A framework for filtering in hidden Markov models with normalized random measures

Filtraggio di modelli di Markov nascosti in presenza di misure aleatorie normalizzate

Filippo ASCOLANI, Antonio LIJOI, Igor PRÜNSTER and Matteo RUGGIERO

Abstract The vast majority of explicitly available posterior characterizations in Bayesian nonparametrics refer to the exchangeable case, a restrictive assumption for time-dependent phenomena. Alternative formulations that accommodate partial exchangeability include hidden Markov models (HMMs), where the exact derivation of the posterior distribution given data collected at past times (*optimal filtering*) remains a challenging task. Here we outline a general framework based on duality for the analysis of HMMs which feature normalized random measures. The posterior tractability is ensured by combining certain projective properties of the infinite-dimensional distributions involved with the existence of a suitable duality relation between the hidden signal and an appropriate death process. Under these conditions, the filtering distributions are all finite mixtures, paving the way for closed form inferential strategies.

Abstract *I risultati analitici disponibili nel campo della statistica Bayesiana nonparametrica riguardano perlopiù il caso scambiabile, in cui la distribuzione deli dati è invariante rispetto a permutazioni finite. Poichè questa assunzione non è adatta per studiare fenomeni dinamici, sono state proposte molte alternative per i modelli di Markov nascosti. In questo lavoro proponiamo una classe generale che è dotata di grande trattabilità analitica, grazie a una relazione di dualità tra il segnale nascosto e un opportuno processo di Markov.*

Filippo ASCOLANI
Bocconi University and BIDSA, Milan, e-mail: filippo.ascolani@phd.unibocconi.it

Antonio LIJOI
Bocconi University and BIDSA, Milan, e-mail: antonio.lijoi@unibocconi.it

Matteo RUGGIERO
University of Torino and Collegio Carlo Alberto, Turin, e-mail: matteo.ruggiero@unito.it

Igor PRÜNSTER
Bocconi University and BIDSA, Milan, e-mail: igor.pruenster@unibocconi.it

Key words: Bayesian nonparametrics, hidden Markov models, duality, optimal filtering, completely random measures, partially observed Markov processes

1 Completely random measures and hidden Markov models

The standard Bayesian nonparametric specification for exchangeable data is

$$Y_i \mid X \overset{\text{iid}}{\sim} X, \quad X \sim \Pi, \tag{1}$$

where observations Y_i live in a Polish space $\mathbb{Y}$ with Borel sigma algebra $\mathscr{Y}$, and Π is a distribution on the space $\mathscr{P}_{\mathbb{Y}}$ of probability measures on $(\mathbb{Y}, \mathscr{Y})$ that plays the role of the prior. A general approach to construct such priors considers suitable transformations of completely random measures (CRMs); see [5, 6]. Denote by $(\Omega, \mathscr{F}, \mathbb{P})$ a probability space and by $\mathbb{M}_{\mathbb{Y}}$ the space of boundedly finite measures on $(\mathbb{Y}, \mathscr{Y})$, with corresponding Borel sigma algebra $\mathscr{M}_{\mathbb{Y}}$.

Definition 1. A measurable function μ from $(\Omega, \mathscr{F}, \mathbb{P})$ to $(\mathbb{M}_{\mathbb{Y}}, \mathscr{M}_{\mathbb{Y}})$, is a *completely random measure* if, for any disjoint collection $A_1, \ldots, A_n \in \mathscr{Y}$, the random variables $\mu(A_1), \ldots, \mu(A_n)$ are mutually independent.

