Nonparametric estimator of the distribution of fitness effects of new mutations

Guillaume Garnier
Under the supervision of Marie Doumic, Marc Hoffmann and Lydia Robert

26/09/2023

Introduction

- All organisms are subject to mutations
- These new traits can change the selective value (fitness) of an individual
- *Fitness* : ability of an individual with a certain genome to survive and reproduce
- How these mutations affect selective value is a central question in evolutionary biology

- The density of the distribution of these effects is called the Distribution of Fitness Effect (DFE)

Introduction

Why study the DFE ?

- DFE is important of these arising mutations define the range of possible evolutionary trajectories a population can follow
- *Study the effects of new mutations in an individual to see if they are beneficial or not*
- *Understanding and quantifying the genetic diversity of human diseases and its future evolution*
- *Predict the consequences of maintaining a small population of animals or plants, as in captive breeding programs*

L'expérimentation

Goal :

Inferring DFE from experimental measurements of selective value over time

What data ?

Two experimental protocols (Robert et al. 2018 [ROR+18])

- See in real time the appearance of mutations in e.coli
- New measurements of cell fitness

$\Longrightarrow$ New data to estimate the DFE

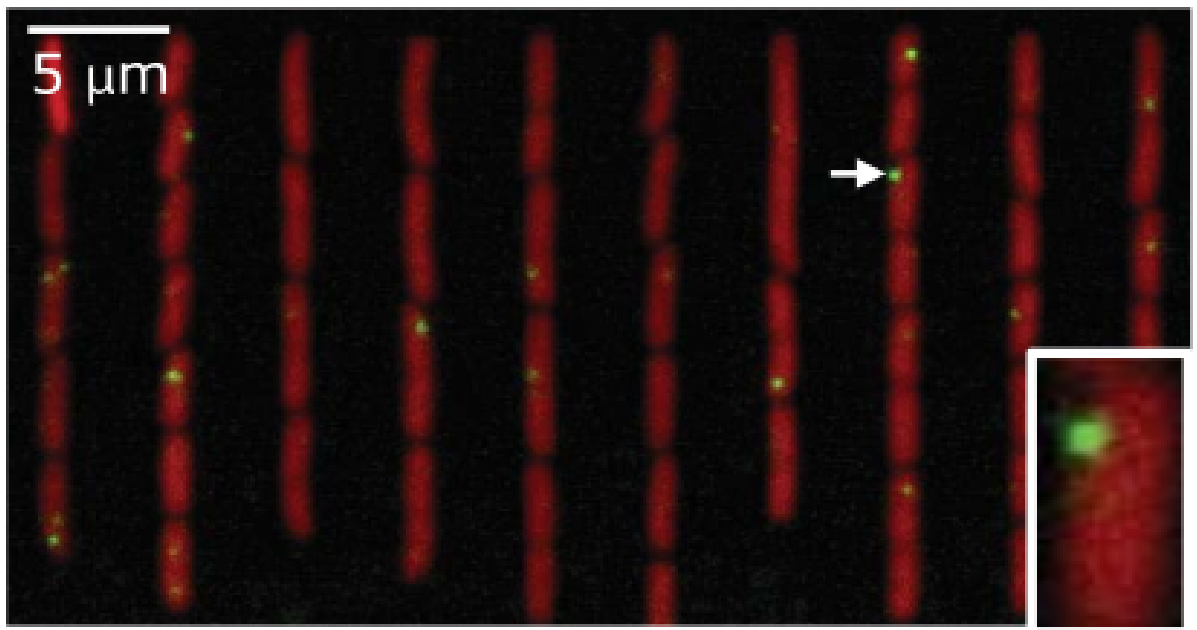
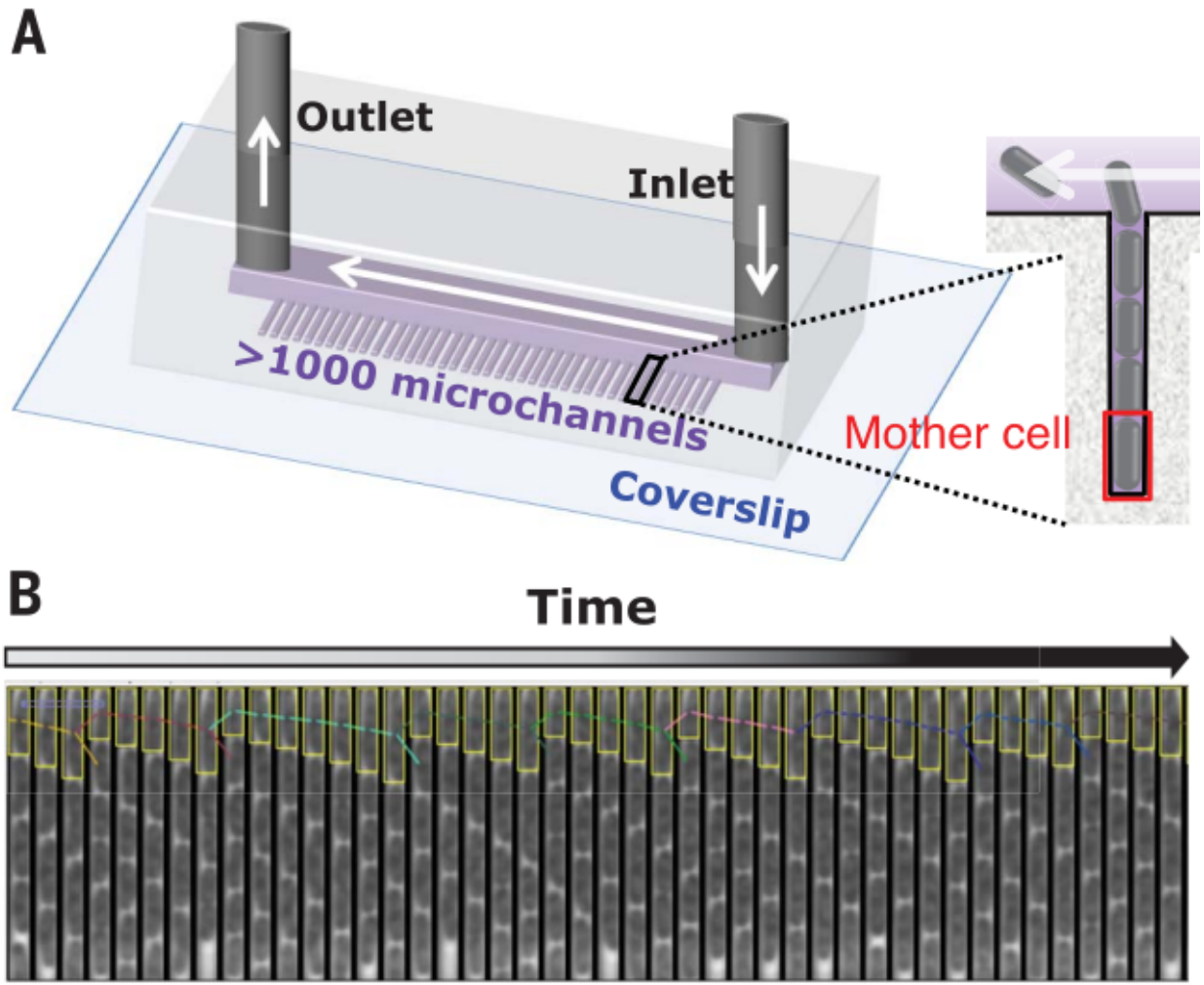


FIGURE – L. Robert and al, Science, 2018

microfluidic Mutation Accumulation (μMA) experiment

- Measuring the fitness of cells
- 1476 parallel and independent channels



Model Building

- A first model. [ROR+18]
- The mutations are deleterious and appear according to a Poisson process $\mathcal{P}(\lambda t)$
- $(W_t)_{t\in\mathbb{R}^+}$ the selective value over time of an individual

$$\boxed{s_i = \frac{W_{t_{i-1}} - W_{t_i}}{W_{t_{i-1}}}\ ,\ i > 0}\ ,$$

  s_i effect of the $\{i\}$–i-th mutation on the fitness of the individual.

