# A Bayesian Density Estimation Algorithm *

Stefanos Manganaris †

*Dept. of Computer Science*
*Vanderbilt University*
*Box 1679, Station B*
*Nashville, TN 37235, U.S.A.*

March 13, 1996

**Abstract**

Density estimation is a central problem in data mining and knowledge discovery, with applications from data visualization and exploratory data analysis to supervised and unsupervised concept learning. This paper presents a simple nonparametric method for univariate density estimation that uses Bayesian inference and the minimum-message length principle to induce appropriate mixture models. Its advantage over existing methods is that it does not depend on ad hoc smoothing parameters or roughness penalties. The algorithm is efficient and robust.

**Keywords:** discovery of probability laws, nonparametric density estimation, Bayesian induction, minimum message-length principle, mixture models.

*Supported in part by a grant from NASA Ames (NAG 2–834; P.I.: Doug Fisher).
†Tel. (615)322–2796, (615)343–8006 (fax). E-mail: `stefanos@vuse.vanderbilt.edu`.

*Submitted to KDD '96*

# 1 Introduction

Given a set of data points forming a sample from an unknown probability distribution function, $f(x)$, the goal of density estimation is the induction of an estimate of $f(x)$. Parametric approaches make assumptions about the unknown distribution and reduce the problem to estimating the parameters of a model; they are usually deductive approaches. Nonparametric approaches make less rigid assumptions and are thus more flexible; they address problems of induction.

Density estimation methods are very useful in knowledge discovery and data mining. For *data visualization*, a plot of the estimated probability density function (pdf) summarizes the statistical behavior of the data. For *exploratory data analysis*, the induced density function reveals statistical peculiarities, such as, the number of modes and skewness characteristics. In *supervised concept learning*, the most accurate classification rule (i.e., the Bayes rule) can be expressed in terms of the probability density functions associated with the classes; density estimation methods can provide estimates for these pdfs. *Unsupervised methods* based on the premise that clusters correspond to modes in an underlying pdf can also rely on density estimates constructed from a sample. Nonparametric methods are of particular interest in all of these tasks, because they do not require a priori distributional assumptions.

Nonparametric density estimation has been studied extensively. Popular methods include the histogram, kernel estimation, nearest neighbor methods, and restricted maximum likelihood methods; good reviews are [Ize91] and [Sil86].

This work started with the observation that ideas from Bayesian model induction could be used for nonparametric density estimation. A Bayesian approach would be interesting, because it would produce estimates of appropriate complexity for the sample, addressing the problem of over- or under-smoothing. The works of Rissanen [Ris87, Ris78] and Wallace *et al.* [WF87, BW70] study the general problem of making inferences from random data using the minimum message-length principle. Cheesman in [Che90] discusses a general framework for model selection, using Bayesian inference and the minimum message-length (MML) principle to assign priors. Pednault in [Ped89] presents a method for inducing piecewise polynomial models in a similar framework. As it turned out, minimum complexity density estimation is not novel; a theoretical study by Barron and Cover [BC91] shows that the MML principle can be used to obtain density estimators with good asymptotic properties, in terms of convergence and consistency. This paper builds on that work and presents a practical nonparametric method that induces accurate univariate density estimates.

## 2 A Bayesian Density Estimation Algorithm

In the Bayesian framework, learning can be viewed as an optimization problem. In the context of density estimation, the goal is to find the most probable density function, given some constraints on its form and a set of independent sample values.

In this work, the unknown density function, $f(x)$, is approximated by a mixture distribution:

$$f(x) \approx \sum_{i=1}^{n} p_i \, g_i(x;\ \mu_i, \sigma_i^2), \tag{1}$$

where $n$ is the number of components in the mixture, $p_i$ is the weight of the $i$-th component, and $g_i(x;\ \mu_i, \sigma_i^2)$ is a Gaussian component with mean $\mu_i$ and variance $\sigma_i^2$. To ensure that the approximation is a bona fide density, it is required that $p_i > 0$ and $\sum_{i=1}^{n} p_i = 1$. The parameters of the mixture (i.e., the number of components, the mixture proportions, $p_i$, and the mean and variance of each component) are to be estimated from a sample of the unknown distribution. This representation for estimates was chosen based on theoretical results that show that any sufficiently smooth probability density function can be expressed as a mixture of Gaussians (occasionally in multiple ways) [Jay94] [DH73]. In other words, the chosen representation is expressive enough to approximate accurately most pdfs encountered in practice.

