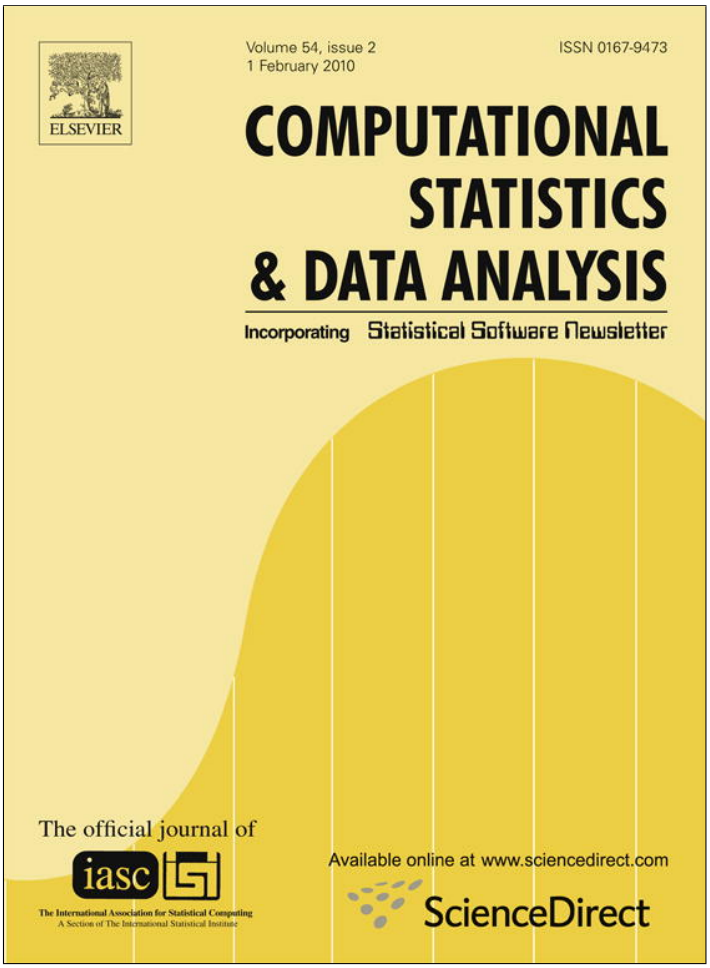

Computational Statistics and Data Analysis 54 (2010) 598–606

Contents lists available at ScienceDirect

Computational Statistics and Data Analysis

journal homepage: www.elsevier.com/locate/csda

# Clustering of discretely observed diffusion processes

Alessandro De Gregorio [a], Stefano Maria Iacus [b,*]

[a] Department of Statistics, Probability and Applied Statitics, University of Rome, P.le Aldo Moro 5, 00185 Rome, Italy

[b] Department of Economics, Business and Statistics, University of Milan, Via Conservatorio 7, 20124 Milan, Italy

ARTICLE INFO

*Article history:*
Received 27 April 2009
Received in revised form 6 October 2009
Accepted 7 October 2009
Available online 12 October 2009

ABSTRACT

A new distance to classify time series is proposed. The underlying generating process is assumed to be a diffusion process solution to stochastic differential equations and observed at discrete times. The mesh of observations is not required to shrink to zero. The new dissimilarity measure is based on the $L^1$ distance between the Markov operators estimated on two observed paths. Simulation experiments are used to analyze the performance of the proposed distance under several conditions including perturbation and misspecification. As an example, real financial data from NYSE/NASDAQ stocks are analyzed and evidence is provided that the new distance seems capable to catch differences in both the drift and diffusion coefficients better than other commonly used non-parametric distances. Corresponding software is available in the add-on package `sde` for the `R` statistical environment.



## 1. Introduction

In recent years, there has been a lot of interest in mining time series data. Although many measures of dissimilarity are available in the literature (see e.g. Liao, 2005, for a review) most of them ignore the underlying structure of the stochastic model which drives the data. Among the few measures which take into account the properties of the data generating model, we can mention Hirukawa (2006) which considers non-Gaussian locally stationary sequences; Piccolo (1990) proposed an AR metrics, Corduas and Piccolo (2008) used this distance to develop a clustering algorithm; Maharaj (1999) extended this metric to the multivariate case and Otranto (2008) adapted it to GARCH models. Caiado et al. (2006) used an approach based on periodograms; Xiong and Yeung (2002) proposed a model based clustering for mixtures of ARMA models. Kakizawa et al. (1998) and Alonso et al. (2006) performed clustering based on several information measures constructed on the estimated densities of the processes.

Needless to say, starting from the Black and Scholes (1973) and Merton (1973) theory, most of the models of modern finance rely on continuous time processes. In particular, the dynamics of underlying process used in option pricing is assumed to be a diffusion process solution to some stochastic differential equation. This paper proposes a new distance which is particularly tailored to discretely observed diffusion processes but not restricted to financial data. Indeed, diffusion processes are basic models in many fields like: physics (Papanicolaou, 1995), astronomy (Schuecker et al., 2001), mechanics (Kushner, 1967), economics (Bergstrom, 1990), geology (Ditlevsen et al., 2002), genetic analysis (Holland, 1976), ecology (Holmes, 2004), cognitive psychology (Tuerlink et al., 2001), neurology (Holden, 1976), biology (Ricciardi, 1977), biomedical sciences (Banks, 1975), epidemiology (Bailey, 1957), political analysis and social processes (Cobb, 1981) and many other fields of science and engineering.

This new dissimilarity is based on a new application of the results by Hansen et al. (1998) on identification of diffusion processed observed at discrete time when the time mesh $\Delta$ between observations is not necessarily shrinking to zero. The

* Corresponding author.
*E-mail addresses:* alessandro.degregorio@uniroma1.it (A. De Gregorio), stefano.iacus@unimi.it (S. Maria Iacus).

