

Model selection of stochastic simulation algorithm based on generalized divergence measures

Abstract

MCMC methods (Monte Carlo Markov Chain) are a class of methods used to perform simulations according to a given probability distribution P . These methods are often used when one has difficulties to directly sample according to the probability distribution P . This distribution is then considered as target and stationary distribution of the generated chain. One can now generate a Markov chain $(X_n)_{n \in \mathbb{N}}$ which, when n is large, one obtains $X_n \sim P$. These MCMC methods consist of several simulation strategies including the *Independent Sampler (IS)*, the *Random Walk of Metropolis Hastings (RWMH)*, the *Gibbs Sampler*, the *Adaptive Metropolis (AM)* and *Metropolis Within Gibbs (MWG)*. Actually *AM* strategy [Haario, Saksman and Tamminen (2001)] does not draw Markov chain but a chain that checks the ergodic property. As to the other algorithms they generate, when running, Markov chains. Each simulation method has a convergence time. It is interesting, with a given target law P , to compare these diverse simulation strategies for determining which one has the best convergence time. Chauveau and Vandekerckhove (2007) have compared *IS* and *RWMH* strategies using the Kullback-Leibler divergence. In our paper we will compare these five simulation methods already mentioned by using generalized divergence measures. These divergence measures are taken in the family of α -divergence measures [Cichocki and Amari (2010)], with a parameter α . It comes the Rényi divergence, Tsallis divergence and D_α divergence.

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

In many areas of science, computer is essential. So computers are used for testing, calculations, simulations ... Beyond performance, the computation time is very important in the choice of calculation

methods. In the field of Statistics, the simulation of random variables are required to perform assessments such integral calculation. Thus we resort to numerical analysis methods or probabilistic methods. It is the latter that concern us in particular. The integral form is rewritten like an expectation $\mathbb{E}(X)$ and use the *Law of Large Numbers* for the Monte Carlo integration and the *Ergodic Theorem* for MCMC methods. We will focus here on MCMC methods with which one performs sampling according to some density f via Markov chains. Among MCMC methods we have Metropolis-Hastings algorithms, Gibbs Sampler and others adaptive and hybrid algorithms.

Thus the major challenge of these algorithms, like many computational technics, is the saving time i.e. their speedy convergence to the stationary distribution with density f . This is due to the fact that the convergence time is a brand of efficiency of any algorithm. This paper investigates the problem of selecting a good strategy simulation. We have several simulation strategies that are *Independence Sampler*, *Random Walk of Metropolis-Hastings*, *Gibbs Sampler*, *Metropolis Within Gibbs* [Latuszyński, Roberts and Rosenthal (2013)], *Adaptive Metropolis* [Haario, Saksman and Tamminen (2001)]... Each of these strategies try, depending on the context, to raise the challenge of saving time. Then we can have a wide range of divergence measures other than the Kullback divergence to appraise the robustness of these different simulation methods.

The objective of this paper is to compare different strategies simulation in order to select the best. Chauveau and Vandekerkhove (2007) have given in their paper a methodology for selecting the best of k candidate strategies with the Kullback-Leibler divergence. We use in this paper generalized divergence measures that are α -divergence [Póczos, and Schneider (2011)] to compare simulation strategies. These divergence measures that we use are α -divergence D_α , the Rényi α -divergence R_α , the Tsallis α -divergence T_α . We know that these divergences are all generalizations of the Kullback-Leibler divergence. We have the Kullback-Leibler divergence when the parameter α of these divergence measures tends to 1.

We will give in Section 2 description of these α -divergence measures and methodology for finding the best between two simulation strategies. We will give a consistent estimator for each divergence measure in Section 3. We will show in Section 4 that our divergence measures have a gaussian asymptotic distribution. Then we have some application examples in order to illustrate our methodology (Section 5). Finally, we do discussion in Section 6 to explain our various examples and see the lesson we can draw thereof.

