# Independent Similarly Distributed Assumptions for Semiparametric Density Estimation

**Tony Jebara, Yingbo Song and Kapil Thadani**
Department of Computer Science, Columbia University
New York, NY 10027, USA
{jebara,yingbo,kapil}@cs.columbia.edu

## 1 Introduction

This article describes a technique for semiparametric density estimation which uses both parametric and non-parametric criteria. Parametric methods assume the functional form of the underlying distribution; these methods can underfit the data and may run into problems if the data is generated by a varying distribution or one that does not match the assumptions. In contrast, non-parametric approaches such as kernel density and Parzen estimation do not rely on parametric assumptions but may be too flexible and are prone to overfitting the data.

The proposed approach utilizes a continuous interpolation between the two extremes of independently distributed (*i.d.*) sampling assumptions and independently identically distributed (*i.i.d.*) sampling assumptions. This approach makes independent *similarly* distributed (*i.s.d.*) sampling assumptions on the data, using a scalar parameter $\lambda$ to trade off parametric and non-parametric properties in order to obtain a better density estimate. This technique is computationally efficient, unimodal and consistent over a wide range of models.

## 2 Approach

Given a dataset $\mathcal{X} = \mathbf{x_1}, \ldots, \mathbf{x_N}$, density estimation seeks to recover $p(\mathbf{x_1}, \ldots, \mathbf{x_N})$. Consider maximum a posteriori (MAP) estimation where parameters $\Theta = \{\theta_1, \ldots, \theta_N\}$ define the marginals. Assuming that the points are *i.d.* gives the joint likelihood as a product of independent singleton marginals $p(\mathbf{x_n}|\theta_\mathbf{n})$. Therefore, obtaining the parameters $\Theta$ for the MAP *i.d.* setup involves maximizing the posterior $p^{i.d.}(\mathcal{X}, \Theta) = \prod_{n=1}^{N} p(\mathbf{x_n}|\theta_\mathbf{n})\mathbf{p}(\theta_\mathbf{n})$. The singleton priors $p(\theta_n)$ can be assumed to be uniform in the case of maximum likelihood (ML) estimation. To obtain the parameters for the *i.i.d.* setup where the assumption is that the samples share the *same* singleton marginal, $p^{i.d.}$ can be maximized subject to constraints that the marginals must be equal.

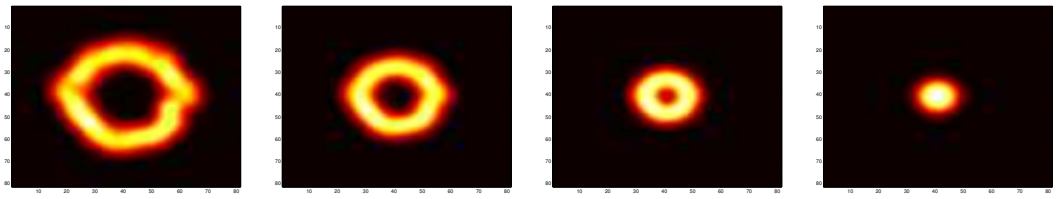
Instead of $N(N-1)/2$ pairwise equality constraints for the $i.i.d.$ setup, penalty functions can be applied across pairs of marginals that reduce the posterior score when they disagree, thereby encouraging cohesion between the models. The level of agreement between two marginals $p(\mathbf{x}|\theta_m)$ and $p(\mathbf{x}|\theta_n)$ can be measured using the Bhattacharya affinity metric between two distributions (Bhattacharyya, 1943) where $\mathcal{B}(p(\mathbf{x}|\theta_m), p(\mathbf{x}|\theta_n)) = \int p^\beta(\mathbf{x}|\theta_m)p^\beta(\mathbf{x}|\theta_n)d\mathbf{x}$. Using $\beta = 1/2$, the affinity is maximal and equal to 1 if and only if $p(\mathbf{x}|\theta_m) = p(\mathbf{x}|\theta_n)$. Bhattacharya affinity is preferred over alternative information divergences such as KL divergence because it has some useful computational and log-concavity properties and can be computed analytically for a wide range of models including hidden Markov models (Jebara *et al.*, 2004). This leads to the following formulation for the posterior score for independent *similarly* distributed (*i.s.d.*) data:

$$p_\lambda(\mathcal{X}, \Theta) \propto \prod_n p(\mathbf{x}_n|\theta_n)p(\theta_n) \prod_{m \neq n} \mathcal{B}^{\lambda/N}(p(\mathbf{x}|\theta_m), p(\mathbf{x}|\theta_n)) \tag{1}$$

