Semiparametric Estimation of a Noise Model with Quantization Errors

Sébastien Li-Thiao-Té

Noname manuscript No.
(will be inserted by the editor)

Semiparametric Estimation of a Noise Model with Quantization Errors

Li-Thiao-Té Sébastien

the date of receipt and acceptance should be inserted later

Abstract The detectors in mass spectrometers are precise enough to count ion events, but in practice large quantization errors affect the observations. To study the statistics of low intensity chemical noise, we model the detector signal as $X = \lfloor \tau N \rfloor$ and estimate both τ and N in a semi-parametric approach where the integer valued random variable N represents the number of ions and τ represents the gain parameter of the detector. When $\tau \leq 1$, we explain why the gain parameter cannot be recovered without a priori information on N. When $\tau > 1$ however, N can be deduced from X and a sufficiently precise estimate of τ. To perform parametric estimation of τ, we first study simple estimators which provide useful upper bounds. We then introduce the concept of *compatible lattices* and we derive an optimal estimator that is independent of the law of N.

Keywords parametric estimation · quantization effects · life sciences signal processing

Mathematics Subject Classification (2000) MSC 62F10 Point estimation (Parametric inference) · 65G30 Interval and finite arithmetic · 65G50 Roundoff error

1 Introduction

1.1 Ion Detectors

Mass spectrometers are instruments that ionize the compounds of a sample, separate the ions, then quantify the ions at each mass to charge ratio. The resulting signal is a histogram that represents ion intensity as a function of the mass to charge ratio of the ions. With sufficient precision in the separation and the mass to charge measurement, the components of the sample can be identified and quantified. Mass spectrometers are widely used for analysing very diverse mixtures, e.g. detecting explosives for

CMLA ENS Cachan CNRS UniverSud, 61 Avenue President Wilson, F-94230 Cachan E-mail: lithiaote@cmla.ens-cachan.fr

airport security or analysing oil products. See [1, 8, 9] for an introduction to mass spectrometry in the life sciences.

We consider detectors similar to microchannel plate detectors that are used in most mass spectrometers [11]. When an ion hits the detector plate, it produces an analog signal that is amplified, quantized, then reported to the computer. The level of quantization is quite high as there may be only $2^{11} = 2048$ levels[1] in some instruments, and small signals as well as chemical noise are strongly affected by quantization effects.

Specific difficulties have appeared with high-throughput analyses of biological material. In particular, biological samples may contain trace amounts of molecules of interest. These are difficult to distinguish from chemical noise which produces patterns similar to real signals [3, 7, 10]. In 2004, [2] suggested Poisson-like behaviour for the ion intensity based on a linear relationship between the mean and variance of the noise. This linear relationship suggests that the amplification factor of the detector may be unaccounted for in the data set.

To study chemical noise in the experimental data, we interpret the amplification factor as an overdispersion parameter in a semi-parametric approach. To study chemical noise in the experimental data, we estimate the amplification factor and an unknown distribution for the chemical noise in a semi-parametric approach. Let N denote the number of chemical noise ions that reach the detector, we consider the following observation model:

$$X = \lfloor \tau N + \varepsilon \rfloor$$

where the noisy signal $\tau N + \varepsilon$ is truncated before observation. τ represents the amplification factor of the detector and ε represents electronic noise. In this paper, we make the assumption that $\varepsilon = 0$ or equivalently that there are only quantization errors in the measurements. The observation model is associated with the statistical structure $(\mathbb{N}, \mathfrak{B}, \mathbb{P}_\theta)$, with $\theta = (\tau, N)$ where τ is a positive real number and N is a probability distribution on $\mathbb{N}$.

We believe a priori that N is Poisson distributed as it models rare events (ion counts). Consequently, τ can be interpreted as an overdispersion parameter affecting Poisson distributed observations. This has been tackled in the framework of double exponential families, as presented by Efron in [5]. In [4], Antoniadis et al use double exponential families in a regression model to analyse diffraction spectra. This corresponds to estimating the regression function μ_i in the model $X'_i = \tau N_i$ where N_i is Poisson distributed with varying parameter μ_i. However, this framework does not explicitly take into account quantization errors and thus provides poor parameter estimates as we will show in Section 4.2. Moreover, we wish to confirm the Poisson hypothesis using non parametric estimation.

Our approach is to first estimate τ given a set of observations of X, then deduce the distribution of N from the estimate. We show that the estimate is precise enough to allow complete disambiguation of the observations.

[1] Single precision floating point numbers have $2^{24} \sim 16.10^6$ levels of precision plus sign and exponent.

1.2 Estimation of the Ion Statistics

With negligible quantization error, the observation model becomes $X = \tau N$. Estimation of τ is trivial; all that is required is to observe the event $\{N = 1\}$ i.e. $\{X = \tau \times 1\}$, or the two events $\{x_1 = \tau i\}$ and $\{x_2 = \tau(i+1)\}$ and compute the difference $x_2 - x_1$. To recover N, it then suffices to consider X/τ. The quantization error may be neglected when $\tau \gg 1$ in the observation model $X = \lfloor \tau N \rfloor$ and the previous estimates provide τ with a precision on the order of the quantization error.

In the general case, we can recover the samples of N from the samples of X when the mapping $x \mapsto \lfloor \tau x \rfloor$ is injective. The inverse mapping is $y \mapsto \lceil y/\tau \rceil$. We call this situation the *distinguishible* case. It occurs if and only if $\tau \geq 1$ (see proof in the Appendix, Prop 6). In this situation, the semi-parametric approach can be separated into parametric estimation of the gain parameter τ then non parametric estimation of the distribution of N from iid samples.

Example 1

```
> data = floor(1.32 * n)
% Distinguishible case
> n
 [1]  1  2  3  4  5  6  7  8  9 10
> data
 [1]  1  2  3  5  6  7  9 10 11 13
```

When τ is smaller than 1, the truncation error merges adjacent values of N. In the following example, the events $\{N = 3\}$ and $\{N = 4\}$ cannot be distinguished in the data set. This is because the corresponding observation is $\{X = 2\}$ in both cases.

