Diffusion processes as models in social sciences. A review and some new challenges

Ilia Negri

Department of Information Technology and Mathematical Methods
University of Bergamo (Italy)
ilia.negri@unibg.it

7th RC33 International Conference on Social Science Methodology
September 1 – 5, 2008
Campus di Monte Sant'Angelo, Università di Napoli Federico II (Italy)

Plan of the talk

- Diffusion processes
  - Definition
  - Properties
  - Examples
- Inference for diffusions
  - Estimation
  - Goodness of fit tests (continuous time observations, discrete time observations)
- Simulation study

Diffusion processes: definition

Let $\{X_t : t \geq 0\}$ a quantitative variable of interest varying continuously over time.

If the trajectory of X_t is nowhere differentiable the classical formulation

$$\frac{dX_t}{dt} = S(X_t) + W_t$$

is ill-defined.

The Itô definition of stochastic differential equation for the process X_t starts from a different point of view.

Let us suppose that we can describe the rate of change in X_t as

$$\mathbf{E}(X_{t+h} - X_t | \mathcal{F}_t) = S(X_t)h + o(h)$$

and we can specify the variance of this rate of change as

$$\mathbf{E}((X_{t+h} - X_t - S(X_t)h)^2 | \mathcal{F}_t) = \sigma(X_t)^2 h + o(h)$$

The functions S and σ^2 (that are measurable functions with respect to $\mathcal{F}_t$, the σ algebra generated by $\{X_s : s \leq t\}$), are called the *drift* and the *diffusion* function of the diffusion process X_t.

Itô's idea: to construct the sample paths of a diffusion from the sample paths of a Brownian motion.

Let W_t, $t \geq 0$ a Brownian motion. We can define a class of diffusion process as the (continuous) solution of a stochastic differential equation (SDE) given by

$$X_t = X_0 + \int_0^t S(X_s)ds + \int_0^t \sigma(X_s)dW_s, \quad t \geq 0$$

or, written in differential form

$$dX_t = S(X_t)dt + \sigma(X_t)dW_t, \quad X_0 = \xi, \quad t \geq 0$$

Diffusion processes: properties

Let us introduce some conditions for the pair of functions (S, σ).

A1. There exists a constant $C > 0$ such that

$$|S(x) - S(y)| \leq C|x - y|, \quad |\sigma(x) - \sigma(y)| \leq C|x - y|.$$

Under **A1**, the SDE has a unique strong solution X_t, $t \geq 0$.

Ergodic diffusion are interesting for their stationary properties.

The process X_t has the ergodic property if there exists a measure μ such that for every function $g \in L^1(\mu)$

$$\mathbf{P}\left(\lim_{T\to\infty} \frac{1}{T}\int_0^T g(X_t)\mathrm{d}t = \int_{\mathbb{R}} g(z)\mu(dz)\right) = 1$$

μ is called *invariant measure* or *invariant distribution*.

Diffusion processes: properties

A2. The diffusion process is regular. The *speed measure*

$$m(\mathrm{d}x) = \frac{1}{\sigma(x)^2 p'(x)} \mathrm{d}x$$

where p is the *scale function*

$$p(x) = \int_0^x \exp\left\{-2\int_0^y \frac{S(v)}{\sigma^2(v)}\,\mathrm{d}v\right\}\mathrm{d}y,$$

is finite and has the second moment.

Under conditions **A1** and **A2** the diffusion process is ergodic and has an invariant density given by

$$f(y) = \frac{1}{G}\frac{1}{\sigma(y)^2}\exp\left\{2\int_0^y \frac{S(v)}{\sigma(v)^2}dv\right\}$$

where G is a normalization constant given by $G = m(\mathbb{R})$.

Diffusion processes: properties

Under some regularity conditions the process X_t whose dynamics is described by the stochastic differential equation

$$dX_t = S(X_t)dt + \sigma(X_t)dW_t, \quad X_0 = \xi, \quad t \geq 0$$

has a stationary distribution of the form

$$f(y) = \frac{1}{G}\frac{1}{\sigma(y)^2}\exp\left\{2\int_0^y \frac{S(v)}{\sigma(v)^2}dv\right\}$$

and it can be used to calculate the probability that the trajectories will pass through a generic interval (a, b) of the real line once the process has reached a *statistical equilibrium*.

