TRIMMED KERNEL MEAN SHIFT

Luca Greco[1] , Giovanna Menardi[2] and Marco Rudelli[2]

[1] University Giustino Fortunato - Benevento
(e-mail: l.greco@unifortunato.eu

[2] Department of Statistical Sciences, University of Padova,
(e-mail: menardi@stat.unipd.it, marco.rudelli@studenti.unipd.it)

ABSTRACT: A robust procedure based on impartial trimming is discussed, aimed to protect nonparametric clustering stemming from kernel mean shift from the deleterious effect of outliers.

KEYWORDS: Clustering, Density estimation, Outliers

1 Introduction

The problem of data contamination, where unexpected points that do not share the pattern of the majority of the data are observed, is known to possibly hinder the validity of inferential procedures. The issue is even more critical in clustering, where the lack of a reference ground-truth to aim at makes even the simplest problem an ill-posed one. Genuine observations forming small clusters can be mistaken with outliers (*swamping*); on the other side, outlying data lying close to each other just by chance can form spurious clusters (*masking*). Moreover, in this setting it is quite difficult to state a working notion of outliers, and robustness is not only data dependent, but rather cluster dependent (Hennig, 2008), which is itself often arbitrary. It then looks clear how contaminated data can compromise or even invalidate unsupervised techniques.

A large amount of work has been done to define robust clustering strategies in the mainstream approaches within the distance- and the model-based approach (see Farcomeni & Greco, 2016, for a review). Conversely, the issue has been largely neglected in the nonparametric framework, where clusters are identified as the domains of attractions of the modes of the underlying density (Stuetzle, 2003). The correspondence between groups and modal regions entails some reasons of attractiveness: clusters are not constrained to predetermined shapes, and resorting to nonparametric methods keeps this flexibility; additionally, the number of clusters is inherent of the data density, hence determined as part of the estimation procedure (see, Menardi, 2016, for a review). However, these very same properties turn out to be pitfalls of nonparametric

methods in the presence of outliers. Actually, outliers can produce spurious modes. In the presence of spurious modes, outliers self-validate themselves, as they can not be declared unlikely with respect to the cluster they have given birth to. Finally: how can one say what is unlikely, with respect to a cluster which can take any shape? In the following, a robust-to-outliers counterpart of the Kernel Mean Shift (KMS, Fukunaga & Hostetler, 1975) for modal detection is discussed, based on an outlyingness criterion specifically designed for the considered framework.

2 Methodology

Let $\mathcal{X} = (x_1, x_2, \ldots, x_n)$ be a sample of size n, with $x_i \in \mathbb{R}^d$, $d \geq 1$. A kernel density estimator is given by $\hat{f}(x) = \frac{1}{n}\sum_{i=1}^{n} K_H(x - x_i)$ where $K_H(x)$ is a d-variate kernel function scaled by a symmetric positive definite $d \times d$ bandwidth matrix H. KMS is an iterative algorithm to identify modal clusters from a kernel density estimate of a set of data. The algorithm recursively shifts each data point to a local weighted mean $m_{K,H}$,

$$m_{K,H}\left(x^{(j)}\right) = x^{(j+1)} - x^{(j)} = \frac{\sum_{i=1}^{n} x_i^{(j)} \nabla K_H(x)\left(x - x_i^{(j)}\right)}{\sum_{i=1}^{n} \nabla K_H(x)\left(x - x_i^{(j)}\right)} \propto \frac{\nabla \hat{f}\left(x_i^{(j)}\right)}{\hat{f}\left(x_i^{(j)}\right)} .$$

until convergence. The weights are normalized gradient vectors of the kernel function. Hence, the mean shift is a gradient ascent algorithm based on a normalised kernel estimator of the gradient.

We propose a robust counterpart of KMS based on impartial trimming (Cuesta-Albertos *et al.*, 1997). The methodology, summarised in Algorithm 1, climbs iteratively via KMS the modes of a trimmed kernel density estimate, obtained by discarding at each iteration a fixed proportion α of data with the lowest densities with respect to the pertaining cluster. Then, the identified clusters allow to update the outlyingness score of each observation and run KMS on a renewed active set. Iterations stop as the trimmed set is not updated. The procedure is impartial since the detection of the trimmed points is a result of the procedure jointly with cluster assignments and it recasts to a trimmed KMS (tKMS). The initial active subset $I^{(0)}$ can be obtained as follows: (a) consider an over-smoothed fitted density; (b) select a proportion of points with the largest fitted densities.

