

# Asymptotic Issues for Bayesian Histograms

## *Aspetti Asintotici di Istogrammi Bayesiani*

Catia Scricciolo<sup>1</sup>

Dipartimento di Scienze Statistiche  
Università degli Studi di Padova  
catia@stat.unipd.it

**Riassunto:** Il problema della stima di densità su  $[0, 1]$  viene considerato adottando un’impostazione bayesiana non parametrica. Per distribuzioni iniziali su misure di probabilità dominate, il funzionale media *a posteriori* fornisce una procedura di stima della densità. Indicazioni sul limite superiore della velocità di convergenza si possono trarre dallo studio della stessa per la distribuzione finale. Definita una misura iniziale che selezioni funzioni costanti a tratti con probabilità uno, si vede che, per densità campionarie dello stesso tipo, la distribuzione finale converge con velocità “quasi parametrica”  $n^{-1/2}(\log n)$ . Per densità più regolari, la convergenza è meno rapida e la velocità differisce al più per un fattore logaritmico da quella dello stimatore frequentista basato sul metodo dell’istogramma. Un miglioramento si consegue definendo una distribuzione iniziale sui poligoni.

**Keywords:** Bayesian histogram, frequency polygon, posterior rate of convergence.

## 1. Introduction

Given a random sample of  $n$  independent observations from an unknown probability distribution absolutely continuous with respect to (w.r.t.) Lebesgue measure  $\lambda$  on  $[0, 1]$ , we consider the problem of consistently estimating the distribution’s density function adopting a Bayesian approach. Let  $\mathcal{F}_{[0,1]} = \{P \in \mathcal{P}_{[0,1]} : P \ll \lambda\}$  be the class of all probability measures on  $[0, 1]$  that are absolutely continuous w.r.t.  $\lambda$ . A prior measure on  $\mathcal{F}_{[0,1]}$  with full weak topological support may be constructed exploiting the approximation property of the “theoretical histogram”, see Abou-Jaoude (1976) and Berlinet *et al.* (1998). Uncertainty in the density may be thus modeled. For some fixed integer  $b > 1$ , let  $\mathcal{K} = \{k = b^m, m \in \mathbb{N}\}$ . Let  $K$  be a random variable (r.v.) with probability mass function  $p(\cdot)$  on  $\mathcal{K}$ , determining the number of bins of equal length over  $[0, 1]$ , which are defined as  $A_{1,k} = [0, 1/k]$ ,  $A_{j,k} = ((j-1)/k, j/k]$ ,  $2 \leq j \leq k$ . Given  $K = k$ , let  $\mathbf{W}_k = (W_{1,k}, \dots, W_{k,k})$  be distributed over the  $(k-1)$ -dimensional simplex  $S_{k-1}$  in  $\mathbb{R}^k$ . For each  $n \in \mathbb{N}$ , given  $K = k$  and  $\mathbf{W}_k = \mathbf{w}_k$ , the r.v.’s  $X_1, \dots, X_n$  are distributed according to the piecewise constant density

$$h_k(x) = \sum_{j=1}^k w_{j,k} k \mathbb{I}_{A_{j,k}}(x), \quad x \in [0, 1],$$

where the  $w_{j,k}$ ’s play the role of the mixing weights for the densities  $k \mathbb{I}_{A_{j,k}}(\cdot)$ , with  $\mathbb{I}_{A_{j,k}}(\cdot)$  denoting the indicator function of the set  $A_{j,k}$ . This prior selects stepwise constant densities with probability one. Still, if for all  $k \in \mathcal{K}$ ,  $p(k) > 0$  and  $\mathbf{w}_k$  has full support on

---

<sup>1</sup>Address for correspondence: Catia Scricciolo, Dipartimento di Scienze Statistiche, Università degli Studi di Padova, Via C. Battisti 241/243, I-35121 Padova, Italy.

