Nonparametric sequential estimation for diffusion processes based on discrete data

S. PERGAMENCHTCHIKOV

Laboratoire de Mathématiques Raphaël Salem
Université de Rouen, France

Joint work with L. GALTCHOUK

Le Mans, Septembre 17 - 18 , 2026

Ce projet est cofinancé par l'Union européenne.
L'Europe s'engage en Normandie
avec le Fonds européen de Développement Régional (FEDER)

RÉGION NORMANDIE

1 Model

2 Pointwise estimation

3 Regression model

4 Model selection

5 Oracle inequalities

6 Conclusion

Big Data diffusion models

De Gregorio and Iacus (2012), Fujimori (2018)

$$\mathrm{d}y_t = \left(g_0(y_t) + \sum_{j=1}^{m} \beta_j \, g_j(y_t) \right) \mathrm{d}t + \mathrm{d}W_t \,, \quad 0 \le t \le T \,,$$

where $(g_j)_{0 \le j \le m}$ are linear independent known functions.

The problem is to estimate $(\beta_j)_{1 \le j \le m}$ on the basis of discrete data

$$(y_{t_j})_{1 \le j \le N} \,, \quad t_j = j\delta \,, \qquad N = [T/\delta] \,, \quad m > N \,.$$

If m is unknown it is not possible to apply the usual methods as LASSO algorithm or Dantzig selector. In this case we set

$$S(x) = g_0(x) + \sum_{j=1}^{m} \beta_j \, g_j(x) \,.$$

Nonparametric model

In this talk we consider nonparametric diffusion model

$$\mathrm{d}y_t = S(y_t)\,\mathrm{d}t + b(y_t)\,\mathrm{d}W_t\,, \quad 0 \le t \le T\,,$$

where $(W_t)_{t\ge 0}$ is a standard Wiener process, the initial value y_0 is a given constant, $S(\cdot)$ is unknown drift function and $b(\cdot)$ is unknown diffusion coefficient.
Such models appear in a number of applications: the geography; signal processing, stochastic control; filtering; spectral analysis; identification of dynamic system; financial mathematics and others. See, for example,
ARATO, KOLMOGOROV, SINAI (1962),
NOVIKOV (1971), LIPTSER and SHIRYAEV (2001),
KARATZAS and SHREVE (1998),
DALALYAN AND KUTOYANTS (2002),
KUTOYANTS (2003)
and etc.

Problem

The goal is to estimate $S(\cdot)$ on the basis of the discrete observations

$$(y_{t_j})_{1\le j\le N}\,, \quad t_j = j\delta\,, \quad N = N(T) = [T/\delta]\,,$$

where the frequency $\delta = \delta_T \in (0,1)$. To simplify the presentation we consider the case when $b(\cdot) \equiv 1$.

The problem is to study the estimation problem w.r.t. *quadratic risk*

$$\mathcal{R}_S(\widehat{S}_N) \;=\; \mathbf{E}_S\,\|\widehat{S}_N - S\|^2 \quad \text{and} \quad \|S\|^2 = \int_{\mathbf{x}_0}^{\mathbf{x}_1} S^2(x)\mathrm{d}x\,,$$

where $\widehat{S}_N$ is an estimator for S, $\mathbf{E}_S$ is the expectation over the observation distribution $\mathbf{P}_S$ corresponding to the model with S.

Kernel estimators

For the estimation of the $S(z)$ at some fixed point $\mathbf{x}_0 < z < \mathbf{x}_1$ usually one uses the kernel estimators. In the complete data case

$$\widehat{S}_T(z) = \frac{\int_0^T \chi\left(\frac{y_t - z}{h}\right) \mathrm{d}y_t}{\int_0^T \chi\left(\frac{y_t - z}{h}\right) \mathrm{d}t},$$

where

$\chi(y) = \chi_{\{|y| \leq 1\}}$ is kernel function and

$h > 0$ is the bandwidth which goes to zero as $T \to \infty$.