- If $(s_i)_i$ are i.i.d

$$\boxed{\frac{W_t}{W_0} = \prod_{i=1}^{N_t}(1 - s_i)\ ,\ N_t \sim \mathcal{P}(\lambda t)}$$

- DFE = probability density of s_i

Model Building

- By taking the logarithm, we have

$$\ln W_t = \sum_{i=1}^{N_t} \ln(1 - s_i)\,, \quad N_t \sim \mathcal{P}(\lambda t)\,, \quad \lambda > 0$$

- It is a compound Poisson process : $X_i \sim \ln(1 - s_i)$ et $Y_t \sim \ln W_t$,

$$\boxed{Y_t = \sum_{i=1}^{N_t} X_i}\,.$$

Model Building

- We want to model the errors in the measurements

$$\frac{W_t}{W_0} = \prod_{i=1}^{N_t} (1 - s_i)\varepsilon_t \,, \quad N_t \sim \mathcal{P}(\lambda t) \,, \quad \lambda > 0 \,,$$

- By taking the logarithm, we have (8). Dans ce cas on a

$$\boxed{Z_t := Y_t + \xi_t = \sum_{i=1}^{N_t} X_i + \xi_t} \,,$$

Model Building

1. Z_t^j : noisy measure in channel $j \in J$ at time t.
2. N_t^j : number of mutation in channel j. $(N_j(t), j \geq 1)$ are *i.i.d* Poisson processes with intensity $\lambda \in (0, \infty)$.
3. X_k^j jump of k-th mutation in channel j. $(X_i^j)_{i,j \geq 0}$ are *i.i.d* with density $f \in L^1(\mathbb{R}) \cap L^2(\mathbb{R})$.
4. ε_t^j represents the measurement noise at time t for channel j. $(\varepsilon_t^j)_{j \geq 0}$ are *i.i.d* and that $\mathbb{E}(\varepsilon_t^j) = 0$.

We consider a noisy compound Poisson process :

$$\boxed{Z_t^j = \left(\sum_{k=1}^{N_t^j} X_k^j \right) + \varepsilon_t^j, \, t \geq 0}.$$

Estimate the DFE

In each model, we want to estimate the probability density of X_i from the observations

Strategy, Tools & Methods

Strategy : We want to estimate the characteristic function of X :

(heuristic) If $\varphi_X(\xi) \simeq \widehat{\varphi}_X(\xi)$, then $f(x) \simeq \widehat{f}(x)$

Indeed, the characteristic function $\varphi_X \to$ Density f of X :

$$f(x) = \frac{1}{2\pi} \int_{\mathbb{R}} \varphi_X(\xi) e^{-ix\xi} d\xi$$

Building the estimator

We consider a noisy compound Poisson process :

$$\boxed{Z_t^j = \left(\sum_{k=1}^{N_t^j} X_k^j \right) + \varepsilon_t^j \, , \, t \geq 0 \, .}$$

For a single channel Z_t^j, the characteristic function is :

$$\boxed{\forall u \in \mathbb{R}, \varphi_{Z_t^j}(u) = e^{-\lambda t + \lambda t \varphi_X(u)} \cdot \varphi_\varepsilon(u)}$$

Building the estimator

Consider two different times $0 < t_1 < t_2$, then

$$\frac{\varphi_{Z_{t_2}}}{\varphi_{Z_{t_1}}} = e^{-\lambda(t_2-t_1)+\lambda(t_2-t_1)\varphi_X(u)}$$

Then

$$\varphi_X(u) = 1 + \frac{1}{t_2 - t_1}\Big(\log \varphi_{Z_{t_2}}(u) - \log \varphi_{Z_{t_1}}(u)\Big)$$

Building the estimator

Consider two different times $0 < t_1 < t_2$, then

$$\frac{\varphi_{Z_{t_2}}}{\varphi_{Z_{t_1}}} = e^{-\lambda(t_2-t_1)+\lambda(t_2-t_1)\varphi_X(u)}$$

Then

$$\varphi_X(u) = 1 + \frac{1}{t_2 - t_1}\Big(\log \varphi_{Z_{t_2}}(u) - \log \varphi_{Z_{t_1}}(u)\Big)$$

