Estimating the division rate from indirect measurements of single cells

Marie Doumic, Adélaïde Olivier, Lydia Robert

Estimating the division rate from indirect measurements of single cells

Marie Doumic* Adélaïde Olivier† Lydia Robert

July 11, 2019

Abstract

Is it possible to estimate the dependence of a growing and dividing population on a given trait in the case where this trait is not directly accessible by experimental measurements, but making use of measurements of another variable? This article adresses this general question for a very recent and popular model describing bacterial growth, the so-called *incremental* or *adder* model. In this model, the division rate depends on the *increment of size* between birth and division, whereas the most accessible trait is the size itself. We prove that estimating the division rate from size measurements is possible, we state a reconstruction formula in a deterministic and then in a statistical setting, and solve numerically the problem on simulated and experimental data. Though this represents a severely ill-posed inverse problem, our numerical results prove to be satisfactory.

Introduction

The field of structured population equations has attracted much interest for more than sixty years, leading to substantial progress in their mathematical understanding. These equations describe a population dynamics in terms of well-chosen traits, which offer a relevant characterization of the individual behaviour. More recently, thanks to considerable progress in experimental measurements, the question of estimating the parameters from single-cell measurements also attracts a growing interest, since it finally allows comparing models model and data, and thus investigating which variable is biologically relevant as a structuring variable - see for instance [30] for the application to age-structured and size-structured models for bacterial growth.

However, the so-called *structuring* variable of the model may be quite abstract ("maturity", "satiety"...), and/or not directly measurable, whereas the quantities that are effectively measured may be linked to the structuring one in an unknown or intricate manner. As an illustration of this idea, we can cite the interesting series of articles by H.T. Banks and co-authors, concerning the estimation of the division rate in data sets where the measured quantity was the fluorescence (carboxyfluorescein succinimidyl ester (CFSE)) of the cells. Initially, they designed a fluorescence-structured model [2], but then the estimated division rates appeared difficult to interpret biologically. Indeed, the fluorescence was artificially added to the cells, thus it was not *structuring*: the

*Sorbonne Université, Inria, Université Paris-Diderot, CNRS, Laboratoire Jacques-Louis Lions, F-75005 Paris, France. Email adress: marie.doumic@inria.fr

†Laboratoire de Mathématiques d'Orsay, Univ. Paris-Sud, CNRS, Université Paris-Saclay, 91405 Orsay, France. Email adress: adelaide.olivier@u-psud.fr

difficulty was to find out *which* variable was really structuring, and how it was related to the measured quantities. This was done successfully by this group by building a model structured in both the true structuring variable - the so-called "cyton model" - and in the label, *i.e.* the measured quantity [3].

From such considerations we can formulate a general question: is it possible to estimate the dependence of a population on a given variable, which is not experimentally measurable, by taking advantage of the measurement of another variable?

In this article, we address this question in a specific setting, namely the growth and division of bacteria. Recently, it was evidenced that for several types of bacteria and yeast cells, the "increment of size", *i.e.* the increase of size of a cell between its birth and its division, provides a better-fitted model than age- or size-structured models [9, 17, 19, 33]. These studies were based on data obtained by time-lapse microscopy and consisting in measurements of single-cells growing and dividing. Such data allows estimating for each cell its lifetime, its size at birth and at division, and its size-evolution through time. We refer to this kind of data as "measurements of dividing cells".

Comparison of models and data, such as performed in the above-mentioned studies, requires time-lapse microscopy data obtained in finely controlled conditions ensuring stable, steady-state growth. In addition, precise image analysis is also required to obtain accurate size measurements of numerous single-cells. Obtaining such data is therefore not straight-forward and can be time-consuming. This can represent an important limitation, for instance for screening strategies where data has to be obtained in many different bacterial strains or experimental conditions. Here we consider the case of data consisting only in instantaneous size measurements of single-cells in a population. Such measurements can be more easily obtained, by microscopy snapshots, or using a flow cytometer or a coulter counter which both allow high-throughput acquisition.

From such data, the question of estimating the division rate in a size-structured model has been studied in a series of papers, in a deterministic [6, 16, 29] or statistical [15] setting. The rates of convergence for the estimates have been proved to correspond to an inverse problem of degree of ill-posedness one, hence worse than the rate of convergence obtained from measurements of dividing cells (corresponding, in a deterministic setting, to a degree of ill-posedness zero, see [14] for a discussion of this heuristics).

In view of the new biological evidence in favour of the incremental model [8, 10, 17, 31, 33], this article is devoted to the same question as in [29] and following articles, but for the incremental model: Can we estimate an increment-dependent division rate from a measurement of the size-distribution of cells? Though formulated in a similar way, this new problem is much more complex, since the observed variable (instantaneous size) is not the structuring variable (size increment).

Let us now give a mathematical definition of the problem under study. First of all, we recall the increment-structured model.

The incremental model for bacterial growth

Let us denote $u(t, a, x)$ the density of cells at time t of size x which have an increment $a = x - y$ between their actual size x and their size at birth y. We denote this increment a as it may be viewed as a kind of age, since it increases monotonically and starts at zero at birth - but an age that would have a link with the size: if $g(x)$ denotes the growth rate of a cell of size x, its "aging rate" is also $g(x)$. We have the following increment-and-size model, as proposed in Taheri-Araghi

et al. [33] for bacteria, and also designed in a different context by Hall *et al.* in [22] :

$$\frac{\partial}{\partial t}u(t,a,x)+\frac{\partial}{\partial a}(g(x)u(t,a,x))+\frac{\partial}{\partial x}(g(x)u(t,a,x))+g(x)B(a,x)u(t,a,x)=0, \quad (1)$$

$$g(x)u(t,0,x)=4\int_0^\infty g(2x)B(a,2x)u(t,a,2x)da, \qquad g(0)u(t,a,0)=0, \qquad u(0,a,x)=u^{in}(a,x). \quad (2)$$

The instantaneous probability to divide is, as in [22], $g(x)B(a,x)$ for a cell of size-increment a and size x. From a modelling point of view, writing the division rate as the product of g and a function B allows us to interpret B as the instantaneous probability to divide *in a unit of growth* instead of a unit of time. This is coherent with the fact that the cell may ignore the "time" and use its growth as a clock. As will be explained below, in the case proposed by Taheri-Araghi *et al.* [33] where $B(a,x)=B(a)$, this is also coherent with a much simpler and more natural underlying piecewise deterministic Markov process (PDMP), where the time of division is a simple renewal process of jump rate $B(a)$, so that the increments of dividing cells are mutually independent and distributed according to the density $f_B(a)=B(a)\exp\big(-\int_0^a B(s)ds\big)$.

Asymptotic behaviour of the incremental model

Under suitable assumptions on the coefficients g and B (for instance the theorem 3.7. of [11] may be adapted and provides a proof for smooth growth and division rates, and most recently [20] studies the case $g(x)=x$, with fairly general division rate B), we have a dominant eigentriplet (λ,U,ϕ) unique solution of

$$\lambda U(a,x)+\frac{\partial}{\partial a}(g(x)U(a,x))+\frac{\partial}{\partial x}(g(x)U(a,x))+g(x)B(a)U(a,x)=0, \quad (3)$$

$$g(x)U(0,x)=4\int_0^\infty g(2x)B(a)U(a,2x)da, \quad g(0)U(a,0)=0, \quad (4)$$

$$\lambda\phi(a,x)-g(x)\frac{\partial}{\partial a}\phi(a,x)-g(x)\frac{\partial}{\partial x}\phi(a,x)+g(x)B(a)\phi(a,x)=2g(x)B(a)\phi(0,\frac{x}{2}), \quad (5)$$

$$\lambda>0, \quad \phi\geq 0, \quad U\geq 0, \qquad \iint U(a,x)dadx=1, \qquad \iint U(a,x)\phi(a,x)dadx=1. \quad (6)$$

Moreover, using for instance the general relative entropy inequalities, we have under some more assumptions (theorem 4.5. in [11])

$$\iint |u(t,a,x)e^{-\lambda t}-\big(\iint u^{in}(a,x)\phi(a,x)dadx\big)U(a,x)|\phi(a,x)dadx\to_{t\to\infty}0.$$

Let us however notice that under the assumption made in [33], namely that $g(x)=x$, this precise asymptotic behaviour fails to happen and a cyclic behaviour is observed, as already proved for the growth-fragmentation equation [5, 21]. This corresponds to an idealised case; in reality, there is always some variability, leading to a growth slightly different from perfectly exponential and

to a division into two slightly unequal parts, see for instance [30]. From a numerical perspective, this slight "imperfection" has to be included, else the cyclic behaviour will perturb the results, see Section 3 for more details.

Mathematical formulation of the inverse problem

From now on, we denote U by U_B to underline the dependence in the unknown division rate. Let us assume that we measure the steady size-distribution, which is modeled by the marginal $U_{B,x}(x) = \int_0^\infty U_B(a,x)da$. Such a measurement may be done for instance via a sample of n cells for which we measure their sizes $(x_1, \cdots, x_n)$, assumed to be the realization of an i.i.d. sample distributed along $U_{B,x}$, in the spirit of Doumic, Hoffmann, Reynaud-Bouret and Rivoirard [15].

We also assume division into two equally-sized daughters, as modelled in (1), and that $g(x)$ and λ are already known from an independent measurement or previous knowledge. Typically, λ may be measured through the time evolution of the total mass, which is classical in biology [27], and under the consensual assumption of exponential growth $g(x) = \tau x$, we have $\lambda = \tau$. In this article, we do not consider the noise in the measurement of λ and g, which could be included in future work.

The problem we want to solve in order to have a fully determined model is:

Given $\big(\lambda, g(\cdot)\big)$ and given measurements of $x \to U_{B,x}(x)$, can we estimate the division rate $a \to B(a)$?

In Section 1, we provide an explicit though intricate formula for the estimation of B from U_B, without taking the noise into account. We then provide a statistical estimator in Section 2, numerically implemented in Section 3 both on simulated data and real data. These numerical results provides us with clues concerning directions for future work, which we comment in the discussion (Section 4).

1 Reconstruction formula in a deterministic setting

Before providing the reconstruction formula for B, let us introduce some useful notation. As standard in the field of renewal processes, we introduce the probability distribution function f_B and the survival function S_B of the increments of dividing cells:

$$f_B(a) := B(a)\exp(-\int_0^a B(s))ds, \qquad S_B(a) := \int_a^\infty f_B(s)ds = e^{-\int_0^a B(s)ds}. \tag{7}$$

Symetrically, we introduce the size-distribution of dividing cells, that we denote $\mathcal{L}_B$

$$\mathcal{L}_B(x) = \int_0^x g(x)B(a)U_B(a,x)da. \tag{8}$$

As shown below, the function $\mathcal{L}_B$ is an important intermediate to formulate B from the measurement of $U_{B,x}$. Though we cannot write it as an explicit function of $U_{B,x}$, it can be obtained in a similar way as the distribution of dividing cells for the size-structured equation, see [16].

Lemma 1. *Let g be a positive continuous function on $(0,\infty)$, $\lambda \geq 0$, and $U_{B,x}$ a positive function on $(0,\infty)$ such that $\lambda U_{B,x} + \frac{d}{dx}(gU_{B,x}) \in L^2(x^p dx)$ with $p \in [0,\infty)\setminus\{3\}$. Then there exists a unique solution $\mathcal{L}_B \in L^2(x^p dx)$ such that*

$$\lambda U_{B,x}(x) + \frac{d}{dx}(gU_{B,x})(x) = 4\mathcal{L}_B(2x) - \mathcal{L}_B(x), \tag{9}$$

and there exists $C_p > 0$ such that

$$\|\mathcal{L}_B\|_{L^2(x^p dx)} \leq C_p \|\lambda U_{B,x} + \frac{d}{dx}(gU_{B,x})\|_{L^2(x^p dx)}.$$

If moreover there exists $U_B \in W^{1,1}((0,\infty)\times(0,\infty))$ solution of the system (3)-(4) such that $U_{B,x} = \int_0^\infty U_B(a,x)da$, then this unique solution coincides with the size-distribution of dividing cells defined by (8).

Proof. The existence, uniqueness and continuity part directly follows from [16] Proposition A.1. If U_B is solution to (3)-(4) , we integrate Equation (3) along the increment a, use the boundary condition (4), and obtain Equation (9) with $\mathcal{L}_B$ defined by (8). ■

With this lemma, we see that the estimation of $\mathcal{L}_B$ from $U_{B,x}$ is an inverse problem of degree of ill-posedness 1 when stated in a space $L^2(x^p dx)$ (degree 3/2 in the framework of a statistical noise [15, 28]), as already known from [29]. Interestingly, we remark that this would remain true also if the division rate would depend on other structuring variables, as soon as the growth rate and the division kernel are known: this allows one to reconstruct the size distribution of dividing cells from the size distribution of all cells, in any framework.

We are now ready to formulate B in terms of $U_{B,x}$, $\mathcal{L}_B$, and the parameters λ and g. This is done in the next proposition.

In all what follows, we denote by f^* the Fourier transform of a function f:

$$f^*(\xi) = \int_{-\infty}^{+\infty} f(x) e^{\mathbf{i}x\xi} dx.$$

Proposition 1. *Let B and g be such that there exists a unique positive eigentriplet (λ, U_B, ϕ_B) solution of the eigenproblem (3)–(6). Let us furthermore assume $\lambda U_{B,x} + \frac{d}{dx}(gU_{B,x}) \in L^2(x^p dx)$ and define $\mathcal{L}_B$ as the unique solution of Equation (9) given by Lemma 1.*

We define f_B and S_B by (7). We define two intermediate functions $\mathcal{N}_B$ and $\mathcal{G}_B$ by

$$\mathcal{G}_B(y) = 4e^{\lambda G(y)}\mathcal{L}_B(2y), \qquad \mathcal{N}_B(y) = g(y)e^{\lambda G(y)}U_{B,x}(y), \tag{10}$$

with $G(x)$ an anti-derivative of $1/g(x)$. We assume that $\mathcal{N}_B^*$ and $\mathcal{G}_B^*$ are the well-defined Fourier transforms of $\mathcal{N}_B$ and $\mathcal{G}_B$.

