METHODS FOR ESTIMATING PRINCIPAL POINTS

Erwin Stampfer and Ernst Stadlober

Institute of Statistics, University of Technology Graz,
Steyrergasse 17/IV, A–8010 Graz, Austria
stampfer@stat.tu-graz.ac.at

Key Words: Self–consistency algorithms, k–means clustering, univariate distributions, Bootstrap methods.

ABSTRACT

Principal points of a distribution have been introduced by Flury (1990) who tackled the problem of optimal grouping in multivariate data. In essence, principal points are the theoretical counterparts of cluster means obtained by a k–means clustering algorithm. There has been considerable effort to find efficient estimation procedures for principal points. It is well known that under certain conditions the k–means estimator is a consistent and asymptotically normal estimator of the population principal points. In this paper some material on principal points is reviewed and new algorithms for the estimation of principal points in univariate distributions (univariate principal points) are proposed. Additionally, the Bootstrap approach is applied to assess the variability of the suggested estimators.

1 INTRODUCTION

Univariate principal points are discrete approximations of random variables with minimal mean squared error. In this section we review some theoretical material underlying this concept (details may be found in Zoppe, 1995). The definitions and results given in this section may apply in general to random vectors, but here we will be concerned with univariate principal points only. Hence our notation will be restricted to random variables. Note that the strength of this scheme especially uncovers in the multivariate case.

Definition 1.1 *The mean squared error of the approximation of* X *by* Y *is defined by* $MSE(X;Y) = \mathbb{E}[(X-Y)^2]$, *where the expectation is taken with respect to the common distribution of the random vector* (X,Y).

It is well known (Pfeiffer, 1990), that for all Borel measurable functions g the conditional expectation $\mathbb{E}[X|Y]$ has the minimal squared error, i.e.
$MSE(X;\mathbb{E}[X|Y]) \leq MSE(X;g(Y))$ for all Borel measurable functions g.
This motivates the following definition.

Definition 1.2 *Let* X *and* Y *be integrable random variables with a common distribution.* Y *is self–consistent for* X, *if* $\mathbb{E}[X|Y] = Y$ *almost surely.*

Some properties of multivariate self–consistent approximations can be found in Tarpey&Flury (1996). Here we are interested in a special type of approximation $Y = h(X)$ of X, where $h(x)$ maps x to a point y of a set $\{y_1, \ldots, y_k\}$ of k points. For convenience we introduce $\mathcal{P}_k$ as the set of vectors $\boldsymbol{y} = (y_1, \ldots, y_k)$, where $y_1 < \ldots < y_k$. Let us define the projection $\Pi_{\boldsymbol{y}}(x)$ by $\Pi_{\boldsymbol{y}}(x) = y_i$ iff $|x - y_i| < |x - y_{i'}|$ $1 \leq i', i \leq k, i' \neq i$, with some breaking rule in the case of ties (i.e. the distance is minimal for two distinct points y_i and $y_{i'}$). We like to optimize over $\mathcal{P}_k$, hence introduce the notation $MSE(F_X;\boldsymbol{y}) = MSE(X;\Pi_{\boldsymbol{y}}(X))$, $\boldsymbol{y} \in \mathcal{P}_k$ and define principal points of X as follows.

Definition 1.3 *Let $X \sim F_X$ with variance σ_X^2 and $\boldsymbol{\zeta}, \boldsymbol{y} \in \mathcal{P}_k$. If*

$$MSE(F_X; \boldsymbol{\zeta}) \leq MSE(F_X; \boldsymbol{y}) \quad \forall \boldsymbol{y} \in \mathcal{P}_k, \tag{1}$$

then $\zeta_1, \ldots, \zeta_k$ are called principal points *of X or F_X, respectively.*

We may call the points $\zeta_1, \ldots, \zeta_k$ self–consistent, if $\Pi_{\boldsymbol{\zeta}}(X)$ is self–consistent for X. There is a nice relationship between self–consistency and the optimizing property of principal points. The following basic result is given in Flury (1990).

Theorem 1.1 *Principal points are self–consistent points.*

The next theorem considers the existence and uniqueness of self–consistent points and is due to Kieffer (1983).

Theorem 1.2 *Let X be a random variable with density f_X and variance σ_X^2. Then the following holds for all $k \in \mathbb{N}$.*

1. *There exists at least one set of principal, hence self–consistent points.*
2. *If in addition f_X is log-concave, then there is exactly one set of self–consistent, hence (see 1.) principal points.*

Theorems 1.1 and 1.2 motivate the so called self–consistency algorithm SCA an iterative procedure designed to produce self–consistent approximations. Its implementation in the case of univariate principal points may be written down as follows.

Algorithm SCA

0. Choose a starting vector $\boldsymbol{y}^{(0)} \in \mathcal{P}_k$.
1. Let $D_i^{(r)}$ denote the set $\left\{x | \Pi_{\boldsymbol{y}^{(r)}}(x) = y_i^{(r)}\right\}$ and $I_{D_i^{(r)}}(x)$ its indicator function. For each $i = 1, \ldots, k$ calculate (if possible)

$$\begin{aligned} y_i^{(r+1)} &= \mathbb{E}\left[X | \Pi_{\boldsymbol{y}^{(r)}}(X) = y_i^{(r)}\right] \\ &= \frac{\int x I_{D_i^{(r)}}(x) dF_X(x)}{\int I_{D_i^{(r)}}(x) dF_X(x)}. \end{aligned} \tag{2}$$

2. Iterate Step 1 until convergence.

Properties of algorithm SCA are discussed by Tarpey (1999) in a more general setting. Two important features are: (i) the mean squared error $MSE(X;\Pi_{\boldsymbol{y}^{(r)}}(X))$ is monotonically decreasing in r, (ii) if $\boldsymbol{y}^{(r)}$ converges to $\boldsymbol{\zeta}$, then $\Pi_{\boldsymbol{\zeta}}(X)$ is self–consistent for X.

