

Model selection of stochastic simulation algorithm based on generalized divergence measures

B. Don Bosco DIATTA¹ and Papa NGOM²

^{1,2}LMA-Laboratoire de Mathématiques Appliquées
Université Cheikh Anta Diop BP 5005 Dakar-Fann Sénégal

Revised Article

Received: XX December 20XX

Accepted: XX December 20XX

Online Ready: XX December 20XX

Abstract

Stochastic simulation methods such as MCMC (Monte Carlo Markov Chain) methods have become a standard tool in Statistics because it may be the only tool available for analysing some classes of probabilistic models or solving problems like the calculation of some complicated integrals, or in the Bayesian analysis, in order to estimate the parameters of Bayesian models. However, model selection under MCMC algorithms has been a subject of intense debate during the recent years. Here, we consider generalized approaches to compare different simulation strategies among which the *Independent Sampler (IS)*, the *Random Walk of Metropolis Hastings (RWMH)*, the *Gibbs Sampler (GS)*, the *Adaptive Metropolis (AM)* and *Metropolis within Gibbs (MWG)*. Specially, once the set of candidate simulation algorithms has been chosen, the statistical analysis allows us to select the best of strategy in the sense of rates of convergence, using some informational criteria like the Rényi divergence measure $R_\alpha(p, q)$, the Tsallis divergence $T_\alpha(p, q)$ and the α -divergence $D_\alpha(p, q)$, where p and q are probability density functions. These divergence measures are taken in the family of α -divergence measures. In different examples borrowed from synthetic models, with target distributions that are one and two-dimensional, we show the consistency and applicability of these α -divergence measures for selection in stochastic simulation.

Keywords: MCMC methods, Metropolis-Hastings algorithm, Gibbs Sampler, Adaptive Metropolis, Metropolis Within Gibbs, simulation strategy, target density, proposal density, stationary distribution, α -divergence.

2010 Mathematics Subject Classification: 60J60, 62F03, 62F05, 94A17

1 Introduction

The stochastic simulation methods (MCMC methods and recent hybrid and adaptive methods) implement many simulation algorithms each having its own convergence speed. But these rates of

Papa NGOM : E-mail: papa.ngom@ucad.edu.sn

convergence are unknown in practice. The study concerning the comparison of stochastic simulation strategies is introduced by Chauveau and Vandekerckhove (2007). These authors have used the Kullback-Leibler divergence measure to compare simulation strategies. The estimator of Kullback divergence proposed by Chauveau and Vandekerckhove (2007) uses several assumptions not always verifiable in practice. These assumptions concern the target densities and proposal densities in their simulation methods.

The contribution of this paper is that it proposes to replace the Kullback divergence by measures named α -divergence measures [Cichocki and Amari (2010)]. These divergence measures are the Rényi divergence $R_\alpha(p, q)$, the Tsallis divergence $T_\alpha(p, q)$ and the α -divergence $D_\alpha(p, q)$ where p and q are probability density functions. These divergence measures have a parameter α . The common property of these divergence measures is that, when the parameter α tends to 1, they converge all to the Kullback divergence measure. Note that in the study made by Chauveau and Vandekerckhove (2007) all the methodology is based on the two well-known strategies of Metropolis-Hastings algorithm (IS and RWMH). Indeed, these authors have assumed assumptions like the Lipschitz condition on the densities p^n of $X_n, n = 1, \dots$, generated from these simulation strategies (IS and RWMH) to establish the conditions of convergence of the Kullback divergence estimator that they have proposed. The methodology presented in our paper has the advantage to show, through graphics, the speed of convergence of the simulation strategies IS and RWMH but also Gibbs Sampler (GS) and others hybrid and adaptive simulation methods (AM and MWG). Added to this, the use of these α -divergence measures gives a variety of divergence measures for different values of the parameter α . Using these α -divergences is particularly advantageous since there is no major requirements on densities p^n from simulation strategies and the target density f . The divergence measures that we use here have a common part that is the integral in the definition of these measures. For our illustrative examples we will use this integral estimator given by Póczos and Schneider (2011).

Now we give more detail about our general study framework. Suppose we have a probability distribution with density function f from which we want to obtain samples. Suppose also that the methods of direct simulations are out of reach. Then we use the MCMC methods and derived methods (adaptive or hybrid simulation methods). We work in this paper with simulation strategies that are Independent Sampler (IS), Random Walk Metropolis-Hastings (RWMH) and Gibbs Sampler (GS) which are MCMC methods but also with Adaptive Metropolis (AM) that is an adaptive method and Metropolis Within Gibbs (MWG) that is hybrid method. These two latter methods (AM and MWG) are explained more fully in Section 4. To implement these different simulation algorithms one needs in each case a proposal distribution that generates samples. Note also that each algorithm, depending on the choice of the proposal law, converges more or less promptly. This is highlighted when one uses a simulation method with two different proposal distributions and compares the convergence time of the two resulting algorithms. We recall that for a given simulation strategy each proposal distribution corresponds to one algorithm to implement. The convergence time of a stochastic simulation algorithm, here, is actually the time that the densities $p^n, n = 1, \dots$ will put to converge to the density f of the target law. The densities p^n are actually densities of elements of the chain $(X_n)_{n \geq 1}$ generated by one simulation strategy ie, at each iteration n , p^n is the density function of the random variable X_n . This is what explains the interest, if we consider for example the Rényi divergence, to calculate for each iteration n the value $R_\alpha(p^n, f)$. For different iterations n quantities $R_\alpha(p^n, f)$ will be represented by a curve; then the interest is to study the evolution of this curve with respect to the value 0. In this respect, consider two simulation strategies S_1 and S_2 for a given target distribution having density function f . We assume that these strategies S_1 and S_2 generate respectively the samples $X_{1,1}, X_{1,2}, \dots, X_{1,n}, \dots$ and $X_{2,1}, X_{2,2}, \dots, X_{2,n}, \dots$. The densities p_1^n and p_2^n are respectively density functions of $X_{1,n}$ and $X_{2,n}$. If we choose for example the Rényi divergence we can compare the curves of values $R_\alpha(p_1^n, f)$ and $R_\alpha(p_2^n, f), n = 1, \dots$. The curve which converges more rapidly to 0 indicates that the corresponding simulation strategy is more efficient.

The paper is organised as follows: in Section 2 we will give the definitions and estimators

of the α -divergence measures and we give methodology for comparing two simulation strategies and how to choose an efficient simulation strategy. We will show in Section 3 that our divergence measures have a gaussian asymptotic distribution. Then we have some application examples in order to illustrate our methodology in Section 4. Finally, in Section 5 we will explain the results of our various illustrative examples.