We have the following reconstruction formula for B in terms of λ, $U_{B,x}$ and g :

$$B(a) = \frac{f_B(a)}{S_B(a)} = \frac{\int_{-\infty}^{+\infty}\left(1 + \mathbf{i}\xi\frac{\mathcal{N}_B^*(\xi)}{\mathcal{G}_B^*(\xi)}\right)e^{-\mathbf{i}a\xi}d\xi}{\int_a^{+\infty}\left(\int_{-\infty}^{+\infty}\left(1 + \mathbf{i}\xi\frac{\mathcal{N}_B^*(\xi)}{\mathcal{G}_B^*(\xi)}\right)e^{-\mathbf{i}s\xi}d\xi\right)ds}, \tag{11}$$

provided that all the inverse Fourier transforms are well defined and that neither $\mathcal{G}_B^*$ nor the denominator vanishes.

Corollary 1. *Under the assumptions of Proposition 1, if $g(x) = \tau x$ we have $\lambda = \tau$, $\mathcal{G}_B(y) = 4y\mathcal{L}_B(2y)$ and $\mathcal{N}_B(y) = \tau y^2 U_{B,x}(y)$.*

Remark 1. *At this stage, the reconstruction formula is formal: to give it a rigorous meaning and ensure its validity, we would have to prove that all the quantities are well-defined, in particular that the Fourier transform $\mathcal{G}_B^*$ never vanishes. This requires a full study* per se*, and is beyond the scope of this work: in another case study, this has been done for instance for the estimation of the fragmentation kernel of the growth-fragmentation equation in the article [13], using the Cauchy integral to prove that a Mellin transform never vanishes, proof adapted to another case in [23]. For these two cases however, the proofs used strongly an explicit formulation of the solution with the use of Mellin or Fourier transforms, thanks to the fact that B was a power law in [13], and constant in [23]. We let it for future work.*

Proof. The aim is to use the classical formula $B(a) = \frac{f_B(a)}{S_B(a)}$, and to find a formulation for f_B in terms of $U_{B,x}$, then express S_B as its integral.

First step. Formulating $U_{B,x}$ in terms of λ, g, $\mathcal{L}_B$ and B.

As done for the study of the eigenvalue problem carried out for instance in [11, 20], we can classically obtain a formulation of $U_B(a,x)$ in terms of $\mathcal{L}_B$. We first write Equation (4) under the form

$$g(x)U_B(0,x) = 4\mathcal{L}_B(2x), \tag{12}$$

and then use the method of characteristics to solve (3) and (12). We define an intermediate function $C(a,x) = g(x)U_B(a,x)$ solution of

$$\frac{\partial}{\partial a}C(a,x) + \frac{\partial}{\partial x}C(a,x) + \big(\frac{\lambda}{g(x)} + B(a)\big)C(a,x) = 0,$$

we define $\tilde{C}(a,x) = C(a,x+a)e^{\int_0^a \left(\frac{\lambda}{g(x+s)}+B(s)\right)ds}$, which satisfies $\frac{\partial}{\partial a}\tilde{C}(a,x) = 0$, so that

$$C(a,x+a)e^{\int_0^a \left(\frac{\lambda}{g(x+s)}+B(s)\right)ds} = C(0,x),$$

which gives for U_B:

$$g(y)U_B(a,y) = g(y-a)U_B(0,y-a)e^{-\int_0^a \left(\frac{\lambda}{g(y-a+s)}+B(s)\right)ds} = 4\mathcal{L}_B\big(2(y-a)\big)e^{-\int_0^a \left(\frac{\lambda}{g(y-a+s)}+B(s)\right)ds}, \tag{13}$$

using (12) for the last equality. We integrate the equation (13) in a and obtain

$$g(y)U_{B,x}(y) = \int_0^y 4\mathcal{L}_B\big(2(y-a)\big)e^{-\int_0^a \left(\frac{\lambda}{g(y-a+s)}+B(s)\right)ds}da.$$

Second step. Formulating two deconvolution problems for S_B and f_B.

Denoting by G an antiderivative of $1/g$, and defining an intermediate function $\mathcal{N}_B$, the previous formula is equivalent to

$$\mathcal{N}_B(y) := g(y)e^{\lambda G(y)}U_{B,x}(y) = 4\int_0^y e^{\lambda G(y-a)}\mathcal{L}_B\big(2(y-a)\big)e^{-\int_0^a B(s)ds}da. \tag{14}$$

We define $\mathcal{G}_B$ by (10), thus Equation (14) is nothing but

$$\mathcal{N}_B(x) = \big[\mathcal{G}_B \star S_B\big](x). \tag{15}$$

This is a deconvolution problem, where S_B is the unknown. If we find estimators of $\mathcal{N}_B$ and $\mathcal{G}_B$, we can reconstruct S_B. Since $f_B = -\frac{d}{da}S_B$, integrating by parts we can transform (15) into a deconvolution problem for f_B.

Third step. Solution of the deconvolution problems by Fourier transform.

Extending all the functions on $\mathbb{R}_-$ by zero (with a slight abuse of notation, we keep the same notation for the function and for its extension), we rewrite (15) in the Fourier domain, and for $\xi \in \mathbb{R}$ such that $\mathcal{G}_B^*(\xi) \neq 0$:

$$\mathcal{N}_B^* = \mathcal{G}_B^* S_B^* \quad \Longrightarrow \; S_B^*(\xi) = \frac{\mathcal{N}_B^*}{\mathcal{G}_B^*}(\xi).$$

Since f_B is a probability density for which we assume continuity around 0 and $f_B(0 = 0$, we can extend it continuously by 0 on $\mathbb{R}_-$, we have $f_B^*(0) = 1$, and and since $f_B = -\frac{dS_B}{da}$ for $a > 0$ we get

$$f_B^* = 1 + \mathbf{i}\xi S_B^* = 1 + \mathbf{i}\xi\frac{\mathcal{N}_B^*(\xi)}{\mathcal{G}_B^*(\xi)},$$

for $\xi \in \mathbb{R}$ such that $\mathcal{G}_B^*(\xi) \neq 0$. The 1 expresses the Fourier transform of the discontinuity of the prolongation of S_B in 0, since $S_B(0^+) = 1$; however, this term should be compensated by $\mathbf{i}\xi\frac{\mathcal{N}_B^*(\xi)}{\mathcal{G}_B^*(\xi)}$, since we have assumed $f_B(0) = 0$.

Fourth step. Inverse Fourier transforms. f_B and S_B are given by the formulae

$$f_B(a) = \frac{1}{2\pi}\int_{-\infty}^{+\infty}\Big(1 + \mathbf{i}\xi\frac{\mathcal{N}_B^*(\xi)}{\mathcal{G}_B^*(\xi)}\Big)e^{-\mathbf{i}a\xi}d\xi \quad \text{and} \quad S_B(a) = \int_a^{+\infty} f_B(s)ds,$$

provided all these quantities exist, and we have proved the formula (11). ■

Remark 2. *An alternative formula for $B(a)$ would be obtained using a direct formula for the survival function,*

$$S_B(a) = \frac{1}{2\pi}\int_{-\infty}^{+\infty}\frac{\mathcal{N}_B^*}{\mathcal{G}_B^*}(\xi)e^{-\mathbf{i}a\xi}d\xi.$$

2 Statistical setting: estimation procedure

Let us assume that we have a sample $X_1, \dots, X_n$ independent and identically distributed according to $U_{B,x}$. This idealizes the case where, for instance, a picture of all individuals at a given time is taken, and their sizes experimentally measured. We give here a procedure to estimate B from such a sample.

2.1 Estimation of $\mathcal{G}_B$

Step 1. Estimation of $U_{B,x}$ by a kernel estimator

$$\widehat{U}_{n,x}(y) = \frac{1}{n}\sum_{j=1}^{n} K_{h_1}(y - X_j)$$

with $K_{h_1}(\cdot) = h_1^{-1}K(h_1^{-1}\cdot)$. Estimation of $D(y) = \big(gU_{B,x}\big)'(y)$ by

$$\widehat{D}_n(y) = \frac{1}{n}\sum_{j=1}^{n} g(X_j)K'_{h_2}(y - X_j).$$

For the choice of the regularization parameters $h_1 = h_{1,n}$ and $h_2 = h_{2,n}$, see Section 3.

Step 2. Inversion of (9) replacing the left-hand side by $\lambda\widehat{U}_{n,x}(y) + \widehat{D}_n(y)$. We obtain $\widehat{\mathcal{L}}_n(y)$ an estimator of $\mathcal{L}_B(y)$. For this step, we follow [6], and concatenate the inverse given in Lemma 1 in $L^2(dx)$ for $x \le \bar{x}$, that we denote $\widehat{\mathcal{L}}_{n,l}(y)$, with the inverse in $L^2(x^4dx)$, denoted $\widehat{\mathcal{L}}_{n,r}(y)$, for $x \ge \bar{x}$, for a given (to be determined numerically) $\bar{x} > 0$: we set

$$\widehat{\mathcal{L}}_n(y) = \widehat{\mathcal{L}}_{n,l}(y)\mathbb{1}_{x\le\bar{x}} + \widehat{\mathcal{L}}_{n,r}(y)\mathbb{1}_{x>\bar{x}}.$$

Step 3. We deduce an estimator of $\mathcal{G}_B(y)$ using its definition (10):

$$\widehat{\mathcal{G}}_n(y) = 4e^{\lambda G(y)}\widehat{\mathcal{L}}_n(2y).$$

2.2 Estimation of the Fourier transforms $\mathcal{G}_B^*$ and $\mathcal{N}_B^*$

Estimation of $\mathcal{N}_B^*$. Recall Equation (14) giving the definition of $\mathcal{N}_B$. Then

$$\mathcal{N}_B^*(\xi) = \int_0^\infty g(x)e^{\lambda G(x)}e^{\mathbf{i}x\xi}U_{B,x}(x)dx,$$

which can be estimated by

$$\widehat{\mathcal{N}_n^*}(\xi) = \frac{1}{n}\sum_{j=1}^{n} g(X_j)e^{\lambda G(X_j)}e^{\mathbf{i}X_j\xi}.$$

When growth is exponential *i.e.* in the specific case $g(x) = \tau x$, we have $\lambda = \tau$ and $G(x) = \ln(x)/\tau$. Thus the previous formula simplifies into

$$\widehat{\mathcal{N}_n^*}(\xi) = \frac{\tau}{n}\sum_{j=1}^{n}(X_j)^2e^{\mathbf{i}X_j\xi}.$$

Estimation of $\mathcal{G}_B^*$. We compute the Fourier transform of $\widehat{\mathcal{G}}_n(y)$ (see Section 3 for the practical details). It gives us an estimator $\widehat{\mathcal{G}}_n^*$ of $\mathcal{G}_B^*(y)$.

2.3 Estimation of the Fourier transform f_B^*

We have an estimator of the Fourier transform of f_B by

$$\widehat{f_n^*}(\xi) = 1 + \mathbf{i}\xi \frac{\widehat{\mathcal{N}_n^*}(\xi)}{\widehat{\mathcal{G}_n^*}(\xi)} \mathbf{1}_{\Omega_n}(\xi)$$

with $\Omega_n = \{\xi \in \mathbb{R}; |\widehat{\mathcal{G}_n^*}(\xi)| \geq \underline{\xi}\}$, with $\underline{\xi} = \underline{\xi_n}$ a well-adapted threshold.

2.4 Inverse Fourier transforms to estimate S_B and f_B.

Estimators of f_B and S_B are

$$\widehat{f}_{n,h_3}(a) = \frac{1}{2\pi} \int_{-1/h_3}^{1/h_3} \widehat{f_n^*}(\xi) e^{-\mathbf{i}a\xi} d\xi \quad \text{and} \quad \widehat{S}_{n,h_3}(a) = \int_a^{\infty} \widehat{f}_{n,h_3}(s) ds$$

with $h_3 = h_{3,n}$ a parameter of regularization to be well-chosen (see Section 3).

2.5 Reconstruction of B.

Finally the division rate B can be estimated at a given point a by

$$\widehat{B}_{n,h}(a) = \frac{\widehat{f}_{n,h}(a)}{\widehat{S}_{n,h}(a) \vee \varpi_2} = \frac{\int_{-1/h}^{1/h} \left(1 + \mathbf{i}\xi \frac{\widehat{\mathcal{N}_n^*}(\xi)}{\widehat{\mathcal{G}_n^*}(\xi)} \mathbf{1}_{\Omega_n}(\xi)\right) e^{-\mathbf{i}a\xi} d\xi}{\int_s^{+\infty} \left(\int_{-1/h}^{1/h} \left(1 + \mathbf{i}\xi \frac{\widehat{\mathcal{N}_n^*}(\xi)}{\widehat{\mathcal{G}_n^*}(\xi)} \mathbf{1}_{\Omega_n}(\xi)\right) e^{-\mathbf{i}s\xi} d\xi \right) ds \vee \varpi}$$

with $\varpi = \varpi_n$ a well-adapted threshold.

Remark 3. *Following Remark 2, an alternative estimator of B would be obtained using*

$$\widehat{S}_{n,h_4}(a) = \frac{1}{2\pi} \int_{-1/h_4}^{1/h_4} \widehat{S_n^*}(\xi) e^{-\mathbf{i}a\xi} d\xi \quad \textit{where} \quad \widehat{S_n^*}(\xi) = \frac{\widehat{\mathcal{N}_n^*}}{\widehat{\mathcal{G}_n^*}}(\xi) \mathbf{1}_{\Omega_n}(\xi)$$

with $h_4 = h_{4,n}$ a regularization parameter to be well-chosen. Note that such a procedure would not guarantee the decaying property of S_B, or yet the fact that $f_B = -S_B'$, $S_B(0) = 1$ and $S_B(\infty) = 0$.

3 Numerical study

3.1 Numerical implementation

For a given estimator $\hat{g}$ of a function g (with real or complex values), we evaluate $\hat{g}$ on a regular grid with mesh Δt and compute the empirical error as

$$e = \frac{\|\hat{g} - g\|_{2,\Delta t}}{\|g\|_{2,\Delta t}}$$

where $\|\cdot\|_{2,\Delta t}$ is the L_2-discrete norm over the numerical sampling.

Choice of the regularization parameters

In the aboveseen formulae, we have distinguished five successive steps, and several regularization parameters which need careful implementation: h_1, h_2 and $\bar{x}$ for the estimation of $\mathcal{L}_B$; h_3 for the integration domain of the inverse Fourier transform; and the thresholds $\underline{\xi}$ and ϖ to avoid explosion when the denominators vanish in the formulae. Each of the five steps have been tested separately, and we discuss below practical ways to determine such regularization parameters.