2 Description of divergence measures and methodology for finding the best between two simulation strategies

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 $p_i(x)$:

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

Basseville (1996) and a universal monotonicity property known under the name of generalized data data processing theorem (Merhav (2010)). Their topological properties are investigated in Csiszár (1967). An extension of the family of ϕ -divergences to squared metric distances is introduced in Vajda (2009).

Definition 2.1. Kullback-Leiblers 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

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

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

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

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

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

2.2 Methodology for finding a good simulation strategy

Chauveau and Vandekerckhove (2007) have studied the strategies of simulation with Kullback-Leibler divergence. Their study was the comparison of simulation strategies by the Kullback-Leibler divergence which allowed to choose the best among candidate strategies. They used majority Metropolis-Hastings (MH) algorithm to explain their method. Thus, with a target density f , proposal density q and an initial distribution p_i^0 corresponding to the strategy i of MH algorithm they assumed several conditions on these densities to provide certain regularity properties, as the Lipschitz property, on the successive densities p_i^n of strategy i . This allowed to obtain a consistent estimate of the Kullback entropy

$$\mathcal{H}(p_i^n) = \int_{\mathcal{X}} p_i^n(x) \log(p_i^n(x)) \lambda(dx)$$

of p_i^n at time n .

In this present paper we propose to study how to find the optimal stochastic simulation algorithm may come from the MH methods (*IS* and *RWMH*), *Gibbs Sampler*, or recent adaptive (*AM*) and hybrid (*MWG*) methods. For that, we will use our three divergence measures. The interest of these divergences is that it subtracts the target density f , the proposal density q and initial density p_i^0 of many assumptions like in [Chauveau and Vandekerckhov (2012)].

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, if we consider for example the Rényi divergence, 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 the best.

- If the $R_\alpha(p_1^n, f)$ values are closer to 0 than $R_\alpha(p_2^n, f)$ values, we can then 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 the s_2 .

p_1^n is density function of X_n from Markov chain generated by strategy s_1

- If the reverse happens, ie $R_\alpha(p_2^n, f)$ values closer to 0 than $R_\alpha(p_1^n, f)$, we will say that the strategy s_2 is better than s_1 .
- If $R_\alpha(p_1^n, f) \simeq R_\alpha(p_2^n, f)$ and are very close to 0, we can say that the two simulation strategies s_1 and s_2 are both efficient.
- If $R_\alpha(p_1^n, f)$ and $R_\alpha(p_2^n, f)$ are far from 0 and slow to approach 0 or do not approach 0, then both strategies are ineffective.

3 Estimators of divergence measures

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

But very often in practice one can encounter, cases where the densities f and p^n are not analytically known. There are also cases where f is known to a multiplicative constant: the Bayesian context. It is often difficult to obtain the analytical form of p^n . For density f even if it is analytically known, we will estimate it. Then we use the methods of nonparametric estimation of probability densities. Póczos, Xiong and Schneider (2012) have proposed in their paper an estimator of probability density based on *k-NN method* (k nearest neighbor).

3.1 k-NN based density estimator

This is a method of density estimation proposed by [Loftsgaarden and Quesenberry (1965)]. Let a set of data training $E = \{x_1, x_2, \dots, x_N\}$ where x_i are independent observations with common probability density function g which is continuous. For a data z , the problem is the estimation of a probability density g at z . The principle is to find among $x_i, i = 1, \dots, N$ the k th nearest neighbor of z considering euclidean distance and we are interested by the distance between the two points ($x_{i,k}$ and z). If $x_{i,k}$ and z are points of $\mathbb{R}^d$ this distance $d(z, x_{i,k})$ allows us to calculate the volume of the d -dimensional hypersphere H about z with radius $r = d(z, x_{i,k})$.

The proposed estimator of g is

$$\hat{g}_{k,N}(z) = \frac{k/N}{V(H(z, d(z, x_{i,k})))} = \frac{k/N}{\pi^{d/2} r^d / \Gamma((d/2) + 1)}$$

where $V(H(z, r)) = \pi^{d/2} r^d / \Gamma((d/2) + 1)$ is the volume of d -dimensional sphere around $z \in \mathbb{R}^d$ with radius $r > 0$, $\Gamma(\cdot)$ is the Gamma function. Loftsgaarden and Quesenberry (1965) suggested after several experiments to take k equals to $\sqrt{N}$. We will round it if necessary to the nearest whole.

