RAIRO Operations Research
RAIRO Oper. Res. **35** (2001) 315-328

FINDING THE PRINCIPAL POINTS OF A RANDOM VARIABLE*

Emilio Carrizosa[1], E. Conde[1], A. Castaño[2] and D. Romero–Morales[1,3]

Communicated by Erol Gelenbe

Abstract. The p-principal points of a random variable X with finite second moment are those p points in $\mathbb{R}$ minimizing the expected squared distance from X to the closest point. Although the determination of principal points involves in general the resolution of a multiextremal optimization problem, existing procedures in the literature provide just a local optimum. In this paper we show that standard Global Optimization techniques can be applied.

Keywords: Principal points, d.c. functions, branch and bound.

1. Introduction

Given a random variable X with finite second moment, consider the function $\Phi : \mathbb{R}^p \longrightarrow \mathbb{R}$,

$$\Phi(c_1, \ldots, c_p) = E\left[\min_{1 \leq i \leq p} (X - c_i)^2\right].$$

Received December, 1999. Accepted April, 2001.

* *The research of the two first authors has been supported by Grant PB96-1416-C02-02 of DGES, Spain.*

[1] Facultad de Matemáticas, Universidad de Sevilla, C/ Tarfia s/n, 41012 Sevilla, Spain.

[2] Departamento de Matemáticas, E.U. Empresariales, Universidad de Cádiz, C/ Por Vera, N. 54, Jerez de la Frontera, Cádiz, Spain.

[3] Faculty of Economics and Business Administration, Maastricht University, P.O. Box 616, 6200 MD Maastricht, The Netherlands.

A p-uple $c^* = (c_1^*, c_2^*, \ldots, c_p^*)$ such that

$$\Phi(c_1^*, \ldots, c_p^*) \leq \Phi(c_1, \ldots, c_p) \qquad \forall (c_1, \ldots, c_p) \in \mathbb{R}^p \tag{1}$$

is said to be a set of *p-principal points* of X [3,4,9,11,12,14–17].

Observe that, for $p = 1$, Φ becomes the classical expected squared distance, thus the unique 1-principal point is the mean of X. This shows that the concept of p-principal point constitutes a generalization to p points of the mean, thus representing a natural way to partition a population into p clusters (according to the attraction regions), see [10].

The literature addressing computational aspects of the problem of determination of principal points is rather scarce, and mainly limited to the statement of sufficient conditions under which the nonlinear equation $\nabla\Phi(c) = 0$ has as unique solution the principal points of X [10,13,17].

This requires assumptions as strong as existence and symmetry of the density function of X, conditions which are unlikely to hold just when the use of principal points is most natural, namely, when X is a mixture of p populations [10]. However, as shown in this paper no assumptions on X other than the existence of its second moment are required in order to solve the problem by standard global optimization procedures.

The rest of the paper is structured as follows. In Section 2 we state some general properties on the function Φ. In Section 3 we show how to construct a bounded polyhedron in $\mathbb{R}^p$ which is known to contain an ε-optimal solution. This polyhedron can be used as starting region for a Branch and Bound procedure. Some conclusions are given in Section 4 to end the paper.

2. General properties

Finding p-principal points amounts to finding an optimal solution to the optimization problem (PP),

$$\inf_{c \in \mathbb{R}^p} \Phi(c). \tag{PP}$$

In this section we address the important, though non-trivial question of existence of principal points, *i.e.*, the attainment of the infimum of (PP). Since the usual sufficient conditions for existence of optimal solutions, namely the compactness of the level sets $\{c \in \mathbb{R}^p : \Phi(c) \leq \alpha\}$, do not hold, an ad-hoc analysis is needed. To do that we start showing that Φ can easily be expressed as a d.c. function (difference of convex functions).

Property 2.1. *Let $f : \mathbb{R}^p \longrightarrow \mathbb{R}$ be the convex quadratic function defined as follows*

$$f(c_1, \ldots, c_p) = \sum_{i=1}^{p} c_i^2. \tag{2}$$

Then $f - \Phi$ is a convex function and $\Phi = f - (f - \Phi)$ defines a d.c. decomposition of Φ.