- The parameter h_1 is either automatically chosen by the kernel smoothing function `ksdensity` of Matlab, or chosen via an adaptive method such as Goldenschluger and Lepski, or Penalized Comparison to Overfitting (PCO) most recently introduced [26, 34]. Similarly for h_2, we can use an adaptive method or choose it *a priori*, in relation with h_1 - for instance, if we have the a priori that $U_{B,x} \in H^1$, the order of an optimal choice for h_1 will be $O(n^{-\frac{1}{3}})$, for h_2 it is $O(n^{-\frac{1}{5}})$, so that we can choose *a priori* $h_2 = h_1^{\frac{3}{5}}$.

- To compute $\bar{x}$, we refer to the previous studies [6, 30]. Since for any estimation $\widehat{U}_{B,x} \in H^1((1+x^p)dx) \cap W^{1,\infty}$ with $p > 3$ the unique solution $\widehat{\mathcal{L}}_B^0$ in $L^2(dx)$ to Equation (9) is also in $L^1 \cap L^\infty$, which is not the case for the unique solution $\widehat{\mathcal{L}}_B^p$ in $L^2(x^p dx)$, we keep the solution in $L^2(x^p dx)$ only for the right-hand tail of the distribution, which leads us to define $\bar{x}$ as

$$\bar{x} := \mathrm{argmin}_{x \geq x^0_{max}} |\widehat{\mathcal{L}}_B^p - \widehat{\mathcal{L}}_B^0|, \qquad x^0_{max} := \mathrm{argmax}_{x \geq 0} \widehat{\mathcal{L}}_B^0.$$

- The key parameter h_3 is chosen in an oracle way, that is to say that we minimise in h the criterion

$$e(h) = \frac{\|\hat{f}_{n,h} - f_B\|_{2,\Delta a}}{\|f_B\|_{2,\Delta a}}.$$

  This oracle choice requires the knowledge of f_B, which is impossible in practice, but our aim here is to learn how well our procedure can do (in ideal conditions, when the tuning parameters can be chosen perfectly). In practice we have set $\underline{\xi} = 0$, since the regularization by h_3 suffices numerically. We chose $\varpi_n = 1/n$.

Use of the regularity properties of the functions

The protocol described in Section 2 does not guarantee important properties of the functions, such that the positivity of f_B and B, and $\int f_B da = 1$. All these characteristics will be enforced in the numerical study to improve qualitatively the results.

3.2 Numerical results on simulated data

In order to evaluate the quality of the reconstruction and the influence of each step of the protocol, we first studied separately each estimation necessary for the reconstruction of B. For the simulations, growth is exponential with rate 1, $g(x) = x$, so that $\lambda_B = 1$. We choose for the division rate $B(a) = a^2$ for $a \geq 0$. We thus immediately deduce formulae for the density f_B and the survival function S_B. We compute numerically the Fourier transform f_B^*. All Fourier and Inverse Fourier transforms were computed by an integral using the same scheme as in [6] (see Technical aspects of Section 4.1 in [6]).

Numerical solution for $U_{B,x}$

To compute numerically the first eigenvector U solution of (3)–(6), we follow the classical first order finite volume scheme proposed for instance in [7, 12] for similar equations, renormalize the solution at each time step, and stop the iterations when the distribution has converged, the error between two successive steps being smaller than the precision desired. An important point is to allow the scheme to be slightly dissipative - which is the case by choosing a regular grid - contrarily to the one proposed for instance in [5], which would give rise to an oscillatory behaviour.

The solution $U_{B,x}$ is computed in our framework ($g(x) = x$ and $B(a) = a^2$) with regular grids: size ranges from 0 to 6 and increment ranges form 0 to 3 with meshes $\Delta x = \Delta a = 6/500$. We require a precision of 0.01%. Besides the stationary distribution in size we compute the size-distribution of dividing cells $\mathcal{L}_B$. By formula (10), we obtain numerically $\mathcal{G}_B$ and $\mathcal{N}_B$. (And we compute numerically the Fourier transforms $\mathcal{G}_B^*$ and $\mathcal{N}_B^*$.)

Protocol 1 – Reconstruction of B when both $U_{B,x}$ and $\mathcal{L}_B$ are given with highest accuracy.

For the reconstruction of B we use directly the numerical and high-resolution solutions of $U_{B,x}$ and $\mathcal{L}_B$ of the previous step. The noise is thus limited to the numerical error, itself very limited thanks to the high resolution of the grid and to the requirement of a very small error in the long-time asymptotics.

See Figure 1 for the different steps of the protocol. See Figure 3 (red curves) and Tables 1 and 2 for results.

Protocol 2 – Reconstruction of B when $U_{B,x}$ is given with highest accuracy but $\mathcal{L}_B$ is unknown.

See Figure 2 for the different steps of the protocol. See Figure 3 (yellow curves) and Tables 1 and 2 for results.

Both of the Protocols 1 and 2 give a satisfactory reconstruction of the division rate B on the range $[0; 2]$, with an error around 8% (Table 2). The estimation deteriorates for an increment of size higher than 2 since the probability for a cell to exceed this increment is less than 10%. Computing the error on the wider range $[0; 2.5]$, we surprisingly observe that Protocol 2 (with an error around 13%) gives a more robust estimation than Protocol 1, which includes fewer statistical unknowns (error around 20%).

Coming back to the first steps (Table 1) we observe that Protocol 2 achieves errors below 5%, whereas Protocol 1 leads to 10% error for the reconstruction of the density f_B. Is this due to error compensation when computing the ratio in the reconstruction formula of Section 2.3? Figure 3d shows that the reconstruction of the Fourier transform of the density is good in modulus for frequency $|\xi| < 5$ for Protocol 2, whereas the reconstruction by Protocol 1 deteriorates from smaller frequencies (around $\xi = \pm 3$). Both protocols underestimate the maximum of the density f_B, but this is amplified using Protocol 1. As a consequence of the bias in the estimation of the density f_B, we observe a bias in the estimation of the division rate B. It is slightly overestimated for increments of size lower than 1 and underestimated beyond.

Protocol 3 – Reconstruction of B when $U_{B,x}$ is reconstructed from $X_1, \dots, X_n$ i.i.d. $\sim U_{B,x}$ and $\mathcal{L}_B$ is given with highest accuracy.

See Figure 4 for the different steps of the protocol. See Figures 5, 8 and 10 for results.

Protocol 4 – Reconstruction of B from $X_1, \dots, X_n$ i.i.d. $\sim U_{B,x}$.

See Figure 6 for the different steps of the protocol. See Figures 7, 8 and 10 for results.

Protocols 3 and 4 have been repeated $M = 100$ times for each tested n ranging from 500 to 50 000. This enables us to obtain empirical confidence intervals (CI) for the estimation of the division rate B and for the different intermediate reconstructions. As expected the computed 95%-CI shrink as n grows (Figure 8). The reconstruction of B is satisfactory on the range $[0; 1.75]$ when $n = 500$, and slightly beyond 2 when $n = 50\,000$. We observe the same bias as the one already mentioned for Protocols 1 and 2. It seems even amplified looking at the mean of the 100 reconstructions, to such an extent that the true division rate $B(a) = a^2$ is on the fringe of the 95%-CI when $n = 50\,000$.

One can plot the mean error over the $M = 100$ reconstructions versus the sample size n in log-log scale (Figure 10). Doing so linear curves are obtained and the extracted-slopes give us the speeds in the decrease of the error (with respect to n) for our different reconstructions. The speeds are surprisingly slightly better for Protocol 4 than for Protocol 3, which is in line with the comparison of Protocols 1 and 2.

As regards Protocol 4, the speed for the estimation of $U_{B,x}$ is close to $n^{-0.4}$, which is expected (indeed $(n_0 + 1)/(2(n_0 + 1) + 1) = 0.4$ with $n_0 = 1$ the order of a Gaussian kernel). For the estimation of $U'_{B,x}$ we expect $(n_0+1)/(2(n_0+1)+2) \approx 0.33$ and we obtain worse (slope of -0.25). The inversion step in order to obtain $\mathcal{L}_B$ (and $\mathcal{G}_B$) does not deteriorate this speed hugely (slope of -0.23). After the computation of the Fourier transform, the speed for the estimation of $\mathcal{G}_B$ is of the same order (slope of -0.25). For the estimation of $\mathcal{N}_B$ we obtain the expected slope -0.5 which corresponds to a parametric speed in $n^{-1/2}$. It is not possible to predict *a priori* the speed of a quotient estimator, and we obtain a slope of -0.24 for the estimation of f_B^*. We obtain a final speed of $n^{-0.16}$ for the estimation of B (for Protocol 4). This is much more difficult to interpret these last speeds since the regularity of $\mathcal{N}_B$ and $\mathcal{G}_B$ comes in. We refer to the Discussion below and to Johannes [25] for general results on the quality of density estimators in a deconvolution problem when the "noise" law is unknown, generalizing results such as Fan [18] when the "noise" law is known. See also the study of Belomestny and Goldenshluger [4] in the case of a multiplicative measurement error leading to the use of Mellin transform techniques (instead of an additive error leading to the use of Fourier transform ones).

Last but not least we observe a saturation of the error for very large n ($n \geq 20\,000$ in Protocol 3 and $n \geq 30\,000$ in Protocol 4). Thus the slopes were computed taking into account this effect when necessary (removing the last points).

3.3 Numerical results on experimental data

We now turn to experimental data of bacterial growth to test the method. In the corresponding experiments cells are followed through time and the joint distribution of instantaneous size and size increment is estimated, as well as the joint distribution of size and size increment of dividing cells. This allows us to compare our results obtained through our indirect method with a direct estimation of the division rate $B(a)$ from kernel density estimation of $f_B(a)$ and $S_B(a)$.

The dataset we analysed comes from a single-cell experimental study on *E. coli* growth, performed by Stewart *et al.* [32], and we used the data analysis performed in [30] (see Methods - data analysis). Following the results of [30], we can assume here that all cells grow with approximately the same growth rate $g(x) = \tau x$ with $\tau = 0.0275 min^{-1}$. Corollary 1 then states that we have $\mathcal{G}_B(y) = 4y\mathcal{L}_B(2y)$ and $\mathcal{N}_B(y) = \tau y^2 U_{B,x}(y)$.

The experimental sample contains $n = 31,333$ measurements of cell sizes. We perform a kernel density estimation with $h_1 = 0.125$ on Figure 11a to display it (Step 1 of the estimation procedure, Section 2.1).

The second step consists in the estimation $\widehat{\mathcal{L}}_n$ of the size distribution of dividing cells $\mathcal{L}_B$, through the numerical solution of Equation (9) (Step 2 of the procedure, Section 2.1). In our case of a rich dataset, we also have access to a sample of $n_d = 1,679$ dividing cells, so that we can compare our estimation with the kernel density estimation of this sample (done here with $h_d = 0.167$): we denote this estimate $\widehat{\mathcal{L}}_{n_d}^{h_d}$, and both are displayed in Figure 11b. We see that this approximation is relatively satisfactory, though far from being perfect. The distance between the two distributions has two main reasons. First, the sample of dividing cells may be noisier than the sample of all cells: since the measurement is done only on time intervals of $2min$, there is an error due to the size growth of the cell during these $2min$. Second, the cells may not all grow at exactly the same growth rate τ (see Figure S4 in [30]) so that Equation (9) is an approximation, see for instance [14] for a more complete model including growth rate variability. However, we also note that this way of computing the size distribution of cells remains valid even if the division rate would depend on other variables, so that Equation (9) allows one to estimate the size distribution of dividing cells from the measurement of a size sample, provided the approximation of homogeneous growth rate is valid, and if this growth rate is measured independently.

For the following steps, we do not have a direct way to compare the Fourier transforms of the intermediate functions $\mathcal{G}_B$ and $\mathcal{N}_B$ to directly measured quantities, but we can estimate f_B (and so $S_B = \int_a^\infty f_B(s)ds$ and $B = \frac{f_B}{S_B}$, as classically done for renewal processes) from the increment-and-size experimental distribution of dividing cells. Let us recall that f_B is *not* equal to the increment distribution of dividing cells, due to the well known bias selection effect [24, 30] of keeping the two daughter cells at each division step. However, for the specific case of a linear growth rate, we have an easy relation between the increment-and-size distribution of dividing cells and f_B. We notice

that the function $U_1(a,x) := \frac{xU(a,x)}{\iint xU(a,x)dadx} \geq 0$, is solution of the system

$$\frac{\partial}{\partial a}(g(x)U_1(a,x)) + \frac{\partial}{\partial x}(g(x)U(a,x)) + g(x)B(a)U_1(a,x) = 0, \tag{16}$$

$$g(x)U_1(0,x) = 2\int_0^\infty g(2x)B(a)U_1(a,2x)da, \quad g(0)U_1(a,0) = 0, \tag{17}$$

$$U \geq 0, \qquad \iint U(a,x)dadx = 1, \tag{18}$$

which defines the increment-and-size distribution of all cells in the conservative case, *i.e.* when only one child is kept at each division. We thus have

$$f_B(a) = \frac{\int \tau xB(a)U_1(a,x)dx}{\iint \tau xB(a)U_1(a,x)dadx} = \frac{\int x(xB(a)U(a,x))dx}{\iint x^2B(a)U(a,x)dadx},$$

and we notice that this formula is nothing but a weighted average of $\frac{xB(a)U(a,x)}{\iint xB(a)U(a,x)dadx}$, which is the increment-and-size distribution of dividing cells taken in experimental conditions, that is, when the two children are kept at each division (for a more detailed explanation on the links between the two points of view, we refer to [14], Section 4.2). All these considerations provide us with the following estimate of f_B from the increment-and-size sample of dividing cells: let us denote $(A_j, X_j)_{1\leq j\leq n_d}$ the 2-dimensional sample of increment A and size X at division, that we assume to behave as if (A_j, X_j) were independent identically distributed according to the probability law $\frac{xB(a)U(a,x)}{\iint xB(a)U(a,x)dadx}$: we have

$$\widehat{f}_{B,n_d}(a) := \frac{1}{n_d}\sum_{j=1}^{n_d} K_{h_d}(a - A_j)X_j.$$

In Figure 12a we compare $\widehat{f}_{B,n}$ to a kernel density estimation of the increment distribution of dividing cells; and finally in Figure 12b we compare our estimator $\widehat{B}_n$ to $\widehat{B}_{n_d} = \frac{\widehat{f}_{B,n_d}}{\int_a^\infty \widehat{f}_{B,n_d}(s)ds \vee \varpi_2}$. If the results may be viewed as qualitatively in agreement, they are not fully satisfactory, and especially not comparable with the results obtained on simulated data. The reasons may be twofold. First, modeling errors: the incremental model is in good agreement with the data but cannot completely describe the full complexity of the biological process, including all potential fluctuations

Second, our problem being severely-ill posed, even if the results on simulated data are of good quality, we did not take into account the experimental noise, which is due to errors and/or imprecision of image analysis, leading to noise in size measurement and division timing, and to the sampling in time (which affects only the measurements on dividing cells), with a time step that is only ten times less than the cell generation time. This means that relatively important differences in the increment distribution of dividing cells / in the increment-structured division rate, may lead to differences on the size distribution of all cells which, for this level of noise, are not significant.

