Identification and Clustering of Discretely Observed Diffusion Processes

S.M. Iacus
University of Milan
(joint work with A. De Gregorio)

Wien University, 6 March 2009

Plan of the talk

- Diffusion processes/SDEs: why, when, where.
- Part I: identification of SDE's via pseudo ϕ-divergences
- Part II: clustering ot SDE's via Markov Operator Distance

(Very) Roughly speaking, given a smooth, non stochastic dynamical system $X_t = X(t)$, its evolution with respect to time can be represented as

$$\frac{\mathrm{d}X_t}{\mathrm{d}t} = b(X_t) \quad \text{or in differential form} \quad \mathrm{d}X_t = b(X_t)\mathrm{d}t$$

A stochastic differential equation models the noise (or the stochastic part) of this system by adding the variation of some stochastic process to the above dynamics, e.g. the Wiener process

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

i.e.

deterministic trend + stochastic noise

In this talk the statistical model is the parametric family of diffusion process solutions of the SDE

$$dX_t = b(\alpha, X_t)dt + \sigma(\beta, X_t)dW_t, \quad X_0 = x_0,$$

$\theta = (\alpha, \beta) \in \Theta_\alpha \times \Theta_\beta = \Theta$, where $\Theta_\alpha \subset R^p$ and $\Theta_\beta \subset R^q$.

The drift and diffusion coefficients are known up to α and β and such that the solution of the SDE exists and the process is also ergodic.

We will consider hypotheses testing via ϕ-divergences and clustering based on discrete time observations from X.

Every time we pass from a static analysis to a dynamic one, modeling via SDEs appears to be effective.

SDEs: when?

Every time we pass from a static analysis to a dynamic one, modeling via SDEs appears to be effective.

Although we can think of using standard time series approach (AR, ARIMA, etc) stochastic differential equations assume that the stochastic process evolves continuously.

Every time we pass from a static analysis to a dynamic one, modeling via SDEs appears to be effective.

Although we can think of using standard time series approach (AR, ARIMA, etc) stochastic differential equations assume that the stochastic process evolves continuously.

In both cases observations come at discrete time. While this is not a problem for time series analysis, additional care is needed for discretely observed continuous time processes. This is a relatively new field of statistics (1993 - today)

Every time we pass from a static analysis to a dynamic one, modeling via SDEs appears to be effective.

Although we can think of using standard time series approach (AR, ARIMA, etc) stochastic differential equations assume that the stochastic process evolves continuously.

In both cases observations come at discrete time. While this is not a problem for time series analysis, additional care is needed for discretely observed continuous time processes. This is a relatively new field of statistics (1993 - today)

Next is an example from economics where discrete vs continuous time modeling matters

In time series analysis, Granger causality is synonym of “ability to predict with minimal variance”. Assume we are given a target time series $Y_t, Y_{t-1}, \ldots$ and the information $\mathcal{F}_t$ generated by two other times series $X_t, X_{t-1}, \ldots$ and Z_t, $Z_{t-1}, \ldots$.

In time series analysis, Granger causality is synonym of "ability to predict with minimal variance". Assume we are given a target time series $Y_t, Y_{t-1}, \ldots$ and the information $\mathcal{F}_t$ generated by two other times series $X_t, X_{t-1}, \ldots$ and Z_t, $Z_{t-1}, \ldots$.

The process X_t is said to "Granger cause" Y_t with respect to $\mathcal{F}_t$ if the variance of the optimal linear predictor of Y_{t+h} based on $\mathcal{F}_t$ has smaller variance than the optimal linear predictor based on $Z_t, Z_{t-1}, \ldots$.

In time series analysis, Granger causality is synonym of "ability to predict with minimal variance". Assume we are given a target time series $Y_t, Y_{t-1}, \ldots$ and the information $\mathcal{F}_t$ generated by two other times series $X_t, X_{t-1}, \ldots$ and Z_t, $Z_{t-1}, \ldots$.

The process X_t is said to "Granger cause" Y_t with respect to $\mathcal{F}_t$ if the variance of the optimal linear predictor of Y_{t+h} based on $\mathcal{F}_t$ has smaller variance than the optimal linear predictor based on $Z_t, Z_{t-1}, \ldots$.

In other words, X_t is Granger causal for Y_t if X_t helps predict Y_t at some stage in the future.

Usually a VAR model is used to test Granger causality

$$\begin{bmatrix} Y_t \\ Z_t \\ X_t \end{bmatrix} = \begin{bmatrix} \mu_1 \\ \mu_2 \\ \mu_3 \end{bmatrix} + \sum_{i=1}^{k} \left\{ \begin{bmatrix} A^i_{11} & A^i_{12} & A^i_{13} \\ A^i_{21} & A^i_{22} & A^i_{23} \\ A^i_{31} & A^i_{32} & A^i_{33} \end{bmatrix} \begin{bmatrix} Y_{t-i} \\ Z_{t-i} \\ X_{t-i} \end{bmatrix} \right\} + \begin{bmatrix} \epsilon_{1t} \\ \epsilon_{2t} \\ \epsilon_{3t} \end{bmatrix}$$

Usually a VAR model is used to test Granger causality

$$\begin{bmatrix} Y_t \\ Z_t \\ X_t \end{bmatrix} = \begin{bmatrix} \mu_1 \\ \mu_2 \\ \mu_3 \end{bmatrix} + \sum_{i=1}^{k} \left\{ \begin{bmatrix} A_{11}^i & A_{12}^i & A_{13}^i \\ A_{21}^i & A_{22}^i & A_{23}^i \\ A_{31}^i & A_{32}^i & A_{33}^i \end{bmatrix} \begin{bmatrix} Y_{t-i} \\ Z_{t-i} \\ X_{t-i} \end{bmatrix} \right\} + \begin{bmatrix} \epsilon_{1t} \\ \epsilon_{2t} \\ \epsilon_{3t} \end{bmatrix}$$

X_t does not Granger cause Y_t with respect to the information generated by Z_t if

$$A_{13}^i = A_{23}^i = 0 \quad \text{and/or} \quad A_{13}^i = A_{12}^i = 0$$

i.e. A^i is lower triangular

Unfortunately, it is not unsual to obtain that “X_t Granger cause Y_t” *and* “Y_t Granger cause X_t” even if it makes sense to expect causality in one direction only.

Unfortunately, it is not unsual to obtain that “X_t Granger cause Y_t” *and* “Y_t Granger cause X_t” even if it makes sense to expect causality in one direction only.

According to McCrorie & Chambers (2006) and others, this *“spurious Granger causality is only a consequence of the intervals in which economic data are generated being finer than the econometrician’s sampling interval.”*

Unfortunately, it is not unsual to obtain that "X_t Granger cause Y_t" *and* "Y_t Granger cause X_t" even if it makes sense to expect causality in one direction only.

According to McCrorie & Chambers (2006) and others, this *"spurious Granger causality is only a consequence of the intervals in which economic data are generated being finer than the econometrician's sampling interval."*

In essence, the underlying process (and causality) evolve continuously rather than discretely.

McCrorie and Chambers (2006) proposed the continuous version of the VAR model

$$\mathrm{d}X(t) = A(\theta)X(t)\mathrm{d}t + \mathrm{d}W_t$$

where $X(t)$ is a n-dimensional diffusion process (i.e. the components of $X(t)$ may be Y_t, X_t, and Z_t in the previous notation), $A(\theta)$ is a $n \times n$ matrix and θ parameters on which tests will be performed and $W(t)$ is a multidimensional Brownian motion.

Discrete sampling of the above continuous time model leads to observations $X_i = X(t_i), t_i = i\Delta, i = 0, \dots, n, n\Delta = T$ which solve exactly the so called Euler scheme

$$X_i = F(\theta)X_{i-1} + \Delta\epsilon_i, \quad i = 1, \dots, n$$

where ϵ is a white noise and

$$F(\theta) = e^{A(\theta)}, \quad \mathrm{Var}(\epsilon) = \Omega(\theta) = \int_0^1 e^{rA(\theta)} \sum e^{rA(\theta)'} \mathrm{d}r$$

Discrete sampling of the above continuous time model leads to observations $X_i = X(t_i), t_i = i\Delta, i = 0, \dots, n, n\Delta = T$ which solve exactly the so called Euler scheme

$$X_i = F(\theta)X_{i-1} + \Delta\epsilon_i, \quad i = 1, \dots, n$$

where ϵ is a white noise and

$$F(\theta) = e^{A(\theta)}, \quad \mathrm{Var}(\epsilon) = \Omega(\theta) = \int_0^1 e^{rA(\theta)} \sum e^{rA(\theta)'} \mathrm{d}r$$

Assuming the above continuous time model instead of the classical time series approach can greatly improve testing of Granger causality (see McCrorie and Chambers, 2006)

In pharmacokinetic/pharmacodynamic (PK/PD) models we have

$$y_{ij}, \quad i = 1, \ldots, N, \quad j = 1, \ldots, n_i$$

repeated measures on the i-th individual at time point t_{ij}

N = number of individuals

n_i = number of measurements for individual i

The response is modeled as a NLME model

$$y_{ij} = f(x_i(t_{ij}), d_i(t_{ij}), \phi_i) + \epsilon_{ij}$$

$$y_{ij} = f(x_i(t_{ij}), d_i(t_{ij}), \phi_i) + \epsilon_{ij}$$

$x_i(\cdot)$: i-th individual state variables (e.g. the amount of drug in the PK experiment)

Non Linear Mixed Effects model

$$y_{ij} = f(x_i(t_{ij}), d_i(t_{ij}), \phi_i) + \epsilon_{ij}$$

$x_i(\cdot)$: i-th individual state variables (e.g. the amount of drug in the PK experiment)

$d_i(\cdot)$: a vector of inputs (e.g. dose administration)

ϕ_i : vector of individual parameters

ϵ_{ij} : model residuals

t_{ij} : time of the measurement

$$y_{ij} = f(x_i(t_{ij}), d_i(t_{ij}), \phi_i) + \epsilon_{ij}$$

$x_i(\cdot)$: i-th individual state variables (e.g. the amount of drug in the PK experiment)

$d_i(\cdot)$: a vector of inputs (e.g. dose administration)

ϕ_i : vector of individual parameters

ϵ_{ij} : model residuals

t_{ij} : time of the measurement

There is a second stage model for ϕ, not relevant to this discussion.

$$y_{ij} = f(x_i(t_{ij}), d_i(t_{ij}), \phi_i) + \epsilon_{ij}$$

The PK dynamics is usually assumed to be regulated by the ordinary differential equation (ODE)

$$\frac{\mathrm{d}x_i(t)}{\mathrm{d}t} = g(x_i(t), d_i(t), \phi_i)$$

$$y_{ij} = f(x_i(t_{ij}), d_i(t_{ij}), \phi_i) + \epsilon_{ij}$$

The PK dynamics is usually assumed to be regulated by the ordinary differential equation (ODE)

$$\frac{\mathrm{d}x_i(t)}{\mathrm{d}t} = g(x_i(t), d_i(t), \phi_i)$$

but applications (Overgaard *et. al*, 2005) shows that this model is inadequate and to capture deviations from the ODE model, the SDEs approach has been proposed (expressed in differential form)

$$\boxed{\mathrm{d}x_i(t) = g(x_i(t), d_i(t), \phi_i)\mathrm{d}t + \sigma_w \mathrm{d}W_t}$$

σ_w may be a function of x, d, t and some other parameters

Divergences are measure of discrepancy between statistical models and ϕ-divergences are defined as follows

$$D_\phi(\theta, \theta_0) = \int_{\mathcal{X}} p(\theta_0, x)\phi\left(\frac{p(\theta, x)}{p(\theta_0, x)}\right)\mu(\mathrm{d}x) = \mathbf{E}_{\theta_0}\phi\left(\frac{p(X, \theta)}{p(X, \theta_0)}\right)$$

and the minimal requirement on the function ϕ is: $\phi(1) = 0$

The ϕ-divergences where introduced by Csiszár (1963) and studied extensively later in Liese and Vajda (1987)

They include most of known other divergences. We discuss some examples in the following

Examples: α-divergences

The α-divergences (Csiszár, 1967, Amari, 1985) are defined as

$$D_\alpha(\theta, \theta_0) = D_{\phi_\alpha}(\theta, \theta_0)$$

with

$$\phi_\alpha(x) = \frac{4\left(1 - x^{\frac{1+\alpha}{2}}\right)}{1-\alpha^2}, \quad -1 < \alpha < 1$$

They are such that $D_\alpha(\theta_0, \theta) = D_{-\alpha}(\theta, \theta_0)$.

The α-divergences include some special cases.

