Electronic Journal of Statistics
Vol. 10 (2016) 683–705
ISSN: 1935-7524
DOI: 10.1214/16-EJS1118

Estimation of a delta-contaminated density of a random intensity of Poisson data

Daniela De Canditiis[†] **and Marianna Pensky**[‡]

I.A.C. "Mauro Picone"
National Research Council
via dei Taurini, 19 Rome, Italy
and
Department of Mathematics and Statistics
University of Central Florida
Orlando FL 32816-1353, USA
e-mail: d.decanditiis@iac.cnr.it; marianna.pensky@ucf.edu

Abstract: In the present paper, we constructed an estimator of a delta contaminated mixing density function $g(\lambda)$ of an intensity λ of the Poisson distribution. The estimator is based on an expansion of the continuous portion $g_0(\lambda)$ of the unknown pdf over an overcomplete dictionary with the recovery of the coefficients obtained as the solution of an optimization problem with Lasso penalty. In order to apply Lasso technique in the, so called, prediction setting where it requires virtually no assumptions on the dictionary and, moreover, to ensure fast convergence of Lasso estimator, we use a novel formulation of the optimization problem based on the inversion of the dictionary elements.

We formulate conditions on the dictionary and the unknown mixing density that yield a sharp oracle inequality for the norm of the difference between $g_0(\lambda)$ and its estimator and, thus, obtain a smaller error than in a minimax setting. Numerical simulations and comparisons with the Laguerre functions based estimator recently constructed by [8] also show advantages of our procedure. At last, we apply the technique developed in the paper to estimation of a delta contaminated mixing density of the Poisson intensity of the Saturn's rings data.

MSC 2010 subject classifications: Primary 62G07, 62C12; secondary 62P35.
Keywords and phrases: Mixing density, Poisson distribution, empirical Bayes, Lasso penalty.

Received August 2015.

Contents

*The authors would like to thank Dr. Joshua Colwell for helpful discussions and for providing the data.
†Supported by "Italian Flagship Project Epigenomic" (http://www.epigen.it/).
‡Supported by National Science Foundation (NSF), grant DMS-1407475.

1. Introduction

Poisson-distributed data appear in many contexts. In the last two decades a large amount of effort was spent on recovering the mean function in the Poisson regression model. In this set up, one observes independent Poisson variables $Y_1, \cdots, Y_n$ where $\mathbb{E}Y_i = \lambda_i = f(i/n)$, $i = 1, \cdots, n$. Here, f is the function of interest which is assumed to exhibit some degree of smoothness. The difficulty in estimating f on the basis of Poisson data stems from the fact that the variances of the Poisson random variables are equal to their means and, hence, do not remain constant as f changes its values. Estimation techniques are either based on variance stabilizing transforms ([3], [10]), wavelets ([1], [2], [11]), Haar frames [13]) or Bayesian methods ([14] and [19]).

The fact that the variance of a Poisson random variable is equal to its mean serves as a common and reliable test that data in question are indeed Poisson distributed. However, in many practical situations, although each of the data value $Y_i \sim Poisson(\lambda_i)$, $i = 1, \cdots, n$, the overall data do not have the Poisson distribution. This is due to the fact that the consecutive values of λ_i are so different from each other that f is not really a function. In this case, in order to account for the extra-variance, it is usually reasonable to assume that λ itself is a random variable with an unknown probability density function g which needs to be estimated.

In particular, below we consider the following problem. Let λ_i, $i = 1, \cdots, n$, be independent random variables that are not observable and have an unknown pdf $g(\lambda)$. One observes variables $Y_i|\lambda_i \sim Poisson(\lambda_i)$, $i = 1, \cdots, n$, that, given λ_i, are independent. Our objective is to estimate $g(\lambda)$, the so called *mixing density*, on the basis of observations $Y_1, \cdots, Y_n$. Here, g can be viewed as the prior density of the parameter λ, so that the model above reduces to an empirical Bayes model where the prior has to be estimated from data.

Estimation of the prior density of the parameter of the Poisson distribution has been considered by several authors. For example, [15] suggested nonparametric maximum likelihood estimator, [20] and [21] studied estimators based on Laguerre polynomials, [22] considered smoothing kernel estimators and [12] investigated Fourier series based estimators of g. All papers listed above provided upper bounds for the mean integrated squared error (MISE); [22] and [12] also presented lower bounds for the MISE over smoothness classes. The common feature of all these estimators is that the convergence rates are very low. In particular, if $n \to \infty$, both [22] and [12] obtained convergence rates of the form

$(\ln n/\ln\ln n)^{-2\nu}$ where ν is the parameter of the smoothness class to which g belongs. The latter seem to imply that there is no hope for accurate estimation of the mixing density g unless the sample sizes are extremely high. On a more positive note, in a recent paper, [8] considered an estimator of g based on expansion of g over the orthonormal Laguerre basis. They showed that if the Laguerre coefficients of g decrease exponentially, then the resulting estimator has convergence rates that are polynomial in n and provided some examples where this happens. Moreover, they proposed a penalty for controlling the number of terms in the expansion and provided oracle inequalities for the estimators of g under various scenarios.

The low convergence rates for the prior density of Poisson parameter are due to the fact that its recovery constitutes a particular case of an ill-posed linear inverse problem. Indeed, let $L^2[0,\infty)$ and ℓ^2 be the Hilbert spaces of, respectively, square integrable functions on $[0,\infty)$ and square integrable sequences. Denote the probability that $Y=l$, $l=0,1,\cdots$, by $P(l)=\mathbb{P}(Y=l)$. Then, introducing a linear operator $Q: L^2[0,\infty)\to\ell^2$, we can present $g(\lambda)$ as the solution of the following equation

$$(Qg)(l)=\int_0^\infty \frac{\lambda^l\, e^{-\lambda}}{l!}g(\lambda)\,d\lambda = P(l),\quad l=0,1,\cdots \tag{1.1}$$

Since exact values of the probabilities $P(l)$ are unknown, they can be estimated by the relative frequencies ν_l, so the problem of recovering g appears as an ill-posed linear inverse problem with the right-hand side measured with error. Solution of equation (1.1) is particularly challenging since g is a function of a real argument while P is an infinite-dimensional vector.

On the other hand, in the last decade a great deal of effort was spent on recovery of an unknown function in regression setting from its noisy observations using overcomplete dictionaries. In particular, if the dictionary is large enough and the function of interest has a sparse representation in this dictionary, then it can be recovered with a much better precision than when it is expanded over an orthonormal basis. Lasso and its versions (see e.g. [5] and references therein) allow one to identify the dictionary elements that guarantee efficient estimation of the unknown regression function. The advantage of this approach is that the estimation error is controlled by the, so called, oracle inequalities that provide upper bounds for the risk for the particular function that is estimated rather than convergence rates designed for the "worst case scenario" of the minimax setting. In addition, if the function of interest can be represented via a linear combination of just a few dictionary elements, then one can prove that it can be estimated with nearly parametric error rate provided certain assumptions on the dictionary hold.

In the present paper, we extend this idea to the case of estimating a mixing density g on the basis of $Y_1,\cdots,Y_n$. However, there is an intrinsic difficulty arising from the fact that the problem above is an ill-posed inverse problem. Currently, one can justify convergence of a Lasso estimator only if stringent assumptions on the dictionary, the, so called, compatibility conditions, are satisfied. In regression set up, as long as compatibility conditions hold, one can

prove that Lasso estimator is nearly optimal. Regrettably, while compatibility conditions may be satisfied for the functions in the original dictionary, they usually do not hold for their images due to contraction imposed by the operator Q. In the present paper, we show how to circumvent this difficulty and apply Lasso methodology to estimation of g. We formulate conditions on the dictionary and the unknown mixing density that yield a sharp oracle inequality for the norm of the difference between $g(\lambda)$ and its estimator and, thus, result in a smaller error than in a minimax setting. Numerical simulations and comparisons with the Laguerre functions based estimator recently constructed by [8] also show advantages of our procedure.

Our study is motivated by the analysis of astronomical data, in particular, the photon counts $Y_i, i = 1, \cdots, n$ that come from sets of observations of the stellar occultations recorded by the Cassini UVIS high speed photometer at different radial points on the Saturn's ring plane. It is well known that the Saturn ring is comprised of particles of various sizes, each on its own orbit about the Saturn. With no outside influences, these photon counts should follow the Poisson distribution, however, obstructions imposed by the particles in the ring cause photon counts distribution to deviate from Poisson. The latter is due to the fact that although, for each $i = 1, \cdots, n$, the photon counts $Y_i \sim Poisson(\lambda_i)$, the values of λ_i, $i = 1, \cdots, n$, are extremely varied and, specifically, are best described as random variables with the unknown underlying pdf $g(\lambda)$.

In addition, if a ring region contains a significant proportion of large particles, those particles can completely block out the light leading to zero photon counts. For this reason, we assume that the unknown pdf g is delta-contaminated, i.e., it is a combination of an unknown mass π_0 at zero and a continuous part, so that $g(\lambda)$ can be written as

$$g(\lambda) = \pi_0 \delta(\lambda) + f(\lambda) \quad \text{with} \quad f(\lambda) = (1 - \pi_0) g_0(\lambda) \tag{1.2}$$

where $g_0(\lambda)$ is an unknown pdf and $\delta(\lambda)$ is the Dirac delta function such that, for any integrable function u one has $\int u(x)\delta(x)dx = u(0)$. Models of the type (1.2) also appear in other applied settings (see, e.g., [16]). However, to the best of our knowledge, we are the first ones to estimate the delta-contaminated density of the intensity parameter of the Poisson distribution. In this setting, we also obtain a sharp oracle inequality for the norm of the difference between $g_0(\lambda)$ and its estimator. We also derive convergence rates for the estimator $\widehat{\pi_0}$ of the mass π_0 at zero. The estimator has also been successfully applied to recovery of delta-contaminated densities of the intensities λ for various sub-regions of the Saturn's rings.