Sequential version

Galtchouk and Pergamenchtchikov (2001) proposed the following sequential kernel estilatoir version

$$\widehat{S}_{\tau}(z) = \frac{1}{H} \int_0^{\tau} \chi\left(\frac{y_t - z}{h}\right) \mathrm{d}y_t \,,$$

where

$$\tau = \inf\left\{ t \geq 0 \,:\, \int_0^t \chi\left(\frac{y_u - z}{h}\right) \mathrm{d}u \geq H \right\}$$

and

H is some positive non random threshold.

Decomposition for sequential estimator

This estimator can be represented as

$$\widehat{S}_\tau(z) = S(z) + B(z) + \frac{1}{\sqrt{H}}\xi(z)\,,$$

where

$$B(z) = \frac{1}{H}\int_0^\tau \chi\left(\frac{y_t - z}{h}\right)(S(y_t) - S(z)\mathrm{d}t$$

and

$$\xi(z) = \frac{1}{\sqrt{H}}\int_0^\tau \chi\left(\frac{y_t - z}{h}\right)\mathrm{d}W_t \quad \sim \quad \mathcal{N}(0,1)\,.$$

In the case of discrete observations

$$\widehat{S}_N(z) = \frac{\sum_{j=1}^{N} \chi_{j,z} \left(y_{t_j} - y_{t_{j-1}}\right)}{\delta \sum_{j=1}^{N} \chi_{j,z}} \quad \text{and} \quad \chi_{j,z} = \chi\left(\frac{y_{t_{j-1}} - z}{h}\right).$$

Through the ergodic theorem for any $S \in \mathcal{S}_{st}$

$$\sum_{j=1}^{N} \chi_{j,z} \quad \sim \quad 2\mathbf{q}_S(z)\, N\, h \quad \text{as} \quad N \to \infty\,, \quad h \to 0\,,$$

where

$$\mathbf{q}_S(z) = \left(\int_{\mathbb{R}} e^{\widetilde{S}(v)} \mathrm{d}v\right)^{-1} e^{\widetilde{S}(z)} \quad \text{and} \quad \widetilde{S}(x) = 2\int_0^x S(v)\mathrm{d}v\,.$$

ATTENTION! The density $\mathbf{q}_S(z)$ is unknown.

Key idea: Using the method proposed by Borisov and Konev (1977) for the parametric model we correct the last term in the sequential version, i.e we set

$$\widehat{S}_\tau(z) = \frac{\sum_{j=N_0+1}^{\tau-1} \chi_{j,z}\left(y_{t_j} - y_{t_{j-1}}\right) + \sqrt{\kappa}\chi_{\tau,z}\left(y_{t_\tau} - y_{t_{\tau-1}}\right)}{\delta H},$$

where

$$\chi_{j,z} = \chi\left(\frac{y_{t_{j-1}} - z}{h}\right) \quad \text{and} \quad \tau = \tau(z) = \inf\left\{k \geq 1 \,:\, \sum_{j=N_0+1}^{k} \chi_{j,z} \geq H\right\}.$$

The correcting coefficient

$$\sum_{j=N_0+1}^{\tau-1} \chi_{j,z} + \kappa\chi_{\tau,z} = H\,.$$

N_0 will be specified later.

Decomposition

One gets that

$$\widehat{S}_\tau(z) = S(z) + B(z) + \frac{1}{\sqrt{\delta H}}\,\eta(z)\,,$$

where

$$\eta(z) = \frac{\sum_{j=N_0+1}^{\tau-1} \chi_{j,z}\,\xi_j + \sqrt{\kappa}\chi_{\tau,z}\,\xi_\tau}{\sqrt{\delta\, H}} \quad \text{and} \quad \xi_j = W_{t_j} - W_{t_{j-1}}\,.$$

