Bayesian nonparametric estimation in the current status continuous mark model

Geurt Jongbloed, Frank van der Meulen and Lixue Pang

Delft University of Technology

March 15, 2021

Abstract

In this paper we consider the current status continuous mark model where, if the event takes place before an inspection time T a "continuous mark" variable is observed as well. A Bayesian nonparametric method is introduced for estimating the distribution function of the joint distribution of the event time (X) and mark variable (Y). We consider a prior that is obtained by assigning a distribution on heights of cells, where cells are obtained from a partition of the support of the density of (X, Y). As distribution on cell heights we consider both a Dirichlet prior and a prior based on the graph-Laplacian on the specified partition. Our main result shows that under appropriate conditions, the posterior distribution function contracts pointwisely at rate $(n/\log n)^{-\frac{\rho}{3(\rho+2)}}$, where ρ is the Hölder smoothness of the true density. In addition to our theoretical results we provide efficient computational methods for drawing from the posterior relying on a non-centred parameterisation and Crank-Nicolson updates. The performance of our computational methods is illustrated in several numerical experiments.

1 Introduction

1.1 Problem formulation

Survival analysis is concerned with statistical modelling of the time until a particular event occurs. The event may for example be the onset of a disease or failure of equipment. Rather than observing the time of event exactly, censoring is common in practice. If the event time is only observed when it occurs prior to a specific (censoring) time, one speaks of right censoring. In case it is only known whether the event took place before an inspection time or not, one speaks of current status censoring. The resulting data are then called current status data.

In this paper we consider the current status continuous mark model where, if the event takes place before an inspection time T, a "continuous mark" variable is observed as well. More specifically, denote the event time by X and the mark by Y. Independent of (X, Y), there is an inspection time T with density function g on $[0, \infty)$. Instead of observing each (X, Y) directly, we observe the inspection time T together with the information whether the event occurred before time T or not. If it did so, the additional mark random variable Y is also observed, for which we assume $P(Y = 0) = 0$. Hence, an observation of this experiment can be denoted by $W = (T, Z) = (T, \Delta \cdot Y)$ where $\Delta = \mathbf{1}_{\{X \le T\}}$ (note that, equivalently, $\Delta = \mathbf{1}_{\{Z>0\}}$). This experiment is repeated n times independently, leading to the observation set $\mathcal{D}_n = \{W_i, i = 1, \dots, n\}$. We are interested in estimating the joint distribution function F_0 of (X, Y) nonparametrically, based on $\mathcal{D}_n$.

An application of this model is the HIV vaccine trial studied by Hudgens, Maathuis & Gilbert (2007). Here, the mark is a specifically defined viral distance that is only observed if a participant to the trial got HIV infected before the moment of inspection.

1.2 Related literature

In this section we review earlier research efforts on models closely related to that considered here.

Survival analysis with a continuous mark can be viewed as the continuous version of the classical competing risks model. In the latter model, failure is due to either of K competing risks (with K fixed) leading to a mark value that is of categorical type. As the mark variable encodes the cause of failure it is only observed if failure has occurred before inspection. These "cause events" are known as competing risks. Groeneboom, Maathuis & Wellner (2008) study nonparametric estimation for current status data with competing risks. In that paper, they show that the nonparametric maximum likelihood estimator (NPMLE) is consistent and converges globally and locally at rate $n^{1/3}$.

Huang & Louis (1998) consider the continuous mark model under right-censoring, which is more informative compared to the current-status case because the exact event time is observed for noncensored data. For the nonparametric maximum likelihood estimator of the joint distribution function of (X, Y) at a fixed point, asymptotic normality is shown.

Hudgens, Maathuis & Gilbert (2007) consider interval censoring case k, $k = 1$ being the specific setting of current-status data considered here. In this paper the authors show that both the NPMLE and a newly introduced estimator termed "midpoint imputation MLE" are inconsistent. However, coarsening the mark variable (i.e. making it discrete, turning the setting to that of the competing risks model), leads to a consistent NPMLE. This is in agreement with the results in Maathuis & Wellner (2008).

Groeneboom, Jongbloed & Witte (2011) and Groeneboom, Jongbloed & Witte (2012) consider the exact setting of this paper using frequentist estimation methods. In Groeneboom, Jongbloed & Witte (2011) two plug-in inverse estimators are proposed. They prove that these estimators are consistent and derive the pointwise asymptotic distribution of both estimators. Groeneboom, Jongbloed & Witte (2012) define a nonparametric estimator for the distribution function at a fixed point by finding the maximiser of a smoothed version of the log-likelihood. Pointwise consistency of the estimator is established. In both papers numerical illustrations are included.

1.3 Contribution

In this paper, we consider Bayesian nonparametric estimation of the bivariate distribution function F_0 in the current status continuous mark model. Approaching this problem within the Bayesian setup has not been done before, neither from a theoretical nor computational perspective. Whereas consistent nonparametric estimators exist within frequentist statistics, convergence rates are unknown. We prove consistency and derive Bayesian contraction rates for the bivariate distribution function of (X, Y) using a prior on the joint density f of (X, Y) that is piecewise constant. For the values on the bins we consider two different prior specifications. The first of these is defined by equipping the bin probabilities with the Dirichlet-distribution. The other specification is close to a logistic-normal distribution (Aitchison and Shen (1980)), where additionally smoothness on nearby bin-probabilities is enforced by taking the precision matrix of the Normal distribution equal to the graph-Laplacian induced by the grid of bins. The graph-Laplacian is a well known method to induce smoothness, see for instance Murphy (2013) (Chapter 25.4) or Hartog & van Zanten (2017) for an application in Bayesian estimation. Full details of the prior specification are in Section 2.2.

Our main result shows that under appropriate conditions, the posterior distribution function contracts pointwisely at rate $(n/\log n)^{-\frac{\rho}{3(\rho+2)}}$, where ρ is the Hölder smoothness of the true density. In this result, we assume that for the prior the bins areas where the density is constant tends to zero at an appropriate (non-adaptive) rate, as $n \to \infty$. The proof is based on general results from Ghosal & Van der Vaart (2017) for obtaining Bayesian contraction rates. Essentially, it requires the derivation of suitable test functions and proving that the prior puts sufficient mass in a neighbourhood of the "true" bivariate distribution. The latter is proved by exploiting the specific structure of our prior.

In addition to our theoretical results, we provide computational methods for drawing from the posterior. For the Dirichlet prior this is a simple data-augmentation scheme. For the graph-Laplacian prior we provide simple code to draw from the posterior using probabilistic programming in the Turing Language under Julia (see Bezanson et al. (2017), Ge, Xu & Ghahramani (2018)). Additionally, a much faster algorithm is introduced that more carefully exploits the structure of the problem. The main idea is to use a non-centred parameterisation (Papaspiliopoulos et al. (2003)) combined with a preconditioned Crank-Nicolson (pCN) scheme (Cf. Cotter et al. (2013)). The performance of our computational methods is illustrated in two examples. Code is available from https://github.com/fmeulen/CurrentStatusContinuousMarks.

1.4 Outline

The outline of this paper is as follows. In section 2 we introduce further notation for the current status continuous mark model and detail the two priors considered. Subsequently, we present the main theorem on the posterior contraction rate in Section 3. The proof of this result is given in Section 4. In Section 5 we present MCMC-algorithms to draw from the posterior and give numerical illustrations, including a simulation study.

1.5 Notation

For two sequences $\{a_n\}$ and $\{b_n\}$ of positive real numbers, the notation $a_n \lesssim b_n$ (or $b_n \gtrsim a_n$) means that there exists a constant $C > 0$, independent of n, such that $a_n \leq Cb_n$. We write $a_n \asymp b_n$ if both $a_n \lesssim b_n$ and $a_n \gtrsim b_n$ hold. We denote by F and F_0 the cumulative distribution functions corresponding to the probability densities f and f_0 respectively. The Hellinger distance between two densities f, g is written as $h^2(f,g) = \frac{1}{2}\int(f^{1/2} - g^{1/2})^2$. The Kullback-Leibler divergence of f and g and the L_2-norm of $\log(f/g)$ (under f) by

$$KL(f,g) = \int f \log\frac{f}{g}, \qquad V(f,g) = \int f\left(\log\frac{f}{g}\right)^2.$$

2 Likelihood and prior specification

2.1 Likelihood

In this section we derive the likelihood for the joint density f based on data $\mathcal{D}_n$. As $W_1, \ldots, W_n$ are independent and identically distributed, it suffices to derive the joint density of $W_1 = (T_1, Z_1)$ (with respect to an appropriate dominating measure). Recall that f denotes the density of (X,Y). Let F denote the corresponding distribution function of (X,Y). The marginal distribution function of X is given by

$$F_X(t) = \int_0^t \int_0^\infty f(u,v)\,\mathrm{d}v\,\mathrm{d}u.$$

Define the measure μ on $[0,\infty)^2$ by

$$\mu(B) = \mu_2(B) + \mu_1(\{x \in [0,\infty) : (x,0) \in B\}), \qquad B \in \mathcal{B} \tag{1}$$

where $\mathcal{B}$ is the Borel $\sigma-$algebra on $[0,\infty)^2$ and μ_i is Lebesgue measure on $\mathbb{R}^i$. The density of the law of W_1 with respect to μ is then given by

$$s_f(t,z) = g(t)\left(\mathbf{1}_{\{z>0\}}\partial_2 F(t,z) + \mathbf{1}_{\{z=0\}}(1 - F_X(t))\right), \tag{2}$$

where

$$\partial_2 F(t,z) = \frac{\partial}{\partial z}F(t,z) = \int_0^t f(u,z)\,\mathrm{d}u.$$

By independence of $W_1, \ldots, W_n$, the likelihood of f based on $\mathcal{D}_n$ is given by $l(f) = \prod_{i=1}^n s_f(T_i, Z_i)$.

