EVALUATING FIT OF DIRICHLET PROCESS USING HELLINGER DISCRIMINATION

PAPA NGOM○, RICHARD EMILION○∗
○ LMA - Université Cheikh Anta Diop - Dakar - Sénégal
∗ MAPMO - Université d’Orléans - France

Abstract. The aim of this work is to extend a Bayesian method for evaluating fit, based on model embeddings proposed by Carota, Parmigiani, and Polson [4]. Our approach is to make estimating updates of posterior distribution of Hellinger’s discrimination between the proposed family of distributions and the true distribution generating the data. This paper addresses the question of consistency of estimating posterior distribution whether the prior distribution is given by Dirichlet Process. We illustrate this method by using Markov Chain Monte Carlo (MCMC) techniques for computing the posterior distribution of the Hellinger metric, when this posterior is a Dirichlet Process model.

1. Introduction

Nonparametric Bayesian methods have been popular and successful in many estimation problems but their relevance in hypothesis testing situations have become of interest only recently. In particular, the testing of model is an exact fit has received considerable attention from Bayesian, e.g, Kass and Rafety [10], Gelman, Meng, Stern [7]. We develop an extended version of the method proposed by Carota, Parmigiani, and Polson [4]. This method is based on embedding a proposed model class into a larger group of models intended to represent uncertainty in the model specification. The posterior distribution is then estimated within the larger class of models and the marginal distribution of the distance to the proposed class is found, with small distances indicating good fit and large distances indicating poor fit. This method has focused on quantifying the lack of fit, not determining whether an assumption exactly describes the actual distribution of the process that generated the data. Viele [16] considers a similar method based on embedding a proposed model in a family of discrete distribution of Kullback-Leibler information for estimating the lack of fit of the proposed model. Suppose we have a proposed family of models \( D_θ \). From an observed set of data, we wish to evaluate the fit of \( D_θ \). The approach consists of embedding the proposed model into a larger family of models, assuming the true process generating the data is within the larger family, and then computing a posterior distribution for the Hellinger’s discrimination which we can denoted \( H(P, D_θ) \), between the true and the proposed models. We consider the approach of estimation the distance \( P \) and the family of models, instead of test. We estimate the process \( P \) by

Key words and phrases. Probabilistic inference, Gaussians mixtures, EM algorithm, divergence statistics, test of goodness fit.
using nonparametric Bayesian methods and suggest how to compute point estimate $H(P, D_\theta)$.

In this paper, our goal is to establish consistency method for estimating $H(P, F_\theta)$. The remainder of this paper is organized as follows: in section 2, we describe some definitions and the method of choosing the Hellinger metric using Dirichlet Processes. Section 3 presents notations definitions and section 4 gives main results about consistency of the posterior distribution. We provide a proof of consistency of Hellinger neighborhood when using discrete Dirichlet Process priors. We propose a simulation study in Section 5, for illustrating the method for evaluating the fit of discrete distribution. We end this paper with a discussion of results in Section 6. The current paper builds on previous unpublished work by Viele [16].

2. Method

The following notations will be used throughout the paper. Let $\lambda$ be a probability measure on a measurable space $(X, B)$, where the $\sigma$–field $B$ is separable. Let $Q$ be the set of all finite measures on $(X, B)$ that are absolutely continuous with respect to $\lambda$.

One the most important issues in statistical inference theory is to distinguish two probability distributions based on some observations. In this respect, various distance measures (metrics) play crucial role. Apart from the relative entropy (Kullback-Leibler entropy)[11], the most fundamental distance is the so-called Hellinger distance[12].

The Hellinger distance is intended to quantify the lack of fit between the $P$ and the null family $D_\theta$. It is all known that absolute continuity with respect to a $\sigma$–finite measure is equivalent to absolute continuity with respect to a probability measure. Let $H(\cdot, \cdot)$ denote the Hellinger metric on $Q$:

\[
H(Q_1, Q_2) = \int \left[ f_1(x)^{1/2} - f_2(x)^{1/2} \right]^2 d\lambda(x)
\]

where $f_i = \frac{dQ_i}{d\lambda}$.

In case that $f_1$ and $f_2$ are discrete probability distributions, their Hellinger distance is

\[
H(Q_1, Q_2) = \sum_x \left[ \sqrt{f_1(x)} - \sqrt{f_2(x)} \right]^2
\]

Inspired by the original Hellinger distance, we introduce a distance between $P$ and the family $D_\theta$ as :

\[
d(P, D_\theta) = \inf_{\theta \in \Theta} H(P, D_\theta)
\]

where $H$ stands for Hellinger discrimination.