We find in [Póczos, and Schneider (2011)] theorem of convergence for density's estimators (k-NN estimators), showing the consistency of these estimators.

Theorem 3.1 (Almost sure convergence in sup norm). *If $\lim_{n \rightarrow \infty} \frac{k(n)}{\log(n)} = \infty$ and $\lim_{n \rightarrow \infty} n/k(n) = \infty$, then $\lim_{n \rightarrow \infty} \sup_x |\hat{g}_{k(n),N}(x) - g(x)| = 0$, almost surely.*

a) *Application to p^n*

Let $X_{1:N} = (X_1, \dots, X_N)$ be a simulated sample according to the law having density p^n with X_i i.i.d (independent identically distributed). We choose one X_i on the sample. Let $\rho_k(X_i)$ the euclidean distance between X_i and its k th nearest neighbor on the sample.

The estimator of p^n at point X_i is

$$\hat{p}_{k,n,N}(X_i) = \frac{k/(N-1)}{V(H(X_i, \rho_k(X_i)))} = \frac{k}{(N-1)c\rho_k^d(X_i)} \quad (3.1)$$

Denoting by $V(H(X_i, \rho_k(X_i))) = c\rho_k(X_i)^d$, where c stands for the volume of a d-dimensional unit ball.

b) *Application to f*

We should also have the i.i.d simulations according to the density f , but we can't directly simulate under density f that is why we apply MCMC simulation algorithms. Then after several iterations (when N is large) we can begin to recover samples. However, we take the precaution that two successive drawn points are separated by N_0 iterations. We get now the sample $Y_{1:M} = (Y_1, \dots, Y_M)$ of random variables which are "independent" under density f . Let $\gamma_k(X_i)$ the euclidean distance between X_i and its k th nearest neighbor on the sample $Y_{1:M}$.

The estimator of f at X_i is

$$\hat{f}_{k,M}(X_i) = \frac{k}{M c \gamma_k^d(X_i)} \quad (3.2)$$

3.2 Construction of estimators

Póczos, and Schneider (2011) have proposed an estimator of the common integral part to our three divergence measures. However, their first estimator is asymptotically biased. They have shown that if one multiply it by a appropriate constant, he will obtain an asymptotically unbiased estimator. Starting from that, we can make our estimators.

a) *Estimator for divergence D_α*

Begin by noting $M_\alpha(p^n, f) = \int (f(x)/p^n(x))^{1-\alpha} p^n(x) dx$ common integral part of our three α -divergence measures. By replacing f and p^n by their respective k -NN estimators we have the following estimator of $D_\alpha(p^n, f)$

$$\begin{aligned} \hat{D}_{\alpha,N,M,k}(p^n, f) &= \frac{1}{\alpha(1-\alpha)} \left(1 - \frac{1}{N} \sum_{i=1}^N \left(\frac{\hat{f}_{k,M}(X_i)}{\hat{p}_{k,n,N}(X_i)} \right)^{1-\alpha} \right) \\ &= \frac{1}{\alpha(1-\alpha)} \left(1 - \frac{1}{N} \sum_{i=1}^N \left(\frac{(N-1)\rho_k^d(X_i)}{M\gamma_k^d(X_i)} \right)^{1-\alpha} \right) \end{aligned}$$

We have now

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

that is an asymptotically biased estimator of $M_\alpha(p^n, f)$ [Póczos, and Schneider (2011)]. When we multiply it by the constant

$$B_{k,\alpha} = \frac{(\Gamma(k))^2}{\Gamma(k-\alpha+1)\Gamma(k+\alpha-1)} \quad (3.3)$$

independent of p^n and f , with $\Gamma(\cdot)$ the Gamma function. We obtain an asymptotically unbiased estimator of $M_\alpha(p^n, f)$. We obtain now a new asymptotically unbiased estimator of $D_\alpha(p^n, f)$