2.2 Prior specification

In this section, we define a prior on the class of all bivariate density functions on $\mathbb{R}^2$. Denote

$$\mathcal{F} = \left\{ f : \mathbb{R}^2 \to [0, \infty) : \int_{\mathbb{R}^2} f(x, y)\, \mathrm{d}x\, \mathrm{d}y = 1 \right\}.$$

For any $f \in \mathcal{F}$, if S denotes the support of f and $\cup_j C_j, j = 1, \ldots, p_n$ is a partition of S, we define a prior on $\mathcal{F}$ by

$$f_{\boldsymbol{\theta}}(x, y) = \sum_j \frac{\theta_j}{|C_j|} \mathbf{1}_{C_j}(x, y), \qquad (x, y) \in \mathbb{R}^2,$$

where $|C|= \mu_2(C)$ is the Lebesgue measure of the set C and $\boldsymbol{\theta} = (\theta_1, \ldots, \theta_{p_n})$. We require that all θ_j are nonnegative and $\sum_j \theta_j = 1$ (i.e. $\boldsymbol{\theta}$ is a probability vector). We consider two types of prior on $\boldsymbol{\theta}$.

1. *Dirichlet.* For a fixed parameter $\alpha = (\alpha_1, \ldots, \alpha_{p_n})$ consider $\boldsymbol{\theta} \sim \text{Dirichlet}(\alpha)$. This prior is attractive as draws from the posterior distribution can be obtained using a straightforward data-augmentation algorithm (Cf. Section 5.1). We will refer to this prior as the *D-prior*.

2. *Logistic-Normal with graph-Laplacian precision matrix.* For a positive-definite matrix Υ, assume that the random vector $\boldsymbol{H}$ satisfies $\boldsymbol{H} \sim N_{p_n}(0, \tau^{-1}\Upsilon^{-1})$, for fixed positive τ. Next, set

$$\theta_j = \frac{\psi(H_j)}{\sum_{j=1}^{p_n} \psi(H_j)}, \quad \text{where} \quad \psi(x) = e^x. \tag{3}$$

That is, we transform $\boldsymbol{H}$ by the softmax function implying that realisations of $\boldsymbol{\theta}$ are probability vectors. The matrix Υ is chosen as follows. The partition $\cup_j C_j$ induces a graph structure on the bins, where each bin corresponds to a node in the graph, and nodes are connected when bins are adjacent (meaning that they are either horizontal or vertical "neighbours"). Let L denote the graph-Laplacian of the graph obtained in this way. This is the $p_n \times p_n$ matrix given by

$$L_{i,i'} = \begin{cases} \text{degree node } i & \text{if } i = i' \\ -1 & \text{if } i \neq i' \text{ and nodes } i \text{ and } i' \text{ are connected .} \\ 0 & \text{otherwise.} \end{cases} \tag{4}$$

We take

$$\Upsilon = L + p_n^{-2} I.$$

We will refer to this prior as the *LNGL-prior* (Logistic-Normal Graph Laplacian).

2.1 Remark Under the Dirichlet prior values of θ_j in adjacent bins are negatively correlated, preventing the density to capture smoothness. This is illustrated in the numerical study Section 5. We would like to take a large number of bins, while not overparametrising. The idea of the graph-Laplacian prior is to induce positive correlation on adjacent bins and thereby specify a prior that produces smoother realisations. The numerical illustrations reveal that the posterior based on the LNGL-prior are less sensitive to the chosen number of bins in the partition, compared to the posterior based on the D-prior.

Hjort (1994) (Section 2.4) has proposed a modification to the Dirichlet prior to induces smoothness among nearby bins. These generalised Dirichlet priors on probability vectors in $\mathbb{R}^k$ take the form

$$\pi(p_1, \ldots, p_{k-1}) \propto p_1^{\alpha_1 - 1} \cdot p_k^{\alpha_k - 1} \exp\left(-\tau \Delta(\boldsymbol{p})\right),$$

where $\boldsymbol{p} := (p_1, \ldots, p_k)$ is a probability vector and $\tau > 0$. Small values of $\Delta(\boldsymbol{p})$ indicates a certain characteristic is present. To connect it to the graph-Laplacian, one choice is indeed to take

$$\Delta(\boldsymbol{p}) = \boldsymbol{p}^T \Upsilon \boldsymbol{p}.$$

Just as for the LNGL-prior, MCMC methods can be used to sample from the posterior in case of the generalised Dirichlet prior. The difference in the two approaches consists of imposing smoothness on the θ_j directly (as in case of the generalised Dirichlet prior), or in terms of H_j, followed by applying the transformation in Equation (3). We feel the approach we take in this paper is conceptually somewhat simpler. Moreover, it allows to use MCMC methods where the prior is obtained in a simple way as the pushforward of a vector of independent standard Normal random variables.

2.2 Remark For both generalised Dirichlet prior and the LNGL-prior, the posterior mode has the usual interpretation of a penalised likelihood estimator. The choice of τ controls the amount of smoothing. In a Bayesian setting, uncertainty on τ is dealt with by employing an additional prior on τ (in the numerical section we will follow this approach). For the LNGL-prior another handle to control smoothness is be obtained by considering Υ^r, where $r \geq 1$. Hartog & van Zanten (2017) study the effect of r in a simulation study in the setting of classification on a graph and advise to take $r = 1$ or $r = 2$ to obtain good performance empirically. In this paper $r = 1$ throughout. However, the numerical methods apply to any value $r \geq 1$.

2.3 Remark One can argue whether the presented prior specifications are truly nonparametric. It is not if one adopts as definition that the size of the parameter should be learned by the data. For that, a solution could be to put a prior on p_n as well. While possible, this would severely complicate drawing from the posterior. As an alternative, one can take large values of p_n (so that the model is high-dimensional), and let the data determine the amount of smoothing by incorporating flexibility in the prior. As the Dirichlet prior lacks smoothness properties, fixing large values of p_n will lead to overparametrisation, resulting in high variance estimates (under smoothing). On the contrary, as we will show in the numerical examples, for the LNGL-prior, this overparametrisation can be substantially balanced/regularised by equipping the parameter τ with a prior distribution. The idea of histogram type priors with positively correlated adjacent bins has recently been used successfully in other settings as well, see for instance Gugushvili et al. (2018), Gugushvili et al. (2019).

2.4 Remark Ting et al. (2010) consider the limiting behaviour of the LNGL-prior under mesh refinement. Unsurprisingly, in the limit it behaves like the "ordinary" Laplace operator which is the infinitesimal generator of a driftles diffusion.

3 Posterior contraction

In this section we derive a contraction rate for the posterior distribution of F_0. Denote the posterior measure by $\Pi_n(\cdot|\mathcal{D}_n)$ (under the prior measure Π_n described in Section 2.2).

3.1 Assumption The underlying joint density of the event time and mark, f_0, has compact support given by $\mathcal{M} = [0, M_1] \times [0, M_2]$ and is ρ-Hölder continuous on $\mathcal{M}$. That is, there exists a positive constant c and a $\rho \in (0, 1]$ such that for any $\boldsymbol{x}, \boldsymbol{y} \in \mathcal{M}$,

$$|f_0(\boldsymbol{x}) - f_0(\boldsymbol{y})| \leq c \, \|\boldsymbol{x} - \boldsymbol{y}\|^{\rho}. \tag{5}$$

In addition, there exist positive constants $\underline{M}$ and $\overline{M}$ such that

$$\underline{M} \leq f_0(x, y) \leq \overline{M}, \quad \text{for all } (x, y) \in \mathcal{M}. \tag{6}$$

3.2 Assumption The censoring density g is bounded away from 0 and infinity on $(0, M_1)$. That is, there exist positive constants $\underline{K}$ and $\overline{K}$ such that $0 < \underline{K} \leq g(t) \leq \overline{K} < \infty$ for all $t \in (0, M_1)$.

3.3 Assumption For the Dirichlet prior, the parameter $\alpha = (\alpha_1, \ldots, \alpha_{p_n})$ satisfies $ap_n^{-1} \leq \alpha_l \leq 1$ for all $l = 1, \ldots, p_n$ and some constant $a \in \mathbb{R}^+$.

Let $\varepsilon_n = (n/\log n)^{-\frac{\rho}{2(\rho+2)}}$ and $\eta_n = \varepsilon_n^{2/3}$. Note that $\varepsilon_n \leq \eta_n$ and $n(\varepsilon_n^2 \wedge \eta_n^2) \to \infty$ as $n \to \infty$. Our main theoretical result is the following theorem.

3.4 Theorem *Consider either of the priors defined in Section 2.2 and impose assumptions 3.1 and 3.3. Fix $(x, y) \in \mathcal{M}$. Then for sufficiently large C*

$$\mathbb{E}_0 \Pi_n(f \in \mathcal{F} \colon |F(x,y) - F_0(x,y)| > C\eta_n \mid \mathcal{D}_n) \to 0, \quad \text{as} \quad n \to \infty$$

provided that all bin areas are equal to $\varepsilon_n^{4/\rho}$.

The proof of this theorem is based on the following two lemmas.

3.5 Lemma *Fix f_0 and g satisfying the conditions in assumption 3.1 and 3.2. If*

$$S_n = \left\{ f \in \mathcal{F} \colon KL(s_{f_0}, s_f) \leq \varepsilon_n^2, V(s_{f_0}, s_f) \leq \varepsilon_n^2 \right\}. \tag{7}$$

then $\Pi_n(S_n) \geq e^{-cn\varepsilon_n^2}$ for some constant $c > 0$.

3.6 Lemma *Fix $(t, z) \in \mathcal{M}$. Define $U_n(t,z) := \{f \in \mathcal{F} \colon |F(t,z) - F_0(t,z)| > C\eta_n\}$. There exists a sequence of test functions Φ_n such that*

$$\begin{aligned} \mathbb{E}_0(\Phi_n) &= o(1), \\ \sup_{f \in U_n(t,z)} \mathbb{E}_f(1 - \Phi_n) &\leq c_1 e^{-c_2 C^2 n \varepsilon_n^2}, \end{aligned} \tag{8}$$

for positive constants c_1, c_2 and constant C appearing in Theorem 3.4.