Finally, we should remark on several other advantages of the approach presented in the paper. First, although in the numerical studies of the paper we are using the gamma dictionary, all theoretical and methodological results of the paper are valid for any type of dictionary functions since it is based on the numerical inversion of dictionary elements. Moreover, the method can be used even if the underlying conditional distribution is different from the Poisson. The estimator exhibits no boundary effects and performs well in simulations delivering small errors. Moreover, since we apply the Tikhonov regularization for

recovering the inverse images of the dictionary elements, our estimator can be viewed as a version of the elastic net estimator described in [23].

The rest of the paper is organized as follows. Sections 2 and 3 present, respectively, the method and the algorithm for the estimation of the density function, while Section 4 studies convergence properties of the estimator. Section 5 investigates precision of the estimator developed in the paper via numerical simulations using synthetic data. Section 6 provides application of the technique proposed in the paper to the occultation data for the Saturn's rings. Finally, Section 7 contains the proofs of the statements presented in the paper.

2. The Lasso estimator of the mixing density

In what follows, we assume that $g_0(\lambda)$ in (1.2) can be well approximated by a dictionary

$$\mathcal{D} = \{\phi_k(\lambda), k = 1, \cdots, p\}.$$

In particular, in our simulations and real-life applications we consider the dictionary that consists of gamma pdfs

$$\phi_k(\lambda) = \gamma(\lambda; a_k, b_k) = \frac{\lambda^{a_k - 1} \exp(-\lambda/b_k)}{b_k^{a_k} \Gamma(a_k)}, \quad k = 1, \cdots, p. \tag{2.1}$$

This is a natural choice since, for a fixed $b_k = b$ and $a_k = 1, 2, 3, \cdots$, this dictionary contains various linear combinations of the Laguerre functions and, hence, its span approximates the $L^2[0, \infty)$ space. Therefore, any square integrable function can be approximated by a linear combination of ϕ_k with a small error. On the other hand, using a variety of scales b_k allows one to accurately represent a function of interest with many fewer terms. Nevertheless, all theoretical results in Sections 2, 3 and 4 are valid for an arbitrary dictionary in $L^2[0, +\infty)$.

If π_0 were known, then, using the dictionary, we would estimate g by

$$\hat{g}(\lambda) = \pi_0 \delta(\lambda) + \hat{f}(\lambda) \quad \text{with} \quad \hat{f}(\lambda) = \sum_{k=1}^{p} \widehat{\theta_k} \phi_k(\lambda),$$

where coefficients θ_k, $k = 1, \cdots, p$, are chosen so to minimize the squared L^2-norm

$$\|g - \hat{g}\|_2^2 = \|g - \pi_0 \delta\|_2^2 + \left\| \sum_{k=1}^{p} \theta_k \phi_k \right\|_2^2 - 2 \sum_{k=1}^{p} \theta_k \langle g - \pi_0 \delta, \phi_k \rangle. \tag{2.2}$$

The first term in formula (2.2) does not depend on coefficients θ_k while the second term is completely known. In order to estimate the last term, note that $\langle g - \pi_0 \delta, \phi_k \rangle = \langle g, \phi_k \rangle - \pi_0 \phi_k(0)$. Moreover, if we found functions $\chi_k \in \ell^2$ such that

$$(Q^* \chi_k)(\lambda) = \sum_{i=0}^{\infty} \frac{e^{-\lambda} \lambda^i}{i!} \chi_k(i) = \phi_k(\lambda), \quad \forall \lambda \in (0, +\infty), \tag{2.3}$$

then, it is easy to check that

$$\begin{aligned}\langle g,\phi_k\rangle &= \int_0^{+\infty} g(\lambda)\sum_{i=0}^{\infty}\frac{e^{-\lambda}\lambda^i}{i!}\chi_k(i)d\lambda = \sum_{i=0}^{\infty}\chi_k(i)\int_0^{+\infty} g(\lambda)\frac{e^{-\lambda}\lambda^i}{i!}d\lambda \\ &= \sum_{i=0}^{\infty}\chi_k(i)P(i) = \mathbb{E}\chi_k(Y). \end{aligned} \tag{2.4}$$

Here, $P(l)$ is the marginal probability function

$$P(l) = \mathbb{P}(Y=l) = \pi_0\,\mathbb{I}(l=0) + \sum_{k=1}^{p}\theta_k U_k(l), \quad l=0,1,2,\cdots \tag{2.5}$$

where $\mathbb{I}(l=0)$ is the indicator that $l=0$ and

$$U_k(l) = \int_0^{+\infty}\frac{e^{-\lambda}\lambda^l}{l!}\,\phi_k(\lambda)d\lambda = \frac{\Gamma(l+a_k)}{\Gamma(a_k)\,l!}\,b_k^l(1+b_k)^{-(l+a_k)} \tag{2.6}$$

Hence, $\langle g,\phi_k\rangle$ can be estimated by

$$\widehat{\langle g,\phi_k\rangle} = n^{-1}\sum_{i=1}^{n}\chi_k(Y_i) = \sum_{l=0}^{\infty}\chi_k(l)\nu_l = \langle\chi_k,\nu\rangle, \quad k=1,\dots,p, \tag{2.7}$$

where

$$\nu_l = n^{-1}\sum_{i=1}^{n}\mathbb{I}(Y_i=l), \quad l=0,1,\cdots \tag{2.8}$$

are the relative frequencies of $Y=l$ and $\mathbb{I}(A)$ is the indicator function of a set A.

There is an obstacle to carrying out estimation above. Indeed, for some k solutions $\chi_k(Y)$ of equations (2.3) may not have finite variances or variances may be too high. In particular, this is true for ϕ_k defined by formula (2.1) whenever $b_k<1$. In order to stabilize the variances we use the Tikhonov regularization. In particular, we replace solution $\chi_k=(Q^*)^{-1}\phi_k$ of equation (2.3) by solution $\widetilde{\psi}_{k,\zeta_k}$ of equation

$$(QQ^*+\zeta_k I)\widetilde{\psi}_{k,\zeta_k} = Q\phi_k, \quad \zeta_k>0, \tag{2.9}$$

where operators Q and Q^* are defined in (1.1) and (2.3), respectively, and I is the identity operator, so that, for any f,

$$(QQ^*f)(j) = \sum_{l=0}^{\infty}\binom{j+l}{l}2^{-(j+l+1)}f(l), \quad j=0,1,\dots$$

Observe that $\mathrm{Var}[\widetilde{\psi}_{k,\zeta_k}(Y)]$ is a decreasing function of ζ_k while the squared bias $(\mathbb{E}\widetilde{\psi}_{k,\zeta_k}-\langle g,\phi_k\rangle)^2$ is an increasing function of ζ_k. Denote $\widehat{\zeta_k}$ the unique solution of the following equation

$$\frac{1}{n}\mathrm{Var}[\widetilde{\psi}_{k,\widehat{\zeta_k}}(Y)] = \left(\mathbb{E}\widetilde{\psi}_{k,\widehat{\zeta_k}}-\langle g,\phi_k\rangle\right)^2 \tag{2.10}$$

and replace $\chi_k(Y)$ in (2.7) by

$$\psi_k(Y) = \widetilde{\psi}_{k,\widehat{\zeta}_k}(Y) \quad \text{with} \quad \sigma_k^2 = \mathrm{Var}[\psi_k(Y)]. \tag{2.11}$$

After the values of $\langle g, \phi_k \rangle$, $k = 1, \cdots, p$, are estimated, the only obstacle for minimizing the right hand side of formula (2.2) is that we do not know the values of $\pi_0 \langle \phi_k, \delta \rangle = \pi_0 \phi_k(0)$. Therefore, we choose a dictionary such that $\pi_0 \phi_k(0) = 0$, $k = 1, \cdots, p$. The latter means that in our numerical studies, unless we know that $\pi_0 = 0$, we choose $a_k > 1$ in (2.1).

In order to identify the correct subset of dictionary functions ϕ_k, we introduce a weighted Lasso penalty. In particular, the vector of coefficients $\hat{\boldsymbol{\theta}}$ with components $\hat{\theta}_k, k = 1, \cdots, p$, can be recovered as a solution of the following optimization problem

$$\hat{\boldsymbol{\theta}} = \underset{\boldsymbol{\theta}}{\mathrm{argmin}} \left\{ \left\| \sum_{k=1}^{p} \theta_k \phi_k \right\|_2^2 - 2 \sum_{k=1}^{p} \theta_k \langle \psi_k, \nu \rangle + \alpha \sum_{k=1}^{p} \sigma_k |\theta_k| \right\}. \tag{2.12}$$

Here, $\sum_{k=1}^{p} \sigma_k |\theta_k|$ is the weighted Lasso penalty and α is the penalty parameter. Note that the right hand side of formula (2.12) is independent of π_0, so the value of $\boldsymbol{\theta}$ can be evaluated.