Note that the criterion function MSE always converges. However, in general this does not imply, that $\boldsymbol{y}^{(r)}$ converges.

A difficult problem is the estimation of self–consistent and principal points for a given sample $x_1,\ldots,x_n$ from distribution F_X. Key references to this topic are Flury (1993) and Tarpey (1997). In Section 2 we discuss several procedures for estimating principal points: the maximum likelihood (ML) estimator, a parametric ($QE1$) and a nonparametric ($QE2$) version of quick estimates based on an asymptotic result of Pötzelberger&Felsenstein (1994), the classical k–means estimate (KM), a smoothed variant (SKM) of KM, and two variants of KM using symmetry properties ($KMS1$, $KMS2$). A mainly untackled problem is the investigation of the variability of the estimates. In Section 3 the Bootstrap approach will be proposed to this end. Section 4 reports on a simulation study comparing the seven estimates in terms of mean squared error. In Section 5 optimal grouping of the *Cherry–Tree* data of Hand&Daly (1993) is considered and our findings are summarized in the final Section 6.

2 ESTIMATION PROCEDURES

Let $X \sim F_X$ be a random variable with density f_X, expectation μ_X and variance σ_X^2. The vector of principal points $\boldsymbol{\zeta} = (\zeta_1,\ldots,\zeta_k)$ of X is assumed to be unique. In general, $\boldsymbol{\zeta}$ will be unknown and has to be estimated based on an iid sample $X_1,\ldots,X_n$ of F_X. As criterion for the quality of an estimator $\hat{\boldsymbol{\zeta}}$ we will use the mean squared error defined by

$$MSE(\hat{\boldsymbol{\zeta}};\boldsymbol{\zeta}) := \mathbb{E}\left[\sum_{i=1}^{k}(\hat{\zeta}_i - \zeta_i)^2\right].$$

2.1 The ML-Estimator

Assume $f_X(x) = f_X(x;\boldsymbol{\theta})$, $\quad \boldsymbol{\theta} \in \Omega \subseteq \mathbb{R}^d$, where $\boldsymbol{\theta}$ is a vector of unknown parameters. To express the dependency of $\boldsymbol{\zeta}$ on $\boldsymbol{\theta}$ we write $\boldsymbol{\zeta} = \boldsymbol{\zeta}(\boldsymbol{\theta}) = (\zeta_1(\boldsymbol{\theta}), \ldots, \zeta_k(\boldsymbol{\theta}))$. Now the usual ML estimator of $\boldsymbol{\theta}$ is denoted by $\hat{\boldsymbol{\theta}}$ and defined by

$$\hat{\boldsymbol{\theta}} = \arg\max_{\theta\in\Omega} \prod_{j=1}^{n} f_X(x_j;\boldsymbol{\theta}). \tag{3}$$

The ML estimator of $\boldsymbol{\zeta}$ is given by $\hat{\boldsymbol{\zeta}} = \boldsymbol{\zeta}(\hat{\boldsymbol{\theta}})$. Usually the calculation of $\hat{\boldsymbol{\zeta}}$ is carried out in two steps; in the first step estimator $\hat{\boldsymbol{\theta}}$ (3) is calculated and in the second step the self–consistency algorithm is applied to $f_X(x;\hat{\boldsymbol{\theta}})$.

2.2 Quick Estimates

The quick estimates rest upon an asymptotic result of Pötzelberger&Felsenstein (1994). Set

$$g(x) = \frac{f_X^{1/3}(x)}{\int_{-\infty}^{\infty} f_X^{1/3}(x)dx}, \tag{4}$$

provided that $g(x)$ exists. Denote by G the corresponding cdf, and let $y_i(k) = G^{-1}\left(\frac{i}{k+1}\right)$ be its $\frac{i}{k+1}$–quantiles. Indicate the vector $(\zeta_1, \ldots, \zeta_k)$ of principal points by $\boldsymbol{\zeta}_k$, $k \geq 1$. Under suitable conditions the sequence $\boldsymbol{y}_k = (y_1(k), \ldots, y_k(k))$ is asymptotically optimal in the following sense

$$\lim_{k\to\infty} \frac{MSE(F_X;\boldsymbol{y}_k)}{MSE(F_X;\boldsymbol{\zeta}_k)} = 1.$$

Even for small k the vector $\boldsymbol{y}_k$ is a good approximation to $\boldsymbol{\zeta}_k$, as Pötzelberger&Felsenstein (1994) demonstrate. This result can be used to get estimates for $\boldsymbol{\zeta}_k$. The idea is to estimate the pdf f_X by $\hat{f}_X$, which can be done parametrically (plugging in the ML estimates) or non–parametrically by density estimation. Then the principal points are estimated by $\hat{\zeta}_i = \hat{G}^{-1}\left(\frac{i}{k+1}\right)$, where $\hat{G}$ is the distribution function belonging to density $\hat{g}$ obtained by replacing f_X in (4) by $\hat{f}_X$. We will call the ML version $QE1$ and the nonparametric variant $QE2$.