Proof. Since

$$\sum_{i=1}^{p}(x - c_i)^2 - \min_{1\le i\le p}(x - c_i)^2 = \max_{1\le i\le p}\sum_{j\neq i}(x - c_j)^2,$$

it follows that

$$f(c_1, \ldots, c_p) - \Phi(c_1, \ldots, c_p) = \int \max_{1\le i\le p}\sum_{j\neq i}(x - c_j)^2 \, \mathrm{d}F(x)$$

which is a convex function. Thus, the results holds by observing that f is a convex function. □

For any $\alpha, \beta \in \mathbb{R} \cup \{\pm\infty\}$, $\alpha \le \beta$, let $S_{[\alpha,\beta]}$ denote the polyhedron

$$S_{[\alpha,\beta]} = \{(c_1, \ldots, c_p) \in \mathbb{R}^p : \alpha \le c_1 \le c_2 \le \ldots \le c_p \le \beta\}.$$

Observe that $S_{[\alpha,\beta]}$ is bounded iff $\alpha > -\infty$ and $\beta < +\infty$. Moreover, when $-\infty < \alpha < \beta < +\infty$, $S_{[\alpha,\beta]}$ is a simplex which, rewritten in terms of its extreme points, leads to the expression

$$S_{[\alpha,\beta]} = \left\{\sum_{i=0}^{p} \lambda_i v^i : \lambda_i \ge 0 \, \forall i, \sum_{i=0}^{p} \lambda_i = 1\right\}, \tag{3}$$

where

$$\begin{aligned} v^0 &= (\alpha, \alpha, \ldots, \alpha, \alpha) \\ v^1 &= (\alpha, \alpha, \ldots, \alpha, \beta) \\ \vdots \quad &\vdots \qquad\qquad \vdots \\ v^p &= (\beta, \beta, \ldots, \beta, \beta). \end{aligned}$$

Due to the symmetry of Φ in its arguments, it follows:

Proposition 2.1. *One has:*

$$\inf_{c\in\mathbb{R}^p} \Phi(c) = \inf_{c\in S_{[-\infty,+\infty]}} \Phi(c).$$

This result can be strengthened when X has compact support. Indeed, one has:

Proposition 2.2. *If X has compact support $[m, M]$, then $S_{[m,M]}$ contains a set of p-principal points.*

Proof. For any $c = (c_1, c_2, \dots, c_p) \in S_{[-\infty,+\infty]}$, define $c^* = (c_1^*, \dots, c_p^*) \in S_{[m,M]}$ as

$$c_i^* = \begin{cases} m, & \text{if } c_i < m \\ M, & \text{if } c_i > M \\ c_i, & \text{else.} \end{cases}$$

By construction one has

$$(c_i - x)^2 \geq (c_i^* - x)^2 \ \forall x \in [m, M], \forall i = 1, 2, \dots, p$$

thus

$$\min_{1 \leq i \leq p} (c_i - x)^2 \geq \min_{1 \leq i \leq p} (c_i^* - x)^2 \ \forall x \in [m, M].$$

Since X has zero mass outside $[m, M]$, one has

$$\Phi(c) = \int_{[m,M]} \min_{1 \leq i \leq p} (c_i - x)^2 \, \mathrm{d}F(x) \geq \int_{[m,M]} \min_{1 \leq i \leq p} (c_i^* - x)^2 \, \mathrm{d}F(x) = \Phi(c^*).$$

This shows that

$$\inf_{c \in S_{[-\infty,+\infty]}} \Phi(c) = \inf_{c \in S_{[m,M]}} \Phi(c). \tag{4}$$

But $S_{[m,M]}$ is compact, and, by Property 2.1, Φ continuous on compact sets, thus there exists some $\bar{c} \in S_{[m,M]}$ such that

$$\Phi(\bar{c}) \leq \Phi(c) \ \forall c \in S_{[m,M]},$$

then, by Proposition 2.1 and (4),

$$\Phi(\bar{c}) \leq \Phi(c) \ \forall c \in \mathbb{R}^p.$$

□

The following example illustrates a simple case of compact support of X and shows how local search methods might yield suboptimal solutions.

Example 2.1. *Let X be the discrete random variable with support at points x_i and probability mass p_i, $i = 1, 2, \dots, 9$, given in Table 1. Figure 1 depicts a detail of the level sets and graph of the corresponding Φ for two principal points, clearly showing the multiextremal character of Φ.*

Following Proposition 2.2 the set $S_{[0,20]}$ contains a set of principal points and consequently, could be used as starting region of a global optimization method. Instead of this, we build a direct AMPL code [5] for solving (PP), taking as starting point in the local-search procedure a random pair in $(0, 20) \times (0, 20)$, as described in the Appendix (Tab. 2).