2 Description and estimators of divergence measures and methodology for finding an optimal simulation method

2.1 Description of divergence measures

Let $(\mathcal{X}, \mathcal{A}, \lambda)$ be an arbitrary measure space with λ being a finite or σ -finite measure. Let also μ_1, μ_2 probability measures on $\mathcal{X}$ such that $\mu_1, \mu_2 \ll \lambda$ (absolutely continuous). Denote the Randon-Nikodym derivatives (densities) of μ_i with respect to λ by $p_i(x)$:

$$p_i(x) = \frac{\mu_i(dx)}{\lambda(dx)}, \quad i = 1, 2.$$

Definition 2.1. The Kullback-Leibler relative divergence (also called relative entropy) between two probability measures μ_1, μ_2 is defined by

$$K(\mu_1, \mu_2) = \int_{\mathcal{X}} p_1(x) \log \left(\frac{p_1(x)}{p_2(x)} \right) \lambda(dx) = \mathbb{E}_{\mu_1} \left[\log \frac{p_1(X)}{p_2(X)} \right] \quad (2.1)$$

We can also write $K(p_1, p_2)$.

We will focus on in this article to a particular family of *Csiszár ϕ -divergence* that is the family of α -divergence measures. The D_α divergence [Cichocki, Lee, Kim and Choi (2008)], the Rényi α -divergence and the Tsallis α -divergence [Cichocki and Amari (2010)] are part of this family.

Definition 2.2 (α -divergence). The α -divergence is defined by

$$D_\alpha(\mu_1, \mu_2) = \frac{1}{\alpha(1-\alpha)} \left(1 - \int_{\mathcal{X}} p_1^\alpha(x) p_2^{1-\alpha}(x) \lambda(dx) \right), \quad \alpha > 0 \text{ and } \alpha \neq 1. \quad (2.2)$$

Definition 2.3 (Rényi α -divergence). Rényi (1961) for the first time gave one generalization of the relative entropy given in (2.1). It is defined by

$$R_\alpha(\mu_1, \mu_2) = \frac{1}{\alpha-1} \log \int_{\mathcal{X}} p_1^\alpha(x) p_2^{1-\alpha}(x) \lambda(dx), \quad \alpha > 0 \text{ and } \alpha \neq 1. \quad (2.3)$$

Definition 2.4 (Tsallis α -divergence). Tsallis α -divergence is defined by

$$T_\alpha(\mu_1, \mu_2) = \frac{1}{\alpha-1} \left(\int_{\mathcal{X}} p_1^\alpha(x) p_2^{1-\alpha}(x) \lambda(dx) - 1 \right), \quad \alpha > 0 \text{ and } \alpha \neq 1 \quad (2.4)$$

In these definitions of divergence measures appears one integral. This integral $M_\alpha(p_1, p_2) = \int_{\mathcal{X}} p_1^\alpha(x) p_2^{1-\alpha}(x) \lambda(dx)$, common to all three measures, is often difficult to calculate. That is why in practice, we need estimators of these divergence measures to assess the distance (similarity) between two density functions.

2.2 Estimators of divergence measures

We first make an estimator of the divergence measure $D_\alpha(p^n, f)$, afterward we can have in the same manner an estimator of Tsallis and Rényi divergences. We work now with λ the Lebesgue measure defined on $\mathcal{X} = \mathbb{R}^d$. Then we have

$$D_\alpha(p^n, f) = \frac{1}{\alpha(1-\alpha)} \left(1 - \int_{\mathbb{R}^d} \left(\frac{f(x)}{p^n(x)} \right)^{1-\alpha} p^n(x) dx \right)$$

We can initially think to write the integral like a mathematical expectation and apply the method of Monte Carlo integration. We would have then

$$D_\alpha(p^n, f) = \frac{1}{\alpha(1-\alpha)} \left(1 - \mathbb{E} \left(\left(\frac{f(X)}{p^n(X)} \right)^{1-\alpha} \right) \right)$$

If f and p^n are such $(f(x)/p^n(x))^{1-\alpha}$ is measurable we can use the *Strong Law of Large Numbers* we will have the following estimator

$$\hat{D}_\alpha(p^n, f) = \frac{1}{\alpha(1-\alpha)} \left(1 - \frac{1}{N} \sum_{i=1}^N \left(\frac{f(X_i)}{p^n(X_i)} \right)^{1-\alpha} \right), \text{ with } X_i \sim p^n$$

which converges almost surely to $D_\alpha(p^n, f)$.

The densities p^n are often untractables because it is difficult to find their analytical form. There are also cases where f is known to a multiplicative constant (bayesian context). However, even if the density f is known, as in some of our examples in Section 4, we will use its estimator like in [Póczos and Schneider (2011)]. Then we use the methods of nonparametric estimation of probability densities. Póczos and Schneider (2011) have proposed in their paper an estimator of probability density based on *k-NN method* (k nearest neighbor). This probability density estimation method has been introduced by [Loftsgaarden and Quesenberry (1965)]. The authors Póczos and Schneider have used it in [Póczos and Schneider (2011)]. If we apply their results to our densities f and p^n , we obtain the following density estimators

$$\hat{p}_{k,n,N}(X_i) = \frac{k/(N-1)}{V(H(X_i, \rho_k(X_i)))} = \frac{k}{(N-1)c \rho_k^d(X_i)} \quad (2.5)$$

$$\hat{f}_{k,M}(X_i) = \frac{k/M}{V(H(X_i, \gamma_k(X_i)))} = \frac{k}{M c \gamma_k^d(X_i)} \quad (2.6)$$

Denote by $V(H(z, r)) = \pi^{d/2} r^d / \Gamma((d/2) + 1)$ is the volume of d-dimensional sphere around $z \in \mathbb{R}^d$ with radius $r > 0$, $\Gamma(\cdot)$ is the Gamma function and c stands for the volume of a d-dimensional unit ball.

Recall some how these density estimators are constructed in [Póczos and Schneider (2011)]. Let $X_{1:N} = (X_1, \dots, X_N)$ be a sample simulated from distribution of density p^n where $X_i, i = 1, \dots, N$ are i.i.d (independent identically distributed). One can choose one realization X_i on the sample $X_{1:N}$ and calculate the value $\rho_k(X_i)$ that is the euclidean distance between X_i and its k th nearest neighbor on the sample. If the samples $Y_{1:M} = (Y_1, \dots, Y_M)$, where the $Y_i, i = 1, \dots, M$ are i.i.d and generated from the density f , then $\gamma_k(X_i)$ is the euclidean distance between X_i and its k th nearest neighbor on the sample $Y_{1:M}$. We can find in [Póczos and Schneider (2011)] convergence results of such types of density estimators. Thus the estimators $\hat{p}_{k,n,N}(X_i)$ and $\hat{f}_{k,M}(X_i)$ are consistent.