Example 2

```
> data = floor(0.68 * n)
% Non distinguishible case
> n
 [1]  1  2  3  4  5  6  7  8  9 10
> data
 [1]  0  1  2  2  3  4  4  5  6  6
```

In the distinguishible case, it is natural to sort and index the observed values in order to determine the mapping $x \mapsto \lfloor \tau x \rfloor$. This is not sufficient in practice because of missing values or outliers which can modify the indexes.

1.3 Observation Set

The gain parameter and the law of N have separate effects on X. In the distinguishible case, the distribution function of X is a transformation of the distribution function of N by the mapping $x \mapsto \lfloor \tau x \rfloor$. The gain parameter and the truncation error only distort the position of each peak, whereas the relative frequencies are unchanged.

Consequently, the support $\mathscr{S}$ of the empirical distribution is sufficient information for estimating τ whereas the empirical frequencies are sufficient information for the distribution of N.

In the non distinguishible case, the set of observed integers is always $\mathbb{N}$ for large samples (see Section 3). As a consequence, τ cannot be estimated based on that set alone. A semi-parametric approach is not feasible either. For instance, we cannot estimate the mean $\mathbb{E}[N]$ but only $\mathbb{E}[X] = \tau\mathbb{E}[N]$. To separate τ and N, we have to provide prior assumptions on the distribution of N like a Poisson parametric family.

In the following, we study properties of the set $\mathscr{S}$ of observed integers. This set can be constructed from the dataset in $\mathscr{O}(n\log(n))$ time using sorting for example. The algorithmic complexity of the following algorithms is governed by the size of $\mathscr{S}$, and in particular, the maximum integer in $\mathscr{S}$.

We focus on the distinguishible case, and perform parametric estimation of the gain parameter from a random set of integers. As the support of the empirical distribution is a sufficient statistic for τ, we use the statistical structure $\left(\Omega = 2^{\mathbb{N}}, \mathfrak{T}, \mathbb{P}_\tau, \tau \in]1,+\infty[\right)$ where Ω is the power set of $\mathbb{N}$ and $\mathfrak{T}$ is the exhaustive σ-algebra on Ω.
$(\mathbb{P}_\tau, \tau \in]1,+\infty[)$ is a parametric family of distributions on Ω that is implicitly generated in the following way. For a fixed integer n and fixed but unknown integer-valued random variable N, $\mathbb{P}_\tau$ is the distribution of the random variable $\mathscr{S}$ which is the set of observed integers in an independent identically distributed sample $(X_1,\ldots,X_n)$ of $X = \lfloor \tau N \rfloor$.

1.4 Organization of the paper

To estimate τ in the distinguishible case, we first provide simple estimators for τ in Section 2. These are later used as a starting point for improved estimators and to restrict the search space for τ.

In Section 3, we define the notion of *compatible values* and provide a few properties of the set of compatible values. In particular, the true parameter τ is a compatible value and is close to the highest compatible value. This leads to an optimal estimator that is described in 3.3.

We show the results of some simulations in section 4 and compare with the Maximum Likelihood Estimator obtained from the Double Poisson Family, an estimator based on linear regression and another one based on Fourier transform.

2 Estimators and Upper Bounds for τ

The results in this section are based on the following idea. Two points in $\mathscr{S}$ are separated by at least $\lfloor \tau \rfloor$. Consequently, when τ is large, then $\mathscr{S}$ is a sparse set, whereas $\mathscr{S}$ is dense when τ is near 1. For instance, there are consecutive points in $\mathscr{S}$ if and only if $\tau \leq 2$ (see Proposition 7 in the Appendix).

A better estimate can be obtained by combining more than 2 consecutive points. Let $[\![x,y]\!]$ denote the set of integers between x and y. If $[\![x,y]\!]$ is a subset of $\mathscr{S}$, then $\tau < 1 + \frac{1}{y-x}$. Consequently, τ can be estimated by $1 + \frac{1}{y-x}$ with a precision on the order of the inverse of the length of the interval $\frac{1}{y-x}$. However, this estimator is strongly affected by missing values in $\mathscr{S}$.

Instead of considering all the segments in $\mathscr{S}$, we propose to use the overall density of the set, which is easier to compute algorithmically. Let $\hat{x} = \lfloor \tau\, \hat{n} \rfloor$ denote the largest integer in $\mathscr{S}$. Then $\tau < \frac{\hat{x}+1}{\hat{n}}$. When $\hat{n}$ is unknown (because of potential missing values), let n denote the number of non zero observed integers i.e. the number of elements in $\mathscr{S}$. Then

$$\tau < \frac{\hat{x}+1}{\hat{n}} \leq \frac{\hat{x}+1}{n}.$$

Consequently, $\frac{\hat{x}+1}{n}$ is an estimate of τ with precision on the order of $1/n$ (without missing values). As it uses the whole data, it is usually more precise than the previous bound. We will use this in the rest of the paper to restrict the search space for τ.

Let us compare the previous bounds on an example. Suppose that $\tau = 1.32$ and $\mathscr{S} = \{1,2,3,5,6,7,9,10,11,13\}$.
As there are consecutive integers in $\mathscr{S}$ we obtain $\tau < 2$.
Using the interval $[\![5,7]\!]$, we obtain $\tau < 1 + 1/2$.
Using the interval $[\![9,11]\!]$, we obtain $\tau < 1 + 1/2$ as well.
The density upper bound is $\tau < 14/10$.

Remark We only provided upper bounds in this section because lower bounds can only be deduced from the integers that cannot be generated in the model. These are difficult to distinguish from missing values, which are integers that can be generated in the model, but do not appear in the set $\mathscr{S}$ of observed integers.

3 Compatible Values

The upper bounds that we proposed in the previous section are easy to compute but rather poor because they only take into account the proportion of observed integers. In this section, we describe an algorithm with higher computational load but which can leverage the information in the location of each observed integer in the data set.