TABLE 1. X with mass at 9 points.

i	1	2	3	4	5	6	7	8	9
x_i	0	2	3	8	10	15	16	18	20
p_i	$\frac{1}{15}$	$\frac{1}{15}$	$\frac{1}{15}$	$\frac{4}{15}$	$\frac{4}{15}$	$\frac{1}{15}$	$\frac{1}{15}$	$\frac{1}{15}$	$\frac{1}{15}$

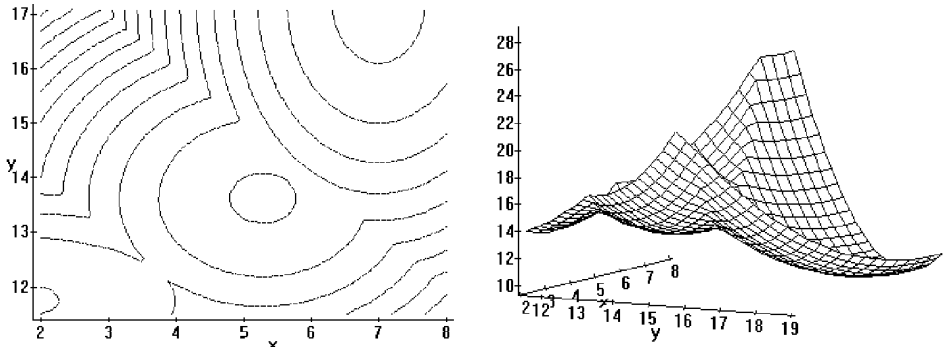

FIGURE 1. Level curves and graph of Φ.

The solution proposed by AMPL as optimal is $(9.733, 9.733)$ (which is not even a local minimum!) with $\Phi(9.733, 9.733) = 30.196$. However, it is easily checked that the 2-principal points for this problem are $(7, 17.25)$, with $\Phi(7, 17.25) = 9.65$.

A Branch and Bound algorithm was implemented using a *bisection subdivision* [7, 8], and replacing (2), the convex component of the objective function, by its linear minorant, as bounding scheme, [8]. It required 280 iterations to detect that an ε-optimal solution ($\varepsilon = 0.0001$) had been found: $(6.99707, 17.25098)$, with $\Phi(6.99707, 17.25098) = 9.650006552$.

3. Existence and localization of ε-optimal solutions

It has been shown in Section 2 that (PP) admits an optimal solution for random variables with bounded support. This result can be extended for unbounded supports. First, a technical lemma is given.

Lemma 3.1. *One has:*

$$\lim_{N\to+\infty} \int_{(-N,N)} \min\{(x+N)^2, (x-N)^2\}\, \mathrm{d}F(x) \; = +\infty. \tag{5}$$

Proof.

$$\begin{aligned}\int_{(-N,N)} \min\{(x+N)^2,(x-N)^2\}\,\mathrm{d}F(x) &\geq \int_{(-\frac{N}{2},\frac{N}{2})} \min\{(x+N)^2,(x-N)^2\}\,\mathrm{d}F(x)\\ &\geq \int_{(-\frac{N}{2},\frac{N}{2})} (N/2)^2\,\mathrm{d}F(x)\\ &= N^2/4\,P[X\in(-N/2,N/2)],\end{aligned}$$

thus

$$\begin{aligned}&\lim_{N\to+\infty}\int_{(-N,N)} \min\{(x+N)^2,(x-N)^2\}\,\mathrm{d}F(x)\\ &\geq \lim_{N\to+\infty} N^2/4\,P[X\in(-N/2,N/2)] = +\infty.\end{aligned}$$

□

Proposition 3.1. *There exists a set of p-principal points.*

Proof. If X has bounded support, the result follows from Proposition 2.2, so that we assume that X has unbounded support.