It leads us to define

$$\widehat{\varphi}^J_X(u) = 1 + \frac{1}{t_2 - t_1}\Big(\log \widehat{\varphi}^J_{Z_{t_2}}(u) - \log \widehat{\varphi}^J_{Z_{t_1}}(u)\Big)$$

with

$$\widehat{\varphi}^{'J}_{Z_\tau}(u) = \frac{1}{J}\sum_{j=1}^{J} i Z_\tau^j e^{iuZ_\tau^j}, \widehat{\varphi}^J_{Z_\tau}(u) = \frac{1}{J}\sum_{j=1}^{J} e^{iuZ_\tau^j},$$

$$\log \widehat{\varphi}^J_{Z_\tau}(u) = \int_0^u \frac{\widehat{\varphi}^{'J}_{Z_\tau}(z)}{\widehat{\varphi}^J_{Z_\tau}(z)} dz$$

Building the estimator

As there is no guarantee that the previous quantities will not explode, a cut-off is added to ensure this.

$$\widetilde{\varphi}_X^J(u) = 1 + \frac{1}{t_2 - t_1}\Big\{ \log \widehat{\varphi}_{Z_{t_2}}^J(u) \cdot \mathbf{1}_{|\log \widehat{\varphi}_{Z_{t_2}}^J(u)| \leq \ln(J)} - \log \widehat{\varphi}_{Z_{t_1}}^J(u) \cdot \mathbf{1}_{|\log \widehat{\varphi}_{Z_{t_1}}^J(u)| \leq \ln(J)} \Big\}$$

We estimate f by Fourier inversion.
For any $m \in (0, \infty)$,

$$\widehat{f}_{m,J}(x) = \frac{1}{2\pi} \int_{-m}^{m} e^{-iux} \widetilde{\varphi}_X^J(u) du \, , \, x \in \mathbb{R}$$

Building the estimator

As there is no guarantee that the previous quantities will not explode, a cut-off is added to ensure this.

$$\widetilde{\varphi}_X^J(u) = 1 + \frac{1}{t_2 - t_1}\Big\{ \log \widehat{\varphi}_{Z_{t_2}}^J(u) \cdot \mathbf{1}_{|\log \widehat{\varphi}_{Z_{t_2}}^J(u)| \leq \ln(J)} - \log \widehat{\varphi}_{Z_{t_1}}^J(u) \cdot \mathbf{1}_{|\log \widehat{\varphi}_{Z_{t_1}}^J(u)| \leq \ln(J)} \Big\}$$

We estimate f by Fourier inversion.
For any $m \in (0, \infty)$,

$$\widehat{f}_{m,J}(x) = \frac{1}{2\pi} \int_{-m}^{m} e^{-iux} \widetilde{\varphi}_X^J(u) du \, , \, x \in \mathbb{R}$$

Here, the choice of m is very important because it defines the frequencies that we keep to apply the inverse Fourier transformation

Theorem : convergence of the estimator

For all reals $0 < t_1 < t_2$ such that $t_2 \leq \frac{1}{4}\log(Jt_2)$
$Jt_1 \to \infty, Jt_2 \to \infty$ as $J \to \infty$ and for any $m < C^J_{t_1,t_2}$, the following inequality holds

$$\mathbb{E}\left(\|\widehat{f}_{m,J} - f\|^2\right) \leq \|f_m - f\|^2 + \sum_{i=1}^{2} \frac{4e^{4t_i}}{J(t_2-t_1)^2} \int_{-m}^{m} \frac{du}{|\varphi_\varepsilon(u)|^2} + \frac{4K_{J,t_1,t_2}}{(t_2-t_1)^2} \cdot \left(\frac{\mathbb{E}[X_i^2]}{Jt_i} + \frac{\mathbb{E}[\varepsilon^2]}{Jt_i^2} + 4\frac{m}{(Jt_i)^2} \right).$$

where K_{J,t_1,t_2} and $C^J_{t_1,t_2}$ depends on m, t_1, t_2 and $\log\varphi_\varepsilon(\cdot)$.

Theorem : adaptative estimator

Question : How to select m ?

