Nonparametric Bayesian Estimation of X/γ-Ray Spectra using a Hierarchical Pólya Tree–Dirichlet Mixture Model

Éric Barat and Thomas Dautremer

CEA Saclay, Electronics and Signal Processing Laboratory, 91191 Gif sur Yvette, France

Abstract. The nonparametric Bayesian estimation of physical spectra is addressed by considering a mixture involving a Dirichlet process in order to capture the hidden discrete nature of peaks and a Pólya tree to tackle the complexity of the hidden continuous background.

Keywords: Bayesian nonparametrics, density estimation, Pólya tree, Dirichlet mixture, stick breaking priors, nuclear physics, x/γ-ray spectra.
PACS: 02.50.Tt, 02.50.-r, 07.85.Nc, 29.30.Kv.

INTRODUCTION

A common problem in quantitative analysis of physical spectra is the ability of separating the data of interest from a smooth background. As reported in [1], there are several methods for background subtraction in the literature. Most of them rely on polynomial or semi-empirical nonlinear fitting procedures. These techniques suffer from unpredictable reliability and require human monitoring. A great improvement is given in [1, 2] where authors follow a probabilistic approach based on prior knowledge – albeit vague – of the observed data model. The key idea in [1, 2] is to use a prior for cubic spline values ensuring that the background is smoother than the "peaks" information. The prior is chosen to emphasize solutions which minimize the integral of the second derivative of the background. Yielding a rigorous framework, this Bayesian approach offers better properties than any *ad hoc* method.

Our motivation in this contribution is basically the same as authors in [1, 2], but we address here the additional problem of ensuring that both background and residual peaks spectrum after background subtraction remain probability measures. We focus thus on the estimation of a two component mixture model. Referring to physical considerations, we think reasonable to consider a latent discrete mixture model underlying the peaks spectrum while background data sit on an absolutely continuous distribution.

Indeed, quantum physics says that energy levels of x/γ-rays lead to a discrete spectrum. The actual number of nuclei and of energy levels, so the number of peaks, is most of the time unknown (specially in exotic nuclei characterization experiments). Any attempt to specify this number could lead to dramatic physical misinterpretations. It is therefore appealing to leave this number open-ended in our model. This important issue has been addressed [3] following a reversible jump MCMC method (see [4]).

Besides theoretical considerations, the experimental setup increases the complexity of an "ideal" quantum levels spectrum. In the field of x/γ-ray spectrometry, the particle detection process induced by semiconductors or scintillators devices entail several kinds of interactions, mainly the photoelectric absorption, pair creation and the Compton scattering effects [5]. Photoelectric absorption and pair creation lead to a discrete energy distribution whereas Compton scattering produce a continuous spectrum. Energies are not directly observed due to detection devices noises which introduce a convolution of both discrete and continuous measures by a parametrized kernel, typically normal with an energy dependent variance. Moreover, other interactions processes (e.g. binding, backscattering, multiple scattering, bremsstrahlung escape) induce more complex features for the observed energy distribution. These phenomena highlight the actual complexity of the overall data distribution. The paper is organized as follows: first, we review two nonparametric Bayesian tools for density estimation. We then give a hierarchical model for physical spectra and sketch out a sampling algorithm which is illustrated on experimental data. Detailed expressions of the posterior distributions involved in the proposed algorithms are defered to a longer contribution.

BAYESIAN NONPARAMETRIC DENSITY ESTIMATION

Thanks to their ability to capture the structure of complex data, nonparametric and semiparametric statistical models appear more and more a relevant alternative to parametric models. Very good tutorials [6, 7, 8, 9] bring out the wide variety of fields covered by the approach. A nonparametric or semiparametric model involve one or more infinite dimensional parameters. This is particularly suitable for estimation in the space of probability measures on $\mathbb{R}$ ($\mathbb{R}^+$ for energy distribution). Nonparametric Bayesian density estimation turns out to be quite appealing in physical sciences. First, it is well known that the Bayesian approach allows to embed our physical prior belief in the analysis. Second, predictive distribution, and more generally spectral inference, can be obtained in a straightforward manner by means of posterior draws.