3.1 Lattices of Integers

In the observation model $X = \lfloor \tau N \rfloor$ where N is integer valued, only specific integers can be generated. Given a strictly positive real number t, let us define the set of

possible values for x as the *lattice associated to* t, i.e. the infinite set of integers $\mathscr{S}_t = \{x = \lfloor tk \rfloor, k \in \mathbb{N}\}$. The set of observed integers $\mathscr{S}$ is also called the *empirical lattice*.

With infinitely many observations, the parameter τ is completely characterized by the empirical lattice as the following proposition shows. This justifies that $\mathscr{S}$ is sufficient information for estimating τ.

Proposition 1 (Equivalence between lattices and numbers) *In the distinguishible case, let* t_1 *and* t_2 *denote two real numbers such that* $t_1 \geq 1$ *and* $t_2 \geq 1$. *Then* $\mathscr{S}_{t_1} = \mathscr{S}_{t_2}$ *if and only if* $t_1 = t_2$.

Proof Obviously, if $t_1 = t_2$ then $\mathscr{S}_{t_1} = \mathscr{S}_{t_2}$. Let us prove the converse, i.e. $\mathscr{S}_{t_1} = \mathscr{S}_{t_2}$ implies $t_1 = t_2$ or equivalently if $t_1 \neq t_2$ then $\mathscr{S}_{t_1} \neq \mathscr{S}_{t_2}$. Suppose that $t_1 < t_2$. There exists $n \in \mathbb{N}$ such that $\lfloor t_1 n \rfloor < \lfloor t_2 n \rfloor$. Either $\lfloor t_2 n \rfloor \notin \mathscr{S}_{t_1}$, in which case $\mathscr{S}_{t_1} \neq \mathscr{S}_{t_2}$, or $\lfloor t_2 n \rfloor = \lfloor t_1 n_1 \rfloor$ with $n_1 > n$. In the latter case, distinguishibility implies that there are strictly more elements in $\mathscr{S}_{t_1} \cap A$ than in $\mathscr{S}_{t_2} \cap A$ where A denotes the set of integers $[\![0, \lfloor t_2 n \rfloor]\!]$.

3.2 The Set of Compatible Values

Proposition 1 is not sufficient for estimating τ because in practice we only observe a finite set $\mathscr{S} \subsetneq \mathscr{S}_\tau$. Consequently we define the notion of compatible lattices and equivalently compatible values. For any positive real t, we say that t is *compatible* with the data if $\mathscr{S} \subset \mathscr{S}_t$. Likewise, for any two sets A and B, A is compatible with B if $B \subset A$. Being compatible with the data set is a necessary condition for a valid estimator of τ.

The set of values that are compatible with the infinite lattice $\mathscr{S}_\tau$ is adequate for estimating τ because of the following proposition.

Proposition 2 τ *is the largest real number in* $\mathscr{C}(\mathscr{S}_\tau)$.

Proof τ is a compatible value, we only have to show that it is the largest.
Let u denote a real number greater than τ, and let α denote a positive real number such that $\tau < \tau + \alpha < u$. We will prove that u is not compatible with τ by constructing an element in $\mathscr{S}_\tau$ that cannot be in $\mathscr{S}_u$.
Let a denote a positive integer such that $a > \frac{1}{\alpha}$, and $n = \lfloor \tau a \rfloor$.
Suppose that $\mathscr{S}_\tau \subset \mathscr{S}_u$ then n belongs to $\mathscr{S}_u$, and there exists a positive integer a' such that $n = \lfloor ua' \rfloor$.
$a' \geq a$ because in the distinguishible case, a and a' correspond to their indices in the sets $\mathscr{S}_\tau$ and $\mathscr{S}_u$ and $\mathscr{S}_\tau \subset \mathscr{S}_u$.
Moreover, as $\tau \leq u$ we have $\lfloor \tau a \rfloor \leq \lfloor ua \rfloor \leq \lfloor ua' \rfloor$. For all three terms to be equal to n in the distinguishible case requires that $a = a'$.
Consequently, both τ and u lie in the interval $[\frac{n}{a}, \frac{n+1}{a}[$. As a result, $|\tau - u| \leq \frac{1}{a}$ which contradicts $a > \frac{1}{\alpha}$.

The set $\mathscr{C}(\mathscr{S}_\tau)$ has an intricate structure. It contains the positive real numbers smaller than 1 and the harmonics $\left\{\frac{\tau}{k}, k \in \mathbb{N}^*\right\}$, but these are not the only values. For example, $4/3$ is compatible with 2 because every even integer can be written as $\lfloor k \times 4/3 \rfloor$, $k \in \mathbb{N}$. Indeed, let k be an even integer. Either k is a multiple of 4, in which case $k = 4i = \lfloor \frac{4}{3} \times 3i \rfloor$, or $k = 4i+2 = \lfloor \frac{4}{3} \times (3i+2) \rfloor$.

3.3 Estimation with a Finite Lattice

In practice, the empirical lattice is finite and can contain missing values and outliers. We say that an integer is *missing* from $\mathscr{S}$ when it is in the theoretical lattice $\mathscr{S}_\tau$, smaller than $\hat{x} = \max \mathscr{S}$, but not in $\mathscr{S}$. The set of compatible values with $\mathscr{S}$ is a finite union of intervals and compatible values are never isolated. As the data contains less information, the true parameter τ is not the supremum of $\mathscr{C}(\mathscr{S})$, but it is still maximal in the following sense.

Proposition 3 *The set of compatible values $\mathscr{C}(\mathscr{S})$ contains exactly $]0,1]$ and intervals of length at least $1/\hat{x}^2$ where $\hat{x} = \max \mathscr{S}$. In particular, if there are no outliers or missing values in $\mathscr{S}$ then τ belongs to the interval $[a,b[$ such that $b = \sup \mathscr{C}(\mathscr{S})$.*

The proof is based on the following two lemmas.