- The dominant terms :
  biais term : $\int_{u\in[-m,m]} |\varphi_X(u)|^2 \,\mathrm{d}u$
  variance term : $\frac{4e^{4t_2}}{J(t_2-t_1)^2} \int_{-m}^{m} \frac{\mathrm{d}u}{|\varphi_\varepsilon(u)|^2}$
- Through differentation, the optimal $\overline{m_J}$ satisfies

$$|\varphi_X(\overline{m_J})|^2 = \frac{4ae^{4t_2}}{J(t_2-t_1)^2}(1+\overline{m_J}^2)\,.$$

then

$$\boxed{\left|\frac{\varphi_X(\overline{m_J})}{\sqrt{(1+\overline{m_J}^2)}}\right|^2 = \frac{4ae^{4t_2}}{J(t_2-t_1)^2}}\,.$$

Theorem : adaptive estimator

It leads us to define the empirical cutoff parameter

$$\widehat{m}_J = \max\left\{u \geq 0 : \left|\frac{\overline{\varphi}_X(u)}{\sqrt{1+u^2}}\right| \geq \frac{\kappa_{J,t_1,t_2}}{\sqrt{J}(t_2-t_1)}\right\} \wedge \left(J(t_2-t_1)^2\right)^{\alpha}, \quad \alpha \in (0,1)$$

where

$$\overline{\varphi}_X^J(u) = \widetilde{\varphi}_X^J(u) \cdot \mathbb{1}\left|\frac{\widetilde{\varphi}_X^J(u)}{\sqrt{1+u^2}}\right| \geq \frac{\kappa_{J,t_1,t_2}}{\sqrt{J}(t_2-t_1)}$$

and $\kappa_J = 2e^{2t_2} + \kappa\sqrt{\ln(J(t_2-t_1)^2)}$, $\kappa > 0$

Theorem : adaptive estimator

For all reals $0 < t_1 < t_2$ such that $t_2 \leq \frac{1}{4}\log(Jt_2)$ and $(m)^\alpha < C^J_{t_1,t_2}$, $Jt_1 \to \infty$, $Jt_2 \to \infty$ as $J \to \infty$. Then,

$$\mathbb{E}\Big[\|\overline{f}_{\widehat{m}_J} - f\|^2\Big] \leq \inf_{m\in[0,m^\alpha_m]}\left\{\|f_m - f\|^2 + C\frac{\ln(J(t_2-t_1)^2)\cdot m\cdot(1+m^2)}{J(t_2-t_1)^2} + \widetilde{C}A\right\} + \left(2+\frac{2\log(J)}{(t_2-t_1)}\right)^2\cdot T_J$$

where

$$A = \sum_{i=1}^{2}\frac{4e^{4t_i}}{J(t_2-t_1)^2}\int_{-m}^{m}\frac{\mathrm{d}u}{|\varphi_\varepsilon(u)|^2} + \frac{4K_{J,t_1,t_2}}{(t_2-t_1)^2}\cdot\left(\frac{\mathbb{E}[X_i^2]}{Jt_i} + \frac{\mathbb{E}[\varepsilon^2]}{rJt_i^2} + 4\frac{m}{(Jt_i)^2}\right)$$

and

$$T_J \leq C_0(J(t_2-t_1)^2)^{\alpha - c(\theta)^2} + \frac{C_1}{J(t_2-t_1)^2} + \frac{C_2}{J(t_2-t_1)^4} \qquad (1)$$

and $c(\theta) = \kappa(t_2-t_1)e^{2t_2}\cdot\frac{d}{\sqrt{1+(m_*)^2}}$ and where C_0, C_1 and C_2 depends on $\mathbb{E}[X_1^2], \mathbb{E}[\varepsilon^2]$ and where C and $\widetilde{C}$ are two constants.