$$\hat{D}_{\alpha, N, M, k}(p^n, f) = \frac{1}{\alpha(1-\alpha)} \left(1 - \frac{1}{N} \sum_{i=1}^N \left(\frac{(N-1)\rho_k^d(X_i)}{M\gamma_k^d(X_i)} \right)^{1-\alpha} \times B_{k, \alpha} \right) \quad (3.4)$$

b) Estimator for Tsallis α -divergence

Applying the same process we obtain an asymptotically unbiased estimator $\hat{T}_{\alpha, N, M, k}(p^n, f)$ of $T_\alpha(p^n, f)$

$$\hat{T}_{\alpha, N, M, k}(p^n, f) = \frac{1}{\alpha-1} \left(\frac{1}{N} \sum_{i=1}^N \left(\frac{(N-1)\rho_k^d(X_i)}{M\gamma_k^d(X_i)} \right)^{1-\alpha} \times B_{k, \alpha} - 1 \right) \quad (3.5)$$

c) Estimator for Rényi divergence

The Rényi divergence measure has a particularity due to the presence of the logarithm function that takes $M_\alpha(p^n, f)$ as argument. We have not shown in this paper that its estimator is unbiased or asymptotically unbiased. However, we propose this consistent estimator

$$\hat{R}_{\alpha, N, M, k}(p^n, f) = \frac{1}{\alpha-1} \log \left(\frac{1}{N} \sum_{i=1}^N \left(\frac{(N-1)\rho_k^d(X_i)}{M\gamma_k^d(X_i)} \right)^{1-\alpha} \times B_{k, \alpha} \right) \quad (3.6)$$

We can check in [Póczos, and Schneider (2011)] that $\hat{M}_{\alpha, N, M, k}(p, q)B_{k, \alpha}$ is asymptotically unbiased, with p and q two probability densities. If we replace p and q respectively by p^n and f , then the following theorem provide conditions under which estimator $\hat{M}_{\alpha, N, M, k}(p^n, f) \times B_{k, \alpha}$ is L^2 consistent. With a function $F(x, p^n, \delta, 1/2)$ defined in [Póczos, and Schneider (2011)] we have

Theorem 3.2 (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 $x \in \mathbb{R}^d$, $\int \int \|x - y\|^\gamma p^n(y)p^n(x)dydx < \infty$, and that f is bounded above. Then*

$$\lim_{N, M \rightarrow \infty} \mathbb{E} \left((\hat{M}_{\alpha, N, M, k}(p^n, f)B_{k, \alpha} - M_\alpha(p^n, f))^2 \right) = 0 \quad (3.7)$$

that is, the estimator is L_2 consistent.

This theorem shows the consistency of $\hat{M}_{\alpha, N, M, k}(p^n, f)B_{k, \alpha}$ which is the estimator of the common integral part of the divergences. Therefrom, estimators $\hat{R}_{\alpha, N, M, k}(p^n, f)$, $\hat{D}_{\alpha, N, M, k}(p^n, f)$ and $\hat{T}_{\alpha, N, M, k}(p^n, f)$ are also consistent.

4 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.

4.1 Asymptotic distribution of D_α and T_α divergence estimators

Theorem 4.1. *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*

$$\hat{D}_{\alpha, N, k}(p^n, f) \xrightarrow{d} \mathcal{N} \left(D_\alpha(p^n, f), \frac{\sigma^2 B_{k, \alpha}^2}{\alpha^2(1-\alpha)^2} \right), \text{ when } N \rightarrow \infty. \quad (4.1)$$

$$\hat{T}_{\alpha,N,k}(p^n, f) \xrightarrow{d} \mathcal{N}\left(T_\alpha(p^n, f), \frac{\sigma^2 B_{k,\alpha}^2}{(\alpha-1)^2}\right), \text{ when } N \rightarrow \infty. \quad (4.2)$$

