Abstract

In this work, we will focus on the Gibbs sampler and we will present how to combine the important sampling within the Gibbs sampling, employing an augmenting function to modify the target distribution of the sampler. The almost sure convergence of the estimates under the new measure will be proved and it will be shown that with respect to new measure the variance of the monte carlo integration can be reduced. More interestingly the change of measure will induce a modification of the Markov chain and for a proper choice of the important sampling function the new chain will show better mixing properties than the original one. All these will be applied to two cases of interest, the first is the smoothing of the unobservable conditional volatility in a stochastic volatility model and the second is the smoothing of the unobservable state in a switching state space model.

Change of Measure in Monte Carlo Integration via Gibbs Sampling

Filippo Altissimo*
Research Department
Bank of Italy
Via Nazionale, 91
Rome, 00184
Italy
ph. +39 06 4792-3441
fax. +39 06 4747-820
email alti.3441@interbusiness.it

First draft: October 1997
This draft: July 1998

*I thank Valentina Corradi, Riccardo Cristadoro, Frank Diebold and the partecipants to a seminar at University of Pennsylvania for helpful comments and suggestions. Errors are only my responsibility.

1. Introduction

The use of simulation methods is becoming widely applied in econometrics and finance in recent years. Among the variety of methods proposed, Markov chain monte carlo and, in particular, Gibbs sampling proved to be the most interesting and powerful one.

Markov chain monte carlo[1] is essentially monte carlo integration using Markov chain. Bayesians and also frequentists need to integrate high dimensional probability distributions to make inference about parameters of the model or to make prediction or to filter variables of interest. In particular, Bayesians need to integrate over the posterior distribution of the model parameters and classics may need to integrate out unobservable variables in the computation of the likelihood.

Monte carlo methods of integration relay upon samples from the objective distribution to average over. Markov chain monte carlo draws those samples by running a properly constructed Markov chain which has the distribution to sample from as the invariant distribution.

Initially, Markov chain was extremely attractive in Bayesian inference since that most problems of interest in that framework can be solved by properly simulating the posterior distribution, and the MCMC proved to be a powerful tool in doing it.

Lately, no Bayesian application of monte carlo Markov chain methods are becoming increasly popular; in the generality of cases they are based on the possibility of using MCMC to calculate probabilities or expectations which otherwise cannot be computed analytically. Applications of those techniques have been presented in the literature to likelihood maximization, by Geyer and Thompson (1992), to simulated EM, by Shepard (1994), and to filtration in nonlinear state space models, by Carter and Kohn (1994) and Dejong and Shepard (1995); a much longer list can be found in Geyer (1996).

In many of the indicated applications straight methods, in particular Gibbs sampling, work surprising well, however, as the complexity of the applications increases the performance in term of speed and accuracy of those methods is likely to be affected.

In fact the presence of poor mixing of the chain which generates the samples, can be an extra source to add on top of other sources of poor performance of monte carlo integration in

[1] Markov Chian Monte Carlo will sometimes be refered as MCMC, here on.

the computation of expectation and probabilities. A large literature has developed refining and improving the performance of monte carlo integration through variance reduction technique and in particular through important sampling methods. However the use of the realizations from a MCMC chain to perform monte carlo integration adds the problem of how to extend those techniques to this new framework and how the problem of the dependence of the realizations induced by the chain can affect the result of the computation.

In this work, we will focus on the Gibbs sampler and we will present how to combine the important sampling within the Gibbs sampling, employing an augmenting function to modify the target distribution of the sampler. Similar to the results of Geweke (1989) in the case of independent monte carlo, the almost sure convergence of the estimates under the new measure will be proved and it will be shown that with respect to new measure the variance of the monte carlo integration can be reduced. More interestingly the change of measure will induce a modification of the Markov chain and for a proper choice of the important sampling function the new chain will show better mixing properties than the original one. We will specialize this to the case of the data augmentation showing that the chain can be modified to obtain one independent in its elements and the property of reversability of the data augmentation delivers a criteria for choosing among augmenting functions.

All these will be applied to two cases of interest. First is the smoothing of the underlying conditional variance in a stochastic volatility model; stochastic volatility models arise naturally as discrete approximation of various diffusion processes in the continuous asset pricing literature and recently this model has become a natural battleground for the application of different MCMC techniques due to the intractability of the likelihood. Second is the smoothing of the unobservable state in a switching state space model.

The work is organized as follows. In section 2 and 3 the ideas of monte carlo integration, important sampling and Gibbs sampling are introduced. Section 4 introduces the idea of the change of measure through a simple example. Section 5 presents in details how to perform the change of measure and section 6 specializes it to the data augmentation. Finally in section 7 the properties of the methodology are investigated by means of two applications.

2. Monte carlo integration and important sampling

In many econometrics problem, it is not uncommon to work in presence of latent factors (or missing observations) which have to be integrated out in the joint distribution of the observable and unobservable variables. In the generality of cases, this step of marginalization relies upon numerical integration of the joint density, given the dimension and the analytical intractability of those integrals. Among the various methods, the monte carlo integration has been increasingly used for its simplicity and for its accuracy compared with deterministic methods of integration in the case of large dimensional integral, as often occurs in economic model when the unobservable variables have dynamic (Geweke, 1995).

The precision of the monte carlo methods of integration depends upon the choice of the proper weighting function, used to select and weight the points at which the function of interest has to be evaluated. The weighting function is called the important sampling density, a term due to Hammersly and Handscomb (1964). The necessity of a proper choice of the important sampling function derives from the fact that the natural important sampling, given directly by the specification of the problem, can be highly inefficient in term of variance of the numerical evaluation of the integral of interest.

Suppose that we are trying to evaluate the following integral

$$I = \int g(x)\, p(x)\, dx \tag{2.1}$$

where $p(x)$ is the probability density of x. Using a brute force approach, let $\{x_i\}_{i=1}^{N}$ be an $i.i.d.$ sample from the *pdf* $p(x)$, then the monte carlo estimator of I is

$$\widetilde{I} = \frac{1}{N} \sum_{i=1}^{N} g(x_i) \tag{2.2}$$

and by a SLLN it will converge almost surely to I if some regularity conditions are satisfied. The asymptotic variance of the estimator $\sqrt{N}\left(\widetilde{I} - I\right)$ is given by

$$\int [g(x) - I]^2\, p(x)\, dx, \tag{2.3}$$

in most cases the value of (2.3) can be sizable.

A way of reducing the variance of the monte carlo integral is to properly define a change of measure through an important sampling function $w(x)$ such that

$$I = \int g(x) \frac{p(x)}{w(x)} w(x) \, dx \tag{2.4}$$

and the important sampling estimator of the integral is given by

$$\widehat{I} = \frac{1}{N} \sum_{i=1}^{N} g(x_i) \frac{p(x_i)}{w(x_i)} \tag{2.5}$$

where $\{x_i\}$ be an $i.i.d.$ sample from the density $w(x)$. $\widehat{I}$ is an unbiased estimator of I and it will converge to it by a SLLN result. The asymptotic variance of the important sampling estimator of I is

$$\int \left[g(x) \frac{p(x)}{w(x)} - I \right]^2 w(x) \, dx. \tag{2.6}$$

The important sampling scheme has two advantages: first it can be helpful in cases in which there are no simple methods for constructing draws from the distribution p and, more relevant, it is possible that a proper choice of the weight function w will induce a reduction in the variance of the estimator of the integral, i.e. (2.6) is smaller than (2.3).

The proper choice of the weight function is an open question which has to be evaluated on a case by case bases. Geweke (1989) pointed out that under a set of regularity conditions the important sampling density which minimizes (2.6) has kernel density equal (i.e. it is proportional) to $|g(x) - I| p(x)$. From this finding, he derives the common rule of thumb which suggests that the important sampling function should have thicker tails then the original density itself.

Sometimes the independent monte carlo method, just described, cannot be implemented either for the difficulties in finding an important sampling function w with satisfactory properties or for the incomplete knowledge of the form of the density p; in those cases a generalization of the independent monte carlo, that has become known as Markov chain monte carlo, can be an useful alternative.

3. The Gibbs Sampler

Markov chain monte carlo methods are simulation techniques that generate a sample from a target distribution through the specification of a transition kernel of a Markov process which

has as invariant distribution the target distribution itself[2].

Markov chain monte carlo methods have an history in mathematical physics dating back to the algorithm of Metropolis et al. (1953). Such algorithm is now known with the name of Metropolis-Hastings algorithm. It allows to simulate complex, nonstandard multivariate distributions and it does not require the knowledge of the normalizing constant of the density itself. In the Metropolis algorithm the chain is constructed so that the following value of the chain is generated from a 'proposal' density (called auxiliary density) and then accepted or rejected according to the value of the target density at the new value respect with the present one.

A particular version of Metropolis-Hastings algorithm[3], originally suited for image reconstruction, was proposed by Geman and Geman (1984). This algorithm, known as Gibbs sampler, allows to draw a random vector from the joint distribution without its full knowledge but knowing only the sequence of full conditional distributions. The subsequent elements of the Markov chain are obtained by sampling elements of the random vector one at the time from the full conditional distributions. This method proved to be of great potentiality in Bayesian statistics and lately it proved to be very successful in the treatment of latent and unobservable variables in the classical framework. Being the focus of the work, the construction of the Gibbs algorithm will be described more carefully.

Consider a random vector, $x \in \Re^m$, with density $p(x)$ and consider a blocking scheme of the vector $(x_{(1)}, ..., x_{(k)})$ such that $x_{(i)} = (x_{i1}, ..., x_{im(i)})$ and $\sum m(i) = m$. The object is to make a draw of x from $p(x)$ but the function is not known or it is unfeasible to draw form it. Instead the conditional distribution of the i block respect to the rest of the variables, $p(x_{(i)}|x_{(-i)})$, is available and it is easy to handle.

The sampler works in a sequential way as follows[4]:

1) Specify a starting value x^0 of the chain and set $x^i = x^0$ and $i = 1$;

[2] A comprehensive reference for MCMC methods is in Gilks, Richardson and Spiegelhalter (1996).

[3] It is possible to show that the Gibbs sampler can be interpreted as a special case of the Metropolis-Hastings algorithm.

[4] Under-scores will mean blocks of the random vector, while upper-scores will indicate iterations of the sampler.

2)

draw $x_{(1)}^{i+1}$ from $p(x_{(1)}^{i+1}|x_{(j)}^{i} \ \ (j>1))$,

...

draw $x_{(i)}^{i+1}$ from $p(x_{(i)}^{i+1}|x_{(j)}^{i+1} \ \ (j<i) \ \ , x_{(j)}^{i} \ \ (j>i))$,

...

draw $x_{(k)}^{i+1}$ from $p(x_{(k)}^{i+1}|x_{(j)}^{i+1} \ \ (j<k))$;

3) set $i = i+1$ and $x^i = x^{i+1}$ and goto step 2.

The iteration of this algorithm thus provides a Markov chain whose transition probability from x to $\widetilde{x}$ is given by

$$K_{GS}\left(x,\widetilde{x}\right) = \prod_{i=1}^{k} p(\widetilde{x}_{(i)}|\widetilde{x}_{(j)} \ \ (j<i)\,, \ x_{(j)} \ \ (j>i)). \tag{3.1}$$

It is easy to see that if x^0 is drawn from $p(x)$ then also x^i for $i>0$ is a draw from $p(x)$. If instead x^0 is not a draw from $p(x)$ then, as $i \to \infty$, and if some regularity conditions are satisfied the distribution of x^i converges to p.

More precisely, p is an invariant distribution to the chain generated by the transition kernel K_{GS} and if the transition kernel is also p-irreducibility and aperiodic (Tierney 1994, Roberts and Smith 1994) then for almost every initial conditions x^0 the conditional distribution $p(x^i|x^0)$ converges to the p in the sense of the total variation norm[5] and p is the unique invariant distribution. Roberts and Smith (1994) provide more primitive conditions on p which assure the result; recently Athreya, Doss and Sethuraman (1996) have weakened those conditions and proved that the probability measure with respect to which the irreducibility needs to hold, is not restricted to be the invariant measure itself; this is useful in the case in which there is no knowledge of the functional form of p.