Lemma 1 *The set of compatible values contains exactly $]0,1]$ and a finite number of intervals of the form $[a,b[$ of length at least $1/\hat{x}^2$ where $\hat{x} = \max \mathscr{S}$.*

Proof Let $t > 1$ denote a compatible value. For each observed value $x \in \mathscr{S}$, there exists an integer n such than $x = \lfloor tn \rfloor$. Consequently, t verifies $t \in [\frac{x}{n}, \frac{x+1}{n}[$. The intersection of the constraints $t \in [\frac{x}{n}, \frac{x+1}{n}[$ for all $x \in \mathscr{S}$ is an interval $t \in [\frac{x_1}{n_1}, \frac{x_2}{n_2}[$. All values $t \in [\frac{x_1}{n_1}, \frac{x_2}{n_2}[$ verify all of the constraints and are thus compatible. The length of this interval is $\frac{x_2}{n_2} - \frac{x_1}{n_1}$ which is at least $1/(n_1 n_2)$. In the distinguishible case, $n_1 < \max \mathscr{S}$ and $n_2 < \max \mathscr{S}$, which implies that the length is at least $1/(\max \mathscr{S})^2$.

Lemma 2 *Let t be a positive real number that is compatible with the empirical lattice. Then $t < \dfrac{\hat{x}+1}{n}$.*

Proof This follows directly from the upper bounds in Section 2. See Proposition 9 in the Appendix.

To complete the proof, it suffices to show that τ belongs to the largest interval.

Proof There are only finitely many intervals of length at least $1/\hat{x}^2$ in $[0, \frac{\hat{x}+1}{n_{max}}]$, so there exists such an interval $[a,b[$.
Let $\mathscr{N}$ denote the set $\mathscr{N} = \{n | \lfloor n\tau \rfloor \in \mathscr{S}\}$, i.e. the set of values for N that generate $\mathscr{S}$. τ belongs to a certain interval $[a',b'[$ which is the intersection of the constraints $\tau \in [\frac{x}{n}, \frac{x+1}{n}[$, for all $x = \lfloor n\tau \rfloor$ in $\mathscr{S}$. We show that $b' = b$, i.e. no positive real is both greater than b' and compatible. Let t such that $b' < t$. For all $n \in \mathscr{N}$, $\lfloor n\tau \rfloor \leq \lfloor nt \rfloor$. As $t \notin [a',b'[$, t breaks at least one of the constraints, that is to say, there is an integer x in $\mathscr{S}$ such that $x = \lfloor n\tau \rfloor < \lfloor nt \rfloor$. x is skipped in $\mathscr{S}_t$ and thus t is not compatible.

The previous proposition suggests that it suffices to find the largest compatible interval to estimate τ, and this is our proposed estimator $\tilde{\tau}$. More precisely, the set $\mathscr{C}(\mathscr{S}) = \cup_{j=1}^{J} [a_j, b_j[$ is a union of J intervals, with (a_j) and (b_j) increasing sequences, then

$$\tilde{\tau} = \frac{a_J + b_J}{2}.$$

We use the following algorithm to compute $\tilde{\tau}$. This also computes the mapping $x = \lfloor \tau n \rfloor \mapsto n$ and the precision.

- compute the set of observed values by sorting the data set and removing multiple occurences
- compute the upper bound $\tau < B = \dfrac{\hat{x}+1}{n}$ where $n = \text{card}\mathscr{S}$
- find an approximation of the largest compatible value t by testing the compatibility of the real numbers $t_k = B - \dfrac{k}{\hat{x}^2}$
- deduce the indexes from t, that is to say for all $x \in \mathscr{S}$, find i such that $x = \lfloor ti \rfloor$
- compute the interval $[a, b[$ as the intersection of the constraints $t \in \left[\dfrac{x}{i}, \dfrac{x+1}{i} \right[$, for all x in $\mathscr{S}$
- return $\dfrac{a+b}{2}$ as an estimator for τ

3.4 Properties of the Estimator

According to the previous results, the estimator performs well when there are no missing values or outliers. Its precision is $(b-a)/2$ and can be computed inside the algorithm. The precision is at least $1/\hat{n}$, but depending on the value of τ it can reach a precision on the order of $1/\hat{n}^2$. In all cases, the precision is better than the density bound, and there is a lower bound.

If there are missing values or outliers, the algorithm may find an interval of compatible values that does not contain τ. For example, if the dataset is $\{0,2,4,6,8\}$, a reasonable estimator would answer 2 and not $\tau = 4/3$ with missing values 1 and 5. In practice, such cases are rare, and are related to arithmetic properties of the set $\mathscr{S}$. However, the largest compatible value is never an erroneous answer to the problem. It is a parcimonious answer in the sense that it is the smallest lattice which may explain the dataset.

The estimator is optimal in the sense that the algorithm finds an interval of positive real numbers that are all plausible. Given a dataset $(x_1 = \lfloor \tau\, i_1 \rfloor, \ldots, x_n = \lfloor \tau\, i_n \rfloor)$ of size n, there is an interval of compatible values that can generate $(x_1, \ldots, x_n)$ from the same realization $(i_1, \ldots, i_n)$ of N. Let $[a_J, b_J[$ with $b_J = \sup \mathscr{C}(\mathscr{S})$, the following proposition holds.

Proposition 4 *Given a realization $(i_1,\dots,i_n)$ of N, all values in $[a_J,b_J[$ generate the same data set $(x_1,\dots,x_n)$, i.e.*

$$\forall t \in [a_J,b_J[,\ \forall j \in [\![1,n]\!],\ x_j = \lfloor \tau\, i_j \rfloor = \lfloor t\, i_j \rfloor$$

Proof As in the proof of Proposition 3, $[a_J,b_J[$ is the intersection of the constraints $x_j = \lfloor t\, i_j \rfloor$.