For example, for $\alpha \to -1$, D_{-1} is the well-known Kullback-Leibler divergence

$$D_{-1}(\theta, \theta_0) = D_{KL}(\theta, \theta_0) = -E_{\theta_0}\left\{\log\left(\frac{p(X,\theta)}{p(X,\theta_0)}\right)\right\}$$

For $\alpha \to 0$, the Hellinger distance (see, e.g., Beran, 1977, Simpson, 1989) can be derived

$$d_H(\theta, \theta_0) = \frac{1}{2}E\left(\sqrt{p(X,\theta)} - \sqrt{p(X,\theta_0)}\right)^2$$

Examples: Rényi divergences

The α-divergence is also equivalent to the Rényi's divergence (Rényi, 1961)

$$R_\alpha(\theta, \theta_0) = \frac{1}{1-\alpha} \log E_{\theta_0} \left(\frac{p(X, \theta)}{p(X, \theta_0)} \right)^\alpha$$

$$D_{KL}(\theta, \theta_0) = \lim_{\alpha \to 1} R_\alpha(\theta, \theta_0)$$

again

$$d_H(\theta, \theta_0) = 1 - \exp\left\{ \frac{1}{2} R_{\frac{1}{2}}(\theta, \theta_0) \right\}$$

Liese and Vajda (1987) generalized Rényi divergences to all real orders $\alpha \neq 0, 1$

$$D_\alpha(\theta, \theta_0) = \frac{1}{\alpha(\alpha - 1)} \log E_{\theta_0} \left(\frac{p(X, \theta)}{p(X, \theta_0)} \right)^\alpha$$

only for $\alpha = \frac{1}{2}$ the divergence is symmetric

$$D_{\frac{1}{2}}(\theta_0, \theta) = D_{\frac{1}{2}}(\theta, \theta_0) = 4 \log \int \sqrt{p(x, \theta) p(\theta_0, x)} \mu(\mathrm{d}x)$$

[known as Bhattacharyya (1946) divergence], otherwise

$$D_\alpha(\theta_0, \theta) = D_{1-\alpha}(\theta, \theta_0)$$

The transformation

$$\psi(R_\alpha) = (\exp\{(\alpha - 1)R_\alpha - 1\}/(1 - \alpha)$$

coincides with the power-divergence introduced by Cressie and Read (1984)

Power divergences D_{ϕ_λ} can be obtained directly from the ϕ-divergences choosing

$$\phi_\lambda(x) = \frac{x^{\lambda+1} - \lambda(x-1) - x}{\lambda(\lambda+1)}, \quad \lambda \in \mathbb{R} - \{0, -1\}$$

...and so forth

Summary of ϕ-divergences (see Pardo, 2006)

$\phi(x)$ with $x = p(\theta, \cdot)/p(\theta_0, \cdot)$	*Divergence*
$x \log x - x + 1$	Kullback-Leibler
$-\log x + -1$	Minimum Discrimination Information
$(x-1)\log x$	J-divergence
$\frac{1}{2}(x-1)^2$	Pearson, Kagan
$\frac{(x-1)^2}{(x+1)^2}$	Balakrishnan & Sanghvi
$\frac{-x^s + s(x-1)+1}{1-s}, \quad s \neq 1$	Rathie & Kannappan
$\frac{1-x}{2} - \left(\frac{1+x^{-r}}{2}\right)^{-1/r} \quad r > 0$	Harmonic mean (Mathai & Rathie)
$\frac{(1-x)^2}{2(a+(1-a)x)} \quad 0 \leq a \leq 1$	Rukhin
$\frac{ax \log x - (ax+1-a)\log(ax+1-a)}{a(1-a)} \quad a \neq 0, 1$	Lin
$\frac{x^{\lambda+1} - x - \lambda(x-1)}{\lambda(\lambda+1)} \quad \lambda \neq 0, -1$	Cressie & Read
$\lvert 1 - x^a \rvert^{1/a} \quad 0 < a < 1$	Matusita
$\lvert 1 - x \rvert^a \quad a \geq 1$	χ-divergence of order a (Vajda) and Total Variation if $a = 1$ (Saks)

Divergences can be used in both hypotheses testing and estimation (see, e.g. Pardo, 2006), here we consider hypotheses testing problems.

For a given sample of n i.i.d. observations $X_1, \dots, X_n$, under standard regularity assumptions on the model and on ϕ, the standard result is that, under $H_0 : \theta = \theta_0$

$$2nD_\phi(\hat{\theta}_n, \theta_0) \Rightarrow \chi^2_d$$

where $\hat{\theta}_n = \hat{\theta}_n(X_1, \dots, X_n)$ is $\sqrt{n}$-consistent and asymptotically gaussian estimator of θ and d is the dimension of θ

For continuous time observations from diffusion processes, Vajda (1990) considered the model

$$\mathrm{d}X(t) = -b(t)X_t\mathrm{d}t + \sigma(t)\mathrm{d}W_t$$

and derived explicit formulas for the Rényi divergence

For continuous time observations from diffusion processes, Vajda (1990) considered the model

$$\mathrm{d}X(t) = -b(t)X_t\mathrm{d}t + \sigma(t)\mathrm{d}W_t$$

and derived explicit formulas for the Rényi divergence

Küchler and Sørensen (1997) and Morales *et al.* (2004) contain several results on the generalized likelihood ratio test statistics and Rényi statistics for exponential families of diffusions

For continuous time observations from diffusion processes, Vajda (1990) considered the model

$$\mathrm{d}X(t) = -b(t)X_t\mathrm{d}t + \sigma(t)\mathrm{d}W_t$$

and derived explicit formulas for the Rényi divergence

Küchler and Sørensen (1997) and Morales *et al.* (2004) contain several results on the generalized likelihood ratio test statistics and Rényi statistics for exponential families of diffusions

Explicit derivations of the Rényi information on the invariant law of ergodic diffusion processes have been presented in De Gregorio and I. (2008)

For continuous time small diffusion processes f-unbiased information criteria have been derived in Uchida and Yoshida (2004) by means of Malliavin calculus. For mixing processes in Uchida and Yoshida (2001)

Review: divergences & inf. criteria

For continuous time small diffusion processes f-unbiased information criteria have been derived in Uchida and Yoshida (2004) by means of Malliavin calculus. For mixing processes in Uchida and Yoshida (2001)

Rivas *et al.* (2005) derived Rényi divergences for discrete time observations from the model $\mathrm{d}X_t = a\mathrm{d}t + b\mathrm{d}W_t$ where a and b are two scalars

Review: divergences & inf. criteria

For continuous time small diffusion processes f-unbiased information criteria have been derived in Uchida and Yoshida (2004) by means of Malliavin calculus. For mixing processes in Uchida and Yoshida (2001)

Rivas *et al.* (2005) derived Rényi divergences for discrete time observations from the model $\mathrm{d}X_t = a\mathrm{d}t + b\mathrm{d}W_t$ where a and b are two scalars

Akaike Information Criteria for discretely observed diffusion processes was derived by Uchida and Yoshida (2005)

Kutoyants (2004) and Dachian and Kutoyants (2008) consider the problem of testing statistical hypotheses for ergodic diffusion models in continuous time

Review: hypotheses testing

Kutoyants (2004) and Dachian and Kutoyants (2008) consider the problem of testing statistical hypotheses for ergodic diffusion models in continuous time

Kutoyants (1984) and Iacus and Kutoyants (2001) consider parametric and semiparametric hypotheses testing for small diffusion processes

Review: hypotheses testing

Kutoyants (2004) and Dachian and Kutoyants (2008) consider the problem of testing statistical hypotheses for ergodic diffusion models in continuous time

Kutoyants (1984) and Iacus and Kutoyants (2001) consider parametric and semiparametric hypotheses testing for small diffusion processes

Negri and Nishiyama (2008) propose a non parametric test based on score marked empirical process for continuous time observations of ergodic diffusions and Masuda *et al.* (2008) analyzed the discrete time case. Lee and Wee (2008) considered the parametric version of this test for a simplified ergodic model. Negri and Nishiyama (2007) studied the same test for continuous and discrete time observations from small diffusion processes.

Aït-Sahalia (1996), Giet and Lubrano (2008) and Chen *et al.* (2008) proposed tests based on the several distances between parametric and nonparametric estimation of the invariant density of discretely observed ergodic diffusion processes

Aït-Sahalia (1996), Giet and Lubrano (2008) and Chen *et al.* (2008) proposed tests based on the several distances between parametric and nonparametric estimation of the invariant density of discretely observed ergodic diffusion processes

(Up to our knowledge) No other option exists for parametric hypotheses testing based on divergences for discretely observed diffusions processes, and this was the motivation for this work

Consider again the ϕ-divergence

$$D_\phi(\theta, \theta_0) = E_{\theta_0} \phi \left(\frac{p(X, \theta)}{p(X, \theta_0)} \right)$$

where $p(X, \theta)$ is the likelihood of the process X under θ.

Let $\phi(\cdot)$ be such that $\phi(1) = 0$. When they exist, define $C_\phi = \phi'(1)$ and $K_\phi = \phi''(1)$.

Consider again the ϕ-divergence

$$D_\phi(\theta, \theta_0) = E_{\theta_0} \phi\left(\frac{p(X,\theta)}{p(X,\theta_0)}\right)$$

where $p(X, \theta)$ is the likelihood of the process X under θ.

Let $\phi(\cdot)$ be such that $\phi(1) = 0$. When they exist, define $C_\phi = \phi'(1)$ and $K_\phi = \phi''(1)$.

In order to get additional properties, in the i.i.d. case $\phi(x)$ is assumed to be convex or decreasing in $x \in (0, 1)$ and increasing for $x > 1$. These conditions are very convenient in the presence of exponential families. We do not ask for these conditions in our framework.

The test statistic

We assume that the process X_t is ergodic for every θ with invariant law μ_θ. The process X_t is observed at discrete times $t_i = i\Delta_n, i = 0, 1, 2, ..., n$, where Δ_n is the length of the steps. We denote the observations by $\mathbf{X}_n := \{X_i = X_{t_i}\}_{0\leq i\leq n}$.

The asymptotic is $\Delta_n \to 0, n\Delta_n \to \infty$ and $n\Delta_n^2 \to 0$ as $n \to \infty$.

Given $\tilde{\theta}_n$ a consistent and asymptotically gaussian estimator for θ, we propose the following test statistics based on the pseudo ϕ-divergence

$$\mathbb{D}_\phi(\tilde{\theta}_n, \theta_0) = \phi\left(\frac{f_n(\mathbf{X}_n, \tilde{\theta}_n)}{f_n(\mathbf{X}_n, \theta_0)}\right)$$

to test $H_0 : \theta = \theta_0$ against $H_1 : \theta \neq \theta_0$.

Here $f_n(\cdot, \theta)$ is an the approximate likelihood of the observed diffusion.

Notice: there is no expect value! Hence, “pseudo” ϕ-divergences

The test statistic

We need an estimator $\tilde{\theta}_n$ such such that: $\Gamma^{-1/2}(\tilde{\theta}_n - \theta_0) \xrightarrow{d} N(0, \mathcal{I}(\theta_0)^{-1})$

where $\mathcal{I}(\theta_0)$ the Fisher information matrix, positive definite and invertible at θ_0

$$\mathcal{I}(\theta_0) = \begin{pmatrix} (\mathcal{I}_b^{kj}(\theta_0))_{k,j=1,\dots,p} & 0 \\ 0 & (\mathcal{I}_\sigma^{kj}(\theta_0))_{k,j=1,\dots,q} \end{pmatrix}$$

with

$$\mathcal{I}_b^{kj}(\theta_0) = \int \frac{1}{\sigma^2(\beta_0, x)} \frac{\partial b(\alpha_0, x)}{\partial \alpha_k} \frac{\partial b(\alpha_0, x)}{\partial \alpha_j} \mu_{\theta_0}(dx)$$

$$\mathcal{I}_\sigma^{kj}(\theta_0) = 2 \int \frac{1}{\sigma^2(\beta_0, x)} \frac{\partial \sigma(\beta_0, x)}{\partial \beta_k} \frac{\partial \sigma(\beta_0, x)}{\partial \beta_j} \mu_{\theta_0}(dx)$$

and Γ the $(p+q) \times (p+q)$ matrix

$$\Gamma = \begin{pmatrix} \frac{1}{n\Delta_n} I_p & 0 \\ 0 & \frac{1}{n} I_q \end{pmatrix}$$

with I_p is the $p \times p$ identity matrix.

Approximation of the density statistics

As in Uchida and Yoshida (2005), consider the following approximation of the likelihood

$$f_n(\theta) = \exp\left\{u_n(\theta)\right\}\,, \quad u_n(\theta) = \sum_{k=1}^{n} u(\Delta_n, X_{i-1}, X_i, \theta)$$

where

$$u(t,x,y,\theta) = -\frac{1}{2}\log(2\pi t) - \log\sigma(y,\beta) - \frac{S^2(x,y,\beta)}{2t} + H(x,y,\theta) + t\tilde{g}(x,y,\theta)\,,$$

with

$$S(x,y,\beta) = \int_x^y \frac{\mathrm{d}u}{\sigma(u,\beta)}\,, \quad H(x,y,\theta) = \int_x^y \frac{B(u,\theta)}{\sigma(u,\beta)}\mathrm{d}u$$

$$\tilde{g}(x,y,\theta) = -\frac{1}{2}\left\{C(x,\theta) + C(y,\theta) + \frac{1}{3}B(x,\theta)B(y,\theta)\right\}$$

$$C(x,\theta) = \frac{1}{2}B^2(x,\theta) + \frac{1}{2}B_x(x,\theta)\sigma(x,\beta), \quad B(x,\theta) = \frac{b(x,\alpha)}{\sigma(x,\beta)} - \frac{1}{2}\sigma_x(x,\beta)$$

Consistent and asymptotically Gaussian estimator

We consider the approximated maximum likelihood estimator $\hat{\theta}_n$ based on the locally Gaussian approximation (see, e.g. Yoshida, 1992), i.e.

$$\hat{\theta}_n = \arg\max_{\theta} \ell_n(\theta)$$

with

$$\ell_n(\theta) = \prod_{i=1}^{n} \frac{1}{\sqrt{2\pi\Delta_n \sigma^2(X_{i-1},\theta)}} \exp\left\{ -\frac{1}{2} \frac{(X_i - X_{i-1} - b(X_{i-1},\theta)\Delta_n)^2}{\Delta_n \sigma^2(X_{i-1},\theta)} \right\}$$

The estimator $\hat{\theta}_n$ satisfies previous convergence assumptions for $\tilde{\theta}_n$

Regularity conditions on the process

i) There exists a constant C such that

$$|b(\alpha_0, x) - b(\alpha_0, y)| + |\sigma(\beta_0, x) - \sigma(\beta_0, y)| \leq C|x - y|.$$

ii) $\inf_{\beta,x} \sigma^2(\beta, x) > 0$.

iii) The process X is ergodic for every θ with invariant probability measure μ_θ. All polynomial moments of μ_θ are finite.

iv) For all $m \geq 0$ and for all θ, $\sup_t E|X_t|^m < \infty$.

v) For every θ, the coefficients $b(\alpha, x)$ and $\sigma(\beta, x)$ are five times differentiable with respect to x and the derivatives are polynomial growth in x, uniformly in θ.

vi) The coefficients $b(\alpha, x)$ and $\sigma(\beta, x)$ and all their partial derivatives respect to x up to order 2 are three times differentiable respect to θ for all x in the state space. All derivatives respect to θ are polynomial growth in x, uniformly in θ.