4 Discussion

In this article, we have proposed an explicit reconstruction formula to estimate the increment-structured division rate of a population dividing by fission into two equal parts. The formula may be easily generalized to other types of division kernels, as done for the size-structured equation in [6]. Based on this formula, we designed and implemented a numerical protocol, which we used to investigate numerically which rates of convergence could be expected. We finally tested the method on experimental data; though our results reveal qualitatively satisfactory, they did not reach the precision obtained for simulated data. This highlighted inherent difficulties of the problem, which deserve to be further investigated. These difficulties are linked to two sources of noise, not yet adressed neither theoretically nor numerically: modelling error and single-cell measurement errors. Investigating the influence of modelling error, *e.g.* heterogeneity of cells with respect to growth rate or fragmentation kernel, or yet dependence of the division rate on another trait different from the increment, could give first insights in this direction. To take into account single-cell measurement errors, we need to add a deconvolution problem to our noise model, assuming for instance that we measure realizations of an i.i.d. sample $Y_i = X_i + \sigma\xi_i$ where X_i are i.i.d. random variables distributed along $U_{B,x}$ and ξ_i are i.i.d. normally distributed random variable, σ being the level of noise.

Let us now discuss some possible variants of the method, and directions to prove estimation inequalities.

Possible variants of the method

Instead of estimating $\mathcal{L}_B$ from $U_{B,x}$ by Lemma 1, using $L^2(dx)$ for $x < \bar{x}$ and $L^2(x^4dx)$ for $x > \bar{x}$, and only then take its Fourier transform, we could use the following lemma 2: first define an estimate of the function Γ_B from the sample, and then solve Equation (19). It seems attractive since the Fourier transform then admits a simple and explicit definition from the sample $(X_1, \cdots X_n)$; but in practice, it appeared difficult to handle and would deserve a full study - in particular to determine, as for $\bar{x}$, a convenient threshold $\bar{\xi}$ to use one or the other of the spaces $L^2(x^p dx)$.

Lemma 2. *Let $\mathcal{G}_B$ defined by (10), and $\mathcal{G}_B^*$ its Fourier transform. Then $\mathcal{G}_B^*$ is solution of the following equation*

$$\mathcal{G}_B^*(2\xi) = \mathcal{G}_B^*(\xi) + \Gamma_B(\xi) \tag{19}$$

with Γ_B defined by

$$\Gamma_B(\xi) = \mathbf{i}\tau\xi \int_0^\infty x^2 e^{\mathbf{i}x\xi} U_{B,x}(x)dx. \tag{20}$$

For $\Gamma_B \in L^2(\xi^p d\xi)$ with $p \geq 0$, this functional equation admits a unique solution $\mathcal{G}_B^* \in L^2(\xi^p d\xi)$.

Proof. At the view of (9) and (10) one immediately gets

$$\mathcal{G}_B^*(\xi) = \int_0^\infty \mathcal{G}_B(x)e^{\mathbf{i}x\xi}dx = \int_0^\infty 4e^{\lambda G(x)}\mathcal{L}_B(2x)e^{\mathbf{i}x\xi}dx = T_1 + T_2$$

with

$$T_1 = \int_0^\infty e^{\lambda G(x)}\big(\lambda U_{B,x}(x) + (gU_{B,x})'(x)\big)e^{\mathbf{i}x\xi}dx \qquad \text{and} \qquad T_2 = \int_0^\infty e^{\lambda G(x)}\mathcal{L}_B(x)e^{\mathbf{i}x\xi}dx.$$

Let us compute the first term:

$$T_1 = \int_0^\infty \lambda e^{\lambda G(x)+\mathbf{i}x\xi} U_{B,x}(x)dx + \Big[(gU_{B,x})(x)e^{\lambda G(x)+\mathbf{i}x\xi}\Big]_{x=0}^{\infty} - \int_0^\infty (gU_{B,x})(x)\big(\tfrac{\lambda}{g(x)} + \mathbf{i}\xi\big)e^{\lambda G(x)+\mathbf{i}x\xi}dx$$
$$= -\mathbf{i}\xi \int_0^\infty e^{\lambda G(x)+\mathbf{i}x\xi} g(x) U_{B,x}(x)dx.$$

In order to treat the second term, we assume the growth is exponential *i.e.* $g(x) = \tau x$. In this special case we have $G(x) = \ln(x)/\tau$ and thus $G(2x) = G(2) + G(x)$. Then

$$T_2 = 2\int_0^\infty e^{\lambda G(2x)}\mathcal{L}_B(2x)e^{\mathbf{i}2x\xi}dx = \tfrac{1}{2}e^{\lambda G(2)}\int_0^\infty 4e^{\lambda G(x)}\mathcal{L}_B(2x)e^{\mathbf{i}x(2\xi)}dx = \mathcal{G}_B^*(2\xi)$$

using at last $e^{\lambda G(2)} = 2$ since $\lambda = \tau$. Gathering the two terms we obtain (19). The existence and uniqueness in $L^2(x^p dx)$ for $p \neq -1$ directly follows from [16] Proposition A.1.applied to the equation transformed for $u(\xi) = \frac{1}{\xi^2}\mathcal{G}_B^*(\xi)$, which satisfies

$$4u(2\xi) - u(\xi) = \frac{\Gamma_B(\xi)}{\xi^2},$$

which admits a unique solution for $\frac{\Gamma_B}{x^2} \in L^2(\xi^q d\xi)$ for $q \neq 3$, which is equivalent to $\Gamma_B \in L^2(\xi^p d\xi)$ with $p \neq -1$. Since we have $\mathcal{G}_B^*(0) = \int_0^\infty 4y\mathcal{L}_B(2y)dy > 0$ (it represents the average size of dividing cells), we are only interested in $p \geq 0$. ■

The function Γ_B can be easily estimated by $\widehat{\Gamma}_n(\xi) = \frac{\mathbf{i}\tau\xi}{n}\sum_{j=1}^n X_j^2 e^{\mathbf{i}X_j\xi}$ truncated for $|\xi| \leq \overline{\xi}$. Then, in the same spirit as for the solution of (9), we could solve Equation (19) by writing

$$\mathcal{G}_B^*(\xi) = -\sum_{k=0}^\infty \Gamma_B(2^k\xi), \qquad \Longrightarrow \qquad \widehat{\mathcal{G}}_n = -\sum_{k=0}^N \widehat{\Gamma}_n(2^k\xi), \qquad \forall \xi \in \mathbb{R},$$

but numerically it happens not to give better estimation results and to be less easily compared with the original function in the space state, so that we preferred the method explained above.

Other variants and improvements of the method would be to constraint the space of solutions for $\widehat{f}_{n,h}$ to the space of probability measures. This is done manually in our procedure, taking the positive and real part of the estimated inverse Fourier transform (see Section 2.3), but constrained optimization or projection on a finite-dimension space approximating probability measures could improve the results. This will be investigated in future work.

Estimation inequalities

To prove estimation inequalities, several difficulties appear, which give a roadmap for further investigation of the problem.

First, we have to prove that the denominator of our ratios in our inverse Fourier transforms, namely $\mathcal{G}_B^*$, does not vanish. This has been done for related problems (estimation of the fragmentation kernel) in two recent papers, by using complex analysis methods (Lemma 1.iii in [23], Theorem 2.i in [13]). In [13], it was the central and most technical point of the study. Here however, the proof carried out in [23] based on the argument principal could probably be adapted.

Second, the deconvolution problem (15) appears as a deconvolution problem with unknown "noise", since $\mathcal{G}_B$ plays the role of the noise. The difficulty is thus to investigate whether $\mathcal{G}_B$ is *ordinary smooth* or *super smooth*, with the following definitions:

- Ordinary smooth of order β: $c_1|t|^{-\beta} \leq |\mathcal{G}_B^*(t)| \leq c_2|t|^{-\beta}$ for any $|t| \geq M$, for positive constants.
- Super smooth of order β: $c_1|t|^{\gamma_1}e^{-c_0|t|^\beta} \leq |\mathcal{G}_B^*(t)| \leq c_2|t|^{\gamma_2}e^{-c_0|t|^\beta}$ for any $|t| \geq M$, for positive constants.

The smoother the "noise", the more ill-posed the problem. Once a given order of magnitude for the decay of $\mathcal{G}_B^*$ assumed, speeds of convergence and orders of magnitude for the choice of the regularization constant h_3 may be classically obtained, see for instance [1] (ch.4, Section 4.2.2). However, assuming a given decay for $\mathcal{G}_B^*$ means that we assume a certain degree of regularity - and no more - for $\mathcal{G}_B$, *i.e.* for $\mathcal{L}_B$, *i.e.* for the unknown B itself. If regularity results exist and can be extended to higher regularity by the chain rule for our equation, see e.g. [11, 20], the reverse is false - results such as: if B is not derivable, then $U_{B,x}$ cannot be twice derivable. This shows the importance of designing *a posteriori* and adaptive methods.

Acknowledgments. A.O. was on leave ("délégation") at the french National Research Centre for Science (CNRS) during the finalization of this work. M.D. has been partly supported by the ERC Starting Grant SKIPPER^{AD} (number 306321). We thank Albert Cohen, Marc Hoffmann and Benoît Perthame for very fruitful discussions.

5 Appendix: Figures and tables

Figures 1, 2, 4 and 6 use the notation $\mathcal{F}$ for the Fourier transform and $\mathcal{F}_h^{-1}$ for the following operator. For a suitable function f,

$$\mathcal{F}_h^{-1}(f)(a) = \frac{1}{2\pi}\int_{-1/h}^{1/h} f(\xi)e^{-ia\xi}d\xi.$$

Reconstruction of	$\mathcal{L}_B$	$\mathcal{G}_B$	$\mathcal{G}_B^*$	f_B^*	f_B	S_B
Numerical sampling	[0;6] $\Delta x = \frac{6}{500}$	[0;6] $\Delta x = \frac{6}{500}$	[-50;50] $\Delta\xi = 0.05$	$[\frac{-1}{4.75}; \frac{1}{4.75}]$ $\Delta\xi = 0.05$	[0;5] $\Delta a = 0.01$	[0;5] $\Delta a = 0.01$
Protocol 1	-	-	-	0.1062	0.1043	0.0395
Protocol 2	0.0478	0.0417	0.0417	0.0470	0.0482	0.0149

Table 1: Errors of Protocols 1 and 2 for the intermediate steps.

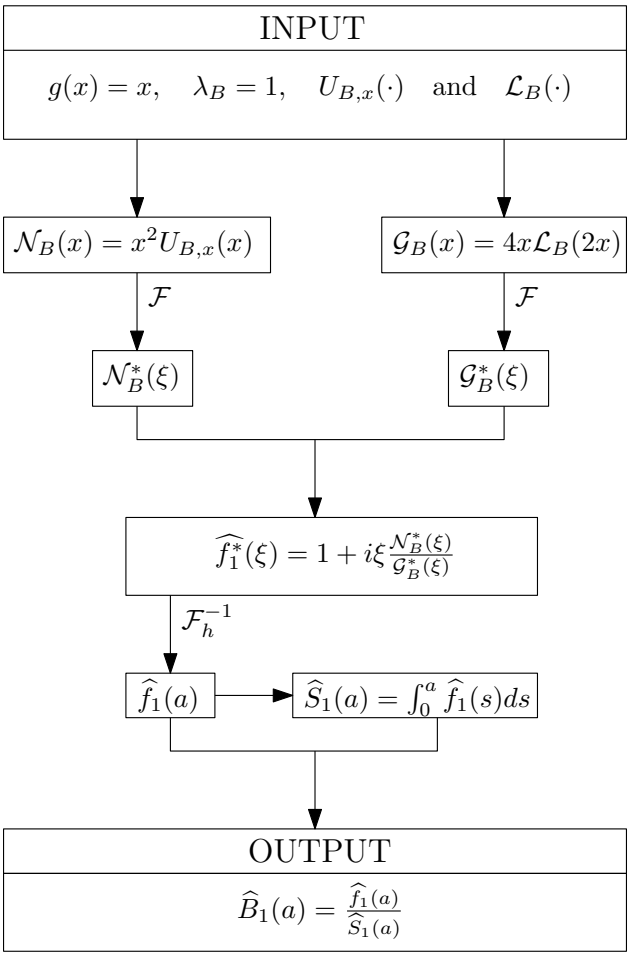


Figure 1: Protocol 1 – Reconstruction of B when both $U_{B,x}$ and $\mathcal{L}_B$ are (almost) exactly known. The oracle choice for h gives us the value $1/4.75$.

Reconstruction of	B	B
Numerical sampling	[0;2] $\Delta a = 0.01$	[0;2.5] $\Delta a = 0.01$
Protocol 1	0.0730	0.2065
Protocol 2	0.0849	0.1321

Table 2: Errors of Protocols 1 and 2 for B in function of the numerical sampling.