$S_{k-1}$ , then every absolutely continuous distribution on  $[0, 1]$  is in the weak support of the prior. This fact makes such a prior suitable for nonparametric density estimation and the pointwise posterior expectation of the random density, i.e.

$$\hat{f}_n(x) = \sum_{k \in \mathcal{K}} \left[ \sum_{j=1}^k \mathbb{E}[W_{j,k} | k, x_1, \dots, x_n] \mathbb{I}_{A_{j,k}}(x) \right] p(k | x_1, \dots, x_n), \quad x \in [0, 1],$$

which need not be a histogram, provides the Bayes' estimate under quadratic loss.

For the purpose of assessing rates of convergence, we assume that given  $K = k$ , the random vector  $\mathbf{W}_k \sim \text{Dir}(\alpha_{1,k}, \dots, \alpha_{k,k})$ , namely  $\mathbf{W}_k$  has a  $k$ -dimensional Dirichlet distribution with parameters  $\alpha_{1,k}, \dots, \alpha_{k,k}$  such that for some positive constant  $M$ ,  $\alpha_{j,k} \leq M$  for all  $j, k$ . This requirement is met by the following specifications:  $\alpha_{j,k} = M\alpha(A_{j,k})$ , with  $\alpha$  a probability measure on  $[0, 1]$ , or  $\alpha_{j,k} = 1$  for all  $j, k$ . The issue of the choice of the Dirichlet parameters is discussed in Petrone (1999). Let  $\Pi$  denote the prior distribution on  $\mathcal{K}_{[0,1]}$  induced by

$$\pi = \sum_{k \in \mathcal{K}} p(k) \pi_k = \sum_{k \in \mathcal{K}} p(k) \text{Dir}(\alpha_{1,k}, \dots, \alpha_{k,k}).$$

Strong consistency of the Bayes' estimator follows from consistency of the posterior distribution with respect to the Hellinger metric  $d_H(f, g) = \{\int (\sqrt{f} - \sqrt{g})^2\}^{1/2}$ , the  $L_2$ -distance between square rooted densities, which has been shown by Barron *et al.* (1999).

The focus here is on determining the rate of convergence of the Bayes' estimator. In fact, it suffices to study the convergence rate of the posterior distribution, that is the speed at which the posterior mass concentrates on shrinking Hellinger neighborhoods of the true density  $f_0$ . The entailed estimator converges at least as fast as the posterior distribution.

Applying the general theory developed by Ghosal *et al.* (2000) and Shen and Wasserman (2001), which provide sufficient conditions for assessing posterior rates, we show that if the sampling density is piecewise constant, then the posterior converges at a rate  $n^{-1/2}(\log n)$ . The convergence cannot be so fast in general. Under mild regularity conditions on  $f_0$ , the rate turns out to be equivalent up to a logarithmic factor to the rate  $n^{-1/3}$  of the frequentist histogram estimator for integrated mean squared error. Different aspects of Bayesian histograms have been studied by Andreev and Arjas (1996), Gasparini (1996) and Hartigan (1996), but to our knowledge, rates of convergence have not been investigated yet.

Results on histograms are a first step towards the study of mixture models based on more smoothed densities. The simplest histogram smoothing is obtained by joining the histogram ordinates at mid-bin points with straight lines. For any  $k \in \mathcal{K}$ , let  $t_{j,k} = j/k$ , for  $j = 0, \dots, k$ , and  $c_{j,k} = (t_{j-1,k} + t_{j,k})/2$ , for  $j = 1, \dots, k$ . The polygon-shaped density is defined as

$$p_k(x) = \begin{cases} h_k(c_{1,k}), & 0 \leq x \leq c_{1,k}, \\ k [h_k(c_{j,k})(c_{j+1,k} - x) + h_k(c_{j+1,k})(x - c_{j,k})], & c_{j,k} < x \leq c_{j+1,k}, \\ h_k(c_{k,k}), & c_{k,k} < x \leq 1. \end{cases}$$

Note that  $p_k(\cdot)$  differs from the frequency polygon in Scott (1985) because it equals the histogram ordinates in the intervals containing the endpoints 0 and 1. Unlike the histogram, the polygon is continuous and its posterior distribution converges at a faster rate.