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

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

We have now

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

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

□

4.2 Asymptotic distribution for Rényi divergence estimator

Theorem 4.2. 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

$$\hat{R}_{\alpha,N,k}(p^n, f) \xrightarrow{d} \mathcal{N}\left(\frac{1}{\alpha-1} \log[M_\alpha(p^n, f)], \frac{\delta^2 B_{k,\alpha}^2}{(\alpha-1)^2 [M_\alpha(p^n, f)]^2}\right) \quad (4.3)$$

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

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

$$\sqrt{N}(\bar{Y}_N B_{k,\alpha} - M_\alpha(p^n, f)) \xrightarrow{d} \mathcal{N}(0, \delta^2 B_{k,\alpha}^2), \quad (4.4)$$

Relation (4.4) 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

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

□

5 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 divergence measures to compare proposal densities corresponding to a given simulation strategy and mainly

to compare two simulation strategies for a given target law. The proposal density or the simulation strategy is considered as optimum if the divergence measure (function of n) between densities p^n and the target density f tends quickly to 0 when n tends to ∞ . As suggested by [Loftsgaarden and Quesenberry (1965)] we will take k equals to the nearest integer of $\sqrt{N-1}$. We will also take $N = M$.

5.1 One-dimensional case

5.1.1 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 the α -divergence measure estimator. Indeed densities p^n will always be estimated here because we can not explicitly know them. This is why even if density f is analytically known, we will take its k -NN estimator in order to be in terms of the divergence estimators presented in [Póczos, and Schneider (2011)].

First we show an example that compares two proposal densities for a given strategy. We choose the *Independence Sampler (IS)*, which is one of the strategies of the Metropolis Hastings algorithm, to compare these densities.

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

For a given simulation strategy the choice of good proposal density is important. Indeed, for achieving satisfactory simulation results it is important to choose a good proposal density. For simplicity, we consider densities that differ by the value of their parameters. Thus we take respective distribution densities of $\mathcal{N}(-3, 2)$ and $\mathcal{N}(0, 3)$; the target density is from normal standard distribution $\mathcal{N}(0, 1)$. It is found by looking at Figure 1 that the dashed curve converges very rapidly to 0 while the solid curve is slow to converge.

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 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*.

5.1.2 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.

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

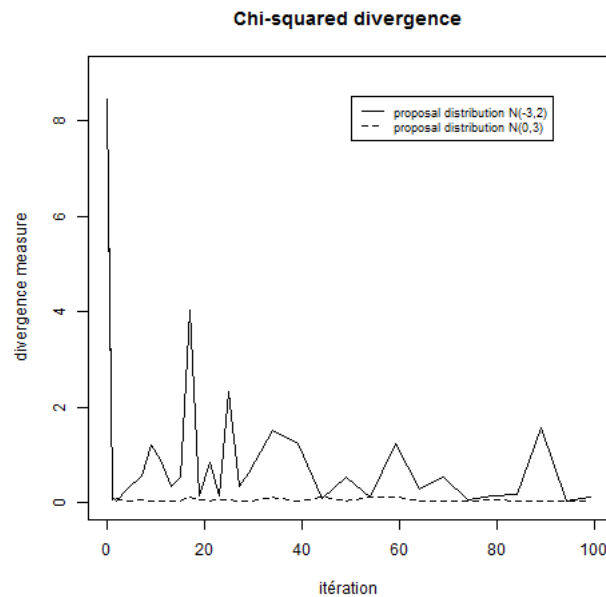


Figure 1: Comparison of two proposal densities, using the Independence Sampler and the Chi-squared divergence (D_α with $\alpha = 2$)

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 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 $f(x) \propto \exp(-x^2)(2 + \sin(5x) + \sin(2x))$.