The remainder of the proof of Theorem 3.4 is standard and follows the general ideas in Ghosal, Ghosh & Van der Vaart (2000).

4 Proof of Lemmas

4.1 Proof of lemma 3.5

Proof The proof consists of 4 steps:

1. constructing a subset Ω_n of S_n for which the prior probability can be bounded from below;
2. deriving this bound for the D-prior;
3. deriving this bound for the LNGL-prior;
4. verifying that the prior mass condition is satisfied for the choice of ε_n in the lemma.

In the proof we will use double-indexation of coefficients, so rather than θ_j we write $\theta_{j,k}$, where (j,k) indexes a particular bin. A similar convention applies to H.

Step 1. We first give a sequence of approximations for f_0. Let δ_n be a sequence of positive numbers tending to 0 as $n \to \infty$. For each n, let $A_{n,j}$, $B_{n,k}$ be sets such that $\cup_{j=1}^{J_n} \cup_{k=1}^{K_n} (A_{n,j} \times B_{n,k})$ is a partition of $\mathcal{M}$, where the integers J_n and K_n are chosen such that $|A_{n,j}| = |B_{n,k}| = \delta_n$ for all j and k.

Let $f_{0,n}$ be the piecewise constant density function defined by

$$f_{0,n}(t,z) = \sum_{j=1}^{J_n} \sum_{k=1}^{K_n} \frac{w_{0,j,k}}{|A_{n,j} \times B_{n,k}|} \mathbf{1}_{A_{n,j} \times B_{n,k}}(t,z), \tag{9}$$

where $w_{0,j,k} = \int_{A_{n,j}} \int_{B_{n,k}} f_0(u,v)\,\mathrm{d}v\,\mathrm{d}u$. That is, we approximate f_0 by averaging it on each bin. Set $\varepsilon_n^2 = \delta_n^\rho$. By Lemma A.2 in the appendix there exists a constant $C > 0$ such that the set defined by

$$\bar{\Omega}_n := \{f \in \mathcal{F} : ||f - f_{0,n}||_\infty \leq C\delta_n^\rho,\ \mathrm{supp}(f) \supseteq \mathcal{M}\}\,. \tag{10}$$

satisfies $\bar{\Omega}_n \subseteq S_n$.

Recall that $p_n = J_n K_n$ denote the total number of bins. According to the prior specifications in Section 2.2, for any $f \in \mathcal{F}$, we parameterize

$$f_{\boldsymbol{\theta}}(x,y) = \sum_{j,k} \frac{\theta_{j,k}}{|A_{n,j} \times B_{n,k}|} \mathbf{1}_{A_{n,j}\times B_{n,k}}(x,y), \qquad (x,y) \in \mathbb{R}^2,$$

where $\boldsymbol{\theta}$ denotes the vector obtained by stacking all coefficients $\{\theta_{j,k}, j = 1,\dots,J_n, k = 1,\dots,K_n\}$. For any $(t,z) \in A_{n,j} \times B_{n,k}$, $j,k \geq 1$, we have

$$|f_{\boldsymbol{\theta}}(t,z) - f_{0,n}(t,z)| = |A_{n,j} \times B_{n,k}|^{-1} |\theta_{j,k} - w_{0,j,k}| \leq \delta_n^{-2} \max_{j,k} |\theta_{j,k} - w_{0,j,k}|.$$

Hence

$$\Omega_n := \left\{ f_{\boldsymbol{\theta}} \in \mathcal{F} \colon \max_{j,k} |\theta_{j,k} - w_{0,j,k}| \leq C\delta_n^{\rho+2} \right\} \subseteq \bar{\Omega}_n \tag{11}$$

and consequently $\Pi_n(S_n) \geq \Pi_n(\Omega_n)$.

Step 2. For the D-prior we have $\boldsymbol{\theta} \sim \mathrm{Dirichlet}(\alpha)$ for fixed $\alpha = (\alpha_1,\dots,\alpha_{p_n})$, where we assume that $ap_n^{-1} \leq \alpha_l \leq 1$ for all $l = 1,\dots,p_n$. By Lemma 6.1 in Ghosal, Ghosh & Van der Vaart (2000), we have

$$\begin{aligned}\Pi_n(\Omega_n) &\geq \Gamma\left(\sum_{l=1}^{p_n} \alpha_l\right) \left(C\delta_n^{\rho+2}\right)^{p_n} \prod_{l=1}^{p_n} \alpha_l \\ &\geq \exp\left(\log\Gamma(a) + p_n \log\left(C\delta_n^{\rho+2}\right) + p_n \log(ap_n^{-1})\right)\end{aligned}$$

As $p_n \asymp \delta_n^{-2}$, $p_n \log \delta_n^{\rho+2} \asymp p_n \log p_n^{-(\rho+2)/2} = -(\rho/2+1)p_n \log p_n$. We conclude that the exponent behaves asymptotically as a multiple of $-p_n \log p_n$.

Step 3. Let $\theta_{j,k} = \frac{\psi(H_{j,k})}{\sum_{j,k}\psi(H_{j,k})}$ as defined in (3).For the LNGL-prior, as will become clear shortly, we need a lower bound on $|\Upsilon|$ and upper bound on the largest eigenvalue of Υ. For the latter, note that all eigenvalues of any stochastic matrix are bounded by 1. For the graph considered here, there exist diagonal matrices U and D such that $D(U+L)$ is a stochastic matrix (simply adding elements to the diagonal to have a matrix with nonnegative elements, followed by renormalising each row to have its elements sum to one. This procedure implies that in the present setting the largest eigenvalue of L is bounded by 8. Therefore, the largest eigenvalue of Υ is bounded by 9. As L has smallest eigenvalue 0, the smallest eigenvalue of Υ equals p_n^{-2}. Hence

$$|\Upsilon| \geq (p_n^{-2})^{p_n} = \exp\left(-2p_n \log p_n\right).$$

We have the following bounds:

$$\begin{aligned}\delta_n^2 \|f - f_{0,n}\|_\infty &\leq \sup_{j,k} |\theta_{j,k} - \theta_{0,j,k}| = \|S(\boldsymbol{H}) - S(\boldsymbol{H}_0)\|_\infty \\ &\leq \|S(\boldsymbol{H}) - S(\boldsymbol{H}_0)\|_2 \leq \|\boldsymbol{H} - \boldsymbol{H}_0\|_2,\end{aligned}$$

assuming S is the softmax function. At the last inequality we use that this function is Lipschitz-continuous with respect to $\|\cdot\|_2$ (Cf. Proposition 4 in Gao and Pavel (2018)). Therefore

$$\underline{\Omega}_n := \{\boldsymbol{H} \,:\, \|\boldsymbol{H} - \boldsymbol{H}_0\|_2 \leq C\delta_n^2 \Delta_n\} \subseteq \{f \,:\, \|f - f_{0,n}\|_\infty \leq C\Delta_n\},.$$

where $\Delta_n := \delta_n^\rho$. Whereas notationally implicit, both sets in the preceding display are sets of ωs induced by the constraints on either $\boldsymbol{H}$ or f. Hence it suffices to lower bound

$$\int \int_{\underline{\Omega}_n} \varphi(\boldsymbol{H}; 0, \tau\Upsilon^{-1}) \,\mathrm{d}\boldsymbol{H} f(\tau)\,\mathrm{d}\tau.$$

We first focus on the inner integral. As the largest eigenvalue of Υ s bounded by 9, we have $\boldsymbol{H}^T\Upsilon\boldsymbol{H} \leq 9\|\boldsymbol{H}\|_2^2$. Hence,

$$\begin{aligned}
\mathcal{I}_n(\tau) &:= \int_{\underline{\Omega}_n} \varphi(\boldsymbol{H}; 0, \tau\Upsilon^{-1})\,\mathrm{d}\boldsymbol{H} \\
&= (2\pi\tau)^{-p_n/2}|\Upsilon|^{1/2}\int_{\underline{\Omega}_n} \exp\left(-\frac{1}{2\tau}\boldsymbol{H}^T\Upsilon\boldsymbol{H}\right)\,\mathrm{d}\boldsymbol{H} \\
&\geq (2\pi\tau)^{-p_n/2}\exp\left(-p_n\log p_n\right)\int_{\underline{\Omega}_n} \exp\left(-\frac{1}{2\tau}9\|\boldsymbol{H}\|^2\right)\,\mathrm{d}\boldsymbol{H}
\end{aligned}$$

If $\boldsymbol{H} \in \underline{\Omega}_n$, then

$$\|\boldsymbol{H}\|^2 \leq \|\boldsymbol{H} - \boldsymbol{H}_0\|^2 + \|\boldsymbol{H}_0\|^2 \leq C^2\delta_n^4\Delta_n^2 + \|\boldsymbol{H}_0\|^2.$$

which means that

$$\begin{aligned}
\mathcal{I}_n(\tau) &\geq (2\pi\tau)^{-p_n/2}\exp\left(-p_n\log p_n\right)\int_{\underline{\Omega}_n} \exp\left(-\frac{1}{2\tau}9\left(C^2\delta_n^4\Delta_n^2 + \|\boldsymbol{H}_0\|^2\right)\right)\,\mathrm{d}\boldsymbol{H} \\
&= (2\pi\tau)^{-p_n/2}\exp\left(-p_n\log p_n\right)\exp\left(-\frac{1}{2\tau}9\left(C^2\delta_n^4\Delta_n^2 + \|\boldsymbol{H}_0\|^2\right)\right)\mathrm{Vol}(\underline{\Omega}_n).
\end{aligned}$$

Hence,

$$\log\mathcal{I}_n(\tau) \geq -\frac{p_n}{2}\log(2\pi\tau) - p_n\log p_n - \frac{1}{2\tau}9\left(C^2\delta_n^4\Delta_n^2 + \|\boldsymbol{H}_0\|^2\right) + \log\mathrm{Vol}(\bar{\Omega}_n). \tag{12}$$

We express the asymptotic behaviour of all terms in the exponent in terms of p_n. To this end, note that $p_n \asymp \delta_n^{-2}$, hence $\Delta_n \asymp p_n^{-\rho/2}$, leading to $\delta_n^4\Delta_n^4 \asymp p_n^{-2}p_n^{-\rho} = p_n^{-\rho-2}$. As $\mathrm{Vol}(\bar{\Omega}_n) \asymp \Delta_n^{p_n}$ we have $\log\mathrm{Vol}(\bar{\Omega}_n) \asymp p_n\log\Delta_n \asymp -\frac{\rho}{2}p_n\log p_n$. So we can conclude that the right-hand-side of (12) behaves as $-p_n\log p_n$ for n large.