We can notice in their respective definitions that divergence measures used here have a common part which is the integral $M_\alpha(p_1, p_2) = \int_{\mathbb{R}^d} p_1^\alpha(x) p_2^{1-\alpha}(x) dx$ where p_1 and p_2 are density functions. As we have already mentioned in the Introduction, Póczos and Schneider (2011) have

given an estimator of $M_\alpha(p_1, p_2)$ that we use in this paper. This estimator is

$$\hat{M}_{\alpha, N, M, k}(p_1, p_2) = \frac{1}{N} \sum_{i=1}^N \left(\frac{\hat{p}_{2, k, M}(X_i)}{\hat{p}_{1, k, N}(X_i)} \right)^{1-\alpha} \times B_{k, \alpha} \quad (2.7)$$

where $\hat{p}_{1, k, N}$ and $\hat{p}_{2, k, M}$ are respective density estimators of p_1 and p_2 ; the constant $B_{k, \alpha}$ is defined as follow

$$B_{k, \alpha} = \frac{(\Gamma(k))^2}{\Gamma(k - \alpha + 1)\Gamma(k + \alpha - 1)} \quad \text{with } \Gamma(\cdot) \text{ the Gamma function} \quad (2.8)$$

Recall that in this paper we are in a context where we have a density f from which we can not have directly i.i.d. samples. Then we use an efficient MCMC method for generating a Markov chain $(X_n)_{n \geq 1}$. The samples that we use to estimate the density f are from this Markov chain and are built as follow: when n is sufficiently large, we choose $X_{n+r}, X_{n+2r}, \dots, X_{n+ir}, \dots, X_{n+Mr}$ where r is an integer for example such as $r \geq 10$. The fact to make this leap of r elements between samples that we choose allows us to obtain samples more or less independent. Then if we replace p_1 by p^n and p_2 by f we can apply the method proposed in [Póczos and Schneider (2011)] to obtain our divergence measures estimators

$$\hat{D}_{\alpha, N, M, k}(p^n, f) = \frac{1}{\alpha(1-\alpha)} \left(1 - \frac{1}{N} \sum_{i=1}^N \left(\frac{(N-1)\rho_k^d(X_i)}{M\gamma_k^d(X_i)} \right)^{1-\alpha} \times B_{k, \alpha} \right) \quad (2.9)$$

$$\hat{T}_{\alpha, N, M, k}(p^n, f) = \frac{1}{\alpha-1} \left(\frac{1}{N} \sum_{i=1}^N \left(\frac{(N-1)\rho_k^d(X_i)}{M\gamma_k^d(X_i)} \right)^{1-\alpha} \times B_{k, \alpha} - 1 \right) \quad (2.10)$$

$$\hat{R}_{\alpha, N, M, k}(p^n, f) = \frac{1}{\alpha-1} \log \left(\frac{1}{N} \sum_{i=1}^N \left(\frac{(N-1)\rho_k^d(X_i)}{M\gamma_k^d(X_i)} \right)^{1-\alpha} \times B_{k, \alpha} \right) \quad (2.11)$$

that are respectively the divergence estimators of $D_\alpha(p^n, f)$, $T_\alpha(p^n, f)$ and $R_\alpha(p^n, f)$.

The following theorem contains result proving the convergence of these divergence estimators. This is the theorem given by Póczos and Schneider (2011) that we have taken up replacing the densities in this theorem by the densities p^n and f .

Theorem 2.1 (L_2 consistency). *We have the following assumptions: $k \geq 2$, $0 < \gamma = 1 - \alpha < (k-1)/2$, p^n is bounded away from 0, p^n is uniformly Lebesgue approximable, $\exists \delta_0$ such that $\forall \delta \in (0, \delta_0)$ $\int F(x, p^n, \delta, 1/2) p^n(x) dx < \infty$, $\int \|x - y\|^\gamma p^n(y) dy < \infty$ for almost all $x \in \mathbb{R}^d$, $\int \int \|x - y\|^\gamma p^n(y) p^n(x) dy dx < \infty$, and that f is bounded above. Then*

$$\lim_{N, M \rightarrow \infty} \mathbb{E} \left(\left(\hat{M}_{\alpha, N, M, k}(p^n, f) - M_\alpha(p^n, f) \right)^2 \right) = 0 \quad (2.12)$$

that is, the estimator is L_2 consistent.

The function $F(x, q, \delta, 1/2)$, with q a probability density, defined in [Póczos and Schneider (2011)]

Once the divergence measures and their estimators described, one needs to know how to use these measures. In the next section (Methodology) we show the methodology of comparing two simulation strategies using these divergence measures.

2.3 Methodology

In this present paper we propose to study how to find an optimal stochastic simulation algorithm may come from Metropolis-Hastings methods (*IS* and *RWMH*), *Gibbs Sampler*, or recent adaptive (*AM*) and hybrid (*MWG*) methods. For that, we will use our α -divergence measures. The interest

of these divergences is that it subtracts the target density f , the proposal density q_i and initial density p_i^0 (corresponding to one simulation strategy S_i) of many assumptions like in [Chauveau and Vandekerckhove (2007)].

We develop here our methodology that we have slightly discussed in the Introduction. The novelty of our approach lies in its ease of implementation because the divergence estimators that we use are built independently of any simulation strategy. This is why we can compare, in addition to IS and RWMH methods, various simulation methods. Among these methods we choose here the GS, AM and MWG strategies that we have also presented in the Introduction. However one can not compare simulation strategies that are not empowered to solve the same problem. For example, the Gibbs Sampler method, usable when the target density f is a function defined on $E \subseteq \mathbb{R}^d$ with $d \geq 2$ and when the conditional densities are available, must not be compared to an one-dimensional IS strategy (target density f is defined on $E \subseteq \mathbb{R}$). Thus, to compare two or more stochastic simulation strategies, one must ensure the existence of a common target distribution. The simulation strategies, some more effective than others, are intended to generate samples $X_1, X_2, \dots, X_n, \dots$ such that when n is large (tends to ∞) then X_n has f as density function, ie X_n is considered as simulated by the law of the density f . It is then appropriate to evaluate at each iteration n the measure of similarity between the density p^n of X_n and the target density f . Here we still use the Rényi divergence measure for illustrating although is not the only divergence measure used in this paper. Our methodology allows to have multiple measures of divergence that we can use as needed. Among them there are well known divergence measures: it is the Hellinger divergence measure $D_{\frac{1}{2}}(p, q)$ and the Chi-square divergence measure $D_2(p, q)$ where p and q are density functions. Thus assume that we have two simulation strategies S_1 and S_2 having respective probability densities p_1^n and p_2^n at time n and a target density f which we want to have samples, we can use these divergence measures to compare S_1 and S_2 . Indeed, we can compare $R_\alpha(p_1^n, f)$ and $R_\alpha(p_2^n, f)$, for each iteration n , to see which of S_1 and S_2 is more efficient. The comparison is made by using curves.

