Optimisation of Smoothing Parameter of Diffeomorphism Kernel Estimate for Bounded Random Data

Molka Troudi[1] and Faouzi Ghorbel[2]
[1]*Laboratoire de Traitement du Signal, de l'Image et Reconnaissance des Formes, ENIT,BP 37, 1002, Tunis, Tunisie*
[2]*Laboratoire Cristal, ENSI, Campus Science Universiatire de la Manouba, 2010, La Manouba, Tunisie*

Keywords: Diffeomorphisme Kernel Estimate, Plug-in Algorithm, Banwidth.

Abstract: The Diffeomorphism Kernel Density Estimator (DKDE) requires the estimation of an optimal value of the bandwidth to ensure a reliable pdf estimation of bounded distributions. In this paper, we suggest to approach the optimal bandwidth value by adapting Plug-in algorithm to DKDE estimator. We will show that the proposal method allows better density estimation in the MISE sense. Otherwise, the Gibbs phenomenon completely disappears. These results are illustrated by some bounded and semi bounded distributions simulations.

1 INTRODUCTION

It is well known that the estimation of the probability density functions (pdf) is an important step in many applications.

In practice, the application of a best estimator improves the systems performances. For examples, the optimal scalar quantification which is based on the pdf estimates is an important step in Signal and image coder. The advanced hashing procedure which is known as an essential task in data basis indexing gives improvement in its performances when the pdf of signal or image features are well estimated. In pattern recognition systems, the application of the Bayesian classification rule needs the determination of the conditional pdf and the mixture one and so on…. The coder parameters, the used features in data base index systems or the shape descriptors could be confined to a bounded or a semi bounded intervals (Ghorbel et al., 2012). The pdf estimate of such bounded or semi bounded attributes which are modeled by a set of random variables, have some convergence problems in its border values known by the Gibbs phenomenon. For these raisons, some authors have recently developed new non parametric pdf estimate methods taking account of the data support. The Diffeomorphism kernel estimate is one of this pdf estimate kinds. In the present work, we propose an improvement of such method by optimizing its smoothing parameter value in the mean of the Mean Integrate Square Error (MISE).

The kernel method is one of the most popular non-parametric pdf estimation methods (Parzen, 1962). Nevertheless, the studied random variables are mostly subject to algebraic constraints (bounded or semi bounded support) which are not respected by kernel method. The orthogonal series estimators studied by Hall (Hall, 1982) represent a first solution to this problem. Unfortunately, a disadvantage related to the Gibbs phenomenon on the bias of these estimators is generally observed. Saoudi et al. (1994) (1997) and Ghorbel (2011) proposed a new attractive method based on the kernel method with an appropriately chosen regular change of variable. Indeed, thanks to a regular diffeomorphism, the pdf is estimated on the natural support of the random variable.

However, the choice of the bandwidth noted by h_N is very important. Several techniques have been proposed for optimal bandwidth selection for the usual Kernel Density Estimation (KDE) method (Jones, Marronl and Seather, 1991) (Bowman and Azzalini, 1997). We focus in this paper on the plug-in method (Hall and Marron, 1987) which gives a good approximation of the optimal bandwidth in the mean integrated square error (MISE) sense. This method achieves approximation of the bandwidth h_N by an iterative approximation of second derivative of the density f, noted by $J(f)$. Thus, a sequence of

Troudi M. and Ghorbel F..
Optimisation of Smoothing Parameter of Diffeomorphism Kernel Estimate for Bounded Random Data.
DOI: 10.5220/0004073600560061
In *Proceedings of the International Conference on Signal Processing and Multimedia Applications and Wireless Information Networks and Systems* (SIGMAP-2012), pages 56-61
ISBN: 978-989-8565-25-9

positive numbers $h_N^{(k)}$ is constructed through the iterations with N as the sample size, and k as the number of iterations. Yet, we propose to adjust the plug-in method to the modified KDE method in order to obtain a better approximation.

The present paper is organized as follows. Section 2 is devoted to recall the Kernel pdf estimate method. In Section 3, the theoretical principles of the modified KDE which is adapted to the probability density functions with a bounded support are presented. The convergence according to mean square error criterion gives a sufficient condition so that the estimator converges in terms of the integrated mean square error (IMSE). An asymptotic study is developed in section 4. So, the expression of the optimal smoothing parameter is presented according to IMSE criterion. In section 5, we describe the different steps of the iterative plug-in algorithm witch converges to the optimal smoothing parameter or bandwidth. Therefore, section 6 is devoted to present some simulations in order to evaluate the performances of the suggested method.

