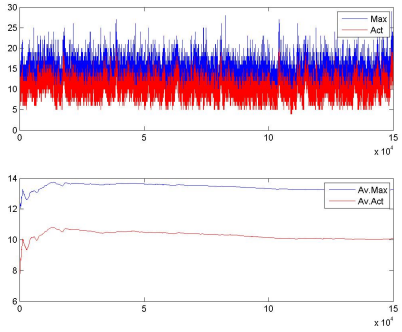
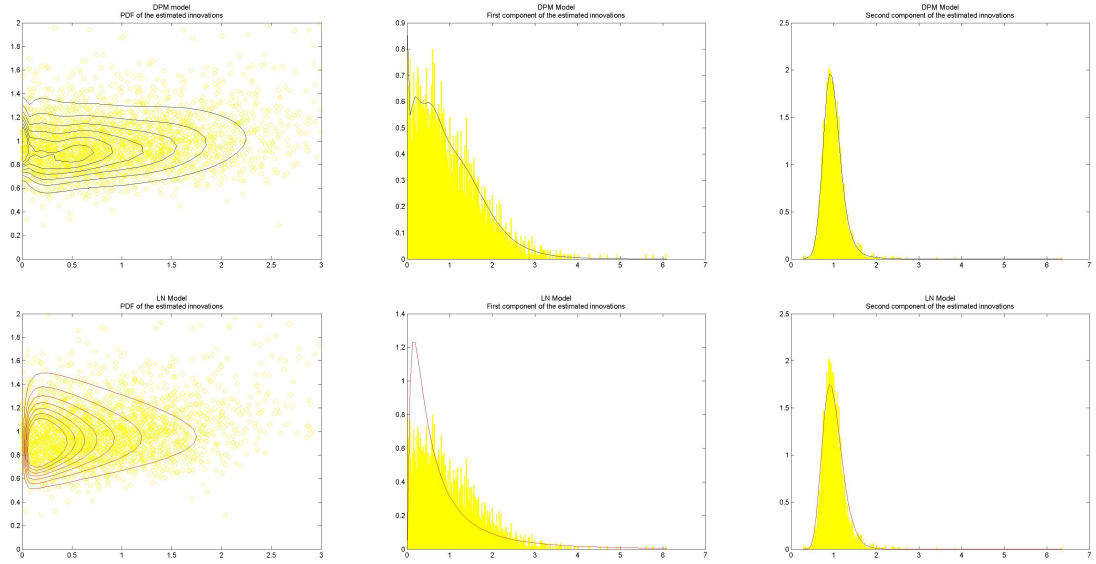


Figure B.3.21: Estimated joint and marginal densities of the innovations over the estimated innovations. On the upper row there are the results obtained with the DPMLN2-vMEM, on the lower row the ones obtained with LN1-vMEM



B.4 Specification 4:

B.4.1 S&P 500

Figure B.4.1: MCMC traces, posterior histograms, ACFs and running averages of the components of the post-processed vector $\boldsymbol{\omega}$. The green lines in the histogram represent the 95% C.I.

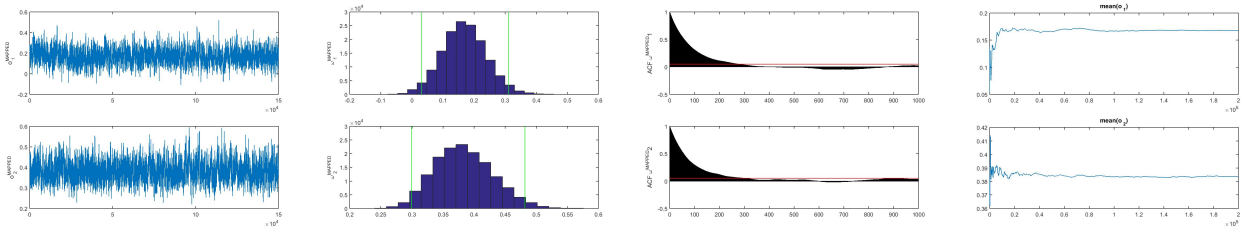


Figure B.4.2: MCMC traces, posterior histogram, ACFs and running averages of the components of the post-processed vector $\boldsymbol{\beta}$. The green lines in the histogram represent the 95% C.I.

