12-1-2014

Information-Theoretic Limits for Density Estimation

James Brofos
Dartmouth College

Dartmouth College Computer Science Technical Report TR2014-765

INFORMATION-THEORETIC LIMITS FOR DENSITY ESTIMATION

By James Brofos*

*Dartmouth College**

This paper is concerned with the information-theoretical limits of density estimation for Gaussian random variables with data drawn independently and with identical distributions. We apply Fano's inequality to the space of densities and an arbitrary estimator. We derive necessary conditions on the sample size for reliable density recovery and for reliable density estimation. These conditions are true simultaneously for both finitely and infinitely dimensional density spaces.

1. Introduction. Given a set of k normal densities $\{f_1, \ldots, f_k\}$ (assumed to be univariate), a single density is selected uniformly at random and n samples are generated from it. The density that was selected (and in particular its index in $\{1, \ldots, k\}$) is assumed to be unknown, though we are provided the samples. Denoting these n samples collectively as $\mathbf{X}^n$, the task is then to estimate, from $\mathbf{X}^n$, which of the k densities was responsible for generating the data. This problem of density recovery from data is one that appears ubiquitously in statistics and related fields. We also treat the problem of density estimation, which, rather than estimating the index of the underlying distribution, estimates the distribution itself.

In this paper, we consider the information-theoretic limits of the density recovery and density estimation problems for Gaussian random variables and samples generated from thm in an i.i.d. manner. We treat this problem very generally, and in fact our results hold for any algorithm and associated computational complexity. The key ingredient of this analysis is Fano's inequality, which has been at the source of many recent exciting developments in theoretical statistics and machine learning.

Throughout this paper we will assume that the distributions are identifiable in the sense that if f_i is parameterized by $\theta_i = \left(\mu_i, \sigma_i^2\right)$, then $\theta_i \neq \theta_j$ whenever $i \neq j$. It is understood that, throughout this paper, probabilities and expectations are computed with respect to the appropriate probability measure induced by a density.

2. Main Results. We establish necessary conditions on the sample size for both reliable density recovery and reliable density estimation associated

with an estimator. We define $\psi : \mathcal{X}^n \to \{1, \ldots, k\}$ to be an estimator of the density index. We begin by introducing two definitions for ψ that will be useful in our analysis.

DEFINITION 2.1 (δ-Suspect and δ-Inconsistent). We say that an estimator for the index of the true density ψ is δ-suspect if

$$\max_{i \in \{1,\ldots,k\}} \mathbb{P}\left[\psi\left(\mathbf{X}^n\right) \neq i\right] \geq \delta. \tag{2.1}$$

Furthermore, the density selected by ψ on the basis of n samples, denoted f_ψ, is said to be δ-inconsistent if

$$\max_{i \in \{1,\ldots,k\}} \mathbb{E}\left[\left|\left|f_\psi - f_i\right|\right|_1\right] \geq \delta. \tag{2.2}$$

The notation $||\cdot||_1$ denotes, as usual, the manhattan norm, which is for probability densities twice the total variation distance.

REMARK 2.1. Lower bounds in 2.1 on the error probability and in 2.2 on the total variation distance are useful primarily for establishing necessary conditions for any density index estimator ψ. In particular, one can establish an upper bound on the right-hand side of either 2.1 or 2.2, say $\delta < \delta'$, and deduce a lower bound on the sample size n that will depend on δ'. This lower bound can then be viewed as a necessary (although certainly not sufficient) condition for any estimator ψ to be reliable in terms of density recovery or density estimation.

We now present our main results as two theorems relating the sample size for density estimation problems for k normal densities with parameterizations $\{\theta_i\}_{i=1}^k$ to notions arising from information theory.

THEOREM 2.1 (Necessary Condition for Density Recovery). *Let Ξ be distributed uniformly at random from $\{1, \ldots, k\}$. Then if $\Xi = i$, we have that $\mathbf{X}_j^n \sim f_i$ for all $j \in \{1, \ldots, n\}$. If ψ is an estimator of Ξ from $\mathbf{X}^n$, then*

$$\max_i \ \mathbb{P}\left[\psi\left(\mathbf{X}^n\right) \neq i\right] \geq 1 - \frac{\frac{n}{k^2}\sum_{a=1}^k \sum_{b=1}^k \mathcal{D}_{KL}\left(f_a, f_b\right) + \log 2}{\log k} \tag{2.3}$$

$$\geq 1 - \frac{n \cdot \max_{a,b}\left\{\mathcal{D}_{KL}\left(f_a, f_b\right)\right\} + \log 2}{\log k}. \tag{2.4}$$

It follows that a necessary condition for ψ to not be a δ-suspect estimator we must have,

$$n > \frac{\log \frac{k}{2} - \delta \log k}{\max_{a,b}\left\{\log \frac{\sigma_b}{\sigma_a} + \frac{\sigma_a^2 + (\mu_a - \mu_b)^2}{2\sigma_b^2} - \frac{1}{2}\right\}}. \tag{2.5}$$

