Available online at www.sciencedirect.com
ScienceDirect
Computational Statistics & Data Analysis 51 (2007) 4814 – 4831
COMPUTATIONAL STATISTICS & DATA ANALYSIS
www.elsevier.com/locate/csda

The random projection method in goodness of fit for functional data

J.A. Cuesta-Albertos[a,∗,1], E. del Barrio[b,1], R. Fraiman[c], C. Matrán[b,1]

[a]*Departamento de Matemáticas, Estadística y Computación, Universidad de Cantabria, Avda. los Castros s.n., 39005 Santander, Spain*
[b]*Departamento de Estadística e Investigación Operativa, Universidad de Valladolid, Spain*
[c]*Departamento de Matemáticas, Universidad de San Andrés, Argentina and Centro de Matemática, Universidad de la República, Uruguay*

Available online 10 October 2006

Abstract

The possibility of considering random projections to identify probability distributions belonging to parametric families is explored. The results are based on considerations involving invariance properties of the family of distributions as well as on the random way of choosing the projections. In particular, it is shown that if a one-dimensional (suitably) randomly chosen projection is Gaussian, then the distribution is Gaussian. In order to show the applicability of the methodology some goodness-of-fit tests based on these ideas are designed. These tests are computationally feasible through the bootstrap setup, even in the functional framework. Simulations providing power comparisons of these projections-based tests with other available tests of normality, as well as to test the Black–Scholes model for a stochastic process are presented.

Keywords: Random projections; Goodness-of-fit tests; Families of distributions; Gaussian distributions; Black–Scholes; Stochastic processes

1. Introduction

The use of projections of very-high-dimensional objects on randomly chosen subspaces is getting an increasing interest as a powerful tool in several applications of Mathematics. For instance, this idea is employed to obtain approximate algorithms in problems of high computational complexity (see, e.g., Vempala, 2004) or, even, randomly chosen projections are starting to be employed as a tool to detect copyright violations on images posted in the Internet (see Lejsek et al., 2005).

Although in such those applications Statistics is at the basis of the procedure, a genuinely statistical analysis of the possibilities of the idea (including the design of methods based on it and a theoretical exploration of their power) has not been carried out yet. We are aware of some results in which random projections have been used to estimate mixtures of distributions (see Gldófalvi, 2002; Vempala and Wang, 2004), but even these papers have not been written from a purely statistical point of view but rather from learning theory.

∗ Corresponding author. Tel.: +34 942201428; fax: +34 942201402.
E-mail address: cuestaj@unican.es (J.A. Cuesta-Albertos).

[1] Research partially supported by the Spanish Ministerio de Educación y Ciencia, grant MTM2005-08519-C02-01 and 02 and by the Consejería de Educación y Cultura de la Junta de Castilla y León, grant PAPIJCL VA074/03.

0167-9473/$ - see front matter
doi:10.1016/j.csda.2006.09.007

A first look at the problem with statistical motivation was Cuesta-Albertos et al. (2006a). To analyze to what extent random projections characterize a probability distribution, the closed cone

$$\mathbb{E}(P, Q) := \{h \in \mathbb{H} : P_h = Q_h\} \tag{1}$$

associated to two Borel distributions, P and Q, on a Hilbert space $\mathbb{H}$ is considered there. In (1), P_h denotes the marginal distribution of P along the h direction, i.e., the law of $\langle X, h\rangle$ if X is an $\mathbb{H}$-valued random element with law P and $\langle \cdot, \cdot \rangle$ is the scalar product in $\mathbb{H}$. In the sequel we will often refer to this distribution as the projection of P on h. With a similar meaning we will denote by P_V the projection of P on a linear subspace V.

A slight modification of one of the main results in Cuesta-Albertos et al. (2006a) gives the following theorem (see Theorem 4.1 in Cuesta-Albertos et al., 2006a). It involves the property of a probability measure, P, being determined by its moments (in the sequel $P \in DM(\mathbb{H})$). A discussion of this property (including sufficient conditions like the so-called *Carleman condition*) can be found in Shohat and Tamarkin (1943) (see also Section 8.4 in Chow and Teicher, 1997). A straightforward result is that an affine transformation of a probability measure determined by its moments is also determined by its moments.

Theorem 1.1 (*Cuesta–Fraiman–Ransford*). *Let $P \in DM(\mathbb{H})$ and Q be Borel probability measures on a separable Hilbert space, $\mathbb{H}$. Let μ be a non-degenerate Gaussian distribution on $\mathbb{H}$. Then $P = Q$ if and only if $\mu[\mathbb{E}(P, Q)] > 0$.*

This result was employed in Cuesta-Albertos et al. (2006b) to obtain some consistent goodness-of-fit test to a fixed distribution.

We emphasize that Theorem 1.1 goes, somehow, in a direction opposite to that of older results in Diaconis and Freedman (1984) where (p. 793) it is claimed that "*For many data sets*, *we show that most projections are nearly the same and approximately Gaussian*". Below we include further comments on this point.

Our goal, in this paper, is to generalize Theorem 1.1 in some senses, to employ these results to obtain goodness-of-fit tests for families of distributions and to make a preliminary study to explore the possibilities of the technique. In particular, taking into account the above quotation from Diaconis and Freedman (1984), we are particularly interested in seeing how this procedure works when applied to the design of goodness-of-fit tests for Gaussianity.

The proposed generalizations can be summarized as follows. First, in Section 2, we present a class of probability measures which can replace the Gaussian measure, μ, in Theorem 1.1. The possibility of choosing suitable (non-Gaussian) measures to increase the power against particular alternatives is still the object of current research.

In Section 3, the distribution P in Theorem 1.1 is replaced by a family of distributions. We consider two different cases. In the simplest situation one single random projection suffices to determine whether a distribution belongs to the family. This is the case for *invariant families* in the sense given in Definition 3.1. In particular, Theorem 3.6 states that if a randomly chosen projection of a distribution is Gaussian, then the distribution is Gaussian.

We want to stress the interest of this result. Let us assume that we are interested in knowing whether a given (multivariate) distribution, P, is Gaussian. Projection pursuit techniques to reject this hypothesis are based on the fact that, if P is not Gaussian, then not every one-dimensional projection is Gaussian. However, since most projections of P are approximately Gaussian, an extensive search is required to find out one of the scarce directions in which the projection of P is clearly not Gaussian. On the other hand, according to Theorem 3.6, this search is not required because, if P is not Gaussian almost every projection of P is not Gaussian. Some simulations which are reported in Section 5.1 support this last point of view.

Section 3.2 focuses on non-invariant families. In this case, one projection is not sufficient to determine inclusion in the family. We show, though, that for a location-scale model with k-dimensional parameter, $(k+1)$ projections suffice. Notice that we are not considering here a projection on a $(k+1)$-dimensional subspace, but $(k+1)$ one-dimensional projections. Of course, it would suffice to handle the $(k+1)$-dimensional projection, but we want to remark that applying, for instance, a $(k+1)$-dimensional Kolmogorov–Smirnov test is, by far, more time consuming than taking the maximum of $(k+1)$ one-dimensional Kolmogorov–Smirnov tests.

Our results are applied in Section 4 to obtain goodness-of-fit tests to families of distributions. Unfortunately, when more than one projection is considered, the goodness-of-fit test is often not distribution free; this is the case for the Kolmogorov–Smirnov statistic that we consider in Section 5. We propose to apply bootstrap to estimate the distribution