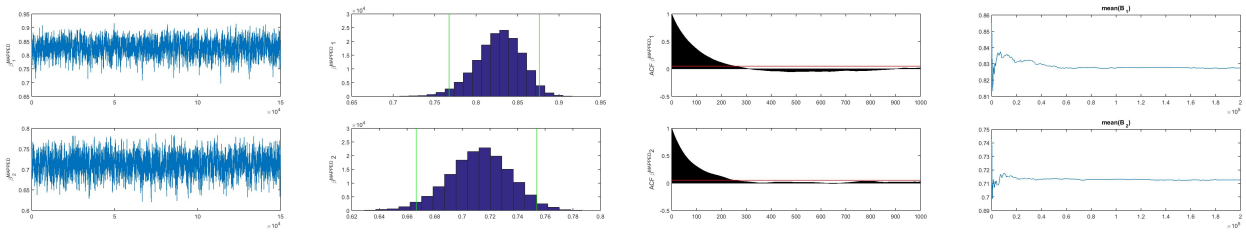


Figure B.4.3: MCMC traces, posterior histograms, ACFs and running averages of the components of the post-processed matrix $\mathbf{A}$. The green lines in the histogram represent the 95% C.I.

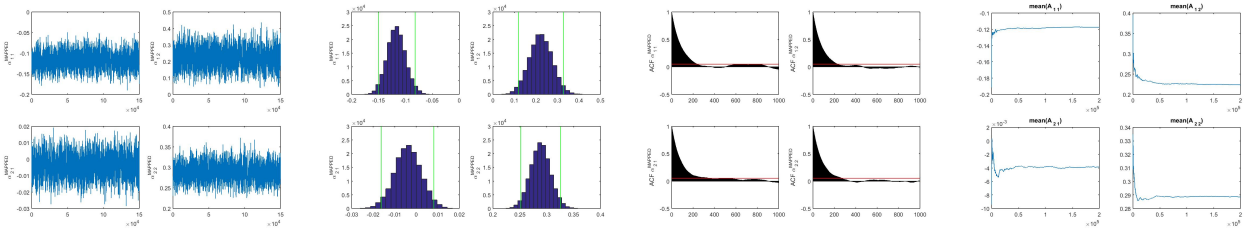


Figure B.4.4: MCMC traces, posterior histograms, ACFs and running averages of the components of the post-processed matrix $\mathbf{A}^{(-)}$. The green lines in the histogram represent the 95% C.I.

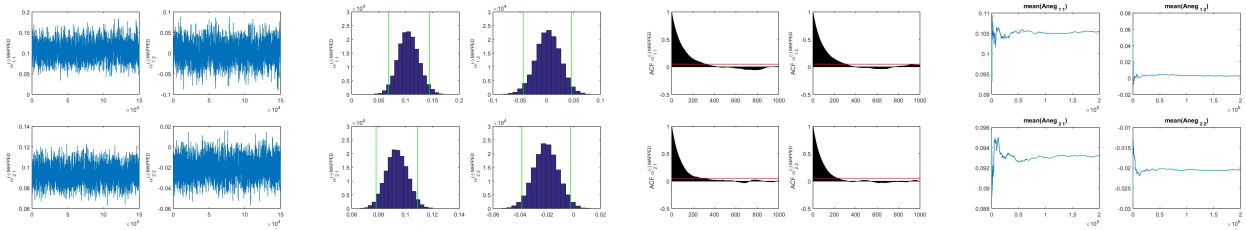


Figure B.4.5: MCMC traces, posterior histograms, ACFs and running averages of the components of the post-processed matrix $\mathbf{F}$. The green lines in the histogram represent the 95% C.I.

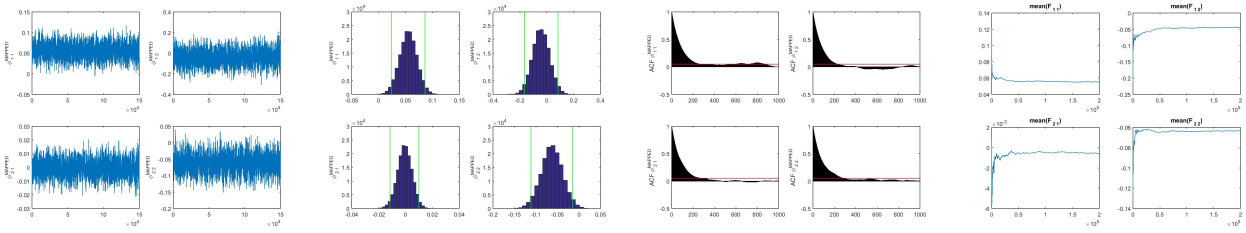


Figure B.4.6: The upper plot shows the traces of total number of components and of the number of active components at each step. The lower plot shows the corresponding running averages.