2.3 Variants of the k–means estimate

The four methods described in the sequel are very similar in nature. First we define the smoothed k–means estimate (SKM) by relation (1), where we replace F_X by the distribution function of the corresponding nonparametric density estimate $\hat{f}_X$. The second nonparametric variant is the classical k–means estimate (KM) as well stipulated by (1), but now F_X is superseded by the empirical distribution function (edf) $\hat{F}_X$, putting mass $1/n$ on each data point.

The next two methods are k–means variants, which require symmetry properties of the distribution or the principal points, respectively. One variant presumes f_X to be *symmetric about its mean.* In this case we have $F_X(x) = 1 - F_X(2\mu_X - x)$ for $x \in \mathbb{R}$, which can be estimated more efficiently by

$$\hat{F}_{X,sym}(x) = \begin{cases} \frac{1}{2n}\sum_{j=1}^{n}\left(I_{(-\infty,x]}(x_j) + I_{(2\mu_X - x,\infty)}(x_j)\right) & x \leq \mu_X \\ 1 - \hat{F}_{X,sym}(2\mu_X - x) & x > \mu_X\ . \end{cases} \quad (5)$$

Estimate (5) is unbiased for all $x \in \mathbb{R}$ having variance

$$\mathbb{V}\left[\hat{F}_{X,sym}(x)\right] \;=\; \frac{1}{2n}\left(F_X(x)(1 - F_X(x)) - F_X^2(x)\right), \quad x \leq \mu_X.$$

The $KMS1$ estimate is defined analogously to the KM estimate, but now substituting $\hat{F}_{X,sym}$ for F_X in (1).

The last variant, denoted by $KMS2$, postulates *symmetry of the principal points* around the expectation μ_X, i.e. in addition to $MSE(\hat{F}_X; \hat{\boldsymbol{\zeta}}) \leq$

$MSE(\hat{F}_X; \boldsymbol{y})$ for all $\boldsymbol{y} \in \mathcal{P}_k$ we require

$$\hat{\zeta}_i = 2\mu_X - \hat{\zeta}_{k-i+1}, \quad i = 1, \ldots, k. \tag{6}$$

For k odd (6) implies $\hat{\zeta}_{\lceil k/2 \rceil} = \mu_X$. Since μ_X is usually unknown it has to be replaced by an estimate $\hat{\mu}_X$ in both variants.

2.4 Computation of the k–means variants

The computation of the estimates SKM, KM, $KMS1$ and $KMS2$ is based on algorithm SCA as described in the introduction. The procedure for the evaluation of the SKM estimate may be called *smoothed k–means algorithm* and employs algorithm SCA to the nonparametric density estimate $\hat{f}_X$. Hence, this algorithm involves the calculation of $\hat{f}_X$ and integrals of type (2), where integration is with respect to $\hat{f}_X$.

The KM estimate can be calculated by the classical k–means algorithm, where algorithm SCA is applied to $\hat{F}_X$. In this case the integrals (2) reappear as finite sums.

Essentially the same applies to the $KMS1$ estimate, after augmentation of the sample $x_1, \ldots, x_n$ by $2\hat{\mu}_X - x_1, \ldots, 2\hat{\mu}_X - x_n$, with $\hat{\mu}_X$ as estimate of μ_X. Obviously, the edf of the augmented sample equals $\hat{F}_{X,sym}$ (5) with μ_X replaced by $\hat{\mu}_X$.

The calculation of the $KMS2$ estimate is somewhat trickier. Define

$$x_j^* = \begin{cases} x_j & x_j > \hat{\mu}_X \\ 2\hat{\mu}_X - x_j & x_j \leq \hat{\mu}_X. \end{cases}$$

and let $\hat{F}_{X^*}$ be the edf based on $x_1^*, \ldots, x_n^*$. For k even one calculates the KM estimate of the $k/2$ principal points $\zeta_1^*, \ldots, \zeta_{k/2}^*$ of X^*. The $KMS2$ estimate is then given by the augmented vector $(-\hat{\zeta}_{k/2}^*, \ldots, -\hat{\zeta}_1^*, \hat{\zeta}_1^*, \ldots, \hat{\zeta}_{k/2}^*)^T$. For k odd one has to take into account that $y_{\lceil k/2 \rceil} = \hat{\mu}_X$. This leads to the following modified variant of algorithm SCA applied to $\hat{F}_{X^*}$.

0. Choose $\boldsymbol{q}^{(0)} \in \mathcal{P}_{\lceil k/2 \rceil}$ with $q_1^{(0)} = \hat{\mu}_X$.

1. For $i = 2, \ldots, \lceil k/2 \rceil$ calculate the conditional integrals $q_i^{(r+1)}$ given by formula (2), plugging in $\hat{F}_{X^*}$ for F_X. Set $\boldsymbol{q}^{(r+1)} = (\hat{\mu}_X, q_2^{(r+1)}, \ldots, q_{\lceil k/2 \rceil}^{(r+1)})$.