Statistical inference can be done to estimate the drift and/or the diffusion coefficients of the diffusion process or directly the invariant density or the distribution function of the invariant distribution of the process.

Diffusion processes: example 1

Consider a *linear feedback system* responding to a randomly changing environment. The state X_t has a goal μ toward which it moves. Its expected rate is proportional to its deviation from μ, so

$$S(x) = \theta(\mu - x)$$

where θ is the strength of the feedback system.

Moreover suppose that the disturbance is constant, so $\sigma^2(x) = \varepsilon$

The SDE describing this system is

$$dX_t = \theta(\mu - X_t)dt + \sqrt{\varepsilon}dW_t, \quad t \geq 0, \quad \varepsilon > 0$$

This process is known as *mean reverting Ornstein-Uhlenbeck* process.

The invariant distribution is Gaussian

$$f(x) = \frac{1}{\sqrt{2\pi\varepsilon/\theta}} \exp\left\{-\frac{(x-\mu)^2}{2\varepsilon/\theta}\right\}$$

with mean μ and variance ε/θ

Diffusion processes: example 2

A model for political polarization.

Let X_t the political persuasion of a subject. The persuasion is measured on the hypothetical axis of conviction. (x=0, extremely liberal, x=1 extremely conservative).

There is a tendency to move toward the average persuasion of the population, so the drift coefficient is

$$S(x) = \theta(\mu - x)$$

but the choice of σ is motivated by the fact that people who had extreme view are much less subject to random fluctuation than those near the center, so the diffusion coefficient has the form

$$\sigma^2(x) = \varepsilon x(1 - x)$$

where ε measure the amount of random change in political persuasion.

Diffusion processes: example 2

The SDE describing the motion of each person through the political spectrum is

$$dX_t = \theta(\mu - X_t)dt + \sqrt{\varepsilon X_t(1 - X_t)}dW_t, \quad t \geq 0, \quad \varepsilon > 0$$

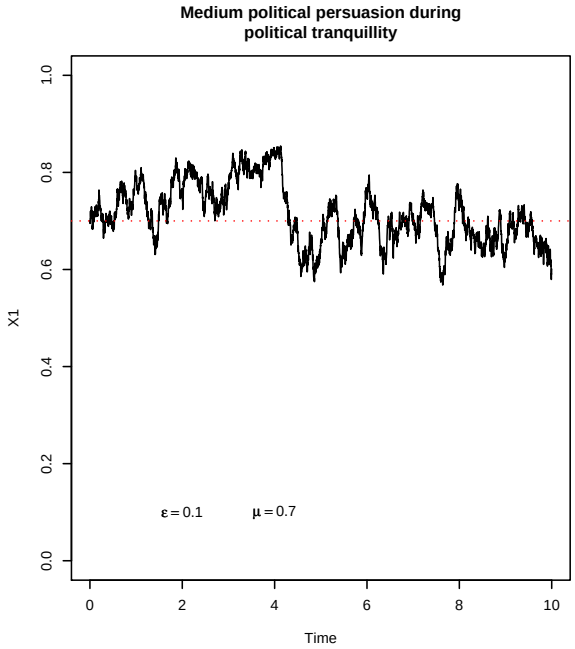
The ratio $\delta = \frac{\varepsilon}{\theta}$ controls the extent of polarization in this model.

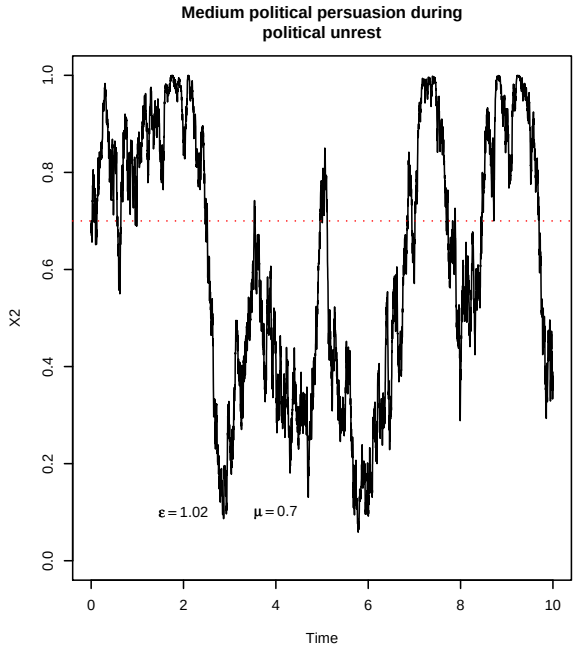
The invariant distribution is of type Beta with density