- If the curve of $R_\alpha(p_1^n, f)$ values is closer to 0 than the curve of $R_\alpha(p_2^n, f)$ values (ie $0 < R_\alpha(p_1^n, f) < R_\alpha(p_2^n, f)$), then we can say that densities p_1^n converge faster than densities p_2^n to the stationary and target density f . Consequently the simulation strategy S_1 associated to p_1^n is more effective than S_2 .
- If the reverse happens, ie the curve of $R_\alpha(p_2^n, f)$ values is closer to 0 than the curve of $R_\alpha(p_1^n, f)$, ($0 < R_\alpha(p_2^n, f) < R_\alpha(p_1^n, f)$) we will say that the strategy S_2 is more efficient than S_1 .
- If the curve of $R_\alpha(p_1^n, f)$ is more or less similar to the curve of $R_\alpha(p_2^n, f)$ and all are very close to 0 ($R_\alpha(p_1^n, f) \simeq R_\alpha(p_2^n, f)$), we can say that the two simulation strategies S_1 and S_2 are equivalent and both efficient.
- If the curve of $R_\alpha(p_1^n, f)$ and the curve of $R_\alpha(p_2^n, f)$ are far from 0 and slow to approach 0 or do not approach 0, then both strategies are ineffective.

After presenting the methodology, the knowledge of the distribution of divergence measures estimators can be useful. However, we will not use these distributions, we only present it.

3 Asymptotic distribution of divergence estimators

We seek to know the asymptotic distribution of our estimators. The asymptotic distribution of these estimators is studied under the assumption of measurability of densities p^n and f . If densities p^n , f and their respective estimators are measurable, we choose to take $N = M$ and have the following results.

3.1 Asymptotic distribution of D_α and T_α divergence estimators

Theorem 3.1. If estimators of densities f and p^n are measurable, X_i are i.i.d and $\sigma^2 = \lim_{N \rightarrow \infty} \text{Var}(\bar{Y}_N) < \infty$, then

$$\hat{D}_{\alpha,N,k}(p^n, f) \xrightarrow{d} \mathcal{N}\left(D_\alpha(p^n, f), \frac{\sigma^2 B_{k,\alpha}^2}{\alpha^2(1-\alpha)^2}\right), \text{ when } N \rightarrow \infty. \quad (3.1)$$

$$\hat{T}_{\alpha,N,k}(p^n, f) \xrightarrow{d} \mathcal{N}\left(T_\alpha(p^n, f), \frac{\sigma^2 B_{k,\alpha}^2}{(\alpha-1)^2}\right), \text{ when } N \rightarrow \infty. \quad (3.2)$$

Proof. We know that X_i are i.i.d, now consider $\bar{Y}_N = \frac{1}{N} \sum_{i=1}^N h_N(X_i)$ with $h_N(X_i) = \left(\frac{(N-1)\rho_k^d(X_i)}{N\gamma_k^d(X_i)}\right)^{1-\alpha}$. Function h_N which is the ratio of two measurable functions is also measurable. We have therefore the independence of $h_N(X_i)$, $i = 1, \dots, N$. We can now apply the *Central Limit Theorem* for having asymptotical normal distribution

$$\frac{\bar{Y}_N - \mathbb{E}(\bar{Y}_N)}{\sqrt{\text{Var}(\bar{Y}_N)}} \xrightarrow{d} \mathcal{N}(0, 1) \implies \bar{Y}_N \times B_{k,\alpha} \xrightarrow{d} \mathcal{N}(M_\alpha(p^n, f), \sigma^2 B_{k,\alpha}^2)$$

We have now

$$\hat{D}_{\alpha,N,k}(p^n, f) \xrightarrow{d} \mathcal{N}\left(D_\alpha(p^n, f), \frac{\sigma^2 B_{k,\alpha}^2}{\alpha^2(1-\alpha)^2}\right), \text{ when } N \rightarrow \infty.$$

$$\hat{T}_{\alpha,N,k}(p^n, f) \xrightarrow{d} \mathcal{N}\left(T_\alpha(p^n, f), \frac{\sigma^2 B_{k,\alpha}^2}{(\alpha-1)^2}\right), \text{ when } N \rightarrow \infty.$$

□

3.2 Asymptotic distribution for Rényi divergence estimator

Theorem 3.2. If estimators of densities f and p^n are measurable, X_i are i.i.d. $\sim p^n$, $\sigma^2 = \lim_{N \rightarrow \infty} \text{Var}(\bar{Y}_N) < \infty$ and $\delta^2 = \lim_{N \rightarrow \infty} \text{Var}(h_N(X)) < \infty$ with $X \sim p^n$, then

$$\hat{R}_{\alpha,N,k}(p^n, f) \xrightarrow{d} \mathcal{N}\left(\frac{1}{\alpha-1} \log[M_\alpha(p^n, f)], \frac{\delta^2 B_{k,\alpha}^2}{(\alpha-1)^2 [M_\alpha(p^n, f)]^2}\right) \quad (3.3)$$

Proof. For determining the asymptotic distribution of the estimator of the Rényi divergence, we use the *Delta method*. Consider now this mean $\bar{Y}_N = \frac{1}{N} \sum_{i=1}^N h_N(X_i)$. We know that X_i are i.i.d. and h_N is measurable, then $h_N(X_i)$ are also i.i.d. The *Central Limit Theorem* gives us the following result

$$\frac{\bar{Y}_N - \mathbb{E}(\bar{Y}_N)}{\sqrt{\text{Var}(\bar{Y}_N)}} \xrightarrow{d} \mathcal{N}(0, 1) \text{ implying,}$$

$$\sqrt{N}(\bar{Y}_N B_{\alpha,k} - M_\alpha(p^n, f)) \xrightarrow{d} \mathcal{N}(0, \delta^2 B_{k,\alpha}^2), \quad (3.4)$$

Relation (3.4) allows us to apply the *Delta method* to obtain

$$\sqrt{N}(\log(\bar{Y}_N B_{k,\alpha}) - \log(M_\alpha(p^n, f))) \xrightarrow{d} \mathcal{N}\left(0, \frac{\delta^2 B_{k,\alpha}^2}{[M_\alpha(p^n, f)]^2}\right).$$

Hence we obtain the result

$$\hat{R}_{\alpha,N,k}(p^n, f) \xrightarrow{d} \mathcal{N}\left(\frac{1}{\alpha-1} \log[M_\alpha(p^n, f)], \frac{\sigma^2 B_{k,\alpha}^2}{(\alpha-1)^2 [M_\alpha(p^n, f)]^2}\right)$$

□

4 Examples

We will now illustrate our methodology with simple examples. That's why we study only the one-dimensional and two-dimensional cases. In the following examples, we will use α -divergence measures to compare firstly proposal densities corresponding to one given simulation strategy. Recall that one simulation strategy can implement several algorithms, each corresponding to one proposal density (or proposal law). One can now try to find among these densities those that allow to obtain better results (the speedy convergence of the algorithm). After this first example (Fig. 1) we will mainly compare different simulation strategies for the others examples. As we have already said, two simulation strategies can not be compared if they do not have the same target density. In all our examples the proposal distributions that we use are just examples to illustrate our methodology. Someone might choose to take others; for example in Example 1 he could choose to change the two proposal laws. The sizes of sample drawn from distributions of the densities p^n and f are equal for all examples, ie $M = N$. The value of k must be known, now empirical experiments have suggested to take k equals to the integer part of $\sqrt{N}$ [Loftsgaarden and Quesenberry (1965)]. In this section we only present the results, the explanations will be made in the Section Discussion.