The paper is thus organized. The main results on the rate of convergence of the posterior distribution for histograms and polygons are presented in Section 2. A few remarks are made in Section 3.

## 2. Main Results

The rate of convergence of the posterior distribution for histograms and polygons is determined. We briefly outline the main idea behind the results stated below whose proofs are herein omitted. We appeal to general theorems on posterior rates. To this aim, we compute the prior concentration rate on Kullback-Leibler type neighborhoods of  $f_0$ , which accounts for the approximation error of a density by histograms or polygons, and provide estimates for the Hellinger packing number of sieve sets  $\mathcal{F}_n \subseteq \mathcal{F}_{[0,1]}$ . The first result deals with the case when the sampling density is piecewise constant.

**Theorem 1** *Suppose  $f_0$  is a piecewise constant density, i.e.  $f_0 = \sum_{j=1}^{k_0} w_{j,k_0}^0 \mathbb{I}_{A_{j,k_0}}$  for some  $k_0 \in \mathbb{N}$  and  $\mathbf{w}_{k_0}^0 \in S_{k_0-1}$ . For constants  $B, \beta > 0$ , let  $0 < p(k) \leq Be^{-\beta k}$  for all  $k \in \mathcal{K}$ . Then, for a constant  $C$  large enough,*

$$\Pi \left( \{P : d_H(f_0, f_P) > Cn^{-1/2}(\log n)\} \mid X_1, \dots, X_n \right) \rightarrow 0$$

as  $n \rightarrow \infty$ , in  $P_0^n$ -probability.

The result is analogous to Theorem 5.1 in Ghosal and van der Vaart (2001) and Theorem 2.2 in Ghosal (2001): when the “true” density is of the same form as the statistical model, the “nearly parametric rate”  $n^{-1/2}(\log n)$  is obtained; otherwise, the rate is slower.

**Theorem 2** *Suppose  $f_0$  is a Lipschitz function bounded away from zero. For constants  $B_i, \beta_i > 0, i = 1, 2$ , let  $B_1 e^{-\beta_1 k} \leq p(k) \leq B_2 e^{-\beta_2 k}$  for all  $k \in \mathcal{K}$ . Then, for a constant  $C$  large enough,*

$$\Pi \left( \{P : d_H(f_0, f_P) > Cn^{-1/3}(\log n)^{5/6}\} \mid X_1, \dots, X_n \right) \rightarrow 0$$

as  $n \rightarrow \infty$ ,  $P_0^\infty$ -almost surely.

Except for the log-factor, the rate is equivalent to that at which the integrated mean squared error (IMSE) for the frequentist estimator defined by the relative frequency of the number of observations falling into each cell converges to zero. In fact, the IMSE is asymptotically  $O(n^{-2/3})$ , see Scott (1979).

In Theorem 2.3, Ghosal (2001) obtained the same rate  $n^{-1/3}(\log n)^{5/6}$  for the Bernstein polynomial prior but assuming that  $f_0$  is bounded away from zero and has bounded second derivative. Under substantially the same regularity conditions on  $f_0$ , the rate is improved in the next theorem by adopting a prior on the polygons, which also is denoted by  $\Pi$ . The rate obtained is equivalent up to a logarithmic factor to the rate of convergence  $n^{-3/4}$  in expected  $\chi^2$ -divergence found by Beirlant *et al.* (2002) for the analogous frequentist estimator. It is driven by the approximation error in  $\chi^2$ -divergence of the histogram part near the endpoints 0 and 1 which is only proportional to  $k^{-3}$  and less accurate than in the middle part.

**Theorem 3** *Suppose  $f_0$  is bounded away from zero and twice continuously differentiable with  $f_0'$  and  $f_0''$  uniformly bounded. For constants  $B_i, \beta_i > 0, i = 1, 2$ , let  $B_1 e^{-\beta_1 k} \leq p(k) \leq B_2 e^{-\beta_2 k}$  for all  $k \in \mathcal{K}$ . Then, for a sufficiently large constant  $C$ ,*

$$\Pi \left( \{P : d_H(f_0, f_P) > Cn^{-3/8}(\log n)^{7/8}\} \mid X_1, \dots, X_n \right) \rightarrow 0$$

as  $n \rightarrow \infty$ ,  $P_0^\infty$ -almost surely.