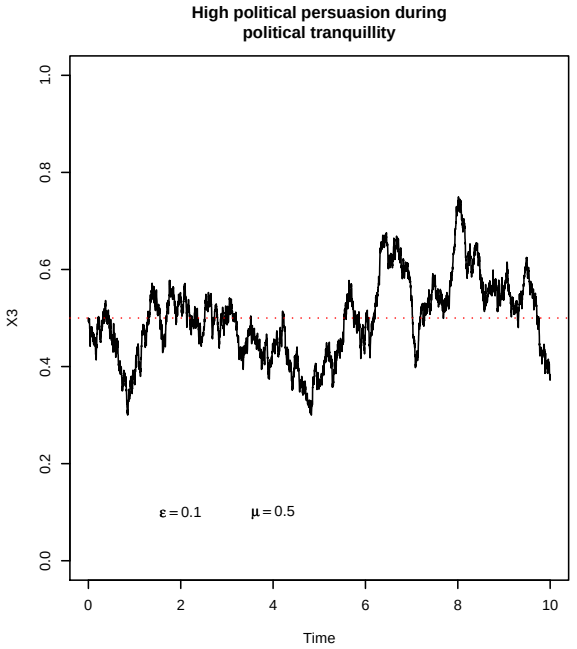
$$f(x) = \frac{\Gamma\left(\frac{1}{\delta}\right)}{\Gamma\left(\frac{\mu}{\delta}\right)\Gamma\left(\frac{1-\mu}{\delta}\right)} x^{-1+\frac{\mu}{\delta}}(1-x)^{-1+\frac{1-\mu}{\delta}}$$

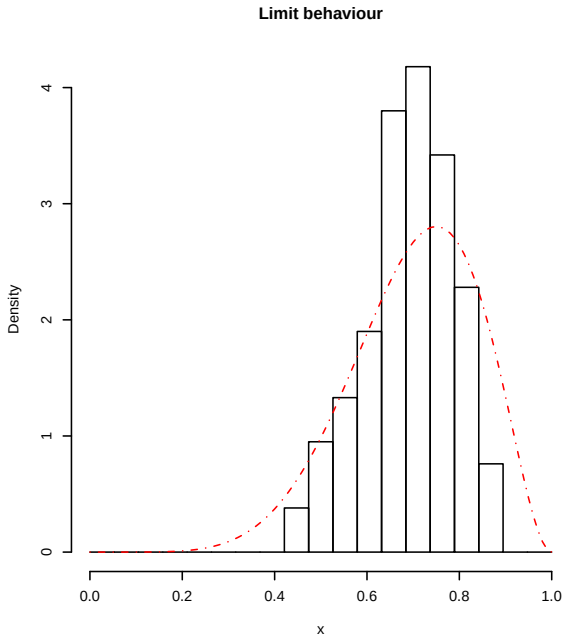
For this model the mean is μ, the mode is $\frac{1-\mu}{1-2\delta}$ and the variance is $\frac{\mu(1-\mu)}{1+\delta}$