In order to implement optimization procedure suggested above, consider matrix $\boldsymbol{\Phi} \in \mathbb{R}^{p \times p}$ with elements $\Phi_{lk} = \langle \phi_k, \phi_l \rangle$, $l, k = 1, \cdots, p$, and define vector $\boldsymbol{\xi}$ in $\mathbb{R}^p$ with components

$$\xi_k = \langle \psi_k, \nu \rangle = \sum_{l=0}^{\infty} \psi_k(l)\, \nu_l = n^{-1} \sum_{i=1}^{n} \psi_k(Y_i). \tag{2.13}$$

Introduce matrix $\mathbf{W}$ such that $\boldsymbol{\Phi} = \mathbf{W}^T \mathbf{W}$ and vector

$$\boldsymbol{\eta} = (\mathbf{W}^T)^+ \boldsymbol{\xi} = \mathbf{W} \left(\mathbf{W}^T \mathbf{W} \right)^{-1} \boldsymbol{\xi}, \tag{2.14}$$

where, for any matrix $\mathbf{A}$, matrix $\mathbf{A}^+$ is the Moore-Penrose inverse of $\mathbf{A}$. Then, the optimization problem (2.12) appears as

$$\widehat{\boldsymbol{\theta}} = \underset{\boldsymbol{\theta}}{\mathrm{argmin}} \left\{ \|\mathbf{W}\boldsymbol{\theta} - \boldsymbol{\eta}\|_2^2 + \alpha \sum_{k=1}^{p} \sigma_k |\theta_k| \right\}. \tag{2.15}$$

Now, consider the problem of estimating the weight π_0. Denote by $\tilde{f}$ the projection of $f(\lambda)$ onto the linear space spanned by the dictionary $\mathcal{D}$ and by $\tilde{\boldsymbol{\theta}}$ the coefficients of this projection. Let $\mathrm{supp}(\tilde{\boldsymbol{\theta}}) = \tilde{J}$. Consider vector $\mathbf{u}$ with components

$$u_k = U_k(0) = \int_0^{\infty} e^{-\lambda} \phi_k(\lambda)\, d\lambda, \quad k = 1, \cdots, p, \tag{2.16}$$

and observe that

$$P(0) = \mathbb{P}(Y = 0) = \pi_0 + \mathbf{u}^T \tilde{\boldsymbol{\theta}} + \Delta \quad \text{with} \quad \Delta = \int_0^{\infty} e^{-\lambda} (\tilde{f}(\lambda) - f(\lambda))\, d\lambda. \tag{2.17}$$

Since $\pi_0 \geq 0$, we estimate π_0 by

$$\hat{\pi}_0 = \max(0,\, \nu_0 - \mathbf{u}^T \hat{\boldsymbol{\theta}}). \tag{2.18}$$

3. Implementation of the Lasso estimator

Formulae (2.15) and (2.18) suggest the following procedure.

The direct algorithm

1. Evaluate sample frequencies ν_l, $l = 0, 1, ...$, given by formula (2.8).
2. Construct functions $\psi_k(Y)$, $k = 1, \cdots, p$, satisfying conditions (2.10) and (2.11).
3. Define a grid α_l, $l = 1, \cdots, L$, of values of α.
4. For each value α_l, $l = 1, \cdots, L$, evaluate a solution $\widehat{\boldsymbol{\theta}}_l$ of optimization problem (2.15) with $\alpha = \alpha_l$.
5. Select $\widehat{\boldsymbol{\theta}} = \widehat{\boldsymbol{\theta}}_l$ which optimizes one of the data driven criteria described in Section 5.
6. Estimate π_0 by $\hat{\pi}_0$ defined in (2.18).
7. Obtain the estimator of g as

$$\hat{g}(\lambda) = \hat{\pi}_0 \delta(\lambda) + \sum_{k=1}^{p} \hat{\theta}_k \phi_k(\lambda). \tag{3.1}$$

In order to implement Lasso estimator, for any ζ_k, we need to obtain a solution $\widetilde{\psi}_{k,\zeta_k}$ of equation (2.9). For his purpose, we introduce a matrix version $\mathbf{Q}$ of operator Q in (1.1). The elements of matrix $\mathbf{Q}$ are Poisson probabilities $\mathbf{Q}_{li} = e^{-x_i}(x_i)^l/(l!)$, where $x_i = ih$, $i = 1, 2...$, are the grid points at which we are going to recover $g(\lambda)$ and h is the step size. Introduce vectors $\boldsymbol{\phi}_k$ and $\widetilde{\psi}_{\mathbf{k},\boldsymbol{\zeta}_\mathbf{k}}$, $k = 1, \ldots, p$, with elements $\phi_k(x_i)$, $i = 1, 2...$, and $\widetilde{\psi}_k(l)$, $l = 0, 1, ...$, respectively. Then, for each $k = 1, \ldots, p$, equation (2.9) can be re-written as

$$\widetilde{\psi}_{\mathbf{k}} = (\mathbf{Q}\mathbf{Q}^T + \zeta_k \mathbf{I})^{-1} \mathbf{Q} \boldsymbol{\phi}_k, \tag{3.2}$$

where $\mathbf{I}$ is the identity matrix. For the sake of finding $\widehat{\zeta_k}$ satisfying (2.10), we create a grid and chose $\widehat{\zeta_k}$ so that to minimize an absolute value of $\widehat{\mathrm{Var}}[\widetilde{\psi}_{k,\zeta_k}(Y)] - \left(\mathbb{E}\widetilde{\psi}_{k,\zeta_k} - \langle g, \phi_k \rangle\right)^2$ where $\widehat{\mathrm{Var}}[\widetilde{\psi}_{k,\zeta_k}(Y)]$ is the sample variance of $\widetilde{\psi}_{k,\zeta_k}(Y)$. After that, we evaluate $\psi_k(Y)$ in (2.11) and replace unknown variances σ_k^2 in (2.11) by their sample counterparts.

Remark 1. (Iterative estimation procedure) In the case when $\pi_0 \phi_k(0) \neq 0$ for some values of $k \in \mathcal{P}$, evaluations above lead to an iterative estimation algorithm. Indeed, in this case, for a given value of π_0, the estimator $\hat{\boldsymbol{\theta}}$ is the solution of the following optimization procedure

$$\widehat{\boldsymbol{\theta}} = \underset{\boldsymbol{\theta}}{\mathrm{argmin}} \left\{ \boldsymbol{\theta}^T \boldsymbol{\Phi} \boldsymbol{\theta} - 2\boldsymbol{\theta}^T (\boldsymbol{\xi} - \hat{\pi}_0 \mathbf{z}) + \alpha \sum_{k=1}^{p} \sigma_k |\theta_k| \right\}, \tag{3.3}$$

where $z_k = \phi_k(0)$. On the other hand, for a given $\hat{\boldsymbol{\theta}}$, the estimator of π_0 is provided by formula (2.18). Combination of (3.3) and (2.18) suggest the estimation procedure described below. However, we emphasize that, unlike the direct algorithm, the iterative procedure comes without any guarantees for the estimation errors.

The iterative algorithm

1. Carry out steps 1, 2 and 3 of the direct algorithm.
2. Choose an initial value $\hat{\pi}_0^{(0)} = \nu_0$ and obtain $\widehat{\boldsymbol{\theta}}^{(0)}$ using steps 4-5 of the direct algorithm with (2.15) replaced by (3.3) evaluated with $\hat{\pi}_0 = \hat{\pi}_0^{(0)}$.
3. For $j = 1, 2, ...,$ set $\hat{\pi}_0^{(j)} = \max[0, \nu_0 - (\widehat{\boldsymbol{\theta}}^{(j-1)})^T \mathbf{u}]$ and obtain $\widehat{\boldsymbol{\theta}}^{(j)}$ by steps 4-5 of the direct algorithm with (2.15) replaced by (3.3) evaluated with $\hat{\pi}_0 = \hat{\pi}_0^{(j)}$. Repeat step 3 until one of the following stopping criteria is met:

   $$(i)\ \hat{\pi}_0^{(j)} = 0; \qquad (ii)\ \|W\widehat{\boldsymbol{\theta}}^{(j)} - W\widehat{\boldsymbol{\theta}}^{(j-1)}\|_2^2 < tol; \qquad (iii)\ j > J_{\max}.$$

   Here *tol* and $J_{\max}$ are, respectively, the tolerance level and the maximal number of steps defined in advance.
4. Obtain the estimator

$$\hat{g}(\lambda) = \hat{\pi}_0 \delta(\lambda) + \sum_{k=1}^{p} \hat{\theta}_k \phi_k(\lambda).$$

4. Convergence and estimation error

Let $\hat{g}(\lambda)$ be given by (3.1). In order to derive oracle inequalities for the error of $\hat{g}(\lambda)$, we introduce the following notations. For any vector $\mathbf{t} \in \mathbb{R}^p$, denote its ℓ^2, ℓ^1, ℓ^0 and ℓ^∞ norms by, respectively, $\|\mathbf{t}\|_2$, $\|\mathbf{t}\|_1$, $\|\mathbf{t}\|_0$ and $\|\mathbf{t}\|_\infty$ and

$$f_{\mathbf{t}} = \sum_{j=1}^{p} t_j \phi_j. \tag{4.1}$$

Similarly, for any function f, denote by $\|f\|_2$, $\|f\|_1$ and $\|f\|_\infty$ its L^2, L^1 and L^∞ norms. Denote $\mathcal{P} = \{1, \cdots, p\}$. For any subset of indices $J \subseteq \mathcal{P}$, subset J^c is its complement in $\mathcal{P}$ and $|J|$ is its cardinality, so that $|\mathcal{P}| = p$. Let $\mathcal{L}_J = \text{Span}\{\phi_j,\ j \in J\}$. If $J \subset \mathcal{P}$ and $\mathbf{t} \in \mathbb{R}^p$, then $\mathbf{t}_J \in \mathbb{R}^{|J|}$ denotes reduction of vector $\mathbf{t}$ to subset of indices J. Recall that $\tilde{f}$ is the projection of $f(\lambda)$ onto the linear space spanned by the dictionary $\mathcal{D}$ and $\tilde{\boldsymbol{\theta}}$ are the coefficients of this projection with $\text{supp}(\tilde{\boldsymbol{\theta}}) = \tilde{J}$.

It turns out that, as long as the sample size n is large enough, estimator $f_{\widehat{\boldsymbol{\theta}}}$ is close to f with high probability, with no additional assumptions. Indeed, the following statement holds.