For any z and positive H and $\mathcal{F}_{N_0}$ measurable

$$\eta(z) \quad \sim \quad \mathcal{N}(0,1)\,.$$

Sequential version

Asymptotic behavior

$$\tau \sim \frac{H}{2q_S(z)h} \quad \text{as} \quad H \to \infty\,.$$

Therefore, to use all N observation one needs to take

$$H = 2q_S(z)hN\,.$$

We use the first N_0 observations to estimate the density. To this end we set

$$N_0 = [N^\gamma] \quad \text{for} \quad 3/4 < \gamma < 1\,.$$

To estimate the density $\mathbf{q}_S$ we will make use of the following kernel estimator

$$\widehat{\mathbf{q}}(z) = \frac{1}{2\,N_0\,h_0} \sum_{j=1}^{N_0} \chi\left(\frac{y_{t_j} - z}{h_0}\right)\,, \quad h_0(T) = \frac{1}{\sqrt{T_0}} \quad \text{and} \quad T_0 = \delta N_0\,.$$

Proposition

For any $a > 0$

$$\lim_{T\to\infty} T^a \max_{\mathbf{x}_0 \le z \le \mathbf{x}_1} \sup_{S\in\mathcal{S}} \mathbf{P}_S \left(|\widehat{\mathbf{q}}(z) - \mathbf{q}_S(z)| > v_T\right) = 0\,,$$

where v_T *is some positive slowly decreasing to zero.*

Therefore, If we choose the threshold H in the following form

$$H(z) = 2\widehat{\mathbf{q}}(z) h \left(N - N_0\right).$$

For any $a > 0$

$$\lim_{T\to\infty} T^a \max_{\mathbf{x}_0 \le z \le \mathbf{x}_1} \sup_{S\in\mathcal{S}} \mathbf{P}_S \left(\tau(z) \ge N\right) = 0\,.$$

Passage to the regression model

To estimate in the quadratic metric we use the partition $(z_l)_{1\le l\le n}$

$$z_l = \mathbf{x}_0 + \frac{l}{n}(\mathbf{x}_1 - \mathbf{x}_0) \quad \text{and} \quad n = [\sqrt{T}]\,.$$

For each point z_l we use the truncated sequential procedure

$$\left(\tau_l\,, \widehat{S}_{\tau_l}(z_l)\right) \quad \text{and} \quad h = \frac{\mathbf{x}_1 - \mathbf{x}_0}{2n}\,,$$

where $\tau_l = \tau(H_l)$ and

$$H_l = H(z_l) = 2\widehat{\mathbf{q}}(z_l)h\,(N - N_0)\,.$$

Passage to the nonparametric regression model

Setting

$$Y_l = \widehat{S}_{\tau_l}(z_l)\,\chi_{\{\tau_l<N\}}$$

we obtain that on the set $\Gamma = \cap_{l=1}^{n}\{\tau_l < N\}$ the heteroscedastic regression model

$$Y_l = S(z_l) + \zeta_l = S(z_l) + B_l + \sigma_l\eta_l\,, \quad 1 \le l \le n\,,$$

where

$$\eta_l = \eta(z_l) \quad \text{and} \quad \sigma_l = \frac{1}{\sqrt{\delta H(z_l)}}\,.$$

Here, $(\eta_l)_{1\le l\le n}$ are i.i.d. $\mathcal{N}(0,1)$ random variables.

We use the model selection procedure proposed by Galtchouk and Pergamenchtchikov (2009) for heteroscedastic regression.
Let $(\phi_j)_{1\le j\le n}$ be an orthonormal basis, i.e.

$$(\phi_i\ ,\ \phi_j)_n = \frac{\mathbf{x}_1 - \mathbf{x}_0}{n}\sum_{l=1}^{n}\phi_i(z_l)\phi_j(z_l) = \mathbf{1}_{\{i=j\}}\ .$$

For example, we can take the trigonometric basis. For this basis we obtain

$$S(z_l) = \sum_{j=1}^{n}\theta_j\,\phi_j(z_l) \quad \text{and} \quad \theta_j = \Big(S, \phi_j\Big)_n\ .$$

To estimate S we have to estimate the Fourrier coefficients $(\theta_j)_{1\le j\le n}$.

$$\widehat{\theta}_j = \big(Y, \phi_j\big)_n = \frac{\mathbf{x}_1 - \mathbf{x}_0}{n}\sum_{l=1}^{n} Y_l\phi_j(z_l)\,,$$

where $Y = (Y_1, \ldots, Y_n)$.