In this work we focus on CRMs without deterministic drift and fixed points, so that μ can be characterized through the *Lévy-Khintchine* representation of its Laplace transform

$$\mathbb{E}\left[e^{-\lambda\mu(A)}\right] = e^{-P_0(A)\psi(\lambda)}, \quad \psi(\lambda) := \int_{\mathbb{R}_+}\int_{\mathbb{Y}} (1 - e^{-\lambda s})\, \nu(\mathrm{d}s, \mathrm{d}y), \quad A \in \mathscr{Y}, \lambda > 0,$$

where ν is a measure on $\mathbb{R}_+ \times \mathbb{Y}$ satisfying $\int_{\mathbb{R}_+}\int_A \min\{1, s\}\, \nu(\mathrm{d}s, \mathrm{d}y) < \infty$, $A \in \mathscr{Y}$, called *Lévy intensity*, and $\psi(\lambda)$ is the *Laplace exponent*. Here we will focus on *homogeneous* intensities $\nu(\mathrm{d}s, \mathrm{d}y) = \rho(\mathrm{d}s)\alpha(\mathrm{d}y)$, where ρ is a measure on $\mathbb{R}_+$ and the intensity of jumps does not depend on the jump location, and we further assume α is a finite non-atomic measure on $\mathbb{Y}$, with normalized version $P_0(\cdot) = \alpha(\cdot)/\alpha(\mathbb{Y})$. Typical examples are the *gamma* process, whereby $\nu(\mathrm{d}s, \mathrm{d}y) = s^{-1}e^{-s}\mathrm{d}s\, \alpha(\mathrm{d}y)$, and the σ-*stable* process, whereby $\nu(\mathrm{d}s, \mathrm{d}y) = (\Gamma(1-\sigma))^{-1}\sigma s^{-1-\sigma}\mathrm{d}s\, \alpha(\mathrm{d}y)$, for $0 < \sigma < 1$.

The following class of models is due to [9].

Definition 2. Let μ be a CRM such that $0 < \mu(\mathbb{Y}) < \infty$ almost surely. Then $p(\cdot) = \mu(\cdot)/\mu(\mathbb{Y})$ is called *normalized random measure with independent increments* (NRMI).

In the following we will further assume $\nu(\mathbb{R}_+ \times \mathbb{Y}) = \infty$ and $\psi(\lambda) < \infty$ for any positive λ, that guarantee almost sure finiteness and positivity of $\mu(\mathbb{Y})$. The class of NRMIs is large and encompasses many priors of interest: for example, the well-known Dirichlet process can be defined as a normalized gamma process. Moreover, the class of NRMIs enjoys a certain degree of analytical tractability, that makes posterior inference feasible (see [4]). In addition, a NRMI p admits the representation as discrete measure

$$p = \sum_{j \geq 1} W_j \delta_{Z_j}, \quad Z_j \overset{\text{iid}}{\sim} P_0,$$

with the weights $\{W_j\}$ independent from the locations $\{Z_j\}$. A sample $(Y_1, \dots, Y_n)$ will therefore yield ties with positive probability, and can be represented by the unique observed values $(Y_1^*, \dots, Y_k^*)$ with associated multiplicities $\mathbf{m} = (m_1, \dots, m_k) \in \mathbb{Z}_+^k$.

A natural extension of (1), that accommodates a partially exchangeable framework where observations are collected at different times $0 = t_0 < t_1 < \dots$, is given by

$$Y_{t_n}^i \mid X_{t_n} \overset{\text{iid}}{\sim} X_{t_n}, i = 1, \dots, n_{t_n}, \quad X \sim Q, \tag{2}$$

where $X = \{X_t : t \geq 0\}$ and Q is the law of a stochastic process indexed by $\mathbb{R}_+$ with state space $\mathscr{P}_{\mathbb{Y}}$. If X_t is a Markov process, then (2) is a *hidden Markov model* (HMM) (see [3]), and we denote by Q_0 the initial distribution of the process and by P_t its transition function. For brevity, let $Y_k := Y_{t_k}$ and $Y_{0:n} := (Y_0, \dots, Y_n)$. The main objects of interest are the *update* operator $\mathscr{U}_Y$, that returns the posterior distribution given observations Y at a fixed time, and the *prediction* operator, defined as applied to measures ξ by

$$\mathscr{P}_t(\xi)(\mathrm{d}x') = \int \xi(\mathrm{d}x) P_t(x, \mathrm{d}x'). \tag{3}$$

The so-called *filtering distribution* $P_n := \mathscr{L}(X_{t_n} \mid Y_{0:n})$ can then be obtained recursively by considering $P_0 = \mathscr{U}_{Y_0}(Q_0)$ and, for $n \geq 1$, $P_n = \mathscr{U}_{Y_n}\left(\mathscr{P}_{t_n - t_{n-1}}(P_{n-1})\right)$.