Numerical Result

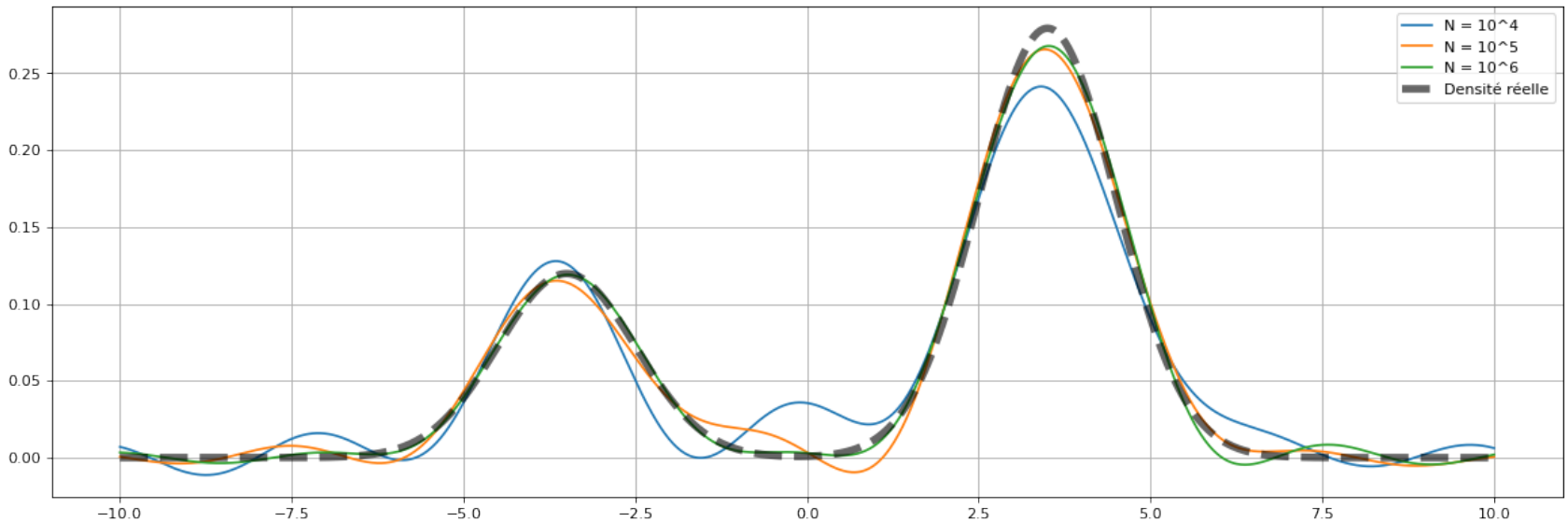


FIGURE – Reconstruction of the $0.3\mathcal{N}(-3.5,1)+0.7\mathcal{N}(3.5,1)$ distribution with J channels, corrupted by a Gaussian noise $\mathcal{J}(0,1)$ with $J \in 10^4, 10^5, 10^6$. $t_1 = 0.1, t_2 = 1, m = 2$