0167-9473/$ – see front matter © 2009 Elsevier B.V. All rights reserved.
doi:10.1016/j.csda.2009.10.005

theory proposed in Hansen et al. (1998) has been used in Kessler and Sørensen (1999) and Gobet et al. (2004) in parametric and non-parametric estimation of diffusion processes, respectively. The theory is based on the fact that, when the process is not observed at high frequency (i.e. $\Delta \not\to 0$) the observed data form a true Markov process for which it is possible to identify the Markov operator $P_\Delta$. The continuous time model is instead characterized by the infinitesimal generator $L_{b,\sigma}$, where $b$ and $\sigma$ are, respectively, the drift and diffusion coefficients of the process $\{X_t, t \geq 0\}$ solution to the stochastic differential equation $\mathrm{d}X_t = b(X_t)\mathrm{d}t + \sigma(X_t)\mathrm{d}W_t$. The two operators – $P_\Delta$ and $L_{b,\sigma}$ – are equivalent in the sense of functional analysis. Therefore, if one can estimate the Markov operator from the data, one can also possibly identify the process and, in particular, the couple $(b, \sigma)$. The identification step of this procedure needs some care (see e.g. Gobet et al., 2004) but this second step is not necessary in our approach. In the present paper, we make use only of the Markov operator to construct a distance between two observed processes (or better, between their Markov operators). Some form of ergodicity or stationarity of the underlying process is usually required, although these hypotheses can be relaxed in several directions, as for example mentioned in Kessler and Sørensen (1999), without affecting the results.

The paper is organized as follows. Section 2 introduces the model and the assumptions. The Markov operator is presented in Section 3. Section 4 studies the performance of the method. First, the behavior of the operator is analyzed on simulated paths with or without perturbation and misspecification for different sample sizes. Finally, real data from the NYSE/NASDAQ are analyzed. All the results include a comparison with three other dissimilarity measures, namely, the Euclidean distance, the Short-Time Series distance and the Dynamic Time Warping distance. Section 5 contains a brief explanation of the software which implements the proposed metric.

## 2. Model and assumptions

Let $I = (l, r)$, $-\infty \leqslant l < r \leqslant +\infty$ be the state space of a time-homogeneous diffusion process $\{X_t, t \geqslant 0\}$ solution to a stochastic differential of the form

$$\mathrm{d}X_t = b(X_t)\mathrm{d}t + \sigma(X_t)\mathrm{d}W_t, \tag{1}$$

with some initial condition $X_0 = x_0$. In Eq. (1) the functions $b : I \to \mathbb{R}$ and $\sigma : I \to (0, \infty)$ represent, respectively, the drift and diffusion coefficients, while $\{W_t, t \geq 0\}$ is a standard Brownian motion.

**Assumption 1.** The drift and diffusion coefficient are such that the stochastic differential equation (1) admits a unique weak solution $X_t$ (see, e.g., Karatzas and Shrevre, 1988).

Let us introduce the scale function and speed measure, defined respectively as

$$s(x) = \exp\left\{-2\int_{\tilde{x}}^{x} \frac{b(y)}{\sigma^2(y)}\mathrm{d}y\right\}, \tag{2}$$

with $\tilde{x}$ any value in the state space $(l, r)$, and

$$m(x) = \frac{1}{\sigma^2(x)s(x)}. \tag{3}$$

**Assumption 2.** We assume that

$$C_0 = \int_l^r m(x)\mathrm{d}x < \infty.$$

Let, $x^*$ be an arbitrary point in the state space of $X_t$ such that

$$\int_{x^*}^{r} s(x)\mathrm{d}x = +\infty, \qquad \int_{l}^{x^*} s(x)\mathrm{d}x = -\infty.$$

If one or both of the above integrals are finite, the corresponding boundary is assumed to be instantaneously reflecting.

If the Assumptions 1 and 2 are satisfied, then the process $X_t$ solution to (1) has an invariant law given by

$$\mu_{b,\sigma}(x) = \frac{m(x)}{C_0} = \frac{\exp\left\{2\int_{\tilde{x}}^{x} \frac{b(y)}{\sigma^2(y)}\mathrm{d}y\right\}}{C_0\sigma^2(x)}. \tag{4}$$

The above conditions are quite standard to perform inference for stochastic differential equations.

## 3. The Markov operator distance

Consider now the regularly sampled data $X_i = X(i\Delta)$, $i = 0, \ldots, N$, from the sample path of $\{X_t, 0 \le t \le T\}$, where $\Delta > 0$ and not shrinking to 0 and such that $T = N\Delta$. The process $\mathbf{X} = \{X_i\}_{i=0,\ldots,N}$ is a Markov process and, under mild regularity conditions, all the mathematical properties of the model are embodied in the transition operator defined as follows

$$P_\Delta f(x) = \mathrm{E}\{f(X_i)|X_{i-1} = x\}.$$

Notice that $P_\Delta$ depends on the transition density from $X_{i-1}$ to $X_i$, so we explicitly put the dependence on $\Delta$ in the notation. This operator is associated with the infinitesimal generator of the diffusion, namely the following operator on the space of continuous and twice differentiable functions $f(\cdot)$

$$L_{b,\sigma} f(x) = \frac{\sigma^2(x)}{2} f''(x) + b(x) f'(x).$$

When the invariant density $\mu = \mu_{b,\sigma}(\cdot)$ of the process $X_t$ exists, the operator is unbounded but self-adjoint negative on $L^2(\mu) = \{f : \int |f|^2 \mathrm{d}\mu < \infty\}$ and the functional calculus gives the correspondence (in terms of operator notation)

$$P_\Delta = \exp\{\Delta L_\mu\}. \tag{5}$$

This relation has been first noticed by Hansen et al. (1998) and Chen et al. (1997). It was then used in statistics to derive estimating functions by Kessler and Sørensen (1999). Indeed, to estimate, parametrically, the coefficients $\sigma(x) = \sigma_\theta(x)$ and $b = b_\theta(x)$ of (1) it suffices to note that

$$L_\theta f(x) = \frac{\sigma_\theta^2(x)}{2} f''(x) + b_\theta(x) f'(x)$$

can be seen as an eigenvalue problem $L_\theta \psi_\theta(x) = \kappa_\theta \psi_\theta(x)$ and the pair $(\kappa_\theta, \psi_\theta)$ satisfies

$$P_\Delta \psi_\theta(X_i) = \mathrm{E}\{\psi_\theta(X_{i+1})|X_i\} = \exp(\kappa_\theta \Delta)\psi_\theta(X_i).$$

If the solution of the above eigenvalue problem exists, it is possible to impose a set of moment conditions from which estimating functions can be obtained. More recently, under a low sampling rate, the result (5) was used to estimate, non-parametrically, the drift and diffusion coefficient by Aït-Sahalia (1996) and Gobet et al. (2004).