Model selection procedure

We estimate the function S on the sieve by the weighted least squares estimator

$$\widehat{S}_\lambda(z_l) = \sum_{j=1}^{n} \lambda(j)\,\widehat{\theta}_j\,\phi_j(z_l)\,\mathbf{1}_\Gamma\,, \quad 1 \le l \le n\,,$$

where the weight vector $\lambda = (\lambda(1), \ldots, \lambda(n))'$ belongs to some finite set $\Lambda \subset [0,1]^n$, the prime denotes the transposition. For any $\mathbf{x}_0 \le t \le \mathbf{x}_1$

$$\widehat{S}_\lambda(t) = \sum_{l=1}^{n} \widehat{S}_\lambda(z_l)\mathbf{1}_{\{z_{l-1}<t\le z_l\}}\,.$$

Minimizing the cost function

$$\widehat{\lambda} = \operatorname{argmin}_{\lambda\in\lambda} J_n(\lambda)\,,$$

yields the model selection procedure

$$\widehat{S}_* = \widehat{S}_{\widehat{\lambda}}\,.$$

Theorem

For any $S \in \mathcal{S}$ *and* $0 < \rho \le 1/12$

$$\mathcal{R}_S(\widehat{S}_*) \le (1 + G(\rho)) \min_{\lambda \in \Lambda} \mathcal{R}_S(\widehat{S}_\lambda) + \frac{\Psi_S(T)}{\rho\, T},$$

where $G(\rho) \to 0$ *as* $\rho \to 0$ *and for any* $\varepsilon > 0$

$$\lim_{T \to \infty} \sup_{S \in \mathcal{S}} \frac{\Psi_S(T)}{T^{\varepsilon}} = 0\,.$$

We chose the parameter ρ as slowly decreasing function of T, i.e. for example, $\rho(T) = (12 + \ln T)^{-1}$.

For the Sobolev ball $\mathcal{S} = \mathcal{W}_{r,k}$ the optimal rate is T^{k_*} and

$$k_* = \frac{2k}{2k+1}\,.$$

Efficiency

We set now the weighted risk as

$$\mathcal{R}_*(\widehat{S}) = \sup_{S\in\mathcal{W}_{r,k}} \upsilon_S\, \mathcal{R}_S(\widehat{S})\,, \quad \upsilon_S = \left(\int_{\mathbf{x}_0}^{\mathbf{x}_1} \frac{1}{\mathbf{q}_S(z)}\,\mathrm{d}z\right)^{-k_*}.$$

The oracle inequality implies

$$T^{k_*}\,\mathcal{R}_*(\widehat{S}_*) \le\ (1+\mathsf{o}(1))\ T^{k_*}\ \min_{\lambda\in\Lambda} \mathcal{R}_*(\widehat{S}_\lambda) + \mathsf{o}(1)\,, \quad T\to\infty\,.$$

Let now $\widehat{S}_{\lambda_0}$ be Pinsker estimator from $\left(\widehat{S}_\lambda\right)_{\lambda\in\Lambda}$. Then

$$T^{k_*}\,\mathcal{R}_*(\widehat{S}_*) \le\ (1+\mathsf{o}(1))\ T^{k_*}\,\mathcal{R}_*(\widehat{S}_{\lambda_0}) + \mathsf{o}(1)\,.$$

Therefore,

$$\lim_{T\to\infty} \frac{\mathcal{R}_*(\widehat{S}_*)}{\mathcal{R}_*(\widehat{S}_{\lambda_0})} \le 1\,.$$