The proposed algorithm searches the space of possible mixture distributions for the one with highest posterior probability. The parameters of the most probable mixture are estimated by minimizing the expression

$$L(M) + \log_2 \left( \frac{1}{\prod_{i=1}^{n} P(v_i|M, I)} \right), \tag{2}$$

where $M$ stands for the mixture, $L(M)$ is the length of its encoding in bits, $v_i,\ i = 1, \ldots, n$, are independent, identically distributed, sample values, and $P(v_i|M, I)$ is the likelihood of $v_i$. Minimizing expression 2 is equivalent to maximizing the probability of the mixture, $P(M|\langle v_1, \ldots, v_n \rangle,\ I)$, (by Bayes' rule, noting that prior probabilities of mixtures are assigned via the minimum message-length principle: the negative logarithm, with base 2, of the a priori probability of a mixture density, $-\log_2 P(M|I)$, is estimated as the length in bits of the shortest possible encoding of $M$ in the framework of prior knowledge.)

The space of mixtures is searched efficiently with a dynamic programming algorithm, similar to the one used by Pednault in [Ped89] (Figure 1). Initially, the sample is sorted into an ordered sequence of sample values. In the first step of the algorithm

**Algorithm 1**

```
1  funct Induce-Mixture(sample) ≡
2      ⟨v_1, ..., v_n⟩ := sort(sample);
3      foreach (k, j) with k ≤ j do
4        mixture(k, j) := Induce-Gaussian(⟨v_k, ..., v_j⟩)
5      od
6      for i := 1 to n do
7          mixture(1, i) := mixture(1, m) + mixture(m + 1, i)
8          where m achieves the following maximization:
9            max_{1≤m<i} P(mixture(1, m) + mixture(m + 1, i)|⟨v_1, ..., v_i⟩, I)
11     od
12 end
```

Figure 1: The algorithm *Induce-Mixture* employs a dynamic-programming strategy to find the most probable mixture of Gaussians that approximates the unknown probability density function from which the sample was drawn.

(lines 3–5), the best Gaussian is induced for each subinterval of the range of possible values (complexity $O(n^2)$, where $n$ is the number of samples). In the second step (lines 6–11), the best mixture is found for successively larger prefixes of the sequence of sample values. Under the assumption that mixture components are well separated, in the sense that each component is associated with a region in which the presence of other components can be ignored, the most probable mixture for a prefix of the sequence can be found by considering the mixtures that are formed by adding (i) all the components of the best mixture for a shorter prefix and (ii) the Gaussian for the remaining subinterval to complete the prefix (proof by contradiction). The assumption essentially allows us to ignore the effects of the long tails of the mixture components outside their corresponding subinterval. The implications of the assumption in practice are examined later. The number of mixtures to consider at each step is equal to the length of the prefix, thus the total complexity of the second step is $O(n^2)$.

The simplest test is probably to apply the algorithm to a sample from a Gaussian distribution. Figure 2 shows the actual pdf (a standard normal, $N(0,1)$), a histogram of the frequencies in the sample, and the induced pdf for various sample sizes. Note,

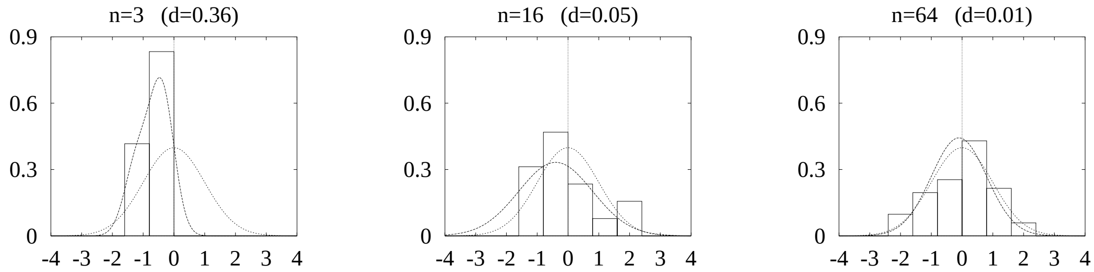


Figure 2: Bayesian estimates of a normal density.

as the size of the sample increases the estimate converges to the actual pdf. The next section examines the robustness of the algorithm in more complex cases.

