Estimating Dimensions of Probability Distributions: Box-Counting and Local Dimension

Klaus Pötzelberger
WU, Institute for Statistics and Mathematics, Vienna, Austria
Klaus.Poetzelberger@wu.ac.at

Abstract

We present consistency results for estimators of the box-counting dimension of the support of probability distributions. The box-counting dimension of the support E is defined via the covering number, the cardinality of a cover of E consisting of cubes of fixed side-length. Accordingly the covering number of a sample allows the definition of an estimator of the box-counting dimension of E. Consistency results for arrays of probability distributions may be applied to the distributions of innovations of Itô processes and allow the construction of consistent estimators of the dimension of the factors, i.e. of the dimension of the Brownian motion driving the process. A first field of application is a nonparametric analysis of high-dimensional data, when structure in the data may be presumed in the sense that the support might be a lower dimensional subset. The second but related area is the analysis of random low-dimensional factors. For distributions with support of exact dimension, an alternative to the empirical box-counting dimension is an empirical estimator of the local dimension. We discuss properties and versions of these estimators.

Keywords: dimension estimator; box-counting dimension; local dimension; quantization dimension.

1. Introduction

The dimension of a probability distribution, i.e. the minimal dimension of a measurable set of full probability, is of interest in many fields of research such as statistics, pure and applied mathematics. For instance the dimension of the attractor of a dynamical system or more specifically, the dimension of the invariant distribution under an iterated function system, is an appealing topic. A second application of estimators of dimension is the analysis of high-dimensional data X where the stochastic properties of X are explained by a vector of factors W in the sense that for a sufficiently smooth (Lipschitz) mapping f, $X = f(W)$ and $s := \dim(W) << \dim(X) =: d$. The third potential application that we have in mind is the analysis of a diffusion process, or more generally, an Itô process (X_t), with $dX_t = \mu_t dt + \sigma_t dW_t$, with (W_t) being Brownian motion (W_t). Here the dimension of the driving Brownian motion has to be estimated. Again, $s := \dim(W_t) < \dim(X) =: d$.

Before defining estimators of "the" dimension of the probability distribution, let us briefly review concepts of dimensions. Concepts of dimensions of sets and probability distributions, frequently discussed in the scientific literature, include the box-counting dimension (Minkowski dimension), the packing dimension, the Hausdorff dimension and, for distributions, the quantization dimension. To define the box-counting dimension of a set $E \subseteq [0,1]^d$, let for $r > 0$, $N(E,r)$ denote the size of the smallest cover of E consisting of r-meshed cubes, i.e. of sets $C = [i_1 r, (i_1+1)r[\times \cdots \times [i_d r, (i_d+1)r[$ with $i_1, \ldots, i_d \in \mathbb{N}$. Then the box-counting dimension of E is

$$\dim_B(E) = \lim_{r \to 0+} \frac{\log N(E,r)}{\log 1/r}. \tag{1}$$

If $\log N(E,r)/\log 1/r$ does not converge, the limsup and the liminf define the upper and the lower box-counting dimension. In this paper we always assume that the box-counting dimension exists. For a Borel probability distribution on $\mathbb{R}^d$ we define

$$\dim_B(P) = \inf\{\dim_B(E) \mid E \text{ Borel and } P(E) = 1\}. \tag{2}$$

For properties of the box-counting dimension and definitions of the Hausfdorff dimension $\dim_{\mathcal{H}}$ and the packing dimension $\dim_{\mathcal{P}}$ see for instance Falconer (1997). For the quantization dimension $\dim_Q$ see Graf and

Luschgy (2000). Results on dimension estimation based on the quantization error are given in Pötzelberger (2000), (2012).
The definition of $\dim_B(P)$ is operational in the sense that if P is replaced by an empirical distribution $\hat{P}_n$ of a sample $X_1, \ldots, X_n \sim P$ then for sequences (r_n) with $r_n \downarrow 0$ suitably and given regularity conditions, then $\hat{s}_n \to s = \dim_B(P)$, where

$$\hat{s}_n = \frac{\log N(\{X_1, \ldots, X_n\}, r_n)}{\log 1/r_n}. \tag{3}$$

Both the Hausdorff and the packing dimension may be defined by using the concept of local, lower and upper, dimension. Let $B(x,r)$ denote the open ball with radius r centered at x. Then $\underline{\text{dimloc}}(P)(x)$ and $\overline{\text{dimloc}}(P)(x)$ are

$$\underline{\text{dimloc}}(P)(x) = \liminf_{r \to 0+} \frac{\log P(B(x,r))}{\log r}, \tag{4}$$

$$\overline{\text{dimloc}}(P)(x) = \limsup_{r \to 0+} \frac{\log P(B(x,r))}{\log r}. \tag{5}$$