THEOREM 2.2 (Necessary Condition for Density Estimation). *Under the same assumptions as in Theorem 2.1, we have for any density estimator f_ψ that the worst case (with respect to $\Xi = i$), is lower bounded as*

$$\mathbb{E}\left[||f_\psi - f_i||_1\right] \geq \frac{\min_{a,b}\left\{2 - 2\sqrt{\frac{2\sigma_a\sigma_b}{\sigma_a^2+\sigma_b^2}}e^{\frac{-1}{4}\frac{(\mu_a-\mu_b)^2}{\sigma_a^2+\sigma_b^2}}\right\}}{2} \times \left(1 - \frac{n \cdot \max_{a,b}\left\{\mathcal{D}_{KL}\left(f_a, f_b\right)\right\} + \log 2}{\log k}\right). \tag{2.6}$$

Denote by α the quantity attaining the minimum in 2.6. Then a necessary condition for f_ψ to not be a δ-inconsistent estimator of the density is

$$n > \frac{(\alpha - 2\delta)\cdot \log k - \alpha \cdot \log 2}{\alpha \cdot \left(\max_{a,b}\left\{\log\frac{\sigma_b}{\sigma_a} + \frac{\sigma_a^2+(\mu_a-\mu_b)^2}{2\sigma_b^2} - \frac{1}{2}\right\}\right)}. \tag{2.7}$$

REMARK 2.2. It should be noted that this same approach can be applied to subsets of the k densities in the event that a more satisfying lower bound on n can be obtained in this fashion. Indeed, if $\mathbf{F}$ denotes a class of densities with a subclass of densities $\mathbf{F}'$, then clearly the number of samples required for reliable density recovery (or estimation) is the maximum of the lower bounds obtained by applying Theorem 2.1 (or Theorem 2.2) to $\mathbf{F}$ and $\mathbf{F}'$ individually.

EXAMPLE 2.1. Suppose we have a set of three normal densities with different means and variances $\{\mathcal{N}(0.89, 0.59), \mathcal{N}(0.90, 0.60), \mathcal{N}(0.91, 0.61)\}$. The bounds in Theorems 2.1 and 2.2 are most useful in "difficult" estimation problems, where the Kullback-Leibler divergence is small. In this example,

$$\max_{a,b}\left\{\log\frac{\sigma_b}{\sigma_a} + \frac{\sigma_a^2+(\mu_a-\mu_b)^2}{2\sigma_b^2} - \frac{1}{2}\right\} = 0.017 \tag{2.8}$$

$$\min_{a,b}\left\{2 - 2\sqrt{\frac{2\sigma_a\sigma_b}{\sigma_a^2+\sigma_b^2}}e^{\frac{-1}{4}\frac{(\mu_a-\mu_b)^2}{\sigma_a^2+\sigma_b^2}}\right\} = 7.55 \times 10^{-5}. \tag{2.9}$$

Therefore, the necessary condition from Theorem 2.1 is $n \geq 18$ for $\delta = 0.10$. Additionally, the necessary condition from Theorem 2.2 is $n \geq 22$ for $\delta = 1 \times 10^{-6}$.

3. Proof of the Main Results. Theorems 2.1 and 2.2 characterize the number of samples required for reliable density recovery and estimation using bounds on information-theoretic quantities. The remainder of the paper is devoted to proving these two theorems. We begin with a quantity well-understood in statistics and information theory that quantifies the similarity between two probability distributions.

DEFINITION 3.1 (Squared Hellinger Distance). The squared Hellinger distance between two probability measures π_1 and π_2 is given as

$$H^2(\pi_1, \pi_2) = \frac{1}{2} \int \left(\sqrt{\frac{d\alpha_1}{d\gamma}} - \sqrt{\frac{d\alpha_2}{d\gamma}} \right)^2 d\gamma. \tag{3.1}$$

The variable γ is a probability measure. In particular, if $\pi_1 \sim \mathcal{N}(\mu_1, \sigma_1^2)$ and $\pi_2 \sim \mathcal{N}(\mu_2, \sigma_2^2)$ then the squared Hellinger distance becomes

$$H^2(\pi_1, \pi_2) = 1 - \sqrt{\frac{2\sigma_1\sigma_2}{\sigma_1^2 + \sigma_2^2}} \exp\left\{ \frac{-1}{4} \cdot \frac{(\mu_1 - \mu_2)^2}{\sigma_1^2 + \sigma_2^2} \right\}. \tag{3.2}$$

REMARK 3.1. The squared Hellinger distance will prove to be an indispensable quantity for our application. It is a well-known result that the squared Hellinger distance is always less than or equal to the total variation distance for two probability measures. Since the total variation distance for normally distributed random variables does not have closed-form, the tractability of the the squared Hellinger distance makes it a natural choice in practical applications as well.

It is easy to see that since the total variation is exactly one-half the manhattan distance, we can introduce a factor of two into the squared Hellinger distance and still obtain a valid lower bound on the manhattan distance.