Besides these advantages there are some difficulties in the use of Bayesian nonparametrics, and analysis has to be undertaken with care [10]. First, construction and elicitation of priors are difficult in infinite dimension space. Second, original MCMC schemes have to be developed to work with infinite dimensional models. As we will show, some truncation mechanisms may be introduced to allow implementation of the nonparametric approach [11]. Finally, general consistency results obtained in parametric Bayesian inference do not exist in the nonparametric context. Consistency can be roughly viewed as a kind of frequentist validation of Bayesian method ensuring that a particular choice of prior is overridden by the observation of a sufficient amount of data [12, 10].

As we characterized the peaks spectrum by a latent discrete model, we look forward a nonparametric prior on discrete measures. This is precisely a feature of the well known Dirichlet process (denoted DP in the followings) introduced by Ferguson [13]. On the other hand, we need to handle priors on continuous distributions to model the latent continuous physical background and will take an interest in Pólya trees (denoted PT), also introduced by Ferguson [14] and developed by Lavine [15] and Mauldin *et al.* [16]. For our purpose, it might be underlined that due to instrumental convolutions both DP and PT will be used as mixing distributions and that the mixture model always admits a probability density function.

Truncated Dirichlet process mixtures

Due to this discreteness of generated measures, DP cannot be used as a prior for estimating a probability density function. Nevertheless, the convolution of the DP by a parametric kernel (e.g. normal, gamma, Weibull kernel) produces continuous densities. Dirichlet process mixtures (DPM), introduced by Antoniak [17] play an important role in nonparametric density estimation and clustering problems.

Despite their elegant hierarchical representation, posterior distribution of DPM are analytically intractable and MCMC techniques are required for inference. An important literature deals with MCMC methods based on Pólya urn Gibbs sampling, see for example [18, 19, 20, 21]. However, this marginalization technique reveals quite slow mixing when working with huge datasets.

This limitation can be avoided by using the blocked Gibbs sampler developed in [11]. One requirement is to work with a truncated approximation of the DP derived from the stick-breaking representation introduced by Sethuraman [22].

The idea is to truncate the infinite summation of Sethuraman representation after a chosen value N. We refer to the work of Ishwaran and James [11] in this section who showed that the finite DP $\mathcal{P}_N$ converges almost surely to a Dirichlet process with mean distribution G_0 and precision parameter α. We define then a truncated Dirichlet mixture for observed data $\mathbf{X} = (X_1, X_2, \ldots, X_n)$,

$$\begin{aligned} (X_i|\mathbf{Z},\mathbf{K}) &\overset{\text{ind}}{\sim} \mu\left(X_i|Z_{K_i}\right) \\ (K_i|\mathbf{p}) &\overset{\text{i.i.d.}}{\sim} \sum_{k=1}^{N} p_k\,\delta_k(\cdot) \\ (\mathbf{p},\mathbf{Z}) &\sim \mu(\mathbf{p}) \times (G_0)^N \end{aligned} \tag{1}$$

where $\mathbf{K} = (K_1, K_2, \ldots, K_n)$ represents the classification vector and $\mathbf{Z} = (Z_1, Z_2, \ldots, Z_N)$ the location vector of the DP : for $i \leq n,\ K_i = j$ if $\{\text{hidden alloted location of } X_i\} = Z_j$. We also define the vector of weights of the finite stick-breaking $\mathbf{p} = (p_1, p_2, \ldots, p_N)$ with distribution $\mu(\mathbf{p})$ such that $V_1, V_2, \ldots$ are i.i.d. $\text{Beta}(1,\alpha)$, $p_1 = V_1$ and for all $k \geq 2,\ p_k = V_k \prod_{i=1}^{k-1}(1-V_i)$. Finally, $\mu\left(X_i|Z_{K_i}\right)$ is the conditional distribution with kernel density $\psi\left(X_i|Z_{K_i}\right)$ and $\delta_\theta(\cdot)$ denotes a discrete measure concentrated at θ.

The form (1) allows direct sampling from $\mathcal{P}_N(\cdot|\mathbf{X})$ by iteratively drawing samples from the conditionals of the blocked variables.

$$\begin{gathered} (\mathbf{Z}|\mathbf{K},\mathbf{X}) \\ (\mathbf{K}|\mathbf{Z},\mathbf{p},\mathbf{X}) \\ (\mathbf{p}|\mathbf{K}) \end{gathered} \tag{2}$$

After convergence of the Markov Chain, the blocked Gibbs sampler produces draws from $(\mathbf{Z},\mathbf{K},\mathbf{p}|\mathbf{X})$. From draws $(\mathbf{Z}^*,\mathbf{K}^*,\mathbf{p}^*)$ we build the random probability measure

$$\mathcal{P}_N^*(\cdot) = \sum_{k=1}^{N} p_k^*\,\delta_{Z_k^*}(\cdot) \tag{3}$$

which is a draw from the posterior $\mathcal{P}_N|\mathbf{X}$. Recording generated $\mathcal{P}_N^*(\cdot)$, we are able to estimate $\mathcal{P}_N|\mathbf{X}$ and its functionals.