## 3 Robustness Study

The second step of the dynamic programming algorithm was justified under the assumption that mixture components are well separated. In this experimental study, the goal is to determine the effects of violating this assumption in the accuracy of the induced pdf.

The separation between two components, $N(\mu_1, \sigma_1^2)$ and $N(\mu_2, \sigma_2^2)$, can be measured by the ratio

$$\mathcal{S} = \frac{\mu_2 - \mu_1}{\sigma_1 + \sigma_2}; \tag{3}$$

component separation is proportional to the difference of their means and inversely proportional to the sum of their standard deviations. A series of experiments was performed in which samples from bimodal distributions with various component-separation ratios were presented to the algorithm. The accuracy of the algorithm was evaluated by comparing the induced pdf to the actual pdf. Probability density functions were compared in terms of their squared Hellinger distance

$$d_H^2(p, g) = \int_{-\infty}^{+\infty} \left(\sqrt{p(x)} - \sqrt{g(x)}\right)^2 dx. \tag{4}$$

The squared Hellinger distance measures the discrepancies between two pdfs and takes on values between zero and two.

The probability density functions used in the experiments consisted of two components. The first component of the mixture was a standardized Gaussian, $N(0, 1)$. The

Table 1: Bayesian Estimation of Two-Component Mixtures

| $p_1$ | $p_2$ | $\sigma_2$ | $\mathcal{S}$ | $E(n)$ | $E(d_H^2)$ | $p_1$ | $p_2$ | $\sigma_2$ | $\mathcal{S}$ | $E(n)$ | $E(d_H^2)$ |
|---|---|---|---|---|---|---|---|---|---|---|---|
| 0.4 | 0.6 | 0.5 | 0.5 | 2.0 | 0.0082 | 0.6 | 0.4 | 0.5 | 0.5 | 1.7 | 0.0145 |
| | | | 1.0 | 2.3 | 0.0095 | | | | 1.0 | 2.0 | 0.0102 |
| | | | 1.5 | 2.0 | 0.0027 | | | | 1.5 | 2.0 | 0.0015 |
| | | 1.0 | 0.5 | 1.0 | 0.0005 | | | 1.0 | 0.5 | 1.3 | 0.0049 |
| | | | 1.0 | 1.0 | 0.0079 | | | | 1.0 | 1.0 | 0.0076 |
| | | | 1.5 | 2.0 | 0.0056 | | | | 1.5 | 2.0 | 0.0063 |
| | | 1.5 | 0.5 | 2.0 | 0.0052 | | | 1.5 | 0.5 | 1.7 | 0.0097 |
| | | | 1.0 | 1.0 | 0.0174 | | | | 1.0 | 2.7 | 0.0096 |
| | | | 1.5 | 2.0 | 0.0102 | | | | 1.5 | 2.3 | 0.0082 |
| 0.5 | 0.5 | 0.5 | 0.5 | 2.0 | 0.0058 | 0.8 | 0.2 | 0.5 | 0.5 | 1.0 | 0.0060 |
| | | | 1.0 | 2.7 | 0.0098 | | | | 1.0 | 1.0 | 0.0112 |
| | | | 1.5 | 2.0 | 0.0015 | | | | 1.5 | 2.0 | 0.0111 |
| | | 1.0 | 0.5 | 1.0 | 0.0020 | | | 1.0 | 0.5 | 1.0 | 0.0032 |
| | | | 1.0 | 1.0 | 0.0079 | | | | 1.0 | 1.7 | 0.0102 |
| | | | 1.5 | 2.7 | 0.0046 | | | | 1.5 | 2.7 | 0.0052 |
| | | 1.5 | 0.5 | 2.0 | 0.0080 | | | 1.5 | 0.5 | 2.3 | 0.0018 |
| | | | 1.0 | 1.3 | 0.0250 | | | | 1.0 | 2.0 | 0.0038 |
| | | | 1.5 | 2.0 | 0.0079 | | | | 1.5 | 2.3 | 0.0091 |