In this paper, we propose to use an estimator of $P_\Delta$, and from this object build a distance between discretely observed diffusion processes.

For a given $L^2$-orthonormal basis $\{\phi_j, j \in J\}$ of $L^2([l, r])$, where $J$ is an index set, following Gobet et al. (2004) it is possible to obtain the matrix $\hat{\mathbf{P}}_\Delta(\mathbf{X}) = [(\hat{P}_\Delta)_{j,k}(\mathbf{X})]_{j,k\in J}$, which is an estimator of $\langle P_\Delta \phi_j, \phi_k \rangle_{\mu_{b,\sigma}}$, where

$$(\hat{P}_\Delta)_{j,k}(\mathbf{X}) = \frac{1}{2N} \sum_{i=1}^{N} \left\{ \phi_j(X_{i-1})\phi_k(X_i) + \phi_k(X_{i-1})\phi_j(X_i) \right\}, \quad j, k \in J. \tag{6}$$

The terms $(\hat{P}_\Delta)_{j,k}$ are approximations of $\langle P_\Delta \phi_j, \phi_k \rangle_{\mu_{b,\sigma}}$, that is, the action of the transition operator on the state space with respect of the unknown scalar product $\langle \cdot, \cdot \rangle_{\mu_{b,\sigma}}$ and hence can be used as "proxy" of the probability structure of the model. Therefore, we introduce the following dissimilarity measure.

**Definition 1.** Let $\mathbf{X}$ and $\mathbf{Y}$ be discrete time observations from two diffusion processes. The Markov Operator distance is defined as

$$d_{MO}(\mathbf{X}, \mathbf{Y}) = \left\| \hat{P}_\Delta(\mathbf{X}) - \hat{P}_\Delta(\mathbf{Y}) \right\|_1 = \sum_{j,k\in J} \left| (\hat{P}_\Delta)_{j,k}(\mathbf{X}) - (\hat{P}_\Delta)_{j,k}(\mathbf{Y}) \right|, \tag{7}$$

where $(\hat{P}_\Delta)_{j,k}(\cdot)$ is calculated as in (6) separately for $\mathbf{X}$ and $\mathbf{Y}$.

Notice that $d_{MO}(\mathbf{X}, \mathbf{Y})$ is the element-wise $L^1$ distance for matrices, not simply a dissimilarity measure (i.e. it also respects the triangular inequality).

**Remark 1.** Like the invariant density $\mu_{b,\sigma}$, the Markov operator itself cannot perfectly identify the underlying process, in the sense that, for some $(b_1, \sigma_1)$ there might exist another couple $(b_2, \sigma_2)$ such that $\mu_{b_1,\sigma_1}(x) = \mu_{b_2,\sigma_2}(x)$. The same considerations apply to the infinitesimal generator and hence to the Markov operator. Nevertheless, the distance $d_{MO}$ helps in finding similarities between two (or more) processes in terms of the action of their Markov operators. The Markov operator also takes into account the probabilistic properties of the observed sequence, which is the natural way to make inference from discretely observed diffusion processes.

## 4. Performance of the Markov operator distance

In this section we use four different distances in both the analysis of synthetic and real data. Due to the fact that $d_{MO}$ is model based but completely nonparametric, we decide to compare its performance against other fully non-parametric metrics only. We avoid the use of metrics which require pre-estimates of parameters, model selection, etc. (e.g. the ARIMA-like metrics, Piccolo, 1990). Although these distances have been proved to be useful when the underlying generating model is a true discrete time series (Otranto, 2008), they fall outside our framework.

We denote by $\mathbf{X} = \{X_i, i = 1, \ldots, N\}$ and $\mathbf{Y} = \{Y_i, i = 1, \ldots, N\}$ two discretely observed data from continuous time diffusion processes. We compare the following distances.

*The Markov-operator distance*

The Markov operator distance $d_{MO}$ is calculated using formula (7). As in Reiß (2003) we deal with a basis 50 orthonormal B-splines on a compact support of degree 10 (see Ramsay and Silverman, 2005). As compact support, we consider the observed support of all simulated diffusion paths enlarged by 10%. In the analysis of synthetic data, the support is just the interval [0, 1].

*Short-time-series distance*

The Short-Time-Series distance proposed by Möller-Levet et al. (1978) is based on the idea to consider each time series as a piecewise linear function and compare the slopes between all the interpolants. It reads as

$$d_{STS}(X, Y) = \sqrt{\sum_{i=1}^{N} \left( \frac{X_i - X_{i-1}}{\Delta} - \frac{Y_i - Y_{i-1}}{\Delta} \right)^2}.$$

This measure is essentially designed to discover similarities in the volatility between two time series regardless of the average level of the process (i.e. one process and a shifted version of it will have zero distance).

*The Euclidean distance*

The usual Euclidean distance

$$d_{EUC}(X, Y) = \sqrt{\sum_{i=1}^{N} (X_i - Y_i)^2}$$

is one of the most used in the applied literature. We use it only for comparison purposes.