2. Iterate Step 1 till convergence – resulting in, say, $\boldsymbol{q} \in \mathcal{P}_{\lceil k/2 \rceil}$.

   The $KMS2$ estimate is given by

$$\hat{\boldsymbol{\zeta}} = (2\hat{\mu}_X - q_{\lceil k/2 \rceil}, \ldots, 2\hat{\mu}_X - q_2, \hat{\mu}_X, q_2, \ldots, q_{\lceil k/2 \rceil}).$$

It can be shown (see Stampfer, 2000, p. 93f) that this algorithm has optimizing properties similar to algorithm SCA. However, note that $\hat{\boldsymbol{\zeta}}$ is not self–consistent with respect to $\hat{F}_X$.

2.5 Choice of the bandwidth h

In the nonparametric case the choice of the bandwidth h is crucial. If $X \sim N(\mu_X, \sigma_X^2)$ and the gaussian kernel is used as estimate $\hat{f}_X$ then the optimal bandwidth is given by

$$h_{opt} = 1.06\, \sigma_X\, n^{-1/5}. \tag{7}$$

Estimating σ_X by the empirical standard deviation s leads to an estimate for h_{opt}. In the case of deviations from the normal distribution with respect to skewness and kurtosis, (7) has the tendency to oversmooth. This can be improved in estimating σ_X by $s_q = (q_{0.75} - q_{0.25})/1.349$ with q_p the p–th empirical quantile. However, this selection tends to oversmooth in the case of bimodal densities. So we follow the recommendation of Silverman (1986) to plug in the adaptive estimate $s_A = \min(s, s_q)$.

3 ASSESSING VARIABILITY BY BOOTSTRAP METHODS

Only a few theoretical results concerning variability of the estimates are available. For both ML and KM estimates consistency and asymptotic normality can be verified under suitable restrictions on the cdf F_X. For the ML method we must demand that in addition to the usual requirements for asymptotic normality of maximum likelihood estimates the principal points are sufficiently smooth functions of the parameters of the distribution. Under suitable conditions on the moments of X the delta method applies (Bhattacharya&Denker, 1990), giving the following result in the case of a normal population.

Lemma 3.1 *Let* $X_1, \ldots, X_n$ *be an iid sample from* $N(\mu_X, \sigma_X^2)$. *Let* $\boldsymbol{\zeta}$ *be the vector of principal points of* X, $\hat{\boldsymbol{\zeta}}$ *its* ML *estimate and* $\boldsymbol{\eta}$ *the vector of principal points of the* $N(0,1)$*–distribution – altogether elements of* $\mathcal{P}_k$, $k \in \mathbb{N}$. *We have*

$$\sqrt{n}(\hat{\boldsymbol{\zeta}} - \boldsymbol{\zeta}) \quad \xrightarrow{d} \quad N(0, \boldsymbol{V}_{\hat{\boldsymbol{\zeta}}}),$$

where $\boldsymbol{V}_{\hat{\boldsymbol{\zeta}}}$ *is a* $(k \times k)$*-matrix with components*

$$(\boldsymbol{V}_{\hat{\boldsymbol{\zeta}}})_{ii'} = \begin{cases} \sigma_X^2 \left(1 + \frac{1}{2}\eta_i^2\right) & i = i' \\ \sigma_X^2 \left(1 + \frac{1}{2}\eta_i \eta_{i'}\right) & i \neq i'. \end{cases}$$

The proof of Lemma 3.1 may be found in Stampfer (2000, p. 97ff). For the KM estimate one may consult Hartigan (1974, 1978) or Pollard (1981, 1982).

In both cases a stochastic model has to be specified completely. Nonparametric estimates for the variability of the suggested estimates are therefore highly desireable. In the following we will briefly describe how the Bootstrap can be used to this end. We may also look at the bias of the estimates. This is elaborated for the KM estimate, but for the other estimates it can be worked out similarly.

Write $\boldsymbol{\zeta} = \boldsymbol{\zeta}(F_X)$ as function of F_X. Following Davison&Hinkley (1997) we regard $\boldsymbol{\zeta}$ as algorithm applied to F_X. The estimate $\hat{\boldsymbol{\zeta}}$ is defined by $\boldsymbol{\zeta}(\hat{F}_X)$, i.e. the same algorithm applied to the edf $\hat{F}_X$. We may also write $\hat{\boldsymbol{\zeta}} = \boldsymbol{\zeta}(X_1, \ldots, X_n)$ and are interested in statistical properties of $\hat{\boldsymbol{\zeta}}$, e.g. the bias

$$\boldsymbol{b}(F_X) = \mathbb{E}[\boldsymbol{\zeta}(X_1, \ldots, X_n) - \boldsymbol{\zeta}(F_X)|F_X] \tag{8}$$

and the variance

$$\boldsymbol{v}(F_X) = \mathbb{V}[\boldsymbol{\zeta}(X_1, \ldots, X_n)|F_X]\,. \tag{9}$$

Since F_X is usually unknown, we cannot calculate the quantities (8) and (9). Following the usual conventions we denote by $X_1^*, \ldots, X_n^*$ an iid sample from $\hat{F}_X$. In (8) and (9) replace F_X by $\hat{F}_X$ and $X_1, \ldots, X_n$ by $X_1^*, \ldots, X_n^*$ to get the *Bootstrap bias- and variance estimates* for $\boldsymbol{\zeta}(\hat{F}_X)$, i.e.