2 A general framework for optimal filtering

We provide general requirements on Q that lead to computing explicitly the filtering distributions. Let $X_t^K := (X_t(A_1), \dots, X_t(A_K))$ be the projection of X_t over an arbitrary measurable partition $A_1, \dots, A_K$ of $\mathbb{Y}$.

We make the following assumptions:

A1 Q_0 is induced by a NRMI with Lévy intensity ν.

A2 Q is such that $\mathscr{L}\left(X_t^K \mid X_0 = x\right) = \mathscr{L}\left(X_t^K \mid X_0^K = x^K\right)$, for any partition of K elements.

A3 Denoting by π^K the distribution of X_0^K and by P_t^K the transition function induced by Q relative to the partition, we assume π^K is reversible with respect to P_t^K, i.e. $\pi^K(\mathrm{d}x) P_t^K(x, \mathrm{d}x') = \pi^K(\mathrm{d}x') P_t^K(x', \mathrm{d}x)$. Moreover, we assume P_t^K admits a strictly positive transition density.

A4 The distribution of X_t^K given $(Y_t^1, \dots, Y_t^n)$ has density $h(x^K, Y^*, \mathbf{m}) \pi^K(x^K)$ for a suitable function $h(\cdot)$, that depends on the unique values Y^* and multiplicities $\mathbf{m} \in \mathbb{Z}_+^k$. We assume X_t^K is *dual* to a time-homogeneous death process M_t on

$\mathbb{Z}_+^k$, that is

$$\mathbb{E}\left[h(X_t^K, Y^*, \mathbf{m}) \mid X_0^K = x^K\right] = \mathbb{E}\left[h(x^K, Y^*, M_t) \mid M_0 = \mathbf{m}\right].$$

We denote by $p_{\mathbf{m},\mathbf{n}}(t)$ the transition probabilities of M_t, with $\mathbf{m} \in \mathbb{Z}_+^k$, $\mathbf{n} \in L(\mathbf{m})$ and $L(\mathbf{m}) = \{\mathbf{n} \mid \mathbf{n} \leq \mathbf{m}\}$, where $\mathbf{n} \leq \mathbf{m}$ if $n_j \leq m_j$ for any $j = 1, \dots, k$.

Example 1. Define the transition $P_t(x, \mathrm{d}x') = e^{-\beta t}\delta_x(\mathrm{d}x') + (1 - e^{-\beta t})Q_0(\mathrm{d}x')$. Then A1–A3 are immediately verified, while A4 reads

$$\mathbb{E}\left[h(X_t^K, Y^*, \mathbf{m}) \mid X_0^K = x^K\right] = e^{-\beta t} h(x^K, Y^*, \mathbf{m}) + 1 - e^{-\beta t},$$

so that $p_{\mathbf{m},\mathbf{m}}(t) = 1 - p_{\mathbf{m},\mathbf{0}}(t) = e^{-\beta t}$.

Example 2. The HMM induced by the Fleming–Viot process (see [8]) satisfies A1–A4 with Q_0 being the law of a Dirichlet process. Indeed, even if its transition function is known up to an infinite series, [7] proved that the projections are dual to a pure death process with rates $(m_j/2)(\alpha(\mathbb{Y}) + |\mathbf{m}| - 1)$, $|\mathbf{m}| = \sum_{j=1}^k m_j$, for jumping from $\mathbf{m}$ to $\mathbf{m} - \mathbf{e}_j$, where $\mathbf{e}_j$ denotes the vector of all zeroes except the j-th element. See [2] for an investigation of the predictive properties of this model.