Conditions on the approximation and identifiability

Let $i = 0, 1, 2, 3$ and ∂_θ^i the partial i-th derivative with respect to θ and similarly for x,

i) $\partial_\theta^i \tilde{h}(x, \theta) = O(|x|^2)$ as $x \to \infty$.

ii) $\inf_x \partial_\theta^i \tilde{h}(x, \theta) > -\infty$

iii) $\sup_\theta \sup_x |\partial_\theta^i \partial_x^5 \tilde{h}(x, \theta)| \leq M < \infty$.

iv) There exists $\gamma > 0$ such that for every θ and $j = 1, \ldots, 4$, $|\partial_\theta^i \partial_x^j \tilde{B}(x, \theta)| = O(|\tilde{B}(x, \theta)|^\gamma)$ as $|x| \to \infty$.

When the coefficients $b(\alpha, x) = b(\alpha_0, x)$ and $\sigma^2(\beta, x) = \sigma^2(\beta_0, x)$ for μ_{θ_0} a.s. for all x, then $\alpha = \alpha_0$ and $\beta = \beta_0$.

Main result [$C_\phi = \phi'(1)$, $K_\phi = \phi''(1)$]

Theorem: Under $H_0 : \theta = \theta_0$ and the asymptotic $n\Delta_n^2 \to 0$, $\Delta_n \to 0$, $n\Delta_n = T \to \infty$ the pseudo-ϕ divergence test statistics is such that

$$\mathbb{D}_\phi(\hat{\theta}_n, \theta_0) \xrightarrow{d} \frac{1}{2}(C_\phi \xi_{p+q} + (C_\phi + K_\phi)\xi_{p+q}^2)$$

where $\xi_{p+q} \sim \chi_{p+q}^2$

Main result $[C_\phi = \phi'(1), K_\phi = \phi''(1)]$

Theorem: Under $H_0 : \theta = \theta_0$ and the asymptotic $n\Delta_n^2 \to 0$, $\Delta_n \to 0$, $n\Delta_n = T \to \infty$ the pseudo-ϕ divergence test statistics is such that

$$\mathbb{D}_\phi(\hat{\theta}_n, \theta_0) \xrightarrow{d} \frac{1}{2}(C_\phi \xi_{p+q} + (C_\phi + K_\phi)\xi_{p+q}^2)$$

where $\xi_{p+q} \sim \chi^2_{p+q}$

Remind that, in the i.i.d. case we have $2nD_\phi(\hat{\theta}_n, \theta_0) \Rightarrow \chi^2_d$

Notice: the limit distribution does not depend on ϕ in the i.i.d. In our approach it does and one can try to characterize the limit. In particular, we can study the power function of the test analytically under contiguous alternatives (not shown here).

For example, consider the α-divergences

$$\phi_\alpha(x) = \frac{4\left(1 - x^{\frac{1+\alpha}{2}}\right)}{1-\alpha^2}$$

and the limit as $\alpha \to -1$, i.e. the Kullback-Leibler divergence, we have

$$\phi(x) = \lim_{\alpha \to -1} \phi_\alpha(x) = -\log(x)$$

for which $C_\phi = -1$ and $K_\phi = 1$. Hence

$$\mathbb{D}_{Kull}(\hat{\theta}_n, \theta_0) = \mathbb{D}_\phi(\hat{\theta}_n, \theta_0) \xrightarrow{d} \frac{1}{2}(C_\phi \xi_{p+q} + (C_\phi + K_\phi)\xi^2_{p+q})$$

reduces to the standard result of the i.i.d. setting.

Idea of the proof

The proof is obtained by means of the δ-method up to second order.

These lemmas are needed to prove convergence

$$\Gamma^{\frac{1}{2}} \nabla_\theta \log \ell_n(\mathbf{X}_n, \theta_0) \xrightarrow{p} N(0, \mathcal{I}(\theta_0)) \qquad \text{(Kessler, 1997)}$$

$$\Gamma^{\frac{1}{2}} \nabla_\theta \log f_n(\mathbf{X}_n, \theta_0) = \Gamma^{\frac{1}{2}} \nabla_\theta \log \ell_n(\mathbf{X}_n, \theta_0) + o_p(1) \qquad \text{(Uchida \& Yoshida, 2005)}$$

$$\Gamma^{\frac{1}{2}} \nabla_\theta^2 \log f_n(\mathbf{X}_n, \theta_0) \Gamma^{\frac{1}{2}} \xrightarrow{p} -\mathcal{I}(\theta_0) \qquad \text{(Uchida \& Yoshida, 2005)}$$

So the result hold for any approximation of the likelihood f_n and appropriate estimator provided that the above lemmas can be proved.

Almost all likelihood approximations available in the literature for SDE's satisfy the assumptions.

- α-divergences

$$\mathbb{D}_\alpha(\hat{\theta}_n, \theta_0) = \phi_\alpha\left(\frac{f_n(\mathbf{X}_n, \hat{\theta}_n)}{f_n(\mathbf{X}_n, \theta_0)}\right)$$

with $\phi_\alpha(x) = 4(1 - x^{\frac{1+\alpha}{2}})/(1-\alpha^2)$, with $C_\alpha = \frac{2}{\alpha - 1}$ and $K_\phi = 1$. With $\alpha \in \{-0.99, -0.90, -0.75, -0.50, -0.25, -0.10\}$;

- power-divergences of order λ

$$\mathbb{D}_\lambda(\hat{\theta}_n, \theta_0) = \phi_\lambda\left(\frac{f_n(\mathbf{X}_n, \hat{\theta}_n)}{f_n(\mathbf{X}_n, \theta_0)}\right)$$

with $\phi_\lambda(x) = (x^{\lambda+1} - x - \lambda(x-1))/(\lambda(\lambda+1))$, with $C_\lambda = 0$, $K_\lambda = 1$. With $\lambda \in \{-0.99, -1.20, -1.50, -1.75, -2.00, -2.50\}$;

- generalized likelihood ratio test statistic

$$\mathbb{D}_{\log}(\hat{\theta}_n, \theta_0) = -\log\left(\frac{f_n(\mathbf{X}_n, \hat{\theta}_n)}{f_n(\mathbf{X}_n, \theta_0)}\right)$$

The Vasicek (VAS) model: $\mathrm{d}X_t = \kappa(\alpha - X_t)\mathrm{d}t + \sigma \mathrm{d}W_t$, where, in finance, σ is interpreted as volatility, α is the long-run equilibrium value of the process and κ is the speed of reversion. Let

$$\theta_0 = (\kappa_0, \alpha_0, \sigma_0^2) = (0.85837, 0.089102, 0.0021854)$$

we consider three different sets of hypotheses for the parameters

model	$\theta = (\kappa, \alpha, \sigma^2)$
VAS_0	$(\kappa_0, \alpha_0, \sigma_0^2)$
VAS_1	$(4 \cdot \kappa_0, \alpha_0, 4 \cdot \sigma_0^2)$
VAS_2	$(\frac{1}{4}\kappa_0, \alpha_0, \frac{1}{4} \cdot \sigma_0^2)$

The Cox-Ingersoll-Ross (CIR) model: $\mathrm{d}X_t = \kappa(\alpha - X_t)\mathrm{d}t + \sigma\sqrt{X_t}\mathrm{d}W_t$. Let

$$\theta_0 = (\kappa_0, \alpha_0, \sigma_0^2) = (0.89218, 0.09045, 0.032742)$$

we consider different sets of hypotheses for the parameters

model	$\theta = (\kappa, \alpha, \sigma^2)$
CIR_0	$(\kappa_0, \alpha_0, \sigma_0^2)$
CIR_1	$(\frac{1}{2} \cdot \kappa_0, \alpha_0, \frac{1}{2} \cdot \sigma_0^2)$
CIR_2	$(\frac{1}{4} \cdot \kappa_0, \alpha_0, \frac{1}{4} \cdot \sigma_0^2)$

This model has a transition density of χ^2-type, hence local gaussian approximation is less likely to hold for non negligible values of Δ_n.

The parameters of the above models, have been chosen according to Pritsker (1998) and Chen *et al.* (2008), in particular VAS_0 corresponds to the model estimated by Aït-Sahalia (1996) for real interest rates data.

The empirical level of the test is calculated as the number of times the test rejects the null hypothesis under the true model, i.e.

$$\hat{\alpha}_n = \frac{1}{M} \sum_{i=1}^{M} \mathbf{1}_{\{\mathbb{D}_\phi > c_\alpha\}} \qquad \text{under} \quad H_0$$

where $\mathbf{1}_A$ is the indicator function of set A, $M = 10,000$ is the number of simulations and c_α is the $(1-\alpha)\%$ quantile of the proper distribution. Similarly we calculate the power of the test under alternative models as

$$\hat{\beta}_n = \frac{1}{M} \sum_{i=1}^{M} \mathbf{1}_{\{\mathbb{D}_\phi > c_\alpha\}} \qquad \text{under} \quad H_1$$

Vasicek. Alpha-div. of order a, $a = -0.99$ = GLRT

model (α, n)	$a = -0.99$	$a = -0.90$	$a = -0.75$	$a = -0.50$	$a = -0.25$	$a = -0.10$
VAS_0 (0.01, 50)	0.01	0.10	0.39	0.62	0.73	0.77
VAS_1 (0.01, 50)	1.00	1.00	1.00	1.00	1.00	1.00
VAS_2 (0.01, 50)	1.00	1.00	1.00	1.00	1.00	1.00
VAS_0 (0.05, 50)	0.04	0.12	0.39	0.62	0.73	0.77
VAS_1 (0.05, 50)	1.00	1.00	1.00	1.00	1.00	1.00
VAS_2 (0.05, 50)	1.00	1.00	1.00	1.00	1.00	1.00
VAS_0 (0.01, 100)	0.01	0.10	0.39	0.63	0.74	0.78
VAS_1 (0.01, 100)	1.00	1.00	1.00	1.00	1.00	1.00
VAS_2 (0.01, 100)	1.00	1.00	1.00	1.00	1.00	1.00
VAS_0 (0.05, 100)	0.04	0.11	0.40	0.63	0.74	0.78
VAS_1 (0.05, 100)	1.00	1.00	1.00	1.00	1.00	1.00
VAS_2 (0.05, 100)	1.00	1.00	1.00	1.00	1.00	1.00
VAS_0 (0.01, 500)	0.02	0.18	0.61	0.83	0.90	0.92
VAS_1 (0.01, 500)	1.00	1.00	1.00	1.00	1.00	1.00
VAS_2 (0.01, 500)	1.00	1.00	1.00	1.00	1.00	1.00
VAS_0 (0.05, 500)	0.07	0.20	0.61	0.83	0.90	0.92
VAS_1 (0.05, 500)	1.00	1.00	1.00	1.00	1.00	1.00
VAS_2 (0.05, 500)	1.00	1.00	1.00	1.00	1.00	1.00

Vasicek. Power-div. of order λ

model (α, n)	$\lambda = -0.99$	$\lambda = -1.20$	$\lambda = -1.50$	$\lambda = -1.75$	$\lambda = -2.00$	$\lambda = -2.50$
VAS_0 (0.01, 50)	0.00	0.00	0.00	0.01	0.02	0.04
VAS_1 (0.01, 50)	0.00	0.99	1.00	1.00	1.00	1.00
VAS_2 (0.01, 50)	0.40	1.00	1.00	1.00	1.00	1.00
VAS_0 (0.05, 50)	0.00	0.00	0.00	0.01	0.03	0.06
VAS_1 (0.05, 50)	0.67	1.00	1.00	1.00	1.00	1.00
VAS_2 (0.05, 50)	0.99	1.00	1.00	1.00	1.00	1.00
VAS_0 (0.01, 100)	0.00	0.00	0.00	0.01	0.02	0.04
VAS_1 (0.01, 100)	0.23	1.00	1.00	1.00	1.00	1.00
VAS_2 (0.01, 100)	0.88	1.00	1.00	1.00	1.00	1.00
VAS_0 (0.05, 100)	0.00	0.00	0.00	0.01	0.03	0.06
VAS_1 (0.05, 100)	1.00	1.00	1.00	1.00	1.00	1.00
VAS_2 (0.05, 100)	1.00	1.00	1.00	1.00	1.00	1.00
VAS_0 (0.01, 500)	0.00	0.00	0.00	0.01	0.03	0.08
VAS_1 (0.01, 500)	1.00	1.00	1.00	1.00	1.00	1.00
VAS_2 (0.01, 500)	1.00	1.00	1.00	1.00	1.00	1.00
VAS_0 (0.05, 500)	0.00	0.00	0.01	0.03	0.06	0.12
VAS_1 (0.05, 500)	1.00	1.00	1.00	1.00	1.00	1.00
VAS_2 (0.05, 500)	1.00	1.00	1.00	1.00	1.00	1.00