2 KERNEL PDF ESTIMATE METHOD

The Kernel pdf Estimate is defined by:

$$\hat{f}_N(x) = \frac{1}{Nh_N}\sum_{i=1}^{N} K\left(\frac{x - X_i}{h_N}\right) \tag{1}$$

where $(X_i)_{1\le i\le n}$ is the observed data with length equal to n. h_N is called the bandwidth and K is a probability density function called the Kernel. K is assumed to be an even regular function with unit variance and zero mean. The evaluation of the performances of estimates methods is usually based on a measure of distance between the true density *f* and its estimate $\hat{f}_N$. Especially common choices are the Integrated Square Error (ISE) and its expected value, the Mean Integrated Square Error (MISE).

$$ISE(f,\hat{f}_N) = D_2\left(\hat{f}_N, f\right) = \int_{-\infty}^{+\infty} \left|\hat{f}_N(x) - f(x)\right|^2 dx$$

$$MISE = E\left[ISE\left(f,\hat{f}_N\right)\right] = E\left[\int_{-\infty}^{+\infty} \left|\hat{f}_N(x) - f(x)\right|^2 dx\right]$$

The minimisation of MISE with respect to the bandwidth, for a fixed size N of the sample, implies the following asymptotic study.

Let us consider the expression of Mean Square Error (MSE):

$$MSE = E\left[\left|\hat{f}_N - f\right|^2\right] = \mathrm{var}\left(\hat{f}_N\right) + \left(f - E\left[\hat{f}_N\right]\right)^2$$

The development of this expression gives the following formula

$$E\left[\left|\hat{f}_N - f\right|^2\right] = \frac{1}{Nh_N}\int K^2(u) f(x - h_N u)\,du + \left[\int K(u)\left(f(x - uh_N) - f\right)du\right]^2 - \frac{1}{N}\left(\int K(u) f(x - h_N u)\,du\right)^2$$

Firstly, let us consider the Taylor pdf expansion:

$$f(x - h_N u) = f(x) - h_N u f'(x) + \frac{u^2}{2} h_N^2 f''(x) - \frac{u^3 h_N^3}{6} f^{(3)}(x - \theta h_N u)$$

where $0 < \theta < 1$.

By using the following notations:

$$M(K) = \int_{-\infty}^{+\infty} K^2(u)\,du \tag{2}$$

and

$$J(f) = \int_{-\infty}^{+\infty} \left(f''(x)\right)^2 dx \tag{3}$$

where f'' is the second derivative of f.

$\Delta(hN)$, which is the Taylor expansion of the MISE (and consequently an approximation of MISE) is given by:

$$MISE \approx \Delta(h_N) = \frac{M(K)}{nh_N} + \frac{J(f)h_N^4}{4}$$

The minimum value of the function $\Delta(h_N)$ is obtained by annulling its derivative $\Delta'(h_N) = 0$.

$$\Delta'(h_N) = -\frac{M(K)}{nh_N^2} + h_N^3 J(f) = 0$$

Therefore, the optimal value of h_N noted by h_N^* becomes:

$$h_N^* = n^{-\frac{1}{5}} \cdot \left(J(f)\right)^{-\frac{1}{5}} \cdot \left(M(K)\right)^{\frac{1}{5}} \tag{4}$$

This minimum value of the MISE is given by the following expression:

$$MISE = \frac{5}{4} N^{-\frac{4}{5}} \left(M(K)\right)^{\frac{4}{5}} \left(J(f)\right)^{\frac{1}{5}} \quad (5)$$