P is of exact lower local dimension, if $\underline{\text{dimloc}}(P)$ is constant a.s. In this case $\underline{\text{dimloc}}(P) = \dim_{\mathcal{H}}(P)$. P is of exact upper local dimension, if $\overline{\text{dimloc}}(P)$ is constant a.s. In this case $\overline{\text{dimloc}}(P) = \dim_{\mathcal{P}}(P)$. In the applications of density estimation that we have in mind, P is of exact lower dimension which is additionally equal to $\dim_B(P)$. Therefore, the local dimension allows an alternative definition of an estimator of $\dim_B(P)$. $\text{dimloc}(P)$ denotes the common value of the upper and the lower local dimension, if they are equal.

2.1. Estimators: I.i.d. Case

We assume that there exist a Borel set $E \subseteq [0,1]^d$ with $P(E) = 1$, $s := \dim_B(P) = \dim_B(E)$. Furthermore, for all $x \in E$, $\text{dimloc}(P)(x) = s$. We define two estimators of s for i.i.d. samples $X_1, \ldots, X_n \sim P$. The first ist the estimator (3), which we denote by $\hat{s}_n^N$. The superscript N indicates that it is defined using the covering number. The second estimator is called $\hat{s}_n^{\text{loc, med}}$ and is defined as follows: For $r = r_n$ let

$$\hat{s}_n^{\text{loc, med}} = \text{median}\{\log \hat{P}_n(C)/\log r \mid C \text{ is } r - \text{meshed cube }\}. \tag{6}$$

For an alternative version of the local estimator a trimmed mean instead of the median to combine the local estimates to the global is chosen: Let mean_p^{t} denote the trimmed mean of a sequence, with trimming proportional to p both on the upper and on the lower tail. Then for $p < 1/4$

$$\hat{s}_n^{\text{loc, t}} = \text{mean}_p^{\text{t}}\{\log \hat{P}_n(C)/\log r \mid C \text{ is } r - \text{meshed cube }\}. \tag{7}$$

A proof of the consistency of the estimator $\hat{s}_n^N$ is given in Pötzelberger (2003): If P has bounded support and for $r = r_n$, ${}^n r^s \to \infty$, then $\hat{s}_n^N$ is weakly consistent. If $\liminf nr^s/\log n > 1$, then $\hat{s}_n^N \to s$ is strongly consistent. The consistency of the local estimators $\hat{s}_n^{\text{loc, med}}$ and $\hat{s}_n^{\text{loc, t}}$ can be shown to hold under the same assumptions, provided P is of exact dimension s.

2.2. Estimators: Itô Case

Pötzelberger (2003) introduces an estimator for the dimension s of the driving Brownian motion (W_t) of an Itô process $(X_t)_{0 \le t \le T}$. Let

$$X_t = X_0 + \int_0^t \mu_u du + \int_0^t \sigma_u dW_u,$$

with processes (μ_t) and (σ_t) satisfying the assumptions given in Pötzelberger (2003). Consider the high-frequency asymptotics: T is fixed, (X_t) is sampled on $0 < t_1 < \cdots < t_n = T$, with, for simplicity, $t_i = Ti/n$. The regularity conditions imply that the process may be approximated by a process with piecewise constant (in t) diffusion parameter $\tilde{\sigma}_t$.