*Dynamic time warping distance*

The Euclidean distance is very sensitive to distortion in time axis and may lead to poor results for sequences which are similar, but locally out of phase (Corduas, 2007). The Dynamic Time Warping (DTW) distance was introduced originally in speech recognition analysis (Sakoe and Chiba, 1978; Wang and Gasser, 1997). DTW allows for non-linear alignments between time series not necessarily of the same length. Essentially, all shifts between two time series are attempted and each time a cost function is applied (e.g. a weighted Euclidean distance between the shifted series). The minimum of the cost function over all possible shifts is the dynamic time warping distance $d_{DTW}$. In our applications, we use the Euclidean distance in the cost function and the algorithm as implemented in the R package `dtw` (Giorgino, 2009).

### 4.1. Analysis of synthetic data

We simulate 23 paths according to the six different models $M_j, j = 1, \ldots, 6$, obtained via the combinations of drift $b_k$ and diffusion coefficients $\sigma_k, k = 1, \ldots, 4$ presented in the following table

| | $\sigma_1(x)$ | $\sigma_2(x)$ | $\sigma_3(x)$ | $\sigma_4(x)$ |
|---|---|---|---|---|
| $b_1(x)$ | M1 | | M4 | |
| $b_2(x)$ | | M2 | M3 | |
| $b_3(x)$ | | M5 | | |
| $b_4(x)$ | | | | M6 |

where

$$b_1(x) = 1 - 2x, \qquad b_2(x) = 1.5(0.9 - x), \qquad b_3(x) = 1.5(0.5 - x), \qquad b_4(x) = 5(0.05 - x)$$

and

$$\sigma_1(x) = 0.5 + 2x(1-x), \qquad \sigma_2(x) = \sqrt{0.55x(1-x)},$$
$$\sigma_3(x) = \sqrt{0.1x(1-x)}, \qquad \sigma_4(x) = \sqrt{0.8x(1-x)}.$$

For each model $M_j$ we simulate a different number $n_j$ of trajectories to have an unbalanced simulation design, i.e. we set $n_1 = 5, n_2 = 3, n_3 = 4, n_4 = 3, n_5 = 4, n_6 = 1$. We explicitly avoid a balanced design (i.e. all $n_i$ equal) to increase the difficulty of the task. Indeed, the cluster algorithm, not only has to identify the number of cluster, but these clusters have different sizes as well. Further, we take one trajectory generated with model $M_1$, say $X^1$, and reverse it around the line $y = 1$, i.e. if $1 - X^1 = \tilde{X}^1$, hence $\tilde{X}^1$ has drift $-b_1(x)$ and the same quadratic variation of $X^1$. So it still belongs to the class $M_1$ with respect to volatility. We also consider an additional trajectory simulated using model $M_1$ but with a different initial value. By the ergodic property of the simulated path, its invariant law still belongs to model $M_1$. Therefore, we have $n_1 = 7$.

We simulate each path using (second) Milstein scheme (see e.g. Kloden et al., 2000 or Iacus, 2008) with time lag $\delta = 1e{-}3$. Observations have been then resampled at rate $\Delta = 0.01$ and observed paths of length $N = 500$ and $N = 1000$ have been used in the analysis in order to capture sample size effects.

Due to the fact that the number of clusters is known in advance (i.e. $K = 6$) we used a cluster similarity index proposed in Gravilov et al. (2000) defined as follows. Given two clustering $C = C_1, \ldots, C_K$ (the real clusters formed by our six models) and $C' = C'_1, \ldots, C'_{K'}$ (the clustering obtained using one of the above distances), we compute the following similarities

$$\mathrm{sim}(C_i, C'_j) = 2\frac{|C_i \cap C'_j|}{|C_i| + |C'_j|}, \quad i = 1, \ldots, K, j = 1, \ldots, K',$$

and the final cluster similarity index is given by the formula

$$\mathrm{Sim}(C, C') = \frac{1}{K}\sum_{i=1}^{K} \max_{j=1,\ldots,K'} \mathrm{sim}(C_i, C'_j). \tag{8}$$

In the application to real data, we also apply the symmetrized version of the index, namely $(\mathrm{Sim}(C, C') + \mathrm{Sim}(C', C))/2$, because the real number of clusters is not known in advance. In formula (8) $K$ and $K'$ may be different, although in our next two experiments they will be the same number. The similarity index will return 0 if the two clusterings are completely dissimilar and 1 if they are the same. The index is not symmetric, so we will always use as argument $C$ the true clustering and as $C'$ the clustering obtained applying one of the four distances. We performed four different experiments. In all cases we run hierarchical clustering with the complete linkage method.

The four experiments are performed according to the following steps:

*Experiment 1: Non perturbed, correctly specified*

Simulate according the above scheme 25 trajectories; calculate the distance matrices $d_{MO}$, $d_{STS}$, $d_{EUC}$ and $d_{DTW}$ and run clustering. Cut the dendrograms into $K = K' = 6$ groups. Calculate the Sim index for each clustering solution.

*Experiment 2: Non perturbed, miss specified*

Simulate according the above scheme 25 trajectories; calculate the four distances and run cluster analysis. Cut the dendrograms into $K' = 5$ groups, real number of groups $K = 6$. Calculate the Sim index for each clustering solution.

*Experiment 3: Perturbed, correctly specified*

Simulate according the above scheme 25 trajectories. Perturbate the experiment, adding 2 trajectories from an ARIMA(1, 0, 1) process with mean 0.5 and AR coefficient 0.9, MA coefficient $= -0.22$, with Gaussian innovations $N(0, 0.01)$ (the parameters of the model are chosen in a way that the simulated trajectories look qualitatively similar to the ones in the Experiment 1). Calculate the four distances, use the same clustering approach as in Experiment 1, set $K = 7$ and cut the dendrograms into $K' = 7$ groups.

*Experiment 4: Perturbed, miss specified*

Proceed as in Experiment 3, set $K = 7$ and cut the dendrograms $K' = 6$ groups.