CIR. Alpha-div. of order a, $a = -0.99$ = GLRT

model (α, n)	$a = -0.99$	$a = -0.90$	$a = -0.75$	$a = -0.50$	$a = -0.25$	$a = -0.10$
CIR_0 (0.01, 50)	0.01	0.14	0.54	0.77	0.85	0.87
CIR_1 (0.01, 50)	0.80	0.98	1.00	1.00	1.00	1.00
CIR_2 (0.01, 50)	1.00	1.00	1.00	1.00	1.00	1.00
CIR_0 (0.05, 50)	0.05	0.16	0.54	0.77	0.85	0.87
CIR_1 (0.05, 50)	0.94	0.98	1.00	1.00	1.00	1.00
CIR_2 (0.05, 50)	1.00	1.00	1.00	1.00	1.00	1.00
CIR_0 (0.01, 100)	0.01	0.13	0.49	0.71	0.79	0.82
CIR_1 (0.01, 100)	0.99	1.00	1.00	1.00	1.00	1.00
CIR_2 (0.01, 100)	1.00	1.00	1.00	1.00	1.00	1.00
CIR_0 (0.05, 100)	0.04	0.15	0.49	0.71	0.79	0.82
CIR_1 (0.05, 100)	1.00	1.00	1.00	1.00	1.00	1.00
CIR_2 (0.05, 100)	1.00	1.00	1.00	1.00	1.00	1.00
CIR_0 (0.01, 500)	0.00	0.06	0.28	0.54	0.69	0.74
CIR_1 (0.01, 500)	1.00	1.00	1.00	1.00	1.00	1.00
CIR_2 (0.01, 500)	1.00	1.00	1.00	1.00	1.00	1.00
CIR_0 (0.05, 500)	0.02	0.08	0.28	0.54	0.69	0.74
CIR_1 (0.05, 500)	1.00	1.00	1.00	1.00	1.00	1.00
CIR_2 (0.05, 500)	1.00	1.00	1.00	1.00	1.00	1.00

model (α, n)	$\lambda = -0.99$	$\lambda = -1.20$	$\lambda = -1.50$	$\lambda = -1.75$	$\lambda = -2.00$	$\lambda = -2.50$
CIR_0 (0.01, 50)	0.00	0.00	0.00	0.01	0.02	0.06
CIR_1 (0.01, 50)	0.00	0.06	0.52	0.75	0.86	0.94
CIR_2 (0.01, 50)	0.00	0.99	1.00	1.00	1.00	1.00
CIR_0 (0.05, 50)	0.00	0.00	0.00	0.02	0.04	0.09
CIR_1 (0.05, 50)	0.00	0.23	0.70	0.85	0.92	0.96
CIR_2 (0.05, 50)	0.06	1.00	1.00	1.00	1.00	1.00
CIR_0 (0.01, 100)	0.00	0.00	0.00	0.01	0.02	0.05
CIR_1 (0.01, 100)	0.00	0.56	0.96	0.99	1.00	1.00
CIR_2 (0.01, 100)	0.00	1.00	1.00	1.00	1.00	1.00
CIR_0 (0.05, 100)	0.00	0.00	0.00	0.02	0.03	0.08
CIR_1 (0.05, 100)	0.00	0.83	0.99	1.00	1.00	1.00
CIR_2 (0.05, 100)	0.97	1.00	1.00	1.00	1.00	1.00
CIR_0 (0.01, 500)	0.00	0.00	0.00	0.00	0.01	0.02
CIR_1 (0.01, 500)	0.00	1.00	1.00	1.00	1.00	1.00
CIR_2 (0.01, 500)	1.00	1.00	1.00	1.00	1.00	1.00
CIR_0 (0.05, 500)	0.00	0.00	0.00	0.01	0.02	0.04
CIR_1 (0.05, 500)	1.00	1.00	1.00	1.00	1.00	1.00
CIR_2 (0.05, 500)	1.00	1.00	1.00	1.00	1.00	1.00

Conclusions

The power divergences are, on average, better than the generalized likelihood ratio test in terms of both empirical level $\hat{\alpha}$ and power $\hat{\beta}$ for the models considered and under the selected alternatives

the α-divergence do not behave very well and only approximate the GLR test at most (i.e. always worse than GLRT)

For the CIR study, all test statistics have, in general, lower power under the alternative CIR_1 than under CIR_2

Power divergences are yet the best test statistics in both cases (CIR and VAS), for $\lambda = (-1.20, -1.50, -1.75)$

Software

The package `sde` for the `R` statistical environment is freely available at `http://cran.R-Project.org`.

It contains the function `sdeDiv` which implements the ϕ-divergence test statistics.

We consider the model

$$dX_t = (\theta_{i1} - \theta_{i2} X_t)dt + \theta_{i3}\sqrt{X_t}dW_t, \quad i = 0, 1$$

with (as, in Pritsker, 1998, and Chen *et al.*, 2008)

$$\theta_0 = (0.0807, 0.8922, 0.1809)$$
$$\theta_1 = (0.0403, 0.8922, 0.1279)$$

```
theta0 <- c(0.0807, 0.8922, 0.1809)
theta1 <- c(0.0403, 0.8922, 0.1279)
```

We simulate under $H_1 : \theta = \theta_1$ and test for $H_0 : \theta = \theta_0$

```
set.seed(123)
X <- sde.sim(X0=rsCIR(1, theta1), N=5000, delta=1e-3,  model="CIR", theta=theta1)
```

after setting up model description

```
b <- function(x,theta) theta[1]-theta[2]*x  # drift coefficient
b.x <- function(x,theta)  -theta[2]
s <- function(x,theta) theta[3]*sqrt(x) # diffusion coefficient
s.x <- function(x,theta) theta[3]/(2*sqrt(x))
s.xx <- function(x,theta) -theta[3]/(4*x^1.5)
```

we choose the power divergences

```
lambda <- -1.75
myphi <-  expression((x^(lambda+1) -x -lambda*(x-1))/(lambda*(lambda+1)))
```

$$\phi(x) = \frac{x^{\lambda+1} - x - \lambda(x-1)}{\lambda(\lambda+1)}$$

Software: example of use

We run the test. Should reject H_0

```
sdeDiv(X=X, theta0 = theta0, phi = myphi, b=b, s=s, b.x=b.x, s.x=s.x, s.xx=s.xx,
       method="L-BFGS-B", lower=rep(1e-3,3), guess=c(1,1,1))

estimated parameters
0.04041047 1.298524 0.1290066

Testing H0 against H1

H0: 0.0807 0.8922 0.1809
H1: 0.04041047 1.298524 0.1290066

Divergence statistic: 2.8492e+151 (p-value=0)
Likelihood ratio test statistic: 930.69 (p-value=1.9486e-201)
```

Software: example of use

We run the test. Should reject H_0

```
sdeDiv(X=X, theta0 = theta0, phi = myphi, b=b, s=s, b.x=b.x, s.x=s.x, s.xx=s.xx,
       method="L-BFGS-B", lower=rep(1e-3,3), guess=c(1,1,1))

estimated parameters
0.04041047 1.298524 0.1290066

Testing H0 against H1

H0: 0.0807 0.8922 0.1809
H1: 0.04041047 1.298524 0.1290066

Divergence statistic: 2.8492e+151 (p-value=0)
Likelihood ratio test statistic: 930.69 (p-value=1.9486e-201)
```

H_0 successfully rejected! Both by power-divergence and GLRT.

Software: example of use

Now we run the test for $H_0 = H_1$, should not reject

```
sdeDiv(X=X, theta0 = theta1, phi = myphi, b=b, s=s, b.x=b.x, s.x=s.x, s.xx=s.xx,
       method="L-BFGS-B", lower=rep(1e-3,3), guess=c(1,1,1))

estimated parameters
0.04041047 1.298524 0.1290066

Testing H0 against H1

H0: 0.0403 0.8922 0.1279
H1: 0.04041047 1.298524 0.1290066

Divergence statistic: 8.7511 (p-value=0.24091)
Likelihood ratio test statistic: 6.883 (p-value=0.075723)
```

Software: example of use

Now we run the test for $H_0 = H_1$, should not reject

```
sdeDiv(X=X, theta0 = theta1, phi = myphi, b=b, s=s, b.x=b.x, s.x=s.x, s.xx=s.xx,
       method="L-BFGS-B", lower=rep(1e-3,3), guess=c(1,1,1))

estimated parameters
0.04041047 1.298524 0.1290066

Testing H0 against H1

H0: 0.0403 0.8922 0.1279
H1: 0.04041047 1.298524 0.1290066

Divergence statistic: 8.7511 (p-value=0.24091)
Likelihood ratio test statistic: 6.883 (p-value=0.075723)
```

clearly H_0 not rejected at 5% by power divergences. Not rejected also by GLRT with suspect p-value

Consider again the nonparametric family of ergodic diffusion process

$$\mathrm{d}X_t = b(X_t)\mathrm{d}t + \sigma(X_t)\mathrm{d}W_t, \quad 0 \leq t \leq T,$$

Let

$$s(x) = \exp\left\{-2\int_{x_0}^{x} \frac{b(y)}{\sigma^2(y)}\mathrm{d}y\right\} \quad \text{and} \quad m(x) = \frac{1}{\sigma^2(x)s(x)}\,.$$

be the *scale* and *speed* measures. $x_0 \in [a, b]$, $[a, b]$ the state space of X.

Consider again the nonparametric family of ergodic diffusion process

$$\mathrm{d}X_t = b(X_t)\mathrm{d}t + \sigma(X_t)\mathrm{d}W_t, \quad 0 \leq t \leq T,$$

Let

$$s(x) = \exp\left\{-2\int_{x_0}^{x} \frac{b(y)}{\sigma^2(y)}\mathrm{d}y\right\} \quad \text{and} \quad m(x) = \frac{1}{\sigma^2(x)s(x)}\,.$$

be the *scale* and *speed* measures. $x_0 \in [a, b]$, $[a, b]$ the state space of X.

Then, the invariant measure of X is

$$\mu_{b,\sigma}(x) = \frac{m(x)}{M}, \quad \text{with} \quad M = \int m(x)\mathrm{d}x$$

Cluster analysis

The discretized observations X_i form a Markov process and all the mathematical properties are embodied in the so-called *transition operator*

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

with f is a generic function, e.g. $f(x) = x^k$.

Notice that P_Δ depends on the transition density between X_i and X_{i-1}, so we put explicitly the dependence on Δ in the notation.

Luckily, there is no need to deal with the transition density, we can estimate P_Δ directly and fully non parametrically.

We assume $n\Delta_n^2 \to 0$, $\Delta_n \to 0$, $n\Delta_n = T \to \infty$.

For a given L^2-orthonormal basis $\{\phi_j, j \in J\}$ of $L^2([a,b])$, where J is an index set, following Gobet *et. al* (2004) it is possible to obtain an estimator $\hat{\mathbf{P}}_\Delta$ of $< P_\Delta \phi_j, \phi_k >_{\mu_{b,\sigma}}$ with entries

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

The terms $(\hat{P}_\Delta)_{j,k}$ are approximations of $< P_\Delta \phi_j, \phi_k >_{\mu_{b,\sigma}}$, that is, *the action of the transition operator on the state space* of X with respect of the unknown scalar product $< \cdot, \cdot >_{\mu_{b,\sigma}}$.

Remind that $\mu_{b,\sigma}$ is the unknown invariat distribution of the process depending on the unknown drift $b(\cdot)$ and diffusion $\sigma(\cdot)$ coefficients but we don't need to specify them.

Markov operator distance

Then, $\hat{\mathbf{P}}_\Delta$ can be used as "proxy" of the probability structure of the model.

Our proposal is to use the distance between two estimated Markov Operators

$$d_{MO}(X,Y) = \sum_{j,k \in J} [(\hat{P}_\Delta)_{j,k}(X) - (\hat{P}_\Delta)_{j,k}(Y)]^2$$

In our examples, we compare d_{MO} against

- the Euclidean distance d_{EUC}
- the Short-Time-Series distance d_{STS}
- and the Dynamic Time Warping distance d_{DTW}

We simulate 10 paths X_i, $i = 1, \ldots, 10$, according to the combinations of drift b_i and diffusion coefficients σ_i, $i = 1, \ldots, 4$ presented in the following table

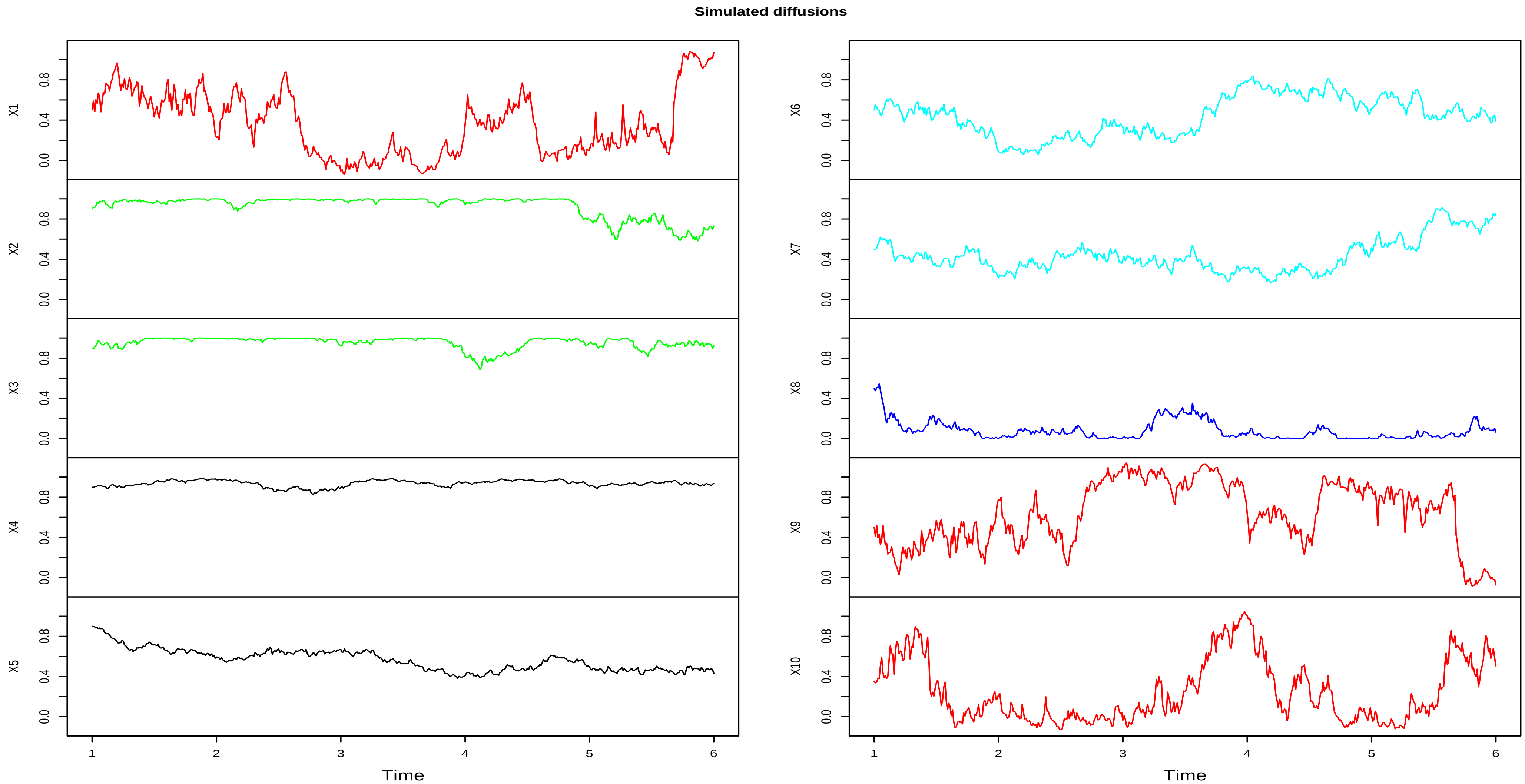
	$\sigma_1(x)$	$\sigma_2(x)$	$\sigma_3(x)$	$\sigma_4(x)$
$b_1(x)$	X10, X1		X5	
$b_2(x)$		X2,X3	X4	
$b_3(x)$		X6, X7		
$b_4(x)$				X8