Since Φ is bounded below

$$\exists \inf_{c\in S_{[-\infty,+\infty]}} \Phi(c) = I \in [0,+\infty). \tag{6}$$

To show that an optimal solution exists, suppose that, on the contrary, such infimum is never attained. Hence,

$$I < \Phi(c) \quad \forall c \in S_{[-\infty,+\infty]}. \tag{7}$$

Hence, by (6) and (7), there exists some unbounded sequence $\{c^n\}_n \subset S_{[-\infty,+\infty]}$ such that $\Phi(c^n)$ converges to I. Let $c_0^n = -\infty$ and $c_{p+1}^n = +\infty$.

By construction of $S_{[-\infty,+\infty]}$, there exist i_0, i_1, with $0 \leq i_0 < i_1 \leq p+1$ such that

$$\begin{array}{ll} c_i^n \longmapsto -\infty & \forall i \leq i_0\\ c_i^n \longmapsto +\infty & \forall i \geq i_1\\ \{c_i^n\}_n \text{ is bounded} & \forall i,\ i_0+1 \leq i \leq i_1-1.\end{array}$$

Moreover, since for each $i \in \{i_0+1,\ldots,i_1-1\}$ the sequence $\{c_i^n\}_n$ is bounded, it has some convergent subsequence. Without loss of generality we assume that each $\{c_i^n\}_n$ $(i = i_0+1,\ldots,i_1-1)$ is convergent.

Two cases may then happen,

1. $\{c^n\}_n$ has at least one component bounded, *i.e.*,

$$i_0 < i_1 - 1; \tag{8}$$

2. all the components of $\{c^n\}_n$ diverge, *i.e.*,

$$i_0 = i_1 - 1. \tag{9}$$

We consider separately the two cases above. Suppose first that (8) holds. Then, for any n one has

$$\begin{aligned}\Phi(c^n) &\geq \sum_{j=1}^{p} \int_{(\frac{c^n_{j-1}+c^n_j}{2}, \frac{c^n_j+c^n_{j+1}}{2}]} (x - c^n_j)^2 \, \mathrm{d}F(x) \\ &\geq \sum_{j=i_0+1}^{i_1-1} \int_{(\frac{c^n_{j-1}+c^n_j}{2}, \frac{c^n_j+c^n_{j+1}}{2}]} (x - c^n_j)^2 \, \mathrm{d}F(x).\end{aligned}$$

Denote by $c^* \in \mathbb{R}^p$ the vector with components

$$c^*_i = \begin{cases} \lim_{n\to+\infty} c^n_{i_0+1}, & \text{if } i \leq i_0 + 1 \\ \lim_{n\to+\infty} c^n_{i_1-1}, & \text{if } i \geq i_1 - 1 \\ \lim_{n\to+\infty} c^n_i, & \text{else.} \end{cases}$$

It then follows that

$$\begin{aligned} I &= \lim_{n\to+\infty} \Phi(c^n) \geq \lim_{n\to+\infty} \sum_{j=i_0+1}^{i_1-1} \int_{(\frac{c^n_{j-1}+c^n_j}{2}, \frac{c^n_j+c^n_{j+1}}{2}]} (x - c^n_j)^2 \, \mathrm{d}F(x) \\ &\geq \Phi(c^*), \end{aligned}$$

which contradicts (7). Hence, the result holds.

Suppose now that (9) holds, thus any component of $\{c^n\}_n$ diverges. Hence, by (5), for any $M > 0$, there exists $N > 0$ such that

$$\int_{(-N,N)} \min\{(x+N)^2, (x-N)^2\} \mathrm{d}F(x) > M.$$

Moreover, given $N > 0$, there exists $n_0 \in \mathbb{N}$ such that, for any $n \geq n_0$,

$$\begin{aligned} c^n_i &< -N \quad \forall i \leq i_0 \\ c^n_i &> N \quad \forall i \geq i_0 + 1. \end{aligned}$$

Hence,

$$\begin{aligned}\Phi(c^n) &\geq \int_{(-N,N)} \min_{1\leq i\leq p} (x - c^n_i)^2 \mathrm{d}F(x) \\ &\geq \int_{(-N,N)} \min\{(x+N)^2, (x-N)^2\} \mathrm{d}F(x) > M,\end{aligned}$$

thus

$$I = \lim_{n \to +\infty} \Phi(c^n) = +\infty,$$

which contradicts (6). Hence, the result holds. □

For the case in which X has zero mass outside an interval $[m, M]$, we have shown in Proposition 2.2 that the search of an optimal solution can be reduced to the bounded polyhedron $S_{[m,M]}$. Now we address the general case and show how to explicitly construct a bounded polyhedron in $\mathbb{R}^p$ of the form $S_{[L,U]}$ that contains an ε-optimal solution c_ε to (PP).