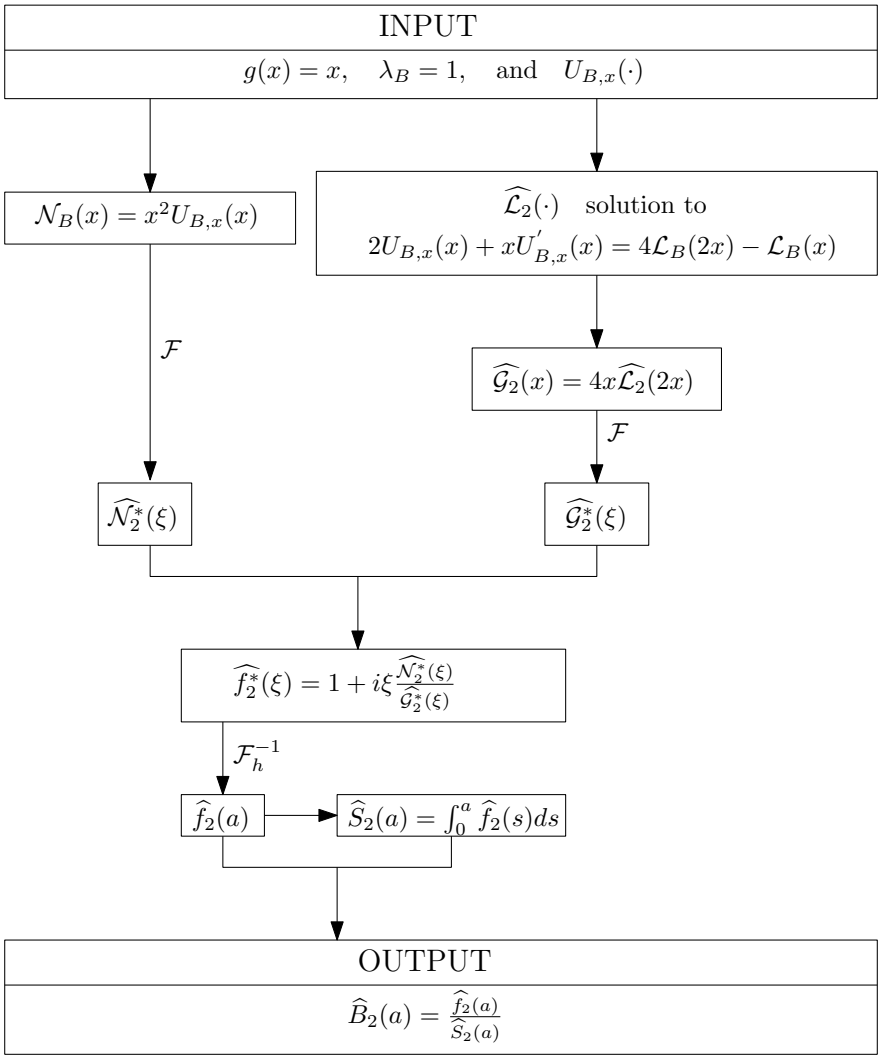
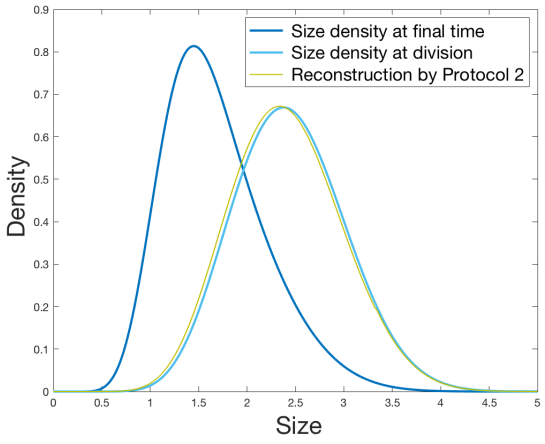
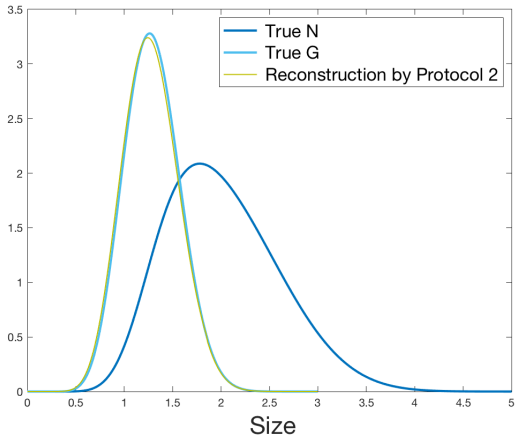


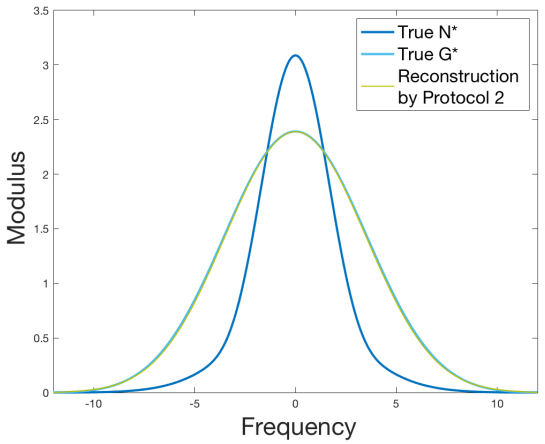
Figure 2: Protocol 2 – Reconstruction of B when $U_{B,x}$ is (almost) exactly known but not $\mathcal{L}_B$. The oracle choice for h gives us the value $1/5$.



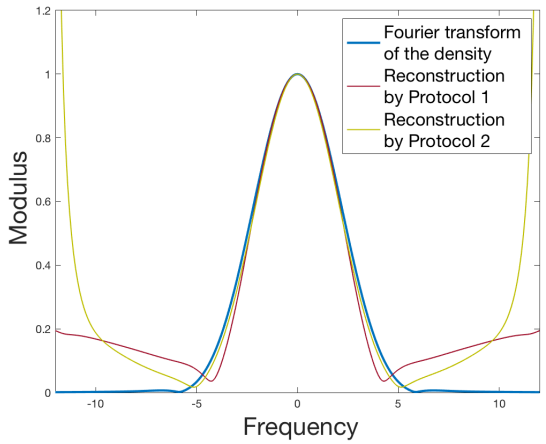
(a) $U_{B,x}$, $\mathcal{L}_B$ and $\widehat{\mathcal{L}}_2$ in function of x



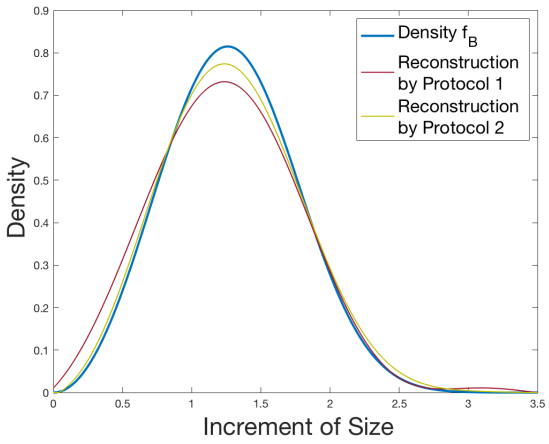
(b) $\mathcal{N}_B$, $\mathcal{G}_B$ and $\widehat{\mathcal{G}}_2$ in function of x



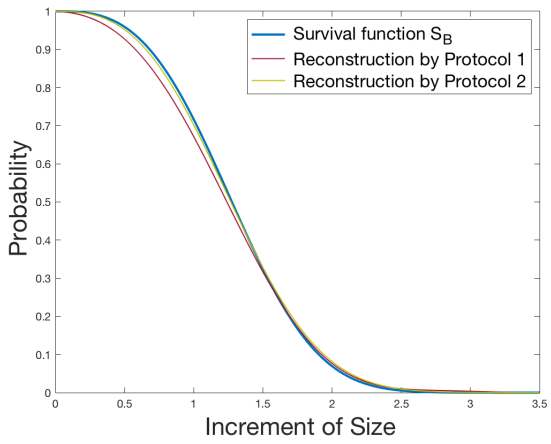
(c) $|\mathcal{N}_B^*|$, $|\mathcal{G}_B^*|$ and $|\widehat{\mathcal{G}}_2^*|$ in function of ξ



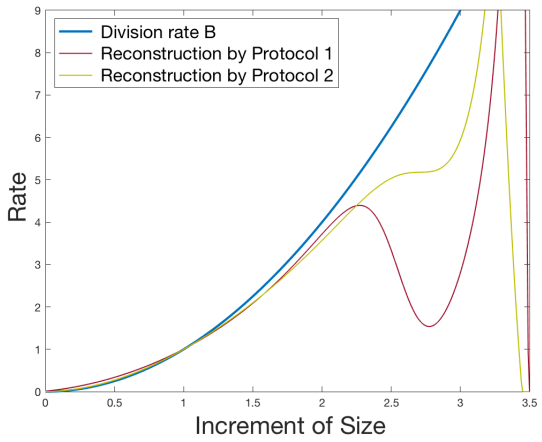
(d) $|f_B^*|$, $|\widehat{f}_1^*|$ and $|\widehat{f}_2^*|$ in function of ξ



(e) f_B, $\widehat{f}_1$ and $\widehat{f}_2$ in function of a



(f) S_B, $\widehat{S}_1$ and $\widehat{S}_2$ in function of a



(g) B, $\widehat{B}_1$ and $\widehat{B}_2$ in function of a

Figure 3: Results of Protocols 1 and 2. (x stands for size, ξ for frequency and a for increment of size)

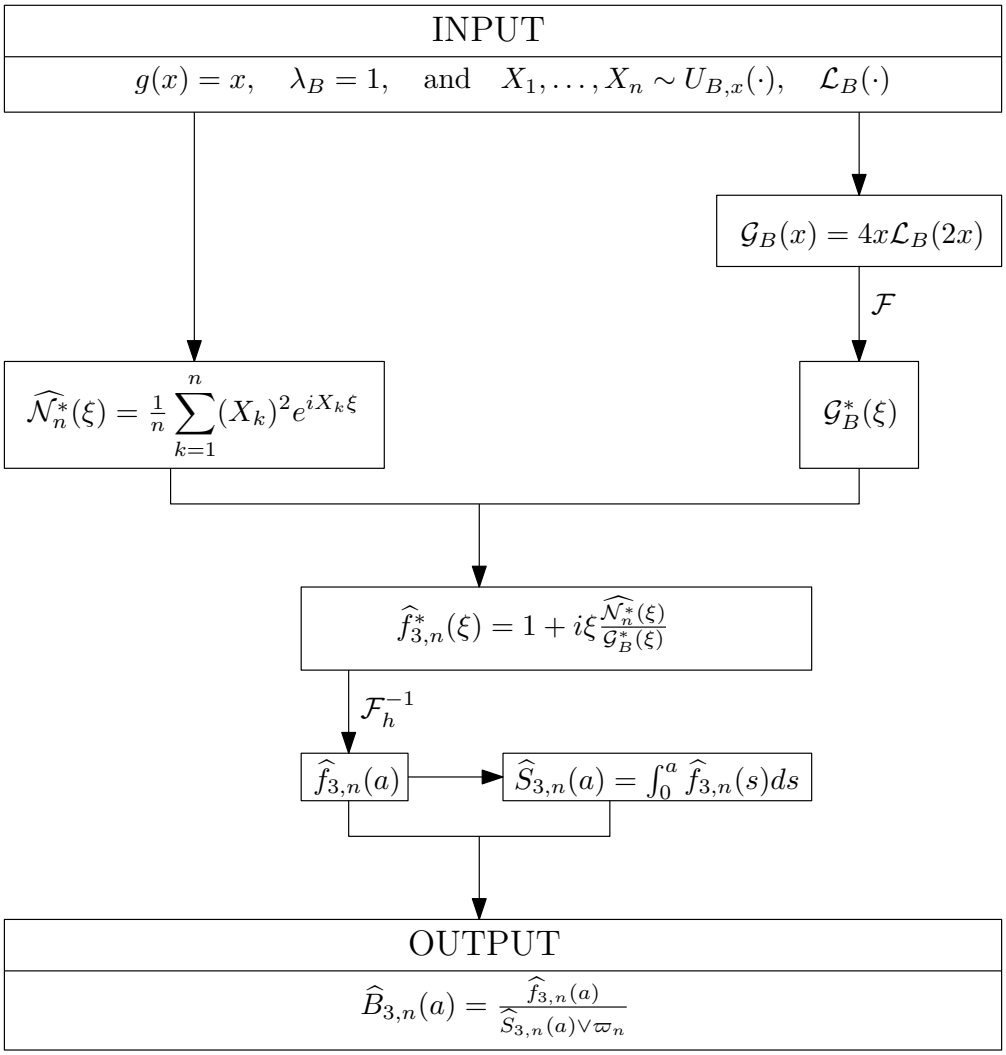
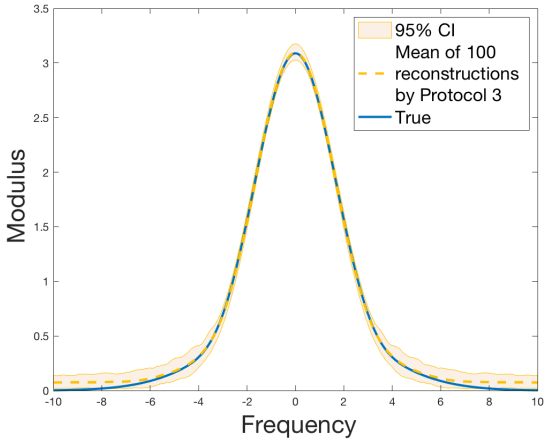
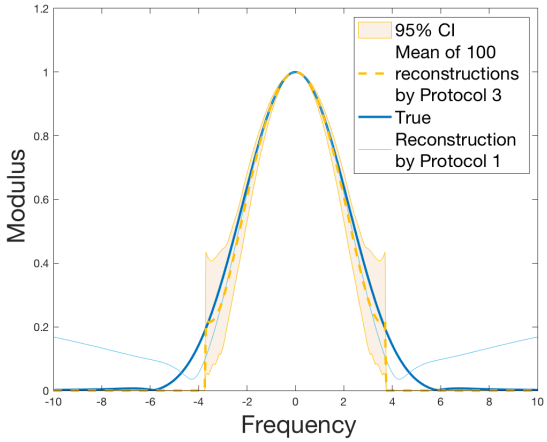


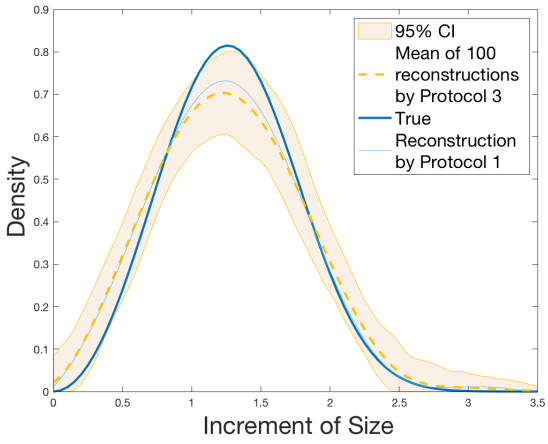
Figure 4: Protocol 3 – Reconstruction of B when $U_{B,x}$ is reconstructed from $X_1, \ldots, X_n$ i.i.d. $\sim U_{B,x}$ but $\mathcal{L}_B$ is (almost) exactly known. The oracle choice for h_3 gives us values that range between $1/3.25$ for $n = 500$ and $1/4.75$ for $n = 50\ 000$. We set $\varpi_n = 1/n$.