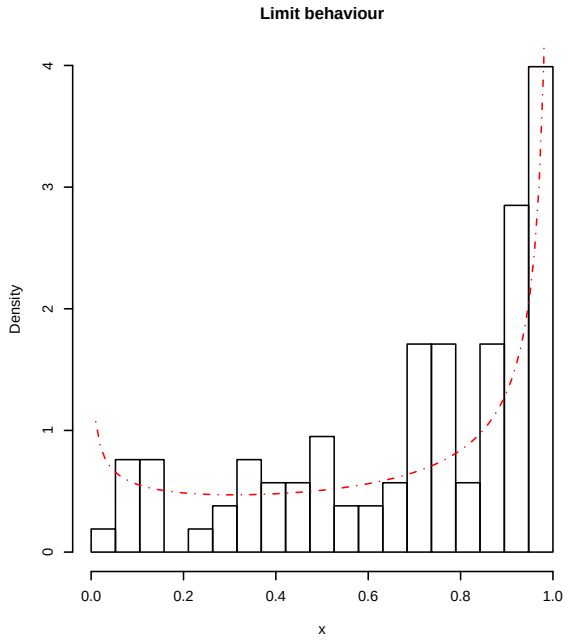
See Cobb, (1981 and 1998) for these and other examples.

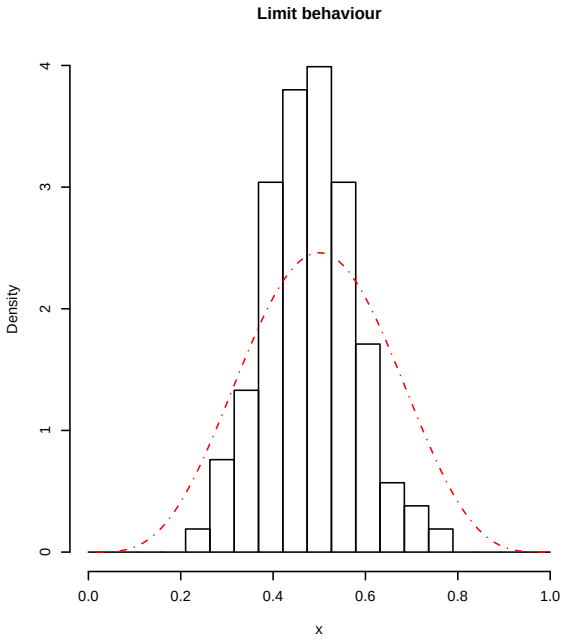

Medium political persuasion during political tranquillity
X1
Time
ε = 0.1
μ = 0.7


Medium political persuasion during political unrest
X2
Time
ε = 1.02
μ = 0.7


High political persuasion during political tranquillity
X3
Time
ε = 0.1
μ = 0.5


Limit behaviour
Density
x


Limit behaviour
Density
x


Limit behaviour
Density
x

Inference for diffusion: estimation

Parameters of these SDE can be estimated by empirical data.

Inference can be based on the observation of the diffusion process at continuous time or at discrete time according to different schemes depending on the nature of the data:

Sampling Scheme 1. Large sample scheme

The process $X = \{X_t : t \in [0, T]\}$ is observed at times $0 = t_0^n < t_1^n < \cdots < t_n^n$, where $h_i = |t_i^n - t_{i-1}^n| = h$ is constant. In this case $hn = T$ and the window of observation increase with n.

Sampling Scheme 2. Rapidly increasing design

The process $X = \{X_t : t \in [0, T]\}$ is observed at times $0 = t_0^n < t_1^n < \cdots < t_n^n$, such that as $n \to \infty$ $t_n^n \to \infty$ and $nh_n^2 \to 0$ where $h_n = \max_{1 \leq i \leq n} |t_i^n - t_{i-1}^n|$.

(See Iacus, 2008, for methods on simulation and estimation under different schemes)

Inference for diffusion: goodness of fit test

Goodness of fit tests play an important role in theoretical and applied statistics.

The origin goes back to the Kolmogorov-Smirnov and the Cramer-Von Mises tests in the i.i.d case.

Despite their importance in applications goodness of fit test for diffusion process has still been a new issue in recent years.

Such test are useful if they are distribution free or asymptotically distribution free.

We present a test for the drift of a diffusion process that is asymptotically distribution free and consistent.

The test is based on discrete time observations and the diffusion process is a nuisance function which is estimated in our test procedure.