where

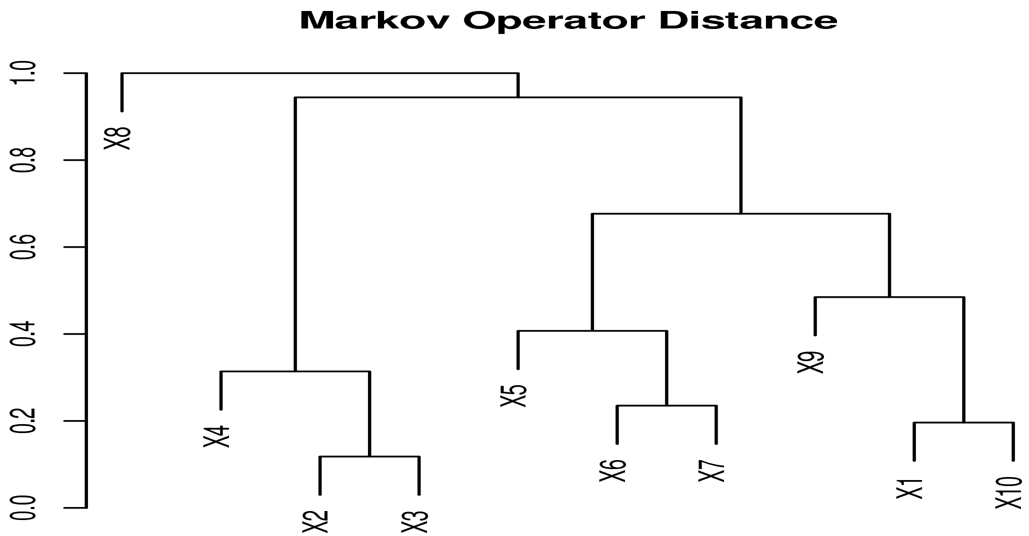
$$b_1(x) = 1 - 2x, \quad b_2(x) = 1.5(0.9 - x), \quad b_3(x) = 1.5(0.5 - x), \quad b_4(x) = 5(0.05 - x)$$

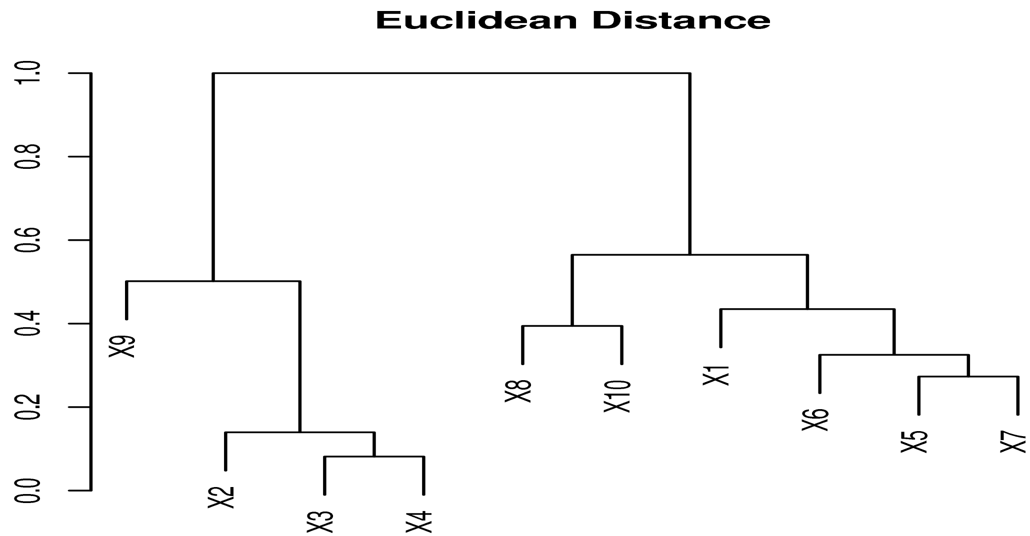
$$\sigma_1(x) = 0.5 + 2x(1 - x), \quad \sigma_2(x) = \sqrt{0.55x(1 - x)}$$
$$\sigma_3(x) = \sqrt{0.1x(1 - x)}, \quad \sigma_4(x) = \sqrt{0.8x(1 - x)}$$

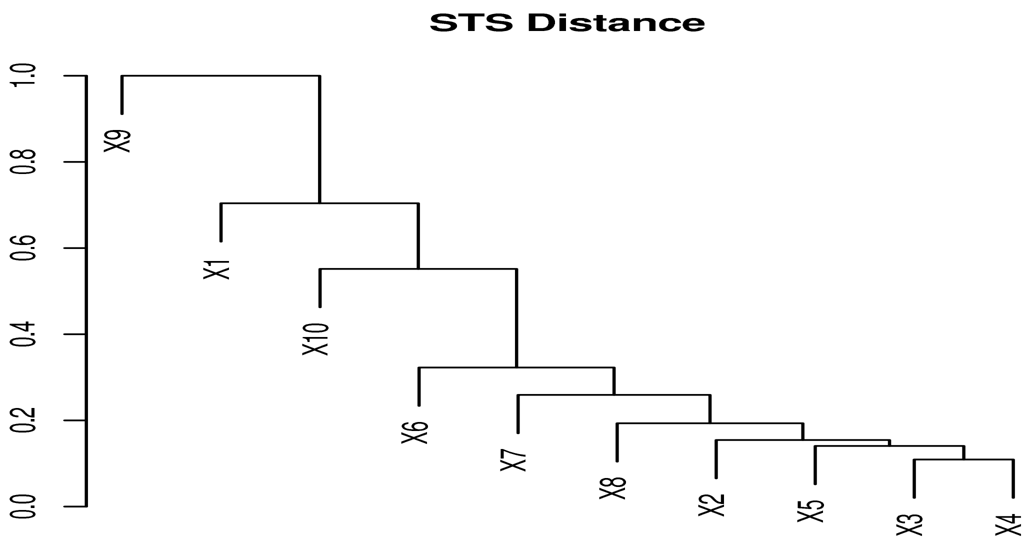
The process X9=1-X1, hence it has drift $-b_1(x)$ and the same quadratic variation of X1 and X10.

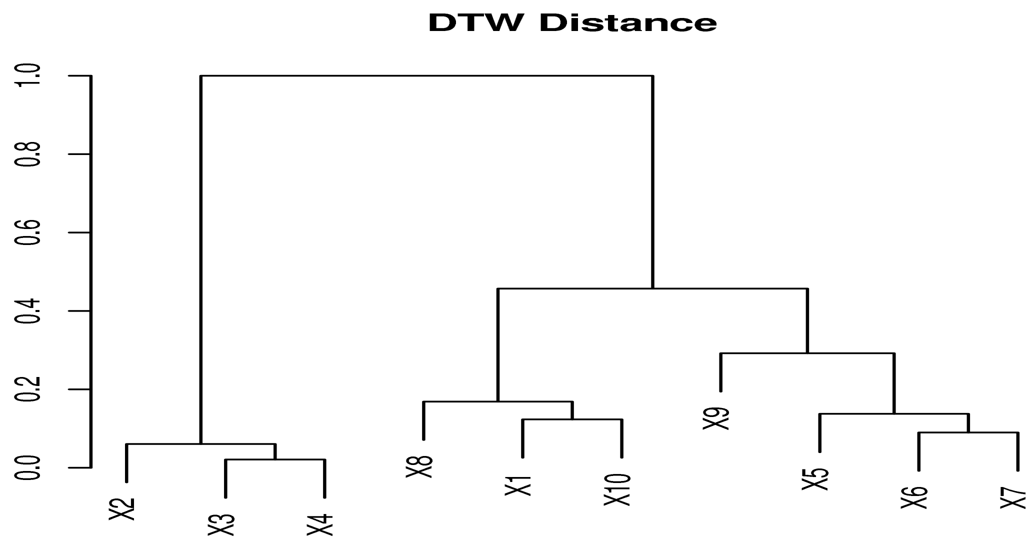

Simulated diffusions
X1
X2
X3
X4
X5
X6
X7
X8
X9
X10
0.0
0.4
0.8
1
2
3
4
5
6
Time

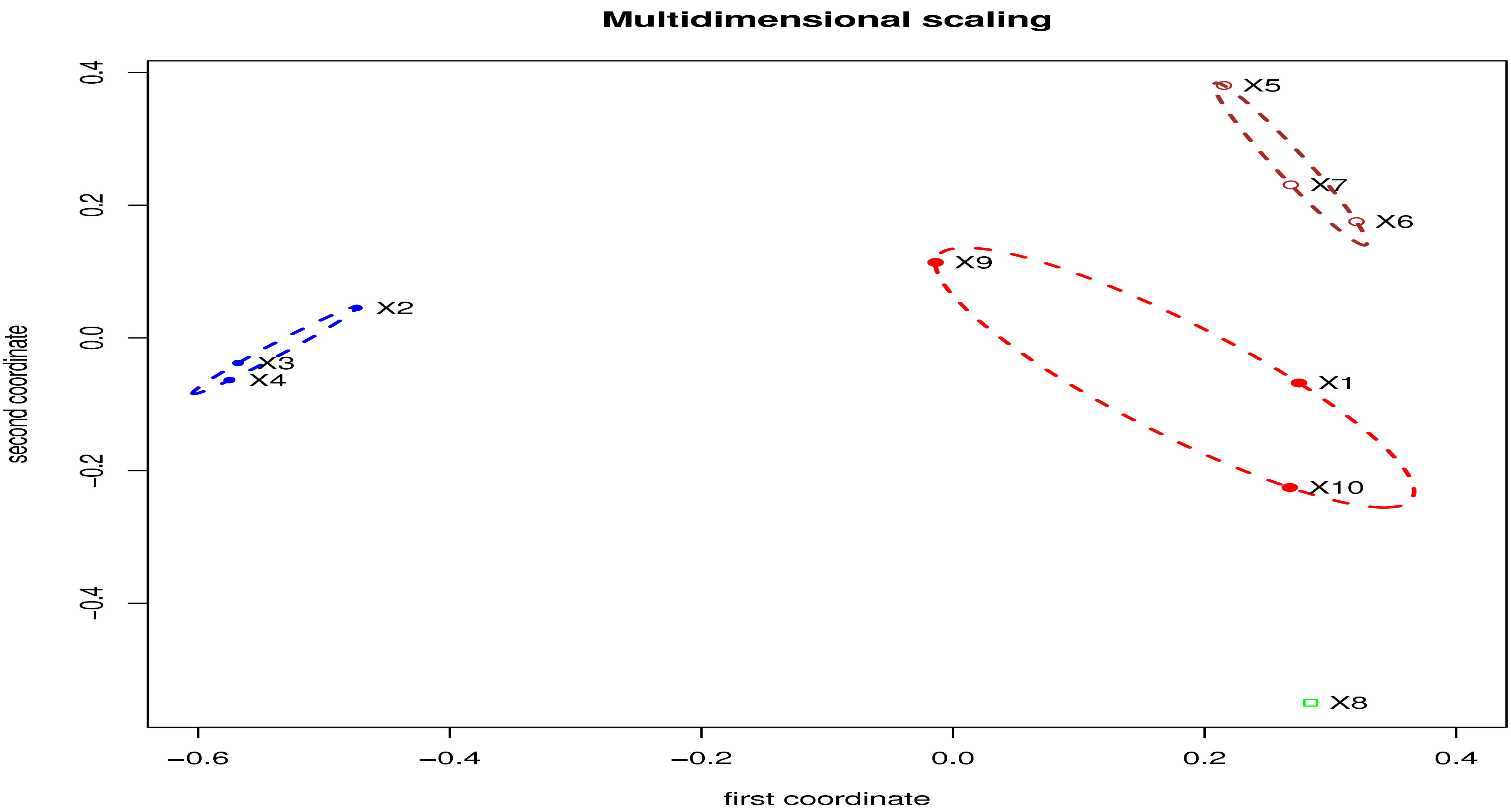
Dendrograms










Multidimensional scaling
X5
X7
X6
X9
X2
X3
X4
X1
X10
X8
0.4
0.2
0.0
−0.2
−0.4
second coordinate
−0.6
−0.4
−0.2
0.0
0.2
0.4
first coordinate