3 DIFFEOMORPHISM KERNEL PDF ESTIMATE METHOD

The Diffomorphism Kernel Density Estimation method (DKDE) (Saoudi et al., 1994) (Saoudi et al., 1997); (Ghorbel, 2011) is based on appropriately chosen regular change of variable. Let $[X_1, X_2,\ldots,X_N]$ be N observations of random variable X and ϕ a C1-difféomorphism from]a , b[to R. The following estimator:

$$\hat{f}_N(x) = \frac{|\phi'(x)|}{Nh_N} \sum_{i=1}^{N} K\left(\frac{\phi(x)-\phi(X_i)}{h_N}\right) \quad (6)$$

is asymptotically unbiased when h_N tends towards 0 and ϕ'(x) tends towards infinity when x tends towards a or b which are the bounds of the interval]a, b[. The expectation of the suggested estimator is estimated by:

$$E\left[\hat{f}_N(x)\right] = \frac{|\phi'(x)|}{Nh_N} \sum_{i=1}^{N} E\left[K\left(\frac{\phi(x)-\phi(X_i)}{h_N}\right)\right]$$

Using the following change of variable, $y = \frac{\phi(x)-\phi(u)}{h_N}$ the expression of variance becomes:

$$E\left[\hat{f}_N(x)\right] = |\phi'(x)| \int_R K(y) \underbrace{fo\phi^{-1}\left(\phi(x)-uh_N\right)\left|\left(\phi^{-1}\right)'_{(\phi(x)-yh_N)}\right|}_{g(x,y)} dy^2$$

Let us compute the variance of this estimator by using the same change of variable:

$$\mathrm{var}\left[\hat{f}_N(x)\right] = \frac{|\phi'(x)|^2}{Nh_N} \int_R K^2(y) \underbrace{fo\phi^{-1}\left(\phi(x)-uh_N\right)\left|\left(\phi^{-1}\right)'_{(\phi(x)-yh_N)}\right|}_{g(x,y)} dy - \frac{1}{N}\left\{E\left[\hat{f}_N(x)\right]\right\}^2$$

The mean square error (MSE) is

$$E\left[\left|\hat{f}_N(x)-f(x)\right|^2\right] = \mathrm{var}\left[\hat{f}_N(x)\right] + \left\{E\left[\hat{f}_N(x)\right]\right\}^2 - 2f(x)E\left[\hat{f}_N(x)\right] + f^2(x)$$

The MSE becomes:

$$E\left[\left|\hat{f}_N(x)-f(x)\right|^2\right] = \frac{|\phi'(x)|^2}{Nh_N}\left\{\int_R K^2(y)g(x,y)dy - h_N\left[\int_R K(y)g(x,y)dy\right]^2\right\} + \left\{\int_R K(y)\left(|\phi'(x)|g_N(x,y)-f(x)\right)dy\right\}^2$$

As (ϕ-1)'is bounded on R, f is assumed to be bounded on]a, b[and K2 is integrable on R, the Lebesque convergence theorem can be easily applied, then the MSE, for a large value of N, becomes equivalent to:

$$E\left[\left|\hat{f}_N(x)-f(x)\right|^2\right] = \frac{|\phi'(x)|f(x)}{Nh_N}\int_R K^2(y)dy - \frac{f^2(x)}{N} + o(h_N)$$

Because of the continuity of ϕ' on R and f on]a, b[, this estimator converges in IMSE for all compact of]a, b[. To obtain this convergence according to the IMSE criterion, the function $\phi'(f)$ have to be integrable on]a, b[because:

$$\int_a^b E\left[\left|\hat{f}_N(x)-f(x)\right|^2\right]dx = \frac{1}{Nh_N}\int_a^b |f(x)\phi'(x)|dx \int_R K^2(y)dy - \frac{1}{N}\int_a^b f^2(x)dx + (b-a)o(h_N)$$

Saoudi and al. (Saoudi et al., 1994) shows that the logarithmic diffeomorphism allows a better convergence of the estimator.

$$\phi_{a,b} :]a,b[\to R, \quad x \to Log\left(\frac{x-a}{b-x}\right)$$

4 ASYMPTOTIC STUDY

The quality of the pdf estimation depends on the choice of the optimal smoothing parameter or bandwidth h_N. The MSE expression can be written as following:

$$E\left[\left|\hat{f}_N(x)-f(x)\right|^2\right]=A_N(x)+B_N(x)-C_N(x)$$

With

$$A_N(x)=\frac{|\phi'(x)|^2}{Nh_N}\int_R K^2(y)\,g(x,y)\,dy$$

$$B_N(x)=\left\{\int_R K(y)\left\lfloor|\phi'|g(x,y)-f(x)\right\rfloor dy\right\}^2$$

$$C_N(x)=\frac{|\phi'(x)|^2}{N}\left\{\int_R K(y)\,g(x,y)\,dy\right\}^2$$

We consider Taylor expansion of the function Hy defined as following in the neighborhood of $\phi(x)$:

$$\phi(x)\xrightarrow{H_y} fo\phi^{-1}(\phi(x)-yh_N)\left|(\phi^{-1})'(\phi(x)-yh_N)\right|$$