Step 4. For both priors, the prior mass condition gives the following condition on ε_n:

$$p_n\log p_n \lesssim n\varepsilon_n^2.$$

As $\delta_n^\rho = \varepsilon_n^2$ we get $p_n \asymp \delta_n^{-2} = \left(\varepsilon_n^{2/\rho}\right)^{-2} = \varepsilon_n^{-4/\rho}$. Hence we need to choose ε_n such that

$$\varepsilon_n^{-4/\rho}\log(1/\varepsilon_n) \lesssim n\varepsilon_n^2.$$

This relationship is satisfied if

$$\varepsilon_n \asymp (n/\log n)^{-\frac{\rho}{4+2\rho}}.$$

□

4.2 Proof of lemma 3.6

Proof We consider different test functions in different regimes of t: $t \in (0, M_1)$ and $t \in \{0, M_1\}$.

First consider $(t, z) \in (0, M_1) \times (0, M_2]$. Define test sequences

$$\Phi_n^+(t,z) = \mathbf{1}\left\{\frac{1}{n}\sum_{i=1}^n \kappa_n^+(t,z;T_i,Z_i) - \int_t^{t+h_n} g(x)F_0(x,z)\,\mathrm{d}x > e_n/2\right\},$$
$$\Phi_n^-(t,z) = \mathbf{1}\left\{\frac{1}{n}\sum_{i=1}^n \kappa_n^-(t,z;T_i,Z_i) - \int_{t-h_n}^{t} g(x)F_0(x,z)\,\mathrm{d}x < -e_n/2\right\},$$

where

$$\kappa_n^+(t,z;T,Z) = \mathbf{1}_{[t,t+h_n]}(T)\mathbf{1}_{(0,z]}(Z),$$
$$\kappa_n^-(t,z;T,Z) = \mathbf{1}_{[t-h_n,t]}(T)\mathbf{1}_{(0,z]}(Z).$$

Note that if $(T, Z) \sim f$, then

$$\begin{aligned}\mathbb{E}_f(\kappa_n^+(t,z;T,Z)) &= \int \mathbf{1}_{[t,t+h_n]}(x)\mathbf{1}_{(0,z]}(u)s_f(x,u)\,\mathrm{d}\mu(x,u)\\ &= \int_t^{t+h_n}\int_0^z g(x)\partial_2 F(x,u)\,\mathrm{d}\mu_2(x,u) = \int_t^{t+h_n} g(x)F(x,z)\,\mathrm{d}x,\end{aligned}$$

where s_f is the density function of (T, Z) defined in (2). By Assumption 3.2,

$$\mathbb{E}_f\left[(\kappa_n^+(t,z;T,Z))^2\right] = \mathbb{E}_f(\kappa_n^+(t,z;T,Z)) \le \int_t^{t+h_n} g(x)\,\mathrm{d}x \le \overline{K}h_n.$$

The same upper bound holds for $\mathbb{E}_f\left[(\kappa_n^-(t,z;T,Z))^2\right]$. By Bernstein's inequality (Van der Vaart (1998), Lemma 19.32),

$$\mathbb{E}_0(\max(\Phi_n^+(t,z),\Phi_n^-(t,z))) \le 2\exp\left(-\frac{1}{16}\frac{ne_n^2}{\overline{K}h_n + e_n/2}\right) = o(1).$$

Let $\{\eta_n\}$ be a sequence of positive numbers tending to zero as $n \to \infty$. For $(t, z) \in [0, M_1] \times (0, M_2]$, define sets

$$U_{n,1}(t,z) = \{f : F(t,z) > F_0(t,z) + C\eta_n\},$$
$$U_{n,2}(t,z) = \{f : F(t,z) < F_0(t,z) - C\eta_n\}.$$

Then $U_n(t, z) = U_{n,1}(t, z) \cup U_{n,2}(t, z)$.

When $f \in U_{n,1}(t, z)$, for any $x \in [t, t + h_n]$, by the monotonicity of F and $f_0 \le \overline{M}$, we have

$$\begin{aligned}F(x,z) - F_0(x,z) &\ge F(t,z) - F_0(t,z) - (F_0(x,z) - F_0(t,z))\\ &\ge C\eta_n - \overline{M}M_2h_n \ge C\eta_n/2.\end{aligned}$$

Then it follows

$$\int_t^{t+h_n} g(x)(F(x,z) - F_0(x,z))\,\mathrm{d}x \ge \frac{C\eta_n}{2}\int_t^{t+h_n} g(x)\,\mathrm{d}x \ge \frac{C\underline{K}}{2}\eta_n h_n = e_n.$$

Hence, for $f \in U_{n,1}$ we have

$$\begin{aligned}\mathbb{E}_f(1 - \Phi_n^+(t,z)) &= \mathbb{P}_f\left(\frac{1}{n}\sum_{i=1}^n \kappa_n^+(t,z|T_i,Z_i) - \int_t^{t+h_n} g(x)F_0(x,z)\,\mathrm{d}x < e_n/2\right)\\ &\le \mathbb{P}_f\left(\frac{1}{n}\sum_{i=1}^n \kappa_n^+(t,z|T_i,Z_i) - \int_t^{t+h_n} g(x)F(x,z)\,\mathrm{d}x \le -e_n/2\right).\end{aligned}$$

Further, Bernstein's inequality gives

$$\mathbb{E}_f(1-\Phi_n^+(t,z)) \le 2\exp\left(-\frac{1}{16}\frac{ne_n^2}{\overline{K}h_n + e_n/2}\right) \le c_1 e^{-c_2 C^2 n\varepsilon_n^2}$$

for some constants $c_1, c_2 > 0$.

When $f \in U_{n,2}(t,z)$, $x \in [t-h_n, t]$, we have

$$\begin{aligned} F(x,z) - F_0(x,z) &\le F(t,z) - F_0(t,z) + F_0(t,z) - F_0(x,z) \\ &\le -C\eta_n + \overline{M}M_2 h_n \le -C\eta_n/2 \end{aligned}$$

and

$$\int_{t-h_n}^{t} g(x)(F(x,z) - F_0(x,z))\,\mathrm{d}x \le -\frac{C\underline{K}}{2}\eta_n h_n = -e_n.$$

Hence for $f \in U_{n,2}$, the type II error satisfies

$$\mathbb{E}_f(1-\Phi_n^-(t,z)) \le \mathbb{P}_f\left(\frac{1}{n}\sum_{i=1}^{n}\kappa_n^-(t,z|T_i,Z_i) - \int_{t-h_n}^{t} g(x)F(x,z)\,\mathrm{d}x \ge e_n/2\right).$$

Using Bernstein's inequality again, we have

$$\mathbb{E}_f(1-\Phi_n^-(t,z)) \le c_1 e^{-c_2 C^2 n\varepsilon_n^2}, \quad \text{for some } c_1, c_2 > 0.$$

For the boundary case $(t,z) \in \{0, M_1\} \times (0, M_2]$. With a similar idea, in order to give non-zero test sequences, we use κ_n^+ define $\Phi_n^+(0,z), \Phi_n^-(0,z)$ and κ_n^- define $\Phi_n^+(M_1,z), \Phi_n^-(M_1,z)$. When $f \in U_{n,1}(0,z)$, using the tests sequence $\Phi_n^+(0,z)$ defined in case $t \in (0, M_1)$, we have

$$\sup_{f \in U_{n,1}(0,z)} \mathbb{E}_f(1-\Phi_n^+(0,z)) \le c_1 e^{-c_2 C^2 n\varepsilon_n^2}.$$

When $f \in U_{n,2}(M_1,z)$, using the tests sequence $\Phi_n^-(M_1,z)$ defined in case $t \in (0, M_1)$, we have

$$\sup_{f \in U_{n,2}(M_1,z)} \mathbb{E}_f(1-\Phi_n^-(M_1,z)) \le c_1 e^{-c_2 C^2 n\varepsilon_n^2}.$$

Note that for any $f \sim \Pi_n$ and $t \in A_{n,j}$, $j = 1, \ldots, J_n$,

$$\int_0^{M_2} f(t,v)\,\mathrm{d}v = |A_{n,j}|^{-1}\sum_{k=1}^{K_n}\theta_{j,k} \le \delta_n^{-1}K_n = M_2. \tag{13}$$

Here we use $\theta_{j,k} \le 1$ and $|A_{n,j}| \ge \delta_n$. When $f \in U_{n,2}(0,z)$, for any $x \in [0, h_n]$, using (13) we have

$$\begin{aligned} F(x,z) - F_0(x,z) &\le F(x,z) - F(0,z) + F(0,z) - F_0(0,z) \\ &\le \int_0^x \int_0^z f(u,v)\,\mathrm{d}v\,\mathrm{d}u - C\eta_n \\ &\le M_2 h_n - C\eta_n \le -C\eta_n/2 \end{aligned}$$

and

$$\int_0^{h_n} g(x)(F(x,z) - F_0(x,z))\,\mathrm{d}x \le -e_n.$$

Define tests sequence

$$\Phi_n^-(0,z) = \mathbf{1}\left\{\frac{1}{n}\sum_{i=1}^n \kappa_n^+(0,z|T_i,Z_i) - \int_0^{h_n} g(x)F_0(x,z)\,\mathrm{d}x < -e_n/2\right\}.$$