Computing Bayes' estimates from previous priors is impossible because of infinite sums. We may consider priors  $\pi_n = \sum_{r=1}^{k_n} p_n(r) \text{Dir}(\alpha_{1,r}^{(n)}, \dots, \alpha_{r,r}^{(n)})$  with  $k_n \rightarrow \infty$  as  $n \rightarrow \infty$ , and  $\sum_{r=1}^{k_n} p_n(r) = 1$  for all  $n$ , while  $\alpha_{j,r}^{(n)} \leq M$  for  $j = 1, \dots, r$  and sufficiently large  $n$  and  $r$ . Under the assumptions of Theorem 2 (Theorem 3), by taking  $c_1(n/\log n)^\gamma \leq k_n \leq c_2(n/\log n)^\gamma$  with  $\gamma = 1/3$  ( $\gamma = 1/4$ ), the rate is  $n^{-1/3}(\log n)^{1/3}$  ( $n^{-3/8}(\log n)^{3/8}$ ).

### 3. Final Remarks

Previous results provide only upper bounds on posterior rates of convergence. Nonetheless, the difference between  $n^{-1/3}(\log n)^{5/6}$  and  $n^{-3/8}(\log n)^{7/8}$  is an indication of a weaker rate, as if smoothing the histogram using the Bernstein polynomial approximation does not improve the rate as, instead, smoothing by means of polygons seems to do. But, it is still too early to draw general conclusions and deeper investigation is necessary.

### References

- Abou-Jaoude S. (1976) Conditions nécessaires et suffisantes de convergence L1 en probabilité de l'histogramme pour une densité, *Annales de l'Institut Henri Poincaré, Section B: Calcul des Probabilités et Statistique*, XII, 213–231.
- Andreev A. and Arjas E. (1996) A note on histogram approximation in Bayesian density estimation, in: *Bayesian Statistics 5*, Bernardo, J.M., Berger, J.O., Dawid, A.P. & Smith, A.F.M. (Eds.), Oxford University Press, 487–490.
- Barron A., Schervish M.J. and Wasserman L. (1999) The consistency of posterior distributions in nonparametric problems, *Annals of Statistics*, 27, 536–561.
- Beirlant J., Berline A., Biau G. and Vajda I. (2002) Divergence-type errors of smooth Barron-type density estimators, *Test*, 11, 191–217.
- Berline A., Vajda I. and van der Meulen E.C. (1998) About the asymptotic accuracy of Barron density estimates, *IEEE Transactions on Information Theory*, 44, 999–1009.
- Gasparini M. (1996) Bayesian density estimation via Dirichlet density processes, *Nonparametric Statistics*, 6, 355–366.
- Ghosal S. (2001) Convergence rates for density estimation with Bernstein polynomials, *Annals of Statistics*, 29, 1264–1280.
- Ghosal S., Ghosh J.K. and van der Vaart A.W. (2000) Convergence rates of posterior distributions, *Annals of Statistics*, 28, 500–531.
- Ghosal S. and van der Vaart A.W. (2001) Entropies and rates of convergence for maximum likelihood and Bayes estimation for mixtures of normal densities, *Annals of Statistics*, 29, 1233–1263.
- Hartigan J.A. (1996) Bayesian histograms, in: *Bayesian Statistics 5*, Bernardo, J.M., Berger, J.O., Dawid, A.P. & Smith, A.F.M. (Eds.), Oxford University Press, 211–222.
- Petrone S. (1999) Random Bernstein polynomials, *The Scandinavian Journal of Statistics*, 26, 373–393.
- Scott D.W. (1979) On optimal and data-based histograms, *Biometrika*, 66, 605–610.
- Scott D.W. (1985) Frequency polygons: theory and application, *Journal of the American Statistical Association*, 80, 348–354.
- Shen X. and Wasserman L. (2001) Rates of convergence of posterior distributions, *Annals of Statistics*, 29, 687–714.