The data set does not contain enough information to distinguish the values in $[a_J,b_J[$. In particular, even if the realization $(i_1,\dots,i_n)$ is given, then the values are not distinguishible. Note that if x_0 is known not to be in $\mathscr{S}_\tau$, then for all integers i, $\tau \geq \frac{x_0+1}{i}$ or $\tau < \frac{x_0}{i}$. These inequalities are not informative because they are already contained in $x = \lfloor ti \rfloor, \forall x \in \mathscr{S}$.

The program is quite fast. First because is relies only on the set $\mathscr{S}$ which is much smaller than the dataset when τ is near 1 and N is independent identically distributed, because repeats of N are discarded. As the following proposition shows, with few missing values, the density bound is precise and the algorithm is quicker. All compatible values can be retrieved by testing $B\hat{x}^2$ numbers.

Proposition 5 *If there are no missing values, the largest compatible value is found after at most $\dfrac{\hat{x}^2}{\hat{n}} \simeq \tau\hat{x}$ steps. With a small number of missing values $k \ll \hat{n}$, the number of steps is on the order of $\tau^2\hat{x}\left(k+\dfrac{1}{\tau}\right)$ where $k = \hat{n} - \mathrm{card}\mathscr{S}$ is the number of missing values.*

Proof The procedure begins at $B = \frac{\hat{x}+1}{n}$, ends before $\frac{\hat{x}}{\hat{n}}$ because $\tau \geq a \geq \frac{\hat{x}}{\hat{n}}$, and proceeds in steps of length $1/\hat{x}^2$. Consequently, there are at most $C = \hat{x}^2(\frac{\hat{x}+1}{n} - \frac{\hat{x}}{\hat{n}})$ steps. Let $k = \hat{n} - \mathrm{card}\mathscr{S}$ denote the number of missing values. We make the following three approximations: $k \ll \hat{n}$, $1 \ll \hat{x}$ and $\tau \simeq \frac{\hat{x}}{\hat{n}}$. Then $C = \hat{x}^2(\hat{x}+1)\left(\frac{k}{\hat{n}\mathrm{card}\mathscr{S}} + \frac{1}{\hat{n}(\hat{x}+1)}\right)$ which can be approximated by $C \simeq \tau^2\hat{x}(k+\frac{1}{\tau})$.

Testing for the compatibility of a real t is linear in the size of $\mathscr{S}$, so the whole procedure is at most quadratic. The full set of compatible values can be obtained in cubic time.

4 Results and Discussion

4.1 Compatible Values Estimator

Figure 1 illustrates the compatible values estimator on a simulated dataset. The dataset $\{6,6,11,5,3,5,2,6,5,13,2,7,7,7,6\}$ is obtained from the observation model $X = \lfloor 1.32 * N \rfloor$ where N is distributed according to a Poisson random variable with mean 5.5. It is first reduced to the lattice $\mathscr{S} = \{2,3,5,6,7,11,13\}$ and is shown at the bottom.

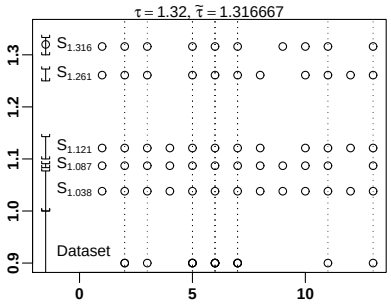

Fig. 1 Comparison of a few compatible lattices and the dataset.

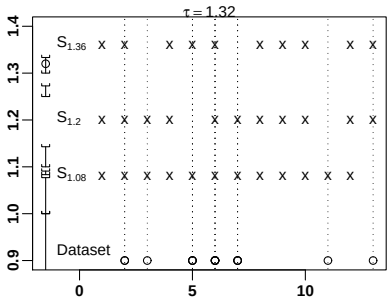

Fig. 2 Comparison of the dataset and a few lattices that are not compatible.

The vertical axis represents values of τ. The set of compatible values is composed of several intervals and represented on the left. For each interval, we select one compatible value t and represent the lattice $\mathscr{S}_t$. All reals in the same interval generate the same lattice, up to $\max(\mathscr{S})$.

For comparison, Figure 2 displays $\mathscr{S}_t$ for several values that are not compatible with the data. For example, 5 and 11 are in the dataset but not in $\mathscr{S}_{1.2}$.

Two sources of variation affect the estimate $\tilde{\tau}$. First, the estimator is not perfect because the dataset is finite. Second, the data set $\mathscr{S}$ is random. Figure 3 shows the performance of the estimator with a fixed dataset ($N \in [\![1,10]\!]$) for several values of τ. The intervals shown correspond to the intervals in $\mathscr{C}(\mathscr{S})$ that contain the largest compatible value.

We can see that the precision of the estimator varies with τ. Only the range $[1,2]$ is shown because the precision only depends on the rest $\tau - \lfloor \tau \rfloor$ modulo 1. Consequently, the absolute precision is roughly constant, whereas the relative precision is $O\left(\frac{1}{\tau}\right)$. With small quantization error ($\tau \gg 1$) the estimation problem is easier.

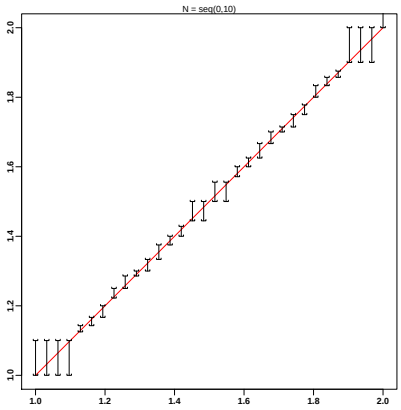

Fig. 3 Length of the maximal interval for several values of τ with $N \in [\![1,10]\!]$.

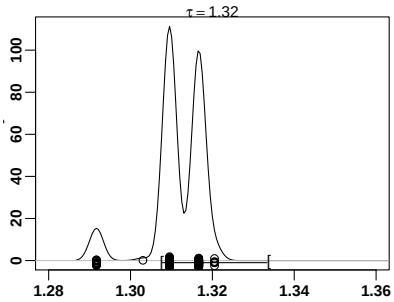

Fig. 4 Kernel density estimate of the distribution of the compatible values estimator on a random dataset.