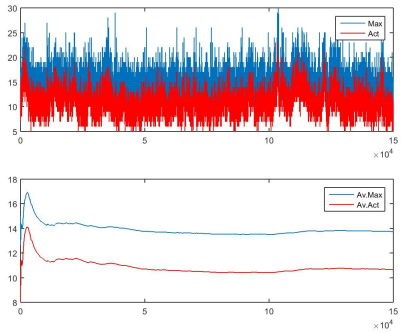
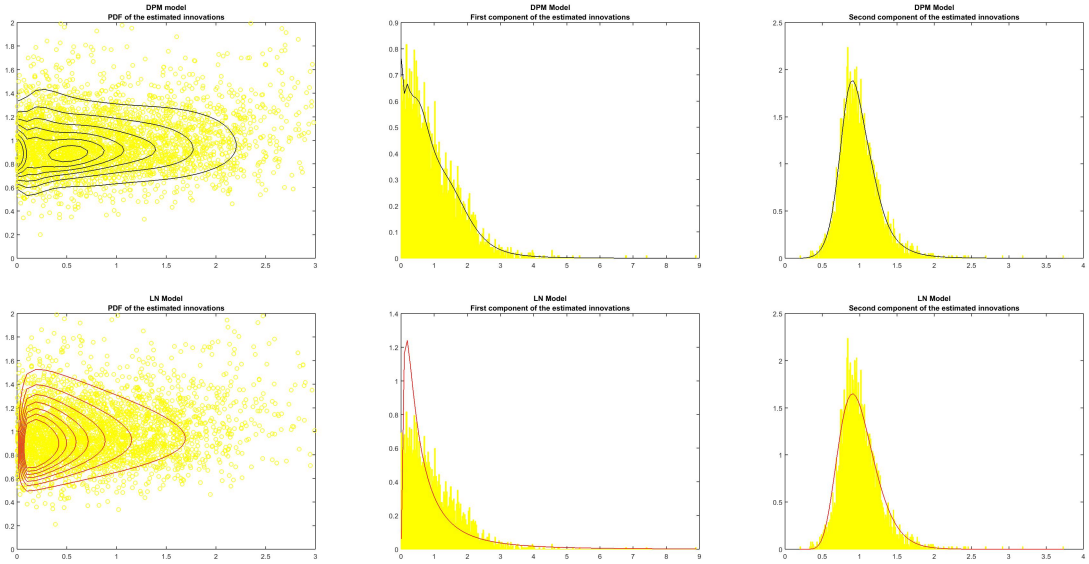


Figure B.4.7: Estimated joint and marginal densities of the innovations over the estimated innovations. On the upper row there are the results obtained with the DPMLN2-vMEM, on the lower row the ones obtained with LN1-vMEM



B.4.2 DJIA

Figure B.4.8: MCMC traces, posterior histograms, ACFs and running averages of the components of the post-processed vector $\boldsymbol{\omega}$. The green lines in the histogram represent the 95% C.I.

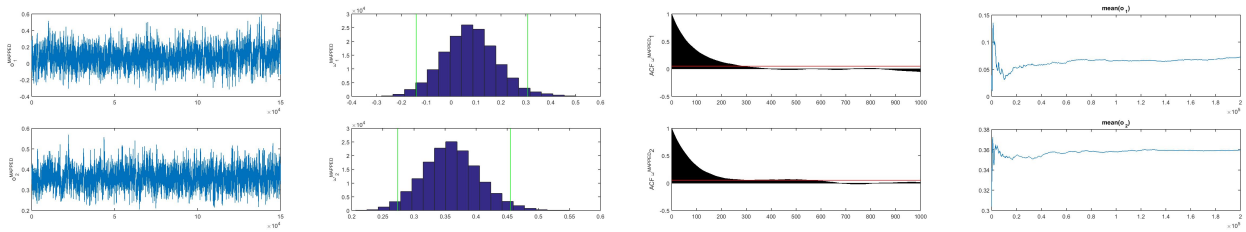


Figure B.4.9: MCMC traces, posterior histogram, ACFs and running averages of the components of the post-processed vector $\boldsymbol{\beta}$. The green lines in the histogram represent the 95% C.I.

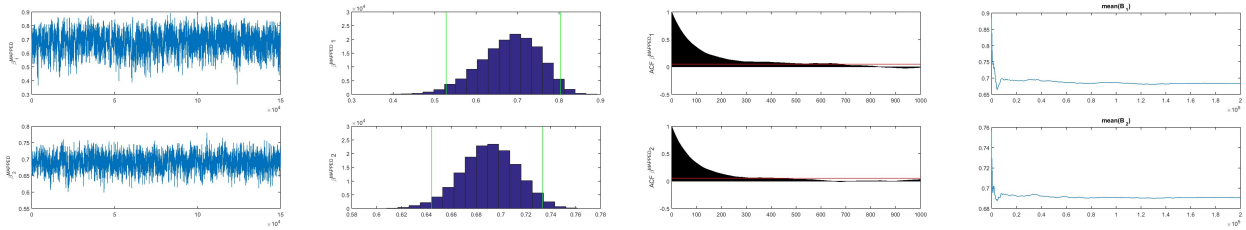


Figure B.4.10: MCMC traces, posterior histograms, ACFs and running averages of the components of the post-processed matrix $\mathbf{A}$. The green lines in the histogram represent the 95% C.I.