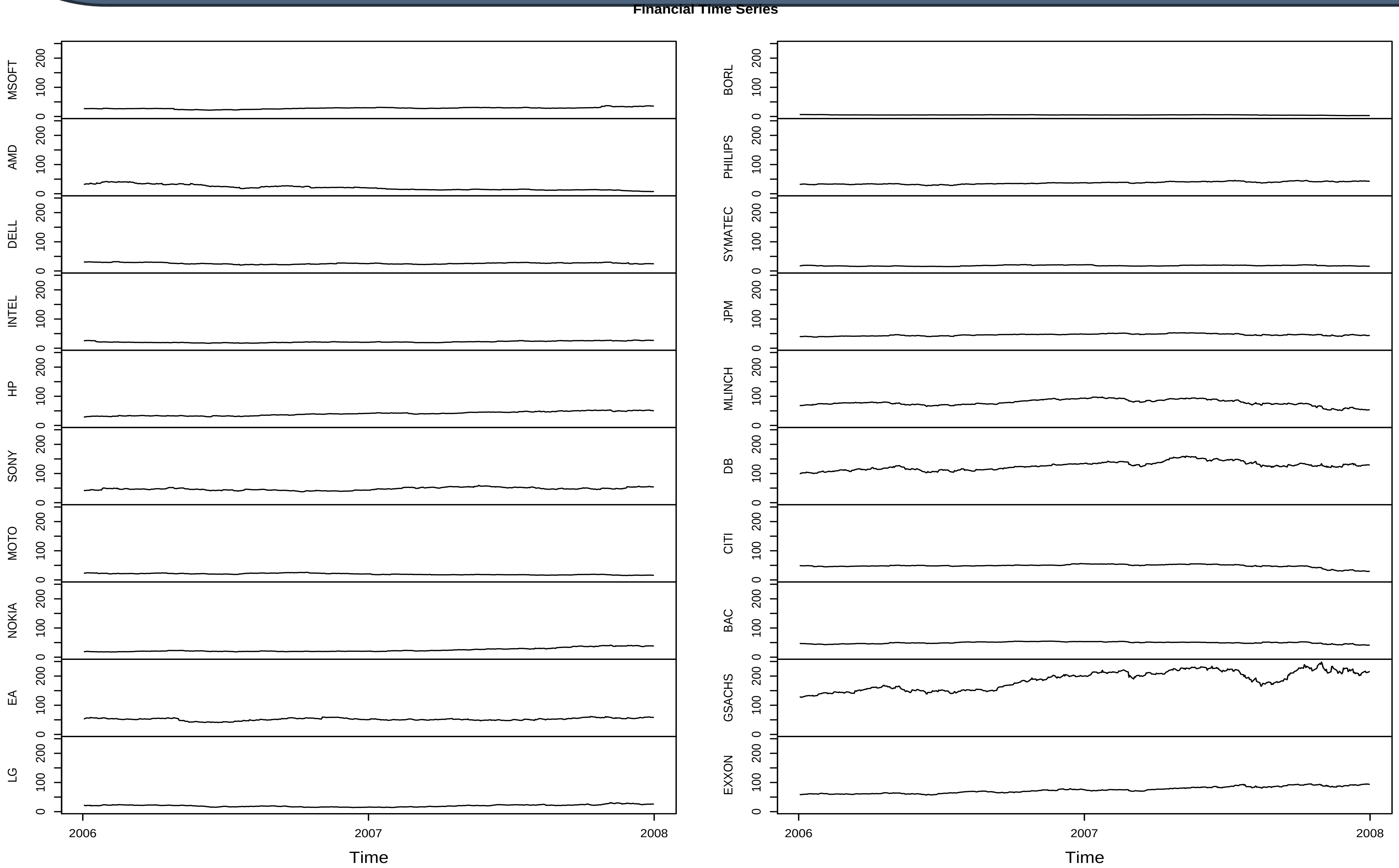
Real Data from NYSE

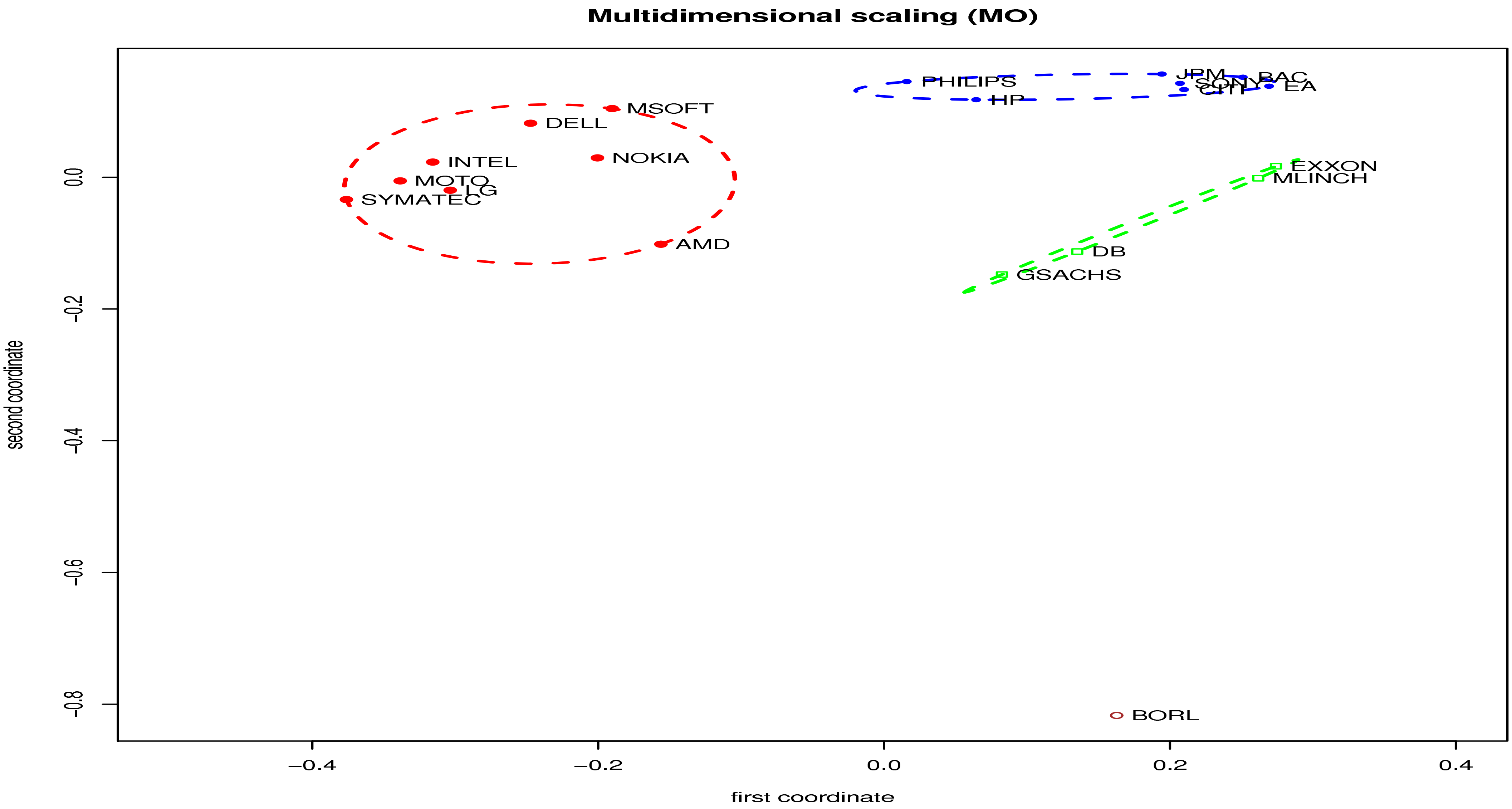
We consider the time series of daily closing quotes, from 2006-01-03 to 2007-12-31, for the following 20 financial assets:

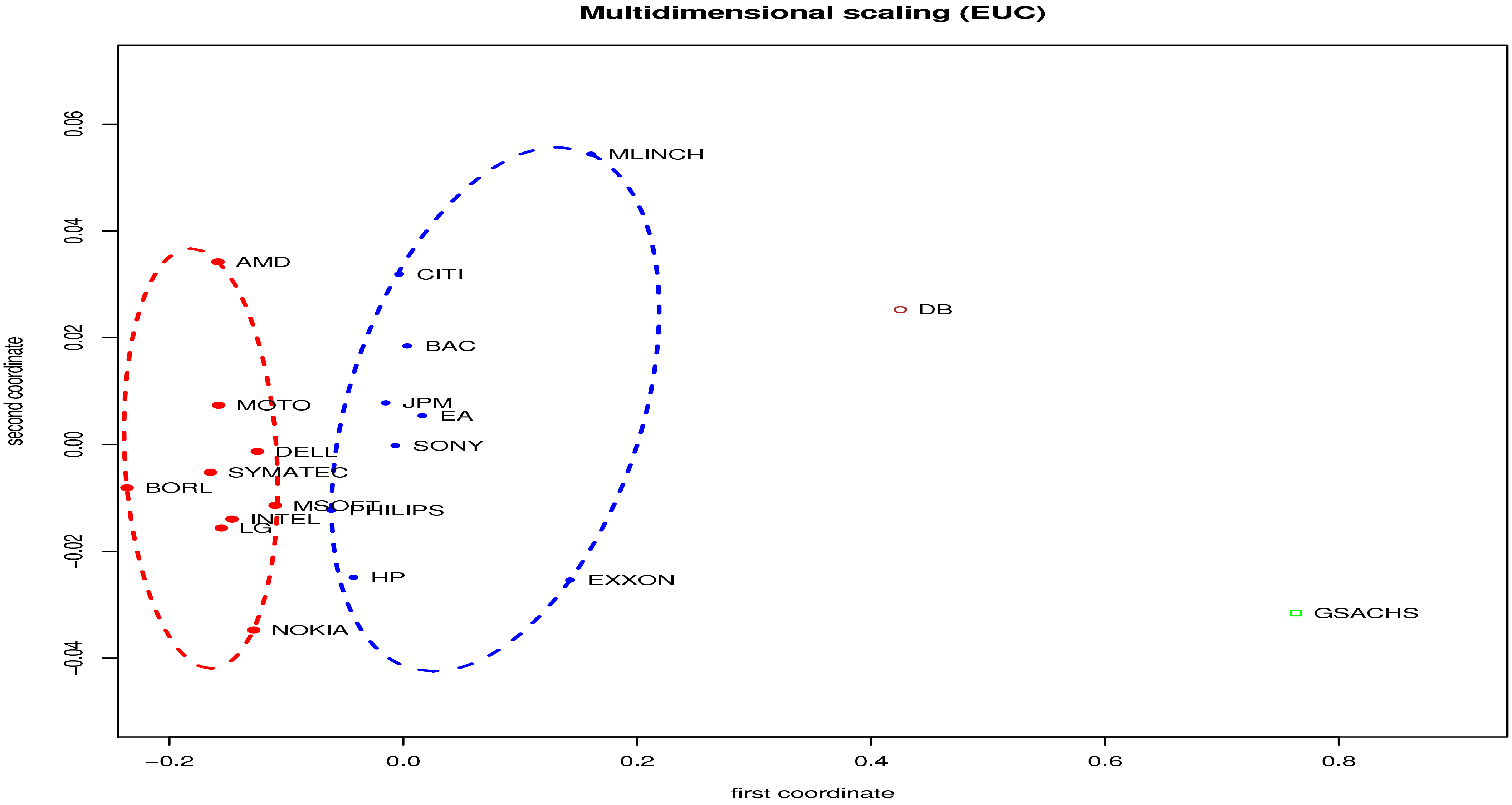
Microsoft Corporation (MSOFT in the plots)	Advanced Micro Devices Inc. (AMD)
Dell Inc. (DELL)	Intel Corporation (INTEL)
Hewlett-Packard Co. (HP)	Sony Corp. (SONY)
Motorola Inc. (MOTO)	Nokia Corp. (NOKIA)
Electronic Arts Inc. (EA)	LG Display Co., Ltd. (LG)
Borland Software Corp. (BORL)	Koninklijke Philips 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)	Exxon Mobil Corp. (EXXON)

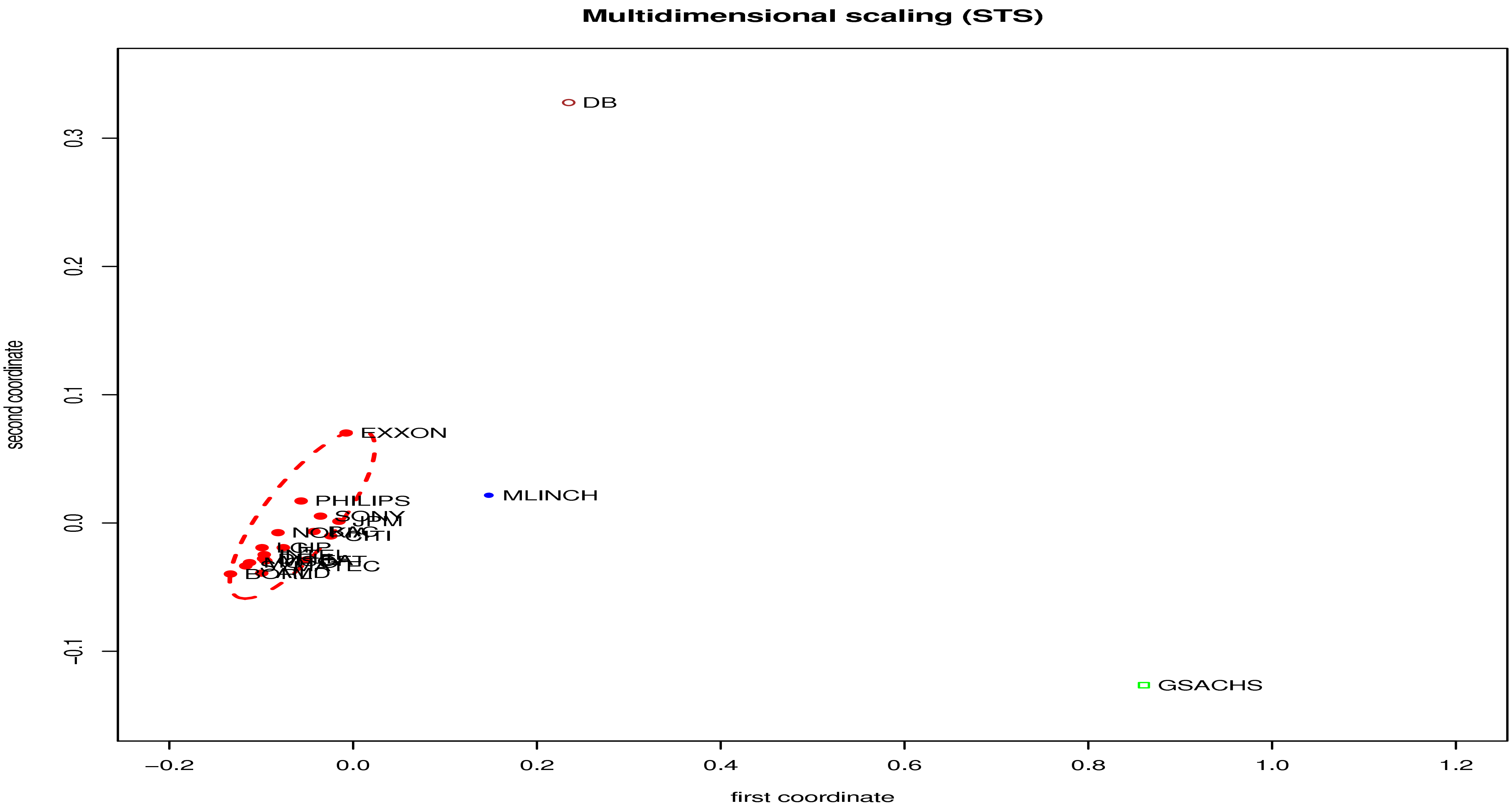
Quotes come from NYSE/NASDAQ. Source Yahoo.com.