parameters of the second component, as well as the mixture proportions, were varied, and the resulting average Hellinger distances for various component-separation ratios are shown in Table 1 (representative cases). In addition to distance, the table includes the average number, $n$, of components in the induced mixtures. The sample size in all experiments was $2,048$. Samples of this size are not expected to differ significantly in their statistical characteristics, thus only three trials were run in each case.

The first observation is that the maximum $d_H^2$ encountered was 0.025, with 94% of the cases resulting, on average, a distance less than 0.015; this suggests that the induced densities are good approximations of the actual densities, even with small component-separation ratios. This can be explained by observing that two components with small separation can be approximated well with *one* appropriate compo-

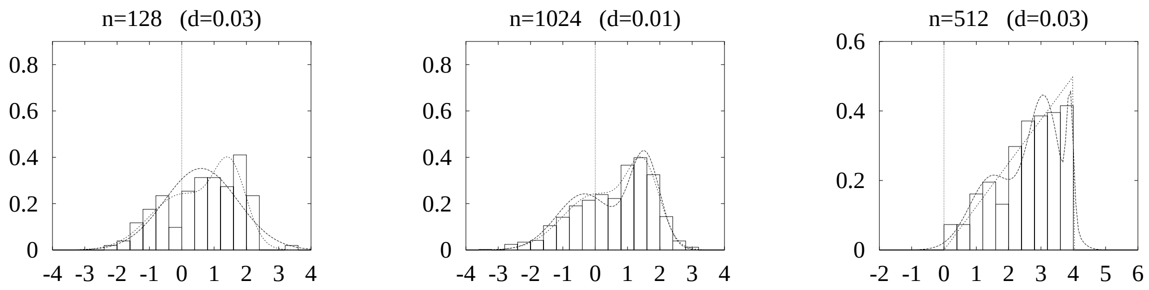


Figure 3: On the left, the underlying pdf is bimodal, yet the algorithm fails to identify two components from the given sample, because of their small separation. In the middle, the size of the sample was increased, resulting in the induction of two components. The actual density is a mixture of $N(0,1)$ and $N(1.5, 0.5^2)$ with weights 0.6 and 0.4 respectively (component-separation ratio is 1.0). The figure on the right is for a different underlying pdf, and shows that the algorithm is not limited to mixture pdfs; in this example, the underlying pdf is triangular, but is approximated by a Gaussian mixture. In all three cases, the squared Hellinger distances are small, suggesting good approximations.

nent. As the separation gets larger, the error increases, until, eventually, there is sufficient evidence to justify the addition of another component in the mixture, at which point the error decreases again. The data support this explanation: in the vast majority of cases, the $d_H^2$ associated with $\mathcal{S}$ equal to 0.5 or 1.5 was less than the $d_H^2$ when $\mathcal{S}$ was equal to 1.0 (all other parameters being equal), suggesting that the transition point, in most cases, was around $\mathcal{S} = 1.0$.

In most cases, the number of components in the mixture was one (when $\mathcal{S}$ was before the transition point) or two (when $\mathcal{S}$ was after the transition point). Occasionally, three components were induced, usually in cases with large separation ratios, where there is a greater potential for sample artifacts to justify an additional component.

## 4 Discussion

The proposed density estimation algorithm is based on Bayesian inference principles and previous results supporting the representation of arbitrary pdfs as mixtures of Gaussians. The implementation in this paper was based on the assumption that mixture components are well separated, but empirical evidence indicates that the induced pdfs are good approximations even when this assumption is violated.