Each experiment is replicated only 100 times and the average value of the cluster similarity index Sim is reported in Table 1 for different sample sizes $N = 500$ (up) and $N = 1000$ (down). The number of replications is limited due to excessively long computational time of the DTW distance in dimension 23. To test the stability of the Monte Carlo results of the first few 100 replications, we drop $d_{DTW}$ from the Monte Carlo analysis and replicate each of the four experiments 5000 times. Table 1 also reports, in parenthesis, the average values but calculated over the 5000 replications.

Experiment 3 corresponds to a perturbation of the diffusion setup with an ARIMA process, while Experiment 4 corresponds to a misspecified setting: there are $K = 7$ real clusters, but we induce misclassification, selecting only $K' = 6$ groups. In Experiment 2 there is only misspecification where the number of real groups $K$ is higher than the number of the groups generated with the cluster $K'$.

As emerges from the analysis of Table 1 we see that all methods perform better in Experiment 1, although a clear ordering – for all experiments – emerges in the different metrics to discover the correct groups: $d_{MO} \prec d_{DTW} \prec d_{EUC} \prec d_{STS}$, where $d_1 \prec d_2$ means: "distance $d_1$ classifies *better* than $d_2$". In the case of perturbation (Experiment 3) one should expect that the

**Table 1**
Results of the simulation experiments. Average values of the Sim index over 100 replications and, with the exclusion of the $d_{DTW}$ distance, 5000 replications (in parentheses). $\Delta = 0.01$, sample size $N = 500$ (up) and $N = 1000$ (bottom).

| Experiment | $d_{MO}$ | $d_{EUC}$ | $d_{STS}$ | $d_{DTW}$ |
|---|---|---|---|---|
| Non perturbed, correctly specified | 0.84 | 0.49 | 0.27 | 0.69 |
| | (0.83) | (0.49) | (0.27) | (–) |
| Non perturbed, miss specified | 070 | 0.44 | 0.24 | 0.60 |
| | (0.69) | (0.43) | (0.24) | (–) |
| Perturbed, correctly specified | 0.81 | 0.45 | 0.39 | 0.65 |
| | (0.81) | (0.45) | (0.39) | (–) |
| Perturbed, miss specified | 0.71 | 0.41 | 0.37 | 0.58 |
| | (0.70) | (0.41) | (0.37) | (–) |
| Non perturbed, correctly specified | 0.94 | 0.51 | 0.27 | 0.69 |
| | (0.93) | (0.50) | (0.26) | (–) |
| Non perturbed, miss specified | 0.75 | 0.45 | 0.24 | 0.63 |
| | (0.75) | (0.45) | (0.24) | (–) |
| Perturbed, correctly specified | 0.91 | 0.47 | 0.39 | 0.67 |
| | (0.90) | (0.46) | (0.39) | (–) |
| Perturbed, miss specified | 0.78 | 0.43 | 0.37 | 0.59 |
| | (0.78) | (0.42) | (0.37) | (–) |

Markov Operator distance should fail to detect the ARIMA group, and instead should not expect any change in performance of the other metrics because they do not assume a particular stochastic structure of the model. But Table 1 shows that all methods are equally affected and $d_{MO}$ looks quite robust. Although there is a decrease of performance of $d_{MO}$ in the misspecified case (Experiment 4), the $d_{MO}$ distance still performs much better than the other competitors. All methods increase performance as the number of observations $N$ increases, but the enhancement of the $d_{MO}$ is particularly remarkable. This is due to the property of the estimator of the Markov Operator, which gets better and better as the sample size increases.

In conclusion, in order to classify diffusion processes by means of discrete time observations, $d_{MO}$ seems to be the best distance.

### 4.2. Analysis of real data

As an example of application of this method to real data, we consider time series of daily closing quotes, from 2006-01-03 to 2007-12-31, for the following 20 financial assets: Microsoft Corporation (MSOFT in the plots), Advanced Micro Devices Inc. (AMD), Dell Inc. (DELL), Intel Corporation (INTEL), Hewlett-Packard Co. (HP), Sony Corp. (SONY), Motorola Inc. (MOTO), Nokia Corp. (NOKIA), Electronic Arts Inc. (EA), LG Display Co., Ltd. (LG), Borland Software Corp. (BORL), Koninklijke Philips Electronics NV (PHILIPS), Symantec Corporation (SYMATEC), JPMorgan Chase & Co. (JMP), Merrill Lynch & Co., Inc. (MLINCH), Deutsche Bank AG (DB), Citigroup Inc. (CITI), Bank of America Corporation (BAC), Goldman Sachs Group Inc. (GSACHS) and Exxon Mobil Corp. (EXXON). Quotes come from NYSE/NASDAQ. Source Yahoo.com. Missing values (the same 19 festivity days over 520 daily data) have been linearly interpolated. These assets come from both electronic hardware, appliance and software vendors or producers, financial institutions of different types, and a petrol company. Fig. 1 represents the 20 paths of the assets all on the same scale in order to make them comparable by visual inspection. It is clear that some titles have larger volatility than others and possibly there are some outliers (e.g. BORL) in terms of both trend and volatility. For example, looking at financial companies, one can notice that MLINCH, DB and GSACHS, although, at different volatility levels, all present the same (cyclic) drift behavior over time. Furthermore, CIT and BAC seem quite close in terms of volatility and drift. But visual inspection alone is not sufficient, so we try to discover clusters using the four distances introduced in Section 4.

In Fig. 2 the dendrogram for the $d_{MO}$ distance, identifies 5 or 6 groups and in particular isolates BORL and "DB + GSACHS" into separate clusters very clearly (the difference between 5 and 6 groups is that, in the 6 groups clustering, "MLINCH + EXXON" are put in a separate cluster). To isolate the BORL asset via the dendrograms of $d_{L2}$ and $d_{DTW}$, we need to cut it into at least 6 groups. The counter effect of this cutting is that DB and GSACHS go into different clusters for these metrics. The metric $d_{STS}$ does not appear to give a sharp indication on how to separate clusters. We have then decided to cut all the dendrograms into 6 groups.