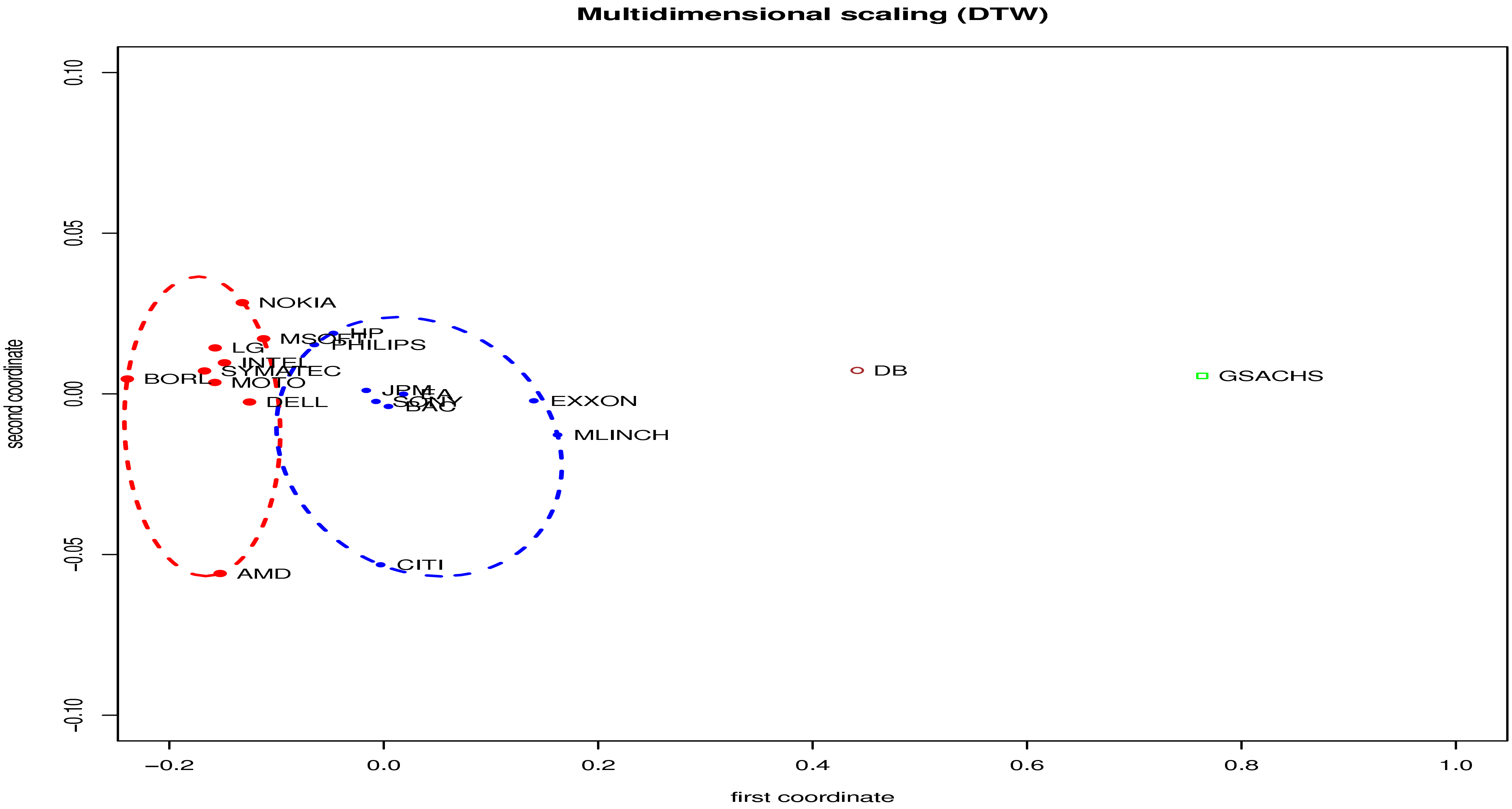
Real Data from NYSE



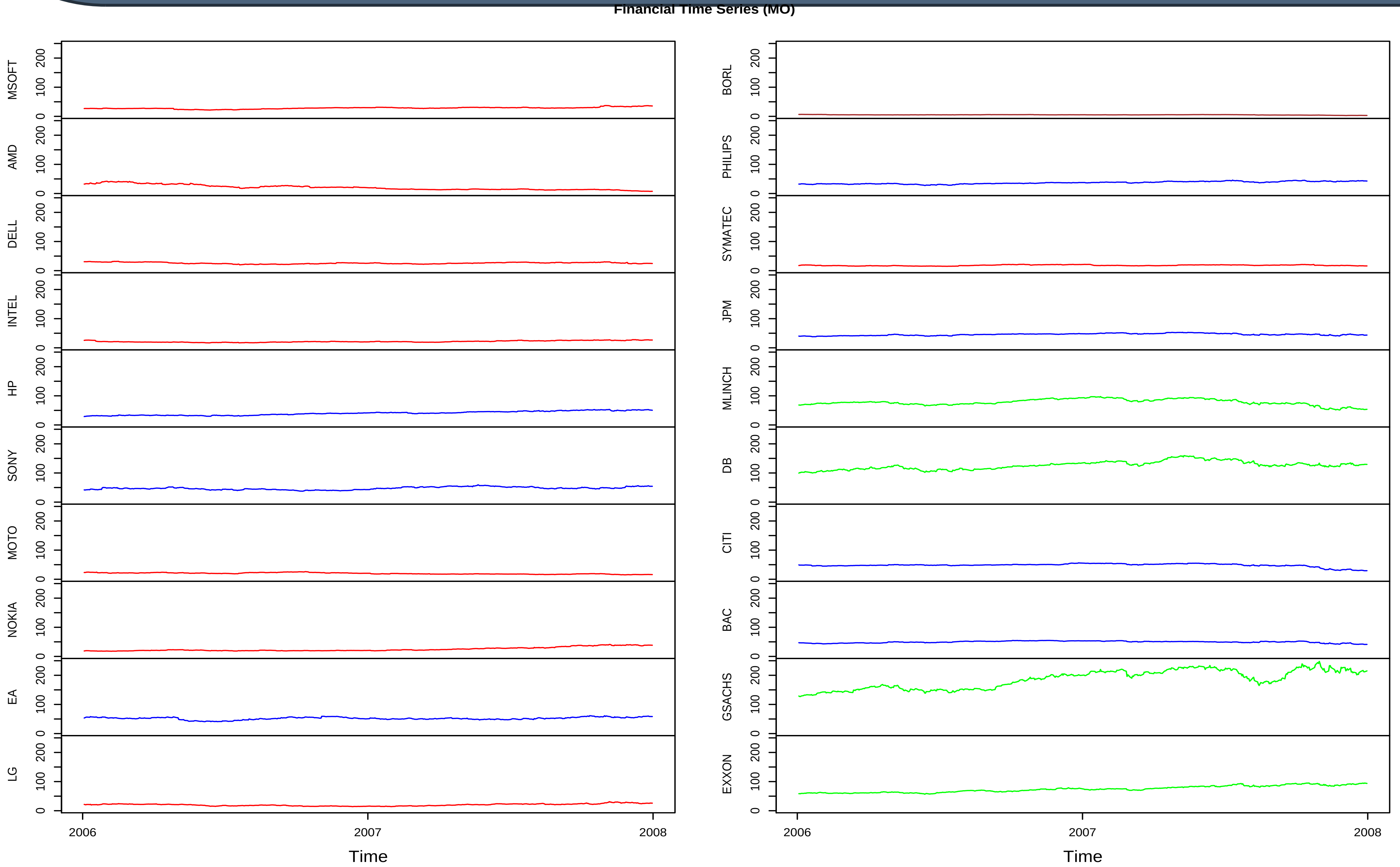

Multidimensional scaling (MO)
second coordinate
first coordinate
0.0
−0.2
−0.4
−0.6
−0.8
−0.4
−0.2
0.0
0.2
0.4
MSOFT
DELL
INTEL
NOKIA
MOTO
LG
SYMATEC
AMD
PHILIPS
HP
JPM
SONY
CITI
BAC
EA
EXXON
MLINCH
DB
GSACHS
BORL


Multidimensional scaling (EUC)
second coordinate
first coordinate
0.06
0.04
0.02
0.00
−0.02
−0.04
−0.2
0.0
0.2
0.4
0.6
0.8
MLINCH
AMD
CITI
DB
BAC
MOTO
JPM
EA
SONY
DELL
SYMATEC
BORL
MSOFT
PHILIPS
INTEL
LG
HP
EXXON
GSACHS
NOKIA

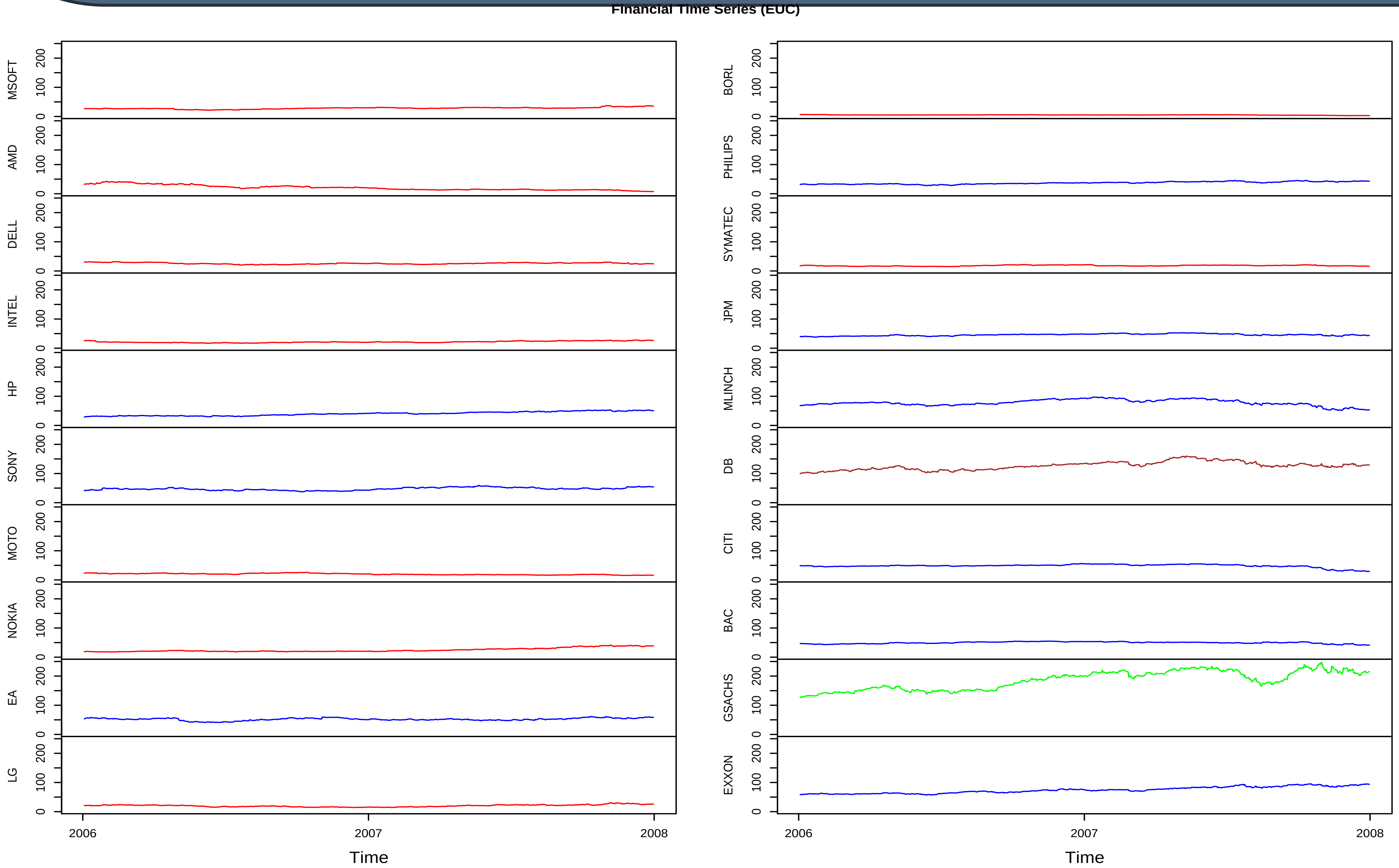

Multidimensional scaling (STS)
DB
EXXON
MLINCH
PHILIPS
GSACHS
first coordinate
second coordinate
−0.2
0.0
0.2
0.4
0.6
0.8
1.0
1.2
-0.1
0.0
0.1
0.2
0.3