Goodness of fit test: the i.i.d observation model

Suppose we observe $X^n = (X_1, \ldots, X_n)$ i.i.d. with distribution function F. To test the simple hypothesis

$$H_0 : F = F_0$$

against any alternative $F_1 \neq F_0$ such that $\sup_{x \in \mathbb{R}} |F_0(x) - F_1(x)| > 0$, we can use the well known Kolmogorov-Smirnov statistic

$$\Delta_n(X^n) = \sup_x \sqrt{T} |\widehat{F}_n(x) - F_0(x)|$$

where

$$\widehat{F}_n(x) = \frac{1}{n} \sum_{j=1}^{n} \mathbf{1}_{\{X_j \leq x\}}$$

is the empirical distribution function.

The Kolmogorov-Smirnov procedure is

$$\phi_n(X^n) = \mathbf{1}_{\{\Delta_n > c_\alpha\}}$$

where c_α is defined by

$$\mathbf{P}\left(\sup_{0\leq t\leq 1} |B_0(t)| > c_\alpha\right) = \alpha$$

and B_0 is a Brownian Bridge, i.e. a continuous Gaussian Process with

$$\mathbf{E}B_0(t) = 0, \quad \mathbf{E}B_0(t)B_0(s) = t \wedge s - ts$$

It turns out that the Kolmogorov-Smirnov test is such that

- it is asymptotically of *size* α
- it is asymptotical distribution-free
- it is consistent (uniformly) against any alternative $F_1 \neq F_0$.

Goodness of fit test for diffusion based on $\widehat{F}_T(x)$

Suppose we want to test

$$H_0 : S = S_0$$

against any alternative $F_1 \neq F_0$. We can introduce the statistic

$$\Delta_T(X^T) = \sup_x \sqrt{T}|\widehat{F}_T(x) - F_{S_0}(x)|$$

where the empirical distribution function based on continuous observation of the diffusion process over $[0, T]$ is

$$\widehat{F}_T(x) = \frac{1}{T}\int_0^T \mathbf{1}_{\{X_t \leq x\}} dt.$$

The decision function is

$$\widehat{\phi}_T(X^T) = \mathbf{1}_{\{\Delta_T(X^T) > c_\alpha\}}$$

where c_α is the solution of the equation

$$\mathbf{P}\left(\sup_x |\eta_{S_0}(x)| > c_\alpha\right) = \alpha$$

and $\eta_{S_0}(x)$, $x \in \mathbb{R}$ is a suitable Gaussian process.

Goodness of fit test for diffusion based on $\widehat{F}_T(x)$

It turns out that the such test

- it is asymptotically of *size* α (Kutoyants 2004)
- it is NOT asymptotical distribution-free
- it is consistent (uniformly) against any alternative $S_1 \neq S_0$ in the sense that $\sup_x |S_1(x) - S_0(x)| > 0$.

Remarks: The fact that $\phi_T(X^T) \in \mathcal{K}_\alpha$ follows from the weak convergence of the process $\eta^T_{S_0}$ to the gaussian process η_{S_0} (Negri 1998) and the continuous mapping theorem.

Remarks: Due to the structure of the covariance function of η_{S_0} the Kolmogorov-Smirnov statistics is not asymptotically distribution free.

Goodness of fit test for diffusion based on $V_T(x)$

To test the simple hypothesis $H_0 : S = S_0$ against any alternative $S_1 \neq S_0$, we introduce the *score marked empirical process*

$$V_T(x) = \frac{1}{\sqrt{T}} \int_0^T \mathbf{1}_{(-\infty,x]}(X_t)(\mathrm{d}X_t - S_0(X_t)\mathrm{d}t)$$

The test based on the following statistical decision function

$$\phi_T^* = \mathbf{1}_{\left\{\frac{1}{\Gamma_0}\sup_{x\in\mathbb{R}}|V_T(x)|>c_\alpha\right\}}$$

where the *critical value* c_α is defined by

$$\mathbf{P}\left(\sup_{0\leq t\leq 1}|B(t)| > c_\alpha\right) = \alpha,$$

is asymptotically distribution free and consistent against any alternative

$$\int_{-\infty}^{+\infty} \mathbf{1}_{(-\infty,x]}(y)(S_0(y) - S_1(y))f_{S_1}(y)\mathrm{d}y \neq 0$$

(Negri and Nishiyama 2008)

Goodness of fit test for diffusion: discrete time observations

The test procedure presented is based on the continuous observation of the process on $[0, T]$.