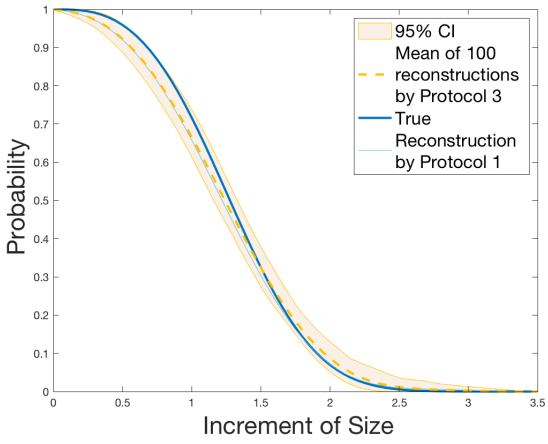
(a) $|\mathcal{N}_B^*|$, $|\widehat{\mathcal{N}_n^*}|$ in function of ξ



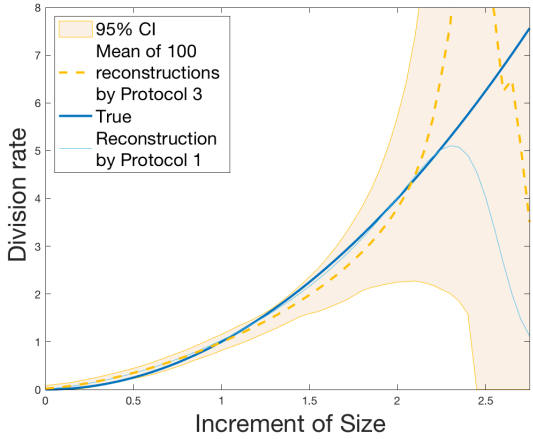
(b) $|f_B^*|$, $|\widehat{f}_1^*|$ and $|\widehat{f}_{3,n}^*|$ in function of ξ



(c) f_B, $\widehat{f}_1$ and $\widehat{f}_{3,n}$ in function of a



(d) S_B, $\widehat{S}_1$ and $\widehat{S}_{3,n}$ in function of a



(e) B, $\widehat{B}_1$ and $\widehat{B}_{3,n}$ in function of a

Figure 5: Results of Protocol 3 for $n = 2000$ and $M = 100$ Monte Carlo samples. (x stands for size, ξ for frequency and a for increment of size)

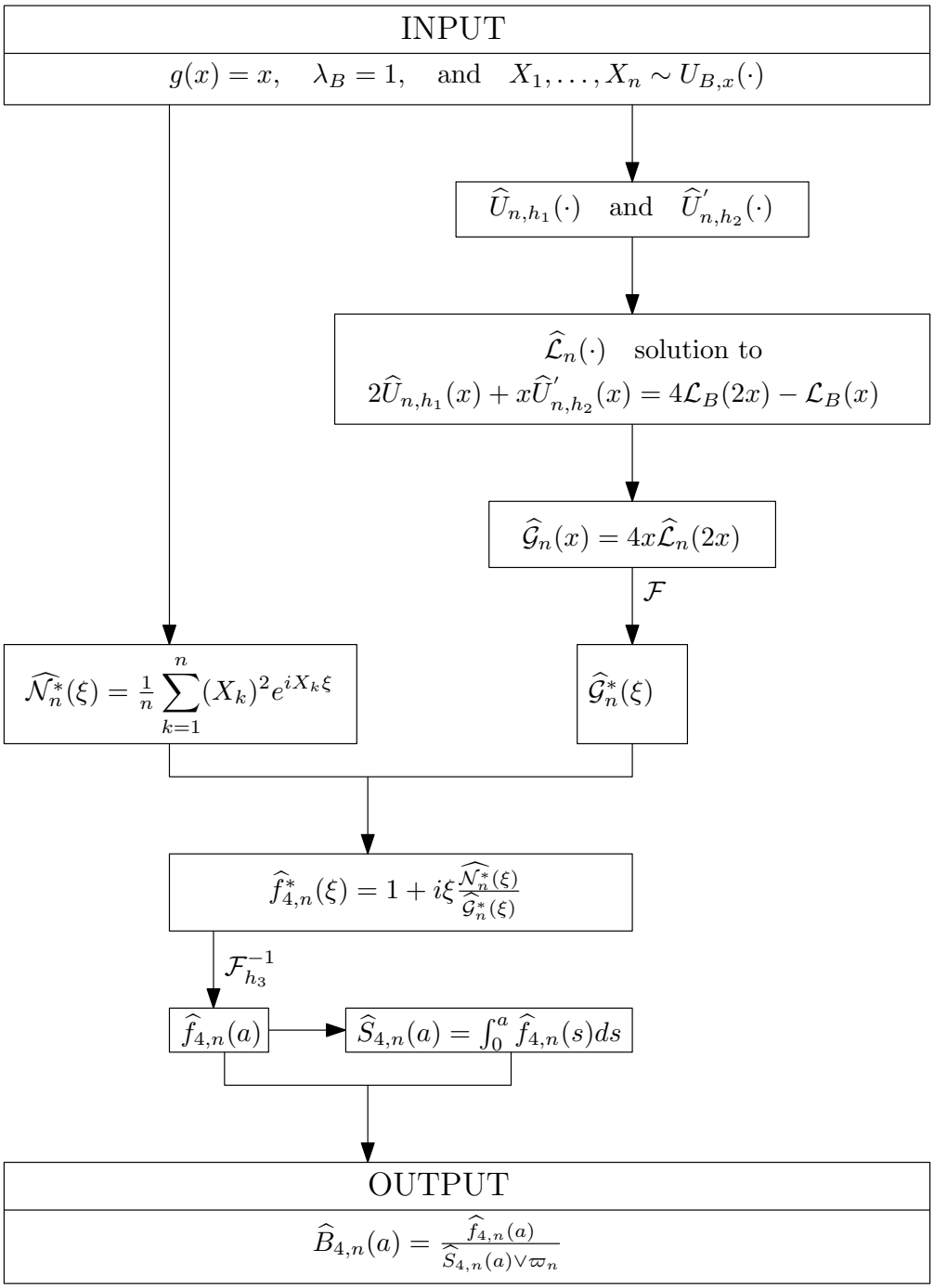
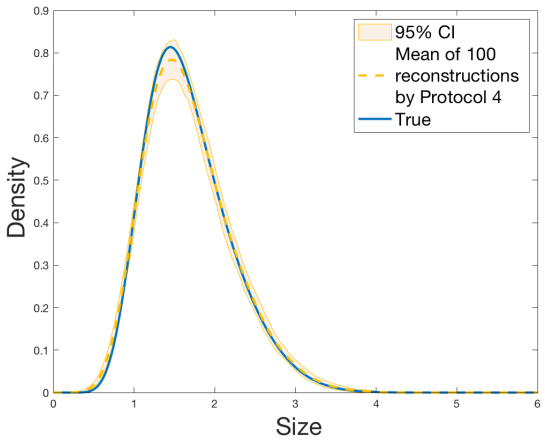
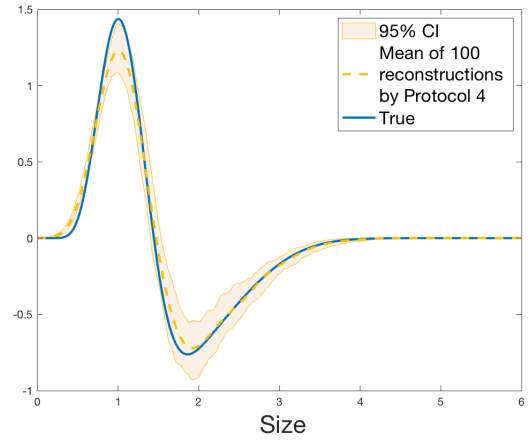


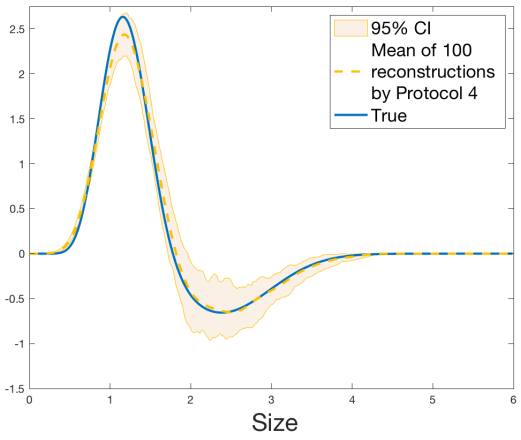
Figure 6: Protocol 4 – Reconstruction of B when both $U_{B,x}$ and $\mathcal{L}_B$ are reconstructed from $X_1, \ldots, X_n$ i.i.d. $\sim U_{B,x}$. The parameter h_1 is automatically chosen by the kernel smoothing function `ksdensity`; h_2 is deduced from h_1. The oracle choice for h_3 gives us values that range between $1/3.25$ for $n = 500$ and $1/4.5$ for $n = 50\ 000$. We set $\varpi_n = 1/n$.



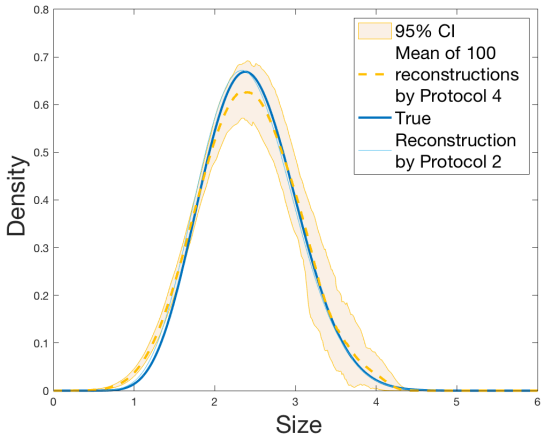
(a) $U_{B,x}$ and $\widehat{U}_{n,h}$



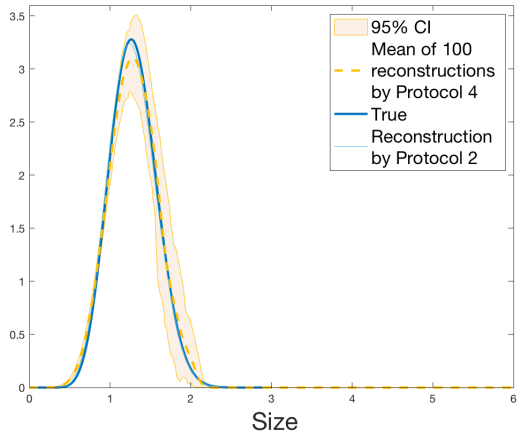
(b) $U^{'}_{B,x}$ and $\widehat{U}^{'}_{n,h'}$



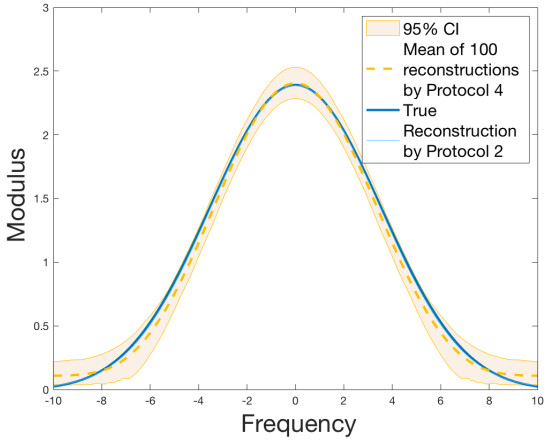
(c) $2U_{B,x} + xU^{'}_{B,x}$ and $2\widehat{U}_{n,h} + \widehat{U}^{'}_{n,h'}$



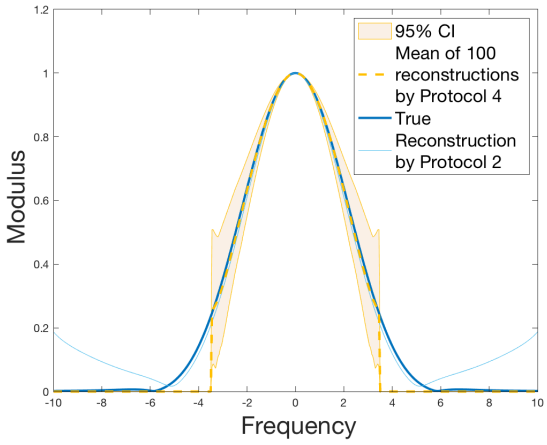
(d) $\mathcal{L}_B$ and $\widehat{\mathcal{L}}_n$ in function of x



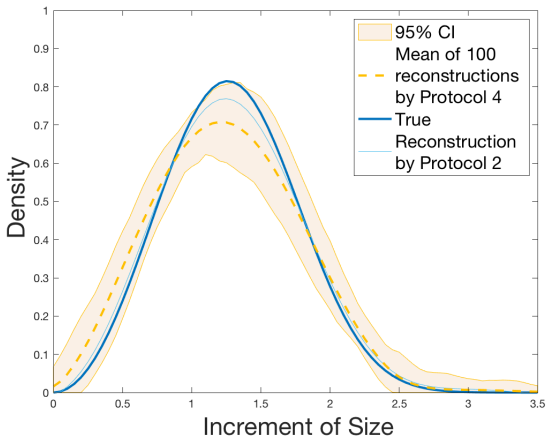
(e) $\mathcal{G}_B$ and $\widehat{\mathcal{G}}_n$ in function of x