The similarity matrix in Table 2 shows that $d_{L2}$ and $d_{DTW}$ form the same groups, i.e. they are essentially the same metric for this data set. The clustering made using $d_{MO}$ is only partially in agreement with $d_{L2}$ and $d_{DTW}$ (84%). The difference is mainly in the placement of the subgroups "HP + PHILIPS" and "MSOFT + DELL". Further, the $d_{MO}$ distance considers "DB + GSACHS" together, which makes sense for this distance, probably because these two time series have the highest volatilities.

EA goes together with SONY in all dendrograms, which is not an unrealistic evidence in that the company essentially produces software for game consoles. Also for CITI, BAC and JPM the methods agree on their placement.

### 4.3. Summary

In conclusion, all but the $d_{STS}$ distance provide similar evidence. Nevertheless, we think $d_{MO}$ easily separates BORL (outlier) and "GSACHS + DB", while, with the other two competitors, in order to separate BORL, we need to force an additional

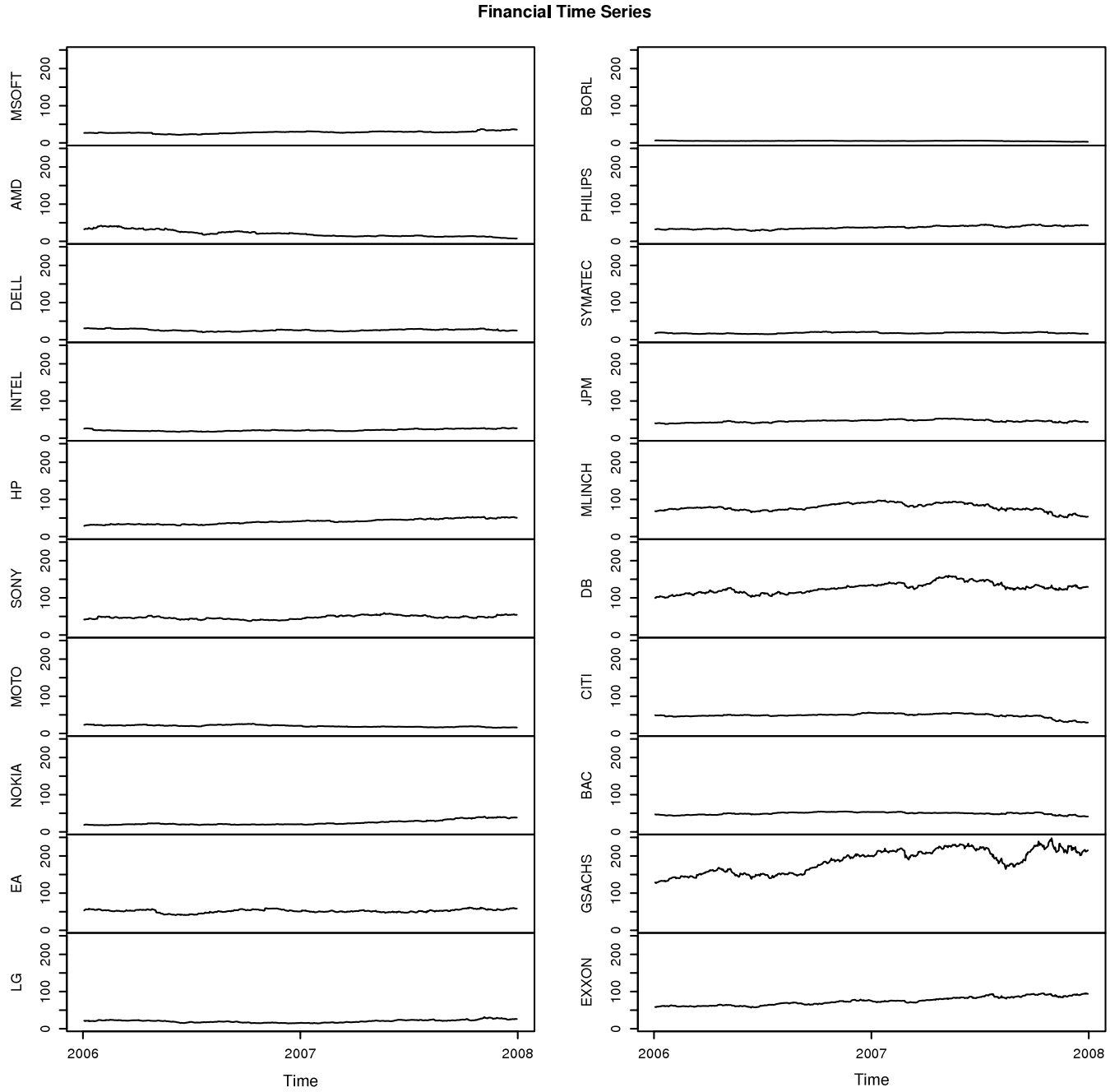


**Fig. 1.** Paths of the 20 assets considered: from 2006-01-03 to 2007-12-31.

**Table 2**
Similarity matrix between the clusters formed by different metrics. Similarity calculated according to the similarity index defined in (8) (left table) and its symmetrized version (right table).

| | $d_{MO}$ | $d_{L2}$ | $d_{STS}$ | $d_{DTW}$ |
|---|---|---|---|---|
| $d_{MO}$ | 1.00 | 0.84 | 0.60 | 0.84 |
| $d_{L2}$ | 0.79 | 1.00 | 0.71 | 1.00 |
| $d_{STS}$ | 0.48 | 0.67 | 1.00 | 0.67 |
| $d_{DTW}$ | 0.79 | 1.00 | 0.71 | 1.00 |

| | $d_{MO}$ | $d_{L2}$ | $d_{STS}$ | $d_{DTW}$ |
|---|---|---|---|---|
| $d_{MO}$ | 1.00 | 0.81 | 0.54 | 0.81 |
| $d_{L2}$ | | 1.00 | 0.69 | 1.00 |
| $d_{STS}$ | | | 1.00 | 0.69 |
| $d_{DTW}$ | | | | 1.00 |

splitting which separates GSACHS and DB. This looks quite unfortunate from a substantial point of view. Of course, this is merely an exercise and the analysis cannot go deeper than this from a simple cluster analysis. In fact, other financial and economics considerations have to be made, but these considerations fall outside the aim of the present paper. Still, the simulation analysis of the previous section seems to ensure that the proposed method is robust to misspecification and perturbation, and quite efficient in the presence of correct specification.