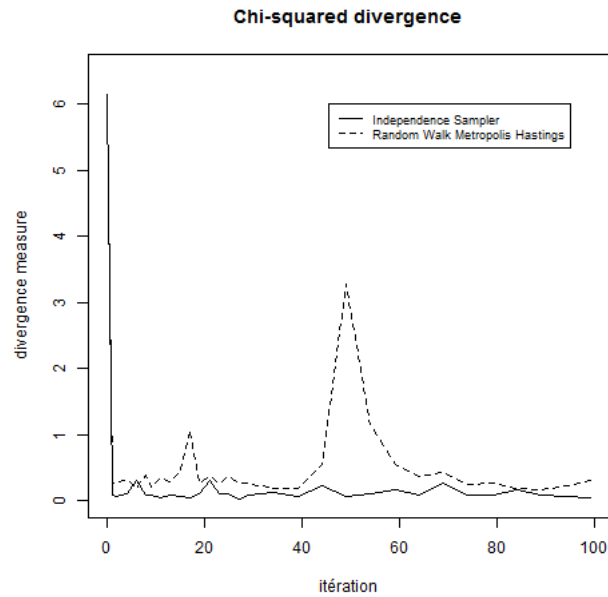


Figure 2: Comparison of two simulation strategies: *IS* vs *RWMH*, using the Chi-squared divergence (D_α with $\alpha = 2$)

5.2 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

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

where

$$\begin{aligned} M &= \frac{\sigma_0^2 \sum_{i=1}^n x_i + \sigma^2 m_0}{\sigma^2 + n\sigma_0^2} \quad \text{and} \quad \Sigma^2 = \frac{\sigma^2 \sigma_0^2}{\sigma^2 + n\sigma_0^2} \\ \sigma^2|m, x &\sim \mathcal{IG}\left(\frac{n}{2} + \alpha, \frac{1}{2} \sum_{i=1}^n (x_i - m)^2 + \beta\right) \end{aligned}$$

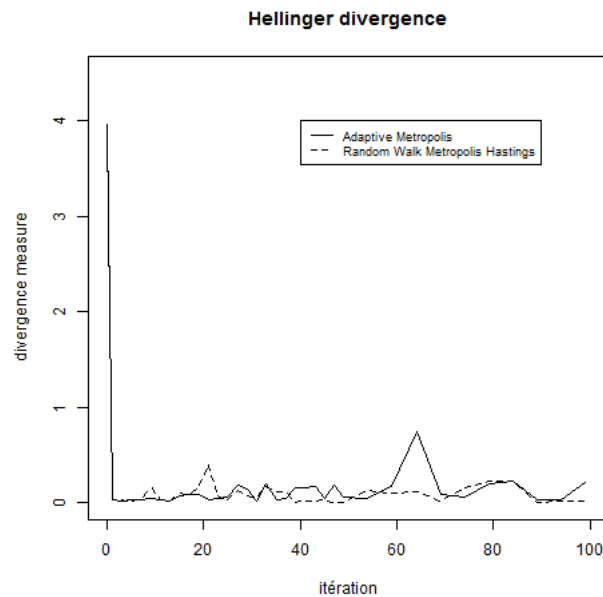


Figure 3: Comparison of two simulation strategies: *AM* vs *RWMH*, using the Hellinger divergence (D_α with $\alpha = 1/2$).

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 6 (Discussion).

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.

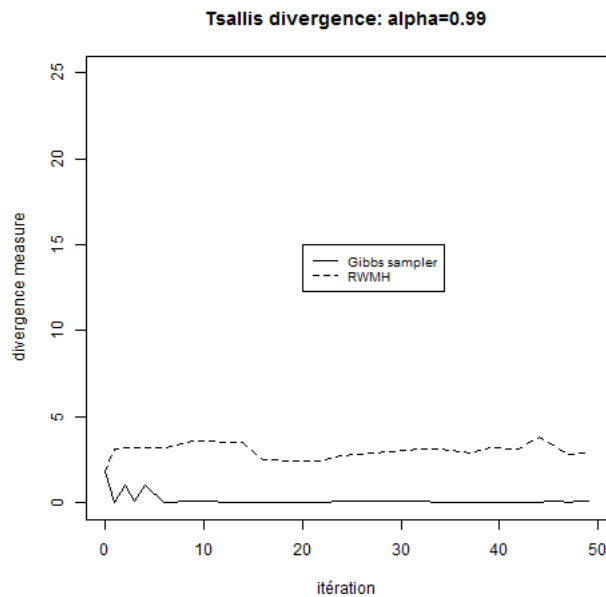


Figure 4: Comparison of two simulation strategies: GS vs RWMH, using the Tsallis divergence with $\alpha = 0.99$.