Details for drawing from the conditional distributions of the Gibbs sampler are given in [11, 23]. The key of the algorithm is the finite dimensionality of the prior which allows to write the model using a finite number of random variables.

Finite Pólya tree process mixtures

In comparison to the profuse literature dealing with Dirichlet mixtures, the Pólya tree process seems to have been neglected for the estimation of probability density functions. However, this random distribution prior offers several nice properties. We refer to Lavine [15] for Pólya tree definition and complete description. First of all, unlike Dirichlet processes, we can choose parameters of PT to give absolute continuity to the generated random distributions by observing the Kraft inequality.

A PT process (denoted $\mathrm{PT}(\mathcal{A},\Pi)$) with parameters $\mathcal{A}$ generates random distributions leaning on a separating binary tree of partitions Π of the measurable space so that sets at level $m+1$ are obtained by a binary split of the sets of level m. The partition construction can be continued indefinitely but since it is not possible to compute with infinite trees, Pólya trees are often simplified by only specifying parameters and constructing the partition until a given finite level M. We denote by $\mathcal{B}_j$ the j^{th} subset at level M.

Unlike DP, Pólya trees are dependent on the underlying partition. Furthermore, even if partitions and parameters are suitably chosen for ensuring continuity of generated random distributions, lack of smoothness appears at partition endpoints. This drawback tends to limit the use of PT in density estimation. To overcome this difficulty, PT mixtures [15, 24, 25] have been developed. The idea is to allow the sequence of partitions and the parameters to depend on a random parameter. We propose here a special kind of randomized finite Pólya tree that shifts the whole partitions tree on a wider space. We assume a non-informative center distribution for the PT such that the Lebesgue measure of all $\mathcal{B}_j$ is equal to λ_M. The shifts are then taken equal to $d\cdot\lambda_M$ with $d\in\{0,\dots,\Delta\}$. Given d, the Pólya tree is denoted : $\mathrm{PT}(\mathcal{A}_d,\Pi_d)$.

A second hierarchical mixture level, based on the convolution with a parametric kernel ψ, may be introduced.

For observations $\mathbf{X}=\{X_1,\dots,X_n\}$, we define $\mathbf{K}=\{K_1,\dots,K_n\}\in\left\{1,\dots,2^M-\Delta\right\}^n$ a classification variable such that $K_i=j$ if $\{\text{hidden alloted subset of } X_i\}=\mathcal{B}_j$.

$$\begin{aligned}
(X_i|\mathbf{K}) &\overset{\text{ind}}{\sim} \mu\left(X_i|\mathcal{B}_{K_i}\right)\\
(K_i|\mathbf{q}) &\overset{\text{i.i.d}}{\sim} \sum_{k=1}^{2^M-\Delta} q_k\delta_k\\
G|d &\sim \mathrm{PT}\left(\mathcal{A}_d,\Pi_d\right),\ q_j=G\left(\mathcal{B}_j\right)\\
d &\sim \frac{1}{1+\Delta}\sum_{l=0}^{\Delta}\delta_l\left(d\right)
\end{aligned}\tag{4}$$

where $\mathbf{q} = \left\{q_j \,:\, j \in \left\{1, \ldots, 2^M - \Delta\right\}\right\}$, PT $(\mathcal{A}_d, \Pi_d)$ is expressed in [26] in this proceedings, and for $j \in \left\{1, \ldots, 2^M - \Delta\right\}$, $\mu\left(X_i | \mathcal{B}_j\right)$ has density $\widetilde{\psi}\left(\cdot | \mathcal{B}_j\right) = \psi \star f_{\mathcal{B}_j}$ where $f_{\mathcal{B}_j}$ is the density of the uniform distribution on the interval $\mathcal{B}_j$ and ψ the kernel density.

Unfortunately, in this case posterior distribution is intractable and we have to settle to sampling techniques.