The following result is easily shown:

Lemma 3.2. *One has:*

- *the function* $R \longmapsto \int_{[R,+\infty)} (x-R)^2 \mathrm{d}F(x)$ *is nonincreasing, and*

$$\lim_{R \to +\infty} \int_{[R,+\infty)} (x-R)^2 \mathrm{d}F(x) = 0; \tag{10}$$

- *the function* $R \longmapsto \int_{(-\infty,R]} (x-R)^2 \mathrm{d}F(x)$ *is nondecreasing, and*

$$\lim_{R \to -\infty} \int_{(-\infty,R]} (x-R)^2 \mathrm{d}F(x) = 0. \tag{11}$$

Lemma 3.2 implies that, for any $\varepsilon_1, \varepsilon_2 > 0$ one can construct real constants $L \leq U$ such that

$$\int_{(-\infty,L]} (x-L)^2 \mathrm{d}F(x) \quad \leq \quad \varepsilon_1 \tag{12}$$

$$\int_{[U,+\infty)} (x-U)^2 \mathrm{d}F(x) \quad \leq \quad \varepsilon_2. \tag{13}$$

Proposition 3.2. *Given* $L \leq U$ *verifying (12)-(13), there exists* $\bar{c} \in S_{[L,U]}$ *which is an* $(\varepsilon_1 + \varepsilon_2)$*-optimal solution of* (PP).

Proof. One just needs to show that, for any $c \in S_{[-\infty,+\infty]}$, there exists $\bar{c} \in S_{[L,U]}$ such that

$$\Phi(\bar{c}) \leq \Phi(c) + \varepsilon_1 + \varepsilon_2.$$

To show this, we first show that for any $c \in S_{[-\infty,+\infty]}$ there exists $\hat{c} \in S_{[-\infty,U]}$, such that

$$\Phi(\hat{c}) \leq \Phi(c) + \varepsilon_1. \tag{14}$$

Let $c = (c_1, c_2, \dots, c_p) \in S_{[-\infty,+\infty]}$. If $c_p \le U$, we are done (take $\hat{c} = c$), thus we can assume that $c_p > U$. Let $c_0 = -\infty$, $c_{p+1} = +\infty$ and let i^* be given by

$$i^* = \max\{i : c_i \le U\},$$

which, by assumption, verifies $i^* < p$.

Let $\hat{c} \in S_{[-\infty,U]}$ be given by

$$\hat{c}_i = \min\{c_i, U\} = \begin{cases} c_i, & \text{if } i \le i^* \\ U, & \text{else.} \end{cases}$$

Then,

$$\begin{aligned}
\Phi(\hat{c}) &\le \sum_{i=1}^{p} \int_{(\frac{c_{i-1}+c_i}{2}, \frac{c_i+c_{i+1}}{2}]} (x - \hat{c}_i)^2 \mathrm{d}F(x) \\
&= \sum_{i=1}^{i^*} \int_{(\frac{c_{i-1}+c_i}{2}, \frac{c_i+c_{i+1}}{2}]} (x - c_i)^2 \mathrm{d}F(x) \\
&\quad + \sum_{i=i^*+1}^{p} \int_{(\frac{c_{i-1}+c_i}{2}, \frac{c_i+c_{i+1}}{2}]} (x - U)^2 \mathrm{d}F(x) \\
&= \Phi(c) - \sum_{i=i^*+1}^{p} \int_{(\frac{c_{i-1}+c_i}{2}, \frac{c_i+c_{i+1}}{2}]} (x - c_i)^2 \mathrm{d}F(x) \\
&\quad + \int_{(\frac{c_{i^*}+c_{i^*+1}}{2}, +\infty)} (x - U)^2 \mathrm{d}F(x).
\end{aligned}$$

Then, either

$$\frac{c_{i^*} + c_{i^*+1}}{2} > U \tag{15}$$

or

$$\frac{c_{i^*} + c_{i^*+1}}{2} \le U. \tag{16}$$

If (15) holds, then by (13),

$$\Phi(\hat{c}) \quad \le \quad \Phi(c) + \int_{(\frac{c_{i^*}+c_{i^*+1}}{2}, +\infty)} (x - U)^2 \mathrm{d}F(x) \le \Phi(c) + \varepsilon_1,$$

and (14) holds.