Hence by the Bernstein's inequality,

$$\mathbb{E}_f(1-\Phi_n^-(0,z)) \le c_1 e^{-c_2 C^2 n\varepsilon_n^2}.$$

By the similar arguments as above, when $f \in U_{n,1}(M_1,z)$, for any $x \in [M_1 - h_n, M_1]$, using (13) we have

$$\begin{aligned} F(x,z) - F_0(x,z) &\ge F(x,z) - F(M_1,z) + F(M_1,z) - F_0(M_1,z) \\ &\ge C\eta_n - \int_{M_1-h_n}^{M_1}\int_0^z f(u,v)\,\mathrm{d}v\,\mathrm{d}u \\ &\ge C\eta_n - M_2 h_n \ge C\eta_n/2 \end{aligned}$$

and

$$\int_0^{h_n} g(x)(F(x,z) - F_0(x,z))\,\mathrm{d}x \ge -e_n.$$

Define tests sequence

$$\Phi_n^+(M_1,z) = \mathbf{1}\left\{\frac{1}{n}\sum_{i=1}^n \kappa_n^-(M_1,z|T_i,Z_i) - \int_{M_1-h_n}^{M_1} g(x)F_0(x,z)\,\mathrm{d}x > e_n/2\right\},$$

hence,

$$\mathbb{E}_f(1-\Phi_n^+(M_1,z)) \le c_1 e^{-c_2 C^2 n\varepsilon_n^2}.$$

To conclude, take $\Phi_n(t,z) = \max(\Phi_n^+(t,z), \Phi_n^-(t,z))$, we derived

$$\begin{aligned} \mathbb{E}_0\Phi_n(t,z) &= o(1), \\ \sup_{f\in U_n(t,z)} \mathbb{E}_f(1-\Phi_n(t,z)) &\le c_1 e^{-c_2 C^2 n\varepsilon_n^2}. \end{aligned}$$

□

5 Computational methods

In this section we present algorithms for drawing from the posterior distribution for both priors described in section 2.2. Contrary to our theoretical contribution, we include a prior on the scaling parameter τ appearing in the LNGL prior. Likewise, for the D-prior we will assume

$$\begin{aligned} \tau &\sim \Pi \\ \theta \mid \tau &\sim \text{Dirichlet}(\tau,\ldots,\tau). \end{aligned}$$

to ensure a fair comparison of simulation results obtained by either of these priors.

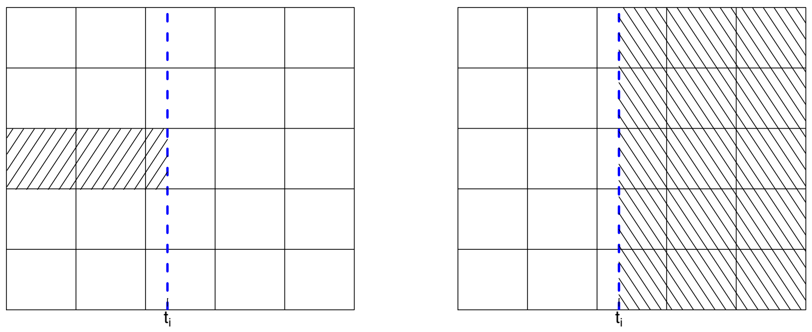

Figure 1: Left: if $x_i \leq t_i$ the mark is observed. Right: if $x_i > t_i$ the mark is not observed.

5.1 Dirichlet prior (D-prior)

First, we consider the case where $\{(X_i, Y_i),\ i = 1, \ldots, n\}$ is a sequence of independent random vectors, with common density f_0 that is piecewise constant on $A_{n,j} \times B_{n,k}$ and compactly supported. This "no-censoring" model has likelihood

$$l(\boldsymbol{\theta}) = \prod_{j,k} \theta_{j,k}^{C_{j,k}},$$

where $C_{j,k} = \sum_i \mathbf{1}\{(X_i, Y_i) \in A_{n,j} \times B_{n,k}\}$ denotes the number of observations that fall in bin $A_{n,j} \times B_{n,k}$. Clearly, the Dirichlet prior is conjugate for the likelihood, resulting in the posterior being of Dirichlet type as well and known in closed form. In case of censoring, draws from the posterior for the Dirichlet prior can be obtained by data-augmentation, where the following two steps are alternated:

1. Given $\boldsymbol{\theta}$ and censored data, simulate the "full data". This is tractable since the censoring scheme tells us in which collection of bins the actual observation can be located. Then one can renormalise the density f restricted to these bins and select a specific bin accordingly and generate the "full data". Cf. Figure 1 for the two types of observations.

2. Given the "full data", draw samples for $\boldsymbol{\theta}$ from the posterior which is of Dirichlet type.

5.2 Logistic Normal Graph Laplacian prior (LNGL-prior)

For the LNGL prior, one could opt for a data-augmentation scheme as well, but its attractiveness is lost since step (2) above is no longer of simple form. Therefore, we propose to bypass data-augmentation in this case. In an initial version of this paper we have proposed to use the probabilistic programming language Turing (see Ge, Xu & Ghahramani (2018)) that is based on the Julia language (see Bezanson et al. (2017)). With modest programming efforts, samples from the posterior can then be obtained by Hamiltonian Monte Carlo methods. For completeness, we present the core of such code in Appendix B. As can be seen, this contains about 10 lines of code and reads intuitively. However, a much faster algorithm can be derived by exploiting structure in the statistical model. The main idea is to use a non-centred parameterisation (Papaspiliopoulos et al. (2003)) combined with a preconditioned Crank-Nicolson (pCN) scheme. The latter scheme dates back to Neal (1998); a more recent exposition can be found in Cotter et al. (2013) (see also Van der Meulen & Schauer (2017) in a different application setting).

Let θ denote the parameter vector (with all $\theta_{j,k}$ stacked). Let $S : \mathbb{R}^p \to \mathbb{R}^p$ denote the "softmax"-function, defined by $S(x_1, \ldots, x_p) = (e^{x_1}, \ldots, x^{x_p})/\sum_{i=1}^p e^{x_i}$ (this is a convenient choice, for our algorithm other mappings of a vector to a probability vector are also allowed). To sample from the prior of (τ, θ), with $\theta \in \mathbb{R}^p$, we sample according to the following scheme

$$\begin{aligned} \tau &\sim \Pi \\ z &\sim \mathrm{N}_p(0, I_p) \\ \theta \mid z, \tau &= S(\bar{U} z \sqrt{\tau}), \end{aligned}$$

where $\bar{U} = U^{-1}$ with U obtained from the Cholesky-decompostion of the graph-Laplacian matrix Υ (with a small multiple of the identity matrix added): $\Upsilon = U^T U$. Hence, we have introduced the random vector z, which is centred in between τ and θ. The algorithm we propose is a Gibbs sampler which iteratively updates τ and z. While our theoretical contribution assumes a prior on τ of Gamma type (see Assumption 3.3), the MCMC-algorithms in this section apply more generally to a prior distribution Π on τ that is supported on the positive halfline. Small values of τ induce more smoothing.

For updating z conditional on τ we use the pCN scheme: pick a tuning parameters $\rho \in [0, 1)$ (typically chosen close to 1) and propose a new value z° for z by setting

$$z^\circ = \rho z + \sqrt{1 - \rho^2} w, \tag{14}$$

where $w \sim \mathrm{N}_p(0, I_p)$, independently of (τ, z). Next, the Metropolis-Hastings (MH) acceptance rule is used to accept the proposal z° with probability $1 \wedge \mathcal{L}(z^\circ, \tau)/\mathcal{L}(z, \tau)$, where $\mathcal{L}(z, \tau)$ denotes the likelihood, evaluated in (z, τ) (note that the prior ratio and proposal ratio cancel, as $\pi(z) q(z^\circ \mid z)$ is symmetric in (z, z°)). The likelihood can be simply and efficiently computed. To see this, consider Figure 1. Observations can be represented either by the left- or right figure. For the i-th observation, we will denote by a_i the vector that contains for each cell of the partition the fraction of the area that is shaded (so most values will be either 0 or 1, only cells intersected by the blue dashed line will have entries in $(0, 1)$). With this notation, it is easy to see that

$$\log \mathcal{L}(z, \tau) = \sum_{i=1}^n \log(\theta^T a_i), \quad \text{with} \quad \theta = S(\bar{U} z \sqrt{\tau}).$$

To efficiently compute the loglikelihood, all that needs to be computed (once), is for each observation the vector of indices corresponding to the shaded boxes and the corresponding area fractions (this enables to exploit sparsity in a_i when computing $\theta^T a_i$). Also note that the Cholesky decomposition of L only needs to be computed once.

For updating τ, conditional on the data and z, we use a MH-step. We draw a proposal τ° according to $\log \tau^\circ \mid \tau \sim \mathrm{N}(\log \tau, \delta^2)$. It is accepted with probability

$$1 \wedge \frac{\mathcal{L}(z, \tau^\circ)}{\mathcal{L}(z, \tau)} \frac{\pi(\tau^\circ)}{\pi(\tau)} \frac{\tau^\circ}{\tau},$$

where π denotes the prior density on τ and the term τ°/τ comes from the Jacobian. Note that a partial conjugacy for updating τ gets lost, even when employing a prior of inverseGamma type.

We conclude that in one iteration of the Gibbs-sampler, it is only required to do a simple MH-step for updating τ, sample z° as in (14) and use the MH-acceptance rule for this step as well.

5.1 Remark In the prior specification we have postulated $\theta \mid z, \tau = S(\bar{U} z \sqrt{\tau})$. Note that the prior on z centers at zero. However, due to translation invariance of the softmax function, the same prior on θ is specified is we take $\theta \mid z, \tau = S(\bar{U} z \sqrt{\tau} + v)$, where v is a vector that is a multiple of the vector with all elements equal to 1. Despite this identifiability issue, we have not encountered severe autocorrelation in traceplots for the LNGL-prior.