Let $H(P, D_\theta)$ be a the distance between $P$ and the family $D_\theta$. In the Bayesian framework, many way of evaluating fit whether $H(P, D_\theta)$ is a "sufficient" model is to test the following hypothesis :
EVALUATING FIT USING HELLINGER DISCRIMINATION

\[
H_0 : H(P, D_\theta) \leq a \\
H_1 : H(P, D_\theta) > a
\]

In this paper, we use the approach of estimation. Our goal is to provide a method of estimating how closely \( D_\theta \) approximates the true model \( P \).

We assess the fit of a null family of discrete distributions \( D_\theta \), where discreteness implies there exists a countably infinite set \( S \) such that \( D_\theta(S) = 1 \) for all \( \theta \).
Let \( X_1, \ldots, X_n \) be \( n \) i.i.d observations from a distribution \( P_\theta \). The \( n \)-fold product of \( P_\theta \) on the product space \((\mathcal{X}^n, \mathcal{B}^n)\) will be denoted \( P_n \) and the infinite product measure will be denoted \( P^\infty \).
We are evaluating the fit of \( D_\theta \), we cannot assume and the true distribution that generating the data, \( P_\theta \), is an element of \( D_\theta \).
Our parameter of interest is the distribution \( P \), that resides in space, which contains every possible value of \( P \). We assume that all the \( D_\theta \) distributions are supported on this space. We define the alternative family of distributions as every possible distribution on the nonnegative integers.

3. Dirichlet Processes

Dirichlet process models go back to Ferguson [6] in a fundamental paper on a Bayesian approach to nonparametric problems. He defines the Dirichlet process (DP) as follows: let \( \mathcal{X} \) be a space and \( \mathcal{B} \) a \( \sigma \)-field of subsets of \( \mathcal{X} \). Let \( D_\theta \) consists of a single discrete distribution, \( D_0 \) defined on the measurable space \((\mathcal{X}, \mathcal{B})\), and we wish to construct a nonparametric prior centered \( D_\theta \).
A random probability measure \( P \) on \((\mathcal{X}, \mathcal{B})\) is a Dirichlet process on \((\mathcal{X}, \mathcal{B})\) with scaling parameter \( \alpha \) and baseline distribution \( D_0 \), denoted \( P \in \text{Dir}(D_0 \times \alpha) \), if for all natural \( k = 1, 2, \ldots, m \), the random vector \((B_1, B_2, \ldots, B_m)\) is measurable partition of \( \mathcal{X} \) with \( P_k = \text{Prob}(B_k) \), the joint distribution of the random probabilities \((P_1, P_2, \ldots, P_m) \in \text{Dir}(\alpha D_0(B_1), \alpha D_0(B_2), \ldots, \alpha D_0(B_m)) \).
A well know result is that the class of \( P \sim \text{Dir}(\alpha D_0) \) is closed, in the sense that if the prior is a DP,

\[
P \sim \text{Dir}(D_0 \times \alpha) \\
X_1, \ldots, X_n/P \sim P
\]
then the posterior distribution of \( P \) is also a DP with base probability measure \((\alpha D_0 + \sum_{i=1}^{n} \delta_{X_i})/(\alpha + n)\) and scaling parameter \( \alpha + n \):

\[
P/X_1, \ldots, X_n \sim \text{Dir}(D_0 \times \alpha + \sum_{i=1}^{n} \delta_{X_i})
\]
where \( \delta_{X_i}(x_j) = 1 \) if \( x_i = x_j \) and 0 otherwise.

**Theorem 3.1.** (Barron [3]) Suppose \( D_\theta \) is a family of distributions and \( P \) is another distribution, which may or may not be an element of \( D_\theta \). Let \( X_1, X_2, \ldots \sim P \), if there exists a prior \( \Pi(\theta) \) on \( \Theta \) and a point \( \theta_p \) such that the posterior distribution \( \Pi_n = \Pi(\theta \setminus X_1, \ldots, X_n) \) is asymptotically carried on \( \theta_p \), in the sense that if \( U \) is an open set containing \( \theta_p \), then \( \lim_{n \to \infty} \Pi_n(U) = 1 \) almost surely \([P^\infty]\).
For $\epsilon > 0$, we define

$$A_\epsilon = \{ P : H(P, D_\theta) \leq \epsilon \}$$

Recall that we denote $X_1, X_2, \ldots, X_n$ by $X_n$.