On the other hand, if (16) holds, then

$$\begin{aligned}\Phi(\hat{c}) \leq \Phi(c) &- \int_{(\frac{c_{i^*}+c_{i^*+1}}{2},U]} (x-c_{i^*+1})^2 \mathrm{d}F(x) \\ &- \sum_{i=i^*+2}^{p} \int_{(\frac{c_{i-1}+c_i}{2},\frac{c_i+c_{i+1}}{2}]} (x-c_i)^2 \mathrm{d}F(x) \\ &+ \int_{(\frac{c_{i^*}+c_{i^*+1}}{2},U]} (x-U)^2 \mathrm{d}F(x) \\ &+ \int_{(U,+\infty)} (x-U)^2 \mathrm{d}F(x).\end{aligned}$$

By definition of i^*, $c_{i^*} \leq U < c_{i^*+1}$. Hence, since by (16), $\frac{c_{i^*}+c_{i^*+1}}{2} \leq U$, it follows that, for any $x \in (\frac{c_{i^*}+c_{i^*+1}}{2}, U]$,

$$|x-U| = U - x \leq c_{i^*+1} - x = |c_{i^*+1} - x|$$

thus,

$$\int_{(\frac{c_{i^*}+c_{i^*+1}}{2},U]} \left((x-U)^2 - (x-c_{i^*+1})^2\right) \mathrm{d}F(x) \leq 0.$$

Hence

$$\Phi(\hat{c}) \leq \Phi(c) + \int_{(U,+\infty)} (x-U)^2 \mathrm{d}F(x) \leq \Phi(c) + \varepsilon_1,$$

and (14) holds.

Now, let $\bar{c} \in S_{[L,U]}$ be given by

$$\bar{c}_i = \max\{L, \hat{c}_i\}.$$

A similar reasoning shows that

$$\Phi(\bar{c}) \leq \Phi(\hat{c}) + \varepsilon_2,$$

thus

$$\Phi(\bar{c}) \leq \Phi(c) + \varepsilon_1 + \varepsilon_2,$$

and the result holds. □

Proposition 3.2 implies that standard Branch and Bound methods [8], can be used taking $S_{[L,U]}$ as starting set to optimize the d.c. function Φ.

Appropriate values L, U can be found by means of Lemma 3.2. Fortunately, this is not hard for most usual distributions, as shown in [1]. As a simple example, one has:

Proposition 3.3. *If X is a normal variable with mean μ and variance σ^2, then the set $S_{[L_\varepsilon, U_\varepsilon]}$, given by*

$$\begin{aligned} L_\varepsilon &= \mu - \sigma\sqrt{2\log\frac{\sigma^2}{\varepsilon}} \\ U_\varepsilon &= \mu + \sigma\sqrt{2\log\frac{\sigma^2}{\varepsilon}} \end{aligned}$$

contains an ε-optimal solution of (PP).

Proof. For any $U \geq \mu$, one has

$$\begin{aligned} \int_{[U,+\infty)} (x-U)^2 \mathrm{d}F(x) &= \int_{[U,+\infty)} (x-U)^2 \frac{1}{\sqrt{2\pi}\sigma} \mathrm{e}^{-\frac{(x-\mu)^2}{2\sigma^2}} \mathrm{d}x \qquad (17) \\ &= \sigma^2 \int_{[0,+\infty)} y^2 \frac{1}{\sqrt{2\pi}} \mathrm{e}^{-\frac{(\sigma y+U-\mu)^2}{2\sigma^2}} \mathrm{d}y \\ &= \sigma^2 \int_{[0,+\infty)} y^2 \frac{1}{\sqrt{2\pi}} \mathrm{e}^{-\frac{\sigma^2 y^2+(U-\mu)^2+2\sigma y(U-\mu)}{2\sigma^2}} \mathrm{d}y \\ &= \sigma^2 \mathrm{e}^{-\frac{(U-\mu)^2}{2\sigma^2}} \int_{[0,+\infty)} y^2 \frac{1}{\sqrt{2\pi}} \mathrm{e}^{-\frac{y^2}{2}} \mathrm{e}^{-\frac{y(U-\mu)}{\sigma}} \mathrm{d}y \\ &\leq \sigma^2 \mathrm{e}^{-\frac{(U-\mu)^2}{2\sigma^2}} \int_{[0,+\infty)} y^2 \frac{1}{\sqrt{2\pi}} \mathrm{e}^{-\frac{y^2}{2}} \mathrm{d}y \\ &= \sigma^2 \frac{1}{2} \mathrm{e}^{-\frac{(U-\mu)^2}{2\sigma^2}}. \qquad (18) \end{aligned}$$