So, the interesting and useful part of the Gibbs sampler scheme is that we do not need to know p but it is enough that we are able to construct a blocking scheme and a sequence of conditional distributions from which we can make the draws in sequence. After a sufficient long

[5] Definition of irreducibility, aperidicity are given in appendix 1; a good reference on the topics is Nummelin (1984).

period of warming up which is function of the degree of dependence the blocks, the chain will converge to its invariant distribution.

In operating the Gibbs sampler three aspects of its implementation are crucial: the blocking scheme, the number of simulation paths and the length of the simulation path.

First, the choice of the blocking scheme is in part constrained by the knowledge of the conditional distributions, but a proper choice of the blocking can be crucial for the performance of the sampler; in fact if some components of the random vector are highly dependent, grouping them together can increase the speed of convergence and reduce the dependence of the Markov chain itself (Lui, Wong and Kong, 1994).

Second key factor is the choice between single or multiple simulation paths, given that it is possible to use a unique long simulation or to restart the chain from different initial conditions. The multiple paths are more cumbersome to perform, on the other hand the realizations of a single path simulation can show dependence, while this is not the case for realization coming from different paths if the initial conditions are chosen in an independent manner.

Finally, both in the single and the multiple paths sampler there is the problem of choosing the length of the simulation and determining how long the warming-up phase has to be, before it is possible to consider that the sampler has reached its invariant distribution. The length of the simulation relates to the degree of precision desired in the computation of the object of interest, while the length of the warming-up is related to the dependence on the initial condition.

In this work, the samplers are run in a single path way, given that this is common practice in the literature and the problem of dependence of the chain is relevant in both ways of sampling: in one case, in term of lasting of the initial condition and in the other in term of variance of the estimates.

Again, the Gibbs sampler gives the possibility of solving situation in which we need to make draws from a distribution of which we have limited knowledge or it is difficult to handle, but whose conditional distributions are easy to work with. The fact that we can simulate from p, does not mean that we can make an efficient draw from the target distribution, this consideration opens the way to the combination of important sampling and Gibbs sampling.

4. A simple example

The combination of Gibbs sampling and monte carlo integration has proved to be a new extremely successful tool particularly in presence of rapid improvement of computer performance. However the use of realization of the sampler to preform integration induces to see how the usual variance reduction techniques can be extended to this setting and how the dependence of the realizations of the chain can affect the precision of the computed integral.

Our aim is to show how the important sampling can be combined with the Gibbs sampler to define a change of measure that improves the overall performance of the algorithm in various respects.

First, a proper choice of the important sampling function could be used to reduce the variance of the monte carlo integration for which the Gibbs sampler draws are made.

Second, the important sampling function can be constructed in such a way that the new Markov chain under the new measure shows lower dependence with respect to the original structure so allowing a faster mixing of the chain to its invariant distribution.

Finally, the modification of the Gibbs sampler chain can ease the job of running the sampler itself if the conditional distribution of the modified model are less cumbersome.

To convoy the idea of the method proposed, a simply and analytically friendly example is worked out. Consider a linear regression model

$$y_i = \alpha + \beta x_i + \varepsilon_i, \quad \varepsilon_i \sim NID(0, \sigma^2)$$

where σ^2 is assumed known and a sample $\{y_i,\ x_i\}_{i=1}^{T}$ is given. The aim is to compute the posterior expectation of β given the sample. It is assumed a flat prior on the parameters. The problem reduces to the computation of

$$\int\int \beta\, p(\alpha, \beta | y, x) d\alpha d\beta \tag{4.1}$$

where $p(\alpha, \beta | y, x)$ is the join posterior of the parameters, which is known to be normal in this case.

Assume that we just learned Gibbs sampler and we want to use it in performing the exercise, even if much easier alternatives are available. The chain is constructed by sampling the two

parameters sequentially from $p(\alpha|\beta, y, x)$ and $p(\beta|\alpha, y, x)$ and the realizations of β are averaged, as

$$\widetilde{\beta} = \frac{1}{N}\sum_{i=1}^{N}\beta_i.$$

The simplicity of the example allows to work out the variance of the estimator. The dependence in the chain is driven by the posterior correlation between α and β in the chain which is

$$\rho_{\alpha\beta} = -\frac{\overline{x}}{\left(\frac{\sum x_i^2}{T}\right)^{\frac{1}{2}}}$$

where $\overline{x}$ is the sample mean of x and it is easy to see, using the results of Liu, Wong and Kong (1994), that the correlation between realization of β at distance τ in the chain is $\rho_{\alpha\beta}^{2\tau}$. So it follows that the asymptotic variance of $\sqrt{N}(\widetilde{\beta} - \beta)$ is given by

$$\begin{aligned} V(\widetilde{\beta}) &= \sum_{\tau=-\infty}^{\infty}\gamma_\beta(\tau) = \sigma_\beta^2\left(1 + 2\sum_{i=1}^{\infty}\rho_{\alpha\beta}^{2i}\right) \\ &= \sigma_\beta^2\left(\frac{1+\rho_{\alpha\beta}^2}{1-\rho_{\alpha\beta}^2}\right). \end{aligned} \tag{4.2}$$

where σ_β^2 is the contemporaneous variance of β[6].

Clearly, high posterior correlation of the parameters induces a large variance of the estimates and a large posterior correlation is function of the ratio of the mean of x with respect to its second moment. Note how the problem of slow mixing can occur in the simplest of the statistical model. The just learned strategy of Gibbs sampling does not prove to be successful in the case in which there is high dependence among the elements of the chain. An idea to improve the performance would be to define a change of measure such that under the new measure the chain shows lower dependence.

In fact it can be possible to modify (4.1) as

[6]Note that the close form of the variance of the estimates derives by the reversability property of the Gibbs sampler in the case in which there are only two blocks in the chain. This property is not present if the number of block is greater than two. A transition kernel $K(x, y)$ with invariant distribution π is reversible if $K(x, y) = K(y, x)$.

$$\int\int \beta\, p(\alpha,\beta|y,x)d\alpha d\beta =$$

$$\int\int \frac{\beta}{f(\alpha,\beta|y,x)} f(\alpha,\beta|y,x)p(\alpha,\beta|y,x)d\alpha d\beta \tag{4.3}$$

where $f(\alpha,\beta|y,x)$ is a proper function of α, β such that $f \cdot p$ is still a kernel density. So under the modified measure, the new density will be proportional to $f \cdot p$ and the function of interest will be $\frac{\beta}{f}$.

Before going on with the example, it is the case to note the difference between what just proposed and the usual important sampling introduced in section (2). In that case, we have a full knowledge of the p function and we define a density w such that we draw from it and weight the realization by the ratio of the original density and w. Here instead we use a function f to augment the original distribution so that the new density is defined as proportional to the product of the original one and the augmenting one. It will be shown in the following section how to draw from this new density.

Going back to the problem, it is possible to find a function f such that the new chain which has as invariant distribution $f \cdot p$ has a lower dependence of the original one and such that the variance of the estimates of β under the new measure

$$\widehat{\beta} = \frac{1}{N}\sum_{i=1}^{N} \frac{\beta_i}{f(\alpha_i,\beta_i|y,x)} \tag{4.4}$$

is lower compared to (4.2). For the semplicity of the example it has been assumed that the constant of integration is known and $\int\int f \cdot p d\alpha d\beta = 1$.

Anticipating future arguments, the choice of the function f will be driven by two consideration, first the possibility that under f the parameters present opposite correlation with respect to the one under p and second that f has sufficient mass on the domain of p. In our example the natural choice is given by a f proportional to a multivariate t density with few degrees of freedom and covariance between the parameters of opposite sign respect to p.

To conclude the example in the following picture, the variance of the estimates of the posterior average of β through Gibbs sampler under the original measure, as in (4.2), and under the new measure are plotted in Fig. (4.1) for different values of $\rho_{\alpha\beta}$ ranging in $(-0.95, 0.95)$, in the case

in which $\beta = .5$, $\alpha = 1$, $E(x^2) = 1$ and in which the f function is a multivariate t with five degrees of freedom and equal variance and covariance of opposite sign respect to p.

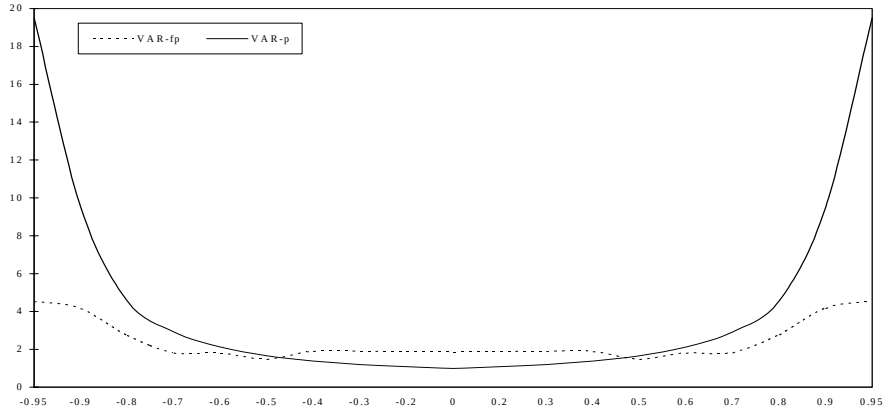

Figure 4.1: Variance of β under the original and the new measure

The U-shaped curve is the variance under the original measure and it is greater than the variance under the modified measure for value of the correlation of the coefficients greater than .5, while for lower value the new measure suffer of the larger variance induced by the t distribution. The simplicity of the proposed example helps to show that running the chain under the original measure can be an unsuccessful strategy if the elements of the chain are highly dependent and this can end up in a large variance of estimates for which the draws are made.

5. Change of measure and the modification of Gibbs sampler

Given the result of the above example, let us try to design a general framework for performing change of measure in MCMC. The problem is to evaluate the following integral

$$I = E_p(g(x)) = \int g(x)\, p(x)\, dx, \tag{5.1}$$

in the case in which the distribution of interest p has an analytically intractable form and also it is not feasible to sample directly, so a MCMC algorithm is called for. The problem of interest is clearly a standard one in econometrics analysis which regularly arises in computing the likelihood function in presence of latent variables or in filtering some unobservable variables. The point is that instead of working under the measure p for the reasons previously highlighted, it could be

"useful" to define a change of measure by an augmenting density function f such that

$$I = E_{fp}(\frac{g(x)}{f(x)}) = \int \frac{g(x)}{f(x)} f(x)\, p(x)\, dx, \tag{5.2}$$

where the new measure is given by $f(x)\, p(x)$ and the function f is the Radon-Nikodým derivative of the new measure with respect to the original one. Where the term "useful" means that the variance of the sample estimates (5.2) will be lower than the one of (5.1) and also that the chain having p as invariant distribution can have worst mixing respect to the chain define under the new measure.

Note that the new measure is the product of the augmenting and the original density and so there are three things which have to be verified in performing the change of variable: first that the function $f(x)\, p(x)$ is itself a density function (or proportional to it), second that we are able to make draws from the new invariant distribution modifying properly the transition kernel of the chain and finally that the important sampling estimate of $\widehat{I}$ actually converges to the desired value.

If both f and p are densities, the necessity of assuring that $f(x)\, p(x)$ is actually the kernel of a density, mainly reduces to check the fact that the functions have a common support with positive measure in which they are both strictly positive. It means that $D \equiv D_f \cap D_p \neq \emptyset$ where $D_f = \{x | f(x) > 0\}$ and $D_p = \{x | p(x) > 0\}$. However the presence of a common support of f and p will not be sufficient to guarantee the convergence of the sample estimates under the new measure to I so the requirement will be strengthened by assuming that $D_p \subset D_f$. This assumption is a regularity condition which is intended to ensure that the new measure has positive mass on the same domain as the original one.