Readers may notice similarities between model (4) and (1). Indeed, we propose a blocked Gibbs sampler inspired from [11] for this hierarchical mixture of PT. For sampling we need to express conditionals for

$$\begin{array}{c} (\mathbf{K} | \mathbf{q}, d, \mathbf{X}) \\ (\mathbf{q}, d | \mathbf{K}) \end{array} \tag{5}$$

The conditional for $(\mathbf{K} | \mathbf{q}, d, \mathbf{X})$ is similar to the allocation step of DP mixtures proposed in [11]. Draws from $(\mathbf{q}, d | \mathbf{K})$ are obtained either involving a Metropolis-Hastings step in the sampler or by first sampling $(d | \mathbf{K})$ then $(\mathbf{q} | \mathbf{K}, d)$. In the latter solution, the step can be dramatically speed up if we assume that PT mixtures are smooth enough so that posterior sampling of d can be assimilated to a discrete uniform distribution. This approximation turns out to be very attractive since our aim is mainly to avoid partition endpoints discontinuities.

Now that probability density mixture priors based on either discrete or absolute continuous mixing distributions are presented we turn on a hierarchical model dedicated to physical spectra inference.

SEMIPARAMETRIC MODEL FOR EXPERIMENTAL SPECTRA ANALYSIS

Dirichlet mixtures meet the peaks spectrum requirements as the latent DP prior generates discrete mixing distributions with eventually infinitely many locations.

Note that there is no guarantee that each DP location corresponds to a radio-nucleus peak location. Indeed, we underlined in the introduction that detection devices may involve, even for photoelectric absorption, other phenomena which, after kernel deconvolution, take away discreteness of energy distribution. But we assume that the DP locations preserve reasonably the energy discrete distribution.

On the other hand, the data belonging to the background spectrum relies on the continuity of the mixing distribution of a PT mixture with the same kernel ψ as DPM. We propose a hierarchical construction where the parameters of the PT may depend on the DP locations. Using such a model, physicists can embed any kind of analytic knowledge about the physical quantity spectrum as the mean distribution of the Pólya tree. This is especially interesting when we consider that many analytic expressions are most of the time parametric approximations of usually more complex phenomena. The complexity may be induced by the physical mechanisms themselves or by interactions with the environment. In one hand, the approach let the analyst rely upon his understanding of the phenomenon by means of an approximated analytical description introduced as the mean distribution of the quantity. In the other hand, his limited degree of belief will be tackled by nonparametric Bayesian random distributions.

Gibbs sampler for physical spectra estimation

The key idea of the approach described here is that by means of finite stick breaking representation for DP processes and finite PT processes, we cast the problem into a random variables formulation where priors of both parts are involved in the allocation phase. The classification vector $\mathbf{K}$ is now extended to cover either DP components or PT subsets.

$$\text{for } i \in \{1,\dots,n\}, \quad K_i = \begin{cases} j & \text{if } \{\text{hidden alloted DP location of } X_i\} = Z_j \\ j+N & \text{if } \{\text{hidden alloted PT subset of } X_i\} = \mathcal{B}_j \end{cases} \tag{6}$$

Using notations of (1) and (4) we propose the following hierarchical model

$$\begin{aligned} (X_i|\mathbf{Z},\mathbf{K},\eta) &\overset{\text{ind}}{\sim} \begin{cases} \mu\left(X_i|Z_{K_i},\eta\right) & \text{if } K_i \leq N \\ \mu\left(X_i|\mathcal{B}_{K_i-N},\eta\right) & \text{otherwise} \end{cases} \\ (K_i|\mathbf{p},\mathbf{q}) &\overset{\text{i.i.d.}}{\sim} w\sum_{k=1}^{N} p_k\,\delta_k(\cdot) + (1-w)\sum_{k=1}^{2^M-\Delta} q_k\delta_{k+N}(\cdot) \\ G|d &\sim \mathrm{PT}\left(\mathcal{A}_d^{\mathbf{p},\mathbf{Z}},\Pi_d\right),\ q_j = G\left(\mathcal{B}_j\right) \\ d &\sim \frac{1}{1+\Delta}\sum_{l=0}^{\Delta}\delta_l(d) \\ (\mathbf{p},\mathbf{Z}) &\sim \mu(\mathbf{p})\times(G_0)^N \\ w &\sim \mathrm{Beta}\left(\nu_P,\nu_B\right) \\ \eta &\sim H \end{aligned} \tag{7}$$

where tree parameters $\mathcal{A}_d^{\mathbf{p},\mathbf{Z}}$ illustate the DP dependence.