Let $U_\varepsilon = \mu + \sigma\sqrt{2\log\frac{\sigma^2}{\varepsilon}}$. Then,

$$\frac{|U_\varepsilon - \mu|}{\sqrt{2}\sigma} \geq \sqrt{\log\frac{\sigma^2}{\varepsilon}},$$

thus

$$\mathrm{e}^{-\frac{(U_\varepsilon-\mu)^2}{2\sigma^2}} \leq \mathrm{e}^{-\log\frac{\sigma^2}{\varepsilon}} = \frac{\varepsilon}{\sigma^2},$$

which this implies that

$$\int_{[U_\varepsilon,+\infty)} (x-U_\varepsilon)^2 \mathrm{d}F(x) \leq \frac{\varepsilon}{2}.$$

Similarly, if one defines $L_\varepsilon = \mu - \sigma\sqrt{2\log\frac{\sigma^2}{\varepsilon}}$, it follows that

$$\int_{(-\infty,L_\varepsilon]} (x-L_\varepsilon)^2 \mathrm{d}F(x) \leq \frac{\varepsilon}{2},$$

and by Proposition 3.2, the result follows. □

In [1] we have found values L_ε and U_ε for other random variables such as gamma, Student's t or Snedecor's F. In particular if X is a gamma variable with density

$$\frac{1}{\Gamma(p)} a^p x^{p-1} \mathrm{e}^{-ax}, \qquad x \geq 0$$

where $a > 0$ and $p \geq 1$, then

$$L_\varepsilon = 0, \qquad U_\varepsilon = \frac{\sqrt{2}\Gamma(p+2)}{\sqrt{\varepsilon} a^2 \Gamma(p)} \cdot \tag{19}$$

Moreover, if the distribution of X is a mixture of (simpler) random variables, one can construct the values L, U from the values L_i, U_i corresponding to the distributions involved in the mixture. This may be of interest for important cases such as the mixture of Erlang laws or more generally Cox distributions [2, 6], commonly used to model queuing systems [6].

Indeed, one has:

Proposition 3.4. *Suppose that the distribution function F is a mixture of distributions functions $F_1, \ldots, F_n$. Let $L_1 \leq U_1, \ldots, L_n \leq U_n$ be scalars such that*

$$\begin{aligned} \int_{(-\infty, L_i]} (x - L_i)^2 \mathrm{d}F_i(x) &\leq \varepsilon_1, \quad i = 1, \ , \ldots, n \\ \int_{[U_i, +\infty)} (x - U_i)^2 \mathrm{d}F_i(x) &\leq \varepsilon_2, \quad i = 1, \ , \ldots, n. \end{aligned}$$

Define $L = \min_{i=1,\ldots,n} L_i$, $U = \max_{i=1,\ldots,n} U_i$; then the set $S_{[L,U]}$ contains an $(\varepsilon_1 + \varepsilon_2)$-optimal solution to (PP).

Proof. First, observe that $F = \sum_{i=1}^n \alpha_i F_i(x)$ for some nonnegative scalars $\alpha_1, \ldots, \alpha_n$, $\sum_{i=1}^n \alpha_i = 1$.

$$\begin{aligned} \int_{(-\infty, L]} (x - L)^2 \mathrm{d}F(x) &= \int_{(-\infty, L]} (x - L)^2 \mathrm{d} \left(\sum_{i=1}^n \alpha_i F_i(x) \right) \\ &= \sum_{i=1}^n \alpha_i \int_{(-\infty, L]} (x - L)^2 \mathrm{d}F_i(x). \end{aligned}$$

By Lemma 3.2,

$$\int_{(-\infty, L]} (x - L)^2 \mathrm{d}F_i(x) \leq \int_{(-\infty, L_i]} (x - L_i)^2 \mathrm{d}F_i(x),$$

thus

$$\int_{(-\infty,L]}(x-L)^2\mathrm{d}F_i(x)=\sum_{i=1}^{n}\alpha_i\int_{(-\infty,L_i]}(x-L_i)^2\mathrm{d}F_i(x)\leq\sum_{i=1}^{n}\alpha_i\varepsilon_1=\varepsilon_1,$$

and (12) follows.