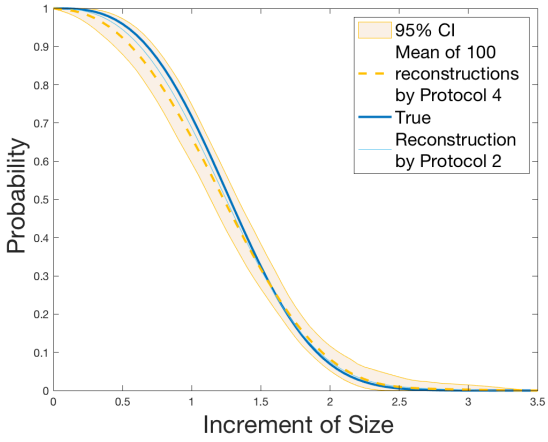
(f) $|\mathcal{G}^*_B|$, $|\widehat{\mathcal{G}}^*_2|$ and $|\widehat{\mathcal{G}}^*_n|$ in function of ξ



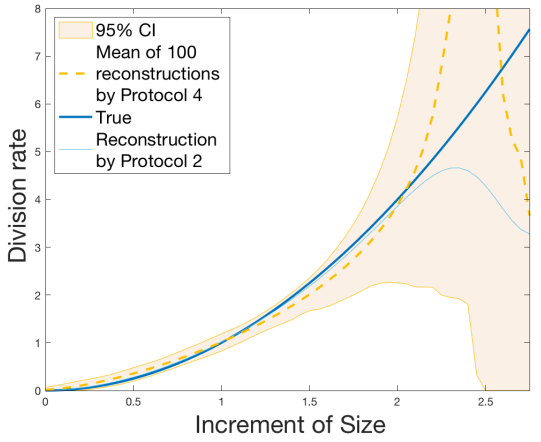
(g) $|f_B^*|$, $|\widehat{f}_2^*|$ and $|\widehat{f}_{4,n}^*|$ in function of ξ



(h) f_B, $\widehat{f}_2$ and $\widehat{f}_{4,n}$ in function of a

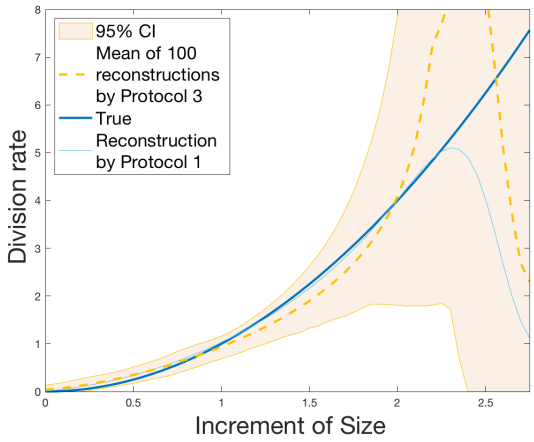


(i) S_B, $\widehat{S}_2$ and $\widehat{S}_{4,n}$ in function of a

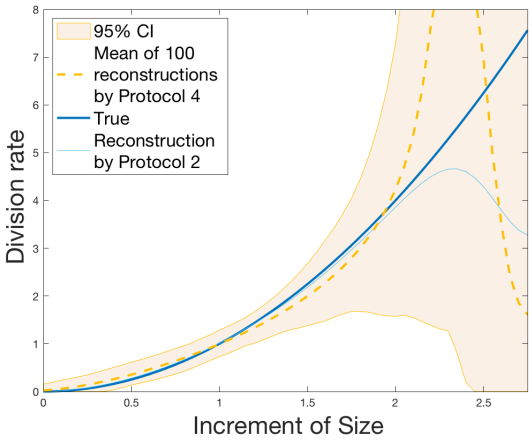


(j) B, $\widehat{B}_2$ and $\widehat{B}_{4,n}$ in function of a

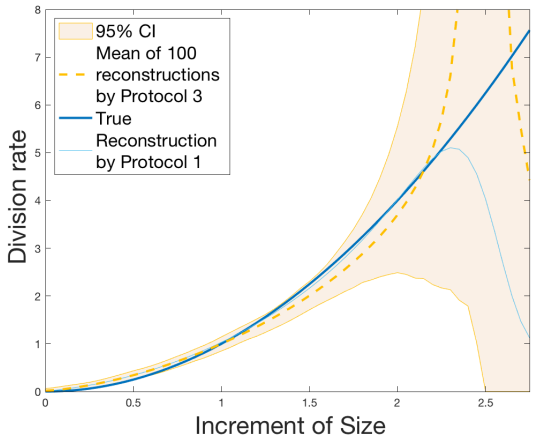
Figure 7: Results of Protocol 4 for $n = 2000$ and $M = 100$ Monte Carlo samples.



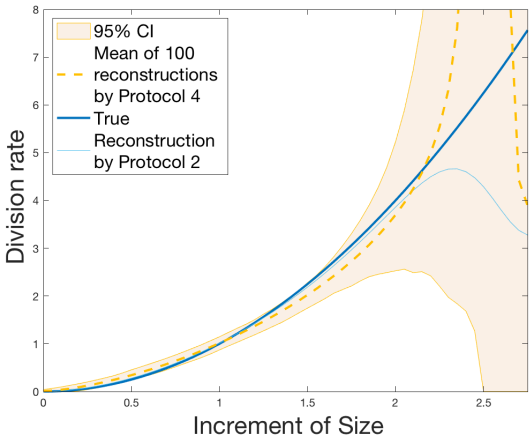
(a) $\widehat{B}_{3,n}$ with $n = 500$



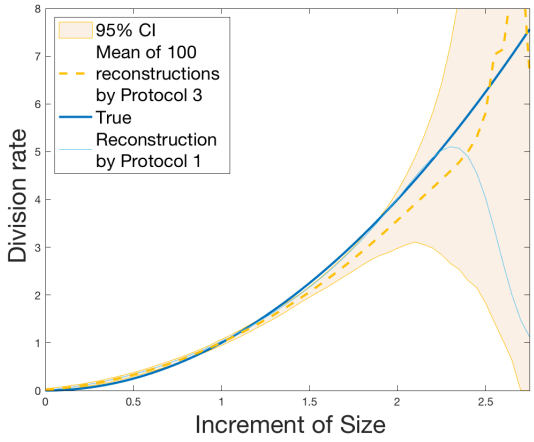
(b) $\widehat{B}_{4n}$ with $n = 500$



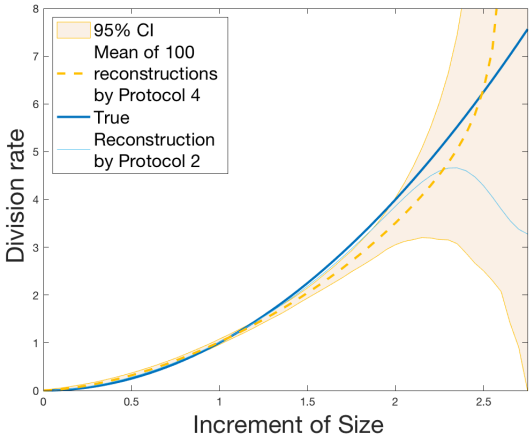
(c) $\widehat{B}_{3,n}$ with $n = 5\ 000$



(d) $\widehat{B}_{4,n}$ with $n = 5\ 000$

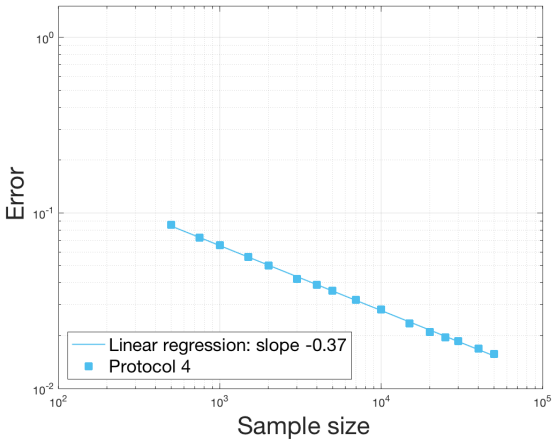


(e) $\widehat{B}_{3,n}$ with $n = 50\ 000$

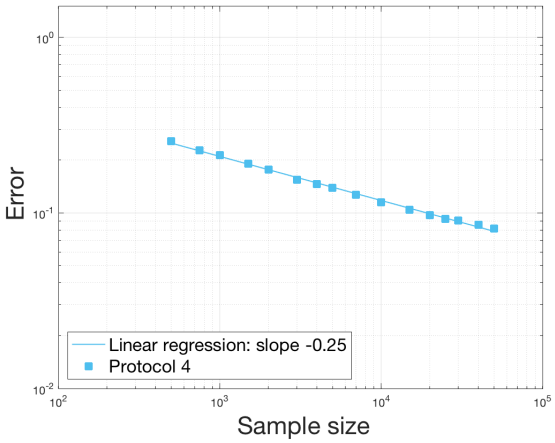


(f) $\widehat{B}_{4,n}$ with $n = 50\ 000$

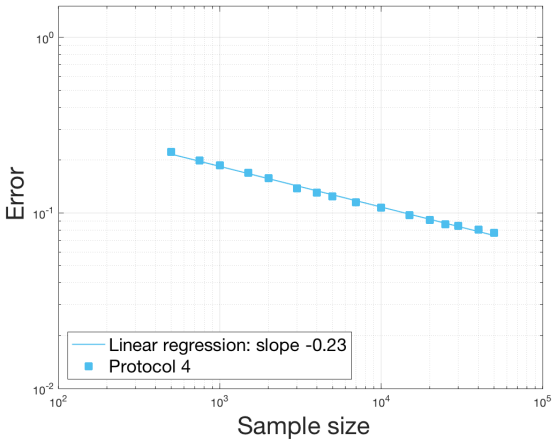
Figure 8: Results of Protocols 3 and 4 – Estimation of the division rate $B(a) = a^2$ in function of the increment of size a for different n (500, 5 000, 10 000) and $M = 100$ Monte Carlo samples.



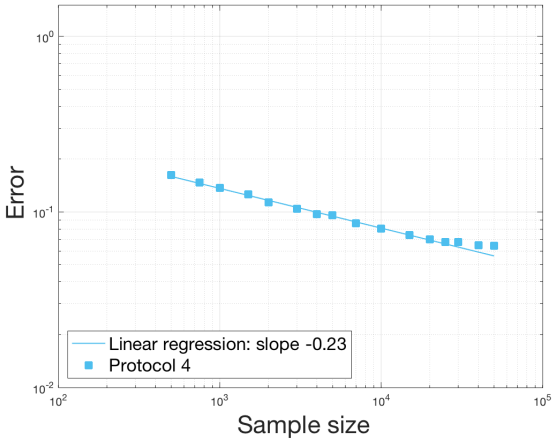
(a) Estimation of $U_{B,x}$



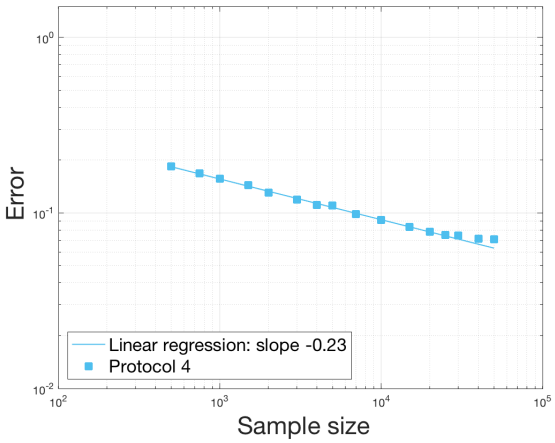
(b) Estimation of $U^{'}_{B,x}$



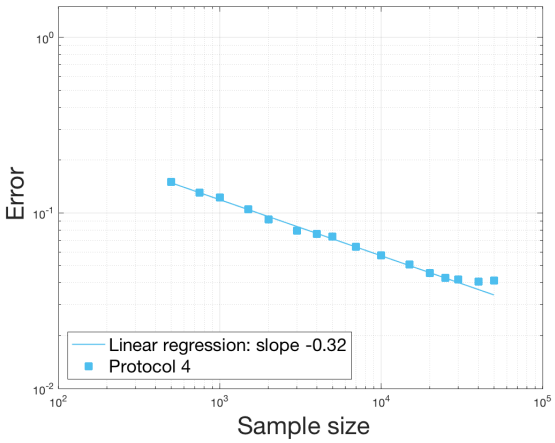
(c) Estimation of $2U_{B,x} + xU^{'}_{B,x}$



