Computational Statistics and Data Analysis 93 (2016) 456–468

Contents lists available at ScienceDirect

Computational Statistics and Data Analysis

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

# Minimum volume peeling: A robust nonparametric estimator of the multivariate mode

T. Kirschstein [a,*], S. Liebscher [a,1], G.C. Porzio [b,2], G. Ragozini [c,3]

[a] Martin-Luther-University Halle-Wittenberg, Gr. Steinstrasse 73, D-6099 Halle (Saale), Germany

[b] University of Cassino and Southern Lazio, Via San Angelo, localita Folcara, I-03043 Cassino, Italy

[c] University of Naples Federico II, Via L. Rodinò 22, I-80138, Napoli, Italy

ARTICLE INFO

*Article history:*
Received 15 May 2014
Received in revised form 29 April 2015
Accepted 30 April 2015
Available online 18 May 2015

*Keywords:*
Robust mode estimation
Skewed distributions
Subset selection
Convex hull

ABSTRACT

Among the measures of a distribution's location, the mode is probably the least often used, although it has some appealing properties. Estimators for the mode of univariate distributions are widely available. However, few contributions can be found for the multivariate case. A consistent direct multivariate mode estimation procedure, called *minimum volume peeling*, can be outlined as follows. The approach iteratively selects nested subsamples with a decreasing fraction of sample points, looking for the minimum volume subsample at each step. The mode is then estimated by calculating the mean of all points in the final set. The robustness of the method is investigated by analyzing its finite sample breakdown point and algorithms to determine minimum volume sets are discussed. Simulation results confirm that using minimum volume peeling leads to efficient mode estimates both in uncontaminated as well as contaminated situations.



## 1. Introduction

Three concurrent measures of location are offered to students within standard statistical textbooks: the arithmetic mean, the median, and the mode. Amongst them, most of the pages are devoted to the mean, some to the median, and few (if any) to the mode. Accordingly, Dalenius (1965) called the mode *the most neglected* parameter some 50 years ago.

However, the mode has many relevant properties that have been well-known in the literature for years. Most importantly, the mode is an appropriate location parameter for skewed distributions. This is why the mode finds application in many different fields where nonnormal or skewed distributions occur. For instance, it is used in biology, medicine, astronomy, and computer sciences (see e.g. the references listed in Hedges and Shah, 2003, and Schwartz and Rozumalski, 2005). Moreover, the mode is a convenient location parameter for truncated distributions. This led Lee (1989) to introduce mode regression where the conditional mode of a response variable is investigated. Finally, it should be mentioned that the value that most likely occurs is an interesting parameter by itself.

Estimating the mode has been extensively studied in the univariate case, and two main approaches (dating back to the 1960s) are available. Parzen (1962) introduced the *indirect* approach by suggesting as mode estimator the value that

* Corresponding author. Tel.: +49 345 55 23424.
*E-mail addresses:* thomas.kirschstein@wiwi.uni-halle.de (T. Kirschstein), steffen.liebscher@wiwi.uni-halle.de (S. Liebscher), porzio@unicas.it (G.C. Porzio), giancarlo.ragozini@unina.it (G. Ragozini).
1 Tel.: +49 345 55 23382.
2 Tel.: +39 0776 299 3448.
3 Tel.: +39 081 2537460.

http://dx.doi.org/10.1016/j.csda.2015.04.012
0167-9473/

maximizes an appropriate nonparametric estimate of the distribution density. On the other hand, *direct* estimators are also available. Chernoff (1964) suggested considering the center of an interval of given length containing the most observations. Venter (1967), instead, suggested adopting as estimator the midpoint of the smallest interval containing a given proportion of data points. Since then, many other authors have presented estimators within these two classes and/or studied their properties. Recently, Bickel and Frühwirth (2006) studied the robustness properties of a univariate Venter-type estimator that turned out to have a 50% breakdown value.

In contrast, few contributions have been made to estimate the mode of a multivariate distribution. An early overview on multivariate mode estimation can be found in Sager (1983). Adopting the indirect approach, Rüschendorf (1977) introduced the idea of using a density estimator, with further developments in Abraham et al. (2003). Recently, Jing et al. (2012) proposed an indirect estimation approach based on polynomial histograms. Some 35 years ago, Sager (1979) defined a Venter-type estimator that exploits an iterative peeling procedure to find the multivariate region with minimum volume. Unfortunately, at that time no efficient algorithms were at hand to determine a minimum volume set. Recently, Hsu and Wu (2013) provided an indirect multivariate mode estimator by extending the univariate estimator of Bickel (2003) to the multivariate setting. This estimator works well for multivariate distributions that can be normalized through a Box–Cox transformation and, in this respect, it is a parametric procedure. It consistently estimates the mode of multivariate log-normal distributions, but fails whenever the Box–Cox transformation does not yield a distribution close to a multivariate normal (see e.g. Hernandez and Johnson, 1980). In addition, it is not a robust estimator. For this reason, *minimum volume peeling*, a nonparametric direct multivariate mode estimation procedure based on Sager's idea is described whose consistency is shown by Sager (1979). Furthermore, it is proven that minimum volume peeling is also highly robust.