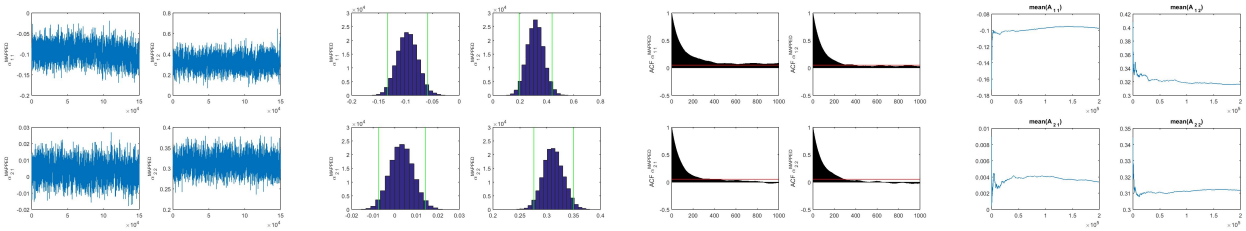


Figure B.4.11: MCMC traces, posterior histograms, ACFs and running averages of the components of the post-processed matrix $\mathbf{A}^{(-)}$. The green lines in the histogram represent the 95% C.I.

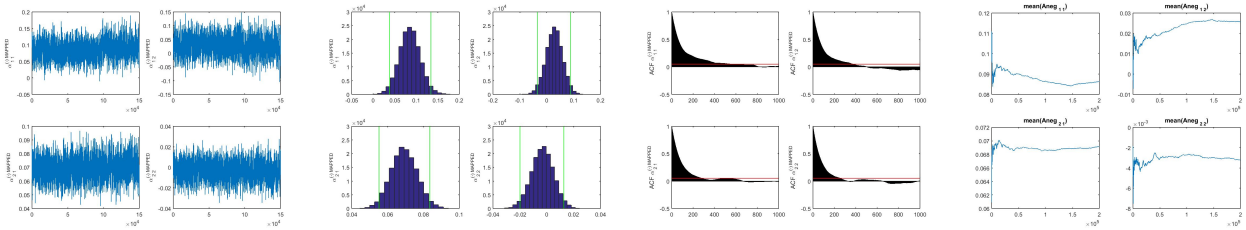


Figure B.4.12: MCMC traces, posterior histograms, ACFs and running averages of the components of the post-processed matrix $\mathbf{F}$. The green lines in the histogram represent the 95% C.I.

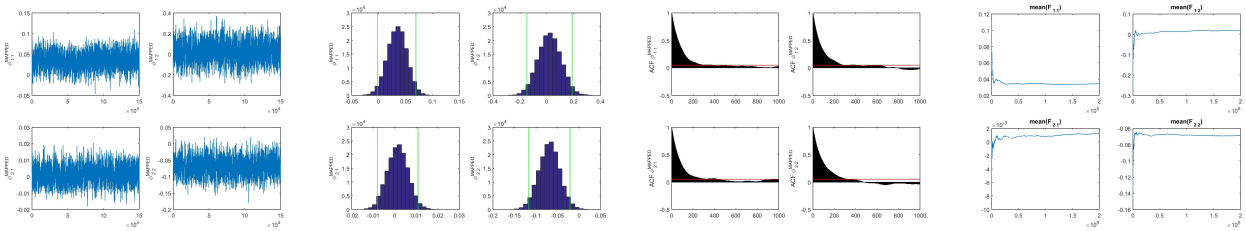


Figure B.4.13: The upper plot shows the traces of total number of components and of the number of active components at each step. The lower plot shows the corresponding running averages.