**Proposition 3.2.** Define the Bayes estimate $D_{\theta_p}$ to be the probability measure,

$$D_{\theta_p}(A) = \int D_\theta(A) \Pi(dP \setminus X_1, X_2, \ldots, X_n) = E[D_\theta(A) \setminus X_n]$$

then

$$\inf_{\theta \in \Theta} H(P, D_\theta) = H(P, D_{\theta_p})$$

**Proof.** Let $\Phi(\cdot) = H(P, \cdot)$ is convex and from Jensen’s inequality :

$$\Phi(D_{\theta_p}(\cdot)) = \Phi\left( \int D_\theta(\cdot) \Pi(dP \setminus X_n) \right) \leq \int \Phi(D_\theta(\cdot)) \Pi(dP \setminus X_n)$$

In this way we obtain the following result :

$$H(P, D_{\theta_p}) \leq \int H(P, D_\theta) \Pi(dP \setminus X_n)$$

(3)

The first term on the right-hand side of (3) is at most $\epsilon$ by the definition of $A_\epsilon$, and the second term goes to 0 a.s $[P^\infty]$ by the Theorem 3.1 and the fact that Hellinger measure is bounded. Since $\epsilon$ is arbitrary, the result follows. $\square$

Our goal is to estimate $h_o = H(P_o, D_\Theta)$ using the posterior distribution of $h = H(P, D_\Theta)$ as $P$ ranges over its posterior distribution within the alternative family. The value $h_o$ quantifies the divergence between the true distribution $P_o$ and the null family of distributions.

Let $D_\Theta$ the parametric family of distributions, we place a nonparametric prior around this family which we write $D_{\text{Dir}}(D_\Theta \times \alpha)$. So, the prior now becomes : $P \sim D_{\text{Dir}}(D_\Theta \times \alpha)$. Hence the posterior distribution of $P$ given $X_1, \ldots, X_n$ can be conveniently written as $D_{\text{Dir}}(\alpha \times D_\Theta + \sum_{i=1}^n \delta_{X_i})$. It may be difficult to obtain the posterior of $H(P, D_\Theta)$ analytically. Use of Dirichlet process has become computationally feasible with the development of Markov chain Monte Carlo methods (MCMC) for sampling from the posterior distribution of $H(P, D_\Theta)$. Methods based on Gibbs sampling can easily be implemented for models based on conjugate prior distributions (Escober [5], Gilks, Richardson and Speigelhalter [8]). For each distribution $P_i$ from the posterior of $P$, we produce $h_i = H(P_i, D_\Theta)$.

4. **Main results**

We give, in this section, conditions that guarantee that the posterior probability of Hellinger distance $h = H(P, D_\Theta)$ between $P$ and $D_\Theta$ tends to a point mass at $h_o = H(P_o, D_\Theta)$.

**Lemma 4.1.** Let $X_1, X_2, \ldots, P_o$ and $\theta_o = \arg\min_{\theta_o} H(P_o, D_\theta)$. Also, for any $P$ in the alternative family define $\theta_P = \arg\min_{\theta_P} H(P, D_\theta)$. Let $\pi$ be a prior over the
alternative class and let \( \pi_n \) be the posterior distribution based on \( X_1, X_2, \ldots, X_n \). Also, define three sets of neighborhoods

\[
U_{1,\delta} = \{ P : |H(P, D_{\theta_0}) - H(P_0, D_{\theta_0})| < \delta \}
\]

\[
U_{2,\delta} = \{ P : |H(P, D_{\theta_0}) - H(P, D_{\theta_0})| < \delta \}
\]

\[
U_{\inf,\delta} = \{ P : \inf_{\theta} H(P, D_{\theta}) - \inf_{\theta} H(P_0, D_{\theta}) < \delta \}
\]

Suppose that, for all \( \epsilon \) and \( \delta \)

1. \( \lim_{\epsilon \to 0} \Pr(\pi_n(U_{1,\delta}) > 1 - \epsilon) = 1 \)
2. \( \lim_{\epsilon \to 0} \Pr(\pi_n(U_{2,\delta}) > 1 - \epsilon) = 1 \)

Then for all \( \epsilon \) and \( \delta \)

\[
\Pr(\pi_n(U_{\inf,\delta}) > 1 - \epsilon) = 1
\]

**Proof.** Let \( H(P, D_{\theta_0}) = \inf_{\theta} H(P, D_{\theta}) \) and \( H(P_0, D_{\theta_0}) = \inf_{\theta} H(P_0, D_{\theta}) \).