4.1 One-dimensional case

4.1.1 Target density f fully known

In the case where the target density f is analytically known, we will still use its k -NN estimator in order to be in terms of the divergence estimators presented in [Póczos and Schneider (2011)]. The densities p^n are also estimated.

First we show an example that compares two proposal densities for a given strategy. We choose the *Independence Sampler (IS)* to compare these densities. The other example compares the IS and RWMH strategies.

a) *Independence Sampler (IS)*: comparison of proposal densities

For a given simulation strategy, the choice of an efficient proposal density (or proposal distribution) is important. Indeed, for achieving satisfactory simulation results it is important to choose an efficient proposal density. For simplicity, we consider densities that differ by the value of their parameters. Assume that the standard normal law $\mathcal{N}(0, 1)$ is the target distribution here, consider that its density function is f . To implement the IS simulation method we propose, for example here, two candidate proposal distributions that are $\mathcal{N}(-3, 2)$ and $\mathcal{N}(0, 3)$. We want to know if these two proposal distributions are all efficient or which of the two is more effective. If densities p_1^n and p_2^n are respectively generated by $\mathcal{N}(-3, 2)$ and $\mathcal{N}(0, 3)$ we compare now for all n the curves $D_2(p_1^n, f)$ and $D_2(p_2^n, f)$. The Figure 1 shows that the curve associated to the law $\mathcal{N}(0, 3)$ converges faster to 0 than the other curve (more details in Section Discussion).

b) *IS - RWMH* : comparison of simulation strategies

After finding a good proposal distribution for each strategy, we compare here the two main strategies of MH algorithm that are *IS* and *RWMH*. It may happen in an experiment that the *IS* strategy trumps *RWMH* strategy and in another experiment the opposite occurs. Everything depends on the instrumental distribution but also the target distribution to some extent, even if the initial law is the same for both strategies.

Here the chosen target distribution is a gaussian mixture $0.4\mathcal{N}(-8, 2) + 0.6\mathcal{N}(0, 6)$. For the *IS* we use the proposal distribution $\mathcal{N}(-2.5, 15)$ whereas for *RWMH* method we propose to take $\mathcal{N}(x, 15)$. This distribution has as mean equal to the current element $X_n = x$. We show the comparison of *IS* and *RWMH* strategies (Fig. 2). Note that the curve associated with *IS* strategy is below curve associated with *RWMH*.

Compare densities is equivalent to compare the respective probability distributions, so we will use interchangeably densities or probability distributions

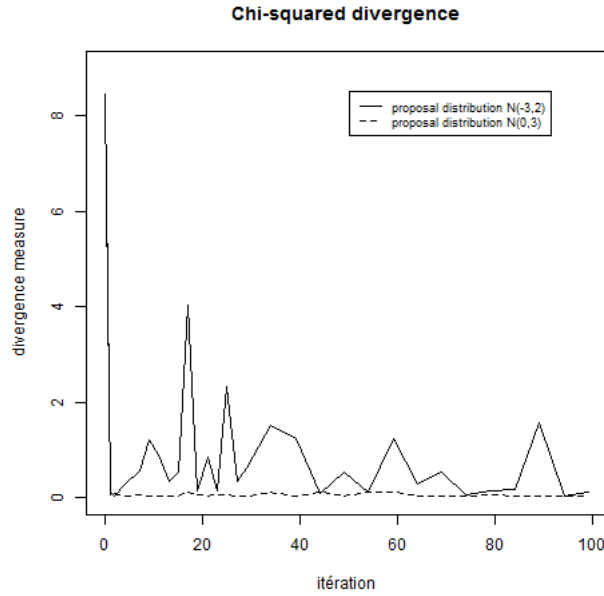


Figure 1: Comparison of two proposal densities, using the Independence Sampler and the Chi-squared divergence (D_α with $\alpha = 2$)

4.1.2 Target density f is not completely known

In most real situations, the density f is not known analytically. This is the case, for example Bayesian context where f is the density of the posterior. Then f is written as $f = c\varphi$ where c is unknown constant.

– *AM - RWMH: comparison of simulation strategies*

Present some *Adaptive Metropolis (AM)* strategy proposed by [Haario, Saksman and Tamminen (2001)]. First recall that the stochastic process generated by this simulation method is not a Markov chain. However it has well ergodicity properties. The assumptions for this are that the target density is bounded from above and has a bounded support.

The target density has a support $E \subset \mathbb{R}^d$. Suppose, that at time t we have sampled the states $X_0, X_1, \dots, X_{t-1}$, where X_0 is the initial state. Then a candidate point Y is sampled from the proposal distribution $q_t(\cdot | X_0, \dots, X_{t-1})$, which now may depend on the whole history $(X_0, X_1, \dots, X_{t-1})$. The candidate point Y is accepted with probability

$$\alpha(X_{t-1}, Y) = \min \left(1, \frac{\pi(Y)}{\pi(X_{t-1})} \right)$$

in which case we set $X_t = Y$, and otherwise $X_t = X_{t-1}$. Observe that the chosen probability for the acceptance resembles the familiar acceptance probability of the Metropolis algorithm for the symmetric case. The proposal distribution $q_t(\cdot | X_0, \dots, X_{t-1})$ employed in the *AM* algorithm is a Gaussian distribution with mean equal to the current point X_{t-1} and covariance

$$C_t = \begin{cases} C_0, & t \leq t_0 \\ S_d \text{Cov}(X_0, \dots, X_{t-1}) + S_d \epsilon I_d, & t > t_0. \end{cases}$$

where S_d is a parameter that depends only on dimension d and $\epsilon > 0$ is a constant that we may choose very small compared to the size of E . Here I_d denotes the d -dimensional identity matrix. The