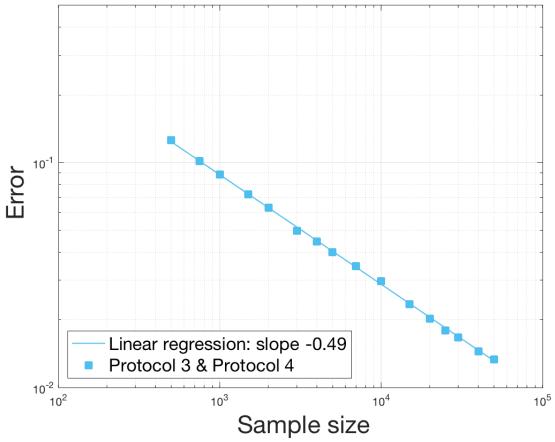
(d) Estimation of $\mathcal{L}_B$



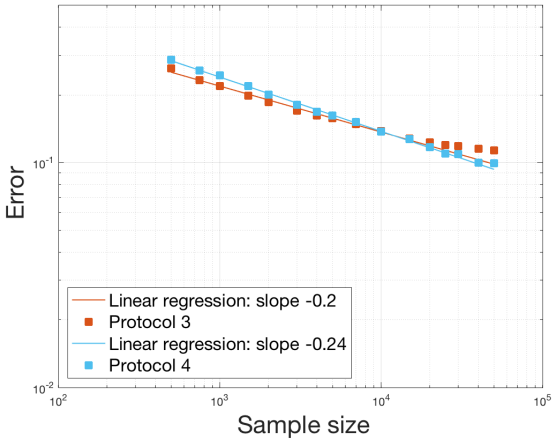
(e) Estimation of $\mathcal{G}_B$



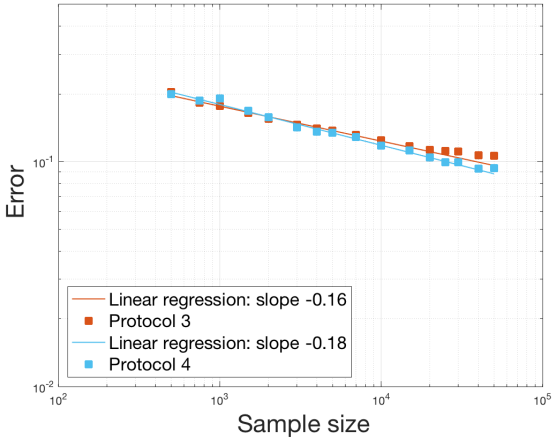
(f) Estimation of $\mathcal{G}^*_B$



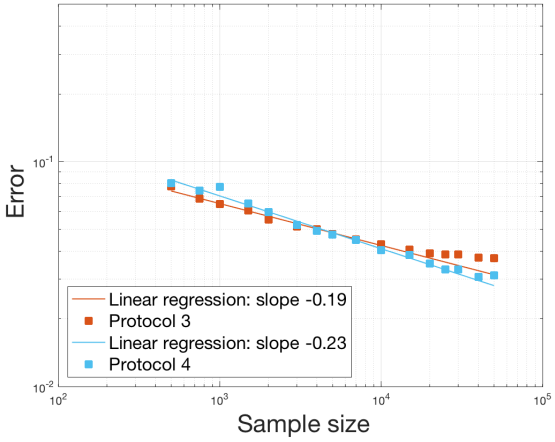
(a) Estimation of $\mathcal{N}_B^*$



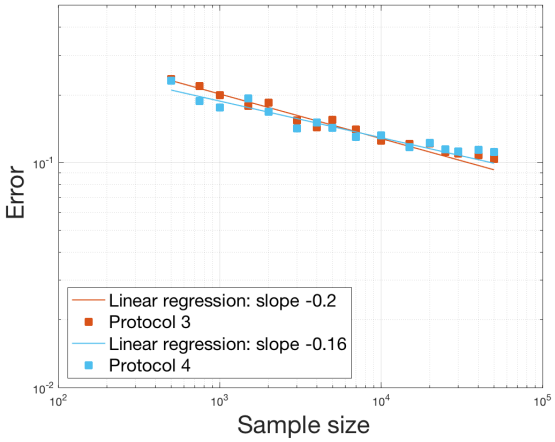
(b) Estimation of f_B^*



(c) Estimation of f_B

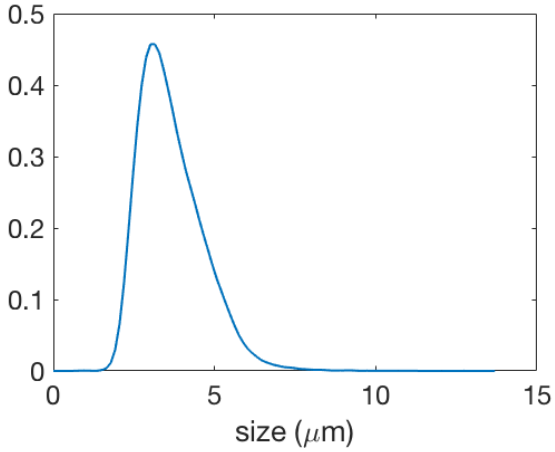


(d) Estimation of S_B

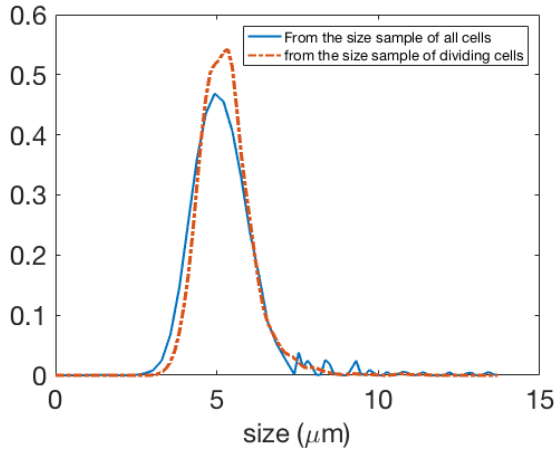


(e) Estimation of B

Figure 10: Results of Protocols 3 and 4 – Reduction of the mean error over $M = 100$ samples (in log-scale) in function of the sample size (from $n = 500$ to $n = 50\,000$). Empirical errors are computed over the following regular grids: (a)-(e) $[0; 6]$, $\Delta x = \frac{6}{500}$; (f)-(h) $[-10; 10]$, $\Delta\xi = 0.05$; (i)-(j) $[0; 2.25]$, $\Delta a = \frac{1}{\sqrt{n}}$; (k) $[0; 2]$, $\Delta a = \frac{1}{\sqrt{n}}$.

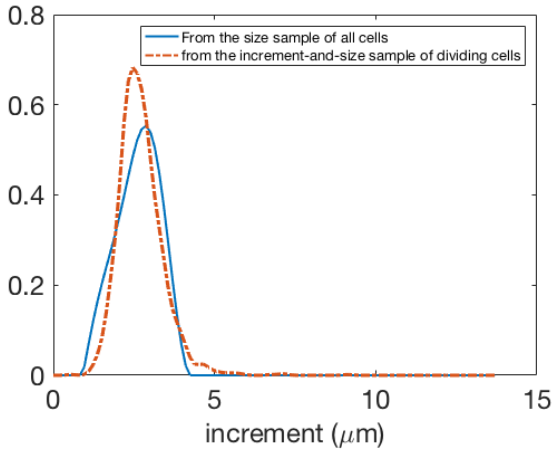


(a) **Estimation of the size distribution** $\widehat{U}_{n,x}$ of $U_{B,x}$ from an experimental sample taken from [32], $n = 31,333$, $h = 0.125$.

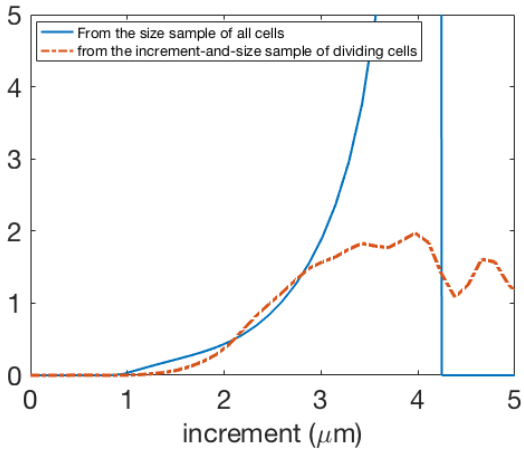


(b) **Estimation of the size distribution of dividing cells $\mathcal{L}_B$:** 1/ estimation $\mathcal{L}_B$ from $\widehat{U}_{n,x}$ (plain blue line, Step 2 of Section 2.1), 2/ estimation from the experimental sample of dividing cells, for which $n_d = 1,679$ and $h_d = 0.167$ (dotted-dashed red line).

Figure 11: **Testing the procedure on experimental data.**
Initial step: estimation of the size distribution



(a) **Estimation of $f_B(a)$**



(b) **Estimation of the division rate $B(a)$**

Figure 12: **Testing the procedure on experimental data.**
Final step: estimation of the increment-structured division rate

References

[1] Johann B. and Antonio L. *Topics in inverse problems*. Publicações Matemáticas do IMPA. [IMPA Mathematical Publication]. Instituto Nacional de Matemática Pura e Aplicada (IMPA), Rio de Janeiro, 25º colóquio brasileiro de matemática. [25th brazilian mathematics colloquium] edition, 2005.

[2] H. T. Banks, K. L. Sutton, W. C. Thompson, G. Bocharov, D. Roose, T. Schenkel, and A. Meyerhans. Estimation of cell proliferation dynamics using cfse data. *Bulletin of mathematical biology*, 73(1):116–150, 2011.

[3] H. T. Banks and W. C. Thompson. A division-dependent compartmental model with cyton and intracellular label dynamics. *Int. J. Pure Appl. Math*, 77:119–147, 2012.

[4] D. Belomestny and A. Goldenshluger. Nonparametric density estimation from observations with multiplicative measurement errors. *arXiv preprint arXiv:1709.00629*, 2017.

[5] E. Bernard, M. Doumic, and P. Gabriel. Cyclic asymptotic behaviour of a population reproducing by fission into two equal parts. *Kinetic and Related Models* , 12(3):551–571, June 2019.

[6] T. Bourgeron, M. Doumic, and M. Escobedo. Estimating the division rate of the growth-fragmentation equation with a self-similar kernel. *Inverse Problems*, 30(2):025007, 2014.

[7] F. B. Brikci, J. Clairambault, and B. Perthame. Analysis of a molecular structured population model with possible polynomial growth for the cell division cycle. *Mathematical and Computer Modelling*, 47(7-8):699–713, 2008.

[8] Clotilde Cadart, Sylvain Monnier, Jacopo Grilli, Pablo J Sáez, Nishit Srivastava, Rafaele Attia, Emmanuel Terriac, Buzz Baum, Marco Cosentino-Lagomarsino, and Matthieu Piel. Size control in mammalian cells involves modulation of both growth rate and cell cycle duration. *Nature communications*, 9(1):3275, 2018.

[9] Manuel Campos, Ivan V Surovtsev, Setsu Kato, Ahmad Paintdakhi, Bruno Beltran, Sarah E Ebmeier, and Christine Jacobs-Wagner. A constant size extension drives bacterial cell size homeostasis. *Cell*, 159(6):1433–1446, 2014.

[10] Maxime Deforet, Dave Van Ditmarsch, and Joao B Xavier. Cell-size homeostasis and the incremental rule in a bacterial pathogen. *Biophysical journal*, 109(3):521–528, 2015.

[11] M. Doumic. Analysis of a population model structured by the cells molecular content. *Math. Model. Nat. Phenom.*, 2(3):121–152, 2007.

[12] M. Doumic. Analysis of a population model structured by the cells molecular content. *Mathematical Modelling of Natural Phenomena*, 2(3):121–152, 2007.

[13] M. Doumic, M. Escobedo, and M. Tournus. Estimating the division rate and kernel in the fragmentation equation. *Annales de l'Institut Henri Poincaré (C) Non Linear Analysis*, 2018.

[14] M. Doumic, M. Hoffmann, N. Krell, and L. Robert. Statistical estimation of a growth-fragmentation model observed on a genealogical tree. *Bernoulli*, 21(3):1760–1799, 2015.

[15] M. Doumic, M. Hoffmann, P. Reynaud, and V. Rivoirard. Nonparametric estimation of the division rate of a size-structured population. *SIAM J. on Numer. Anal.*, 50(2):925–950, 2012.

[16] M. Doumic, B. Perthame, and J.P. Zubelli. Numerical solution of an inverse problem in size-structured population dynamics. *Inverse Problems*, 25(4):045008, 2009.

[17] Ye-Jin Eun, Po-Yi Ho, Minjeong Kim, Salvatore LaRussa, Lydia Robert, Lars D Renner, Amy Schmid, Ethan Garner, and Ariel Amir. Archaeal cells share common size control with bacteria despite noisier growth and division. *Nature microbiology*, 3(2):148, 2018.

[18] J. Fan. On the optimal rates of convergence for nonparametric deconvolution problems. *The Annals of Statistics*, 19(3):1257–1272, 1991.

[19] Anouchka Fievet, Adrien Ducret, Tâm Mignot, Odile Valette, Lydia Robert, Romain Pardoux, Alain R Dolla, and Corinne Aubert. Single-cell analysis of growth and cell division of the anaerobe desulfovibrio vulgaris hildenborough. *Frontiers in microbiology*, 6:1378, 2015.

[20] P. Gabriel and H. Martin. Steady distribution of the incremental model for bacteria proliferation. working paper or preprint, 2018.

[21] G. Greiner and R. Nagel. Growth of cell populations via one-parameter semigroups of positive operators. In J. Goldstein, S. Rosencrans, and G. Sod, editors, *Mathematics Applied to Science*, pages 79 – 105. Academic Press, 1988.

[22] A. J. Hall, G. C. Wake, and P. W. Gandar. Steady size distributions for cells in one-dimensional plant tissues. *Journal of Mathematical Biology*, 30(2):101–123, 1991.

[23] V. H. Hoang, T. M. Pham Ngoc, V. Rivoirard, and V. C. Tran. Nonparametric estimation of the fragmentation kernel based on a PDE stationary distribution approximation. working paper or preprint, 2017.

[24] M. Hoffmann and A. Olivier. Nonparametric estimation of the division rate of an age dependent branching process. *Stochastic Processes and their Applications*, December 2015.

[25] J. Johannes. Deconvolution with unknown error distribution. *The Annals of Statistics*, 37(5A):2301–2323, 2009.

[26] C. Lacour, P. Massart, and V. Rivoirard. Estimator selection: a new method with applications to kernel density estimation. *Sankhya A*, 79(2):298–335, 2017.

[27] J. Monod. The growth of bacterial cultures. *Annual Reviews in Microbiology*, 3(1):371–394, 1949.

[28] M. Nussbaum and S. Pereverzev. The degrees of ill-posedness in stochastic and deterministic noise models. *Preprint WIAS 509*, 1999.

[29] B. Perthame and J.P. Zubelli. On the inverse problem for a size-structured population model. *Inverse Problems*, 23(3):1037–1052, 2007.

[30] L. Robert, M. Hoffmann, N. Krell, S. Aymerich, J. Robert, and M. Doumic. Division in Escherichia coli is triggered by a size-sensing rather than a timing mechanism. *BMC Biology*, 12(1):17, 2014.

[31] Ilya Soifer, Lydia Robert, and Ariel Amir. Single-cell analysis of growth in budding yeast and bacteria reveals a common size regulation strategy. *Current Biology*, 26(3):356–361, 2016.

[32] E. Stewart, R. Madden, G. Paul, and F. Taddei. Aging and death in an organism that reproduces by morphologically symmetric division. *Curr. Biol.*, 20(12):1099–103, 2010.

[33] Sattar T.-A., S. Bradde, J. T. Sauls, N. S. Hill, P. A. Levin, J. Paulsson, M. Vergassola, and S. Jun. Cell-size control and homeostasis in bacteria. *Current Biology*, 11679(1-7), 2015.

[34] S. Varet, C. Lacour, P. Massart, and V. Rivoirard. Numerical performance of penalized comparison to overfitting for multivariate kernel density estimation. *arXiv preprint arXiv:1902.01075*, 2019.