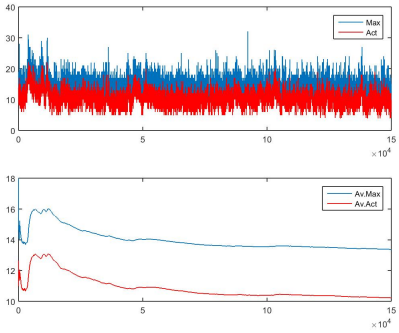
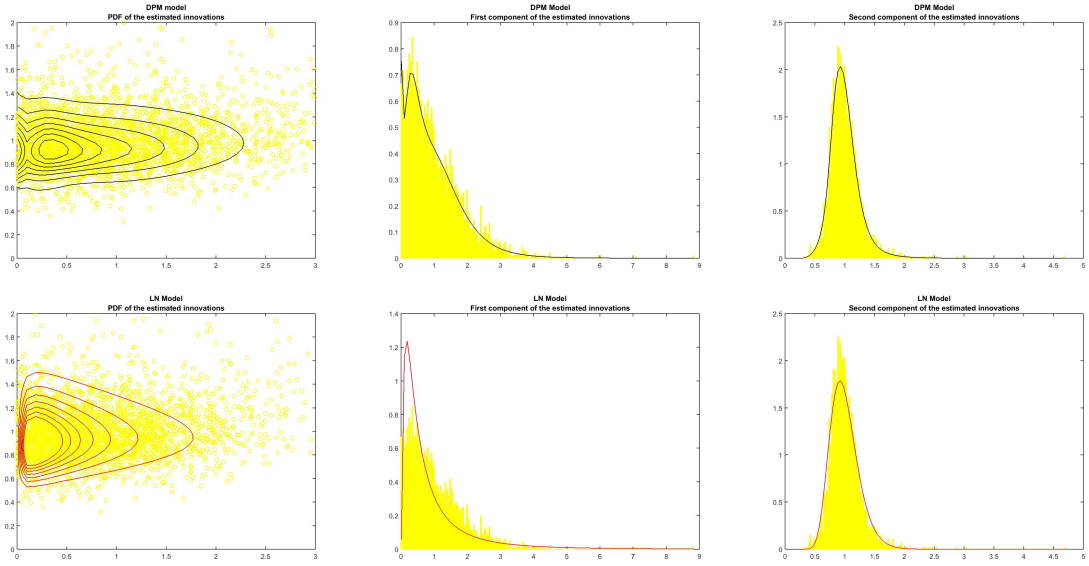


Figure B.4.14: Estimated joint and marginal densities of the innovations over the estimated innovations. On the upper row there are the results obtained with the DPMLN2-vMEM, on the lower row the ones obtained with LN1-vMEM



B.4.3 FTSE 100

Figure B.4.15: MCMC traces, posterior histograms, ACFs and running averages of the components of the post-processed vector $\boldsymbol{\omega}$. The green lines in the histogram represent the 95% C.I.

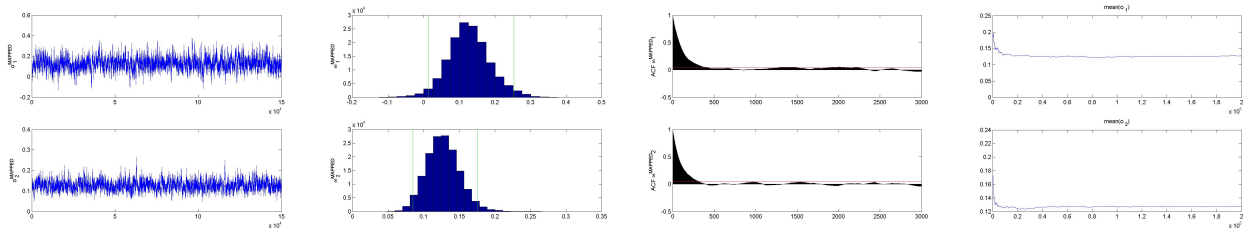


Figure B.4.16: MCMC traces, posterior histogram, ACFs and running averages of the components of the post-processed vector $\boldsymbol{\beta}$. The green lines in the histogram represent the 95% C.I.

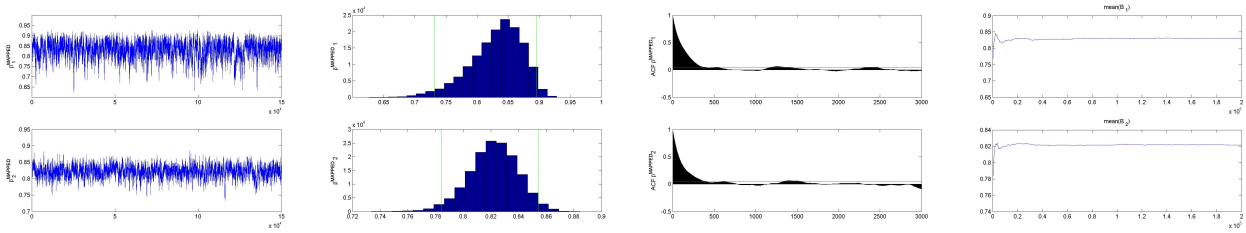


Figure B.4.17: MCMC traces, posterior histograms, ACFs and running averages of the components of the post-processed matrix $\mathbf{A}$. The green lines in the histogram represent the 95% C.I.