5.2 Remark The same algorithm applies to more general censoring schemes. For example, suppose there are two checkup times, and the mark is only observed if the event took place in between those checkup times. This setting just corresponds to a different type of shading of boxes in Figure 1 but otherwise does not implicate any change to the numerical setting. Hence, the interval-censored continuous-mark model (see for instance Maathuis & Wellner (2008)) is covered by our computational methods.

5.3 Numerical examples

In the following simulations, we compare the priors based on the Dirichlet distribution and the LNGL-prior. Updating τ can be done by incorporating a MH-step similar as for the LNGL-prior.

In each of the reported results, $20,000$ MCMC iterations were used, posterior means were computing after discarding the initial $1/3$ of the iterations as burnin. The algorithms were tuned such that the MH-steps have acceptance probability of about $0.25-0.5$. For analysing a data set of size $n=200$ with 100 bins, computing times are about 3 seconds for $20,000$ iterations, using a Macbook-pro 2 GHz Quad-Core Intel Core i5 with 16 GB RAM. Note that the complexity of the algorithm scales with the number of bins and not with the sample size. From experiments with a large number of bins it appears that the acceptance rate for the pCN-step does not deteriorate, as also seen in other settings where pCN is utilised (Cf. (Cotter et al. (2013))). Julia code (Bezanson et al. (2017)) is available from https://github.com/fmeulen/CurrentStatusContinuousMarks. Computed Wasserstein distances are computed using the transport library in R.

Traceplots (not included here) confirm that in our experiments the chain on (θ,τ) mixes well. Only in case of the D-prior, somewhat larger autocorrelation is observed in the chain for τ. The prior measure Π is taken to be the standard Exponential distribution for both priors.

We will consider the following data generated as independent replications of (X,Y), where (X,Y) is drawn from the mixture experiment

$$(X,Y)\sim\begin{cases}(U,V) & \text{with probability } 0.3\\ (1-U,V) & \text{with probability } 0.7\end{cases},$$

with (U,V) having density

$$f(u,v)=(3/8)(u^2+v)\mathbf{1}_{[0,1]\times[0,2]}(u,v).$$

In this case, it is easy to sample (X,Y) by first sampling from the marginal distribution of X and then from the density of $Y\mid X$. In both cases the inverse of the cumulative distribution function can be computed in closed form. We assume that the censoring random vairable $T\sim\sqrt{U}$ where U is uniformly distributed on $[0,1]$. This implies that the density of T is given by $t\mapsto 2t\mathbf{1}_{[0,1]}(t)$. For the LNGL-prior we take $\Upsilon=L+N^{-2}I$ where L is defined in (4) and N is the number of cells in the partition used to cover the support of the density.

5.3.1 Results for one dataset

We sample 200 observations from the model. The true density with observations superimposed is depicted in Figure 2. Horizontally/vertically we took 25/50 bins, yielding equally sized bins. Note that the number of bins is way larger than the sample size; smoothing/penalisation being enforced by the prior. In Figure 3 we compare the performance of the posterior mean estimator under both prior specifications. Smoothing induced by the LNGL-prior is quite apparent and the posterior mean estimate is visually more appealing. The latter is not surprising as the data-generating density is smooth.

5.3.2 Varying the number of bins

In Figures 4 and 5 we ran the algorithm on the same dataset using coarser/refined binning.

For each combination of prior/binning, we compute the Wasserstein distance (based on the ℓ_1/cityblock distance) between the probability vector of bin-masses of the estimate and the probability vector of bin

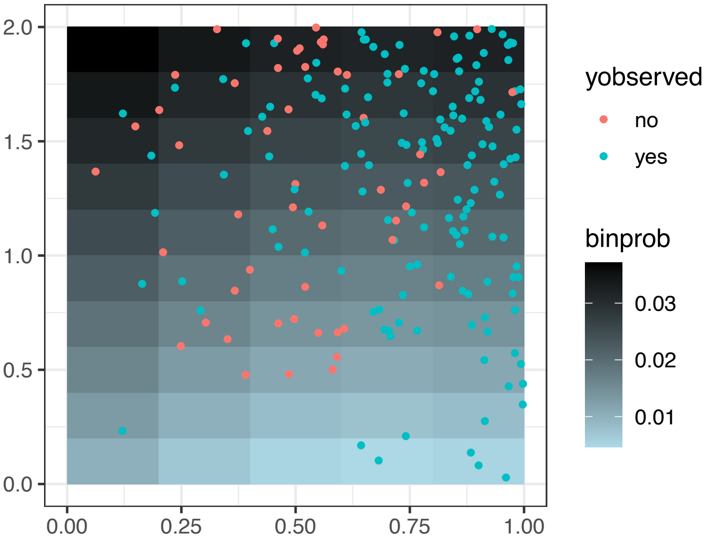

Figure 2: For each cell in the partition of $[0, 1] \times [0, 2]$, the probability of the cell is coloured. Superimposed are points representing the data, where the coordinate along the horizontal axis is the censoring time, and the coordinate along the vertical axis is the mark-variable y. If for a particular point y is observed/unobserved (cyan/red colour), then this means the event time is to the left/right of the censoring time t. The height of the red points is latent in the observations.

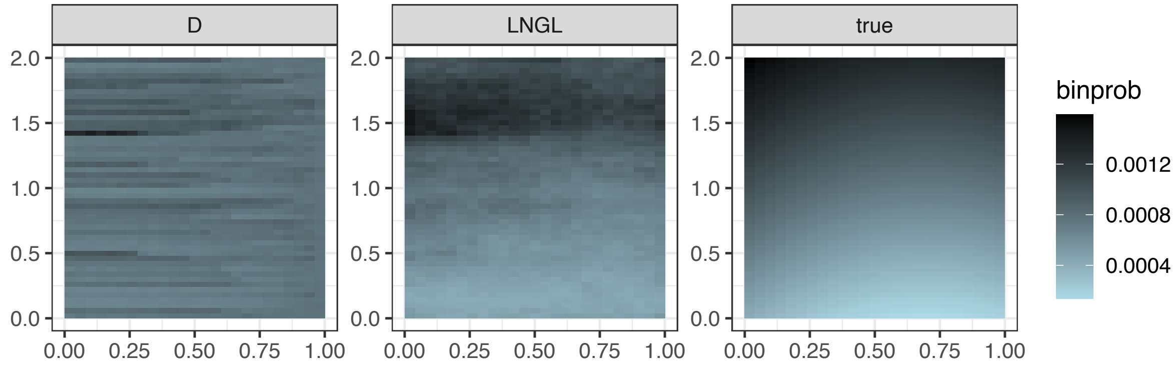

Figure 3: Left and middle: posterior mean probabilities for D- and LNGL-priors respectively. Right: true posterior probabilities. Horizontally 25 bins, vertically 50 bins.

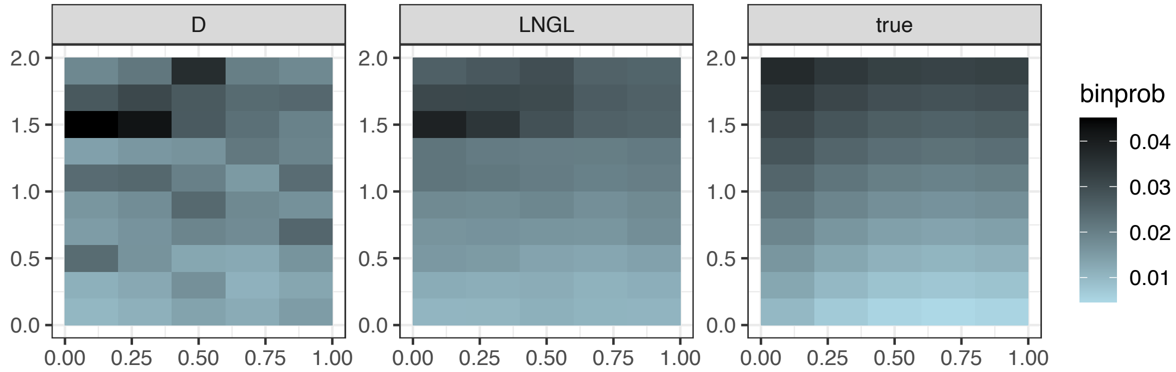

Figure 4: Left and middle: posterior mean probabilities for D- and LNGL-priors respectively. Right: true posterior probabilities. Horizontally 5 bins, vertically 10 bins.

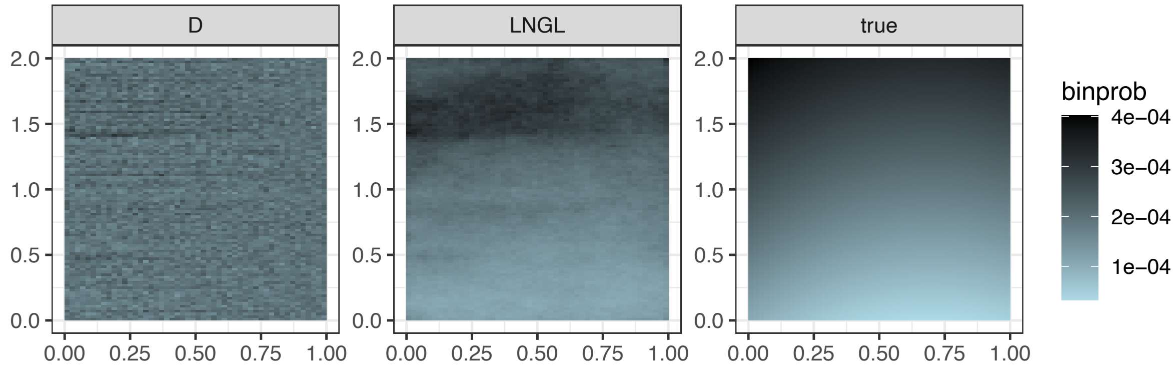

Figure 5: Left and middle: posterior mean probabilities for D- and LNGL-priors respectively. Right: true posterior probabilities. Horizontally 100 bins, vertically 200 bins.

masses of the true data-generating distribution. For a definition and motivation for using this distance we refer to Chapter 2.1 in Panaretos and Zemel (2020).

prior/bins	5/10	25/50	50/100
D	0.118	0.224	0.239
LNGL	0.069	0.076	0.082

Here “5/10” for example means: 5 bins in the horizontal direction 10 bins in the vertical direction. Especially with a large number of bins, the performance of the posterior mean using the D-prior is visually unappealing. Inspection of the traceplot for the parameter τ in this setting reveals large uncertainty. The LNGL-prior on the other hand seems quite robust in performance, once a sufficiently large number of bins is chosen.