In this case, by the triangle inequality:

\[
|H(P, D_{\theta_0}) - H(P_0, D_{\theta_0})| \leq |H(P, D_{\theta_0}) - H(P, D_{\theta_0})| + |H(P, D_{\theta_0}) - H(P_0, D_{\theta_0})|
\]

Therefore, \( U_{1,\delta/2} \cup U_{2,\delta/2} \subset U_{\inf,\delta} \) and the result follows. \( \square \)

Typically, the condition(1) corresponds to the consistency of \( H(P, D_{\theta_0}) \), which means that consistency at \( \theta_0 \), not the entire family \( D_{\theta} \).

Hence, with the standard results in parametric alternative families, one can have without difficulty the condition(1). In other words, the Hellinger divergence between two probability measures is often a continuous transformation of the parameters, and therefore the posterior of \( P \) concentrates around the distribution \( P_0 \) \( (P \) converges to \( \delta_{P_0}) \), then the posterior distribution of \( H(P, D_{\theta}) \) should converge to \( H(P_0, D_{\theta}) \).

Let \( d_{\theta} = \frac{dD_{\theta}}{dx} \), we can see that the neighborhood \( U_{2,\delta} \) is those \( P \) satisfying:

\[
|H(P, D_{\theta_0}) - H(P, D_{\theta_0})| = \left| \int \left[ \left( \frac{d\theta_0 - d\theta_0}{dP(x)} - 2 \left( \frac{d^{1/2}\theta_0(x) - d^{1/2}\theta_0(x)}{(dP(x))^{1/2}} \right) \right) \right] dP(x) \right| < \delta
\]

Often condition (2) of Lemma(4.1) is easy to verify. Usually, for a Dirichlet Process prior, \( \theta_0 \) converges to \( \theta_0 \), resulting, condition(2) being satisfied.

Note that the standard asymptotic results are not usually suitable for nonparametric settings, since the parameter is infinite dimensional.

For Dirichlet Process prior, we will prove condition(1) of Lemma(4.1) by using Theorem(4.2).

Next we move on case where \( X = 1, 2, 3, \ldots \) The set \( P(X) \) of probability measures on \( X \). Let \( D, P \) and \( Q \) be elements of \( P(X) \) with \( D \) and \( Q \) unknown but fixed. \( X_1, X_2, \ldots \) is a sequence of \( X \)-valued random variables that are independent and identically distributed as \( P_0 \). It is convenient to consider \( X_1, X_2, \ldots \) as the coordinate random variable defined on \( \Omega = (X^\infty, A^\infty) \) and \( P^\infty \) as the iid product measure defined on \( \Omega \).

We place a Dirichlet process on \( P \) with base measure \( \gamma Q \) for \( \gamma \geq 0 \) \( (P \sim D(\gamma Q)). \) Hence, the posterior distribution of given \( X_1, X_2, \ldots, X_n, \pi_n(P/X) \) (recall that we denote \( X_1, X_2, \ldots, X_n \) by \( X \)), is therefore (Ferguson [6], a Dirichlet Process with
base measure \((\gamma Q + \sum_{i=1}^{n} \delta_{X_i})\), which can be written \(\pi_n(P/X) \sim \mathcal{D}(\gamma Q + n\hat{P}_n)\), where 

\(\hat{P}_n\) is the empirical distribution of the first \(n\) observation. We consider \(\hat{p}_n\) to be the empirical proposition of \(X_1, X_2, \ldots, X_n\) that equal to \(i\).

Let \(\hat{P}_n\) be a randomly drawn distribution from \(\pi(P/X)\) and \(\tilde{p}_n\) be the posterior probability \(\hat{P}_n\) assigns to \(i\).

Assume there exists an \(\eta\) such that, for \(b_i = \max(p_i^n, q_i^n)\), the sums \(\sum_i p_i^{1-2\eta}\), \(\sum_i q_i^{1-2\eta}\), and \(\sum b_i\) are all finite.