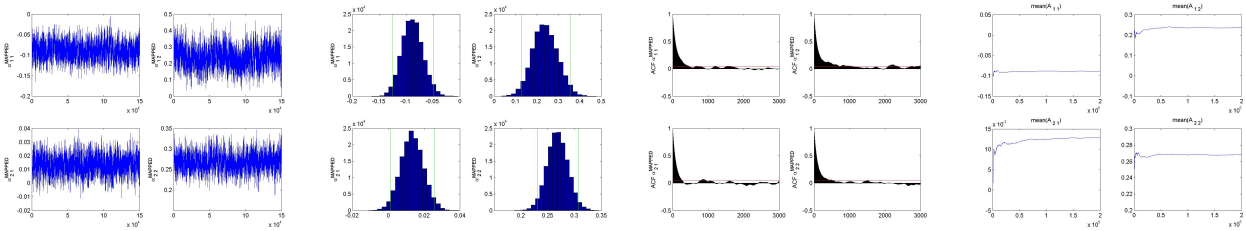


Figure B.4.18: MCMC traces, posterior histograms, ACFs and running averages of the components of the post-processed matrix $\mathbf{A}^{(-)}$. The green lines in the histogram represent the 95% C.I.

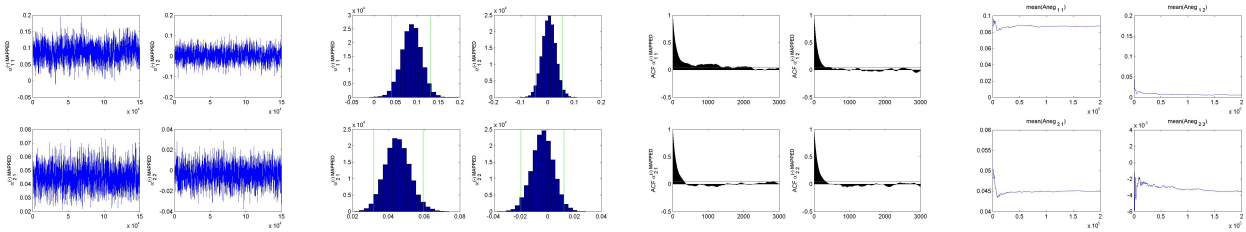


Figure B.4.19: MCMC traces, posterior histograms, ACFs and running averages of the components of the post-processed matrix $\mathbf{F}$. The green lines in the histogram represent the 95% C.I.

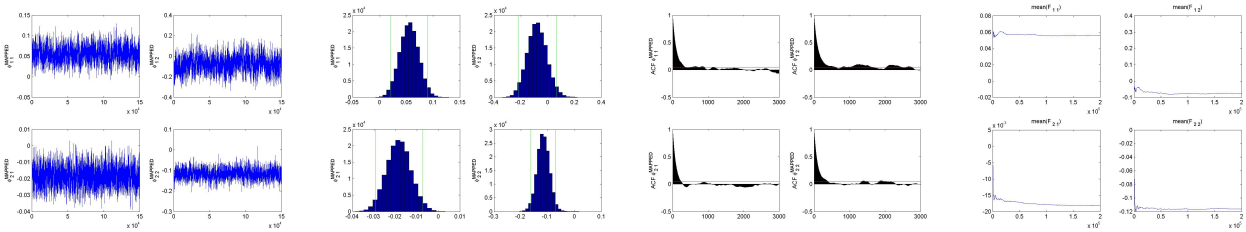


Figure B.4.20: The upper plot shows the traces of total number of components and of the number of active components at each step. The lower plot shows the corresponding running averages.