The change of measure induces also a modification of the target distribution from which it is necessary to sample from and consequently it is necessary a modification of the transition kernel generating the chain. Here it is proposed an algorithm which sample from the new measure $p(x)f(x)$; consider the same blocking structure as before and iterate the following algorithm:

1) Specify a starting value x^0 of the chain and set $x^i = x^0$ and $i = 1$;

2)

draw $x_{(1)}^{i+1}$ from $p(x_{(1)}^{i+1}|x_{(j)}^{i}\ (j>1))f(x_{(1)}^{i}|x_{(j)}^{i}\ (j>1))$,

...

draw $x_{(i)}^{i+1}$ from $p(x_{(i)}^{i+1}|x_{(j)}^{i+1}\ (j<i)\,,\ x_{(j)}^{i}\ (j>i))f(x_{(i)}^{i}|x_{(j)}^{i+1}\ (j<i)\,,\ x_{(j)}^{i}\ (j>i))$,

...

draw $x_{(k)}^{i+1}$ from $p(x_{(k)}^{i+1}|x_{(-k)}^{i+1})f(x_{(k)}^{i+1}|x_{(-k)}^{i+1})$;

3) set $i=i+1$ and $x^i=x^{i+1}$ and goto step 2.

The densities in step (2) of the algorithm are actually kernels of densities and they are the result of the product of two conditionals $p(\cdot|\cdot)$ and $f(\cdot|\cdot)$. It is possible that in a variety of cases the product of the marginals has known form that can be sampled directly; but in general this is not the case and other methods have to be applied. Here three methods are presented which can be implemented in performing random drawing from the kernel densities in step (2).

The first one is a straight acceptance/rejection method based on finding an auxiliary function $h(\cdot)$ and a constant M such that $\frac{p(\cdot|\cdot)f(\cdot|\cdot)}{h(\cdot)}\leq M<\infty$. The goodness of the methods is based on the ability of the auxiliary function h to mimic $p(\cdot|\cdot)f(\cdot|\cdot)$; operatively, the algorithm should work as follows:

a) make a draw $x_{(i)}$ form $h(x_{(i)})$;

b) draw u from a uniform distribution defined on the interval zero-one;

c) accept $x_{(i)}$ if u is less then $\frac{p(x_{(i)}|x_{(-i)})f(x_{(i)}|x_{(-i)})}{Mh(x_{(i)})}$, otherwise goto (a).

The choice of the function h becomes critical, however a sensible and easy choice for h is given by f so that the acceptance probability reduces to $\frac{p}{M}$.

Second alternative strategy, which can be implemented, is the adaptive/rejection method proposed by Gilks and Wild (1994) for log-concave density function. The method is similar to the acceptance/rejection one with the difference that the function h is properly constructed using tangent method so to form a tight envelope of the original function. The method applies to log-concave densities and if both p and f are log-concave, then their product is log-concave too and the adaptive/rejection procedure can be applied to the new measure.

The third method is given by combining the original Gibbs sampler with a Metropolis-Hastings independent step, in which the transition probability of the block j does not depend on its previous realization. This can be performed at each step of (2) by drawing from a transitional kernel $q(x_{(j)}^{i+1}|x_{(-j)})$ and then accept $x_{(j)}^{i+1}$ with probability $\alpha\left(x_{(j)}^{i+1}, x_{(j)}^{i}|x_{(-j)}\right)$ where

$$\alpha\left(x_{(j)}^{i+1}, x_{(j)}^{i}|x_{(-j)}\right) = \min\left\{\frac{p(x_{(j)}^{i+1}|x_{(-j)})f(x_{(j)}^{i+1}|x_{(-j)})q(x_{(j)}^{i}|x_{(-j)})}{p(x_{(j)}^{i}|x_{(-j)})f(x_{(j)}^{i}|x_{(-j)})q(x_{(j)}^{i+1}|x_{(-j)})}, 1\right\},$$

if the $x_{(j)}^{i+1}$ element is accepted then the chain moves to draw the block $(j+1)$ having in block (j) the element $x_{(j)}^{i+1}$, on the contrary having in (j) the element $x_{(j)}^{i}$; so at each loop of the Gibbs sampler, it is possible that some elements of x are updated while others are not. Again the natural choice for the transition probability is to use $f(\cdot|\cdot)$ so the acceptance probability reduces to the ratio of the p at the new and old realization of the chain, i.e.

$$\alpha\left(x_{(j)}^{i+1}, x_{(j)}^{i}|x_{(-j)}\right) = \min\left\{\frac{p(x_{(j)}^{i+1}|x_{(-j)})}{p(x_{(j)}^{i}|x_{(-j)})}, 1\right\}.$$

The iteration of steps (1-3) of the above algorithm generate a sample $\{x_i\}$ of a stochastic process which has transition kernel from x to $\widetilde{x}$ equal to

$$\begin{array}{l} K_{IGS}\left(x, \widetilde{x}\right) = \\ \prod_{i=1}^{k} \frac{1}{c_i} p(\widetilde{x}_{(i)}|\widetilde{x}_{(j)} \ \ (j<i)\,,\, x_{(j)} \ \ (j>i)\,) f(\widetilde{x}_{(i)}|\widetilde{x}_{(j)} \ \ (j<i)\,,\, x_{(j)} \ \ (j>i)) \end{array} \tag{5.3}$$

where $c_i \equiv \int p(x_{(i)}|x_{(j)}^{i+1} \ \ (j<i)\,,\ x_{(j)}^{i} \ \ (j>i)) f(x_j|x_{(j)}^{i+1} \ \ (j<i)\,,\ x_{(j)}^{i} \ \ (j>i)) dx_{(i)}$, if the draw in step (2) are performed without using the Metropolis steps; in the presence of the Metropolis step the transition kernel will be equal to

$$\begin{array}{l} K_{IGSM}\left(x, d\widetilde{x}\right) = \\ \prod_{i=1}^{k} \{q(\widetilde{x}_{(i)}|\widetilde{x}_{(j)} \ \ (j<i)\,,\, x_{(j)} \ \ (j>i)\,) \alpha\left(\widetilde{x}_{(i)}, x_{(i)}|\widetilde{x}_{(j)} \ \ (j<i)\,,\, x_{(j)} \ \ (j>i)\right)) d\widetilde{x}_i \\ +(1-\int q(\widetilde{x}_{(i)}|\cdot)\alpha(\widetilde{x}_{(i)}, x_{(i)}|\cdot) d\widetilde{x}_i) \delta_x\left(d\widetilde{x}\right)\} \end{array} \tag{5.4}$$

where δ_x is a Dieracht function to account the case in which the element of the chain does not move.

It is clear from the above description, that the modification of the target density does not imply an increase in the difficulties and computational time of running the sampler given that

all described strategies are quickly implemented and easy to perform. In particular, if the acceptance/rejection method or the Metropolis step are used with auxiliary function $h(\cdot)$ or transition probability $q(\cdot)$ which can be routinely sampled, then the algorithm simply consists in sampling from $h(\cdot)$ or $q(\cdot)$ and then the draw must be accepted or rejected in function of a proper weighting as indicated above.

The prove that the two algorithms deliver the desired result, is given in the following corollary in which it is shown that the transition kernels of the two algorithms have the new measure $f(x)\,p(x)$ as invariant distribution.

Proposition 1

If the transition kernel $K_{IGS}(\widetilde{x}, x)$ and $K_{IGSM}(\widetilde{x}, x)$ are respectively as in (5.3) and (5.4) then $f(x)p(x)$ is proportional to an invariant distribution of those transition kernels.

The proof of the proposition is straight and it is given in the appendix 1. The new measure is an invariant distribution of the constructed chain, however it is clearly desirable that under the new measure, the chain still converges to its invariant distribution and that the estimates still satisfy an ergodic theorem. Various work in the literature provided conditions which assures the ergodicity (Roberts and Smith 1994, Tierney 1994, Athreya, Doss and Sethuraman 1996) or geometric ergodicity (Chan, 1993) of the Gibbs sampler and the same is true for the Metropolis-Hastings algorithm (Mergersen and Tweedie, 1996).

Now we will show that if the augmenting function is sufficiently well behaving, the modification of the chain will not prevent its convergence under mild regularity conditions. Given the description of the algorithms and given a sample $\{x_i\}_{i=1}^{N}$, it is now possible to state the following strong law of large number result

Proposition 2

a) If $D_p \subset D_f$, $E_p|g| < \infty$ and if $\{x_i\}$ is an $pf-$irreducible Markov chain with transition kernels K_{IGS} or K_{IGSM} then

$$\widehat{I} = \frac{\sum_{i=1}^{N} \frac{g(x^i)}{f(x^i,\theta)}}{\sum_{i=1}^{N} \frac{1}{f(x^i,\theta)}} \overset{a.s.}{\to} I$$

and

$$\frac{1}{N}\sum_{i=1}^{N}\frac{1}{f(x^i,\theta)} \overset{a.s.}{\to} \frac{1}{c}.$$

where $c \equiv \int f(x)p(x)dx$.

b) If the density p and f are lower semi continuous at zero, if $\int f(x)dx_i$ and $\int p(x)dx_i \ \forall i$ are locally bounded and D_p is a connected set with $D_p \subset D_f$ then K_{IGS} is $pf-$irreducible and aperiodic.

c) If the chain with transition $\prod_{i=1}^{k} q(\widetilde{x}_{(i)}|\widetilde{x}_{(j)}\ (j<i)\,,\, x_{(j)}\ (j>i))$ is $pf-$irreducible and the probability $\prod_{i=1}^{k}\alpha\left(\widetilde{x}_i, x_i|\widetilde{x}_{(j)}\ \ (j<i)\,,\ x_{(j)}\ \ (j>i)\right)$ is positive for every $\widetilde{x}, x \in D_p$ then K_{IGSM} is $pf-$irreducible.

The proof of the proposition and the definitions are provided in the appendix 1. Three things deserve to be noted. First, in point (a) there is the necessity of computing the constant of integration of the new measure which is unknown due to the fact that it is constructed in a multiplicative way. The conditions of point (b) are extremely weak, the lower semi continuity is an extremely weak condition which is meant to assure probability mass around component points and around points in the domain D, while connectedness rules out peculiar cases in which the density has positive mass on two separate regions of the domain. In point (c), instead, the irreducibility of the chain of interest follows from the irreducibility of the chain generated by the transition probability q, but given that the transition q is chosen this property can be easily verified.

The conditions of proposition (2) are indeed completely operative only in the case in which the functional form of p is known and this is not always true, but they give the idea that the modification of the chain has generally no effect on the convergence property of the chain itself if the augmenting function f is well behaving. However it is possible that the modified chain has better properties than the original one in term of rate of convergence. This is one of the main motivation of our work and we will show later in some examples that a proper choice of the augmenting function can increase the speed to convergence of the sampler and in some cases it will allow even an immediate convergence.

The result of proposition (2) shows that under the new measure the estimates will still satisfy a strong law of large number, however the result does not provide a clear cut criteria to choose the

function f. The lack of a criterion for choosing the augmenting function follows by the fact that the Gibbs kernel is no reversible.[7] In the absence of reversability, known sufficient conditions for central limit theorems are strong and difficult to establish from verifiable fundamentals. In fact in the absence of reversability the estimates of the variance is possible under the uniform ergodicity of the chian which is implied by the uniform boundedness below of the transition kernel and this condition is extremelly hard to hold. Therefore there is not a know sufficient condition for approximation of the variance term of the central limit theorem and this implies that the variance cannot be used as criterion for the proper choice of f.

However two intuitive criteria have to be used as guideline. First element is given by the fact that f has to have sufficient mass on the domain of interest, otherwise this will be reflected in the variance of the estimates. Second consideration, the function f has to be tailored to the function p to undo the correlation among the block of the chain; this general rules have to be specialized on a case by case base.

6. Data Augmentation

In the previous section the idea of running the chain with respect to a modified measure has been proposed but the section concludes without a clear criterion for discriminating among different augmenting function f. A step further can be done in the case of the data augmentation as called by Tanner and Wong (1987), which essentially refers to the case in which there are only two blocks in the Gibbs chain.

The data augmentation is of some interest because on one side it allows to strength the result of the previous section and, on the other side, it covers a variety of interesting models recently proposed in the econometric literature such as the mixture models and the Partially Non Gaussian State Space (PNGSS) models. This terminology has been introduced by Shepard (1994) and it represents a class of nonlinear state space models which have the property that conditional on the realization of a latent regime variable, the model is still linear and gaussian. Various interesting specifications recently proposed belong to this class, such as: the dynamic factor model with regime switching for measuring business cycle proposed by Diebold and Rude-

[7] A revearsible version of the Gibbs chian has been proposed by Roberts (1995).

busch (1994), the model with switching volatility by Kim (1994) and the one analyzed by Billio and Monfort (1995). Section (7.2) provied some applications.