Notice that, except for A1, the above requirements regard the finite-dimensional projections, typically more tractable especially in terms of transition functions. In this respect, note that NRMIs includes three classes of random measures for which the distribution of the projections is known explicitly (Dirichlet, normalized inverse-Gaussian and normalized stable processes).

3 Main results

In this Section we show how the tractability of NRMIs can be combined with duality in A4 to prove explicit a priori and a posteriori properties.

3.1 Prior properties

The first result shows that the invariance property of the projections extend to the distribution Q_0 itself.

Proposition 1. *Consider* (2) *with Q satisfying A1–A4. Then Q_0 is the invariant measure for the stochastic process with transition P_t.*

Proof. It follows from A3, since $\mathscr{L}(X_t^K)(\mathrm{d}x') = \int \pi^K(\mathrm{d}x) P_t^K(\mathrm{d}x, \mathrm{d}x') = \pi^K(\mathrm{d}x')$ and the fact that random measures are characterized by their finite dimensional distributions. □

Hence if $X_0 \sim Q_0$, $X_t \sim Q_0$ as well, and before conditioning on the data the same Bayesian nonparametric model for exchangeable data as in (1) is propagated to each time t.

Since Q is the law of a collection of random probability measures, one is immediately interested in the support properties. The *weak support* is the smallest closed set in the Borel sigma algebra $\mathscr{B}\{\mathscr{P}_{\mathbb{Y}}^{\mathbb{R}_+}\}$, generated by the product topology of weak convergence, and can be seen as a measure of flexibility: indeed, each neighborhood of an element of the support has positive probability under Q. The next proposition shows that our proposal yields a full weak support, relative to the support of P_0.

Proposition 2. *Let $\mathbb{S}$ be the support of P_0. Then the weak support of a model satisfying A1–A4 is given by $\mathscr{P}_{\mathbb{Y}}^{\mathbb{R}_+}(\mathbb{S})$.*

Proof. The marginal law Q_0 has weak support $\mathscr{P}_{\mathbb{Y}}(\mathbb{S})$. Then the result follows as for Proposition 1 in [2] using A2. □

When dealing with temporal data, it is often of interest to quantify the dependence between measures at different times. A simple way consists in using the correlation between the observables, that in this case can be computed exactly, as the next result highlights.

Proposition 3. *Consider* (2) *with Q satisfying A1–A4. Then*

$$Corr\left(Y_t^i, Y_{t+s}^j\right) = -p_{1,1}(s) \int_{\mathbb{R}_+} u \left\{ \frac{d^2}{du^2} \psi(u) \right\} e^{-\psi(u)}\, du.$$

In particular $Corr\left(Y_t^i, Y_{t+s}^j\right) \geq 0$ for any t and s.

Proof. Since X_t is almost surely discrete we have $\mathrm{Corr}(Y_t^i, Y_{t+s}^j) = \mathbb{P}(Y_t^i = Y_{t+s}^j) = p_{1,1}(s)\mathbb{P}(Y_t^i = Y_t^j)$. The latter is recovered from a reasoning similar to Proposition 2 in [4]. □

Considering for instance Example 1, with Q_0 being the law of a Dirichlet process, the formula reduces to $\mathrm{Corr}(Y_t^i, Y_{t+s}^j) = e^{-\beta t}/(\alpha(\mathbb{Y}) + 1)$.

3.2 Posterior properties

As shown in Proposition 1, at each fixed time t we have a sampling model as in (1) with the marginal law Q_0 in place of Π. We denote by $H(\cdot \mid \mathbf{m})$ the associated posterior distribution given data Y, with unique values $Y_1^*, \dots, Y_k^*$ and multiplicities $\mathbf{m}$. In the notation of Section 1, $\mathscr{U}_Y(Q_0)(\mathrm{d}x) = H(\mathrm{d}x \mid \mathbf{m})$.

The next result shows that the prediction operator yields a finite mixture of such distributions.