Let \(c_n\) be a sequence such that \(c_n > 1\) for all \(n\), \(\lim_n c_n = \infty\), and \(\lim_n \frac{c_n}{n^{1/2}} = 0\)

Let \(F_{\infty}\) be the event that \(\hat{\pi}_n \leq p_i + \frac{b_i}{c_n}\) and let \(K_n = \bigcap_{i=0}^{\infty} F_{\infty}\).

Let \(U_{\varepsilon} = \{F \in \mathcal{P}(X) : |H(F, D) - H(F, D)| < \varepsilon\}\). 

Note that the assumption concerning the three sums is satisfied if \(D, Q, P_0\) are all of the form \(\exp\{-\varphi(i)\}\), where \(\varphi(i)\) is a finite degree polynomial in \(i\).

**Theorem 4.2.** For all \(\varepsilon > 0\), \(\lim_{n \to +\infty} P_n(\pi_n(U_{\varepsilon}) > 1 - \varepsilon) = 1\)

**Proof.** of theorem 4.2

**Lemma 4.3.** Let \(\hat{P}_n\) be a randomly drawn distribution from \(\pi_n(P/X)\), and let \(\tilde{p}_n\) be the probability \(\hat{P}_n\) assigns to \(i\), then

\[
\mathbb{E}(\tilde{p}_n) = \frac{np_i + \gamma q_i}{n + \gamma} \quad \text{and} \quad V(\tilde{p}_n) \leq \frac{(2 + \gamma)p_i + (\gamma + \gamma^2)q_i}{n}
\]

**Proof.** Let \(\hat{\pi}_m\) to be the empirical distribution of \(X_1, \ldots, X_n\) that are equal to \(i\). We have

\[n\hat{\pi}_m \sim Bin(n, \ p_i)\]

and

\[\tilde{p}_n|\hat{\pi}_m \sim Beta(n\hat{\pi}_m + \gamma q_i, \ n(1 - \hat{\pi}_m) + \gamma(1 - q_i))\]

Therefore

\[
\mathbb{E}(\tilde{p}_m) = \mathbb{E}[\mathbb{E}(\tilde{p}_n|\hat{\pi}_m)] = \mathbb{E}\left[\frac{n\hat{\pi}_m + \gamma q_i}{n + \gamma}\right] = \frac{np_i + \gamma q_i}{n + \gamma}
\]

and

\[
V(\tilde{p}_n) = \mathbb{E}[V(\tilde{p}_m|\hat{\pi}_m)] + V[\mathbb{E}(\tilde{p}_n|\hat{\pi}_m)]
\]

\[
= \mathbb{E}\left[\frac{(n\hat{\pi}_m + \gamma q_i)(n(1 - \hat{\pi}_m) + \gamma(1 - q_i))}{n + \gamma^2(n + \gamma + 1)}\right] + V\left[\frac{(np_i + \gamma q_i)}{(n + \gamma)}\right]
\]

\[
= \mathbb{E}\left[\frac{n^2\hat{\pi}_m(1 - \hat{\pi}_m)n\gamma \hat{\pi}_m(1 - \hat{\pi}_m) + \gamma^2 q_i(1 - q_i)}{(n + \gamma)^2(n + \gamma + 1)}\right] + \frac{np_i(1 - p_i)}{(n + \gamma)2}
\]

\[
= \frac{n^2\hat{\pi}_m(1 - \hat{\pi}_m)n\gamma \hat{\pi}_m(1 - \hat{\pi}_m) + \gamma^2 q_i(1 - q_i)}{(n + \gamma)^2(n + \gamma + 1)} + \frac{np_i(1 - p_i)}{(n + \gamma)2}
\]
The variance can be upper bounded by
\[
V(\tilde{p}_{in}) \leq \frac{n(n-1)p_i + n\gamma(p_i + q_i) + \gamma^2 q_i}{(n+\gamma)^2(n+\gamma+1)} + \frac{np_i}{(n+\gamma)^2} \\
\leq \frac{n^2 p_i + n\gamma(p_i + q_i) + \gamma^2 q_i}{n^3} + \frac{np_i}{n^2} \\
\leq \frac{2p_i}{n} + \frac{2(p_i + q_i) + \gamma^2 q_i}{n^2} \\
\leq \frac{(2 + \gamma)p_i + (\gamma + \gamma^2)q_i}{n}
\]

\[\square\]