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.

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. Accepte the candidate Y with probability

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

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.

6 Discussion

6.1 Indepence 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 $D_2(p_2^n, f)$.

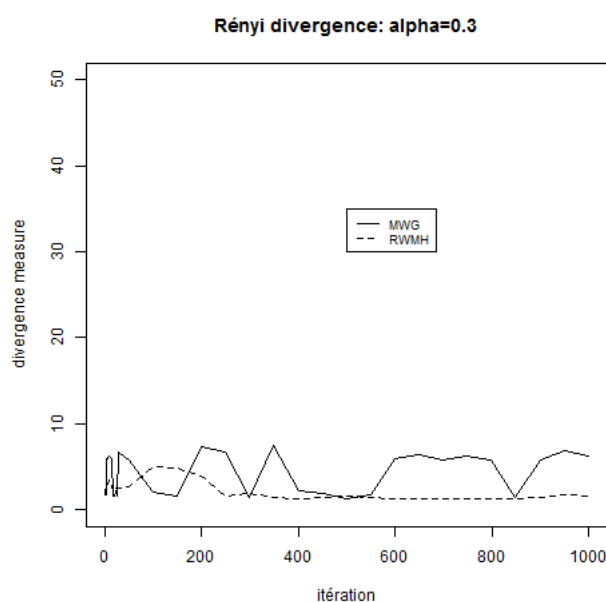


Figure 5: Comparison of two simulation strategies: *MWG* vs *RWMH*, using the Rényi α -divergence with $\alpha = 0.3$.

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.

6.2 Independence Sampler - RWMH

We have earlier said that the two curves in Figure 2 converge 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.

6.3 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.

6.4 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.

6.5 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.

6.6 Conclusion

We have shown that with various divergence measures one can compare two different simulation strategies for determining the best. We need these comparisons because after convergence, the computation time determines the effectiveness of a simulation algorithm. To achieve this, we have used α -divergence measures that we have described in Section 2. We have proposed estimators for our α -divergence measures in Section 3. We have given asymptotic distribution of D_α , T_α and R_α divergence estimators (Section 4). We recall that we had a Gaussian asymptotical distribution. Then in Section 5, we have given some examples for implementing our diverse simulation strategies and thus illustrate our methodology. Finally we have discussed in Section 6 of the causes that make that one simulation strategy is more powerful than the other.

References

- Chauveau D. and Vandekerkhov P. (2007). How to compare MCMC simulation strategies? *hal-00019174 (version 3)*.
- Chauveau D. and Vandekerkhov P. (2012). Smoothness of Metropolis-Hastings algorithm and application to entropy estimation. *ESAIM: Probability and Statistics* (17), 419-431.
- Cichocki A., Lee H., Y. D. Kim and Choi S., (2008). Non-negative matrix factorization with α -divergence. *Pattern Recognition Letters*.
- Cichocki A. and Amari S. (2010). Families of Alpha- Beta- and Gamma- Divergences: Flexible and Robust Measures of Similarities. *Entropy*,1-41.
- Haario H., Saksman E. and Tamminen J. (2001). An adaptive Metropolis algorithm. *Bernoulli* 7,2,223-242.

- 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.
- Loftsgaarden D. O. and Quesenberry C. P. (1965). A Nonparametric Estimate of a Multivariate Density Function. *Annals of Mathematical Statistics*, 36,3, 1049–1051.
- Mehta2
- 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.
- Ngom P. Ntep B. (2012). Minimum Penalized Hellinger Distance for Model Selection in Small Samples. *Open Journal of Statistics* 2,(4), 369-382.
- Póczos B., and Schneider J. *Proceedings of the 14th International Conference on Artificial Intelligence and Statistics (AISTATS)*, 15,609-617.
- 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.
- Van Erven T. and Harremoës P. (2010). Rényi Divergence and Majorization. *arXiv: 1001.4448v3 [cs.IT]*.

©2011 Author1 & Author2; 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.