Some Applications of Univariate Nonlinear Principal Components

Alcune applicazioni delle componenti principali nonlineari univariate

Aldo Goia, Ernesto Salinelli
Dipartimento SEMeQ, Università del Piemonte Orientale “A. Avogadro”, Novara
e-mail: aldo.goia@eco.unipmn.it

Keywords: dependence structure, Poincaré inequality, principal components, Sarmanov family, splines estimates, goodness-of-fit test

1. Definition and properties

Let \( \Upsilon (D) \) be the set of absolutely continuous random variables (r.v.) \( X \) with zero mean, finite variance \( \mathbb{E} [X^2] \) and density \( f_X \) having support on the closure of the interval \( D = (a, b) \), \(-\infty \leq a < b \leq +\infty \). Denoting by \( \dot{L}^2_{f_X} \) the weighted Sobolev space of transformations \( u : D \to \mathbb{R} \) belonging to

\[
\dot{L}^2_{f_X} = \{ u : \mathbb{E}[u(X)] = 0, \mathbb{E}[(u(X))^2] < +\infty \}
\]

with first derivative square integrable (in a weak sense), we introduce the following:

Definition 1. The r.v. \( Z_j = \varphi_j(X) \) is the \( j \)-th nonlinear principal component of \( X \in \Upsilon (D) \) if \( \varphi_j \) is a solution of the maximum problem:

\[
\begin{align*}
\max_{u \in \dot{L}^2_{f_X}} & \quad \mathbb{E}[u(X)^2] \\
\text{subject to} & \quad \mathbb{E}[u(X)\varphi_s(X)] = 0, \quad s = 1, 2, \ldots, j-1, \quad j > 1.
\end{align*}
\]

Under the technical assumption \( (1/f_X) \in L^1_{\text{loc}} (D) \), and having set, for any \( u \in \dot{L}^2_{f_X} \),

\[
\mathcal{F}[u] = \mathbb{E}[u(X)^2] \quad \text{and} \quad \mathcal{Q}[u] = \mathbb{E}[(u'(X))^2],
\]

we show that solving problem (1) corresponds to maximize the reciprocal of the so-called Rayleigh quotient \( \mathcal{Q}[u]/\mathcal{F}[u] \):

\[
\mathcal{F}[\varphi_j] = \lambda_j = \max \left\{ \frac{\mathcal{F}[u]}{\mathcal{Q}[u]} : u \in \dot{L}^2_{f_X}, \mathbb{E}[u(X)\varphi_h(X)] = 0, \ 1 \leq h \leq j-1 \right\}.
\]

This is equivalent to find the \( j \)-th eigenfunction \( \varphi_j \), with corresponding eigenvalue \( \lambda_j \), of the covariance operator \( G \) associated to \( X \), or determining the \( j \)-th solution of a (second order) differential boundary value problem. We point out some results on the existence of transformations \( \varphi_j \), also in terms of the moment generating function of \( X_j \). Referring to classical results on self-adjoint, compact operators on Hilbert spaces, we show that, under suitable assumptions, \( G \) admits countably many real eigenvalues \( \lambda_1 > \lambda_2 > \lambda_3 > \cdots > \lambda_j > \cdots \), which are positive, simple, and they have 0 as unique limit point; the corresponding eigenfunctions \( \varphi_j \) are mutually orthogonal and form a complete set in \( \dot{L}^2_{f_X} \). Furthermore, we prove that the first transformation \( \varphi_1 \) is strictly monotone, it is odd if \( f_X \) is even, and it characterizes the distribution of \( X \). We relate our results with the literature concerning the so-called Chernoff inequality, giving also some explicit examples of analytic computation of the \( \varphi_j \)'s.
2. Estimation procedure

Let us consider a sample \( \{ X_i, i = 1, 2, \ldots, n \} \) of i.i.d. r.v.s drawn from \( X \). Denoting by \( S_{k,d} \) the \( k + d \) dimensional linear space of spline functions of order \( d \) defined on \( D \), and by \( \hat{W}_{1}^{2} \) the space of spline functions \( u \in S_{k,d} \) such that \( \frac{1}{n} \sum_{i=1}^{n} u(X_i) = 0 \), a.s., and setting \( F_n[u] = \frac{1}{n} \sum_{i=1}^{n} (u(X_i))^2 \) and \( Q_n[u] = \frac{1}{n} \sum_{i=1}^{n} (u'(X_i))^2 \), we define the estimator \( \hat{\lambda}_{j,k,n} \) of \( \lambda_j \) as

\[
\hat{\lambda}_{j,k,n} = \max \left\{ \frac{F_n[u]}{Q_n[u]} : u \in \hat{W}_{1}^{2}, \frac{1}{n} \sum_{i=1}^{n} u(X_i) \hat{\varphi}_{h,k,n}(X_i) = 0 \text{ a.s., } 1 \leq h \leq j - 1 \right\},
\]

and the estimator of \( \varphi_j \) as the maximizer \( \hat{\varphi}_{j,k,n} \in \hat{W}_{1}^{2} \) corresponding to \( \hat{\lambda}_{j,k,n} \). Problem (2) can be converted into a generalized eigenvalue problem and then solved by using some available computer packages. Under suitable hypotheses we also obtain some asymptotic results. The performances of the estimator are also analyzed by a simulation study.

3. Statistical applications

We discuss several possible statistical applications of NLPCs. First we use the characterization properties of the variance of the first NLPC transformation to define some goodness-of-fit test with a statistic based on \( \hat{\lambda}_{1,k,n} \). In particular we focus on three examples: testing the hypothesis that a distribution is Uniform, testing that a density is Wigner against the hypothesis that it is an other unimodal symmetric one, and testing normality. The power of the tests is analyzed via simulation for various sample sizes and for many alternatives, and they are compared with those of the Kolmogorv-Smirnov test.

The second application is based on the fact that the set of NLPC transformations represent a “special” basis: this enables us to analyze the dependence structure between two r.v.s by estimating suitable indexes of correlation and of dependence. Finally, we show that the recalled property of NLPC transformations allows us to define some classes of bivariate distributions with fixed marginals: this idea is illustrated by obtaining some densities in the class of the Sarmanov-Lee family, which is a generalization of the popular Farlie-Gumbel-Morgenstern family.

References