5.3.3 Monte-Carlo study

To compare the performance of the posterior mean estimator under both prior specifications we conduct a Monte-Carlo study. In each simulation run we

- simulate a dataset;
- compute the posterior mean under both the D- and LNGL-prior;
- compute the Wasserstein distance (based on the ℓ_1/cityblock distance) between the probability vector of bin-masses of the estimate and the probability vector of bin masses of the true data-generating distribution.

We considered samples sizes $100, 250, 500$ and took the Monte-Carlo sample size equal to 100. In figures 6 and 7 we give two visualisations of the results. As expected, with large sample size the distance tends to be smaller. Additionally, the LNGL-prior appears to outperform the D-prior for this choice of data-generating distribution.

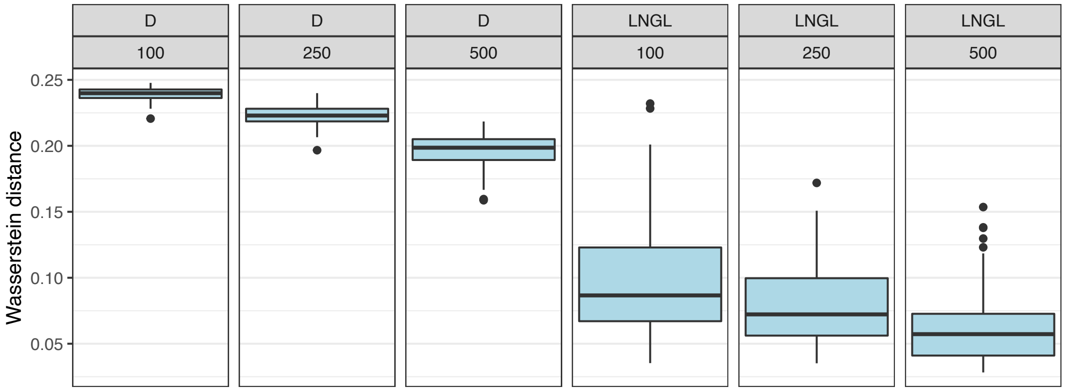

Figure 6: Simulation study. Wasserstein distance, averaged over 100 Monte-Carlo samples for samples sizes $100, 250, 500$ and both the D- and LNGL-prior.

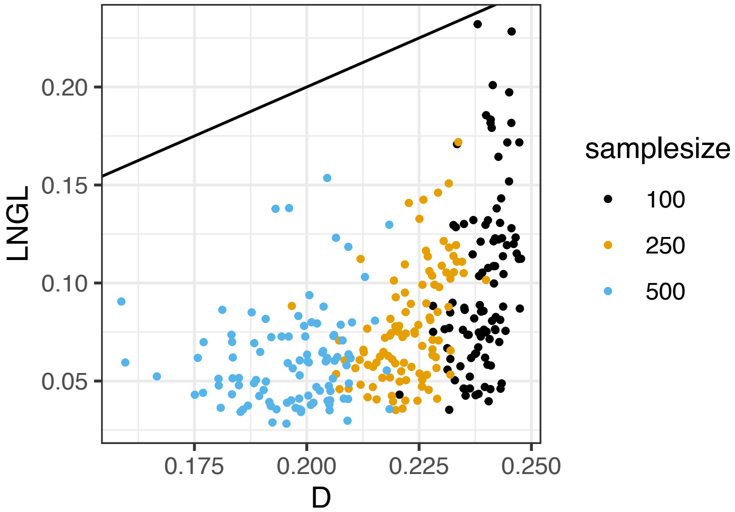

Figure 7: Simulation study. Wasserstein distance for simulated datasets for samples sizes $100, 250, 500$ and both the D- and LNGL-prior. Added is the line with intercept 0 and slope 1.

6 Discussion

The main theoretical contribution of this paper is Theorem 3.4. We expect the the rate in this theorem to be suboptimal. We conjecture ε_n to be the optimal rate, suggesting that Lemma 3.5 is sufficiently sharp. However, due to the tests constructed in Lemma 3.5 we obtain rate $\varepsilon^{2/3}$ in our main result. We postpone further investigations to future research, where we also hope to obtain rates in global metrics.

The derived rate is non-adaptive. Adaptation can be achieved by employing a prior on the number of bins, see Chapter 10 in Ghosal & Van der Vaart (2017). Computationally, such an approach is less attractive, as it requires a sampler on the larger space of the union of partitions. Instead, in our numerical work we have chosen to start off from a larger number of small bins, and have a prior on the scaling parameter τ take care of the regularisation. This yields an easily implementable and efficient computational scheme.

A Technical proofs

A.1 Lemma *Let f_1 and f_2 be bivariate density functions on $\mathcal{M} = [0, M_1] \times [0, M_2]$. Assume the density of the censoring time, g, is bounded. Then there exists a constant $C > 0$, independent of f_1 and f_2 such that*

$$\|s_{f_1} - s_{f_2}||_1 \leq C\|f_1 - f_2\|_\infty.$$

Proof Say g is bounded by $\bar{K}$. If $(t, z) \in \mathcal{M}$, then

$$\begin{aligned}
|s_{f_1}(t,z) - s_{f_2}(t,z)| &= \left| g(t) \left(\mathbf{1}_{\{z>0\}} \int_0^t (f_1(u,z) - f_2(u,z))\, \mathrm{d}u \right.\right. \\
&\qquad \left.\left. + \mathbf{1}_{\{z=0\}} \int_t^{M_1} \int_0^{M_2} (f_1(u,v) - f_2(u,v))\, \mathrm{d}v\, \mathrm{d}u \right) \right| \\
&\leq \bar{K} \max \left\{ \int_0^t |f_1(u,z) - f_2(u,z)|\, \mathrm{d}u, \int_t^{M_1} \int_0^{M_2} |f_1(u,v) - f_2(u,v)|\, \mathrm{d}v\, \mathrm{d}u \right\} \\
&\leq \bar{K}\, |\mathcal{M}|\, \|f_1 - f_2\|_\infty.
\end{aligned}$$

This implies

$$\|s_{f_1} - s_{f_2}||_1 = \int_{\mathcal{M}} |s_{f_1} - s_{f_2}|\, \mathrm{d}\mu \leq \bar{K}\, |\mathcal{M}|\, \mu(\mathcal{M})\, \|f_1 - f_2\|_\infty. \tag{15}$$

□

A.2 Lemma *Impose Assumption 3.1. Let $f_{0,n}$ be as defined in (9). Define set*

$$\Omega_n := \{ f \in \mathcal{F} : ||f - f_{0,n}||_\infty \leq C\delta_n^\rho,\ supp(f) \supseteq \mathcal{M} \}.$$

If $f \in \Omega_n$, then for sufficiently large n there exists a constant $C_1 > 0$ such that

$$KL(s_{f_0}, s_f) \leq C_1 \delta_n^\rho, \quad V(s_{f_0}, s_f) \leq C_1 \delta_n^\rho.$$

Proof The proof is based on Lemma B.2 in Ghosal & Van der Vaart (2017), which gives the following inequalities

$$\begin{aligned}
KL(s_{f_0}, s_f) &\leq 2h^2(s_{f_0}, s_f)\|s_{f_0}/s_f\|_\infty, \\
V(s_{f_0}, s_f) &\leq 2h^2(s_{f_0}, s_f)\|s_{f_0}/s_f\|_\infty.
\end{aligned} \tag{16}$$

Therefore, for $f \in \Omega_n$ we bound $h^2(s_{f_0}, s_f)$ and $||s_{f_0}/s_f||_\infty$. Substituting these bounds in (16) finishes the proof.

Step 1: showing that $h^2(s_{f_0}, s_f) \lesssim \delta_n^\rho$. By the definition of $f_{0,n}$ in (9), for any $(t,z) \in A_{n,j} \times B_{n,k}$,

$$\begin{aligned} |f_{0,n}(t,z) - f_0(t,z)| &= \left| |A_{n,j} \times B_{n,k}|^{-1} \int_{A_{n,j}} \int_{B_{n,k}} f_0(u,v) dv du - f_0(t,z) \right| \\ &\leq |A_{n,j} \times B_{n,k}|^{-1} \int_{A_{n,j}} \int_{B_{n,k}} |f_0(u,v) - f_0(t,z)| \, \mathrm{d}v \, \mathrm{d}u \\ &\leq \max_{(u,v) \in A_{n,j} \times B_{n,k}} |f_0(u,v) - f_0(t,z)|. \end{aligned}$$

By assumption (5) on f_0, we have

$$\max_{(u,v) \in A_{n,j} \times B_{n,k}} |f_0(u,v) - f_0(t,z)| \leq c \max_{(u,v) \in A_{n,j} \times B_{n,k}} ||(u,v) - (t,z)||^\rho \leq L(2\sqrt{2}\delta_n)^\rho.$$

Hence

$$||f_{0,n} - f_0||_\infty = \max_{j,k} \left| \max_{(t,z) \in A_{n,j} \times B_{n,k}} |f_{0,n}(t,z) - f_0(t,z)| \right| \leq c(2\sqrt{2}\delta_n)^\rho. \tag{17}$$

Applying Lemma A.1 with $f_1 \equiv f_0$ and $f_2 \equiv f \in \Omega_n$ gives

$$||s_{f_0} - s_f||_1 \leq C(\|f_0 - f_{0,n}\|_\infty + \|f_{0,n} - f\|_\infty) \lesssim \delta_n^\rho.$$

where we used (17) to bound the first term on the right-hand-side. Using the inequality $h^2(f_1, f_2) \leq \frac{1}{2}||f_1 - f_2||_1$, we then have

$$h^2(s_{f_0}, s_f) \leq \frac{1}{2}||s_{f_0} - s_f||_1 \lesssim \delta_n^\rho. \tag{18}$$

Step 2. showing that for sufficiently large n there exists a constant $\tilde{c} > 0$ such that $||s_{f_0}/s_f||_\infty \leq \tilde{c}$.