The new test procedure is based on discrete time observation which is more realistic in applications.

Precisely the sampling scheme is the following

Sampling Scheme. The process $X = \{X_t : t \in [0, T]\}$ is observed at times $0 = t_0^n < t_1^n < \cdots < t_n^n = T$, such that as $n \to \infty$ $t_n^n \to \infty$ and $h_n = O(n^{-2/3}(\log n)^{1/3})$ (which implies $nh_n^2 \to 0$) where $h_n = \max_{1 \le i \le n} |t_i^n - t_{i-1}^n|$.

This condition implies $h_n \to 0$ so we can assume $h_\varepsilon \le 1$ without loss of generality.

Goodness of fit test for diffusion: test statistic

Our test statistics is based on the random field $U^n = \{U^n(x); x \in [-\infty, \infty]\}$ defined by $U^n(-\infty) = 0$ and

$$U^n(x) = \frac{1}{\sqrt{t_n^n}} \sum_{i=1}^{n} \psi_k^n(X_{t_{i-1}^n})[X_{t_i^n} - X_{t_{i-1}^n} - S_0(X_{t_{i-1}^n})|t_i^n - t_{i-1}^n|]$$

for $x \in (x_{k-1}^n, x_k^n],\ 1 \le k \le m(n) + 1$.

We call it the *smoothed score marked empirical process* based on discrete observation.

This U^n is an approximation of the random field $V^n = \{V^n(x); x \in [-\infty, \infty]\}$ defined by

$$V^n(x) = \frac{1}{\sqrt{t_n^n}} \int_0^{t_n^n} 1_{(-\infty,x]}(X_t)[dX_t - S_0(X_t)dt],$$

which is the *score marked empirical process* based on continuous observation.

Fix a number $\alpha \in (0,1)$, the test procedure is defined by the following statistical decision function

$$\phi_n^* = \mathbf{1}_{\left\{\frac{\sup_{x\in[-\infty,+\infty]}|U^n(u)|}{\widehat{\Sigma}^n} > c_\alpha\right\}},$$

where the *critical value* c_α is defined by

$$\mathbf{P}\left(\sup_{0\leq t\leq 1}|B(t)| > c_\alpha\right) = \alpha,$$

and $\widehat{\Sigma}^n$ is the consistent estimator prosed for

$$\Sigma_{S,\sigma} := \sqrt{\int_{-\infty}^{\infty}\sigma(z)^2 f_{S,\sigma}(z)dz} > 0$$

defined by

$$\widehat{\Sigma}^n = \sqrt{\frac{1}{t_n^n}\sum_{i=1}^{n}|X_{t_i^n} - X_{t_{i-1}^n}|^2}.$$

It turns out (Masuda, Negri and Nishiyama, 2008) that such test

- it is asymptotically of *size* α
- it is asymptotical distribution-free
- it is consistent (uniformly) against any alternative $S_1 \neq S_0$ in the sense that $\int_{-\infty}^{x_S}(S(z) - S_0(z))f_{S,\sigma}(z)dz \neq 0$ for some $x_S \in (-\infty, \infty]$.

Remark: one may think that it is more natural to consider the random field

$$\widetilde{U}^n(x) = \frac{1}{\sqrt{t_n^n}} \sum_{i=1}^{n} 1_{(-\infty,x]}(X_{t_{i-1}^n})[X_{t_i^n} - X_{t_{i-1}^n} - S_0(X_{t_{i-1}^n})|t_i^n - t_{i-1}^n|] \quad x \in [-\infty, +\infty]$$

instead of U^n. At least in our proof, the uniform approximation is due to the continuity of the function $z \rightsquigarrow \psi_k^n(z)$, so it does not seem easy to translate the result for V^n into that for $\widetilde{U}^n$