Figure 4 shows the distribution of $\tilde{\tau}$ when the dataset $\mathscr{S}$ is the result of 15 samples of $X = \lfloor \tau N \rfloor$ where $\tau = 1.32$ and N is distributed according to a Poisson random variable with mean 5.5. The distribution of the estimator value is obtained from the 200 repeats shown in the bottom of the plot thanks to a kernel estimate, even if the distribution is a sum of Dirac point masses. The interval shows the interval obtained with the (complete) dataset $\{\lfloor 1.32 * n \rfloor, n \in [\![1,13]\!]\}$.

4.2 The Double Poisson Family

In this section, we briefly recall the results from [5], and deduce an estimator for our model. Let $g_\mu(y) = e^{-\mu}\mu^y/y!$ denote the distribution function of a Poisson random variable with mean μ. The double Poisson distribution with parameters θ, μ is defined as:

$$\begin{aligned} f_{\theta,\mu}(y) &= c(\theta,\mu)\theta^{1/2}\left\{g_\mu(y)\right\}^\theta \left\{g_y(y)\right\}^{1-\theta} \\ &= c(\theta,\mu)\left(\theta^{1/2}e^{-\theta\mu}\right)\left(\frac{e^{-y}y^y}{y!}\right)\left(\frac{e\mu}{y}\right)^{\theta y} \end{aligned}$$

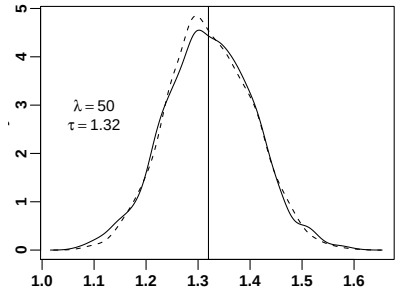

Fig. 5 Kernel density estimate of the distribution of the double poisson estimate of τ with flooring noise (solid line) and without (dotted line).

where c is a normalization constant.

Maximum Likelihood Estimation leads to the following estimators. Let $(Y_1,\ldots,Y_n)$ be independent identically distributed random variables with distribution $f_{\theta,\mu}$, then

$$\hat{\mu} = \frac{1}{n}\sum_{i=1}^{n} Y_i$$
$$\hat{\theta} = \frac{n}{2\sum_{i=1}^{n} I(Y_j,\mu)}$$

where $I(\mu_1,\mu_2) = \mu_1(\log(\mu_1) - \log(\mu_2)) - (\mu_1 - \mu_2)$.

Let $Y_{\theta,\mu}$ be a random variable with distribution function $f_{\theta,\mu}$, then according to [5] $Y_{\theta,\mu}$ has approximately the same distribution as X/θ where X is Poisson distributed with mean $\mu\theta$. With Poisson distribution for the ion counts, our observation model becomes $Y = \lfloor \tau N \rfloor$ where N is Poisson distributed with mean λ. Consequently, estimates for τ and λ can be deduced from $\hat{\theta}$ and $\hat{\mu}$ with the following relations:

$$\hat{\tau} = \frac{1}{\hat{\theta}}$$
$$\hat{\lambda} = \hat{\mu}\hat{\theta}$$

The double Poisson distribution is a correct approximation of the distribution of X/θ for large μ, and in that case, $\hat{\tau}$ and $\hat{\lambda}$ are unbiased estimates of τ and λ. The standard deviation of $\hat{\tau}$ is $\frac{\tau\sqrt{2}}{\sqrt{n}}$. Figure 5 shows the distribution of $\hat{\tau}$ with flooring noise, i.e. in the model $Y = \lfloor \tau N \rfloor$ (solid line) and without flooring noise in the model $Y = \tau N$ (dotted line). The plot was generated with 2000 repeats with data sets of size 500. We observe a large standard deviation compared with the compatible values estimator, even on a much larger data set. With large values of μ, truncation has limited effect on the estimate.

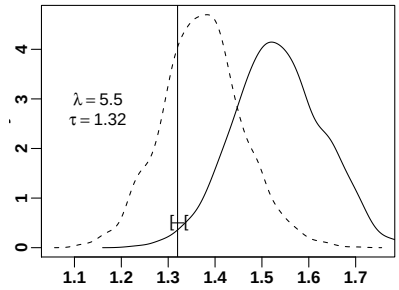

Fig. 6 Kernel density estimate of the distribution of the double poisson estimate of τ with flooring noise (solid line) and without (dotted line).

For modeling rare ion count events, we need to study the estimators with small values of λ and τ. In that case, $\hat{\tau}$ is strongly biased for both models as shown on Figure 6. This implies that the approximation is not suited to this range of parameters, and that the flooring noise makes a significant difference there. Figure 6 was generated using 2000 repeats with data sets of size 500. For comparison, we show the optimal interval obtained by the compatible values estimator on the data set $\{\lfloor 1.32 * n \rfloor, n \in [\![1,13]\!]\}$.

4.3 Fourier Estimator

From the set $\tau\mathbb{N}$ we can construct the signal $f : t \mapsto \sum_{k\in\mathbb{N}} \delta(x-\tau k)$ where δ denotes the Dirac function, that is to say a periodic series of pulses. The period τ may thus be estimated using Fourier transform. Likewise, we define the estimator $1/\tau_F$ as the maximum of the Fourier transform of the quasi-periodic signal $f : t \mapsto \sum_{k\in\mathbb{N}} \delta(x - \lfloor \tau k \rfloor)$.

As τ can be seen as a quasi-period, our estimation problem is closely linked to the "harmonic retrieval problem". Many approaches have been proposed in that domain and the main focus is on the estimation of the Power Spectral Density [6]. However, the signal is usually perturbed by additive noise whereas in this paper we consider a distortion of the time axis.