Theorem 1. *Consider model* (2) *with Q satisfying A1–A4. Then*

$$\mathscr{P}_t\left(H(dx,\boldsymbol{m})\right) = \sum_{\boldsymbol{n}\in L(\boldsymbol{m})} p_{\boldsymbol{m},\boldsymbol{n}}(t)H(dx,\boldsymbol{n})$$

Proof. Given an arbitrary partition, from A2 and A3 we have $\mathscr{L}(X_t^K \mid \mathbf{m})(\mathrm{d}x') = \int P_t^K(x,\mathrm{d}x')h(\mathrm{d}x,Y^*,\mathbf{m})\pi^K(\mathrm{d}x) = \pi^K(\mathrm{d}x')\mathbb{E}\left[h(X_t^K,Y^*,\mathbf{m}) \mid X_0^K = x'\right]$. The result now follows from A4. □

For instance, in the case of Example 1, it reads $\mathscr{P}_t\left(H(\mathrm{d}x,\mathbf{m})\right) = e^{-\beta t}H(\mathrm{d}x,\mathbf{m}) + (1-e^{-\beta t})Q_0(\mathrm{d}x)$.

Thanks to linearity of the prediction operator, the filtering distributions can be derived explicitly.

Theorem 2. *Consider* (2) *with Q satisfying A1–A4. Given unique values $Y_1^*,\dots,Y_k^*$ it holds*

$$\mathscr{L}(X_0 \mid Y_0) = H(dx \mid \boldsymbol{n}_0),$$

with $\boldsymbol{n}_0$ multiplicities of Y_0. Moreover, there exist $M_n \subset \mathbb{Z}_+^k$ and weights $w_{\boldsymbol{n}}$ such that

$$\mathscr{L}\left(X_{t_n} \mid Y_{0:n}\right) = \textstyle\sum_{\boldsymbol{n}\in M_n} w_{\boldsymbol{n}} H(dx \mid \boldsymbol{n}).$$

Proof. Since prediction operator (3) is linear, we apply the same reasoning of Proposition 2.3 in [7]. □

Since Q_0 is a NRMI, the posterior distribution $H(\cdot \mid \mathbf{n})$ is analytically tractable, at least conditionally to a suitable latent variable (see Theorem 2 in [4]). Thus, thanks to the finiteness of the mixture, devising conditional or marginal algorithms for sampling becomes a feasible operation. The results will be detailed and developed in [1].

References

1. Ascolani, F., Lijoi, A., Prünster, I. and Ruggiero, M.: Optimal filtering for hidden Markov models featuring normalized random measures. Work in progress.
2. Ascolani, F., Lijoi, A. and Ruggiero, M.: Predictive inference with Fleming–Viot-driven dependent Dirichlet processes. Bayesian Anal. (2020)
3. Cappé, O., Moulines, E. and Ryden, T.: Inference in Hidden Markov Models. Springer, New York (2005)
4. James, L. F., Lijoi, A. and Prünster, I.: Posterior analysis for normalized random measures with independent increments. Scand. J. Stat. **36**(1), 76–97 (2009)
5. Kingman, J.F.C.: Completely Random Measures. Pacif. J. Math. **21**, 59–78 (1967)
6. Lijoi, A. and Prünster, I.: Models beyond the Dirichlet process. In *Bayesian Nonparametrics*, pp. 80–130, Cambridge Univ. Press, Cambridge (2010)
7. Papaspiliopoulos, O. and Ruggiero, M.: Optimal filtering and the dual process. Bernoulli. **20**(4), 1999–2019 (2014)
8. Papaspiliopoulos, O., Ruggiero, M. and Spanó, D.: Conjugacy properties of time-evolving Dirichlet and gamma random measures. Electron. J. Stat. **10**(2), 3452–3489 (2016)
9. Regazzini, E., Lijoi, A. and Prünster, I.: Distributional results for means of normalized random measures with independent increments. Ann. Stat. **31**(2), 560–585 (2003)