$$\boldsymbol{b}(\hat{F}_X) = \mathbb{E}\left[\boldsymbol{\zeta}(X_1^*, \ldots, X_n^*) - \boldsymbol{\zeta}(\hat{F}_X)|\hat{F}_X\right] = \mathbb{E}\left[\boldsymbol{\zeta}(X_1^*, \ldots, X_n^*)|\hat{F}_X\right] - \hat{\boldsymbol{\zeta}}, \tag{10}$$

and

$$\boldsymbol{v}(\hat{F}_X) = \mathbb{V}\left[\boldsymbol{\zeta}(X_1^*, \ldots, X_n^*)|\hat{F}_X\right]. \tag{11}$$

Theoretically, one can evaluate (10) and (11) exactly, but in most cases this is to complicated and one resorts to Monte Carlo Methods to obtain an approximate solution.

1. Generate R Bootstrap replicates $x_{1r}^*, \ldots, x_{nr}^*$, $r = 1, \ldots, R$, from $\hat{F}_X$. Apply algorithm SCA to the Bootstrap replicates, i.e. calculate the vector $\hat{\boldsymbol{\zeta}}_r^* = \boldsymbol{\zeta}(x_{1r}^*, \ldots, x_{nr}^*)$, $r = 1, \ldots, R$.

2. Approximate (10) by $\bar{\boldsymbol{\zeta}}^* - \hat{\boldsymbol{\zeta}}$, where $\bar{\boldsymbol{\zeta}}^* = \frac{1}{R}\sum_{r=1}^{R} \hat{\boldsymbol{\zeta}}_r^*$ and (11) by
$$\frac{1}{R-1}\sum_{r=1}^{R}(\hat{\boldsymbol{\zeta}}_r^* - \bar{\boldsymbol{\zeta}}^*)(\hat{\boldsymbol{\zeta}}_r^* - \bar{\boldsymbol{\zeta}}^*)^T.$$

4 SIMULATION RESULTS

In this section we present the results of a simulation study where we compare the suggested estimates ML, $QE1$, $KMS1$, $KMS2$, SKM, KM and $QE2$ in terms of mean squared error. (In what follows V denotes one out of these methods.) This was done for $k = 2, 3, 4, 5$ principal points and for each k with five different sample sizes $n(k, r)$, $r = 1, \ldots, 5$, shown in the following table.

$n(k,r)$	1	2	3	4	5
2	20	40	60	80	200
3	30	50	70	90	210
4	40	60	80	100	220
5	50	70	90	110	230

For each of these 20 combinations of (r, k) we generated $m = 2500$ samples of size $n(k, r)$ from a known distribution F_X (and known vector of principal points $\boldsymbol{\zeta}$) and calculated the seven estimates $\hat{\boldsymbol{\zeta}}(V, k, r)$. We compare the estimates by calculating

$$\widehat{MSE}(\hat{\boldsymbol{\zeta}}(V,k,r);\boldsymbol{\zeta}) = \frac{1}{m}\sum_{l=1}^{m}\sum_{i=1}^{k}\left(\hat{\zeta}_i^{(l)}(V,k,r) - \zeta_i\right)^2, \tag{12}$$

i.e. the Monte Carlo approximation of $MSE(\hat{\boldsymbol{\zeta}}(V, k, r); \boldsymbol{\zeta})$. Furthermore the relative efficiency of the estimates with respect to the ML estimate is approximated by

$$\widehat{Eff}\left(\hat{\boldsymbol{\zeta}}(V,k,r);\hat{\boldsymbol{\zeta}}(ML,k,r)\right) = \frac{\widehat{MSE}(\hat{\boldsymbol{\zeta}}(V,k,r);\boldsymbol{\zeta})}{\widehat{MSE}(\hat{\boldsymbol{\zeta}}(ML,k,r);\boldsymbol{\zeta})}. \tag{13}$$

The simulation study has been carried out with the package `S-PLUS 2000` (see Venables&Ripley, 1999) under Windows NT on a Pentium III–333MHz. All programs to perform the simulations were written in `S-PLUS 2000`. Uniform random numbers were generated by the `S-PLUS` function `runif` based on

Marsaglia´s super–duper generator. Normal random variates were produced by `rnorm` using the ratio–of–uniforms method of Kinderman and Monahan (1977).

4.1 Normal distribution

In the following experiment we sampled from $N(0,1)$. Table I contains the approximated mean squared error (12) and the relative efficiency according to (13). The values in Table I may be interpreted as follows.

As expected the ML estimate is most efficient whereas te KM estimate appears inferior.

The $QE1$ estimate is similar to the ML estimate. The (nonparametric) $QE2$ estimate turns out to be more efficient than the KM estimate, and for most sample sizes it is even more efficient than the $KMS1$ and $KMS2$ estimates. The SKM estimate performs similar to $KMS1$ and $KMS2$.

For $k = 2$ the $KMS1$ and $KMS2$ estimates give results comparable to the parametric estimates. However, for larger k they quickly deteriorate. $KMS2$ always performs better than $KMS1$, the latter yields approximately half of the mean squared error of the classical k–means estimate KM.

TABLE I. $X \sim N(0,1)$. Estimates (12) for pairs of k and n; for the smallest and the largest sample size also the values of estimate (13).
Theoretical values (as far as available) in brackets.