Numerical Result

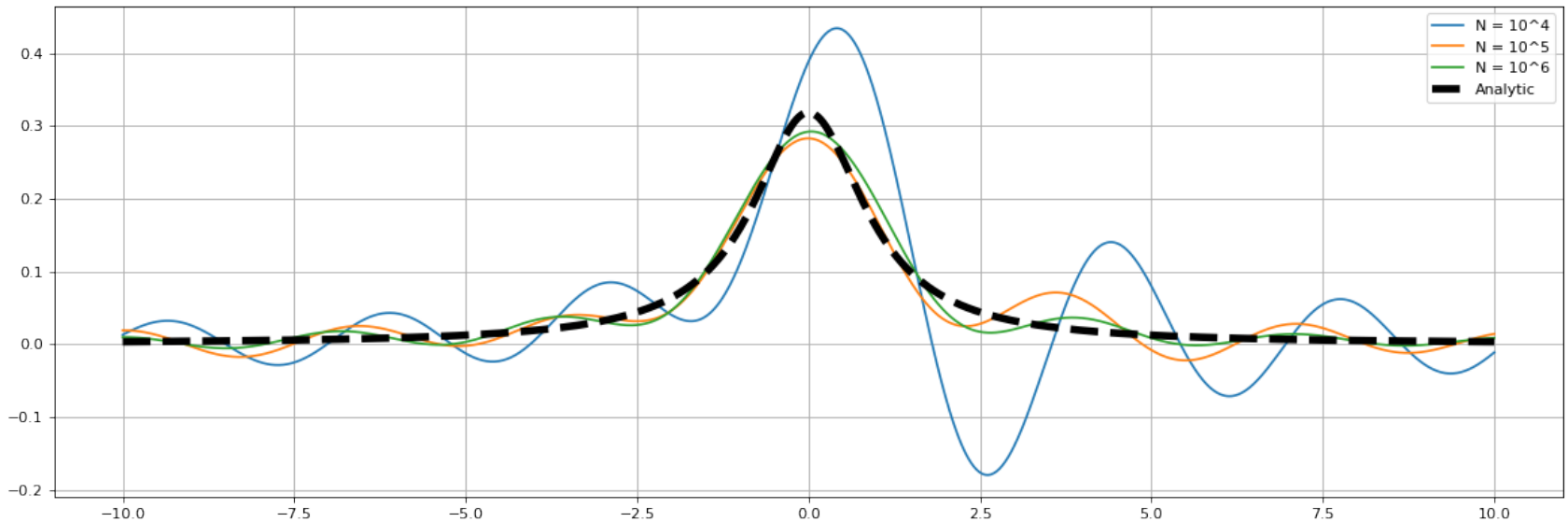


FIGURE – Reconstruction of the Cauchy $\mathcal{C}(0,1)$ distribution with J channels, corrupted by a Gaussian noise $\mathcal{N}(0,1)$ with $J \in 10^4, 10^5, 10^6$. $t_1 = 0.1, t_2 = 1, m = 2$

Numerical Result

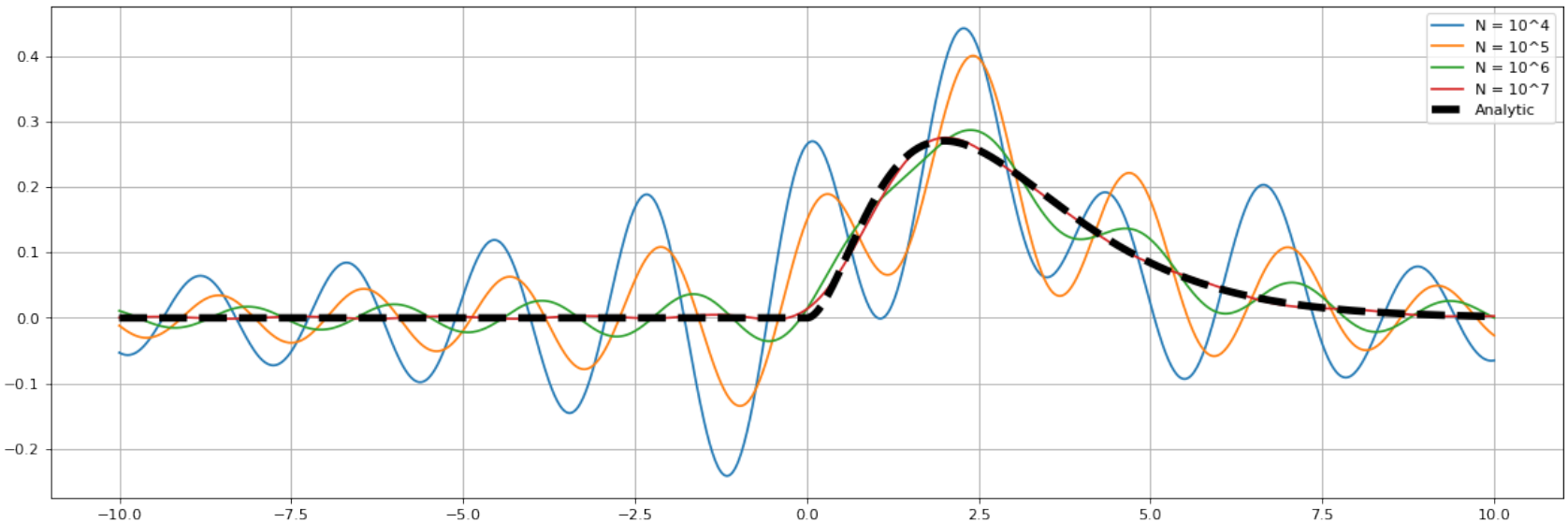


FIGURE – Reconstruction of the Gamma $\Gamma(3)$ distribution with J channels, corrupted by a Gaussian noise $\mathcal{J}(0,1)$ with $J \in 10^4, 10^5, 10^6, 10^7$. $t_1 = 0.1, t_2 = 1, m = 3$

Perspective

- Apply the numerical methods on experimental data.
- Is this estimator minimax ?
  i.e Is it the best estimator among all possible estimators ?

References I

Lydia Robert, Jean Ollion, Jérôme Robert, Xiaohu Song, Ivan Matic, and Marina Elez, *Mutation dynamics and fitness effects followed in single cells*, Science **359** (2018), no. 6381, 1283–1286.