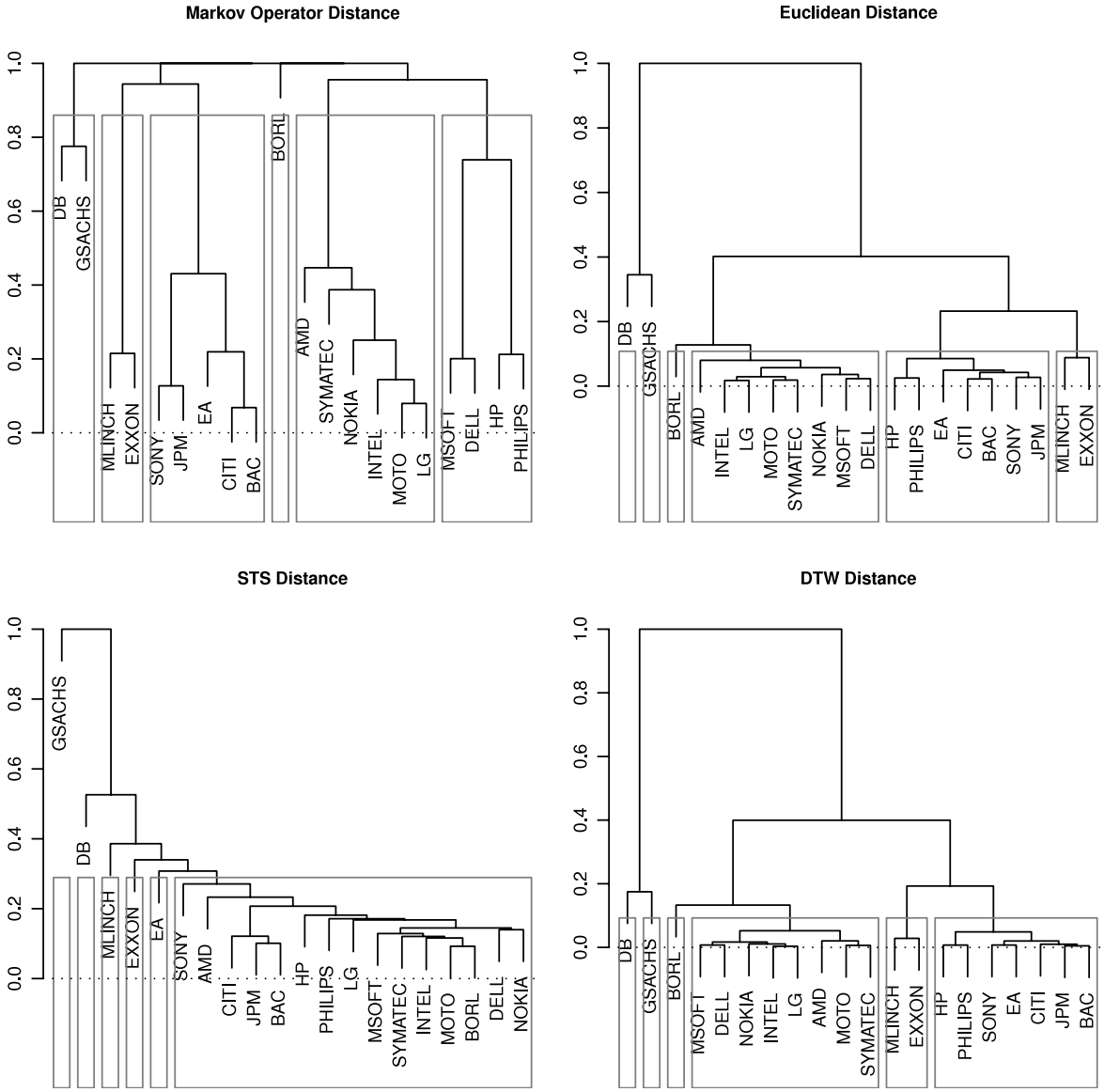


**Fig. 2.** Clustering according to different distances. Distances normalized to 1 just for graphical representation. Although the markers of the terminal nodes go below the zero line (see e.g. right-bottom plot), the final nodes are obtained cutting the dendrogram above the zero line, which is represented as a dotted line just to help visualization.

## 5. Software

This metric is freely available as the function `MOdist` in the add-on package `sde` for the R statistical environment (R Development Core Team, 2009). The function `MOdist` outputs an object of class `dist` which can be used as input to any clustering procedure implemented in R. The software can be installed from within R with command `install.packages("sde")`. The following code is an example of use in a R session:

```
> library(sde)
> data(quotes)
> plot(quotes)
> d <- MOdist(quotes)
> cl <- hclust(d)
> plot(cl)
```

In the above R code, `data(quotes)` loads the NYSE/NASDAQ data of previous section.

## Acknowledgments

We wish to thank the Associate Editor and three anonymous referees for their helpful suggestions which improved the first version of this paper.

## References


Aït-Sahalia, Y., 1996. Nonparametric pricing of interest rate derivative securities. Econometrica 64, 527–560.

Alonso, A.M., Berrendero, J.R., Hernández, A., Justel, A., 2006. Time series clustering based on forecast densities. Computational Statistics & Data Analysis 51 (2), 762–776.

Bailey, N., 1957. The Mathematical Theory of Epidemics. In: Lecture Notes in Biomathematics, Springer-Verlag, Griffin, London.

