Nonparametric estimator of the distribution of fitness effects of new mutations

Guillaume Garnier, Ph.D Student (Inria / Sorbonne Université)

Advisors : Marie Doumic, Marc Hoffmann, Lydia Robert

Introduction and Data

- All organisms are subject to mutations
- These new traits can change the selective value (fitness) of an individual
  We call *Fitness* the ability of an individual with a certain genome to survive and reproduce
- **How these mutations affect the 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)**

- **Probabilistic Model** :
  1. Z_t^j represents the noisy measure of the fitness of the cell in channel $j \in J$ at time t.
  2. N_t^j represents the number of times the cell in channel j has mutated. $(N_j(t), j \geq 1)$ are *i.i.d* Poisson processes with intensity $\lambda \in (0, \infty)$.
  3. X_k^j represents the effect of the k-th mutation on the cell 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:

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

Statement of the problem: Estimate the density of X_i from observations of Z_t on each channel $j \in J$

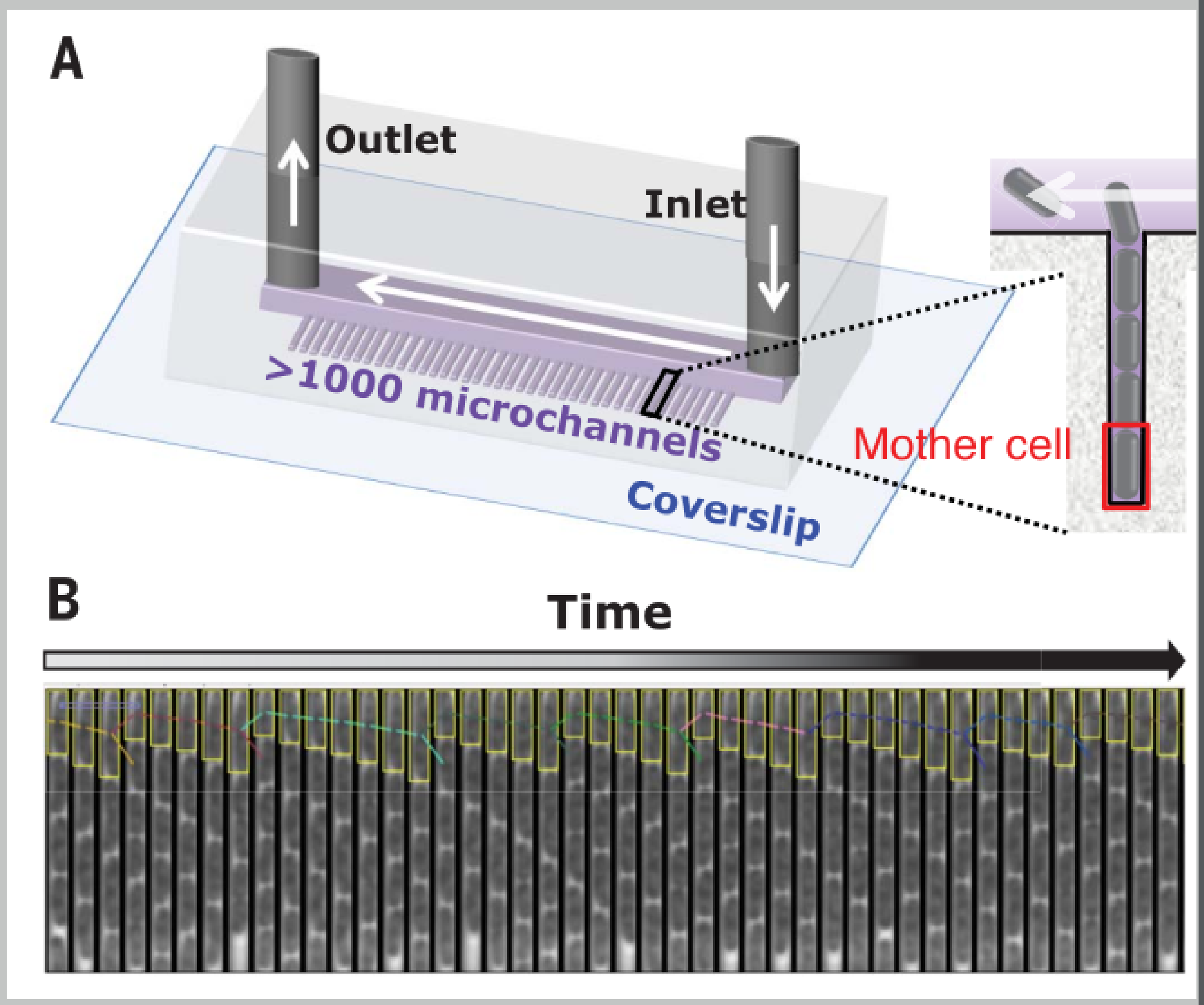

Figure 1:Measurement of the evolution of the fitness of several cell lines over time

Robert et al., 2018

Combine two classical problems in non-parametric inference.