The limit distribution of the test statisitcs

The distribution function of supremum over $t \in [0,1]$ of $|W(t)|$ is the limit as n tends to infinity of the supremum of a normalised poisson process with very high intensity

$$\mathbf{P}\left(\frac{\sup_{t\in[0,1]}|\xi_n(t)-nt|}{\sqrt{n}} < x\right),$$

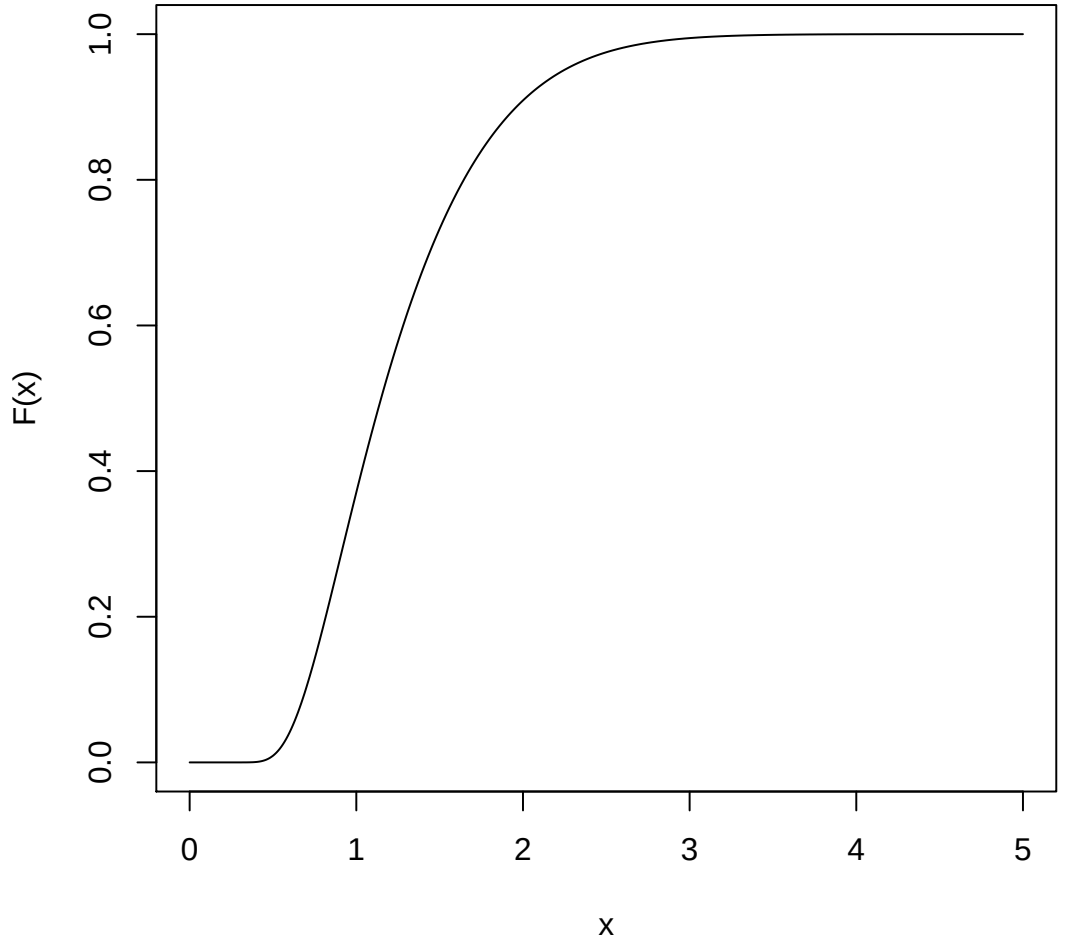
where $\xi_n(t)$ with $t \in [0,1]$ is a Poisson process with intensity n.

The distribution function has been recently calculated from the formula

$$F(x) = \frac{4}{\pi}\sum_{n=0}^{+\infty}\frac{(-1)^n}{2n+1}\exp\left\{-\frac{\pi^2(2n+1)^2}{8x^2}\right\}$$

These formula is easy to implement to find the values of the distribution function corresponding to a selection of p-values. Note that for values of x less than 5, it is only necessary to sum the first 5 terms of the formula to obtain a result accurate to 8 decimal places.

Distribution function of the $\sup_{t\in[0,1]} |W(t)|$.