We use the following algorithm:

- sample the signal f at the points $x_i = i$ for the integers i in $[\![0, \max(\mathscr{S})]\!]$
- compute the Discrete Fourier Transform
- compute an upper bound using Proposition 9 : $\tau < B = \dfrac{\hat{x}+1}{n}$
- find the frequency with highest absolute Fourier coefficient
- return the corresponding period (inverse of the frequency)

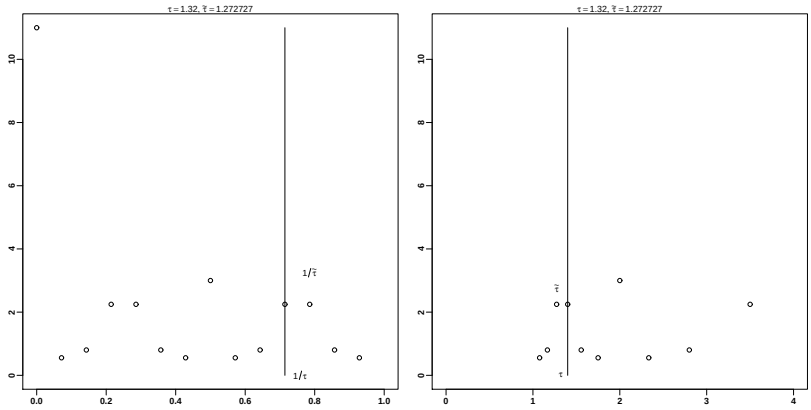

Fig. 7 Fourier transform of the quasi-periodic signal, in Fourier space (left) and period space (right). The vertical line shows the upper bound from Proposition 9.

This estimator has a precision that corresponds to the sampling rate in time space around the true value. In the Fourier space, the sampling rate is uniform with steps of length $1/\max(\mathscr{S})$ which is equivalent to $1/(\tau \times \hat{x})$. In the time space, as $P = 1/f$, then $\Delta P = -\Delta f/f^2$ and the sampling rate is non uniform. For $f = 1/\tau$ we obtain the precision of the Fourier estimator as $\tau/\hat{x}$. This suggests that the precision decreases with τ. However, the signal frequency $1/\tau$ is near $\hat{x}/\max(\mathscr{S})$ which is one of the sampling points. As a result, in practice, the absolute precision is on the order of $1/\hat{x}$ and independent of τ.

Figure 7 shows in the frequency and period space the Fourier transform of the quasi-periodic signal obtained from the dataset $N \in [\![1,10]\!]$. The vertical line corresponds to the upper bound from Proposition 9.

Remark When oversampling by a factor k, i.e. sampling at the points $x_i = \dfrac{i}{k}$ for the integers i in $[\![0, \max(\mathscr{S}) \times k]\!]$, the harmonics of 1 Hz increase in magnitude. Therefore it is necessary to weed out the frequencies above 1 Hz in the distinguishible case. Moreover, oversampling increases the maximum frequency that can be represented in the Fourier space and does not improve the precision of the estimator.

On a random dataset, the Fourier estimator suffers greatly from missing values. Figure 8 shows the distribution of τ_F with 200 simulations and a dataset of size 15 where N is distributed according to a Poisson random variable with mean 5.5. The precision of the estimator is much worse than the compatible values estimator (see the plotted interval). The Fourier estimate τ_F is compatible with the dataset in only about 1% of the simulations.

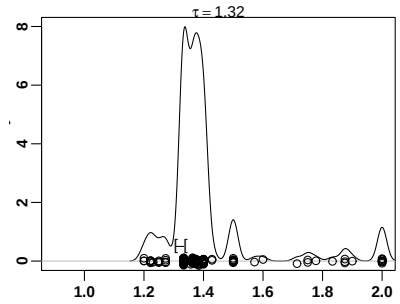

Fig. 8 Kernel density estimate of the distribution of the Fourier estimator.

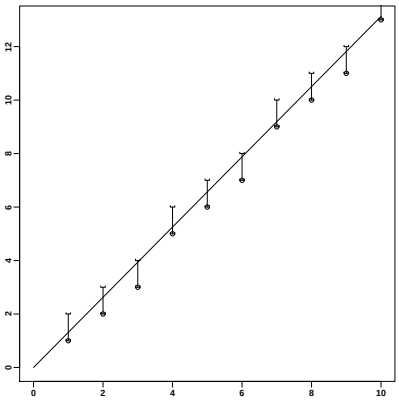

Fig. 9 The linear regression estimator on a dataset without missing values.

4.4 Linear Regression Estimator

The observation model $X = \lfloor \tau N \rfloor$ may be written $X = \tau N + \varepsilon$ where ε is an error term. Even if ε is not Gaussian, linear regression can yield a reasonable estimate of the regression coefficient τ as Figure 9 shows.

We use the following algorithm:

- compute the empirical lattice $\{x_i\}$ by sorting and removing duplicates in the dataset
- compute the indexes $\{n_i\}$ according to the sorting index
- fit a regression line of the form $x_i = an_i + 0.5$
- return a

Figure 9 shows the linear regression estimator on the dataset $\mathscr{S} = \lfloor \tau [\![1,10]\!] \rfloor$ (no missing values). For each element in the dataset, if the regression line intersects the length 1 interval then the estimate is compatible with the data point.

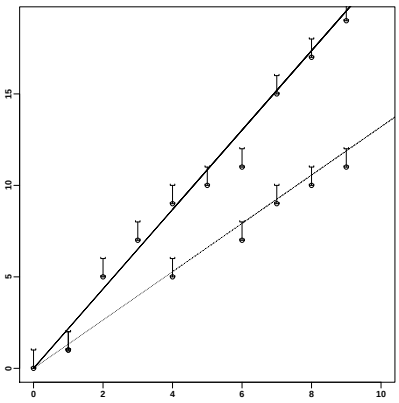

Fig. 10 The linear regression estimator in the case of missing values. The compatible values estimator is shown in dotted line.

Note that the truncation error is not centered. Consequently, we compute the regression coefficient in the the model $X+0.5=\tau N+\varepsilon$. For the same reason, the regressors are below the regression line.