We have assumed that each of the k distributions are parameterized by vectors $\{\theta_i\}_{i=1}^k$. For our purposes it will suffice to lower bound the (non-trivial) total variation for all distributions using the squared Hellinger distance. We see that a global lower bound, and the corresponding indices in $\{1, \ldots, k\}$, is given as

$$(i^\star, j^\star) = \underset{i,j \in \{1,\ldots,k\}: i \neq j}{\arg\min} \; 1 - \sqrt{\frac{2\sigma_i\sigma_j}{\sigma_i^2 + \sigma_j^2}} \exp\left\{ \frac{-1}{4} \frac{(\mu_i - \mu_j)^2}{\sigma_i^2 + \sigma_j^2} \right\}. \tag{3.3}$$

It makes no difference as to how ties are arbitrated and we do not discuss the possibility further.

As it will turn out, we will also require an upper bound on the Kullback-Leibler divergence between all the k densities. In the case of Gaussian densities, the Kullback-Leibler has a simple closed-form. In particular, we have for Gaussian probability measures π_1 and π_2,

$$\mathcal{D}_{\mathrm{KL}}\left(\pi_1, \pi_2\right) = \log \frac{\sigma_2}{\sigma_1} + \frac{\sigma_1^2 + \left(\mu_1 - \mu_2\right)^2}{2\sigma_2^2} - \frac{1}{2}. \tag{3.4}$$

This quantity is easy to maximize as a function of the densities f_i and f_j (noting, of course, that due to the asymmetry of the Kullback-Leibler divergence, $\mathcal{D}_{\mathrm{KL}}\left(f_i, f_j\right) \neq \mathcal{D}_{\mathrm{KL}}\left(f_j, f_i\right)$). We denote the indices achieving this maximum as the *ordered* tuple (i°, j°).

THEOREM 3.1 (Fano's Inequality for Statistics). *Let Ξ be distributed uniformly at random in the set $\{1, \ldots, k\}$. Let a realization of Ξ be the index i. Then the density of $\mathbf{X}_j^n$ is f_i which has parameterization θ_i for all $j \in \{1, \ldots, n\}$. Then if $\psi\left(\mathbf{X}^n\right)$ is an estimator of Ξ based on the available data, we have*

$$\max_i \ \mathbb{P}\left[\psi\left(\mathbf{X}^n\right) \neq i\right] \ \geq \ 1 - \frac{\mathcal{I}\left(\Xi : \mathbf{X}^n\right) + \log 2}{\log k} \tag{3.5}$$

$$\geq \ 1 - \frac{n\beta + \log 2}{\log k}. \tag{3.6}$$

Here the parameter β is chosen such that $\forall\ i, j$ the Kullback-Leibler divergence for densities f_i and f_j is upper bounded like $\mathcal{D}_{KL}\left(f_i, f_j\right) \leq \beta$. It should be noted that the bound in 3.6 follows from standard entropy bounds for collections of random variables.

COROLLARY 3.1. *A more general, but markedly less popular, version of Fano's inequality emerges when one considers the total variation distance. The only additional requirement compared to the "vanilla" version in Theorem 3.1 is a lower bound on the (non-trivial) total variation distance across all density pairs. If we denote this lower bound by α the the general form of Fano's inequality becomes,*

$$\max_i \mathbb{E}\left[||f_\psi - f_i||_1\right] \geq \frac{\alpha}{2}\left(1 - \frac{n\beta + \log 2}{\log k}\right). \tag{3.7}$$

We use the notation f_ψ to indicate an density estimated by the estimator ϕ on the basis of n samples drawn in an i.i.d. fashion from f_i.

At this point the proof of Theorems 2.1 and 2.2 requires only a little algebra to derive.

PROOF. Notice from first principles that both 3.5 and 3.7 are decreasing functions of n. It should be apparent that both can be crushed to zero in the limit as $n \to \infty$. Plainly, if we set either of the right-hand side quantities in 3.5 or 3.7 to be less than any $\delta > 0$, then the bound can be satisfied by a particular choice of n. By rearranging we have for the density recovery case,

$$n > \frac{\log \frac{k}{2} - \delta \log k}{\beta}. \tag{3.8}$$

Similarly for the density estimation case,

$$n > \frac{(\alpha - 2\delta) \cdot \log k - \alpha \log 2}{\alpha\beta}. \tag{3.9}$$

The desired inequalities can be obtained by the substitutions,

$$\alpha = 2 - 2\sqrt{\frac{2\sigma_{i^\star}\sigma_{j^\star}}{\sigma_{i^\star}^2 + \sigma_{j^\star}^2}} \exp\left\{\frac{-1}{4} \cdot \frac{(\mu_{i^\star} - \mu_{j^\star})^2}{\sigma_{i^\star}^2 + \sigma_{j^\star}^2}\right\} \tag{3.10}$$

$$\beta = \log \frac{\sigma_{j^\circ}}{\sigma_{i^\circ}} + \frac{\sigma_{i^\circ}^2 + (\mu_{i^\circ} - \mu_{j^\circ})^2}{2\sigma_{j^\circ}^2} - \frac{1}{2}, \tag{3.11}$$

This gives the claimed necessary conditions on the sample size n. □

E-MAIL: james.a.brofos.15@dartmouth.edu
DECEMBER 10, 2014