Theorem 1. *Let $\pi_0 \phi_k(0) = 0$, for $k = 1, \cdots, p$, and let τ be any positive constant. If $n \geq N_0$ and $\alpha \geq \alpha_0$, where*

$$N_0 = \frac{16}{9}(\tau+1)\log p \max_{1\le k\le p}\left[\frac{\|\psi_k\|_\infty^2}{\sigma_k^2}\right] \quad \textit{and} \quad \alpha_0 = (2\sqrt{(\tau+1)\log p}+1)\, n^{-1/2}, \tag{4.2}$$

then with probability at least $1-2p^{-\tau}$*, one has*

$$\|f_{\widehat{\boldsymbol{\theta}}} - f\|_2^2 \le \inf_{\mathbf{t}}\left[\|f_{\mathbf{t}} - f\|_2^2 + 4\alpha\sum_{j=1}^p \sigma_j|t_j|\right] \tag{4.3}$$

where $\widehat{\boldsymbol{\theta}}$ *is the solution of optimization problem (2.15).*

Theorem 1 provides the so called "slow" Lasso rates. In order to obtain faster convergence rates and also to ensure that $\hat{\pi}_0$ is close to π_0 with high probability, we impose the so called compatibility condition on the dictionary $\phi_k, k = 1,\cdots,p$. In particular, denoting $\boldsymbol{\Upsilon} = \mathrm{diag}(\sigma_1,\cdots,\sigma_p)$ and considering the set of p-dimensional vectors

$$\mathcal{J}(\mu, J) = \{\mathbf{d}\in\mathbb{R}^p:\ \|(\boldsymbol{\Upsilon}\mathbf{d})_{J^c}\|_1 \le \mu\|(\boldsymbol{\Upsilon}\mathbf{d})_J\|_1\}, \quad \mu>1, \tag{4.4}$$

we assume that the following condition holds:

(A) Matrices $\boldsymbol{\Phi}$ and $\boldsymbol{\Upsilon}$ are such that for some $\mu>1$ and any $J\subset\mathcal{P}$

$$\kappa^2(\mu,J) = \min\left\{\mathbf{d}\in\mathcal{J}(\mu,J),\ \|\mathbf{d}\|_2\ne 0:\ \frac{\mathbf{d}^T\boldsymbol{\Phi}\mathbf{d}\cdot\mathrm{Tr}(\boldsymbol{\Upsilon}_J^2)}{\|(\boldsymbol{\Upsilon}\mathbf{d})_J\|_1^2}\right\} > 0. \tag{4.5}$$

Observe that, in the regression setup, $\boldsymbol{\Upsilon}$ is the identity matrix, and condition **A** reduces to the compatibility condition for general sets formulated in Section 6.2.3 of [5]. If one has an orthonormal basis instead of an overcomplete dictionary, then matrix $\boldsymbol{\Phi}$ is an identity matrix and, due to the Cauchy inequality, $\kappa^2(\mu,J)\ge 1$ for any μ and J. On the other hand, for an orthonormal basis, the bias $\|f_{\mathbf{t}} - f\|_2$ in (4.3) may be large.

Under Assumption **A**, one can prove "fast" convergence rates for $\hat{f}$ as well as obtain the error bounds for $\hat{\pi}_0$.

Theorem 2. *Let* $\pi_0\phi_k(0) = 0$*, for* $k = 1,\cdots,p$*,* τ *be any positive constant and Assumption* **A** *hold. Let* $\alpha = \varpi\alpha_0$ *where* α_0 *is defined in (4.2) and* $\varpi \ge (\mu+1)/(\mu-1)$*. If* $n\ge N_0$ *where* N_0 *is defined in (4.2), then with probability at least* $1-2p^{-\tau}$*, one has*

$$\|f_{\widehat{\boldsymbol{\theta}}} - f\|_2^2 \le \inf_{J\subseteq\mathcal{P}}\left[\|f - f_{\mathcal{L}_J}\|_2^2 + \frac{(1+\varpi)^2(2\sqrt{(\tau+1)\log p}+1)^2}{\kappa^2(\mu,J)\ n}\sum_{j\in J}\sigma_j^2\right], \tag{4.6}$$

where $f_{\mathcal{L}_J} = proj_{\mathcal{L}_J} f$*. Moreover, with probability at least* $1-4p^{-\tau}$*, one has*

$$\begin{aligned}(\hat{\pi}_0-\pi_0)^2 \ &\le\ \frac{2\tau\log p}{n} \\ &+\ \inf_{J\subseteq\mathcal{P}}\left[\|f - f_{\mathcal{L}_J}\|_2^2 + \frac{(1+\varpi)^2(2\sqrt{(\tau+1)\log p}+1)^2}{\kappa^2(\mu,J)\ n}\sum_{j\in J}\sigma_j^2\right].\end{aligned} \tag{4.7}$$

5. Numerical simulations

In order to evaluate the accuracy of the proposed estimator we carried out a simulation study where we tested performance of the proposed estimator under various scenarios. In order to assess precision of the estimator, for each of the scenarios, we evaluate the relative integrated error of $\hat{g}$ defined as

$$\Delta_g = \|g - \hat{g}\|_2^2 / \|g\|_2^2 \tag{5.1}$$

where the norm is calculated over the grid $x_i = ih$ with $i = 0, 1, \ldots$, if $\hat{\pi}_0 = \pi_0 = 0$ and $i = 1, 2, \ldots$, otherwise. In addition, we study prediction properties of $\widehat{g}$. In particular, we define the estimated frequencies

$$\hat{\nu}_l = \int_0^\infty \frac{\lambda^l}{l!} e^{-\lambda} \, \widehat{g}(\lambda) d\lambda = \hat{\pi}_0 \mathbb{I}(l = 0) + (1 - \hat{\pi}_0) \sum_{k=1}^{p} \hat{\theta}_k U_k(l), \quad \text{for } l = 0, 1, 2... \tag{5.2}$$

where $U_k(l)$ are given in equation (2.6). The, we evaluate

$$\Delta_\nu = \|\boldsymbol{\nu} - \hat{\boldsymbol{\nu}}\|_2^2 / \|\boldsymbol{\nu}\|_2^2, \tag{5.3}$$

i.e. Δ_ν evaluates the squared relative ℓ^2 distance between the vectors of predicted and of observed frequencies. For the estimator proposed in this paper, we tested various computational schemes that differ by the strategies for selecting the penalty parameter α_l in step 5 of the direct algorithm. In particular, we considered the following options.

> OPT : This estimator is obtained by using the direct algorithm presented in Section 3, where in step 5 parameter α_l is chosen by minimizing $\|g - \hat{g}\|_2^2$, i.e. the squared ℓ^2 distance between the true and the estimated function. Of course, this estimator represents only a benchmark for the proposed procedure since it is not really applicable in practice because it requires knowledge of g.
>
> DD_{l2} : This estimator is obtained by using the direct algorithm presented in Section 3, where in step 5 parameter α_l is chosen by minimizing Δ_ν, given in (5.3), i.e. the squared relative ℓ^2 distance between the predicted and the observed frequencies.
>
> DD_{like} : This estimator is obtained by using the direct algorithm presented in Section 3, where in step 5 parameter α_l is chosen by maximizing the likelihood function as suggested in [6]. In particular, since $\hat{\boldsymbol{\nu}} = (\hat{\nu}_0, \hat{\nu}_1, \ldots,)$ given in (5.2) and $\boldsymbol{\nu} = (\nu_0, \nu_1, \ldots)$ given in (2.8) are, respectively, the predicted and the observed frequencies, the likelihood function can be written as $L(\boldsymbol{\nu}|\hat{\boldsymbol{\nu}}) = \prod_{l=0}^{M} \hat{\nu}_l^{\nu_l}$, where $M = \max_i Y_i$.

For the sake of comparison we also define

> NDE : This is the Nonparametric Density Estimator presented in [8], for which the authors kindly provided the code.

The set of test functions represents different situations inspired by the real data problem described in the next Section. In particular, we consider the following nine test functions:

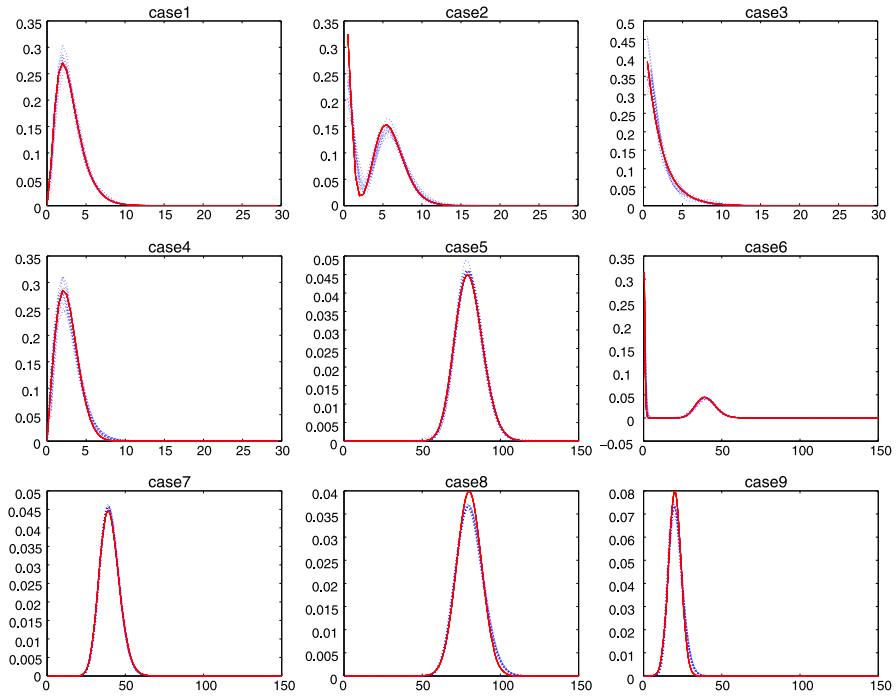

FIG 1. *The true density (red) and DD_{like} estimators (blue) obtained in the first 10 simulation runs with sample size $n = 5000$*

1. the gamma density $g(\lambda) = \Gamma(\lambda; 3, 1)$
2. the mixed gamma density $g(\lambda) = 0.3\Gamma(\lambda; 3, 0.25) + 0.7\Gamma(\lambda; 10, 0.6)$
3. the exponential density $g(\lambda) = \Gamma(\lambda; 1, 2)$
4. the Weibull density $g(\lambda) = \theta p^{-\theta} x^{\theta-1} \exp -(x/\theta)^{\theta}\, \mathbb{I}(x > 0)$, with $p = 3$ and $\theta = 2$
5. the Gaussian density $g(\lambda) = N(\lambda; 80, 1)$
6. the mixed gamma density $g(\lambda) = 0.3\Gamma(\lambda; 2, 0.3) + 0.7\Gamma(\lambda; 40, 1)$
7. the delta contaminated gamma density $g(\lambda) = 0.3\delta(\lambda) + 0.7\Gamma(\lambda; 40, 1)$
8. the delta contaminated Gaussian density $g(\lambda) = 0.2\delta(\lambda) + 0.8N(\lambda; 80, 8^2)$
9. the delta contaminated Gaussian density $g(\lambda) = 0.2\delta(\lambda) + 0.8N(\lambda; 20, 4^2)$

The first four test functions have been analyzed in [8] and represent cases where most of the data is concentrated near zero. The fifth test function corresponds to the situation where most of the data is concentrated away from zero. The last four test functions represent the mixtures of the two previous scenarios. All nine densities are showed in Figure 1.