Banks, H., 1975. Modeling and Control in the Biological Sciences. In: Lecture Notes in Biomathematics, vol. 6. Springer-Verlag, Berlin.

Bergstrom, A., 1990. Continuous Time Econometric Modeling. Oxford University Press, Oxford.

Black, F., Scholes, M., 1973. The pricing of options and corporate liabilities. The Journal of Political Economy 81 (3), 637–654.

Caiado, J., Crato, N., Peña, D., 2006. A periodogram-based metric for time series classification. Computational Statistics & Data Analysis 50 (10), 2668–2684.

Chen, X., Hansen, L., Scheinkman, J., 1997. Shape preserving spectral approximation of diffusions. Working Paper.

Cobb, L., 1981. Stochastic differential equations for the social sciences. In: Cobb, L., Thrall, M. (Eds.), Mathematical Frontiers of the Social and Policy Sciences. Westview Press, pp. 1–26.

Corduas, M., 2007. Dissimilarity criteria for time series data mining. Quaderni di Statistica 9, 107–129.

Corduas, M., Piccolo, D., 2008. Time series clustering and classification by the autoregressive metric. Computational Statistics & Data Analysis 52, 1860–1872.

Ditlevsen, P., Ditlevsen, S., Andersen, K., 2002. The fast climate fluctuations during the stadial and interstadial climate states. Annals of Glaciology 35, 457–462.

Giorgino, T., 2009. Computing and visualizing dynamic time warping alignments in r: The DTW package. Journal of Statistical Software 31, 1–24.

Gobet, E., Hoffmann, M., Reiß, M., 2004. Nonparametric estimation of scalar diffusions based on low frequency data. The Annals of Statistics 32, 2223–2253.

Gravilov, M., Anguelov, D., Indyk, P., Motwani, R., 2000. Mining the stock market; which measure is best?. In: Proceedings of the 6th International Conference on Knowledge Discovery and Data Mining, pp. 487–496.

Hansen, L., Scheinkman, J., Touzi, N., 1998. Spectral methods for identifying scalar diffusions. Journal of Econometrics 86, 1–32.

Hirukawa, J., 2006. Cluster analysis for non-Gaussian locally stationary processes. Int. Journal of Theoretical and Applied Finance 9, 113–132.

Holden, A., 1976. Models for Stochastic Activity of Neurones. Springer-Verlag, New York.

Holland, C., 1976. On a formula in diffusion processes in population genetics. Proceedings of the American Mathematical Society 54, 316–318.

Holmes, E., 2004. Beyond theory to application and evaluation: Diffusion approximations for population viability analysis. Ecological Applications 14, 1272–1293.

Iacus, S., 2008. Simulation and Inference for Stochastic Differential Equations. With R Examples. Springer, New York.

Kakizawa, Y., Sumway, R.H., Taniguchi, M., 1998. Discrimination and clustering for multivariate time series. Journal of the American Statistical Association 93, 328–340.

Karatzas, I., Shrevre, S., 1988. Brownian Motion and Stochastic Calculus. Springer-Verlag, New York.

Kessler, M., Sørensen, M., 1999. Estimating equations based on eigenfunctions for a discretely observed diffusion process. Bernoulli 5, 299–314.

Kloden, P., Platen, E., Schurz, H., 2000. Numerical Solution of SDE through Computer Experiments. Springer, Berlin.

Kushner, H., 1967. Stochastic Stability and Control. Academic Press, New York.

Liao, T., 2005. Clustering of time series data — A survey. Pattern Recognition 38, 1857–1874.

Maharaj, E.A., 1999. Comparison and classification of stationary multivariate time series. Pattern Recognition 32, 1129–1138.

Merton, R., 1973. Theory of rational option pricing. Bell Journal of Economics and Management Science 4, 141–183.

Möller-Levet, C., Klawonn, F., Cho, K.-H., Wolkenhauer, O., 1978. Dynamic programming algorithm optimization for spoken work recognition. IEEE Transactions on Acoustic, Speech and Signal Processing 26, 143–165.

Otranto, E., 2008. Clustering heteroskedastic time series by model-based procedures. Computational Statistics & Data Analysis 52, 4685–4698.

Papanicolaou, G., 1995. Diffusions in random media. In: Keller, J.B., McLaughin, D., Papanicolaou, G. (Eds.), Surveys in Applied Mathematics. pp. 205–255.

Piccolo, D., 1990. A distance measure for classifying ARIMA models. Journal of Time Series Analysis 11, 153–164.

R Development Core Team,, R: A Language and Environment for Statistical Computing. Vienna, Austria, 2009. URL http://www.R-project.org. ISBN 3-900051-07-0.

Ramsay, J.O., Silverman, B.W., 2005. Functional Data Analysis. Springer, New York.

Reiß, M., 2003. Simulation results for estimating the diffusion coefficient from discrete time observation. Available at http://www.mathematik.hu-berlin.se/~reiss/sim-diff-est.pdf.

Ricciardi, L., 1977. Diffusion Processes and Related Topics in Biology. In: Lecture Notes in Biomathematics, Springer, New York.

Sakoe, H., Chiba, S., 1978. Dynamic programming algorithm optimization for spoken work recognition. IEEE Transactions on Acoustic, Speech and Signal Processing 26, 143–165.

Schuecker, P., Böhringer, H., Arzner, K., Reiprich, T., 2001. Cosmic mass functions from Gaussian stochastic diffusion processes. Astronomy and Astrophysics 370, 715–728.

Tuerlink, F., Maris, E., Ratcliff, R., De Boeck, P., 2001. A comparison of four methods for simulating the diffusion process. Behavior Research Methods, Instruments, Computers 33, 443–456.

Wang, K., Gasser, T., 1997. Alignment of curves by dynamic time warping. Annals of Statistics 25, 1251–1276.

Xiong, Y., Yeung, D., 2002. Mixtures of ARMA models for model-based time series clustering. In: Proceedings of the IEEE International Conference on Data Mining, pp. 717–720.