In the data augmentation the integral of interest in (5.1) reduces to the

$$\int g\left(x_{(1)}, x_{(2)}\right) p\left(x_{(1)}, x_{(2)}\right) dx_{(1)} dx_{(2)}$$

where $x_{(1)}$, $x_{(2)}$ are the two blocks which will be sampled. Following the result of the previous section, a change of measure through an augmenting function $f(x_{(1)}, x_{(2)})$ is operated such that

$$\int \frac{g\left(x_{(1)}, x_{(2)}\right)}{f(x_{(1)}, x_{(2)})} f(x_{(1)}, x_{(2)}) p\left(x_{(1)}, x_{(2)}\right) dx_{(1)} dx_{(2)}.$$

The interesting aspect of the data augmentation is that the general results of the previous section assume a particular strength here. In fact it is possible to define a change of measure such that under the new measure the random variables $x_{(1)}$, $x_{(2)}$ are independent and so the subsequent draws of the block of the chain are indeed independent; this is established in the following;

Proposition 3

Let $f(x_{(1)}, x_{(2)}) = \frac{f_1(x_{(1)}) f_2(x_{(2)})}{p\left(x_{(1)}|x_{(2)}\right)}$ where $f_1(x_{(1)})$, $f_2(x_{(2)})$ are two given densities then, respect to the measure $f(x_{(1)}, x_{(2)}) p(x_{(1)}, x_{(2)})$, the variable $x_{(1)}$, $x_{(2)}$ are independent, and the conditional distribution are $f(x_{(1)}|x_{(2)}) p(x_{(1)}|x_{(2)}) \propto f_1(x_{(1)})$ and $f(x_{(2)}|x_{(1)}) p(x_{(1)}|x_{(2)}) \propto f_2(x_{(2)}) \frac{p\left(x_{(2)}|x_{(1)}\right)}{p\left(x_{(1)}|x_{(2)}\right)}$.

The prove is delivered in the appendix 1. The result proves that there is a feasible change of measure such that under the new measure the element of the chain are indeed independent and second it provide a way of simulating from the marginal of the new measure. In fact for one of the two blocks, namely $x_{(1)}$, it consists only in sampling out of the selected function $f_1(x_{(1)})$, while for the second block it is necessary to simulate from $f_2(x_{(2)}) \frac{p\left(x_{(2)}|x_{(1)}\right)}{p\left(x_{(1)}|x_{(2)}\right)}$. This last step can be performed both thought an acceptance/rejection and an adaptive/rejection methods, as described in the previous section. If those methods cannot be implemented, it is possible to use a Metropolis Hasting step but in this case the chain is not any more independent, even if dependence introduced by a Metropolis step can be less sizable then the one generally generated by the Gibbs chain.

So it has been established a transformation of the original problem through a change of measure, under which the elements of the chain are independent, so it has been possible to gain more strength to the result of previous section. Few considerations are worthy.

The possibility of gaining again independence of the blocks or even of reducing the dependence of the blocks would be crucial when the monte carlo integration is a step of a parameter estimation. In fact critical element to prove consistency of the parameter estimates in presence on monte carlo integration to evaluate the objective function is the uniform convergence of the monte carlo estimates of the objective function respect the parameters of interest. But if the monte carlo integration is performed averaging samples coming from a Gibbs sampler then those draws show dependence and the degree of dependence is likely to be function of the same parameters which are supposed to be estimated. This fact can possibly preclude uniform convergence and so the consistency of the estimates. So in that case the possibility of transforming the integral in an integration respect independent draws can be extremely helpful.

Second, the independence of the blocks of the chain under the new measure brings backs the problem into the framework of Geweke (1989) and his results still apply. The asymptotic variance of $\sqrt{N}\left(\widehat{I}-I\right)$ will be given by

$$\sigma^2 \equiv E_{fp}\left(\left(\frac{g(x)}{f(x)}-I\right)^2\right) = c\int (g(x)-I)^2 \frac{p(x)}{f(x)}dx \tag{6.1}$$

where $x=\left(x_{(1)},x_{(2)}\right)$, $c=\int p(x)f(x)dx$ and given a sample $\{x^i\}_{i=1}^N$ from the new measure the variance can be estimated as

$$\widehat{\sigma}^2 = \frac{\sum\limits_i \left(g(x_i)-\widehat{I}\right)^2 \frac{1}{f(x_i)^2}}{\left(\frac{1}{\sqrt{N}}\sum\limits_i \frac{1}{f(x_i)}\right)^2}. \tag{6.2}$$

The estimate of the variance delivers also a ready to apply criteria for the proper choice and parametrization of the function f_1 and f_2, in fact the parametrization of the functions will be choose such that (6.1) will be minimized. However, as Geweke pointed out, a full minimization can be even more cumbersome of the original problem we are trying to solve; a more feasible strategy would be to find a f_1 and f_2 such that variance of the monte carlo integral under the

new measure is reduced compare to the one of $\sqrt{N}\left(\widetilde{I}-I\right)$ under the original measure that it is given by

$$\sigma^2 = E_p\left(g_0(x)-I\right)^2 + 2\sum_{\tau=1}^{\infty} E_p\left(g_0(x)-I\right)\left(g_\tau(x)-I\right)$$

where τ in g_τ means the distance of different realizations in the chain.[8]

However even if the choice of the optimal f_1 and f_2 is important, we share the view of Gilks and Roberts (1996) which regards rapid mixing as a higher ideal than small standard errors and so the search for the optimal f_1 and f_2 can be not worthy, in particular now that under the new measure the blocks of the chain are indeed independent.

Finally note is that this approach is in some sense similar to the recent work of Elliot et al. (1996) on change of measure to solve nonlinear and non gaussian state space model. So it is based on defining a new measure under which the model is *iid* and the filtering can be easily performed and then the filtered result is again transformed into the original one. Here the idea is somehow similar, given that we perform a change of measure to run the sampler in a more efficient way.

7. Application

In the previous section, it has been proposed the idea that a proper reparametrization of the Gibbs sampler chain can be worthy. The idea has been motivated in term of gaining better mixing properties and in term of lower variance of the sample estimates. Here two applications of the method are presented. The first one is the smoothing of the unobservable conditional variance in a stochastic volatility model as an application of the general method of section (5). While the second application is the smoothing of the latent variable in a Partially Non Gaussian State Space model.

[8] The estimation of the variance under the original measure is possible due to the reversablity property of the data augmentation, property which is not share in general by the Gibbs sampler.

7.1. Stochastic volatility models

Here we aim to show an application of the change of measure in the smoothing of the underlying variance in stochastic volatility models (SV here on). Stochastic volatility models arise naturally as discrete approximation of various diffusion processes of interest in the continuous asset pricing literature. But they found limited application mainly due to the difficulties in handling them with respect to alternative ways of modelling the conditional volatility (see Ghysel, Harvey and Renault (1995) for a survey). However basic $GARCH$ models do not have some of the interesting features which are present in the SV models such as better matching of the second moment properties of data on returns, the higher degree of excess kurtosis and the possibility of leverage effect.

Recently the issue of prediction and inference in these models has found large attention as a natural application of Markov chain methods. The stochastic volatility models has become a battleground for the application of different MCMC techniques; among the others Jacquier, Polson and Rossi (1994) (JPR, here on), Kim and Shephard (1994) and Shephard (1994). Jacquier, Polson and Rossi used a combination of Gibbs sampling and Metropolis independent chain to perform Bayesian analysis on the parameters and the underling stochastic volatility. Kim and Shephard argue that the multi steps algorithm of Jacquier, Polson and Rossi can be very slow for some parametrization and they proposed a griddy Gibbs sampler which consists in approximating the non normal distribution of the disturbance in the measurement equation by a mixture of normal which allows to perform the sampler in a single move[9].

Let consider the same simple stochastic volatility model of JPR and Shephard, which is given by

$$\begin{aligned} \varepsilon_t &= h_t^{1/2} u_t \qquad (7.1) \\ \ln h_t &= \alpha + \delta \ln h_{t-1} + \sigma_v v_t \\ & \qquad \begin{pmatrix} u_t \\ v_t \end{pmatrix} \sim NID(0, I_2), \end{aligned}$$

the sample is of dimension T, h_T indicates the whole vector of unobservable volatility and

[9] Mahieu and Schotman (1994) extended this approach by introducing more degree of freedom in the mixture of normals where the parameters are estimated rather then fixed a priori.

$\theta = \{\alpha, \delta, \sigma_v\}$ are the parameters that are considered given for the purpose of the exercise. The two innovations are assumed independent. Consider the problem of smoothing the unobservable volatility given the observed innovation ε_T, so the problem consists in computing the following integral

$$E(h_T|\varepsilon_T, \theta) = \int h_T p(h_T|\varepsilon_T, \theta) dh. \tag{7.2}$$

The difficulties arise by the fact that the density $p(h_T|\varepsilon_T, \theta)$ is known only up to a constant and there are no obvious methods to sample directly form it. In fact

$$\begin{aligned} p(h_T|\varepsilon_T, \theta) &= \frac{p(h_T, \varepsilon_T|\theta)}{p(\varepsilon_T|\theta)} = \frac{p(\varepsilon_T|h_T, \theta)p(h_T|\theta)}{p(\varepsilon_T|\theta)} \\ &\propto p(\varepsilon_T|h_T, \theta)p(h_T|\theta) \\ &\propto \prod_{t=1}^{T} h_t^{-3/2} \exp(-\frac{\sum_{t=1}^{T} \varepsilon_t^2}{2h_t}) \exp(-\frac{\sum_{t=1}^{T} (\ln h_t - \alpha - \delta \ln h_{t-1})^2}{2\sigma_v^2}) \end{aligned} \tag{7.3}$$

A solution proposed in the literature is to consider a Markov Chain method which draws each element of h_T one at the time (multi steps sampler) and so it constructs a sequence of $\{h_T^i\}_{i=1}^N$, where N is the number of iteration of the sampler. To implement this strategy, the starting point is the conditional density of the individual element of the vector h_T which, for $t = 2, ..., T-1$, is given as

$$p(h_t|h_{(-t)}, \varepsilon_T, \theta) \propto h_t^{-3/2} \exp(-\frac{\varepsilon_t^2}{2h_t}) \exp(-\frac{(\ln h_t - \mu)^2}{2\sigma^2}) \tag{7.4}$$

where $\mu = \frac{\alpha(1-\delta)+\delta(\ln h_{t-1}+\ln h_{t+1})}{1+\delta^2}$ and $\sigma^2 = \frac{\sigma_v^2}{1+\delta^2}$.

To draw from the above density kernel two alternative are feasible. The first one is to use a transformation of variables and to work directly with the conditional distribution of $H_t = \ln(h_t)$, as

$$p(H_t|H_{(-t)}, \varepsilon_T, \theta) \propto \exp(-\frac{\varepsilon_t^2}{2} \exp(-H_t)) \exp(-\frac{(H_t - \mu^*)^2}{2\sigma^2}) \tag{7.5}$$

where $\mu^* = \mu - .5\sigma^2$, and to sample H_t either by acceptance/rejection or by the adaptive rejection methods of Gilks and Wind (1992). Alternatively, it is possible to apply the JPR algorithm consisting in introducing a Metropolis independence step into the multi move Gibbs sampler. All these different methods of sampling will produce a multi steps sampler whose iteration will generate a chain having as invariant the distribution of interest (7.3).

Shephard and Kim, following the result of Lui, Kong and Wong (1994), argued that due to the high correlation of the elements of h_T, it is likely that sampling the elements separately will induce a slow mixing and it will affect the calculation of the smoothed value of h. To avoid this they introduced the mixture approximation for the distribution of $ln(\varepsilon_t^2)$ which allows to speed up the sampler by drawing all the elements of h_T vector all at once. The drawbacks of this method is twofold; on one side it is an approximation and it is likely that, as pointed out by JPR, the mixture of normal is not a good proxy, in particular in term of tail behavior, and, on the other side, the proposed method increases the dimension of the integral and it is clearly necessary that all the states of the mixture are visited a sufficient number of time by the sampler to assure a good performance of the algorithm.