It implies that there exists a positive number θ less than 1 such that:

$$H_y(\phi(x)-yh_N)=H_y(\phi(x))-yh_N H_y'(\phi(x))$$
$$+\frac{y^2h_N^2}{2}H_y''(\phi(x))-\frac{y^3h_N^3}{6}H_y''(\phi(x)-\theta yh_N)$$

The following approximations are deduced from the computation of the successive derivatives of the function Hy in ϕ (x):

$$A_N(x)\approx\frac{|\phi'(x)|f(x)}{Nh_N}M(K)$$

$$B_N(x)=\frac{h_N^4}{4[\phi'(x)]^8}F^2(x)$$

$$C_N(x)=\frac{[f(x)]^2}{N}$$

with

$$F(x)=\left[f(x)\left[3\phi''(x)^2-\phi'(x)\phi'''(x)\right]\right]$$
$$-3f'(x)\phi'(x)\phi''(x)+f''(x)[\phi'(x)]^2 \qquad (7)$$

The asymptotical study of IMSE gives:

$$D^2(\hat{f}_N,f)=\int_R[A_N(x)+B_N(x)-C_N(x)]dx$$
$$\approx\frac{M(K)}{Nh_N}\int_R|\phi'(x)|f(x)\,dx+\frac{h_N^4}{4}\int_R\frac{F^2(x)}{[\phi'(x)]^8}dx$$

If M_ϕ and J_ϕ exists, we have:

$$M_\phi(K)=M(K)\int_R|\phi'(x)|f(x)\,dx \qquad (8)$$

and

$$J_\phi(f)=\int_R\frac{F^2(x)}{[\phi'(x)]^8}dx \qquad (9)$$

The optimal value of h_N noted by h_N^* can be deduced by minimization of IMSE.

$$h_N^*=[M_\phi(K)]^{\frac{1}{5}}[J_\phi(f)]^{-\frac{1}{5}}N^{-\frac{1}{5}} \qquad (10)$$

5 PLUG-IN DIFFEOMORPHISM KERNEL ESTIMATE ALGORITHM

Several methods are proposed in the literature for selecting optimal bandwidth parameter. The best known of these include rules of thumb, oversmoothing, least squares cross-validation, direct plug-in methods, solve-the-equation plug-in method, and the smoothed bootstrap (Jones et al., 1991); (Bowman and Azzalini, 1997); (Hall and Marron, 1987). We focus in this paper on the direct plug-in method applied to the kernel diffeomorphism application. Such a method is an iterative algorithm which converges to the optimal bandwidth.

Following, let's recall the steps of the plug-in algorithm.

Step 1: Arbitrary initialization of $M_\phi(K)$. For the experimentations of section 6, we chose to give to $M_\phi(K)$, the $M(K)$ value according to equation (2).

Step 2: Arbitrary initialization of $J_\phi^{(0)}(f)$ in order to determinate $h_N^{(0)}$ (equation (10)).

Step 3: Estimation of the pdf $f^{(0)}$ using $h_N^{(0)}$ and equation (6).

Step 4: At the k[th] iteration, estimation of $M_\phi(K)$ (equation (8)), $\left(f^{(k)}\right)'$ and $\left(f^{(k)}\right)''$.

Step 5: Estimation of $J_\phi\left(f^{(k)}\right)$ (equation (9)) and deduction of $h_N^{(k)}$ (equation (10)).

Step 6: Estimation of $f^{(k)}$ (equation (6)).

Step 7: Stopping the algorithm is conditional on a low relative difference betwee $h_N^{(k)}$ and $h_N^{(k-1)}$ (less than 1%).

6 PLUG-IN DIFFEOMORPHISM KERNEL ESTIMATE PERFORMANCES

In this section, we intend to compare the plug-in kernel diffeomorphism pdf estimator with the fast plug-in kernel pdf estimator which have been published in a previous work (Troudi et al., 2008). Three distributions are estimated: an exponential distribution (E(X) =1) which is semi bounded and defined on R+, a beta distribution (parameters = (2,2)) which is bounded and defined on [0, 1] and an uniform distribution defined on [0, 0.1].