The main difficulty in the linear regression is that the values of the regressor variable N are unknown. In the distinguishible case, it is possible to reconstruct them when there are no missing values, i.e. $\mathscr{S}_\tau \cap [\![0,n]\!] = \mathscr{S}$ where $n = \max \mathscr{S}$. Otherwise, the regressors will be shifted and that affects strongly the estimate. Figure 10 shows such a case. The regressors inferred in the linear regression estimator and the regression line are shown in solid line. For comparison, the true regressors are displayed in dotted line. The compatible values estimator finds the true regressors and its regression line is shown in dotted line.

5 Conclusion

In the observation model $X = \lfloor \tau N \rfloor$, the parameter τ can be reliably estimated independently from N. This allows the full recovery of the statistics of N prior to modeling. The structure of N may then be studied at length afterwards.

The estimator based on compatible values is optimal and reasonably quick. It is resistant to missing values in practice, and in the worst case returns an acceptable (parcimonious) answer without hypotheses on the law of N.

Unfortunately, this estimator only takes into account truncation noise, and yields poor results on real data. We are currently pursuing an extension of the model that mixes electronic noise and truncation effects.

Compared to the other three estimators, the compatible values estimator performs much better but also more slowly. The Double Poisson Family is simply not a suitable model in our range of parameters, but there is room for improvement for the other estimators. For example, the main difficulty in the linear regression is computing the indexes. With some knowledge about the law of N, quantile regression could be applied.

The Fourier estimator suggests a strong relationship with the harmonic retrieval problem, although the signal is not periodic. Although the truncation error considered in this paper is very different from Gaussian errors usually considered in harmonic retrieval, some algorithms from that field may make a better compromise between speed and precision for the current problem.

References

1. R. Aebersold and M. Mann. Mass spectrometry-based proteomics. *Nature*, 422:198–207, 2003.
2. M. Anderle, S. Roy, H. Lin, C. Becker, and K. Joho. Quantifying reproducibility for differential proteomics: noise analysis for protein liquid chromatography-mass spectrometry of human serum, 2004.
3. V.P. Andreev, T. Rejtar, H.S. Chen, E.V. Moskovets, A.R. Ivanov, and B.L. Karger. A universal denoising and peak picking algorithm for LC-MS based on matched filtration in the chromatographic time domain. *Anal Chem*, 75(22):6314–26, 2003.
4. A. Antoniadis, J. Berruyer, and A. Filhol. Estimation semi-paramétrique dans les familles doublement poissonniennes et application aux spectres de diffraction. *Revue de Statistique Appliquée*, 47(3):57–80, 1999.
5. B. Efron. Double exponential families and their use in generalized linear regression. *J. AM. STAT. ASSOC.*, 81(395):709–721, 1986.
6. M.H. Hayes. *Statistical Digital Signal Processing and Modeling*. John Wiley & Sons, Inc. New York, NY, USA, 1996.
7. A.N. Krutchinsky and B.T. Chait. On the nature of the chemical noise in MALDI mass spectra. *Journal of the American Society for Mass Spectrometry*, 13(2):129–134, 2002.
8. CS Lane. Mass spectrometry-based proteomics in the life sciences. *Cellular and Molecular Life Sciences (CMLS)*, 62(7):848–869, 2005.
9. G. Siuzdak. *Mass spectrometry for biotechnology*. Academic Press San Diego, Calif, 1996.
10. W. Windig, J.M. Phalp, and A.W. Payne. A noise and background reduction method for component detection in liquid chromatography/mass spectrometry. *Anal. Chem*, 68(20):3602–3606, 1996.
11. J.L. Wiza. Microchannel Plate Detectors. *Nucl. Instrum. Methods*, 162:587–601, 1979.

Appendix

Proposition 6 *The mapping $x \mapsto \lfloor \tau x \rfloor$ is injective if and only if $\tau \geq 1$.*

Proof If $\tau = 1$, the mapping is the identity function. Suppose $\tau > 1$ and let n_1 and n_2 denote two (positive) integers such that $n_1 < n_2$. Then $\tau n_2 - \tau n_1 > \tau > 1$ and $\lfloor \tau n_2 \rfloor > \lfloor \tau n_1 \rfloor$. When $\tau < 1$, the mapping is not injective because $\lfloor \tau \times 1 \rfloor = \lfloor \tau \times 0 \rfloor = 0$.

Upper Bounds on τ

Proposition 7 (Any two observations) *Let x and y be two distinct elements of the set $\mathscr{S}$ of observed values. Then $\tau < 1 + |x - y|$.*

Proof Let i and j be the values of N corresponding to x and y i.e. $x = \lfloor \tau i \rfloor$ and $y = \lfloor \tau j \rfloor$. Then we have the inequalities: $x \leq \tau i < x + 1$, $y \leq \tau j < y + 1$, and thus $\tau(j - i) < y - x + 1$. Assuming $x < y$, we obtain $\tau < \frac{y-x+1}{j-i} < y - x + 1$.

Proposition 8 (Observed intervals) *Let $[\![x,y]\!]$ denote the set of integers between x and y. If $[\![x,y]\!]$ is a subset of $\mathscr{S}$, then $\tau < 1 + \dfrac{1}{y-x}$.*

Proof Using the same notations as in the proof of Proposition 7, $\tau < \frac{y-x+1}{j-i} < \frac{y-x}{j-i} + \frac{1}{j-i} < 1 + \frac{1}{j-i}$ because in the distinguishible case the number of elements in $[\![x,y]\!]$ is $y - x + 1 = j - i + 1$.

Proposition 9 (Density Upper Bound) *Let $\hat{x} = \lfloor \tau \hat{n} \rfloor$ denote the largest integer in $\mathscr{S}$. Then $\tau < \dfrac{\hat{x}+1}{\hat{n}}$. When $\hat{n}$ is unknown (because of potential missing values), let n denote the number of non zero observed integers. Then $\tau < \dfrac{\hat{x}+1}{\hat{n}} \leq \dfrac{\hat{x}+1}{n}$.*