However it is a fact that the high dependence of the blocks can jeopardize the performance of the sampler. In the stochastic volatility, high dependence of the blocks is associated with values of δ close to unity and σ_v close to zero and in many applications of interest the estimated values of the parameters end up in those regions.

To reach a better understanding of the effect of slow convergence, two experiments are reported by running the multi steps sampler using an acceptance/rejection algorithm to draw from the conditional distributions in (7.5) with a normal density as envelope density with mean and variance set to maximize the acceptance probability[10]. The first experiment reports the average of 1000 independent sampler each with initial condition $h_T^0 = 1$ ($T = 100$) and its aim is to show the last of the initial conditions in the sampler. The second is based on a single multi steps sampler and the realizations of the draw of h_{50} are used to compute its ACF function (50.000 replication are used, after discarding the first 10.000). Figure (7.1) and (7.2) reports the two set of experiments for two different parameters: $\delta = .9$, $\sigma_v = .2$, $\alpha = 0$; $\delta = .99$, $\sigma_v = .1$, $\alpha = 0$.

[10] The algorithm is a modification of the one given by Geweke (1995) and by Kim, Shephard and Chib (1996), in which different mean and variance of the normal envelope have been setted. Kim, Shephard and Chib motivated their choice in term of gaining higher acceptance rate for tail observations.

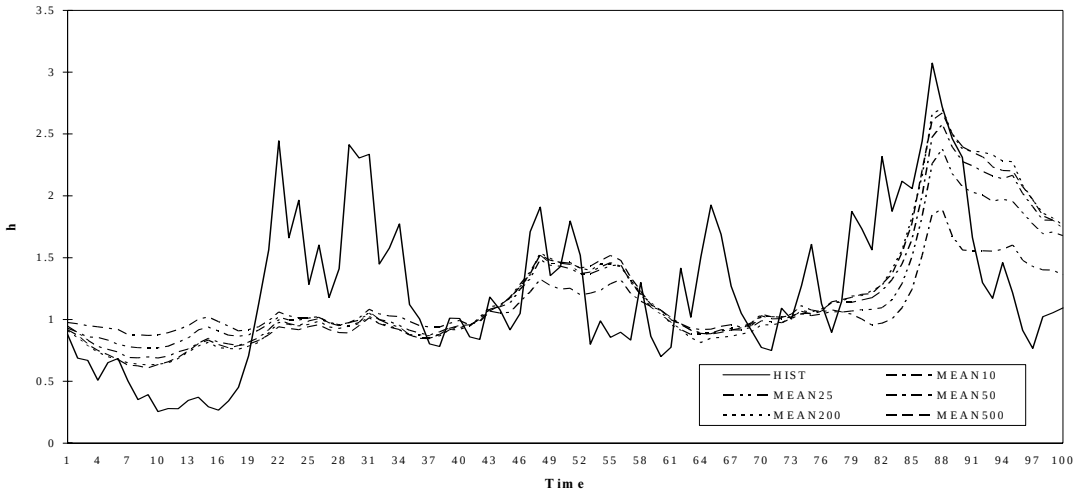

Figure 7.1.1: Average of 1000 independent sampler.

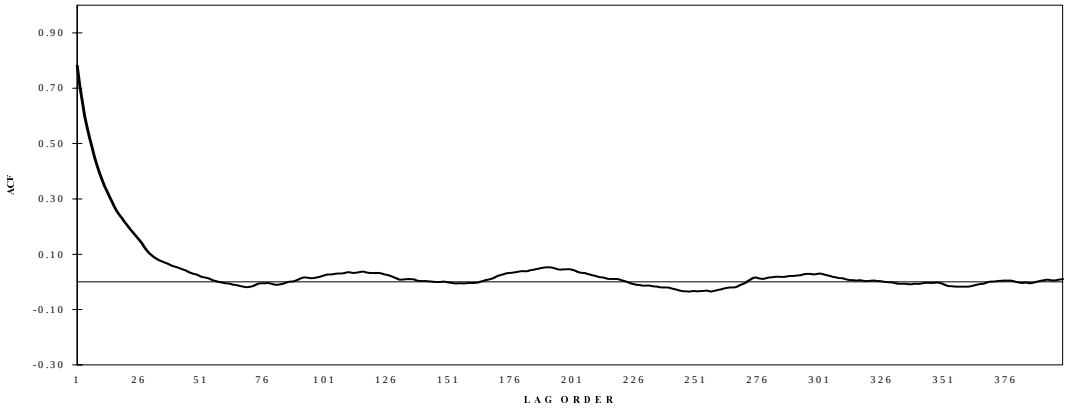

Figure 7.1.2: Autocorrelation function.

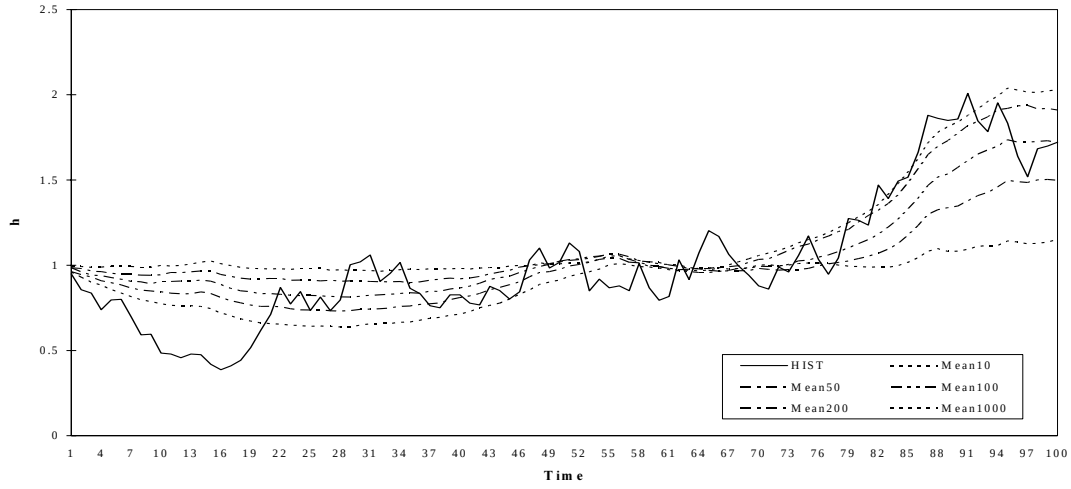

Figure 7.2.1: Average of 1000 independent sampler.

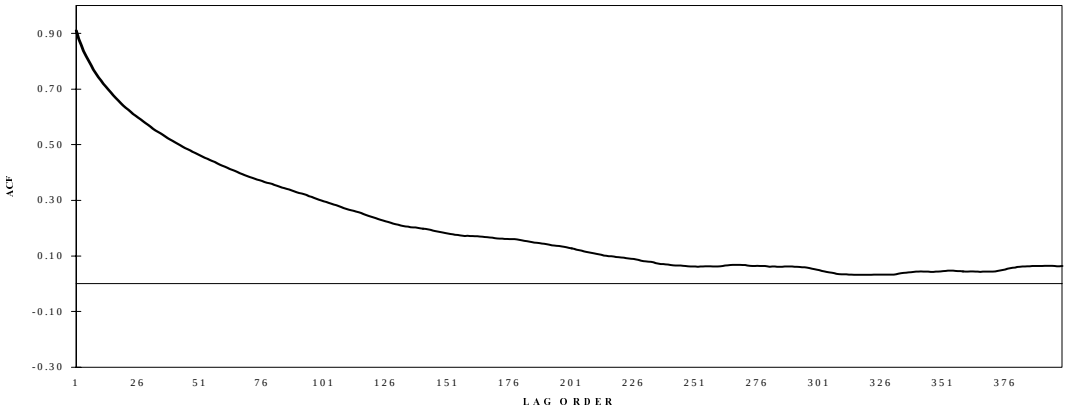

Figure 7.2.2: Autocorrelation function.

As pointed by Shephard and as it is clearly evident from the above figures, the dependence in the sampler is large and persistent as the last of the initial conditions and the ACF function show. This can affect the performance of the use of the sampler in various directions, among the others, the standard deviation of the sample estimates and the applicability in the parameters estimation.

Here the change of measure method proposed in section (5) is applied to the above example as an other way of handling the problem of the dependence. Consider the expectation in (7.2)

respect to $H_T = \ln(h_T)$, the expression under the change of measure can be written as

$$E(h_T|\varepsilon_T, \theta) = \int \frac{\exp(H_T)}{f(H_T|\theta_0)} p(H_T|\varepsilon_T, \theta) f(H_T|\theta_0) dH. \tag{7.6}$$

In the choice of the augmenting function, it has been looked for a density with sufficient mass on the domain of p and such that the conditional of the f would be useful in undoing the dependence present in the conditional distributions under the original measure. For this reason, a multivariate student t with degree of freedom ν_0 has been chosen as augmenting function with parameters $\theta_0 = (\delta_0, \sigma_0^2)$ as[11]

$$f(H_T|\nu_0, \theta_0) \propto (1 + \frac{H_T' V H_T}{\nu_0})^{-\frac{T+\nu_0}{2}} \tag{7.7}$$

where $V = \frac{1}{\sigma_0^2} RR'$ and R is a $T \times T$ matrix as

$$R = \begin{bmatrix} \sqrt{1-\delta_0^2} & 0 & \cdots & \cdots & 0 \\ -\delta_0 & 1 & 0 & & \vdots \\ 0 & -\delta_0 & 1 & 0 & \vdots \\ \vdots & 0 & -\delta_0 & 1 & 0 \\ 0 & \cdots & 0 & -\delta_0 & 1 \end{bmatrix}$$

The nice property of the multivariate t is that the conditional distribution $f(H_t|H_{(-t)}, \nu_0, \theta_0)$ is univariate student t with $(T + \nu_0 - 1)$ degrees of freedom and mean equal to $m_t = \frac{\delta_0}{1+\delta_0^2}(H_{t+1} + H_{t-1})$ and variance $s^2 = \frac{\nu_0}{T+\nu_0-3} \frac{(1+q)\sigma_0^2}{1+\delta_0^2}$ where $q = H_T' V H_T - \frac{\sigma_0^2}{1+\delta_0^2}(H_t - m_t)^2$. The fact that the conditional mean of the function f is of the same form of the mean of the conditional density under the original measure, is the reason why this specification has been chosen.

Under the new measure, the sampler will have invariant distribution $p \cdot f$ and it will be run by drawing from the distribution of H_t conditional $H_{(-t)}$ on respect to the new measure,

$$\begin{gathered} p(H_t|H_{(-t)}, \varepsilon_T, \theta) f(H_t|H_{(-t)}, \theta_0) \propto \\ \exp(-\tfrac{\varepsilon_t^2}{2} \exp(-H_t)) \exp(-\tfrac{(H_t-\mu^*)^2}{2\sigma^2})(1 + \tfrac{1}{(T+\nu_0-1)s^2}(H_t - m_t))^{-\frac{T+\nu_0}{2}}. \end{gathered} \tag{7.8}$$

The draws from the above distribution can be performed by a Metropolis independence step using as auxiliary function a student t distribution with degrees of freedom ν_q, fixed mean μ_q and variance σ_q, as indicated in section (5).

[11] A good reference on multivariate t is appendix B in Zellner (1971).

To compare the property of the described change of measure, the experiments performed before are repeated with the following parameters specification: in the first case $\delta_0 = -.9$, $\sigma_0 = .2$, $\nu_0 = 5$, $\nu_q = 5$, $\mu_q = 0$ and $\sigma_q = .5$; in the second one $\delta_0 = -.99$, $\sigma_0 = .1$, $\nu_0 = 5$, $\nu_q = 5$, $\mu_q = 0$ and $\sigma_q = .5$. The parameters value has been chosen so to undo the dependence present under the original measure between the elements of the chain and therefore the means of the densities $p(\cdot|\cdot)$ and $f(\cdot|\cdot)$ have been settled equal in absolute value but opposite in sign. The choice of the mean and variance of the transition distribution are be chosen so to have a proper envelope of the product of the two conditional distributions in (7.8).