6.1 Exponential Distribution

The estimation of the beta pdf by plug-in kernel pdf estimator (KDE) is presented in figure 1. Figure 2 represents this estimation by the plug-in diffeomorphism kernel pdf estimator (DKDE) which allows obviously a better estimation with an important reduction of Gibbs phenomenon. These results are corroborated by MISE values which are presented in table 1.

6.2 Beta Distribution

Figures 3 and 4 shows that the estimation of beta distribution pdf by the plug-in diffeomorphism kernel pdf estimator gives better results than those obtained by usual plug-in kernel pdf estimator. The Gibbs phenomenon is eliminated and the smoothing seems to be better. The MISE values versus the sample size presented in table 1 confirm these observations.

Table 1: MISE values versus sample size.

MISE for exponential pdf		
Sample size	Plug-in KDE	Plug-in DKDE
1000	3.79*10-5	7.76*10-6
2000	2.68*10-5	6.01*10-6
3000	2.00*10-5	5.51*10-6
4000	9.06*10-6	3.13*10-6
5000	7.80*10-6	2.81*10-6
MISE for beta pdf		
Sample size	Plug-in KDE	Plug-in DKDE
1000	0.0080	0.0058
2000	0.0044	0.0036
3000	0.0033	0.0027
4000	0.0024	0.0022
5000	0.0019	0.0018
MISE for uniform pdf		
Sample size	Plug-in KDE	Plug-in DKDE
1000	0.2502	0.1457
2000	0.1903	0.0746
3000	0.1651	0.0561
4000	0.1546	0.0445
5000	0.1451	0.0359

6.3 Uniform Distribution

The uniform distribution is known by its difficulties to be estimated. The Plug-in Diffeomorphism Kernel pdf estimate gives better results than the conventional Plug-in Kernel pdf estimate as it's shown in figures 5 and 6. Although the uniform distribution is well known by its difficulties in estimating, the DKDE method allows a better MISE values as is shown in table 1.

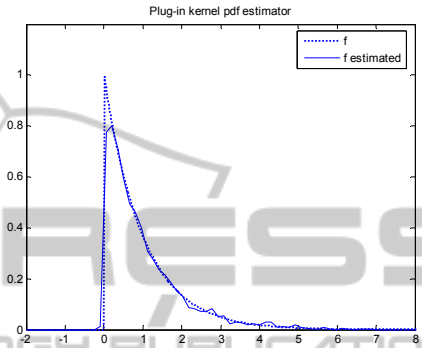

Figure 1: Pdf estimation of an exponential distribution by Plug-in Kernel Density estimator (KDE).

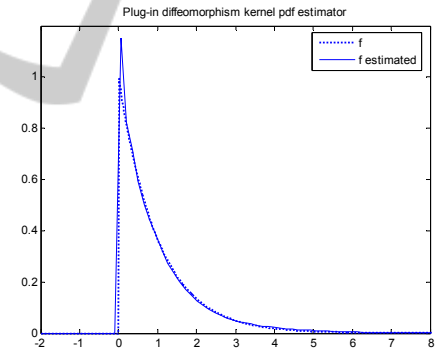

Figure 2: Pdf estimation of an exponential distribution by Plug-in Diffeomorphism Kernel Density Estimator (DKDE).

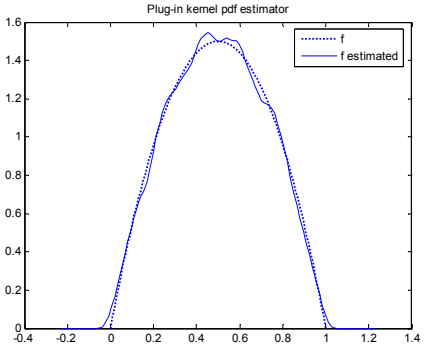

Figure 3: Pdf estimation of a beta distribution by Plug-in Kernel Density estimator (KDE).