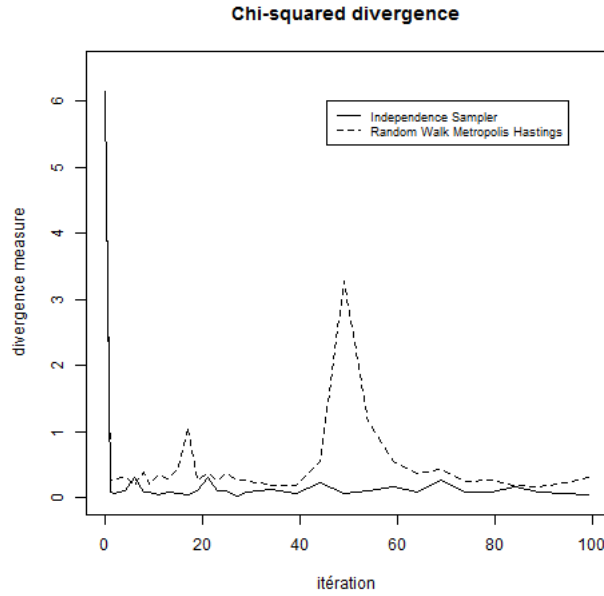


Figure 2: Comparison of two simulation strategies: *IS* vs *RWMH*, using the Chi-squared divergence (D_α with $\alpha = 2$)

covariance C_t may be viewed as a function of t variables from $\mathbb{R}^d$ having values in uniformly positive definite matrices.

For these two simulation methods (*AM* and *RWMH*) we see that the respective divergence measures are very closely and very quickly all tend to 0 (Fig. 3). Here divergence used is the *Hellinger divergence* measure $D_{1/2}$. The target density is $f(x) \propto \exp(-x^2)(2 + \sin(5x) + \sin(2x))$.

4.2 Two-dimensional case

We consider here the only case of a known function up to a constant. We choose now a sample $X_1, \dots, X_n$ which are i.i.d such that $X_i \sim \mathcal{N}(m, \sigma^2)$. So we have the following likelihood

$$L(x|m, \sigma^2) \propto (\sigma^2)^{(-n/2)} \exp\left(-\frac{1}{2\sigma^2} \sum_{i=1}^n (x_i - m)^2\right),$$

the prior distributions are

$$\begin{aligned} m &\sim \mathcal{N}(m_0, \sigma_0^2) \\ \sigma^2 &\sim \mathcal{IG}(\alpha, \beta), \end{aligned}$$

the full posterior density is known up to a constant

$$\Pi(m, \sigma^2|x) \propto (\sigma^2)^{-\frac{n}{2} - (\alpha+1)} \exp\left(-\frac{1}{2\sigma^2} \sum_{i=1}^n (x_i - m)^2 - \frac{(m - m_0)^2}{2\sigma_0^2} - \frac{\beta}{\sigma^2}\right),$$

the conditional distributions of parameters are

$$m|\sigma^2, x \sim \mathcal{N}(M, \Sigma^2)$$

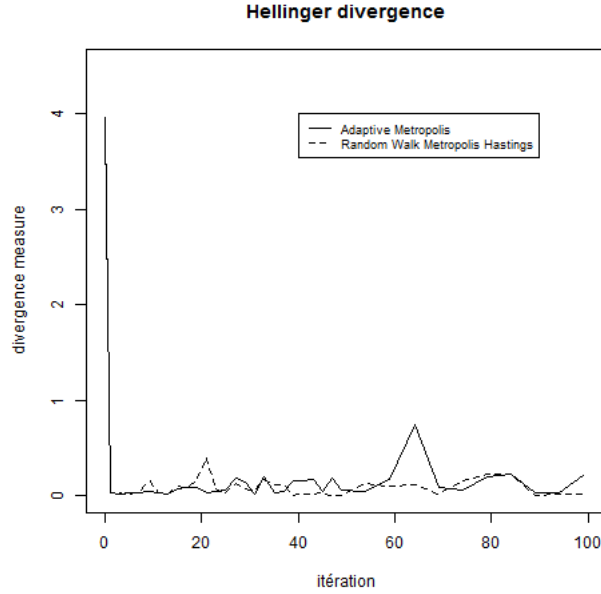


Figure 3: Comparison of two simulation strategies: *AM* vs *RWMH*, using the Hellinger divergence (D_α with $\alpha = 1/2$).

where

$$M = \frac{\sigma_0^2 \sum_{i=1}^n x_i + \sigma^2 m_0}{\sigma^2 + n\sigma_0^2} \quad \text{and} \quad \Sigma^2 = \frac{\sigma^2 \sigma_0^2}{\sigma^2 + n\sigma_0^2}$$

$$\sigma^2 | m, x \sim \mathcal{IG}\left(\frac{n}{2} + \alpha, \frac{1}{2} \sum_{i=1}^n (x_i - m)^2 + \beta\right)$$

So let's compare firstly *RWMH* and *Gibbs Sampler (GS)* and secondly we will compare *RWMH* and *MWG*. $\Pi(m, \sigma^2 | x)$ will be our target density for the two cases.

a) *RWMH - GS: comparison of simulation strategies*

If *RWMH* is applied in dimension $d \geq 1$, the *GS* is only applied in dimension $d > 1$. However, we study *GS* here only in dimension 2. Note that this method (*Gibbs Sampler*) has been used by Geman (1984) to generate observations from a Gibbs distribution (Boltzmann distribution). It is a particular form of the MCMC method because of its effectiveness. It is widely used in Bayesian analysis. Thus to simulate according to a probability density $f(\theta)$ with $\theta = (\theta_1, \dots, \theta_p)$ one can use the following idea

Initialisation: generating a vector $\theta = (\theta_1, \dots, \theta_p)$ according to a initial proposal law Π_0 .

Simulate following the conditional distributions

$$\theta_i | \theta_1, \dots, \theta_{i-1}, \theta_{i+1}, \dots, \theta_p \sim f_i(\theta_i | \theta_1, \dots, \theta_{i-1}, \theta_{i+1}, \dots, \theta_p)$$

$i = 1, 2, \dots, p$.

We see that the curve corresponding to *RWMH* strategy is well above curve representing the *GS* strategy (Fig. 4). The two curves have only one common point that is their origin. As mentioned in the legend, the dashed curve is associated to *RWMH* while the solid curve is associated to the *GS*. However, the curves denote divergence measures, we will talk about it in Section 5 (Discussion).