Goodness of fit test for diffusion: a simulation study

In this section we observe finite-sample performance of our test statistics through numerical experiments. For true (data-generating) process we adopt the Ornstein-Uhlenbeck diffusion starting from the origin:

$$X_t = \int_0^t (-2X_s)ds + W_t. \tag{1}$$

For simplicity we here focus on the equidistant sampling case, that is, $h_n = t_i^n - t_{i-1}^n$ for every $i \leq n$.

We are going to observe the following (a) and (b), in both of which we will take $x_k^n = -n + \frac{k}{n}$ for $k = 1, 2, \ldots, 2n^2$, and $b_n = \frac{1}{100} n^{-1/4} \mathrm{Log} n$:

(a) asymptotic behavior of test statistics with $S_0(x) = -2x$;

(b) asymptotic behavior of test statistics with with $S_0(x) = -4x$.

Throughout we take the significance level to be 0.05. Then we see that $F(x) = 0.95$ for $x \doteq 2.24$, hence the critical region is $\{x > 2.24\}$, and:

if we denote our test statistics with

$$\mathcal{T}^n = \frac{\sup_{x\in[-\infty,\infty]} |U^n(x)|}{\widehat{\Sigma}^n}$$

and

(a) $\mathcal{T}_0^n := \mathcal{T}^n$ with $S_0(x) = -2x$;

(b) $\mathcal{T}_1^n := \mathcal{T}^n$ with $S_0(x) = -4x$.

what we expect is

- $P(\mathcal{T}_0^n > 2.24)$ should tend to 0.05 in (a);
- $P(\mathcal{T}_1^n > 2.24)$ should tend to 1.0 in (b).

For several different terminal time t_n^n and sampling frequency h_n, we simulate 1000 independent copies of a discrete sample trajectory of (1) to obtain, say $\{(\mathcal{T}_0^{n,l}, \mathcal{T}_1^{n,l})\}_{l=1}^{1000}$. We then compute:

- the empirical size (e.s.) defined by $\sharp\{l : \mathcal{T}_0^{n,l} > 2.24\}/1000$, the sample proportion of making incorrect rejections of the null;

- the empirical power (e.p.) defined by $\sharp\{l : \mathcal{T}_1^{n,l} > 2.24\}/1000$, the sample proportion of making successful rejections of the null.

	$t_n^n = 10$		$t_n^n = 20$		$t_n^n = 50$	
h_n	e.s.	e.p.	e.s.	e.p.	e.s.	e.p.
0.1	0.043	0.438	0.059	0.633	0.067	0.943
	$(n = 100)$		$(n = 200)$		$(n = 500)$	
0.05	0.050	0.412	0.047	0.596	0.060	0.911
	$(n = 200)$		$(n = 400)$		$(n = 1000)$	

Finally, it is conjectured that the same result would hold also for $\widetilde{U}^n$; The next Table shows results of experimental trials for this test statistics. Though in this case our theory cannot confirm the same asymptotic behaviors of $\mathcal{T}_0^n$ and $\mathcal{T}_1^n$ as before, from the table we may expect that this is also the case.

	$t_n^n = 10$		$t_n^n = 20$		$t_n^n = 50$	
h_n	e.s.	e.p.	e.s.	e.p.	e.s.	e.p.
0.1	0.054	0.479	0.047	0.662	0.063	0.938
	$(n = 100)$		$(n = 200)$		$(n = 500)$	
0.05	0.042	0.455	0.059	0.613	0.038	0.894
	$(n = 200)$		$(n = 400)$		$(n = 1000)$	

References

- Coob (1981, 1998), Stochastic Differential Equation for the Social Sciences, in *Mathematical Frontiers of the Social and Political Sciences*, Westview Press.

- Negri, (1998). Stationary distribution function estimation for ergodic diffusion process. *Stat. Inf. for Stoch. Processes*.

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

- Iacus, (2008). *Simulation and Inference for Stochastic Differential Equations*, Springer.

- Negri and Nishiyama, (2008). Goodness of fit test for ergodic diffusion processes. *Ann. Inst. Statist. Math.*

- Masuda, Negri and Nishiyama (2008). Goodness of fit test for ergodic diffusions: an innovation martingale approach. (submitted).