A similar reasoning shows that (13) holds. □

4. Concluding remarks

In this note we have shown that the problem of finding p principal points of a random variable X can be tackled with standard tools of Global Optimization: the objective function is d.c. (and a d.c. decomposition is available), and a simplex known to contain an ε-optimal solution is easy to construct. This contrasts with the literature on the field, where only local search procedures have been suggested.

5. Appendix

Table 2. AMPL code.

```
#
# Finding 2-principal points of a discrete variable with
# support in [LOWLIMIT,UPLIMIT]

param LOWLIMIT;
param UPLIMIT;
param SIZESUP; #size of the support
param PROBAB{1..SIZESUP};
param POINT{1..SIZESUP}; #point[i] has probability probab[i]
param P; #number of principal points sought

var C{1..P}:= Uniform(LOWLIMIT,UPLIMIT); # principal points

minimize OBJECTIVE:
sum{i in 1..SIZESUP}
      (PROBAB[i] * min{j in 1..P} ((C[j]-POINT[i])^ 2));
subject to INSIMPLEX{i in 1..P-1}:  C[i] <= C[i+1];
subject to INSUPPORT1:  C[1] >= LOWLIMIT;
subject to INSUPPORT2:  C[P] <= UPLIMIT;
```

References

[1] E. Carrizosa, E. Conde, A. Castaño, I. Espinosa, I. González and D. Romero–Morales, Puntos principales: Un problema de Optimización Global en Estadística, Presented at *XXII Congreso Nacional de Estadística e Investigación Operativa*. Sevilla (1995).

[2] D.R. Cox, A use of complex probabilities in the theory of stochastic processes, in *Proc. of the Cambridge Philosophical Society*, Vol. 51 (1955) 313-319.

[3] B. Flury, Principal points. *Biometrika* **77** (1990) 33-41.

[4] B. Flury and T. Tarpey, Representing a Large Collection of Curves: A Case for Principal Points. *Amer. Statist.* **47** (1993) 304-306.

[5] R. Fourer, D.M. Gay and B.W. Kernigham, *AMPL, A modeling language for Mathematical Programming*. The Scientific Press, San Francisco (1993).

[6] E. Gelenbe and R.R. Muntz, Probabilistic Models of Computer Systems-Part I. *Acta Inform.* **7** (1976) 35-60.

[7] R. Horst, An Algorithm for Nonconvex Programming Problems. *Math. Programming* **10** (1976) 312-321.

[8] R. Horst and H. Tuy, *Global Optimization. Deterministic Approaches*. Springer-Verlag, Berlin (1993).

[9] S.P. Lloyd, Least Squares Quantization in PCM. *IEEE Trans. Inform. Theory* **28** (1982) 129-137.

[10] L. Li and B. Flury, Uniqueness of principal points for univariate distributions. *Statist. Probab. Lett.* **25** (1995) 323-327.

[11] K. Pötzelberger and K. Felsenstein, An asymptotic result on principal points for univariate distribution. *Optimization* **28** (1994) 397-406.

[12] S. Rowe, An Algorithm for Computing Principal Points with Respect to a Loss Function in the Unidimensional Case. *Statist. Comput.* **6** (1997) 187-190.

[13] T. Tarpey, Two principal points of symmetric, strongly unimodal distributions. *Statist. Probab. Lett.* **20** (1994) 253-257.

[14] T. Tarpey, Principal points and self-consistent points of symmetric multivariate distributions. *J. Multivariate Anal.* **53** (1995) 39-51.

[15] T. Tarpey, L. Li and B. Flury, Principal points and self-consistent points of elliptical distributions. *Ann. Statist.* **23** (1995) 103-112.

[16] A. Zoppè, Principal points of univariate continuous distributions. *Statist. Comput.* **5** (1995) 127-132.

[17] A. Zoppè, On Uniqueness and Symmetry of self-consistent points of univariate continuous distribution. *J. Classification* **14** (1997) 147-158.