Efficiency properties

Galtchouk and Pergamenchtchikov (2011) shown that

$$\lim_{T\to\infty} T^{k_*} \inf_{\widetilde{S}_T} \mathcal{R}_*(\widetilde{S}_T) = \lim_{T\to\infty} T^{k_*} \mathcal{R}_*(\widehat{S}_{\lambda_0}) = l_* \,,$$

where

$$l_* = (2k+1) r^{1-k_*} \left(\frac{k}{\pi(k+1)(2k+1)} \right)^{k_*}$$

and, therefore,

$$\lim_{T\to\infty} \frac{\inf_{\widetilde{S}_T} \mathcal{R}_*(\widetilde{S}_T)}{\mathcal{R}_*(\widehat{S}_{\lambda_0})} = 1 \,.$$

Therefore, the model select procedure is efficient, i.e.

$$\lim_{T\to\infty} \frac{\inf_{\widetilde{S}_T} \mathcal{R}_*(\widetilde{S}_T)}{\mathcal{R}_*(\widehat{S}_*)} = 1 \,.$$

It should be emphasized, that this efficiency property is obtained without use the regularity parameters r and k, i.e. in adaptive setting.

What's new?

- By making use of the sequential point wise procedure proposed by Galtchouk and Pergamenshchikov (2015) we passe to the heteroscedastic regression model.
- By making use the approach proposed by Galtchouk and Pergamenshchikov (2009) we construct the model selection procedure based on the weight least square estimators for any orthonormal basis in $\mathbf{L}_2[0,1]$, not only trigonometric.
- The Pinsker constant for this model is found in explicit form.
- We obtained the sharp non asymptotic oracle inequalities for the simple empiric risks and for the robust risks as well.
- ONLY THE SEQUENTIAL APPROACH ALLOWS TO CONSTRUCT the efficient estimator!

References

- ARATO, M., KOLMOGOROV, A. N., SINAI Ya. G. (1962) Evaluation of the parameters of a complex stationary Gauss–Markov process. *Doklady Akademii Nauk SSSR*, **146** (4), 747–750.
- DE GREGORIO, A., IACUS, S.M. (2012) Adaptive LASSO-type estimation for multivariate diffusion processes. *Econom. Theory*, **28** (4), 838 -860.
- FUJIMORI, K. (2019) The Danzing selector for a linear model of diffusion processes. - *Statistical Inference for Stochastic Processes*, **22**, 475 - 498.
- DALALYAN, A.S. and KUTOYANTS, Yu. A. (2002) Asymptotically efficient trend coefficient estimation for ergodic diffusion. *Math Methods Stat*, **11** (4):402–427
- KUTOYANTS, Yu.A. (2003) *Statistical Inferences for Ergodic Diffusion Processes.* Springer, Berlin.

References

- GALTCHOUK, L. AND PERGAMENSHCHIKOV, S. (2011) Adaptive sequential estimation for ergodic diffusion processes in quadratic metric. *Journal of Nonparametric Statistics*, **23**(2), 255-285.
- GALTCHOUK, L. AND PERGAMENSHCHIKOV, S. (2022) Adaptive efficient analysis for big data ergodic diffusion models. *Statistical Inference for Stochastic Processes* **25**, 127–158
- LIPTSER, R. S. and SHIRYAEV, A. N. (2001) *Statistics of Random Processes II* Springer-Verlag Berlin Heidelberg, Chapter 17
- NOVIKOV, A.A. (1971) Sequential estimation of the parameters of the diffusion processes, *Theory Probab. Appl.*, **16** , 394–396.
- PINSKER, M.S. (1981) Optimal filtration of square integrable signals in gaussian noise. *Problems Inform. Transmission*, **17**, 120-133.

THANK YOU VERY MUCH

FOR YOUR ATTENTION !!!

CONGRATULATIONS!

DEAR **YURY**,

HAPPY BIRTHDAY

AND

WISHING YOU A LONG LIFE !!!