Algorithm 1 Iteration r of tKMS

Optimization Step

Evaluate the kernel density estimate over the active set $I^{(r)}$ of size $n - \lfloor n\alpha \rfloor$

$$\hat{f}^{(r)}(x) = \frac{1}{n - \lfloor n\alpha \rfloor} \sum_{i \in I^{(r)}} K_H(x - x_i)$$

Run KMS to identify the modes of $\hat{f}^{(r)}(x)$, and get a partition of $\mathcal{X}$ in clusters $\left\{ C_m^{(r)} \right\}_m$, each with cardinality $n_m^{(r)}$

Let $m = m_i$ if $x_i \in C_m^{(r)}$

Trimming Step

Compute $\hat{g}_i^{(r)} = \hat{g}_{m_i}^{(r)}(x_i), i = 1, 2, \ldots, n$ with

$$\hat{g}_m^{(r)}(x) = \frac{1}{n_m^{(r)}} \sum_{x_i \in C_m^{(r)}} K_H(x - x_i)$$

Update $I^{(r+1)}$ by ruling out from $\mathcal{X}$ the $\lfloor n\alpha \rfloor$ points with the lowest of $\hat{g}_i^{(r)}$.

3 Examples

We illustrate the effectiveness of the proposed methodology, as well as the drawbacks of classical KMS in the presence of contamination, through some synthetic examples. Figure 1 gives the results from running both KMS and tKMS on a pair of bivariate data structured in three clusters in the presence of background noise. While essentially identifying the true clusters, in both examples KMS also detects spurious modes, wheres tKMS recovers the underlying clustering structure and trimmed points are not assigned to any cluster.

References

CUESTA-ALBERTOS, J. A., GORDALIZA, A., & MATRÁN, C. 1997. Trimmed k-means: an attempt to robustify quantizers. *The Annals of Statistics*, **25**(2), 553–576.

FARCOMENI, A., & GRECO, L. 2016. *Robust methods for data reduction*. CRC press.

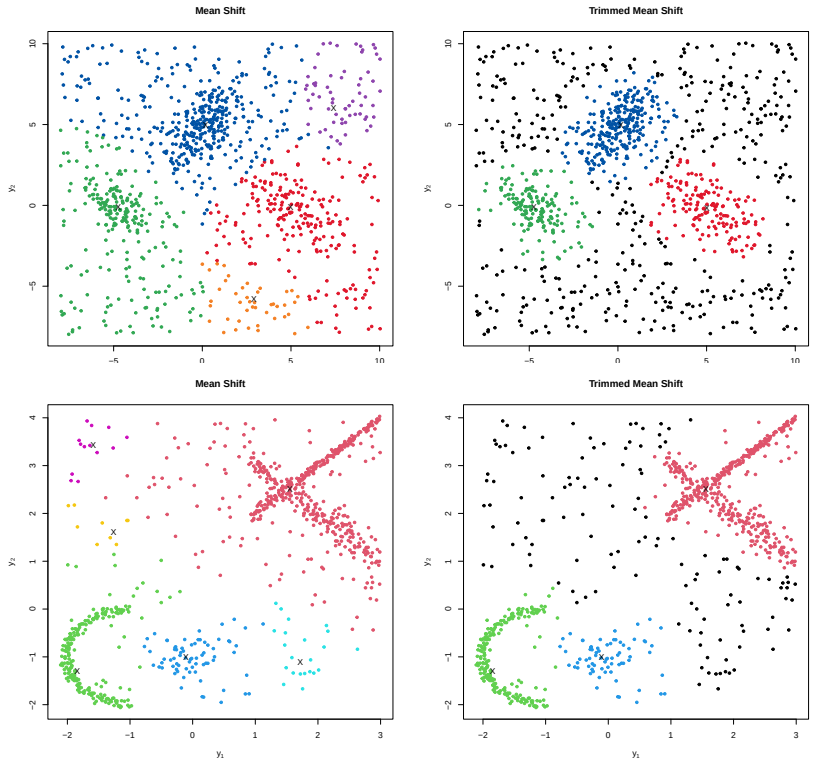

Figure 1. *Classification from KMS (left) and tKMS (right) for two synthetic data exhibiting three variously shaped clusters (one for each row). The identified clusters are denoted by different colors, while the estimated modes are denoted by X. Trimmed points are identified in black.*

FUKUNAGA, K., & HOSTETLER, L. 1975. The estimation of the gradient of a density function, with applications in pattern recognition. *IEEE Transactions on information theory*, **21**(1), 32–40.

HENNIG, C. 2008. Dissolution point and isolation robustness: robustness criteria for general cluster analysis methods. *Journal of multivariate analysis*, **99**(6), 1154–1176.

MENARDI, G. 2016. A review on modal clustering. *International Statistical Review*, **84**(3), 413–433.

STUETZLE, WERNER. 2003. Estimating the cluster tree of a density by analyzing the minimal spanning tree of a sample. *Journal of classification*, **20**(1), 25–47.