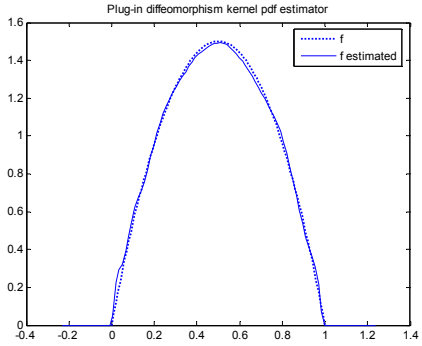

Figure 4: Pdf estimation of a beta distribution by Plug-in Diffeomorphism Kernel Density estimator (DKDE).

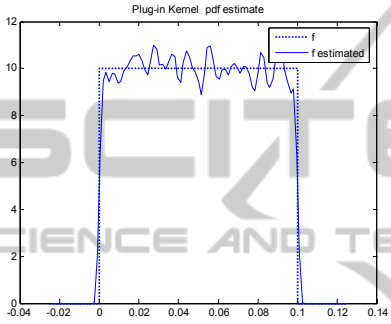

Figure 5: Pdf estimation of a uniform distribution by Plug-in Kernel Density estimator (KDE).

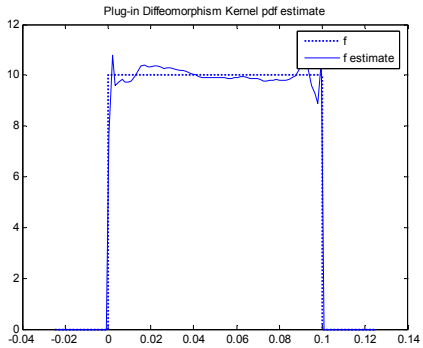

Figure 6: Pdf estimation of an exponential distribution by Plug-in Diffeomorphism Kernel Density Estimator (DKDE).

7 CONCLUSIONS

In this work, we have generalized the plug-in algorithm which adjusts the smoothing parameter of the kernel pdf estimate, to the diffeomorphism kernel estimate version. Such modified plug-in algorithm comes from the optimization of the MISE of this estimate. This generalization gives a more complicated iterative algorithm since the values of two parameters depending on the unknown pdf have to be approximated along iterations instead of only one parameter on the classical plug-in. It is important to note that the convergence is obtained for the proposed algorithm. By simulations concerning different kinds of distributions confined to bounded or semi bounded supports, we illustrate the better performance of the proposed Plug-in Diffeomorphism Kernel pdf estimate in the sense of MISE.

In our future works, we intend to study the case of multivariate bounded support distributions. We also test this well performance estimate in real data.

REFERENCES

Ghorbel, F., Derrode, S., Alata, O., 2012. Récentes avancées en reconnaissance de forme statistique. *Arts-Pi editions*, Tunis.

Parzen, E., 1962. On estimation of a probability density function and mode. *Annals of mathematical statistics*, 33, pp. 1065-1076.

Hall, P., (1982). Comparison of two orthogonal series methods of estimating a density and its derivatives on interval. *J. Multivariate anal*., 12, pp. 432 – 449.

Saoudi, S., Ghorbel, F., Hillion, A., (1994). Non parametric probability density function estimation on a bounded support: applications to shape classification and speech coding. *Applied Stochastic Models and Data Analysis*, 10, pp. 215-231.

Saoudi, S., Ghorbel, F., Hillion, A., (1997). Some statistical properties of the Kernel-diffeomorphism estimator. *Applied Stochastic Models and Data Analysis*,10, pp. 39-58.

Ghorbel, F., (2011). *Vers une approche mathématique unifiée des aspects géométriques et statistiques de la reconnaissance de formes planes*. Arts-Pi éditions, Tunis, 2d edition.

Jones, M. C., Marron, J. S., Seather, S. J., (1996). A brief survey of bandwidth selection for density estimation. *J. Amer. Stat. Assoc*., 91, pp. 401 – 407

Bowman, A.W., Azzalini, A.,(1997). *Applied Smoothing Techniques for Data Analysis*. Oxford University Press.

Hall, P., Marron, J. S., (1987). Estimation of integrated squared density derivatives. *Statistics &Probability letters*, 6, pp. 109 – 115.

Troudi, M., Alimi, A. M., Saoudi, S., (2008). Analytical Plug-in Method for Kernel Density Estimator Applied to Genetic Neutrality Study. *Eurasip Journal of advances in Signal Processing (Eurasip-JASP)*, 2008, Article ID 739082, 8 pages doi: 10.1155/2008/739082.