First note that

$$\left\| \frac{s_{f_0}}{s_f} \right\|_\infty \leq \max \left\{ \left\| \frac{\partial_2 F_0}{\partial_2 F} \right\|_\infty, \left\| \frac{1 - F_{0,X}}{1 - F_X} \right\|_\infty \right\} \leq \left\| \frac{f_0}{f} \right\|_\infty \tag{19}$$

By Assumption 3.1, there exists $\underline{M}$ such that $f_0(t,z) \geq \underline{M}$ for $(t,z) \in \mathcal{M}$. Since $f \in \Omega_n$ we have $f(t,z) \geq f_{0,n}(t,z) - C\delta_n^\rho$. By Equation (17), there exists a k (depending on c and ρ) such that

$$|f_0(t,z) - f_{0,n}(t,z)| \leq k\delta_n^\rho.$$

Therefore

$$\frac{f_0(t,z)}{f(t,z)} \leq \frac{f_0(t,z)}{f_{0,n}(t,z) - C\delta_n^\rho} \leq \frac{f_{0,n}(t,z) + k\delta_n^\rho}{f_{0,n}(t,z) - C\delta_n^\rho}$$

For n sufficiently large we will have $C\delta_n^\rho \leq \underline{M}/2$ and therefore the denominator of the right-hand-side can be lower bounded by $\underline{M}/2$. This implies boundedness of $\|f_0/f\|_\infty$ for n sufficiently large.

□

B Programming details in the Turing language

For each observation, indexed by $i \in \{1, \ldots, n\}$, we compute the vector of indices corresponding to the shaded area in Figure 1, as well as the fraction of the area that is shaded. This means that there is an n-dimensional vector `ci`, where each element is of type (**C**ensoring**I**nformation). The structure **C**ensoring**I**nformation has two fields: `ind` and `fracarea`, holding indices and areafractions respectively. If `L` is the graph-Laplacian (with a small multiple of the identity matrix added), the model is specified as follows:

```
@model GraphLaplacianMod(ci,L) = begin
    tau ~ Exponential(1.0)
    H ~ MvNormalCanon(L/tau)
    Turing.@addlogprob! loglik(H, ci)
end
```

Here the loglikelihood is calculated using

```
function loglik(H, ci)
    theta = softmax(H)
    ll = 0.0
    @inbounds for i in eachindex(ci)
        c = ci[i]
        ll += log(dot(theta[c.ind], c.fracarea))
    end
    ll
end
```

References

Aitchison, J. and Shen, S.M. (1980) *Logistic-normal distributions:Some properties and uses*, Biometrika **67**(2), p. 261–3444. https://doi.org/10.1093/biomet/67.2.261

Betancourt, M. (2018). A conceptual introduction to Hamiltonian Monte Carlo. arXiv:1701.02434.

Bezanson, J., Edelman, A., Karpinski, S. and Shah, V. B. (2017). Julia: A Fresh Approach to Numerical Computing. *Society for Industrial and Applied Mathematics.* **59**, p. 65-98.

Chen, D., Sun, J. and Peace, K. E. (2013). *Interval-Censored Time-to-Event Data: Methods and Applications*. Chapman and Hall/CRC Biostatistics Series.

Choudhuri, N., and Ghosal, S. and Roy, A. (2007). Nonparametric binary regression using a Gaussian process prior. *Statistical Methodology.* **4**, p. 227–243.

Cotter, S., Roberts, G., Stuart, A. and White, D. (2013). *MCMC Methods for Functions: Modifying Old Algorithms to Make Them Faster*. Statistical Science, **28**(3), p. 424-446.

Gao, B. and Pavel, L. (2018) *On the Properties of the Softmax Function with Application in Game Theory and Reinforcement Learning*, arXiv:1704.00805

Ge, H., Xu, K., and Ghahramani, Z. (2018). Turing: A Language for Flexible Probabilistic Inference. *Proceedings of the Twenty-First International Conference on Artificial Intelligence and Statistics*, PMLR: **84**, p. 1682-1690.

Ghosal, S., Ghosh, J.K. and Van der Vaart, A.W. (2000). Convergence rates of posterior distributions. *Ann. Statist.* **28**, p. 500–531.

Ghosal, S., van der Vaart, A. (2007). Posterior convergence rates of Dirichlet mixtures at smooth densities. *Ann. Statist.* **2**, p. 697–723.

Ghosal, S. and Van der Vaart, A.W. (2017). *Fundamentals of Nonparametric Bayesian Inference.* Cambridge University Press.

Groeneboom, P., Jongbloed, G. and Witte, B. I. (2011). Smooth Plug-in Inverse Estimators in the Current Status Mark Model. *Scandinavian Journal of Statistics.* **39**, p. 15–33.

Groeneboom, P., Jongbloed, G. and Witte, B. I. (2012). A maximum smoothed likelihood estimator in thecurrent status continuous mark model. *J. Nonparametr. Stat.* **24**, p. 85–101.

Groeneboom, P., Maathuis, M. H. and Wellner, J. A. (2008). Current status data with competing risks: Consistency and rates of convergence of the MLE. *Ann. Statist.* **36**, p. 1031–1063.

Gugushvili, S., van der Meulen, F. H., Schauer, M. R. and Spreij, P. (2019). Nonparametric Bayesian volatility estimation. (eds) *2017 MATRIX Annals, MATRIX Book Series.* **2**, p. 279–302.

Gugushvili, S., van der Meulen, F. H., Schauer, M. R. and Spreij, P. (2018). Bayesian wavelet de-noising with the Caravan prior. arXiv:1810.07668, to appear in ESAIM.

Hartog J. and van Zanten, H. (2017). Nonparametric Bayesian label prediction on a graph. arXiv:1612.01930

Hjort, N.L. (1994) Bayesian Approaches to Non- and Semiparametric Density Estimation. In Bayesian Statistics **5**. Proceedings of the Fifth Valencia International Meeting, June 5-9, 1994, edited by J. M. Bernardo, J. O. Berger, A. P. Dawid, and A. F. M. Smith,

Hoffman, M. D. and Gelman, A. (2014). The No-U-turn Sampler: Adaptively Setting Path Lengths in Hamiltonian Monte Carlo. *J. Mach. Learn. Res.* **15**, p. 1593–1623.

Huang, Y. and Louis, T. A. (1998). Nonparametric estimation of the joint distribution of survival time and mark variables. *Biometrika.* **85**, p. 7856-7984.

Hudgens, M. G., Maathuis, M. H. and Gilbert, P. B. (2007). Nonparametric estimation of the joint distribution of a survival time subject to interval censoring and a continuous mark variable. *Biometrics* **63**, p. 372-380.

Maathuis, M. H. and Wellner, J. A. (2008). Inconsistency of the MLE for the joint distribution of interval-censored survival times and continuous marks. *Scand. J. Statist.* **35**, p. 83-103.

Lenk, P.J. (1988). *The logistic normal distribution for Bayesian, non-parametric, predictive densities* .J. Amer. Statist. Assoc.**83**(402), p. 509516, MR0971380.

Van der Meulen, F.H. and Schauer, M. (2017) *Bayesian estimation of discretely observed multi-dimensional diffusion processes using guided proposals.* Electron. J. Statist. **11**(1), p. 2358–2396. doi:10.1214/17-EJS1290, `https://projecteuclid.org/euclid.ejs/1495850628`

Moala, F. A. and O'Hagan, A. (2010). Elicitation of multivariate prior distributions: A nonparametric Bayesian approach. *Journal of Statistical Planning and Inference.* **140**, p. 1635–3758.

Murphy, K. *Machine Learning: a Probabilistic Perspective.* MIT Press.

Neal, R. M.(1998). *Regression and classifica-tion using Gaussian process priors.* Available at `http://www.cs.toronto.edu/~radford/valencia.abstract.html`

Panaretos, V. and Zemel, Y. (2020) *An Invitation to Statistics in Wasserstein Space*, Springer.

Papaspiliopoulos, O. Roberts, G.O. and Sköld, M. *Non-centered parame-terizations for hierarchical models and data augmentation (with discussion)*, in BayesianStatistics **7** (Tenerife 2002), p. 307326. Oxford Univ. Press, New York.MR2003180

Robert C. P., Elvira V., Tawn N., and Wu C. (2018). Accelerating MCMC algorithms *Wiley Interdiscip. Rev. Comput. Stat.* **10**, e1435.

Ting, D., Huang, L. and Jordan, M.I. (2010) *An Analysis of the Convergence of Graph Laplacians*, In Proceedings of the 27th International Conference on Machine Learning (ICML), Haifa, Israel, 2010.

van de Meent, J. W., Paige, B., Yang, H. and Wood, F. (2018). An Introduction to Probabilistic Programming. arXiv:1809.10756.

Van der Vaart, A.W. (1998). *Asymptotic Statistics.* Cambridge University Press.

Van der Vaart, A.W. and Van Zanten, J.H. (2007) *Bayesian inference with rescaledGaussian process priors*, Electronic Journal of Statistics **1**, p. 433-448.