The balance between the DP peaks spectrum and the PT background is handled by the variable w which is beta$\left(\nu_P,\nu_B\right)$ distributed.

An additional nuisance parameter η with distribution H is introduced to tackle instrumental parameters like the energy dependence of the spectrum resolution.

Since model (7) belongs to the same class as (1) and (4) a blocked Gibbs sampler for spectrum inference can thus be constructed in the same way and will successively draw samples from

$$\begin{aligned} &(\mathbf{Z}|\mathbf{K},\eta,\mathbf{X}) \\ &(\mathbf{K}|\mathbf{Z},\mathbf{p},\mathbf{q},d,w,\eta,\mathbf{X}) \\ &(\mathbf{q}|\mathbf{K},\mathbf{p},\mathbf{Z},d) \\ &(d|\mathbf{K}) \\ &(\mathbf{p}|\mathbf{K}) \\ &(w|\mathbf{K}) \\ &(\eta|\mathbf{K},\mathbf{Z},\mathbf{X}) \end{aligned} \tag{8}$$

Applications

We present results of the described Gibbs sampler (8) on an experimental spectrum of a Uranium oxide ore sample. Whole observed data correspond to a single binned histogram on the range [0 KeV, 2.5 MeV]. When only binned data are observed, the algorithm can be adapted with minor modifications. The kernel densities of both parts become probability mass functions of the observed bins by integration of ψ and $\widetilde{\psi}$ over the binwidth and the allocation step of the Gibbs sampler becomes a multinomial distribution where we break down simultaneously all the counts of a given bin. This binned version appears computationally attractive for huge datasets.

We focus on two distinct regions of interest. The first region [400 KeV, 630 KeV] (R1) exhibits a high background. Remark that a common problem for physicists is the “detection limit” [5] of small peaks on such a background. The second range [2.25 MeV, 2.47 MeV] (R2) presents a situation with a smaller dataset.

We analyze these two regions separately. For both spectra all parameters are taken identical : DP concentration parameter $\alpha = 1$ and $N = 100$. We use a canonical PT (see [15, 6]) with $M = 10$, $\mathcal{A} = \{a_m = 6^m : m \leq M\}$ and $\Delta = 128$. We averaged 20000 draws of the Gibbs sampler after 10000 burn-in iterations. Conditional means of the mixing distributions are plotted on Figure 1 and Figure 2.

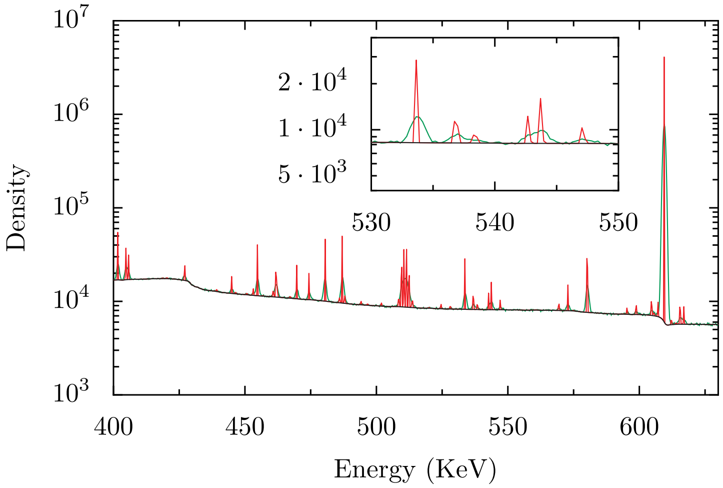

FIGURE 1. Uranium oxide (UO_2) energy spectrum. (R1) region : histogram (green), PT background spectrum (black), DP peaks spectrum (red).

CONCLUSION

We proposed a hierarchical model for physical spectra which allows efficient Gibbs sampling. As expected by the mixed DP / PT model, experimental results show separation capabilities of the algorithm. In addition, posterior draws of DP and PT random mixing distributions produce separated nonparametric deconvoluted spectra. A consistency study of the estimator will be developed in future works.