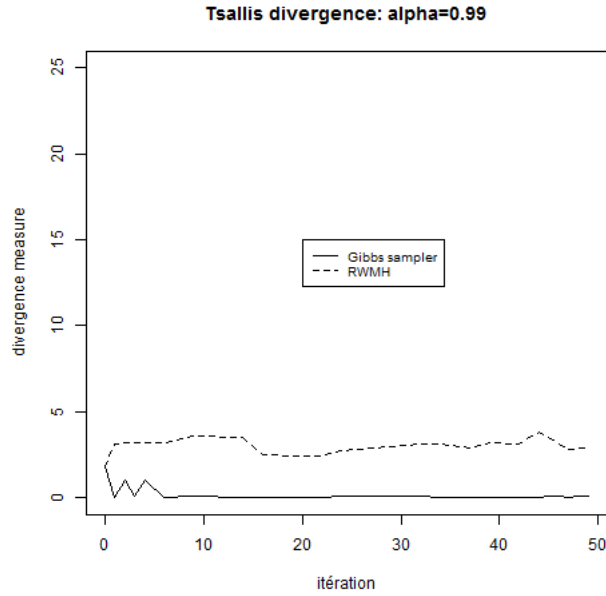


Figure 4: Comparison of two simulation strategies: GS vs RWMH, using the Tsallis divergence with $\alpha = 0.99$.

b) RWMH - MWG: comparison of simulation strategies

Metropolis Within Gibbs (MWG) is a hybrid simulation method that combines stages of the *Gibbs Sampler* and *Metropolis Hastings* method. It is used in some cases where we use GS and have conditional distributions for which we can't directly sample. There are several versions of this sampler, so we present the following.

Assume that $\pi(\cdot)$ is the target density, $\pi(\cdot|z_{-i})$ denote now the conditional distribution of $Z|Z_{-i} = z_{-i}$ where $Z \sim \pi$. $X_n := (X_{n,1}, \dots, X_{n,d})$; $X_{n,-i} := (X_{n,1}, \dots, X_{n,i-1}, X_{n,i+1}, \dots, X_{n,d})$; $\alpha := (\alpha_1, \dots, \alpha_d)$. Now we have the following algorithm.

Algorithm 4.1. 1. Choose coordinate $i \in \{1, \dots, d\}$ according to selection probabilities α , that is, with $\mathbb{P}(i = j) = \alpha_j$.
2. Draw $Y \sim q(X_{n-1,i}, \cdot)$.
3. Accept the candidate Y with probability

$$\min \left(1, \frac{\pi(Y|X_{n-1,-i})q(X_{n-1,i}, Y)}{\pi(X_{n-1,i}|X_{n-1,-i})q(X_{n-1,i}, Y)} \right)$$

and set $X_n := (X_{n-1,1}, \dots, X_{n-1,i-1}, Y, X_{n-1,i+1}, \dots, X_{n-1,d})$
otherwise reject Y and set $X_n = X_{n-1}$.

Starting from a common point, the curves stay away from the value 0 (Fig. 5). The solid curve is associated with *MWG* strategy and the dashed curve associated with *RWMH* strategy. Here we use the Rényi divergence measure for doing comparison between strategies.

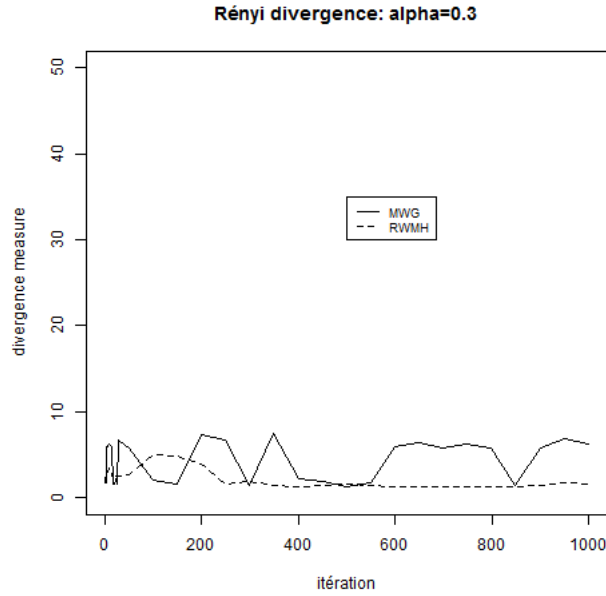


Figure 5: Comparison of two simulation strategies: *MWG* vs *RWMH*, using the Rényi α -divergence with $\alpha = 0.3$.

5 Discussion

5.1 Independence Sampler : comparison of proposal densities

The two curves represent measurements of *Chi-square divergence* D_2 . Indeed, the dashed curve indicates the difference between the densities p_1^n from the *IS* strategy with proposal law $\mathcal{N}(0, 3)$ and the target density f of $\mathcal{N}(0, 1)$, noted by $D_2(p_1^n, f)$. The solid curve denotes the difference between the densities p_2^n produced by the *IS* strategy with proposal distribution $\mathcal{N}(-3, 2)$ and the target distribution $\mathcal{N}(0, 1)$, noted by $D_2(p_2^n, f)$.

The divergence measures $D_2(p_1^n, f)$ and $D_2(p_2^n, f)$ are functions of the number of iterations n and have the same origin. The fact that they have the same starting point can be explained by relevance to begin with a same state x_0 or same initial law π_0 to compare two simulation strategies. The two curves overlap up to the first iteration. Indeed, the *IS* strategy with $\mathcal{N}(0, 3)$ is more efficient because this distribution has a mean $m = 0$ like the mean of target distribution. Its variance $\sigma^2 = 3$ is greater than the variance of the target distribution which is $\sigma^2 = 1$. It follows that its support covers the support of the target density. The other strategy has a proposal distribution $\mathcal{N}(-3, 2)$. This law has a mean $m = -3$ and a variance $\sigma^2 = 2$ that makes its density function is shifted to the left and is therefore not adequate to cover the support of the target density. So that the chain generated by this simulation strategy is slow to converge to its target distribution.

5.2 Independence Sampler - RWMH

The two curves in Figure 2 converge all quickly to 0. Here, we also use the *Chi-square divergence* for comparing. The dashed curve represents the divergence measure between the densities p^n of *RWMH* strategy and target density $f(x) = 0.4f_1(x, -8, 2) + 0.6f_2(x, 0, 6)$ which is

from the Gaussian mixture that we have described above. Similarly, the solid curve depicts divergence measure between the densities p^n from the *IS* strategy and the target density $f(x)$. We notice that the solid curve is below the dashed curve. This shows that *IS* strategy is more efficient than the *RWMH* strategy. The dashed curve makes a big jump to reach the 3.28 value at the 50th iteration. This shows that this *RWMH* algorithm is not still stable meaning that convergence has not yet.

After their common starting point, the difference between the two curves appears in the first iteration. From there, the curve of *IS* is closer to 0. This is explained by the right choice of its proposal density from $\mathcal{N}(-2.5, 15)$. With a mean $m = -2.5$ and a variance $\sigma^2 = 15$, this density function is centered relatively to the target density $f(x)$. Thus the support of the target density is well covered by this proposal density.

Regarding the *RWMH* strategy, it has a normal proposal distribution at each iteration n with a mean equal to the current state $X_n = x_n$ and same variance as in the other case $\sigma^2 = 15$. This normal proposal density is centered on the current state. Consequently the convergence of the algorithm *RWMH* depends greatly on the initial state of the chain generated. In this example (Fig. 2) the generated chain very quickly approaches the stationary distribution but still remains unstable with these jumps that perform the curve associated.