**Proposition 4.4.** For any \(\varepsilon > 0\), \(\lim_{n \to +\infty} P_n(H(\tilde{P}_n, D) > H(P, D) + \varepsilon) = 0\)

**Proof.** Using lemma 4.4 and Chebychev’s inequality:

\[
P_n(F^c_{in}) = P_n(\tilde{p}_{in} > p_i + \frac{b_i}{c_n}) = P_n(\tilde{p}_{in} - \frac{np_i}{n + \gamma} > p_i + \frac{b_i}{c_n} - \frac{np_i}{n + \gamma}) \\
\leq P_n\left(\left|\tilde{p}_{in} - \frac{np_i}{n + \gamma}\right| > \frac{b_i}{c_n} + \frac{\gamma(p_i - q_i)}{n + \gamma}\right) \\
\leq \frac{\frac{\gamma c_n p_i}{n + \gamma}}{\left(\frac{b_i}{c_n} + \frac{\gamma(p_i - q_i)}{n + \gamma}\right)^2}
\]

Let \(N\) be an integer such that \(\frac{\gamma c_n}{n + \gamma} < 1\) and \(q_i^q - q_i > \frac{q_i^q}{2}\) for all \(i > N\) (such an \(N\) exists by the construction of \(c_n\) and the fact that \(Q \in \mathcal{F}\)). Then

\[
P_n(K^c_n) = \sum_{i=0}^{N} P_n(F^c_{in}) + \sum_{i=N+1}^{\infty} P_n(F^c_{in}) \\
\leq \sum_{i=0}^{N} P_n(F^c_{in}) + \frac{\epsilon^2 c_n}{n} \sum_{i=N+1}^{\infty} \frac{p_i(2 + \gamma) + q_i(\gamma^2 + \gamma)}{b_i + \gamma c_n(p_i - q_i)} \\
\leq \sum_{i=0}^{\infty} P_n(F^c_{in}) + \frac{\epsilon^2 c_n}{n} \left(\sum_{i>N; p_i \geq q_i} \frac{p_i(2 + \gamma)}{b_i + \gamma c_n(p_i - q_i)} + \sum_{i>N; q_i \geq p_i} \frac{q_i(2 + \gamma)}{b_i + \gamma c_n(p_i - q_i)}\right) \\
\leq \sum_{i=0}^{\infty} P_n(F^c_{in}) + \frac{\epsilon^2 c_n}{n} \left(\sum_{i>N; p_i \geq q_i} \frac{p_i(2 + \gamma)}{p_i^q + \gamma c_n(p_i - q_i)} + \sum_{i>N; q_i \geq p_i} \frac{q_i(2 + \gamma)}{q_i^q + \gamma c_n(p_i - q_i)}\right)
\]
Using the assumptions defining $N$, we find this quantity is:

$$\leq \sum_{i=0}^{N} P_n(F_{in}) + \left( \frac{2 + 2\gamma + \gamma^2}{n} \right) \left( \sum_{i>N, p_i \geq q_i} p_i^{1-2\eta} + \sum_{i>N/q_i > p_i} q_i^{1-2\eta} \right)$$

The first sum converges to 0 as $n$ increases because it involves a finite number of parameters $p_0, \ldots, p_N$. Each $\tilde{p}_{in}$ converges to $p_i$ at rate $n^{-1/2}$, which is faster than $1/c_n$ by assumption. In addition, the two sums inside the brackets are finite by assumption. Therefore, the entire quantity tends to 0 as $n$ increases, resulting in $P_n(K_n)$ tending to 1 as $n$ increases. When $K_n$ occurs,

$$H(\tilde{P}, D) = \sum_{i=0}^{\infty} (\tilde{p}_{in}^{1/2} - d_i^{1/2})^2$$

$$\leq \sum_{i=0}^{\infty} \left[ (p_{in} + \frac{b_i}{c_n})^{1/2} - d_i^{1/2} \right]^2$$

$$\leq \sum_{i=0}^{\infty} \left[ (p_{in} + \frac{b_i}{c_n}) - d_i^{1/2} (p_{in} + \frac{b_i}{c_n})^{1/2} \right]$$

$$\leq \sum_{i=0}^{\infty} p_{in} + \sum_{i=0}^{\infty} d_i + \sum_{i=0}^{\infty} \frac{b_i}{c_n} - 2 \sum_{i=0}^{\infty} d_i^{1/2} (p_{in} + \frac{b_i}{c_n})^{1/2}$$

$$\leq H(P, D) + \frac{1}{c_n} \sum_{i=0}^{\infty} b_i$$

The sum in the last term is finite by assumption, hence for all $\varepsilon$,

$$\lim_{n \to \infty} P_n \left( H(\tilde{P}, D) > H(P, D) + \varepsilon \right) = 0$$

\[\square\]