Compared to kernel estimation, probably the most popular method, the Bayesian approach has the advantage that it produces curves as smooth as the evidence in the data suggests, without relying on external information. Kernel estimation depends critically on a smoothing parameter, for the determination of which there is as yet no universally accepted approach (some approaches are discussed in [Sil86]). On the other hand, this Bayesian algorithm is computationally more expensive: $O(n^2)$ vs. $O(n \log n)$ for an FFT implementation of kernel estimation ([Sil86], sec. 3.5). The nearest neighbor method also depends on a smoothing parameter. Histograms are sensitive to the choice of origin and bin width (see for example [Sil86], p. 11), and result in pdfs that also require smoothing. Restricted maximum likelihood methods are similar to Bayesian methods, in that they both treat density estimation as an optimization problem. Maximum likelihood methods seek to find the pdf with highest likelihood while imposing restrictions on either the class of densities or the form of the likelihood function. While in some cases the imposed restrictions can be interpreted as priors, yielding a Bayesian interpretation for these methods, there is usually no formal basis for deciding the form of these restrictions, in contrast to the priors assigned by an MML approach. For example, Good and Gaskins [GG71] use a roughness measure based on the square root of the pdf that penalizes its slope and/or curvature.

Derin [Der87] and Redner and Walker [RW84] address parametric versions of mixture density estimation; in their work, the number of components in the mixture is known a priori.

Probability densities are often multivariate. As the number of variables in the joint pdf increases, samples become sparser in the rapidly expanding multidimensional space, making accurate density estimation difficult. While Bayesian estimation principles can be used in higher-dimensional spaces, the particular dynamic programming algorithm used here does not scale well and is limited to univariate densities.

Density estimation is a central problem in data mining and knowledge discovery, with applications from data visualization and exploratory data analysis to supervised and unsupervised concept learning. A simple and efficient algorithm has been presented for nonparametric density estimation. Compared to other methods, its main advantage, stemming from the MML principle, is that it is *objective:* it does not depend on ad hoc smoothing parameters or roughness penalties.

## Acknowledgments

I thank Doug Fisher for his comments on drafts of this paper.

## References

[BC91] A. R. Barron and T. M. Cover. Minimum complexity density estimation. *IEEE Trans. on Information Theory*, 37(4), 1991.

[BW70] D. M. Boulton and C. S. Wallace. A program for numerical classification. *The Computer Journal*, 13(1):63–69, 1970.

[Che90] Peter Cheeseman. On finding the most probable model. In J. Shrager and P. Langley, editors, *Computational Models of Discovery and Theory Formation*, chapter 3, pages 73–95. Morgan Kaufmann, 1990.

[Der87] Haluk Derin. Estimating components of univariate Gaussian mixtures using Prony's method. *IEEE Trans. Pattern Analysis and Machine Intelligence*, 9(1):142–148, January 1987.

[DH73] Richard O. Duda and Peter E. Hart. *Pattern Classification and Scene Analysis*. John Wiley & Sons, 1973.

[GG71] I. J. Good and R. A. Gaskins. Nonparametric roughness penalties for probability densities. *Biometrica*, 58:255–277, 1971.

[Ize91] A.J. Izenman. Recent developments in nonparametric density estimation. *J. Amer. Statist. Assoc.*, 86(413):205–224, 1991.

[Jay94] E. T. Jaynes. Probability theory: The logic of science. Available at `http://omega.albany.edu:8008/JaynesBook.html`, 1994.

[Ped89] Edwin P.D. Pednault. Some experiments in applying inductive inference principles to surface reconstruction. In *Proc. of the Eleventh International Joint Conf. on Artificial Intelligence*, pages 1603–1609, 1989.

[Ris78] Jorma Rissanen. Modeling by shortest data description. *Automatica*, 14:445–471, 1978.

[Ris87] Jorma Rissanen. Stochastic complexity. *Journal of the Royal Statistical Society, Series B*, 49(3):223–239, 1987.

[RW84] Richard A. Redner and Homer F. Walker. Mixture densities, maximum likelihood, and the EM algorithm. *SIAM Review*, 26(2):195–239, April 1984.

[Sil86] B. Silverman. *Density Estimation for Statistics and Data Analysis*. Chapman and Hall, New York, 1986.

[WF87] C. S. Wallace and P. R. Freeman. Estimation and inference by compact coding. *Journal of the Royal Statistical Society, Series B*, 49(3):252–265, 1987.