	ML	$QE1$	$KMS1$	$KMS2$	SKM	KM	$QE2$
k = 2							
$n = 20$	.13	.13	.19	.13	.23	.48	.20
40	.06	.07	.09	.07	.11	.20	.10
60	.04	.05	.05	.04	.08	.13	.07
80	.03	.04	.04	.03	.06	.09	.05
200	.01 (.01)	.02 (.02)	.02	.01	.03	.03 (.03)	.03
20	1	1.02	1.48	1.02	1.73	3.60	1.46
200	1	1.38 (1.53)	1.15	1.08	1.92	2.46 (2.37)	2.08
k = 3							
$n = 30$	.15	.15	.35	.18	.33	.75	.22
50	.09	.10	.20	.11	.20	.57	.14
70	.07	.07	.14	.08	.13	.42	.11
90	.05	.06	.10	.06	.11	.28	.08
210	.02 (.02)	.03 (.03)	.04	.03	.05	.10 (.09)	.04
30	1	1.02	2.32	1.17	2.18	5.00	1.47
210	1	1.29 (1.43)	1.90	1.19	2.42	4.67 (4.13)	2.10
k = 4							
$n = 40$	.16	.17	.49	.35	.40	.90	.25
60	.10	.11	.33	.23	.29	.80	.17
80	.08	.09	.26	.16	.21	.71	.14
100	.07	.07	.22	.14	.17	.60	.11
220	.03 (.03)	.03 (.04)	.08	.044	.08	.25 (.19)	.06
40	1	1.02	2.99	2.12	2.46	5.50	1.52
220	1	1.17 (1.36)	2.79	1.52	2.65	8.72 (6.34)	2.21
k = 5							
$n = 50$	.18	.18	.64	.37	.45	.99	.28
70	.12	.13	.49	.25	.36	.91	.20
90	.10	.10	.43	.18	.29	.88	.17
110	.08	.08	.36	.15	.23	.83	.14
230	.04 (.04)	.04 (.05)	.16	.07	.12	.53 (.33)	.08
50	1	1.02	3.64	2.11	2.56	5.67	1.58
230	1	1.14 (1.31)	4.30	1.86	3.13	14.32 (8.94)	2.22

4.2 Robustness against misspecification of the model

In practice the data generating mechanism is unknown and one is forced to make assumptions about the distribution F_X. In the sequel we investigate the estimates with respect to their robustness against a specific misspecification $\tilde{F}$ of the stochastic model. This is done as follows.

1. Simulate samples $x_1, \ldots, x_n$ from a given distribution F_X with the vector of principal points $\boldsymbol{\zeta}$.
2. Estimate $\boldsymbol{\zeta}$ under distribution $\tilde{F}$.
3. Estimate the mean squared error (12) of the estimate $\hat{\boldsymbol{\zeta}}$ under $\tilde{F}$ from the vector of principal points $\boldsymbol{\zeta}$ of F_X.

We take $\tilde{F}(x) = \Phi_{\mu,\sigma^2}(x)$ with density ϕ_{μ,σ^2}, i.e. a normal distribution with parameters μ and σ^2. But in fact F_X is determined as a scale mixture of normal distributions with density

$$f_X(x) = \tfrac{1}{2}\phi_{0,20/11}(x) + \tfrac{1}{2}\phi_{0,2/11}(x), \tag{14}$$

where the components are scaled such that $\mathbb{V}[X] = 1$. The left panel of Figure 1 shows density (14) and $\phi_{0,1}(x)$. The right panel contains a histogram and a density estimate of a sample of size 100 drawn from (14).

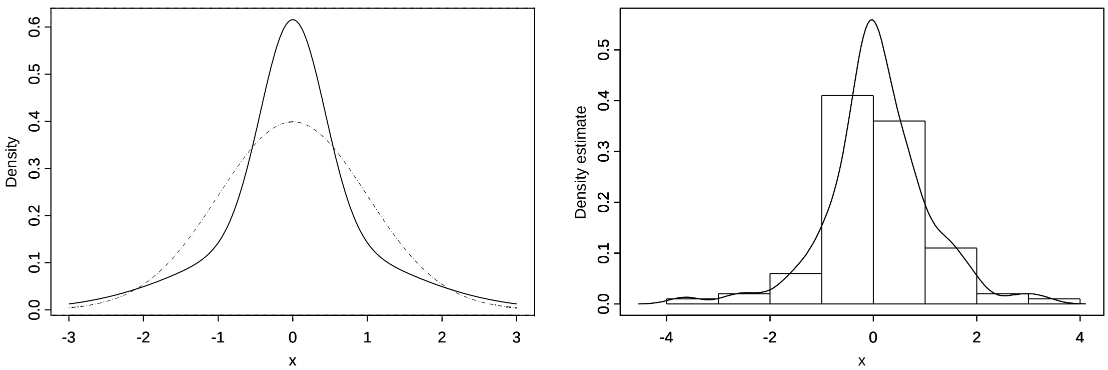

FIG. 1. Left panel: Density function (solid line) of the scale mixture (14) compared with $\phi_{0,1}(x)$. Right panel: Histogram and density estimate of a sample of size 100.

TABLE II. $X \sim f_X$ from (14). Estimates (12) for pairs of k and n; for the smallest and the largest sample size also the values of estimate (13). Best values in italics.