Tables 1, 3 and 5 below display the average values of Δ_g defined in (5.1) while Tables 2, 4 and 6 report Δ_ν defined in (5.3) (with the standard deviations in parentheses) over 100 different realizations of data $Y_i \sim Poisson(\lambda_i)$, $i = 1, .., n$, where $n = 10000$ for Tables 1 and 2, $n = 5000$ for Tables 3 and 4

TABLE 1
Average values of Δ_g (with their standard deviations in parentheses) over 100 simulation runs with $n = 10000$

test case	OPT	DD_{l2}	DD_{like}	NDE
case1	0.0007 (0.0010)	0.0023 (0.0028)	0.0022 (0.0030)	0.1183 (0.8307)
case2	0.0471 (0.0197)	0.2214 (0.0503)	0.0507 (0.0305)	0.0613 (0.0716)
case3	0.0142 (0.0398)	0.0191 (0.0127)	0.0138 (0.0087)	0.0190 (0.0399)
case4	0.0043 (0.0021)	0.0054 (0.0032)	0.0061 (0.0052)	0.0298 (0.0657)
case5	0.0042 (0.0033)	0.0023 (0.0029)	0.0014 (0.0021)	1.0000 (0.0000)
case6	0.0793 (0.0247)	0.4318 (0.0554)	0.0839 (0.0241)	0.3383 (0.0085)
case7	0.0067 (0.0012)	0.0009 (0.0008)	0.0008 (0.0008)	-
case8	0.0060 (0.0010)	0.0068 (0.0009)	0.0069 (0.0010)	-
case9	0.0085 (0.0013)	0.0099 (0.0014)	0.0111 (0.0026)	-

TABLE 2
Average values of Δ_ν (with their standard deviations in parentheses) over 100 simulation runs with $n = 10000$

test case	OPT	DD_{l2}	DD_{like}	NDE
case1	0.0011 (0.0009)	0.0006 (0.0005)	0.0007 (0.0005)	0.0675 (0.5470)
case2	0.0013 (0.0007)	0.0088 (0.0024)	0.0014 (0.0010)	0.0228 (0.0431)
case3	0.0002 (0.0001)	0.0006 (0.0005)	0.0003 (0.0003)	0.1130 (0.0653)
case4	0.0014 (0.0009)	0.0011 (0.0006)	0.0012 (0.0009)	0.0205 (0.0530)
case5	0.0045 (0.0010)	0.0043 (0.0010)	0.0044 (0.0009)	1.0000 (0.0000)
case6	0.0465 (0.1402)	0.0288 (0.0045)	0.0041 (0.0020)	0.4376 (0.0618)
case7	0.0013 (0.0002)	0.0006 (0.0002)	0.0006 (0.0002)	-
case8	0.0039 (0.0009)	0.0018 (0.0004)	0.0018 (0.0004)	-
case9	0.0035 (0.0006)	0.0019 (0.0007)	0.0020 (0.0006)	-

TABLE 3
Average values of Δ_g (with their standard deviations in parentheses) over 100 simulation runs with $n = 5000$

test case	OPT	DD_{l2}	DD_{like}	NDE
case1	0.0006 (0.0008)	0.0038 (0.0049)	0.0030 (0.0046)	0.2424 (1.5002)
case2	0.0590 (0.0343)	0.2106 (0.0549)	0.0640 (0.0428)	0.2048 (0.2196)
case3	0.0148 (0.0097)	0.0251 (0.0310)	0.0178 (0.0119)	0.0309 (0.0650)
case4	0.0055 (0.0019)	0.0074 (0.0051)	0.0086 (0.0060)	0.0493 (0.1123)
case5	0.0069 (0.0052)	0.0044 (0.0048)	0.0024 (0.0036)	1.0000 (0.0000)
case6	0.0830 (0.0277)	0.4068 (0.0856)	0.0879 (0.0266)	0.3456 (0.0110)
case7	0.0077 (0.0035)	0.0023 (0.0030)	0.0023 (0.0030)	-
case8	0.0065 (0.0021)	0.0074 (0.0020)	0.0075 (0.0021)	-
case9	0.0096 (0.0023)	0.0114 (0.0023)	0.0128 (0.0029)	-

and $n = 1000$ for Tables 5 and 6. We chose the grid step $h = 0.5$. The dictionary was constructed as a collection of the gamma pdfs (2.1) where parameters (a_k, b_k) belong to the Cartesian product of vectors $a = [2, 3, 4, \cdots, 150]$ and $b = [0.1, 0.15, \cdots, 0.9, 0.95]$, hence, $\phi_k(0) = 0$ and $p = 2682$. For this dictionary, $\max_{1\le k\le p}\left[\|\psi_k\|_\infty^2/\sigma_k^2\right] = 146.95$, so that condition (4.2) holds with $\tau = 3.85$ and $\tau = 1.43$ for $n = 10000$ and $n = 5000$, respectively, and is not valid for $n = 1000$. However, as simulation results show, our estimator shows good performance even when assumption (4.2) is violated.

As it is expected, performances of all estimators deteriorate when n decreases, although not very significantly. For a fixed sample size, estimator OPT is the most precise in terms of Δ_g as a direct consequence of its definition, however, estimator DD_{like} is always comparable. Estimator DD_{l2} has similar performance to DD_{like} except for cases 2 and 6 where the underlying densities are bimodal and, hence, data can be explained by a variety of density mixtures.

TABLE 4
Average values of Δ_ν (with their standard deviations in parentheses) over 100 simulation runs with $n = 5000$

test case	OPT	DD_{L2}	DD_{like}	NDE
case1	0.0017 (0.0016)	0.0010 (0.0007)	0.0012 (0.0007)	0.1393 (1.0084)
case2	0.0020 (0.0013)	0.0090 (0.0029)	0.0022 (0.0014)	0.1184 (0.1475)
case3	0.0004 (0.0002)	0.0009 (0.0013)	0.0006 (0.0004)	0.1131 (0.0948)
case4	0.0022 (0.0014)	0.0016 (0.0008)	0.0018 (0.0013)	0.0346 (0.0859)
case5	0.0087 (0.0019)	0.0084 (0.0018)	0.0085 (0.0018)	1.0000 (0.0000)
case6	0.0377 (0.1361)	0.0285 (0.0068)	0.0057 (0.0030)	0.4608 (0.0849)
case7	0.0020 (0.0003)	0.0013 (0.0003)	0.0013 (0.0003)	-
case8	0.0052 (0.0011)	0.0032 (0.0008)	0.0032 (0.0008)	-
case9	0.0044 (0.0010)	0.0031 (0.0010)	0.0031 (0.0009)	-

TABLE 5
Average values of Δ_g (with their standard deviations in parentheses) over 100 simulation runs with $n = 1000$

test case	OPT	DD_{l2}	DD_{like}	NDE
case1	0.0040 (0.0097)	0.0221 (0.0258)	0.0176 (0.0207)	0.3004 (0.9331)
case2	0.0992 (0.0760)	0.1973 (0.0718)	0.1335 (0.0983)	0.5370 (0.0960)
case3	0.0533 (0.0889)	0.0753 (0.0838)	0.0662 (0.0894)	0.1127 (0.3912)
case4	0.0069 (0.0014)	0.0178 (0.0183)	0.0179 (0.0135)	0.1393 (0.3381)
case5	0.0170 (0.0108)	0.0217 (0.0253)	0.0152 (0.0223)	1.0000 (0.0000)
case6	0.1223 (0.0710)	0.3151 (0.1409)	0.1270 (0.0759)	0.4479 (0.2572)
case7	0.0133 (0.0115)	0.0102 (0.0118)	0.0098 (0.0118)	-
case8	0.0142 (0.0137)	0.0156 (0.0139)	0.0156 (0.0154)	-
case9	0.0121 (0.0073)	0.0163 (0.0110)	0.0160 (0.0101)	-

TABLE 6
Average values of Δ_ν (with their standard deviations in parentheses) over 100 simulation runs with $n = 1000$

test case	OPT	DD_{l2}	DD_{like}	NDE
case1	0.0063 (0.0042)	0.0043 (0.0026)	0.0047 (0.0027)	0.1458 (0.5377)
case2	0.0076 (0.0051)	0.0117 (0.0047)	0.0084 (0.0048)	0.3498 (0.0937)
case3	0.0021 (0.0022)	0.0031 (0.0033)	0.0027 (0.0037)	0.2149 (0.3773)
case4	0.0075 (0.0068)	0.0046 (0.0029)	0.0048 (0.0032)	0.0967 (0.3578)
case5	0.0427 (0.0091)	0.0411 (0.0090)	0.0416 (0.0090)	1.0000 (0.0000)
case6	0.0157 (0.0044)	0.0304 (0.0127)	0.0166 (0.0067)	0.5344 (0.1001)
case7	0.0072 (0.0018)	0.0067 (0.0018)	0.0067 (0.0018)	-
case8	0.0154 (0.0038)	0.0143 (0.0038)	0.0142 (0.0037)	-
case9	0.0120 (0.0034)	0.0111 (0.0032)	0.0111 (0.0032)	-

In conclusion, apart from OPT which is not available in the case of real data, estimator DD_{like} turns out to be the most accurate in terms of both Δ_g and Δ_ν. For completeness, Figures 1 and 2 exhibit some of the reconstructions obtained using estimator DD_{like} in the case of $n = 5000$.