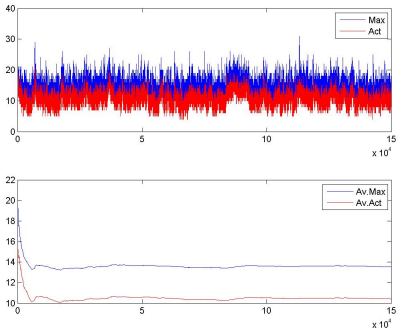
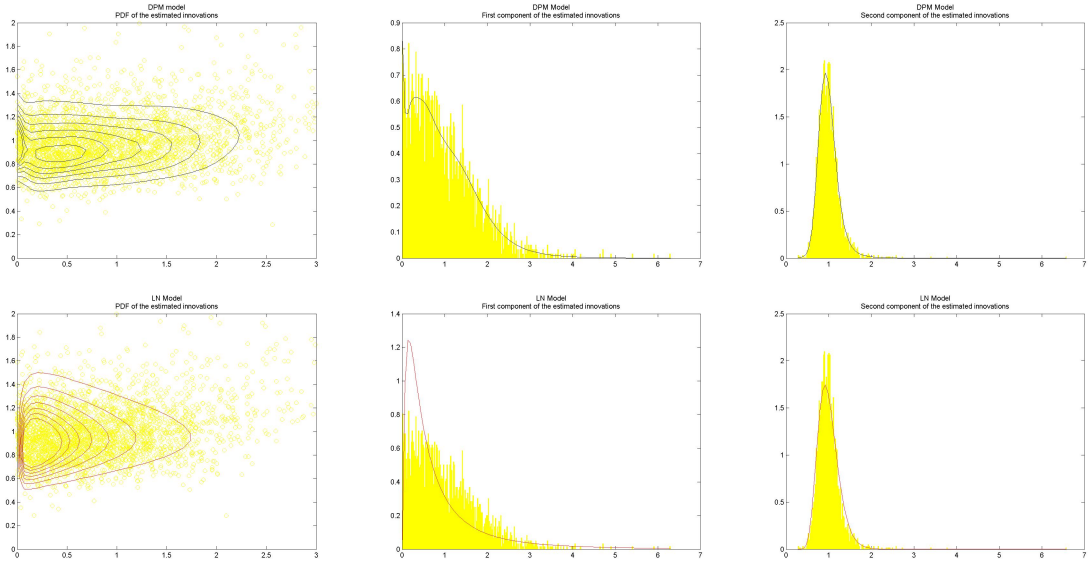


Figure B.4.21: Estimated joint and marginal densities of the innovations over the estimated innovations. On the upper row there are the results obtained with the DPMLN2-vMEM, on the lower row the ones obtained with LN1-vMEM



Appendix C

Codes and data

The Matlab codes used for the analyses in Chapters 5 and 6 and the dataset from the Oxford Man Institute "Realized Library" (Shephard and Sheppard (2010)) are available for download at the following link:

`https://dl.dropboxusercontent.com/u/8383093/Codes.rar`

Bibliography

Andersen, T., and T. Bollerslev (1998): "Answering the skeptics: Yes, standard volatility models do provide accurate forecasts," *International Economic Review*, pp. 885–905.

Andersen, T., T. Bollerslev, F. Diebold, and H. Ebens (2001): "The distribution of realized stock return volatility," *Journal of financial economics*, 61(1), 43–76.

Andersen, T., T. Bollerslev, F. Diebold, and P. Labys (2001): "The distribution of realized exchange rate volatility," *Journal of the American statistical association*, 96(453), 42–55.

Antoniak, C. (1974): "Mixtures of Dirichlet processes with applications to Bayesian nonparametric problems," *The Annals of Statistics*, pp. 1152–1174.

Barndorff-Nielsen, O. E., P. R. Hansen, A. Lunde, and N. Shephard (2008): "Designing realized kernels to measure the ex post variation of equity prices in the presence of noise," *Econometrica*, 76, 1481–1536.

Blackwell, D., and J. MacQueen (1973): "Ferguson distributions via Pólya urn schemes," *The annals of statistics*, pp. 353–355.

Brownlees, C. T., F. Cipollini, and G. M. Gallo (2011): "Intra-daily volume modeling and prediction for algorithmic trading," *Journal of Financial Econometrics*, 9(3), 489–518.

——— (2012): "Multiplicative Error Models," in *Handbook of Volatility Models and Their Applications*, pp. 223–247. Wiley.

Chou, R. (2005): "Forecasting financial volatilities with extreme values: the conditional autoregressive range (CARR) model," *Journal of Money, Credit and Banking*, 37(3), 561–582.

Cifarelli, D., and E. Regazzini (1979): "Considerazioni generali sull'impostazione Bayesiana di problemi non parametrici. Le medie associative nel contesto del processo aleatorio di Dirichlet parte I," *Rivista di matematica per le scienze economiche e sociali*, 2(2), 39–52.

——— (1990): "Distribution functions of means of a Dirichlet process," *The Annals of Statistics*, pp. 429–442.

Cipollini, F., R. Engle, and G. Gallo (2006): "Vector multiplicative error models: representation and inference," Discussion paper, National Bureau of Economic Research.

——— (2013): "Semiparametric vector MEM," *Journal of Applied Econometrics*, 28(7), 1067–1086.