	ML	$QE1$	$KMS1$	$KMS2$	SKM	KM	$QE2$
k = 2							
$n = 20$	.16	.15	.61	*.14*	1.02	1.71	.28
40	.09	.08	.38	*.07*	.56	1.30	.16
60	.06	.05	.26	*.05*	.41	1.00	.11
80	.05	.03	.18	*.03*	.31	.77	.08
200	.03	.01	.07	*.01*	.13	.32	.05
20	1.00	.91	3.64	*.88*	6.07	10.18	1.64
200	1.00	.57	2.12	*.45*	3.93	9.63	1.36
k = 3							
$n = 30$	*.42*	.48	.72	.44	1.09	1.57	.70
50	.32	.39	.42	*.27*	.69	1.16	.54
70	.28	.35	.29	*.19*	.53	.94	.45
90	.26	.33	.23	*.16*	.40	.77	.40
210	.22	.29	.09	*.07*	.18	.27	.29
30	1.00	1.14	1.70	*1.03*	2.57	3.70	1.65
210	1.00	1.32	.40	*.32*	.81	1.20	1.29
k = 4							
$n = 40$	*.55*	.61	.95	.68	1.07	1.76	.82
60	*.45*	.52	.68	.46	.82	1.54	.63
80	.42	.49	.59	*.38*	.66	1.50	.55
100	.40	.48	.47	*.29*	.50	1.29	.49
220	.34	.42	.14	*.09*	.22	.67	.32
40	1.00	*1.12*	1.72	1.24	1.94	3.20	1.49
220	1.00	1.24	.43	*.27*	.64	1.98	.96
k = 5							
$n = 50$	*.62*	.69	1.30	.69	1.16	1.94	.87
70	.54	.61	1.08	*.44*	1.00	1.84	.69
90	.49	.56	1.01	*.37*	.89	1.85	.60
110	.45	.52	.93	*.30*	.79	1.83	.51
230	.39	.47	.47	*.14*	.37	1.36	.32
50	1.00	*1.11*	2.10	1.12	1.88	3.13	1.41
230	1.00	1.20	1.19	*.36*	.96	3.49	.83

The results presented in Table II may be summarized as follows. ML and $QE1$ are based on distributional assumptions. As these are not fulfilled their efficiency deteriorates in comparison to the efficiency against the other estimates in the first experiment.

$KMS1$ and $KMS2$ perform reasonably well, especially for n large. The behaviour of SKM and $QE2$ is not too bad, but the KM (and the SKM) estimate behaves extremely bad for $k = 2$.

We also investigated the effect of misspecification for samples from density

$$f_X(x) = \tfrac{1}{2}\phi_{-\sqrt{2/3},1/3}(x) + \tfrac{1}{2}\phi_{\sqrt{2/3},1/3}(x),$$

i.e. a mixture of normals with different means, but equal variances. With respect to the performance of the proposed estimators the results according to $k \geq 3$ where similar to conclusions in the scale mixture setting. This is in contrast to the case $k = 2$ where the k–means estimates KM and SKM had small mean squared errors.

5 REAL DATA – OPTIMAL GROUPING

We use the *Cherry–Tree* data of Hand&Daly (1993), containing a sample of $n = 31$ cherry trees of *Allegheny National Forest, Pennsylvania.* For each tree the *diameter*, the *volume* and the *height* are registered. We look for an optimal grouping of the trees based on the height measurements (in feet) only. The task of optimal grouping or stratification dates back to the contributions of Cox (1950), Dalenius (1950) and Dalenius&Gurney (1950). A proper application may be found in Flury&Tarpey (1993). Figure 2 shows a histogram of the data along with a suitable nonparametric density estimate.

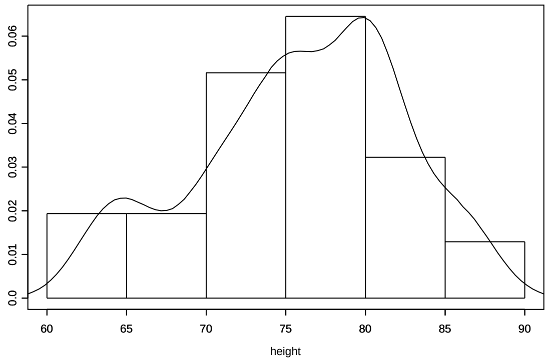

FIG. 2. Histogram of *height* (in feet) along with density estimate.

Table III below contains estimates of principal points for $k = 2, 3, 4$ and the Bootstrap estimates for bias and variance based on $R = 100$ Bootstrap replicates. It suffices to concentrate on the results concerning the estimates ML, KM and $QE2$, since the other ones (listed in columns 2 to 4 of Table III) resemble the outcomes for ML and $QE2$.

TABLE III. Exact and estimated principal points of *height* for $k = 2$ (first block), $k = 3$ (middle block), $k = 4$ (last block). Bootstrap estimates for bias and variance after |.