The paper is organized as follows. In the next section Sager's estimation procedure is described from a peeling perspective, whereas its robustness properties are studied in Section 3. Section 4 describes approaches to determine a subset of given size with minimum volume. One is tailor-made for Sager's procedure and minimizes the volume of the subset's convex hull. The others rely on minimizers of elliptic bodies (Lopuhaä and Rousseeuw, 1991), namely the Minimum Covariance Determinant estimator (MCD) and Minimum Volume Ellipsoid estimator (MVE). Section 5 contains the results of a simulation study providing insights into the performance of both approaches under various conditions. A brief conclusion finishes the paper.

## 2. Sager's multivariate mode estimation procedure

A point $\boldsymbol{\theta} \in \mathbb{R}^d$ is said to be the mode of a $d$-variate continuous density function $f$ if $f(\mathbf{X}) < f(\boldsymbol{\theta}), \forall\, \mathbf{X} \neq \boldsymbol{\theta}, \mathbf{X} \in \mathbb{R}^d$. If $f$ is a strongly unimodal distribution with unique mode $\boldsymbol{\theta}$, it follows that $f$ decreases along any ray emanating from $\boldsymbol{\theta}$ and that the greatest concentration of probability occurs around it (see e.g. Sager, 1979).

A direct estimator of the multivariate mode $\boldsymbol{\theta}$ is any point $\hat{\boldsymbol{\theta}}$ that belongs either to the set with the maximum fraction of points in a given volume (the most dense set) or to the minimum volume set containing a given fraction of points. Focusing on the latter idea, the main issue becomes to efficiently and correctly select such a set. Sager's procedure to estimate the mode can be summarized by the following steps.

1. Given a data set, look for the subset of a certain size with minimal volume.
2. Discard the points not belonging to this set.
3. Repeat steps 1–2 until a *sufficiently* small subset remains.
4. A point estimate of the mode is the arithmetic mean of the points selected in the last iteration.

It is worth noting that this *minimum volume peeling* procedure is similar to the contemporarily published convex hull peeling (Barnett, 1976). There, a sequence of nested sets is determined by successively deleting the convex hull (CH) of the preceding set. However, while convex hull peeling leads to the multivariate median, minimum volume peeling yields the multivariate mode.

Formally, let $\mathcal{X} = (\mathbf{X}_1, \mathbf{X}_2, \ldots, \mathbf{X}_n)$ be a random sample of size $n$ from a $d$-variate continuous distribution $f$ with mode $\boldsymbol{\theta}$, and let $0 < p < 1$. Assume all points are in general position.

At Step 1, given the fraction of points $n_1 = \lfloor n \cdot p \rfloor$, select a subset $\mathcal{X}_1^* \subset \mathcal{X}$, with $\mathcal{X}_1^* = \underset{\mathcal{X}_1^j}{\arg\min}\, CHV(\mathcal{X}_1^j)$, where $\mathcal{X}_1^j$ is any subset of size $n_1$ of $\mathcal{X}$, and $CHV(\mathcal{A})$ is the volume of the convex hull of the set $\mathcal{A}$. At Step 2, $\mathcal{X}_1^*$ is retained and further peeled.

Thus, repeating steps 1 and 2, at the $k$-th iteration, the points $\mathbf{X}_i \in \mathcal{X}_k^*$ are selected such that

$$\mathcal{X}_k^* = \underset{\mathcal{X}_k^j:\, \mathcal{X}_k^j \subset \mathcal{X}_{k-1}^* \wedge |\mathcal{X}_k^j| = n_k}{\arg\min} CHV(\mathcal{X}_k^j),$$

where $n_k = \left\lfloor n_{(k-1)} \cdot p \right\rfloor$. The peeling procedure stops when the subset size $n_k$ degenerates to less than $d+1$ points. This is because all subsets containing less than $d+1$ points possess a unique convex hull volume of 0. Hence, the algorithm stops if $\lfloor n_k \cdot p \rfloor < d+1$.