5.3 Adaptive Metropolis - RWMH

For comparing these two simulation algorithms, we use the *Hellinger divergence* measure ($D_{1/2}$). The solid curve then describes the divergence measure between densities p^n from *AM* and the target density $f(x) \propto \exp(-x^2)(2 + \sin(5x) + \sin(2x))$. The dashed curve describes the divergence between the densities p^n from *RWMH* and the target density $f(x)$. The two curves are the same between the initial state and the first iteration. From there, each curve doesn't move away from each other, but coexist near value 0. Thus, the two corresponding simulation strategies are all very efficient.

The effectiveness of *AM* strategy is explained, by the fact that it adapts its proposal density to the target density. The adjustment mechanism is performed at the variance of proposal density. Thus, if $X_1, X_2, \dots, X_{n-1}$ have already been simulated and one desires to obtain the point (or vector) X_n at time n , he generates it according to a Gaussian proposal distribution having mean equal to the current state X_{n-1} and covariance matrix described above.

Recall that we use the *AM* strategy here in one dimension. Thus, the variance $\sigma^2 = 5$ chosen for our instrumental Gaussian between the initial time and time $t = 15$ (this is the period prior to the iterative update of the variance) yields samples which describe almost the support of the target density (stationary distribution). Added to this, from the 16th iteration the adjustments on the variance allow us to have good samples. Which explains why *AM* algorithm has a very fast convergence. With respect to *RWMH* algorithm, with same variance $\sigma^2 = 5$ for the proposal density, we also obtain good samples that converge very quickly to the stationary distribution.

5.4 Gibbs Sampler - RWMH

We study here two simulation methods in 2 dimensional case by the use of Tsallis divergence with $\alpha = 0.99$. The dashed curve (*RWMH*) is above the solid curve (*GS*) and far from the value 0 after the first 50 iterations. It shows, here, that our *RWMH* strategy is ineffective. This inefficiency is due to the bad choice of covariance matrix

$$M = \begin{bmatrix} 12 & 8 \\ 8 & 15 \end{bmatrix}$$

of the proposal distribution. This matrix which is implementation parameter is not optimum. The solid curve (*GS*) tends rapidly to 0 (after the 7th iteration). The strength of this sample is mainly due to the

fact that the components of the vector X_n are generated directly from the conditional distributions of the target distribution.

5.5 MWG - RWMH

Here we compare the previous *RWMH* strategy and *MWG* algorithm. Then we compare a classical Metropolis algorithm (*RWMH*) to a hybrid algorithm that combines Metropolis algorithm and Gibbs Sampler (*MWG*). The latter is used in specific situations. Indeed *MWG* method is often used in cases where one wants to use the Gibbs Sampler and for some conditional distributions he can't directly simulate. Then one introduce in the algorithm a few steps of *RWMH* sampler, offering instrumental distributions having for target laws the conditional distributions. In our case (Fig. 5) we used a version that systematically applies the steps of the Metropolis sampler to all conditional distributions, so that the algorithm will delay to converge.

We use here the Rényi divergence with $\alpha = 0.3$ to evaluate the two simulation strategies. The *RWMH* strategy is that used in the previous comparison, therefore it is ineffective. As for the *MWG* strategy it puts usually a long time to converge and this is the case here. This is the reason that the two curves are all far from 0 even after 1000 iterations (Fig 5), reflecting here the ineffectiveness of these two simulation algorithms.

5.6 Conclusion

We have shown that with various divergence measures one can compare two different simulation strategies for determining the best. These comparisons lead us to determine the effectiveness or no of a stochastic simulation algorithm. For doing this study, our paper has been structured as follows. In Section 2 we have described the divergence measures we have used and estimators of these divergence measures. We also have given a methodology to choose an effective strategy simulation. We have given the asymptotic distributions of D_α , T_α and R_α divergence estimators in Section 3. We recall that we had a Gaussian asymptotical distribution. Then in Section 4, we have given some examples for implementing our diverse simulation strategies and thus illustrate our methodology. Finally we have discussed in Section 5 the causes that make that one simulation strategy is more powerful than the other.

Acknowledgment

This research was supported, in part, by grants from Cheikh Anta Diop University and SIMONS-NLAGA Project . We are grateful to many seminar participants and to anonymous referees for comments.

References

- [1] D. Chauveau and P. Vandekerkhove (2007). How to compare MCMC simulation strategies? . *hal.inria.fr/hal-00019174/ (version 3), arXiv:math/0605263v1[math.st]*.
- [2] Chauveau D. and Vandekerkhov P. (2012). Smoothness of Metropolis-Hastings algorithm and application to entropy estimation. *ESAIM: Probability and Statistics* (17), 419-431.
- [3] Cichocki A., Lee H., Y. D. Kim and Choi S., (2008). Non-negative matrix factorization with α -divergence. *Pattern Recognition Letters*.

- [4] Cichocki A. and Amari S. (2010). Families of Alpha- Beta- and Gamma- Divergences: Flexible and Robust Measures of Similarities. *Entropy*,1-41.
- [5] Haario H., Saksman E. and Tamminen J. (2001). An adaptive Metropolis algorithm. *Bernoulli* 7,2,223-242.
- [6] Latuszyński K., Roberts G. O. and Rosenthal J. S. (2013). Gibbs sampler and related MCMC methods. *The Annals of Applied Probability* 23,(1), 66-98.
- [7] Loftsgaarden D. O. and Quesenberry C. P. (1965). A Nonparametric Estimate of a Multivariate Density Function. *Annals of Mathematical Statistic*,36,3, 1049–1051.
Mehta2
- [8] Mehta B., Verma N. K. and Sircar P. (2011). Performance analysis of alpha divergence in nonnegative matrix factorization of monaural musical sounds. *International Journal of Engineering, Science and Technology* 3(6), 275–282.
- [9] Ngom P. Ntep B. (2012). Minimum Penalized Hellinger Distance for Model Selection in Small Samples. *Open Journal of Statistic* 2,(4),369-382.
- [10] Póczos B., and Schneider J. (2011) On the Estimation of α -Divergences. *Proceedings of the 14th International Conference on Artificial Intelligence and Statistics (AISTATS)*, 15,609-617.
- [11] Póczos B., Xiong J. and Schneider J. (2012). Nonparametric Divergence Estimation with Applications to Machine Learning on Distributions *arXiv:1202.3758v1 [cs.LG]*, 599–608.
- [12] Van Erven T. and Harremos P. (2010). Rényi Divergence and Majorization. *arXiv: 1001.4448v3 [cs.IT]*.

©2014 B. Don Bosco DIATTA & Papa NGOM; This is an Open Access article distributed under the terms of the Creative Commons Attribution License <http://creativecommons.org/licenses/by/2.0>, which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited.