Journal of Statistical Planning and Inference 167 (2015) 102–122

Contents lists available at ScienceDirect

Journal of Statistical Planning and Inference

journal homepage: www.elsevier.com/locate/jspi

Optimal estimators of principal points for minimizing expected mean squared distance

Shun Matsuura [a,*], Hiroshi Kurata [b], Thaddeus Tarpey [c]

[a] Faculty of Science and Technology, Keio University, Yokohama, Kanagawa, Japan

[b] Graduate School of Arts and Sciences, The University of Tokyo, Tokyo, Japan

[c] Department of Mathematics and Statistics, Wright State University, Dayton, OH, USA

ARTICLE INFO

Article history:
Received 22 October 2014
Received in revised form 29 April 2015
Accepted 21 May 2015
Available online 29 May 2015

Keywords:
Elliptical distributions
k-means clustering
Location-scale family
Normal distribution
Principal curves and surfaces
Self-consistency
t-distribution

ABSTRACT

k-Principal points of a random variable are *k* points that minimize the mean squared distance (MSD) between the random variable and the nearest of the *k* points. This paper focuses on finding optimal estimators of principal points in terms of the expected mean squared distance (EMSD) between the random variable and the nearest principal point estimator. These estimators are compared with nonparametric and maximum likelihood estimators. It turns out that a minimum EMSD estimator of *k*-principal points of univariate normal distributions is determined by the *k*-principal points of the *t*-distribution with $n+1$ degrees of freedom, where *n* is the sample size. Extensions of the results to location-scale families, multivariate distributions, and principal surfaces are also discussed.

1. Introduction

1.1. Principal points

A set of *k-principal points* (Flury, 1990) of a *p*-dimensional distribution is a set of *k* points that optimally approximates or summarizes the distribution in terms of squared error loss. More precisely, let $\boldsymbol{X}$ be a *p*-dimensional random vector whose components have finite second moments. Then, *k* points $\boldsymbol{\gamma}_1^*, \ldots, \boldsymbol{\gamma}_k^* \in \Re^p$ are called *k*-principal points of $\boldsymbol{X}$ if the *k* points satisfy

$$E\left[\min_{j=1,\ldots,k} \|\boldsymbol{X} - \boldsymbol{\gamma}_j^*\|^2\right] \leq E\left[\min_{j=1,\ldots,k} \|\boldsymbol{X} - \boldsymbol{\gamma}_j\|^2\right]$$

for any *k* points $\boldsymbol{\gamma}_1, \ldots, \boldsymbol{\gamma}_k \in \Re^p$, where $\Re^p$ denotes the *p*-dimensional Euclidean space. Let

$$MSD(\boldsymbol{\gamma}_1, \ldots, \boldsymbol{\gamma}_k) = E\left[\min_{j=1,\ldots,k} \|\boldsymbol{X} - \boldsymbol{\gamma}_j\|^2\right], \tag{1}$$

* Corresponding author.
E-mail addresses: matsuura@ae.keio.ac.jp (S. Matsuura), kurata@waka.c.u-tokyo.ac.jp (H. Kurata), thaddeus.tarpey@wright.edu (T. Tarpey).

http://dx.doi.org/10.1016/j.jspi.2015.05.005
0378-3758/

Table 1
k-Principal points of $N(0, 1)$ for $k = 1, \ldots, 10$.

k	δ_1^*	δ_2^*	δ_3^*	δ_4^*	δ_5^*	δ_6^*	δ_7^*	δ_8^*	δ_9^*	δ_{10}^*	MSD
1	0										1
2	$-\sqrt{2/\pi}$	$\sqrt{2/\pi}$									$1-2/\pi$
3	−1.2240	0	1.2240								0.1902
4	−1.5104	−0.4528	0.4528	1.5104							0.1175
5	−1.7241	−0.7646	0	0.7646	1.7241						0.0799
6	−1.8936	−1.0001	−0.3177	0.3177	1.0001	1.8936					0.0580
7	−2.0334	−1.1881	−0.5606	0	0.5606	1.1881	2.0334				0.0440
8	−2.1519	−1.3439	−0.7560	−0.2451	0.2451	0.7560	1.3439	2.1519			0.0345
9	−2.2547	−1.4764	−0.9188	−0.4436	0	0.4436	0.9188	1.4764	2.2547		0.0279
10	−2.3451	−1.5913	−1.0578	−0.6099	−0.1996	0.1996	0.6099	1.0578	1.5913	2.3451	0.0229

which is called the mean squared distance (MSD). k-Principal points are defined as k points that minimize $MSD(\boldsymbol{\gamma}_1, \ldots, \boldsymbol{\gamma}_k)$. We notice that if $p = 1$, then the MSD (1) can be re-expressed as

$$E\left[\min_{j=1,\ldots,k}(X-\gamma_j)^2\right],$$

where X is a univariate random variable and $\gamma_1, \ldots, \gamma_k \in \Re$. This paper mainly focuses on principal points of univariate distributions ($p = 1$), although extensions to multivariate distributions ($p \geq 2$) are also discussed.

The mean of a distribution is the optimal (in terms of squared error loss) single point approximation to a probability distribution. k-Principal points can be viewed as a generalization of the mean from one point to k points and in fact, the 1-principal point of a random vector $\boldsymbol{X}$ is always given by the mean $E[\boldsymbol{X}]$. The k-principal points $\gamma_1^*, \ldots, \gamma_k^*$ of a univariate normal distribution $N(\mu, \sigma^2)$ can be expressed in the form

$$\gamma_j^* = \mu + \delta_j^*\sigma, \quad j = 1, \ldots, k, \tag{2}$$

where $\delta_1^*, \ldots, \delta_k^*$ are the k-principal points of the univariate standard normal distribution $N(0, 1)$. The corresponding MSD can be obtained by multiplying the MSD for k-principal points of the $N(0, 1)$ distribution by σ^2. For reference, the k-principal points $\delta_1^*, \ldots, \delta_k^*$ of the univariate standard normal distribution $N(0, 1)$ and their MSDs for $k = 1$ to 10 are shown in Table 1. The $k = 2$-principal points of $N(0, 1)$ are $\pm\sqrt{\frac{2}{\pi}}$; for $k > 2$, the principal points must be determined numerically. (Note that Table 1 reproduces results previously published, for instance, in Table 1 for $k = 1, \ldots, 5$, in Flury (1990) and Table 5.1 for $k = 1, \ldots, 8$, in Graf and Luschgy (2000).) From Table 1, we see that the k-principal points are symmetric about the mean, and dense near the mean but are sparse near the tails which mirrors the normal density function that is concentrated symmetrically about the mean with low probability in the tails. Table 2 shows the k-principal points of the t-distribution with ν degrees of freedom for $\nu = 3, 4, 6, 11$ and $k = 1, 2, 3, 5, 10$. (Note that the second moment of the t-distribution on $\nu = 1$ and 2 degrees of freedom is not defined and hence neither are the principal points for $\nu = 1, 2$.) We note in Section 2.1 that the k-principal points of a normal distribution are unique for all values of k and the k-principal points of a t-distribution (with degrees of freedom ≥ 3) are also unique for $k = 1$ and 2. However, the uniqueness of $k > 2$ principal points of t-distributions has not been proven. Hence, the results of Table 2 are not ensured to give actual k-principal points with $k \geq 3$, but other sets of k points giving smaller MSD than the k points in Table 2 have not been found. The results of Tables 1 and 2 indicate that the k-principal points of the t-distributions are more spread than those of $N(0, 1)$ and the principal points of the t-distribution with smaller degrees of freedom are more spread than the k-principal points of the t-distribution with larger degrees of freedom, which reflects the fact that the t-distributions have heavier tails than $N(0, 1)$ and the t-distributions with smaller degrees of freedom have heavier tails than the t-distributions with larger degrees of freedom.

1.2. Connections to self-consistency, optimal partitioning, k-means clustering, and vector quantization

For a random vector $\boldsymbol{X}$, and a set of k points $\boldsymbol{\gamma}_1, \ldots, \boldsymbol{\gamma}_k$, the $MSD(\boldsymbol{\gamma}_1, \ldots, \boldsymbol{\gamma}_k)$ (1) can be re-expressed as

$$E\left[\sum_{j=1}^{k}\|\boldsymbol{X}-\boldsymbol{\gamma}_j\|^2 I(\boldsymbol{X}\in C_j)\right],$$

where $I(\cdot)$ denotes the indicator function and

$$C_j = \{\boldsymbol{x}\in\Re^p \mid \|\boldsymbol{x}-\boldsymbol{\gamma}_j\| < \|\boldsymbol{x}-\boldsymbol{\gamma}_l\|,\ l=1,\ldots,j-1,\ \|\boldsymbol{x}-\boldsymbol{\gamma}_j\| \leq \|\boldsymbol{x}-\boldsymbol{\gamma}_l\|,\ l=j+1,\ldots,k\}, \quad j=1,\ldots,k. \tag{3}$$

Principal points are closely related to the notion of *self-consistency* (Tarpey and Flury, 1996). A set of k points $\boldsymbol{\gamma}_1, \ldots, \boldsymbol{\gamma}_k$ is called a set of *k-self-consistent points* of a random vector $\boldsymbol{X}$ if the following equations are satisfied:

$$E[\boldsymbol{X}|\boldsymbol{X}\in C_j] = \boldsymbol{\gamma}_j, \quad j = 1, \ldots, k.$$