Finally, we should mention that NDE is a projection estimator that uses only the first few Laguerre functions. For this reason, it fails to adequately represent a density function that corresponds to the situation where values λ_i, $i = 1, \cdots, n$, are concentrated away from zero, as it happens in case 5 (where NDE returns zero as an estimator) and case 6 (where NDE succeeds in reconstructing only the first part of the density near zero). Also, note that NDE errors are not displayed for cases 7, 8 and 9 since this estimator is not defined for delta contaminated densities.

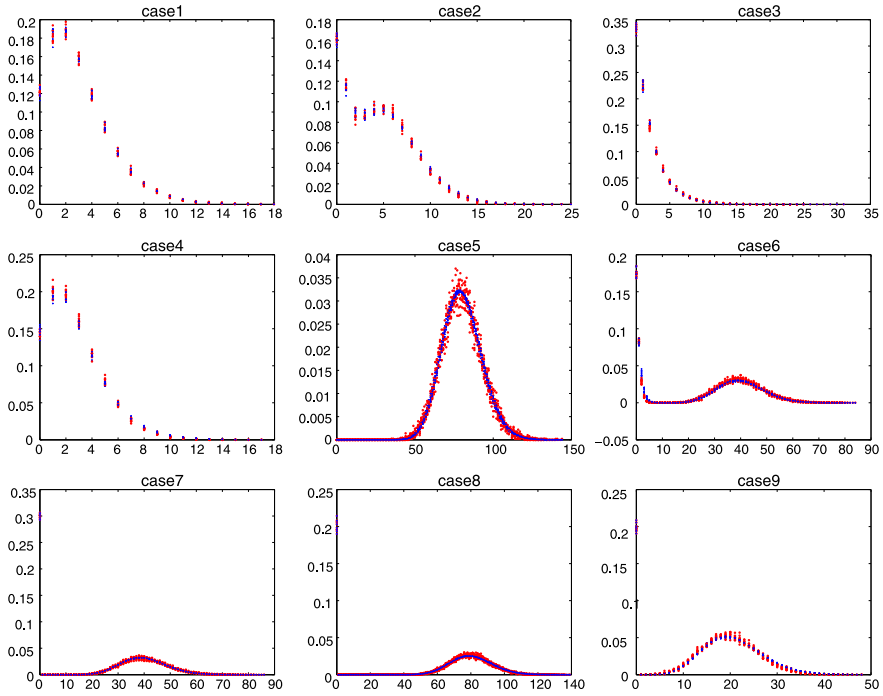

FIG 2. *Sample frequencies (red) and estimated frequencies (blue) obtained in the first 10 simulation runs with sample size* $n = 5000$

6. Application to evaluation of the density of the Saturn ring

The Saturn's rings system can be broadly grouped into two categories: dense rings (A, B, C) and tenuous rings (D, E, G) (see the first panel of Figure 3). The Cassini Division is a ring region that separates the A and B rings. The study of structure within Saturn's rings originated with Campani, who observed in 1664 that the inner half of the disk was brighter than the outer half. Furthermore, in 1859, Maxwell proved that the rings could not be solid or liquid but were instead made up of an indefinite number of particles of various sizes, each on its own orbit about Saturn. Detailed ring structure was revealed for the first time, however, by the 1979 Pioneer and 1980-1981 Voyager encounters with Saturn. Images were taken at close range, by stellar occultation (observing the flickering of a star as it passes behind the rings), and by radio occultation (measuring the attenuation of the spacecraft's radio signal as it passes behind the rings as seen from Earth) (see, e.g., [9] and [7]). By analyzing the intensity of star light while it is passing through Saturn's rings, astronomers can gain insight into properties that telescopes cannot visually determine. Each sub-region in the rings has its own associated distinct distribution of the density and sizes of the particles constituting the sub-region. This distribution uniquely determines the amount of light which is able to pass from a star (behind the rings) to the photometer.

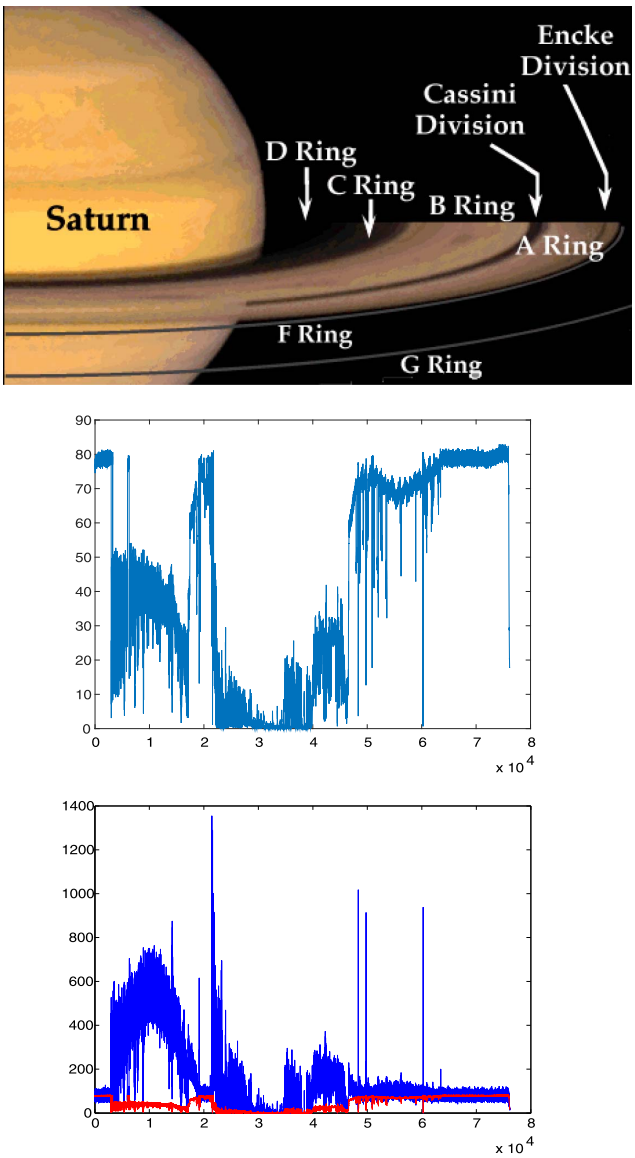

FIG 3. *The first panel: Names of Saturn's rings, courtesy* `science.nasa.gov`. *The second panel: the means of the binned total data set (100 observations per bin). The third panel: the means (red) and the variances (blue) of the binned data*

Our data Y_i, $i = 1, \cdots, n$, come from sets of observations of stellar occultations recorded by the Cassini UVIS high speed photometer and contains $n = 7\,615\,754$ photon counts at different radial points, located at 0.01-0.1 kilometer increments, on the Saturn's rings plane (see the second panel of Figure 3). With no outside influences, these photon counts should follow the Poisson distri-

bution, however, obstructions imposed by the particles in the rings cause their distribution to deviate from Poisson. Indeed, if data were Poisson distributed, then its mean would be approximately equal to its variance for every sub-region. However, as the third panel of Figure 3 shows, observations Y_i have significantly higher variances than means. The latter is due to the fact that, although for each $i = 1, \cdots, n$, the photon counts Y_i are $Poisson(\lambda_i)$, the values of λ_i, $i = 1, \cdots, n$, are extremely varied and, specifically, cannot be modeled as the values of a continuous function. In fact, intensities λ_i, $i = 1, \cdots, n$, are best described as random variables with an unknown underlying pdf $g(\lambda)$.

In addition, if the ring region contains a significant proportion of large particle, those particles can completely block of the light leading to zero photon counts. For this reason, we allow $g(\lambda)$ to possibly contain a non-zero mass at $\lambda = 0$, hence, being of the form (1.2). The shape of $g(\lambda)$ allows one to determine the density and distribution of the sizes of the particles of a respective sub-region of the Saturn rings. This information, in turn, should shed light on the question of the origin of the rings as well as how they reached their current configuration.

In order to identify sub-regions of the Saturn rings with different properties, we segmented the data using the method presented in [4] which is designed for partitioning of complicated signals with several non-isolated and oscillating singularities. In particular, we applied the Gabor Continuous Wavelet Transform (see, e.g. [17]) to the data and selected the highest scale where the number of wavelet modulus maxima takes minimum value. At this scale, we segmented the signal by the method proposed in [4]. We obtained a total of 1531 intervals of different sizes. Figures 4 and 5 refer to six distinct sub-regions of the rings. The left panels of both figures show raw data. The right panels exhibit the sample and the estimated frequencies obtained by DD_{like} estimator for six different intervals that are representative of different portions of the data set.

Note that in Figure 4, for all three data segments, the estimated parameter $\hat{\pi}_0 = 0$. This is not true for the first and the second panels of Figure 5 where $\hat{\pi}_0 = 0.5059$ and $\hat{\pi}_0 = 0.2463$, respectively. The values of Δ_ν, defined in (5.3), obtained for the six data segments are, respectively, 0.0128, 0.0159, 0.0022, 0.0229, 0.003 and 0.0095, and are consistent with the values obtained in simulations. Both, the right panels in Figures 4 and 5 and the values of Δ_ν, confirm the ability of the estimator developed in the paper to accurately explain the Saturn's rings data.

7. Proofs

Proofs of Theorems 1 and 2 are based on the following statement which is a trivial modification of Lemma 2 of [18].