**Proposition 4.5.** For any $\varepsilon > 0$, \(\lim_{n \to \infty} P_n (H(\tilde{P}_n, D) \geq H(P, D) - \varepsilon) = 1\)

**Proof.** Since $H(\tilde{P}, D)$ is finite by assumption, for any $\varepsilon$ we may find a $k$ such that

$$\sum_{i=0}^{k} (\tilde{p}_{in}^{1/2} - d_i^{1/2})^2 > H(P, D) - \frac{\varepsilon}{2} \quad \text{and} \quad \sum_{i=k+1}^{\infty} d_i < \left( \frac{\varepsilon}{4} \right)^2$$

Let

$$H(\tilde{P}, D) = \sum_{i=0}^{k} (\tilde{p}_{in}^{1/2} - d_i^{1/2})^2 + \sum_{i=k+1}^{\infty} (\tilde{p}_{in}^{1/2} - d_i^{1/2})^2$$

Let $f : [0, +\infty[ \to \mathbb{R}$ be a function defined as follows: $f(x) = (1 - \sqrt{x})^2$. We have:

$$\sum_{i} (\tilde{p}_{in}^{1/2} - d_i^{1/2})^2 = \sum_{i} \tilde{p}_{in} f \left[ \frac{d_i}{\tilde{p}_{in}} \right] \geq f \left( \sum_{i} \tilde{p}_{in} \frac{d_i}{\tilde{p}_{in}} \right)$$
The inequality follows from Jensen’s inequality and the convexity of $f$.

$$H(\hat{P}, D) \geq \sum_{i=0}^{k} (\hat{p}_{i+n}^{1/2} - d_{i}^{1/2})^2 + \left(1 - \sqrt{\sum_{i=k+1}^{\infty} d_{i}}\right)^2$$

$$\geq \sum_{i=0}^{k} (\hat{p}_{i+n}^{1/2} - d_{i}^{1/2})^2 - 2 \left(\sqrt{\sum_{i=k+1}^{\infty} d_{i}}\right)^{1/2}$$

$$\geq \sum_{i=0}^{k} (\hat{p}_{i+n}^{1/2} - d_{i}^{1/2})^2 - \frac{\varepsilon}{2}$$

(4)

Since $p_0, \ldots, p_k$ are finite set of parameters and each $\hat{p}_{i+n}$ converges almost surely to $p_i$,

$$\lim_{n \to +\infty} P_n \left(\sum_{i=0}^{k} (\hat{p}_{i+n}^{1/2} - d_{i}^{1/2})^2 \geq \sum_{i=0}^{k} (p_{i}^{1/2} - d_{i}^{1/2})^2 - \frac{\varepsilon}{2}\right) = 1$$

Combining this with the equation (4),

$$\lim_{n \to +\infty} P_n \left(H(\hat{P}_{n}, D) \geq H(P, D) - \varepsilon\right) = 1$$

□

Proposition 4.4 and proposition 4.5 give the result. □

5. Simulated study

5.1. Implementation. Measures drawn from a Dirichlet process turn out to be discrete with probability one [6]. This property is made explicit in the stick-breaking construction due to Sethuraman [14].

In the stick-breaking representation, $P \sim \text{Dir}(D_0 \times \alpha)$ can be written as an infinite sum of spikes:

$$P = \sum_{i}^{\infty} \pi_{i} \delta_{X_{i}} \quad \text{where} \quad X_{i}/\alpha, D_0 \sim D_0 \quad \text{and} \quad \pi_{i} = \beta_{k} \Pi_{j=1}^{k-1} (1 - \beta_{j})$$

with $\beta_{k}/\alpha, D_0 \sim \text{Beta}(1, \alpha)$.

Here $D_0$ is a measure on $X_i$ and $\alpha$ determine how closely the histogram of spikes represents $D_0$.

Sethuraman [14] showed that $P$ as defined in this way is a random probability measure distributed according to $\text{Dir}(D_0 \times \alpha)$.