The parameter $\lambda$ adjusts the importance given to the similarity between pairs of marginals. If $\lambda \to 0$, the marginals are unconstrained as in the *i.d.* setup. If $\lambda \to \infty$, the marginals are constrained to be equal as in the *i.i.d.* setup.

Since the log of the *i.s.d.* posterior described in Equation 1 is the sum of the prior distributions, the *i.d.* log-probabilities which are log-concave in the parameters and data, and the Bhattacharyya affinities which are log-concave in the parameters of *jointly* log-concave distributions (Prekopa, 1973), the *i.s.d.* log-posterior is also log-concave for jointly log-concave distributions and log-concave prior distributions. Most members of the exponential family of distributions, such as fixed-variance Gaussians, satisfy the above criteria and yield log-concave *i.s.d.* log-posteriors. Since this implies unimodality, the parameter estimation can be optimized using a variety of different iterative algorithms[1]. It can also be shown that the *i.s.d.* posterior with $\beta = 1/2$ is consistent in the case of Gaussian distributions although the proof is omitted here due to space constraints.

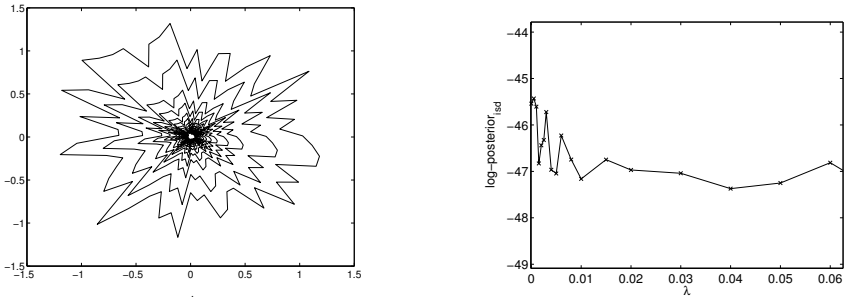
## 3 Experiments



(a) $\lambda$=0.05 (b) $\lambda$=0.95 (c) $\lambda$=2.90 (d) $\lambda$=7.40

Figure 1: Interpolating between *i.d.* and *i.i.d.* density estimation using the $\lambda$ parameter.

Figure 1 shows the interpolation that the *i.s.d.*-based method introduces between the Parzen density estimation technique and ML over a 2-dimensional noisy ring dataset by slowly varying $\lambda$.




(a) (b)

Figure 2: (a) Trace of the Gaussian mean positions as they interpolate between the *i.d.* and *i.i.d.* solutions on another instance of the noisy ring dataset (b) Log-posteriors observed for estimation under *i.s.d.* assumptions for HMMs over varying $\lambda$.

Local minima are possible when *i.s.d.* assumptions are used with models that aren't jointly log-concave, such as hidden Markov models (HMMs). Figure 2(b) shows log-posteriors when HMMs are used to model handwriting strokes under leave-one-out cross-validation with 10 folds. Although non-zero $\lambda$ values result in improved log-likelihoods, procedures for recovering the optimal $\lambda$ remain the subject of ongoing work.

## 4 References


Bhattacharyya, A. (1943). On a measure of divergence between two statistical populations defined by their probability distributions. Bull. Calcutta Math Soc.

Jebara, T., Kondor, R., & Howard, A. (2004). Probability product kernels. Journal of Machine Learning Research, 5, 819 844.

Prekopa, A. (1973). On logarithmic concave measures and functions. Acta. Sci. Math., 34, 335 343.


[1]The update rule for the Gaussian case with $\theta$ as the mean is: $\theta_n = \frac{N\mathbf{x}_n + \lambda/2 \sum_{m \neq n} \tilde{\theta}_m}{N + \lambda(N-1)/2}$