Lemma 1. (Pensky (2016)). *Let f be the true function and $f_{\boldsymbol{\theta}}$ be its projection onto the linear span of the dictionary $\mathcal{L}_{\mathcal{P}}$. Let $\boldsymbol{\Upsilon}$ be a diagonal matrix with components σ_j, $j = 1, \cdots, p$. Consider solution of the weighted Lasso problem*

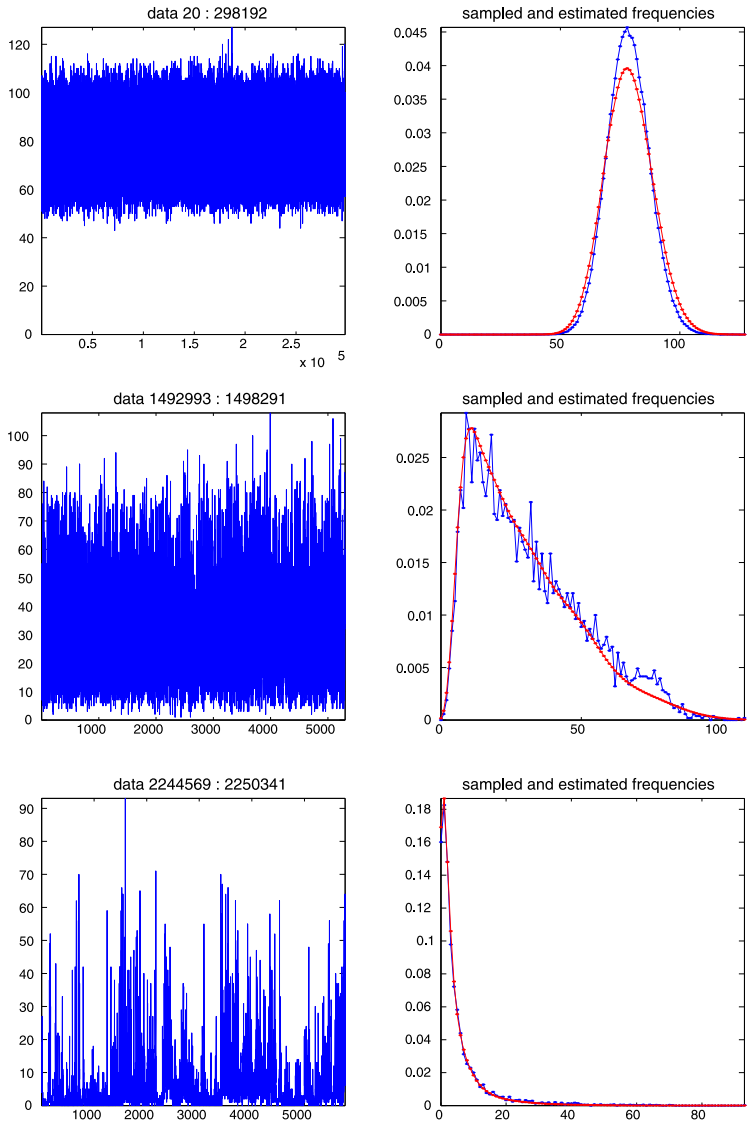

FIG 4. *Left panels: segments of data. Right panels: sample frequencies (blue) and estimated frequencies with the penalty parameter obtained by* DD_{like} *criterion (red).* $\Delta_\nu = 0.0128$ *(top panel),* $\Delta_\nu = 0.0159$ *(middle panel),* $\Delta_\nu = 0.0022$, $\hat{\pi}_0 = 0$ *for all three cases. (bottom panel)*

$$\widehat{\boldsymbol{\theta}} = \arg\min_{\mathbf{t}} \left\{ \mathbf{t}^T \mathbf{W}\mathbf{W}^T \mathbf{t} - 2\mathbf{t}^T \widehat{\boldsymbol{\beta}} + \alpha \|\boldsymbol{\Upsilon}\mathbf{t}\|_1 \right\}. \tag{7.1}$$

with $\boldsymbol{\Phi} = \mathbf{W}^T\mathbf{W}$, $\boldsymbol{\beta} = \boldsymbol{\Phi}\boldsymbol{\theta}$ *and*

$$\widehat{\boldsymbol{\beta}} = \boldsymbol{\beta} + \sqrt{\varepsilon}\boldsymbol{\Upsilon}\boldsymbol{\eta} + \mathbf{h}, \quad \boldsymbol{\eta}, \mathbf{h} \in \mathbb{R}^p, \tag{7.2}$$

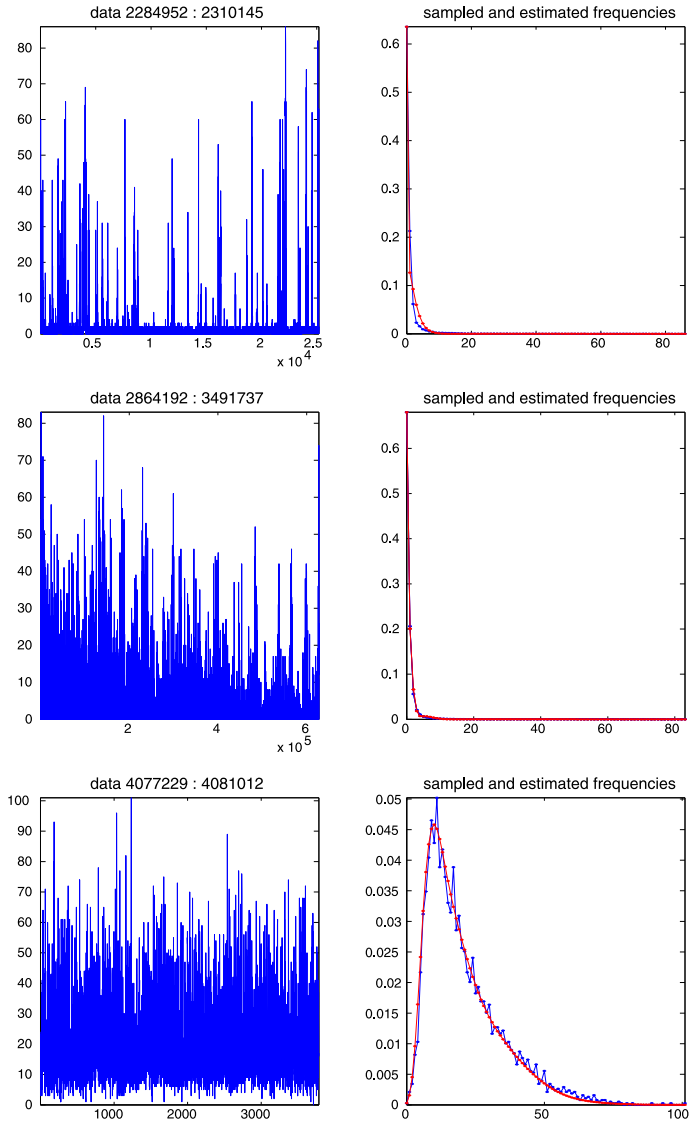

FIG 5. *Left panels: segments of data. Right panels: sample frequencies (blue) and estimated frequencies with the penalty parameter obtained by* DD_{like} *criterion (red).* $\Delta_\nu = 0.0229$ *and* $\hat{\pi}_0 = 0.5059$ *(top panel),* $\Delta_\nu = 0.003$ *and* $\hat{\pi}_0 = 0.2463$ *(middle panel),* $\Delta_\nu = 0.0095$ *and* $\hat{\pi}_0 = 0$ *(bottom panel)*

where $\mathbf{h}$ *is a nonrandom vector,* $\mathbb{E}\boldsymbol{\eta} = 0$ *and components* η_j *of vector* $\boldsymbol{\eta}$ *are random variables such that, for some* $K > 0$ *and any* $\tau > 0$*, there is a set*

$$\Omega = \left\{ \omega : \max_{1 \le j \le p} |\eta_j| \le K\sqrt{(\tau+1)\log p} \right\} \quad \textit{with} \quad \mathbb{P}(\Omega) \ge 1 - 2p^{-\tau}. \tag{7.3}$$

Denote

$$C_h = \max_{1\le j\le p}\left[\frac{|h_j|}{\sigma_j\sqrt{\varepsilon\log p}}\right], \qquad C_\alpha = K\sqrt{\tau+1}+C_h. \tag{7.4}$$

If $\alpha_0 = C_\alpha\sqrt{\varepsilon\log p}$ *then, for any* $\tau>0$ *and any* $\alpha\ge\alpha_0$*, with probability at least* $1-2p^{-\tau}$*, one has*

$$\|f_{\widehat{\boldsymbol{\theta}}}-f\|_2^2 \le \inf_{\mathbf{t}}\left[\|f_{\mathbf{t}}-f\|_2^2+4\alpha\|\boldsymbol{\Upsilon}\mathbf{t}\|_1\right]. \tag{7.5}$$

Moreover, if Assumption **A** *holds and* $\alpha=\varpi\alpha_0$ *where* $\varpi\ge(\mu+1)/(\mu-1)$*, then for any* $\tau>0$ *with probability at least* $1-2p^{-\tau}$*, one has*

$$\|f_{\widehat{\boldsymbol{\theta}}}-f\|_2^2 \le \inf_{\mathbf{t},J\subseteq\mathcal{P}}\left[\|f_{\mathbf{t}}-f\|_2^2+4\alpha\|(\boldsymbol{\Upsilon}\mathbf{t})_{J^c}\|_1+\frac{(1+\varpi)^2C_\alpha^2}{\kappa^2(\mu,J)}\varepsilon\log p\sum_{j\in J}\nu_j^2\right]. \tag{7.6}$$

Proof of Theorem 1. Let vectors $\mathbf{b}$ and $\boldsymbol{\xi}$, respectively, have components $b_k=\langle\phi_k,f\rangle$ and ξ_k defined in (2.13). It is easy to see that

$$\xi_k-b_k=\frac{1}{n}\sum_{i=1}^n[\psi_k(Y_i)-\mathbb{E}\psi_k(Y_i)]+H_k \quad\text{with}\quad H_k=\mathbb{E}\widetilde{\psi}_{k,\widehat{\zeta}_k}-b_k \tag{7.7}$$