Engle, R. (1982): "Autoregressive conditional heteroscedasticity with estimates of the variance of United Kingdom inflation," *Econometrica*, 50(4), 987–1007.

——— (2002): "New frontiers for ARCH models," *Journal of Applied Econometrics*, 17(5), 425–446.

Engle, R., and G. Gallo (2006): "A multiple indicators model for volatility using intra-daily data," *Journal of Econometrics*, 131(1), 3–27.

Engle, R., G. Gallo, and M. Velucchi (2009): "A MEM-based analysis of volatility spillovers in East Asian financial markets," .

Engle, R., G. Gallo, and M. Velucchi (2012): "Volatility spillovers in East Asian financial markets: a MEM-based approach," *Review of Economics and Statistics*, 94(1), 222–223.

Engle, R., and J. Russell (1998): "Autoregressive conditional duration: a new model for irregularly spaced transaction data," *Econometrica*, 66(5), 1127–1162.

Ferguson, T. (1973): "A Bayesian analysis of some nonparametric problems," *The annals of statistics*, pp. 209–230.

Gallo, G., and M. Velucchi (2009): "Market interdependence and financial volatility transmission in East Asia," *International Journal of Finance & Economics*, 14(1), 24–44.

Giovannetti, G., and M. Velucchi (2011): "A MEM analysis of African financial markets," in *New Perspectives in Statistical Modeling and Data Analysis*, pp. 319–328. Springer.

Johnson, N., S. Kotz, and N. Balakrishnan (1997): *Discrete multivariate distributions*. Wiley New York.

——— (2000): *Continuous multivariate distributions*. Wiley New York.

Jørgensen, B., and S. J. Knudsen (2004): "Parameter orthogonality and bias adjustment for estimating functions," *Scandinavian Journal of Statistics*, 31(1), 93–114.

Kalli, M., J. Griffin, and S. Walker (2011): "Slice sampling mixture models," *Statistics and computing*, 21(1), 93–105.

Kim, S., N. Shephard, and S. Chib (1998): "Stochastic volatility: likelihood inference and comparison with ARCH models," *The Review of Economic Studies*, 65(3), 361–393.

Lanne, M. (2006): "A mixture multiplicative error model for realized volatility," *Journal of Financial Econometrics*, 4(4), 594–616.

Lijoi, A., R. E. (2004): "Means of a Dirichlet process and multiple hypergeometric functions," *Annals of probability*, pp. 1469–1495.

Liu, C., R. D. W. Y. (1998): "Parameter expansion to accelerate EM: the PX-EM algorithm," *Biometrika*, 85(4), 755–770.

Liu, J.S., W. Y. (1999): "Parameter expansion for data augmentation," *Journal of the American Statistical Association*, 94(448), 1264–1274.

MANGANELLI, S. (2005): "Duration, volume and volatility impact of trades," *Journal of Financial markets*, 8(4), 377–399.

MULIERE, P., AND L. TARDELLA (1998): "Approximating distributions of random functionals of Ferguson-Dirichlet priors," *Canadian Journal of Statistics*, 26(2), 283–297.

PARKINSON, M. (1980): "The extreme value method for estimating the variance of the rate of return," *Journal of Business*, 53, 61–65.

PITMAN, J. (2002): "Combinatorial stochastic processes," Discussion paper, U.C. Berkeley Department of Statistics.

REGAZZINI, E., A. LIJOI, AND I. PRÜNSTER (2003): "Distributional results for means of normalized random measures with independent increments," *Annals of Statistics*, pp. 560–585.

ROBERTS, G., AND J. ROSENTHAL (2007): "Coupling and ergodicity of adaptive Markov chain Monte Carlo algorithms," *Journal of applied probability*, pp. 458–475.

SETHURAMAN, J. (1994): "A constructive definition of Dirichlet priors," *Statistica Sinica*, 4, 639–650.

SHEPHARD, N., AND K. SHEPPARD (2010): "Realising the future: forecasting with high-frequency-based volatility (HEAVY) models," *Journal of Applied Econometrics*, 25(2), 197–231.

SOLGI, R., AND A. MIRA (2013): "A Bayesian Semiparametric Multiplicative Error Model with an Application to Realized Volatility," *Journal of Computational and Graphical Statistics*, 22(3), 558–583.

TAYLOR, S. (1987): "Modelling financial time series," .

VAN DYK, D.A., M. X. (2001): "The art of data augmentation," *Journal of Computational and Graphical Statistics*, 10(1).

WALKER, S. (2007): "Sampling the Dirichlet mixture model with slices," *Communications in Statistics: Simulation and Computation*, 36(1), 45–54.