The result of the experiment are reported in the following figures (7.3) and (7.4).

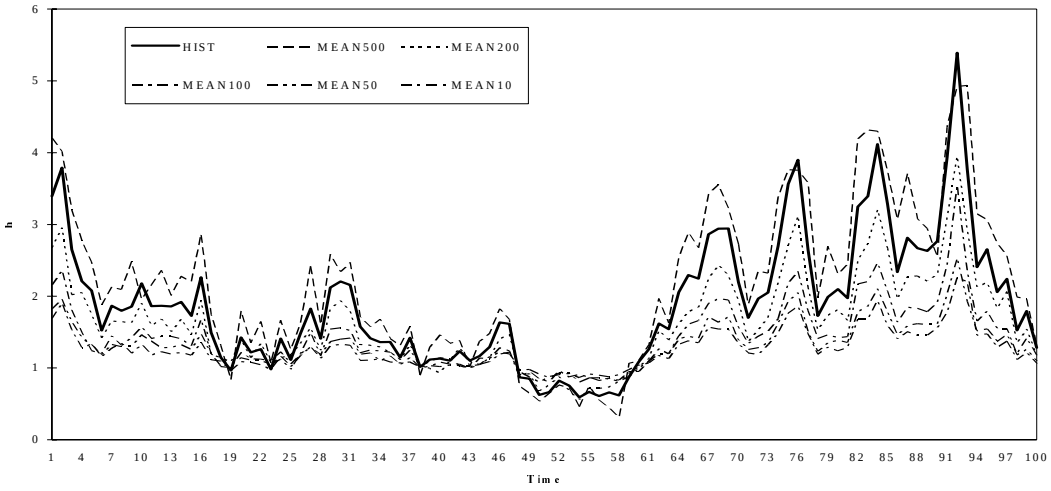

Figure 7.3.1: Average of 1000 independent sampler.

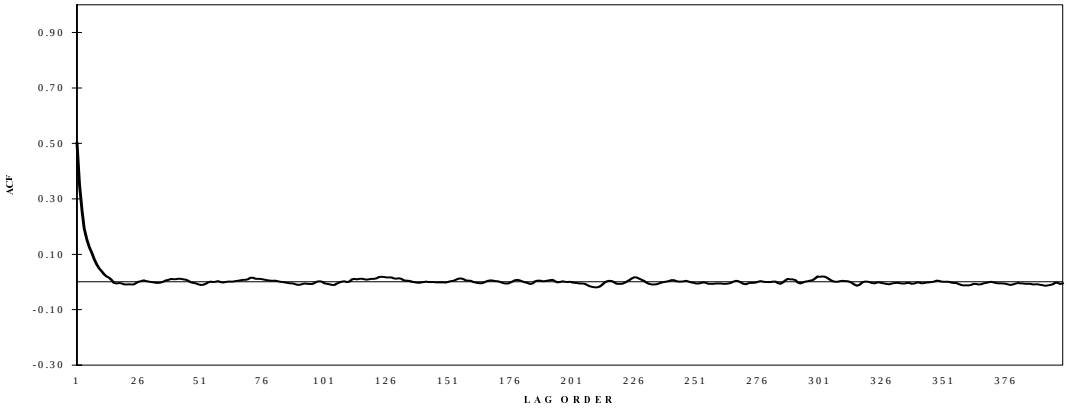

Figure 7.3.2: Autocorrelation function.

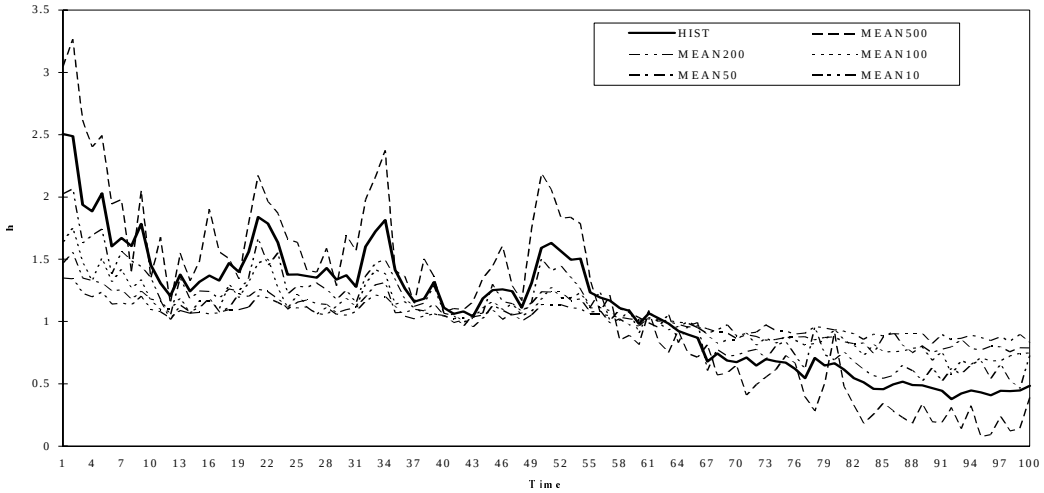

Figure 7.4.1: Average of 1000 independent sampler.

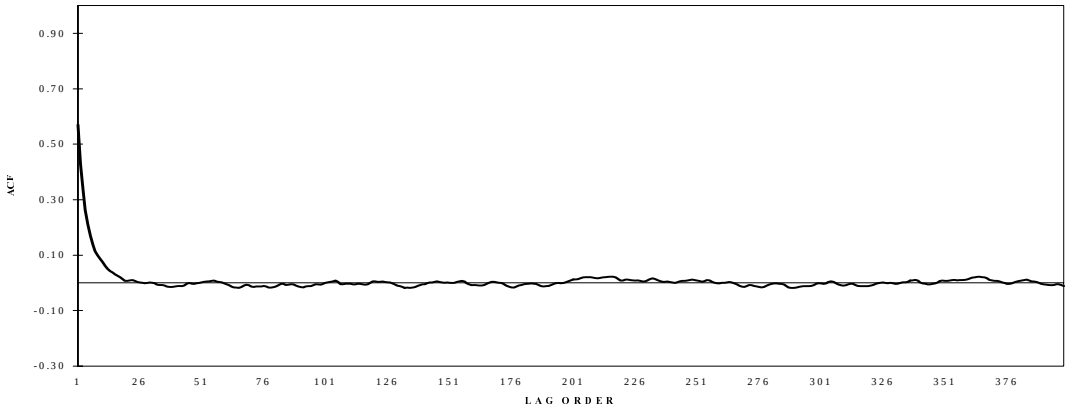

Figure 7.4.2: Autocorrelation function.

By comparing these results to those of figure 7.1 and 7.2, it is evident that the chain under the new measure has much lower dependence and the initial condition lasts less. These properties allow to compute the integral in (7.6) by running the chain in a multi step mode but with higher precision and lower degree of dependence than the original one and without using any approximation or increasing the dimension of the integral.

7.2. Partially Non Gaussian State Space Model

This definition is due to Shephard (1994) and it refers to a class of state space models in which it is possible to introduce nonlinearity ad non gaussianity, while still preserving some ease in the analysis. The model can be represent as

$$\begin{aligned} y_t &= \mu(s_t) + A(s_t)z_t + B(s_t)\varepsilon_t \\ z_t &= \nu(s_t) + C(s_t)z_{t-1} + D(s_t)\eta_t \end{aligned} \tag{7.9}$$

where y is the observable variable of dimension m, z is an unobservable state $(n \times 1)$ and s is a scalar Markov chain process defining the regime. The innovation ε and η are independent standard gaussian. The interesting aspect of the model is that conditional on the realizations of the regime variable, s, the model is still linear and gaussian while this is not true unconditional of the regime.

Different specification recently proposed in the literature can be reconnected to the above framework, among the others: the nonlinear dynamic factor model for measuring business cycle of Diebold and Rudebusch (1994), the model of conditional switching variance of Kim (1994) or the discrete analog of the continuous models for trading time deformation in finance by Ghysel, Gourieroux and Jasiak (1996).

The difficulties in the estimation and filtering of this model is due to the fact that the joint distribution of the states Z and the regimes S conditional on the observables is of unknown form. In the literature, different estimation and filtration scheme have been proposed. Kim (1994) introduced an approximation filter based on cutting the memory of the filter itself about the past history of the regime after a fixed number of lags. Kim's filter proved to work well in models in which the state in time t depends only on a finite number of past realization of the regime. Billio and Monfort (1995) derived a filtering and smoothing algorithm based on finding a sequence of important sampling functions which optimize the performance of the integration of the likelihood respect to the unobservable at one and two steps ahead; however their method proved to work only in the case in which the regime variable does not cause the state variable in a Granger sense.

Despite the intractability of the joint density of the states and the regimes, it has to be noted that the conditional distribution of the states w.r.t. the regime and the observable and

the conditional of the regime w.r.t. the states and the observable have a much friendly form and it is feasible to simulate them. So it is evident that another possible methods to approach the problem of estimation and inference in this class of models is to use a Gibbs sampler, as it has been suggested by Shephard (1994) and Carter and Kohn (1994). In fact the Gibbs sampler can generate draws from the join distribution $p(Z_T, S_T|Y_T)$, which can be used in the computation of the likelihood and in the smoothing of the state. The sampler consists in performing sequential draws from $p(Z_T|S_T, Y_T)$ and $p(Z_T|S_T, Y_T)$.

First, the density of the states conditional on the regimes and the observable is given by

$$p(Z_T|Y_T, S_T) = p(z_T|Y_T, S_T) \prod_{t=1}^{T-1} p(z_t|Y_T, S_t, z_{t+1}) \tag{7.10}$$

where the capital letter means all the history up to time t. The density $p(z_T|Y_T, S_T)$ is given by running the Kalman filter up to time T, while $p(z_t|Y_T, S_t, z_{t+1})$ can be worked out by running the filter backward; so all the densities in (7.10) are normals and can be routinely sampled. However the backward filter implies the computation of the inverse matrix of the forecast variance of the states which can be computationally problematic if the draws are repeated extensively; Dejong and Shephard (1994) proposed a simulation smoother which allows to sample from (7.10) without using matrix inversion. The simulation smoother of Dejong and Shephard will be used in the example proposed.

The second step consists in simulating from the density of the regime conditional to the states and the observables, $p(S_T|Y_T, Z_T)$. In general, this is not an easy task but in the case in which s_t is a scalar Markov chain the algorithm described by Carter and Kohn (1994) allows to sample directly all the vector S_T from the conditional $p(S_T|Y_T, Z_T)$[12].

So given the structure of the model, the sampler is a data augmentation one based on drawing sequentially the regime vector from $p(S_T|Y_T, Z_T)$ and the state vector from $p(Z_T|Y_T, S_T)$.

Note that here we confine ourself to work in the case of qualitative regime which is also the case considered in all applications of the PNGSS models previously indicated; however this does not mean that the framework is not applicable to more general specification; but in those cases the sampling of the state S will not be possible in a single move, as the algorithm of Carter and Kohn allows, and the sampler would not reduce to a data augmentation scheme.

[12] This algorithm is described in appendix 2.