Let $m < n$, divide $[0,T]$ into m subintervals of equal length. Let $R_i = (X_{t_i} - X_{t_{i-1}})/\sqrt{n}$ be the scaled innovations and $Y_i = F(R_i)$, with F a smooth transformation $F : \mathbb{R}^d \to [0,1]^d$. Again partition $Y_1, \ldots, Y_n$ into m sets of size n_i according to the partition $[0,T]$ into its m subintervals,

$$\{Y_1, \ldots, Y_n\} = \{Y_1, \ldots, Y_{n_1}\} \cup \{Y_{n_1+1}, \ldots, Y_{n_1+n_2}\} \cup \cdots \cup \{Y_{n-n_m+1}, \ldots, Y_n\}. \tag{8}$$

For each subset $\{Y_{n_1+\cdots+n_{i-1}+1}, \ldots, Y_{n_1+\cdots+n_i}\}$ let $\hat{N}_{n,i}$ denote the covering number and finally, let $\hat{N}_n$ be their mean and $\hat{s}_n^N = \log \hat{N}_n / \log 1/r$.
$\hat{s}_n^N$ is consistent, if for $r = r_n$, $m = m_n$,

$$\lim_{n\to\infty} \frac{\log n}{r^2 m} = 0, \tag{9}$$

$$\lim_{n\to\infty} \frac{m}{n r^s} = 0. \tag{10}$$

(9) and (10) hold if, for instance, $r = n^{-\alpha}$ and $m = \lfloor n^\beta \rfloor$ with $0 < \alpha < 1/(2+s)$ and $2\alpha < \beta < 1 - s\alpha$.
To define a "local dimension" version, consider again the partition (8) and let $\hat{s}_{n,i}^{\text{loc}}$ denote the estimator (6) or (7) for $\{Y_{n_1+\cdots+n_{i-1}+1}, \ldots, Y_{n_1+\cdots+n_i}\}$. Combine the estimates $\hat{s}_{n,i}^{\text{loc}}, \ldots, \hat{s}_{n,m}^{\text{loc}}$ by taking either the median (version 1) or a trimmed mean (version 2).

3. Discussion

The consistency of the estimators holds, at least in the i.i.d. case, if $nr^s \to \infty$. This is a minimal assumption and implies that the absolute frequency of data in each cube goes to infinity. However, whether the estimate deviates significantly from the dimension s is still a hard problem to decide, it depends not only on the sample-size. The estimators are biased, with a considerable, even huge bias in certain cases. The choice of $r = r_n$ is crucial. Only if information about the bias is available or an upper bound for s is given, a reasonable estimation is possible.
When s is known to be an integer, the estimator is rounded to the positive integers. Instead of the mean squared error the probability/proportion of false classification (PFC) is the criterion for the performance of the estimator. Consider the following simple example. Let $d = 20, s = 6$. $Z_1, \ldots, Z_s$ are independent and uniformly distributed on $[0,1]$ and $Z_{s+1} = \cdots = Z_d = Z_1$. Furthermore, $X = f(Z)$. Properties of the estimators depend crucially on f, even if f is smooth.
Let $X_i = Z_i^p$ with $p \geq 1$. In this case rather small sample-sizes n lead to exact estimates. For instance, if $n = 2^9$, then a simulation with 1000 repetitions gave proportions of false classification of 0 for $\hat{s}_n^N$, 0.2% for $\hat{s}_n^{\text{loc, med}}$ and 1.7% for $\hat{s}_n^{\text{loc, t}}$.
If $X_i = (Z_i^2 + \log(1+Z_i))(1+\sin(2\pi Z_i))$, the situation changes dramatically. The estimator $\hat{s}_n^N$ gives proportions of false classification up to 100% for sample sizes up to 2^{14}. $\hat{s}_n^{\text{loc, med}}$ does much better, but is always beaten by $\hat{s}_n^{\text{loc, t}}$. The following table gives the PFC in some cases (again 1000 repetitions).

n	$\hat{s}_n^{\text{loc, med}}$	$\hat{s}_n^{\text{loc, t}}$
3000	23.8%	1.6%
3500	15.8%	1.7%
4500	8.3%	1.5%

References

Graf S. & Luschgy H. (2000). *Foundations of Quantization for Probability Distributions.* Lecture Notes in Mathematics 1730. Springer, Berlin.

Falconer K. J. (1997). *Techniques in Fractal Geometry.* John Wiley.

Pötzelberger K. (2000). The consistency of the empirical quantization error. *Mathematical Methods of Statistics 9* (2000), 199-207.

Pötzelberger K. (2003). Estimating the dimension of factors of diffusion processes. *Statistics and Decisions 21*, 171 - 184.

Pötzelberger K. (2012). Consistency of the empirical quantization error. *AIP Conference Proceedings, vol. 1479*, 435-437.