Applying Bernstein inequality, for any $x>0$, obtain

$$\mathbb{P}\left(\left|n^{-1}\sum_{i=1}^n[\psi_k(Y_i)-\mathbb{E}\psi_k(Y_i)]\right|\ge\frac{x\sigma_k}{\sqrt{n}}\right)\le 2\exp\left(-x^2\left[2+\frac{4x\sigma_k\|\psi_k\|_\infty}{3\sqrt{n}}\right]^{-1}\right).$$

Using the fact that $A/(B+C)\ge\min(A/(2B),A/(2C))$ for any $A,B,C>0$, under condition $n\ge N_0$, derive

$$\mathbb{P}\left(\left|n^{-1}\sum_{i=1}^n[\psi_k(Y_i)-\mathbb{E}\psi_k(Y_i)]\right|\ge xn^{-1/2}\sigma_k\right)\le 2\exp\{-(x^2/4)\}. \tag{7.8}$$

Choosing $x=2\sqrt{(\tau+1)\log p}$ and recalling that, according to (2.10), $|H_k|=n^{-1/2}\sigma_k$, gather that $\mathbb{P}(|\xi_k-b_k|>n^{-1/2}\sigma_k[2\sqrt{(\tau+1)\log p}+1])\le 2p^{-(\tau+1)}$, so that

$$\Omega_1=\left\{\omega:\max_{1\le k\le p}\left[\frac{|\xi_k-b_k|}{\sigma_k}\right]\le\frac{2\sqrt{(\tau+1)\log p}+1}{\sqrt{n}}\right\}\quad\text{with }\mathbb{P}(\Omega_1)>1-2p^{-\tau}. \tag{7.9}$$

Then, validity of Theorem 1 follows directly from Lemma 1 with $\varepsilon=n^{-1}$, $\eta_k=\xi_k/\sigma_k$ and $K=2$.

Proof of Theorem 2. Validity of inequality (4.6) follows from (7.9) and Lemma 1 with $\varepsilon=\sigma/\sqrt{n}$, $K=2$, $h_j=n^{-1/2}\sigma_j$ and $C_h=(\log p)^{-1/2}$.

In order to establish upper bounds for $(\hat{\pi}_0 - \pi_0)^2$, note that due to (2.17) and (2.18) and $\pi_0 \geq 0$, one has

$$|\hat{\pi}_0 - \pi_0| \leq |\mathbb{P}(Y = 0) - \nu_0| + |\mathbf{u}^T(\hat{\boldsymbol{\theta}} - \tilde{\boldsymbol{\theta}}) + \Delta|. \tag{7.10}$$

For any $\tau > 0$, by Hoeffding inequality obtain $\mathbb{P}\{|\mathbb{P}(Y=0)-\nu_0| \leq \sqrt{\tau n^{-1} \log p}\} \geq 1 - 2p^{-\tau}$. Let Ω_1 be the set on which $|\mathbb{P}(Y = 0) - \nu_0| \leq \sqrt{\tau n^{-1} \log p}$. Then, $\mathbb{P}(\Omega_1) \geq 1 - 2p^{-\tau}$.

Now, let Ω_2 be the set on which inequality (4.6) holds and $\mathbb{P}(\Omega_2) \geq 1 - 2p^{-\tau}$. Observe that

$$\mathbf{u}^T(\hat{\boldsymbol{\theta}} - \tilde{\boldsymbol{\theta}}) = \int_0^\infty e^{-\lambda} \sum_{j=1}^{p} (\tilde{\theta}_j - \hat{\theta}_j)\phi_j \lambda)\, d\lambda = \int_0^\infty e^{-\lambda}(\hat{f}(\lambda) - \tilde{f}(\lambda))d\lambda$$

where $\tilde{f}$ is the projection of $f(\lambda)$ onto the linear space spanned by the dictionary $\mathcal{D}$. Therefore, by the definition of Δ in (2.17) obtain that

$$|\mathbf{u}^T(\hat{\boldsymbol{\theta}} - \tilde{\boldsymbol{\theta}}) + \Delta| \leq \left| \int_0^\infty e^{-\lambda}(\hat{f}(\lambda) - f(\lambda))\, d\lambda \right| \leq \|\hat{f} - f\|_2/\sqrt{2}. \tag{7.11}$$

Hence, on a set $\Omega = \Omega_1 \cap \Omega_2$ with $\mathbb{P}(\Omega) \geq 1 - 4p^{-\tau}$, using (7.10) and (7.11), obtain (4.7) which completes the proof.

Acknowledgments

Marianna Pensky was partially supported by National Science Foundation (NSF), grant DMS-1407475. Daniela De Canditiis was entirely supported by the "Italian Flagship Project Epigenomic" (http://www.epigen.it/). The authors would like to thank Dr. Joshua Colwell for helpful discussions and for providing the data. The authors also would like to thank SAMSI for providing support which allowed the author's participation in the 2013-14 LDHD program which was instrumental for writing this paper.

References

[1] Antoniadis, A., Sapatinas, T. (2004). Wavelet shringkage for natural exponential families with quadratic variance functions, *Biometrika*, **88**, 805–820. MR1859411

[2] Besbeas, P., De Feis, I., Sapatinas, T. (2004). A comparative simulation study of wavelet shrinkage estimators for Poisson counts, *International Statistical Review*, **72**, 209–237.

[3] Brown, L. D., Cai, T. T., Zhou, H. (2010). Nonparametric regression in exponential families, *Ann. Statist.*, **38**, 2005–2046. MR2676882

[4] Bruni, V., De Canditiis, D., Vitulano, D. (2012). Time-scale energy based analysis of contours of real-world shapes, *Mathematics and Computer in Simulation*, **82**, 2891–2907. MR2997151

[5] Bühlmann, P., van de Geer, S. (2011). *Statistics for High-Dimensional Data: Methods, Theory and Applications*, Springer. MR2807761
[6] Chow, Y.-S., Geman, S., Wu, L.-D. (1983). Consistent cross-validated density estimaton, *Ann. Statist.*, **11**, 25–38. MR0684860
[7] Colwell, J. E., Nicholson, P. D., Tiscareno, M. S., Murray, C. D., French, R. G., Marouf, E. A. (2009). The Structure of Saturn's Rings, in *Saturn from Cassini-Huygens*, Dougherty, M., Esposito, L., Krimigis, S. Eds., Springer.
[8] Comte, F., Genon-Catalot, V. (2015). Adaptive Laguerre density estimation for mixed Poisson models, *Electr. Journ. Statist.*, **9**, 1113–1149. MR3352069
[9] Esposito, L. W., Barth, C. A., Colwell, J. E., Lawrence, G. M., McClintock, W. E., Stewart, A. I. F., Keller, H. U., Korth, A., Lauche, H., Festou, M. C., Lane, A. L., Hansen, C. J., Maki, J. N., West, R. A., Jahn, H., Reulke, R., Warlich, K., Shemansky, D. E., Yung, Y. L. (2004). The Cassini ultraviolet imaging spectrograph investigation, *Space Sci. Rev.*, **115**, 294–361.
[10] Fryzlewicz, P., Nason G. P. (2004). A Haar-Fisz algorithm for Poisson intensity estimation, *Journ. Computat. Graph. Statis.*, **13**, 621–638. MR2087718
[11] Harmany, Z., Marcia, R., Willett, R. (2012). This is SPIRAL-TAP: Sparse Poisson Intensity Reconstruction ALgorithms: Theory and Practice, *IEEE Trans. Image Processing*, **21**, 1084–1096. MR2951281
[12] Herngartner, N. W. (1997). Adaptive demixing in Poisson mixture models, *Ann. Stat.*, **25**, 917–928. MR1447733
[13] Hirakawa, K., Wolfe, P. J. (2012). Skellam shrinkage: Wavelet-based intensity estimation for inhomogeneous Poisson data, *IEEE Trans. Inf. Theory*, **58**, 1080–1093. MR2918011
[14] Kolaczyk, E. D. (1999). Bayesian multiscale models for Poisson processes, *Journ. Amer. Statist. Assoc.*, **94**, 920–933. MR1723303
[15] Lambert, D., Tierney, L. (1984). Asymptotic properties of maximum likelihood estimates in the mixed Poisson model, *Ann. Statist.*, **12**, 1388-1399. MR0765931
[16] Lord, D., Washington, S. P., Ivan, J. N. (2005). Poisson, Poisson-gamma and zero-inflated regression models of motor vehicle crashes: Balancing statistical fit and theory, *Accid. Anal. Prevent.*, **37**, 35–46.
[17] Mallat, S. (2009). *A Wavelet Tour of Signal Processing: The Sparse Way*, 3rd ed. Elsevier. MR2479996
[18] Pensky, M. (2016). Solution of linear ill-posed problems using overcomplete dictionaries, *Ann. Stat.*, to appear.
[19] Timmermann, K. E., Nowak, R. D. (1999). Multiscale modeling and estimation of Poisson processes with application to photon-limited imaging, *IEEE Trans. Inf. Theory*, **45**, 846–862. MR1682515
[20] Walter, G. (1985). Orthogonal polynomials estimators of the prior distribution of a compound Poisson distribution, *Sankhya, Ser. A*, **47**, 222–230. MR0844023

[21] Walter, G., Hamedani, G. (1991). Bayes empirical Bayes estimation for natural exponential families with quadratic variance function, *Ann. Statist.*, **19**, 1191–1224. MR1126321

[22] Zhang, C.-H. (1995). On estimating mixing densities in discrete exponential family models, *Ann. Statist.*, **23**, 929–947. MR1345207

[23] Zou, H., Hastie, T. (2005). Regularization and variable selection via the elastic net, *JRSS, Ser. B*, **67, Part 2**, 301–320. MR2137327