- Deconvolution
- Decompounding

Statistical Strategy and statistical Results

- **Strategy:** 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 $\varphi_X \longrightarrow$ Density f of X:

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

- **Theorem :** 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 $\log \varphi_\varepsilon(\cdot)$.
Futhermore, **m can be chosen in an optimal way from data**

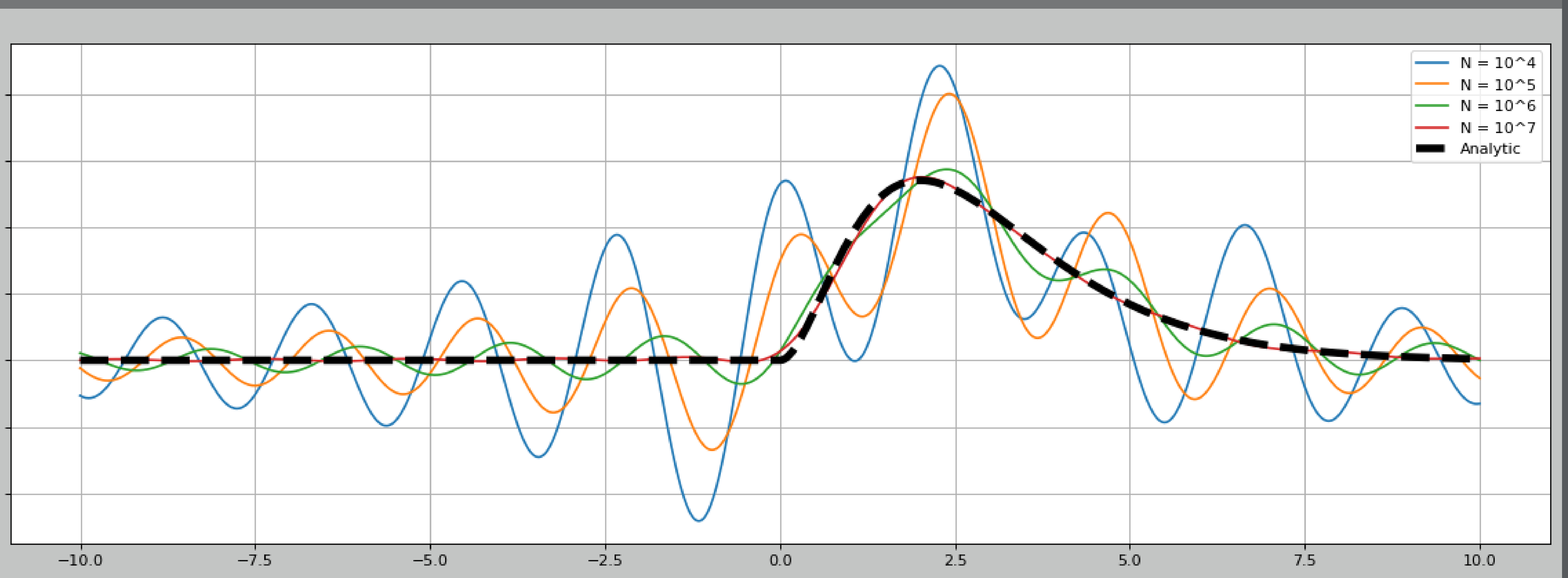

Figure 2:Reconstruction of the Gamma $\Gamma(3)$ distribution with J channels, corrupted by a Gaussian noise $\mathcal{J}(0,1)$ with $J \in 10^4, 10^5, 10^6, 10^7$ and $t_1 = 0.1, t_2 = 1, m = 3$

The estimator converges to f when $J \to \infty$.

- Perspectives:
  1. Is this estimator minimax? (*i.e the "best" estimator among all estimators*)
  2. Use several times ?

Building the estimator

- We write the characteristic function of the process on a single channel Z_t^j. For all $t \in \mathbb{R}_+$, we have

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

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

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

- It leads us to define

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

$$\widehat{\varphi}'^J_{Z_\tau}(u) = \frac{1}{J} \sum_{j=1}^{J} i Z_\tau^j e^{iuZ_\tau^j} \, , \, \widehat{\varphi}^J_{Z_\tau}(u) = \frac{1}{J} \sum_{j=1}^{J} e^{iuZ_\tau^j} \, , \, \log \widehat{\varphi}^J_{Z_\tau}(u) = \int_0^u \frac{\widehat{\varphi}'^J_{Z_\tau}(z)}{\widehat{\varphi}^J_{Z_\tau}(z)} dz$$

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

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

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

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

References

[1] Lydia Robert, Jean Ollion, Jérôme Robert, and al.
Mutation dynamics and fitness effects followed in single cells.
Science, 359(6381):1283–1286, 2018.

[2] Marie Doumic and Miguel Escobedo.
Time asymptotics for a critical case in fragmentation and growth-fragmentation equations.
Kinetic and Related Models, 9(2):251–297, 2016.

[3] Céline Duval and Johanna Kappus.
Nonparametric adaptive estimation for grouped data.
Journal of Statistical Planning and Inference, 182:12–28, 2017.

[4] Céline Duval and Johanna Kappus.
Adaptive procedure for fourier estimators : application to deconvolution and decompounding.
Electronic Journal of Statistics, 13(2):3424–3452, 2019.