ML	$QE1$	$KMS1$	$KMS2$	SKM	KM	$QE2$
70.9|.1,1.3	71.2	70.5	70.9	69.3	68.0|1.2,1.4	71.0|.3,1.1
81.0|-.0,1.0	80.7	80.7	81.0	80.9	79.3|1.6,2.4	80.9|-.1,0.8
68.2|.0,1.5	68.5	69.2	69.0	66.1	66.2|0.1, 1.9	68.2|.3,1.5
76.0|.0,1.0	76.0	77.5	76.0	75.5	74.6|0.3,1.7	76.2|.0,1.0
83.7|-.0,1.2	83.4	85.2	82.9	83.5	81.8|0.3,1.5	83.2|-.2,0.9
66.3|.2,1.8	66.7	66.0	66.0	64.6	64.5|0.5,1.6	66.3|.7,1.5
73.1|.1, 1.3	73.2	72.6	72.9	72.6	70.8|1.5,1.7	73.4|.2,1.0
78.8|-.1,1.1	78.7	78.8	79.1	78.8	75.6|2.6,2.1	79.0|.0,0.8
85.6|-.2,1.3	85.2	86.0	86.0	85.3	81.8|2.2,1.9	84.6|.0,1.0

All estimates except KM are of comparable size. The KM estimates are shifted downwards. A positive bias which is even severe for $k = 2$ and $k = 4$, is present for all values of KM. This is in contrast to the rather negligible biases of ML and $QE2$. Additionally, the variances of the KM estimates turn out to be larger than the variances of the alternate methods $QE2$ and ML being of similar order. From this perspective we may prefer the estimate $QE2$ (or alternatively SKM) since it requires only weak assumptions for F_X. The Bootstrap approach yields also confidence intervals for the principal points and a correlation matrix for the corresponding estimates. The confidence intervals may be used as criterion for the final choice of k. Unless they are not clearly different, the grouping result may be regarded as unstable. In our example some of the 95% confidence intervals for $k = 4$ are overlapping and one may argue to use rather $k \leq 3$. The Bootstrap estimate of the correlation matrix shows the expected structure. All estimates are positively correlated and the correlation being higher for neighbouring principal points.

6 SUMMARY

In this paper we review some material on univariate principal points and particular methods for estimating them. It has been demonstrated that the well known nonparametric k–means estimate is rather unreliable in comparison with other methods suggested in this paper. Our quick estimates based on an asymptotic result of Pötzelberger&Felsenstein (1994) and our estimates employing nonparametric density estimation seem to be reasonable alternatives. The nonparametric estimates of the variability that we obtained by Bootstrap methods give sensible results.

BIBLIOGRAPHY

Bhattacharya, R. and Denker, M. (1990). *Asymptotic Statistics*, DMV Seminar, 14, Birkhäuser, Basel.

Cox, D.R. (1950). "Note on grouping," *Journal of the American Statistical Association*, 52, 543–547.

Dalenius, D. (1950). "The problem of optimum stratification," *Skandinavian Aktuarial Journal*, 33, 203–213.

Dalenius, D. and Gurney, M. (1950). "The problem of optimum stratification II," *Skandinavian Aktuarial Journal*, 34, 133–148.

Davison, A.C. and Hinkley, D.V. (1997). *Bootstrap Methods and their Application*, Cambridge University Press, Cambridge.

Flury, D.B. (1990). "Principal points," *Biometrika*, 77, 33–41.

Flury, D.B. (1993). "Estimation of principal points," *Journal of Applied Statistics*, 42, 139–151.

Flury, D.B. and Tarpey, T. (1993). "Representing a large collection of curves: A case for principal points," *The American Statistician*, 47, 304–306.

Hand, D.J. and Daly, F. (1993). *Handbook of Small Data Sets*, Chapman and Hall, London.

Hartigan, J.A. (1974). *Clustering Algorithms*, Wiley, New York.

Hartigan, J.A. (1978). "Asymptotic distributions for clustering criteria," *The Annals of Statistics*, 6, 117–131.

Kieffer, J.C. (1983). "Uniqueness of locally optimal quantizer for log-concave density and convex error weighting function," *IEEE Transactions on Information Theory*, 29, 42–47.

Kinderman, A.J. and Monahan, J.F. (1977). "Computer generation of random variables using the ratio of uniforms deviates," *ACM Transactions on Mathematical Software*, 3, 257–260.

Pfeiffer, E.P. (1990). *Probability for Applications*, Springer, New York.

Pollard, D. (1981). "Strong consistency of k–means clustering," *The Annals of Statistics*, 9, 135–140.

Pollard, D. (1982). "A central limit theorem for k–means clustering," *The Annals of Probability*, 10, 919–926.

Pötzelberger, K. and Felsenstein, K. (1994). "An asymptotic result on principal points for univariate distributions," *Optimization*, 28, 397–406.

Silverman, B. W. (1986). *Density Estimation for Statistics and Data Analysis*, Chapman and Hall, London.

Stampfer, E. (2000). Prototypen von Verteilungen, PhD Dissertation, 168 pp., Technische Universität Graz, Austria, accessible by `www.cis.tu-graz.ac.at/stat/dthesis/stam00.zip`.

Tarpey, T. and Flury, B. (1996). "Self–consistency: A fundamental concept in statistics," *Statistical Science*, 11, 229–243.

Tarpey, T. (1997). "Estimating principal points of univariate distributions," *Journal of Applied Statistics*, 24, 499–512.

Tarpey, T. (1999). "Self–consistency algorithms," *Journal of Computational and Graphical Statistics*, 8, 889–905.

Venables, W.N. and Ripley, B.D. (1999). *Modern Applied Statistics with S-PLUS*, 3^{rd} edition, Springer, New York.

Zoppe, A. (1995). "Principal points of univariate continuous distributions," *Statistics and Computing*, 5, 127–132.