5.2. Prior and computation. One would prefer to have some control over the prior mass placed on distributions close to $D_0$.

For example, let $P$ be distributed as a Dirichlet Process with base probability measure $D_0$ and closeness parameter $\alpha : P \sim \text{Dir}(D_0 \times \alpha)$

We choose in first the special case $\alpha = 1$, which corresponds to the correctly specified prior : $P \sim \text{Dir}(P(5) \times 1)$.

We simulated 20,000 distributions from the prior with $D_0 = \text{Pois}(5)$ and computed the Hellinger metric (1) for each.

The figure 1 shows the histogram of results.

We note that most of the mass is concentrated between 0.1 and 0.4. Just few
Hellinger distance less than 0.1.

We choose four different values of the parameter $\alpha$:
- two values before $\alpha = 1$,
- two values after $\alpha$.

The results confirm that if $\alpha$ takes small value, the prior places little mass in neighborhoods close to the null class.

Figures 2 gives histograms. For instance, with $\alpha < 1$, (Figure 2 a, b, with $\alpha = 0.1$ and $\alpha = 0.5$), the prior moves away from the null class.

On the other hand, the prior places considerable mass in neighborhoods near zero, for large values of $\alpha$.

Letting $\alpha > 1$ (Figure 2 c, d, with $\alpha = 5$ and $\alpha = 15$), allows one to choose the prior with concentration mass in neighborhoods close to the null class.

Here, We propose the prior of the Hellinger metric : $H(P, D_0)$ for $\alpha = 20$. This prior distribution is much closer to 0 than when the parameter $\alpha$ is in $[0, 1]$.

Avoiding a mode away from a distance of 0 is useful. Inexamining the posterior distribution of Hellinger distance, a mode away from 0 may be taken as evidence the null class is not the correct model. The present of this characteristic in the prior, don’t allows us to conclude whether a mode in the posterior distribution is determined by the data or by the prior. Constructing a prior that avoids modes allows cleaner interpretation of the posterior distribution.

We consider here, for evaluating fit, a Poisson model of distributions from several alternatives. We choose various sets of experiments in which the data are generated from mixture of a Poisson family of distributions.
We simulate two distributions Pois(5) and Pois(10), hence the data generating process has the density:

$$h(\pi) = (1 - \pi)\text{Pois}(5) + \pi\text{Pois}(10)$$

where $\pi$ is set to some specific value for each set of experiments. In each set of experiments, 100 random samples are drawn from this mixture of distributions, and then evaluate the fit of the Poisson family for each of the datasets using the method described in section 2.

We simulated 20,000 posterior distributions from the case where the alternatives correspond to a misspecified model based on mixture of distributions, and compute Hellinger distance for each. We consider the cases where $\pi = 0.2$, $\pi = 0.4$, $\pi = 0.6$, $\pi = 0.8$ and the limiting case $\pi = 1$.

The results of our five sets of experiments are presented in Figure 4.

The two first sets of experiments correspond approximatively to the null hypothesis of our proposed model. In last cases, data come exclusively from the alternative specified as the Poisson distribution Pois(10).

The posterior distributions for Hellinger distance does seem to converge to the correct value if the proposed model is correct. This result is theoretically demonstrated in Lemma 4.1.

The datasets in the last following examples are not simulated from a Poisson family distributions.

Here, the set of experiments corresponds to the Geometric distribution $G(p)$. We note that the posterior distributions of Hellinger distance, clearly, continue to be...
distinctly different from 0, when the probability of success in each trial takes different and increasing values \((p = 0.2\) and \(p = 0.8\)).

6. Conclusion and extension

In summary, by analogy with the classical notions, we have introduced the Hellinger distance on a countably infinite space consisting of all probability densities. We have investigated an fundamental property, in particular, the consistent estimate of \(\text{Inf}H(P, D_\theta)\). We have noted that the resulting tool can be considered as a method for evaluating goodness fit questions, in relation with Dirichlet Process, consisting to quantify the inaccuracy of a discrete distribution. LEaning on the MCMC algorithms, the posterior distributions of Hellinger distance are easy to implement. We have illustrated with some simulated data examples. Similar to the approach described by Verdinelli and Wasserman [15], the theorem
3.1 suggests that it would be useful to extend this method by using Bayes Factor for testing whether the Hellinger metric is less than any particular value.

References