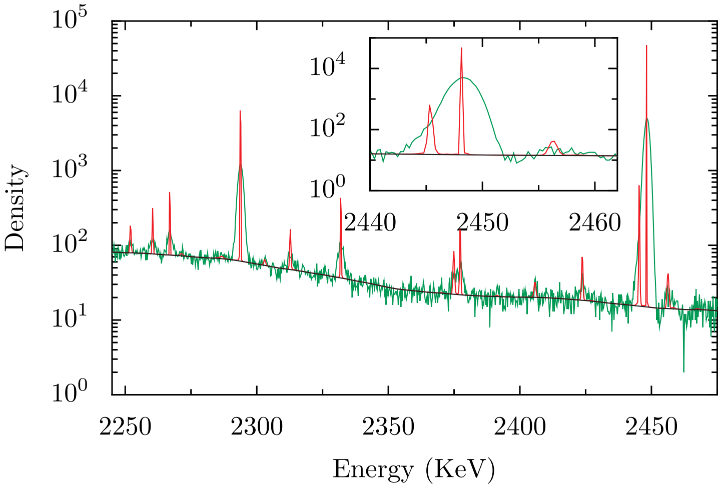

FIGURE 2. Uranium oxide (UO_2) energy spectrum. (R2) region : histogram (green), PT background spectrum (black), DP peaks spectrum (red).

REFERENCES

1. W. von der Linden, V. Dose, J. Padayachee, and V. Prozesky, *Phys. Rev. E* **59**, 6527–6534 (1999).
2. R. Fischer, K. M. Hanson, V. Dose, and W. von der Linden, *Phys. Rev. E* **61**, 1152–1160 (2000).
3. R. Fischer, and V. Dose, "Physical mixture modeling with unknown number of components," in *Bayesian Inference and Maximum Entropy Methods in Science and Engineering*, 2002, vol. 617, pp. 143–154.
4. P. J. Green, *Biometrika* **82**, 711–732 (1995).
5. G. F. Knoll, *Radiation detection and measurement*, Wiley, 1989, 2 edn.
6. N. Choudhuri, S. Ghosal, and A. Roy, *Handbook of statistics* **25**, 373–414 (2005).
7. Z. Gharamani, "Tutorial on Nonparametric Bayesian Methods," in *UAI*, 2005.
8. R. M. Neal, "Bayesian Methods for Machine Learning," in *NIPS 2004*, 2004.
9. M. I. Jordan, "Dirichlet processes, Chinese restaurant processes and all that," in *NIPS 2005*, 2005.
10. J. K. Ghosh, and R. V. Ramamoorthi, *Bayesian Nonparametrics*, Springer, 2003.
11. H. Ishwaran, and L. F. James, *J. Am. Stat. Assoc.* **96**, 161–173 (2001).
12. A. Barron, M. J. Schervish, and L. Wasserman, *Ann. Statist.* **27**, 536–561 (1999).
13. T. S. Ferguson, *Ann. Statist.* **1**, 209–230 (1973).
14. T. S. Ferguson, *Ann. Statist.* **2**, 615–629 (1974).
15. M. Lavine, *Ann. Statist.* **20**, 1222–1235 (1992).
16. R. D. Mauldin, W. D. Sudderth, and S. C. Williams, *Ann. Statist.* **20**, 1203–1221 (1992).
17. C. E. Antoniak, *Ann. Statist.* **2**, 1152–1174 (1974).
18. M. D. Escobar, *J. Am. Stat. Assoc.* **89**, 268–277 (1994).
19. M. D. Escobar, and M. West, *J. Am. Stat. Assoc.* **90**, 577–588 (1995).
20. S. N. Mac Eachern, and P. Müller, *J. Comput. Graph. Stat.* **7**, 223–238 (1998).
21. S. Jain, and R. Neal, *J. Comput. Graph. Stat.* **13**, 158–182 (2004).
22. J. Sethuraman, *Stat. Sinica* **4**, 639–650 (1994).
23. H. Ishwaran, and L. F. James, *J. Comput. Graph. Stat.* **11**, 508–532 (2002).
24. T. Hanson, and W. O. Johnson, *J. Am. Stat. Assoc.* **97**, 1020–1033 (2002).
25. S. M. Paddock, F. Ruggeri, M. Lavine, and M. West, *Stat. Sininica* **13**, 443–460 (2003).
26. E. Barat, T. Dautremer, and T. Trigano, "Nonparametric Bayesian estimation of censored counter intensity from the indicator data," in *Bayesian Inference and Maximum Entropy Methods in Science and Engineering*, 2006.