Having obtained random samples from $p\left(Z_T, S_T|Y_T\right)$, these samples can be used to compute the smoothed expectation of the states or the realization can be used in the expectation step of the Simulated EM algorithm in the estimation of the parameters as indicated by Shephard and Carter and Kohn. Suppose that the aim is to compute the smoothed value of the state

$$\begin{aligned} z_{t|T} &= \int\int z_t p\left(Z_T, S_T|Y_T\right) dZdS \\ &= \int E(z_t|S_T, Y_T) p\left(S_T|Y_T\right) dS \end{aligned} \tag{7.11}$$

then the realizations $\left\{Z_T^i, S_T^i\right\}_{i=1}^N$ of the sampler can be used to estimate (7.11) either by the empirical estimator

$$\widetilde{z}_{t|T} = \frac{1}{N}\sum z_t^i \tag{7.12}$$

or, more efficiently, by the Rao-Blackwell form

$$\widetilde{z}_{t|T}^{RB} = \frac{1}{N}\sum E(z_t|Y_T, S_T^i). \tag{7.13}$$

Following the consideration in the previous section, the chain generated by sampling sequentially the states and the regimes vector will show dependence. As stressed in various occasion the dependence is not desirable and it can induce large variance of the estimates. Following the result of section (6), it is possible to define a change of measure such that under it the state and the regime variable are indeed independent. So consider an augmenting function of the type

$$f\left(Z_T, S_T|Y_T\right) = \frac{f_1\left(Z_T|Y_T\right) f_2\left(S_T|Y_T\right)}{p\left(Z_T|S_T, Y_T\right)}$$

where the function f_1 and f_2 are properly chosen.. Appling the above change of measure into (7.11), it follows that

$$z_{t|T} = \int \frac{E(z|S_T, Y_T)}{f_2\left(S_T|Y_T\right)} f_2\left(S_T|Y_T\right) p\left(S_T|Y_T\right) dS;$$

the sample from $f_2\left(S_T|Y_T\right) p\left(S_T|Y_T\right)$ can be drawn by a Metropolis independent step given that

$$f_2\left(S_T|Y_T\right) p\left(S_T|Y_T\right) \propto \frac{f_2\left(S_T|Y_T\right) p\left(S_T|\overline{Z}_T, Y_T\right)}{p\left(\overline{Z}_T|Y_T, S_T\right)}$$

where $\overline{Z}$ is any given history and the independent step can be performed by using as auxiliary function the $p\left(S_T|\overline{Z}_T, Y_T\right)$ itself. In this case the i iteration of the algorithm consists in

simulating S_T^i from $p\left(S_T|\overline{Z}_T,Y_T\right)$ and then accepting the draw with probability

$$\alpha(S_T^i, S_T^{i-1}) = \min(\frac{\frac{f_2(S_T^i|Y_T)}{p(\overline{Z}_T|Y_T,S_T^i)}}{\frac{f_2(S_T^{i-1}|Y_T)}{p(\overline{Z}_T|Y_T,S_T^{i-1})}}, 1).$$

The iteration of the Metropolis step will generate a sample $\{S_T^i\}_{i=1}^N$ of regime vector from which the smoothed realizations are computed as

$$\widehat{z}_{t|T} = \frac{1}{N}\sum \frac{E(z_t|Y_T, S_T^i)}{f_2\left(S_T^i|Y_T\right)}, \tag{7.14}$$

it is likely that the estimated $\widehat{z}_{t|T}$ will show lower variance and the draws of S_T^i lower dependence than the one in (7.12) and (7.13).

The above methodologies are present in the following contest; consider the univariate specification of the model in (7.10)

$$\begin{aligned} y_t &= \mu + \alpha z_t + \sigma\varepsilon_t \\ z_t &= \nu_1 s_t + c z_{t-1} + (d_0 + d_1 s_t)\eta_t \end{aligned}$$

where ε_t, η_t are independent standard normal innovations and s_t is a zero-one two state markov process with transition probability $P(s_t = 0|s_{t-1} = 0) = p$ and $P(s_t = 1|s_{t-1} = 1) = q$. The parametrization is the following. The transition probabilities are $p = .9$ and $q = .7$, so there is a high persistent state and a low persistent state. In the observation equation $\mu = .5$, $\alpha = .7$ and $\sigma = .3$; while in the state equation $\nu_1 = .4$, $c = .8$, $d_0 = .6$ and $d_1 = .2$.

Given the parameters the smoothed value of the state is computed as above, however at this stage the augmenting function f has assumed constant aiming to highlight only the gain coming from a lower dependence of the chain if the sampler is run following the result of proposition 3. All the three possible methodologies has been performed as indicated in (7.12), (7.13) and in (7.14); as a benchmark also the smoothed value through the Kim's filter has been performed.

The result of the experiment over a sample of 100 observations are reported in the following tables. In table 1.1 the average RMSE of the three estimators over the generated samples are reported together with the RMSE of the Kim's filtered value

in the case in which the sampler has run for 5000 replication after other 2000 of warming up. Note that for the Metropolis independence step the acceptance rate has been around 60 per cent.

	RMSE
$\widetilde{z}_{t\|T}$	0.702
$\widetilde{z}^{RB}_{t\|T}$	0.506
$\widehat{z}_{t\|T}$	0.355
Kim's filter	0.389

Table 7.1: Root Mean Squared Error

	Gibbs	Metropolis
$t = 25$	0.179	0.143
$t = 50$	0.265	0.202
$t = 75$	0.133	0.115
$t = 100$	0.183	0.152

Table 7.2: Asymptotic Variance

While table 1.2 reports the estimates of the asymptotic variance of $s_{t|T}$ for $t = 25, 50, 75, 100$ over the realization of the sampler in the case of the plain Gibbs sampler and in the case of the Metropolis algorithm

The proposed method clearly outperform the Gibbs one both in term of RMSE and in term of lower dependence in the simulated variables; however the result follows just by the use of a Metropolis step and more improvement are possible designing a proper auxiliary function f.

8. Conclusions

In this work the idea of change of measure in performing Markov chain monte carlo integration is investigated. In particular the focus has been posed on the Gibbs sampler and how the change of measure modifies the algorithm to simulate from the target density.

In the Markov chain monte carlo the change of measure has a twofold importance. On one side, as in the independent monte carlo, the change of measure can induce a lower contemporaneous variance of the estimates if there is a proper choice of the augmenting function. On the other side, the modification of the target density can improve the mixing properties of the

Markov chain with respect to the chain under the original measure.

The idea has been specialized to two cases of interest, the stochastic volatility and the nonlinear state space model, which illustrate the possible gain in working under the change of measure. However, even if the proposed methods have shown to have comparative advantage compared of working under the original measure, the problem of the choice of the augmenting function is still an open one and the function f has to be tailored on a case by case basis.

9. Appendix 1

Proof of proposition 1.

To short the notation define $h(x) = \frac{f(x)p(x)}{c}$, $p_M(\cdot|\cdot) = q(\cdot|\cdot)\alpha(\cdot|\cdot)$ and $r(\cdot|\cdot) = (1 - \int q(\cdot|\cdot)\alpha(\cdot|\cdot)d\widetilde{x}_j)$. Consider first the case of K_{IGS} and without lose of generality set $k = 3$. The proof follows by the fact that $h(x) = \frac{p(\widetilde{x}_i|\widetilde{x}_{(-i)})f(\widetilde{x}_i|\widetilde{x}_{(-i)})}{c_i} h(x_{(-i)})$ where $h(x_{(-i)})$ is the marginal of the subvector $x_{(-i)}$. In fact

$\int K_{IGS}(x, \widetilde{x})\, h(x)dx =$

$\int \prod_{i=1}^{3} \frac{1}{c_i} p(\widetilde{x}_{(i)}|\widetilde{x}_{(j)} \ (j < i),\ x_{(j)} \ (j > i)) f(\widetilde{x}_{(i)}|\widetilde{x}_{(j)} \ (j < i), x_{(j)} \ (j > i)) h(x)dx =$

$\frac{p(\widetilde{x}_{(3)}|\widetilde{x}_{(-3)})f(\widetilde{x}_{(3)}|\widetilde{x}_{(-3)})}{c_3}$

$\int \frac{p(\widetilde{x}_{(2)}|\widetilde{x}_{(1)}, x_{(3)})f(\widetilde{x}_{(2)}|\widetilde{x}_{(1)}, x_{(3)})}{c_2} \int \frac{p(\widetilde{x}_{(1)}|x_{(j)} \ j>1) f(\widetilde{x}_{(1)}|x_{(j)} \ j>1)}{c_1}$

$\int h(x_{(1)}|x_{(j)} \ j > 1)dx_{(1)} h(x_{(2)}|x_{(3)})dx_{(2)} h(x_{(3)})dx_{(3)} =$

$\frac{p(\widetilde{x}_{(3)}|\widetilde{x}_{(-3)})f(\widetilde{x}_{(3)}|\widetilde{x}_{(-3)})}{c_3}$

$\int \frac{p(\widetilde{x}_{(2)}|\widetilde{x}_{(1)}, x_{(3)})f(\widetilde{x}_{(2)}|\widetilde{x}_{(1)}, x_{(3)})}{c_2} \int h(\widetilde{x}_{(1)}|x_{(j)} \ j > 2) h(x_{(2)}|x_{(3)})dx_{(2)}$

$h(x_{(3)})dx_{(3)} =$

$\frac{p(\widetilde{x}_{(3)}|\widetilde{x}_{(-3)})f(\widetilde{x}_{(3)}|\widetilde{x}_{(-3)})}{c3} \int h(\widetilde{x}_{(1)}, \widetilde{x}_{(2)}|x_{(3)}) h(x_{(3)})dx_{(3)} = \frac{p(\widetilde{x}_{(3)}|\widetilde{x}_{(-3)})f(\widetilde{x}_{(3)}|\widetilde{x}_{(-3)})}{c_3} h(\widetilde{x}_{(-3)}) = h(\widetilde{x})$

In the case of K_{IGSM} the prove will be given for $k = 2$ to avoid excess of notation; the kernel is

$$\int K_{IGSM}\left(x,\widetilde{x}\right)h(x)dx=$$

$$\int\prod_{i=1}^{2}\left\{p_M(\widetilde{x}_i,x_i|\widetilde{x}_{(j)}\left(j<i\right),x_{(j)}\left(j>i\right))+r(x_i|\widetilde{x}_{(j)}\left(j<i\right),x_{(j)}\left(j>i\right))\delta_x\left(d\widetilde{x}\right)\right\}h(x)dx=$$

$$=\int p_M(\widetilde{x}_1,x_1|x_2)p_M(\widetilde{x}_2,x_2|\widetilde{x}_1)h(x)dx+\int r(x_1|x_2)r(x_2|\widetilde{x}_1)\delta_x\left(d\widetilde{x}_2\right)\delta_x\left(d\widetilde{x}_1\right)h(x)dx$$

$$\int r(x_1|x_2)p_M(\widetilde{x}_2,x_2|x_1)h(x)\delta_x\left(d\widetilde{x}_1\right)dx+\int p_M(\widetilde{x}_1,x_1|x_2)r(x_2|\widetilde{x}_1)\delta_x\left(d\widetilde{x}_2\right)h(x)dx=$$

$$=\int p_M(x_1,\widetilde{x}_1|x_2)h(\widetilde{x}_1|x_2)p_M(\widetilde{x}_2,x_2|\widetilde{x}_1)h(x_2)dx+r(\widetilde{x}_1|\widetilde{x}_2)r(\widetilde{x}_2|\widetilde{x}_1)h(\widetilde{x})+$$

$$\int r(x_1|x_2)p_M(x_2,\widetilde{x}_2|x_1)h(\widetilde{x}_2|x_1)h(x_1)\delta_x\left(d\widetilde{x}_1\right)dx+$$

$$\int p_M(x_1,\widetilde{x}_1|x_2)r(x_2|\widetilde{x}_1)h(\widetilde{x}_1|x_2)h(x_2)\delta_x\left(d\widetilde{x}_2\right)dx=$$

$$=h(\widetilde{x})\int p_M(x_1,\widetilde{x}_1|x_2)p_M(x_2,\widetilde{x}_2|\widetilde{x}_1)dx+r(\widetilde{x}_1|\widetilde{x}_2)r(\widetilde{x}_2|\widetilde{x}_1)h(\widetilde{x})+$$

$$h(\widetilde{x})\int r(\widetilde{x}_1|x_2)p_M(x_2,\widetilde{x}_2|x_1)dx+h(\widetilde{x})\int p_M(x_1,\widetilde{x}_1|x_2)r(\widetilde{x}_2|\widetilde{x}_1)dx=h(\widetilde{x})$$

The fourth equality holds for the reversability property of each of the single metropolis steps $p_M(x_1,\widetilde{x}_1|x_2)h(\widetilde{x}_1|x_2)=p_M(\widetilde{x}_1,x_1|x_2)h(x_1|x_2)$, while the last equality derive by the definition of r. ■

Before proof of proposition 2, some definitions are needed.

A kernel K is π-irreducible if, for all $x\in D=\{x:\pi(x)>0\}$, $\pi(A)>0$ implies that for some $t\geq 1$ $P(x_t\in A|x_0=x)>0$.