Multidimensional scaling (DTW)
second coordinate
first coordinate
NOKIA
MSOFT
HP
PHILIPS
LG
INTEL
SYMATEC
BORL
MOTO
DELL
JPM
SONY
BAC
EXXON
MLINCH
CITI
AMD
DB
GSACHS
0.10
0.05
0.00
−0.05
−0.10
−0.2
0.0
0.2
0.4
0.6
0.8
1.0

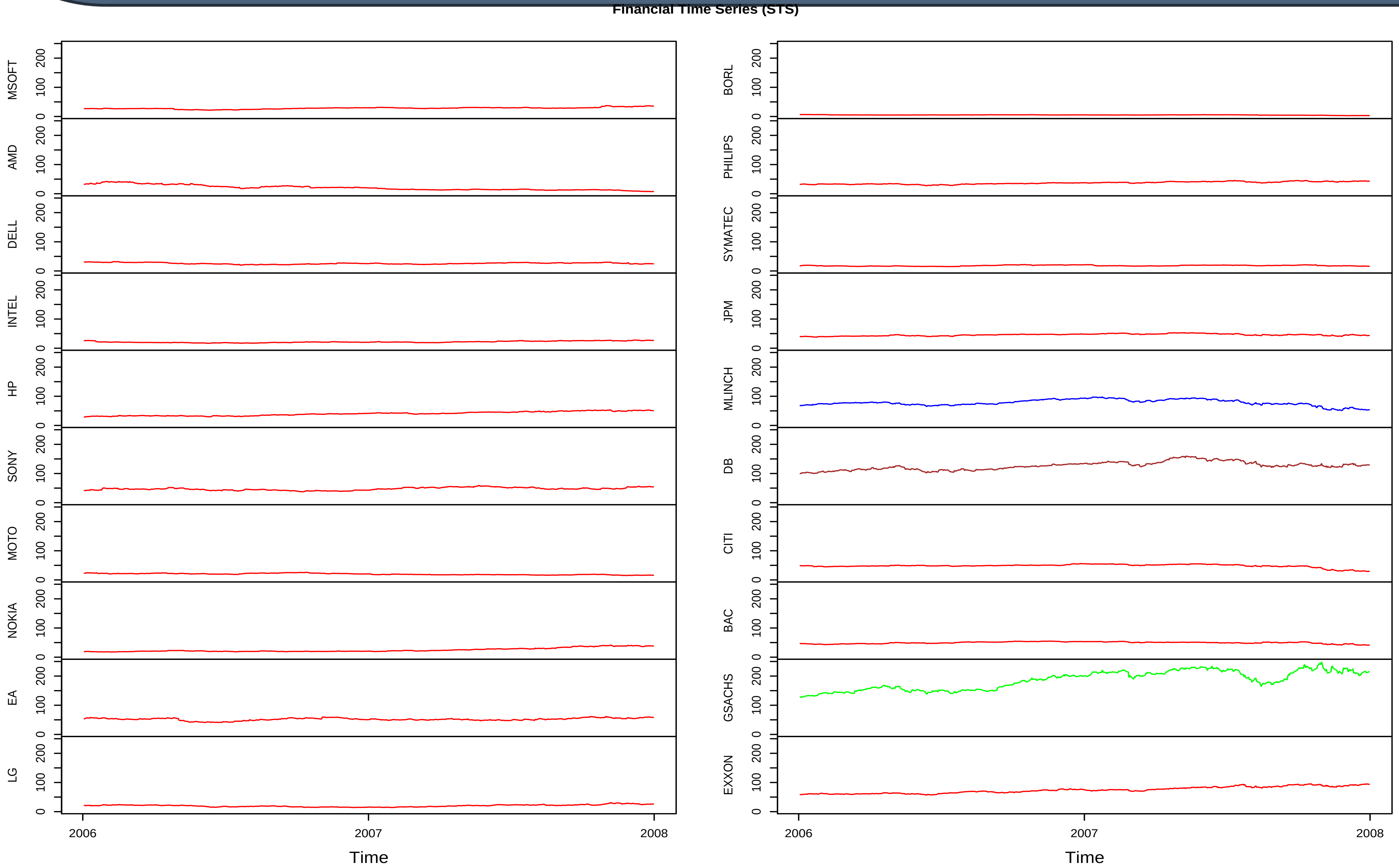
Data clustered according to MO distance



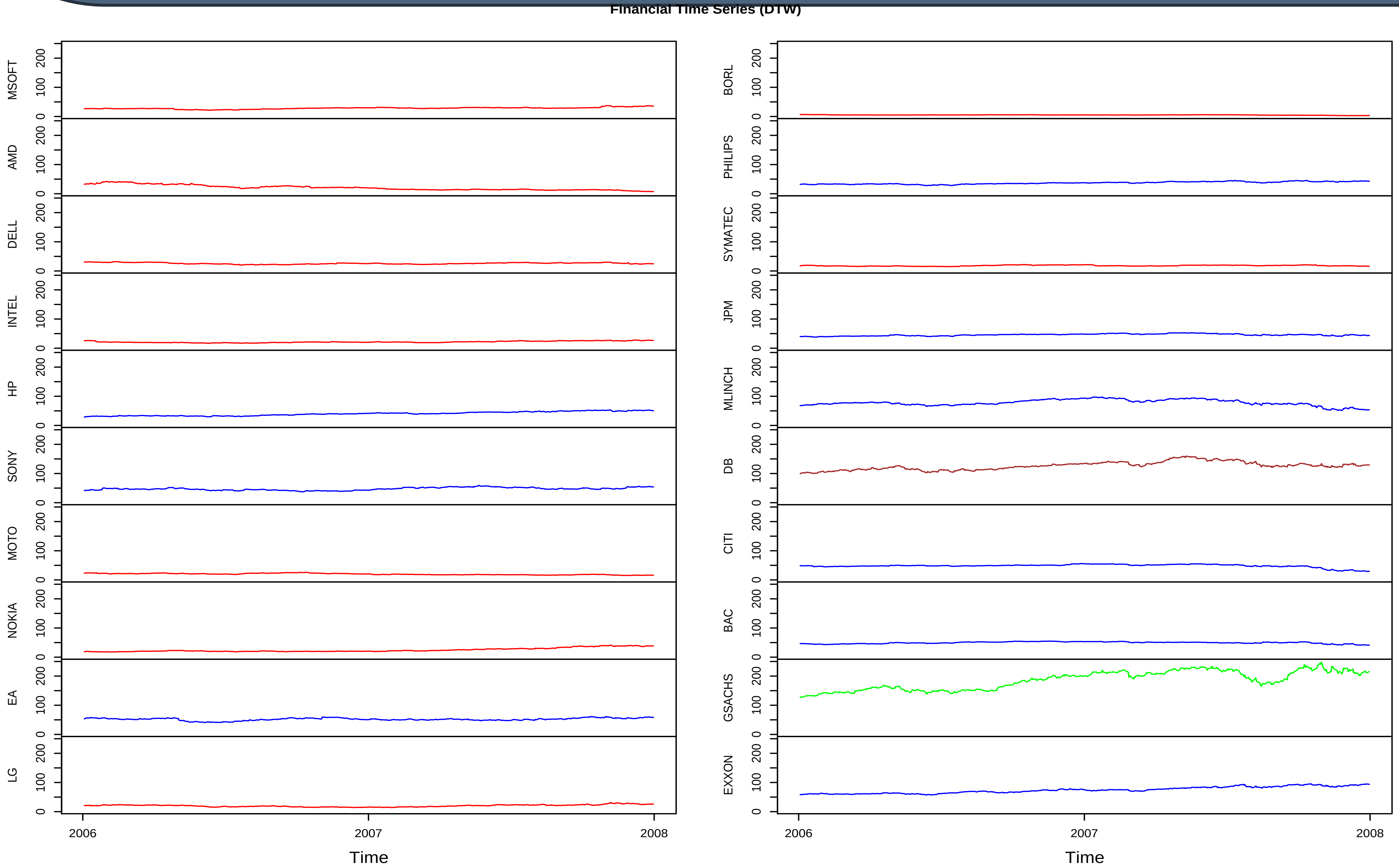
Data clustered according to EUC distance



Data clustered according to STS distance



Data clustered according to DTW distance



The package `sde` for the `R` statistical environment is freely available at `http://cran.R-Project.org`.

It contains the function `MOdist` which calculates the Markov Operator distance and returns a `dist` object.

```
data(quotes)

d <- MOdist(quotes)
cl <- hclust( d )
groups <- cutree(cl, k=4)
plot(quotes, col=groups)

cmd <- cmdscale(d)
plot( cmd, col=groups)
text( cmd, labels(d) , col=groups)
```

Aït-Sahalia, Y. (1996) Testing continuous-time models of the spot interest rate, *Rev. Financial Stud.*,

Amari, S. (1985) *Differential-Geometrical Methods in Statistics*, Lecture Notes in Statist., Vol. **28**, New York, Springer.

Bhattacharyya, A. (1946) On some analogues to the amount of information and their uses in statistical estimation, *Sankhya*, **8**, 1-14.

Beran, R.J. (1977) Minimum Hellinger estimates for parametric models, *Annals of Statistics*, **5**, 445-463.

Chen, S.X., Gao, J., Cheng, Y.T. (2008) A test for model specification of diffusion processes, *Ann. Stat.*, **36**(1), 167-198.

Cressie, N., Read, T.R.C. (1984) Multinomial goodness of fit tests, *J. Roy. Statist. Soc. Ser B*, **46**, 440-464.

Csiszár, I. (1963) Eine Informationstheoretische Ungleichung und ihre Anwendung auf den Beweis der Ergodizität von Markoffschen Ketten. Publ. Math. Inst. Hungar. Acad. Sci. Ser. A **8**, 85-108.

Csiszár, I. (1967) On topological properties of f-divergences, *Studia Scientific Mathematicae Hungarian*, **2**, 329-339.

Dachian, S., Kutoyants, Yu. A. (2008) On the goodness-of-t tests for some continuous time processes, in *Statistical Models and Methods for Biomedical and Technical Systems*, 395-413. Vonta, F., Nikulin, M., Limnios, N. and Huber-Carol C. (Eds), Birkhuser, Boston.

De Gregorio, A., Iacus, S.M. (2008) Rényi information for ergodic diffusion processes. *Information Sciences*, **179**, 279-291.

Giet, L., Lubrano, M. (2008) A minimum Hellinger distance estimator for stochastic differential equations: An application to statistical inference for continuous time interest rate models, *Computational Statistics and Data Analysis*, **52**, 2945-2965.

Küchler, U., Sørensen, M. (1997) *Exponential Families of Stochastic Processes*, Springer, New York.

Kutoyants, Y. (2004) *Statistical Inference for Ergodic Diffusion Processes*, Springer-Verlag, London.

Iacus, S.M., Kutoyants, Y. (2001) Semiparametric hypotheses testing for dynamical systems with small noise, *Mathematical Methods of Statistics*, **10**(1), 105-120.

Jager, L., Wellner, J.A. (2007) Goodness-of-fit tests via phi-divergences, *Annals of Statistics*, **35**(5), 2018-2053.

Relevant references

Lee, S., Wee, I.-S. (2008) Residual emprical process for diffusion processes, *Journal of Korean Mathematical Society*, **45**(3), 2008.

Liese, F., Vajda, I. (1987) *Convex Statistical Distances*, Tuebner, Leipzig.

Masuda, H., Negri, I., Nishiyama, Y. (2008) Goodness of fit test for ergodic diffusions by discrete time observations: an innovation martingale approach, *Research Memorandum 1069, Inst. Statist. Math.*, Tokyo.

Morales, D., Pardo, L., Pardo,, M.C., Vajda, I. (2004) Rényi statistics in directed families of exponential experiments, *Statistics*, **38**(2), 133-147.

Negri, I., Nishiyama, Y. (2008) Goodness of fit test for ergodic diffusion processes, to appear in *Ann. Inst. Statist. Math.*.

Negri, I. and Nishiyama, Y. (2007). Goodness of t test for small diffusions based on discrete observations, *Research Memorandum 1054, Inst. Statist. Math.*, Tokyo.

Pardo, L. (2006) *Statistical Inference Based on Divergence Measures*, Chapman & Hall/CRC, London.

Pritsker, M. (1998) Nonparametric density estimation and tests of continuous time interest rate models, *Rev. financial Studies*, **11**, 449-487.

Rényi, A. (1961) On measures of entropy and information, *Proceedings of the Fourth Berkeley Symposium on Probability and Mathematical Statistics*, Vol 1, University of California, Berkeley, 547-461.

Rivas, M.J., Santos, M.T., Morales, D. (2005) Rényi test statistics for partially observed diffusion processes, *Journal of Statistical Planning and Inference*, **127**, 91-102.

Simpson, D.G. (1989) Hellinger deviance tests: Efficinecy, breakdown points, and examples, *J. Amer. Statist. Assoc.*, **84**, 107-113.

Uchida, M., Yoshida, N. (2001) Information Criteria in Model Selection for Mixing Processes, *Statistical Inference for Stochastic Processes*, **4**, 73-98.

Uchida, M., Yoshida, N. (2004) Information Criteria for Small Diffusions via the Theory of Malliavin-Watanabe, *Statistical Inference for Stochastic Processes*, **7**, 35-67.

Uchida, M., Yoshida, N. (2005) AIC for ergodic diffusion processes from discrete observations, preprint MHF 2005-12, march 2005, *Faculty of Mathematics, Kyushu University*, Fukuoka, Japan.

Yoshida, N. (1992) Estimation for diffusion processes from discrete observation, *J. Multivar. Anal.*, **41**(2), 220-242.