An m-cycle for an irreducible chain with transition kernel K is a collection of disjoint sets $\{E_0,E_1,...,E_{m-1}\}$ such that $\Pr(x,E_j)=1$ for $j=(i+1)\bmod(m)$ and all $x\in E_i$. The period d is the largest m for which an m-cycle exists. The chain is aperiodic if $d=1$.

A function $f:\Re^n\to\Re$ is lower semicontinuos at 0 if, for all x with $h(x)>0$, there exists an open neighborhood $N_x\ni x$ and $\varepsilon>0$ such that for all $y\in N_x$ $h(y)\geq\varepsilon>0$.

Proof of proposition 2.

a) Given corollary (1) and the $pf-$irredibility, the result follows by theorem (4.3) in Tierney (1996) and by the fact that if a $g(x)$ is continuous and $x_n \xrightarrow{as} x$ then $g(x_n) \xrightarrow{as} g(x)$.

b) The fact that the original density and the augmenting one are lower semicontinuous implies that also pf is lower semicontinuous. The locally boundedness of $\int p(x)f(x)dx_i \ \forall i$ follows by the boundedness of the single functions and by the fact that $p(x)f(x) = p(x_i|x_{(-i)})$ $f(x_i|x_{(-i)})p(x_{(-i)})f(x_{(-i)})$. Finally given that $D_p \subset D_f$ then the condition of theorem (2) of Roberts and Smith (1994) are satisfied and the kernel K_{IGS} is $pf-$irreducible and aperiodic.

c) Let K_q the transition kernel relative to the functions q and let $\alpha(\widetilde{x}, x) = \prod_{i=1}^{k} \alpha(\widetilde{x}_i, x_i|\widetilde{x}_{(j)}$ $(j < i)\,,\ x_{(j)}\ (j > i))$ be the probability that all the blocks of the chain move. The prove is similar to theorem (3) of Roberts and Smith and it is based on showing that the irreducubility property of K_q implies the irreducibility of $K_{IGSM.}$ Let define $V_x^{(t)} = \left\{\widetilde{x} : K_q^{(t)}(x, \widetilde{x}) > 0\right\}$ and $U_x^{(t)} = \left\{\widetilde{x} : K_{IGSM}^{(t)}(x, \widetilde{x}) > 0\right\}$. Assume that $U_x^{(t)} \supseteq V_x^{(t)}$ and let $z \in V_x^{(t+1)}$, so

$$\int\limits_{U_x^{(t)}} K_q^{(t)}(x, y)K_q(y, z) > 0. \tag{a.1}$$

Suppose that $z \notin U_x^{(t+1)}$, then the support of the function $K_q^{(t)}(x, \cdot)K_{IGSM}(\cdot, z)$ has measure zero. Given that $\alpha(\widetilde{x}, x)$ is positive for every $\widetilde{x}, x \in D_p$, then it implies that also the support of $K_{IGSM}^{(t)}(x, \cdot)K_q(\cdot, z)$ has measure zero, which is in contradiction with (a.1). So it must be that $U_x^{(t+1)} \supseteq V_x^{(t+1)}$ and given that $U_x^{(1)} \supseteq V_x^{(1)}$ then the $pf-$irreducibility of K_{IGSM} follows. ■

Proof of proposition 3.

The new measure will be $f(x_{(1)}, x_{(2)})p(x_{(1)}, x_{(2)}) = \frac{f_1(x_{(1)})f_2(x_{(2)})}{p\left(x_{(1)}|x_{(2)}\right)}p(x_{(1)}, x_{(2)}) = f_1(x_{(1)})$ $f_2(x_{(2)})p(x_{(2)})$, so the two blocks of the vector are independent. The conditional distribution $f(x_{(1)}|x_{(2)})$ is proportional to $\frac{f_1(x_{(1)})}{p\left(x_{(1)}|x_{(2)}\right)}$ and the conditional and the marginal of x_1 under the new measure will be $f_1(x_{(1)})$, while the conditional $f(x_{(2)}|x_{(1)})$ is proportional to $\frac{f_2(x_{(2)})}{p\left(x_{(1)}|x_{(2)}\right)}$ and so the conditional and the marginal of x_2 under the new measure is $f_2(x_{(2)})\frac{p\left(x_{(2)}|x_{(1)}\right)}{p\left(x_{(1)}|x_{(2)}\right)} =$ $f_2(x_{(2)})\frac{p\left(x_{(2)}\right)}{p\left(x_{(1)}\right)} \propto f_2(x_{(2)})p(x_{(2)})$. ■

10. Appendix 2

Here it is illustrated how to generate S_T from the conditional density $p(S_T|Z_T, Y_T)$ using the algorithm of Carter and Kohn (1994).

The first step of the algorithm consists in using recursive filtering equation to calculate $p(s_t|Z_t, Y_t)$. Given the observable variables Y_T, it is true that

$$p(s_t|Z_t, Y_t) \propto p(y_t|z_t, s_t)p(z_t|z_{t-1}, s_t)p(s_t|Z_{t-1}, Y_{t-1})$$

and, summing over the realization of s_t, the normalizing costant is computed so the probability is given by

$$p(s_t|Z_t, Y_t) = \frac{p(y_t|z_t, s_t)p(z_t|z_{t-1}, s_t)p(s_t|Z_{t-1}, Y_{t-1})}{\sum_{s_t} p(y_t|z_t, s_t)p(z_t|z_{t-1}, s_t)p(s_t|Z_{t-1}, Y_{t-1})}. \tag{a.2}$$

Given the transition probability $p(s_t|s_{t-1})$, it follows that

$$p(s_t|Z_{t-1}, Y_{t-1}) = \sum_{s_{t-1}} p(s_t|s_{t-1})p(s_{t-1}|Z_{t-1}, Y_{t-1}). \tag{a.3}$$

and so, for $t = 1...T$, the loop of (a.2) and (a.3) provides the desired probability for every t.

To sample from $p(S_T|Z_T, Y_T)$, note that

$$p(S_T|Z_T, Y_T) = p(s_T|Z_T, Y_T) \prod_{t=1}^{T-1} p(s_t|Z_t, Y_t, s_{t+1})$$

and

$$p(s_t|Z_t, Y_t, s_{t+1}) = \frac{p(s_{t+1}|s_t, Z_t, Y_t)p(s_t|Z_t, Y_t)}{p(s_{t+1}|Z_t, Y_t)}. \tag{a.4}$$

So the last observation is sampled from $p(s_T|Z_T, Y_T)$ and afterward the other realizations for $t < T$ are sampled backward form (a.4).

References

[1] Athreya, K.B., Doss, H. and Sethuraman J. (1996) "On the Convergence of the Markov Chain Simualtion Method" *Annals of Statistics*, 24, 69-100.

[2] Billio, M., and Monfort, A. (1995) "Switching State Space Models" working paper CREST-INSEE, n.9557.

[3] Carter, C.K., and Kohn, R. (1994) "On Gibbs Sampling for State Space" *Biometrika*, 81, 541-553.

[4] Chib, S., and Greenberg, E. (1996) " Markov Chain Monte Carlo Simulation Methods in Econometrics" *Econometric Theory*, 12, 409-431.

[5] Dejong, P., and Shepard, N. (1995) "The Simulation Smoother for Time Series Models" *Biometrika*, 82, 339-350

[6] Diebold, F.X., and Rudebusch, G.D. (1996) "Measuring Business Cycle: a Modern Prospective" *Review of Economics and Statistics*, 78, 67-77.

[7] Elliot, R.J., Aggoun, L. and Moore , J.B. (1994) *Hidden Markov Models,* Springer-Verlag.

[8] Geman, S., and Geman, D. (1984) "Stochastic Relaxation, Gibbs Distributions and the Bayesian Restoriation of Images" *IEEE Trans. Pattn. Annal. Mach. Intel.*, 6, 721-741.

[9] Geyer, C.J., (1996) "Estimation and Optimization of Functions", in Gilks, W.R., Richardson, S. and Spiegelhalter, D.J., *Markov Chain Montecarlo in Practice*, Chapman & Hall.

[10] Geyer, C.J., and Thomposon, E.A. (1992) " Constained Monte Carlo Maximum Likelihood for Dependent Data", *Journal of the Royal Satistical Societ*y, B.

[11] Geweke, J., (1989) "Bayesian Inference in Econometrics Models Using Monte Carlo Integration" *Econometrica*, 57, 1317-1240.

[12] Geweke, J., (1995) "Monte Carlo Simulation and Numerical Integration" Federal Reserve Bank of Minneapolis, Staff Report 192.

[13] Gilks, W.R.„ and Wild, P. (1994) " Adaptive Rejection Sampling for Gibbs Sampling" *Appl. Statistics*, 41, 337-348.

[14] Gilks, W.R., Richardson, S. and Spiegelhalter, D.J. (1996) *Markov Chain Montecarlo in Practice*, Chapman & Hall.

[15] Ghysel,E., Gourieroux, C., and Jasiak, J., (1995) "Trading Patterns, Time Deformation and Stochastic Volatility in Foreign Exchange markets" CIRANO working paper, 95-42.

[16] Ghysel, E., Harvey, A., and Renault, E. (1995) "Stochastic Volatility" working paper, CORE.

[17] Hammersly, J.M., and Handscomb, D.C. (1964) *Monte Carlo Methods*, Methuen and Company, London.

[18] Jacquier, E., Polson, N.G., and Rossi, P.E. (1994) " Bayesian Analysis of Stochastic Volatility Models" *Journal of Business and Economic Statistics*, 12, 371-417.

[19] Kim, C.J. (1994) "Dynamic Linear Models with Markov Switching" *Journal of Econometrics*, 60, 1-22.

[20] Kim, S. and Shephard, N., (1994) "Comment on 'Bayesian Analysis of Stochastic Volatility Models' " *Journal of Business and Economic Statistics*, 12, 371-417.

[21] Kim, S., Shephard, N., and Chib, S. (1996) "Stochastic Volatility: Likelihood Inference and Comparison with Arch Models" mimeo, Nuffield College, Oxford.

[22] Lui, J.S., Wong, W.W. and Kong, A. (1994) "Covaraince Structure of the Gibbs Sampler with Application to the Comparison of Estimators and Augmentation Schemes" *Biometrica*, 81, 27-40.

[23] Mahieu, R., and Schotman, P. (1997) "Stochastics Volatility and the Distribution of Exchange Rate News" *Journal of Applied Econometrics*, forthcoming

[24] Mengersen, K.L., and Tweedie, R.L. (1996) "Rates of Convergence of the Hastings and Metropolis Algorithm" *Annals of Statistics*, 24, 101-121.

[25] Metropolis, N.A., Rosenbluth, A.W., Rosenbluth, M.N., Teller, A.H., Teller, E. (1953) "Equation of State Calculation by Fast Computing Machines" *Journal of Chemical Physics*, 21, 1087-1092.

[26] Nummelin, E. (1983) *General Irreducible Markov Chains and Non-Negative Operators*, Cambridge University Press.

[27] Roberts, G.O. (1995) "Methods for Estimating L^2 Convergence of Markov Chian Monte Carlo" in D.Barry, K.Chaloner and J. Geweke, *Bayesian Analysis in Statistics and Econometrics: Essay in Honor of A.Zellner*, Wiley, New York.

[28] Roberts, G.O. and Smith, A.F.M. (1994) "Bayesian Computation via the Gibbs sampler and Related Markov Chain Monte Carlo Methods" *Journal of the Royal Satistical Society*, B, 55, 3-24.

[29] Shephard, N. (1994) "Patial non-Gaussian State Space" Biometrika, 81, 115-131.

[30] Tanner, M.A., and Wong, W.H. (1987) "The Calculation of Posterior Distributions by Data Augmentation" *Journal of The American Statistical Association*, 81, 82-86.

[31] Tierney, L. (1994) " Markov Chians for Exploring Posterior Distributions" *Ann. Statistics*, 22, 1701-1762.

[32] Tierney, L. (1996) "Introduction to General Markov Chain Theory" in Gilks, W.R., Richardson, S. and Spiegelhalter, D.J., *Markov Chain Montecarlo in Practice*, Chapman & Hall.

[33] Zellner, A., (1971) *An Introduction to Bayesian Inference in Econometrics*, Wiley, New York.