Distance estimation in clustering

Distance estimation

In [1], distance estimation for squared Euclidean distances was presented. Expected squared euclidean distance (ESD) assumes that data vectors are multivariate Gaussian or a mixture of Gaussian distributed [1, 2]. The assumption is not strict because the central limit theorem states that nearly any continuous distribution can crudely approximate the normal distribution [1]. Further, the Missing-at-Random (MAR) assumption is followed, meaning that the missingness can be accounted for by observed variables. The computation of ESD requires estimating the conditional covariance matrix and conditional mean vector from a distribution where missing values are conditioned with observed ones. Parameters can be estimated using the expectation maximization algorithm (EM) with a maximum negative log-likelihood convergence criterion. Regarding to the multivariate Gaussian distributed data, the detailed equations (Eq.(1)-(9)) and steps (Algorithm 1) are given in [1]. In [3], the computation of the Expected Euclidean distance (EED) is presented. The computation of the EED follows the same basic procedure as given in [1]. However, variances between pairs of data vectors are needed because Gamma distributed Nakagami model requires variance as an input parameter (see details in [3]). The variance is obtainable using non-central moments of normal distribution. Note that weights are unnecessary if data is assumed to follow multivariate Gaussian instead of a mixture of Gaussian. The EED is an output value of the Nakagami model.

Prototype-based clustering with distance estimation

The prototype-based clustering consists of two main steps, an initialization step, and a local refinement step. The initial locations of the prototypes are selected during the initialization step. The K-means++ algorithm is commonly used in initialization. The algorithm uses distances to previously selected prototypes for computing new prototypes. In the local refinement step, prototypes are updated according to the membership of associated data vectors in a cluster. The steps of prototype-based clustering with distance estimation are given in Algorithm 1.

Algorithm 1. Clustering based on distance estimation

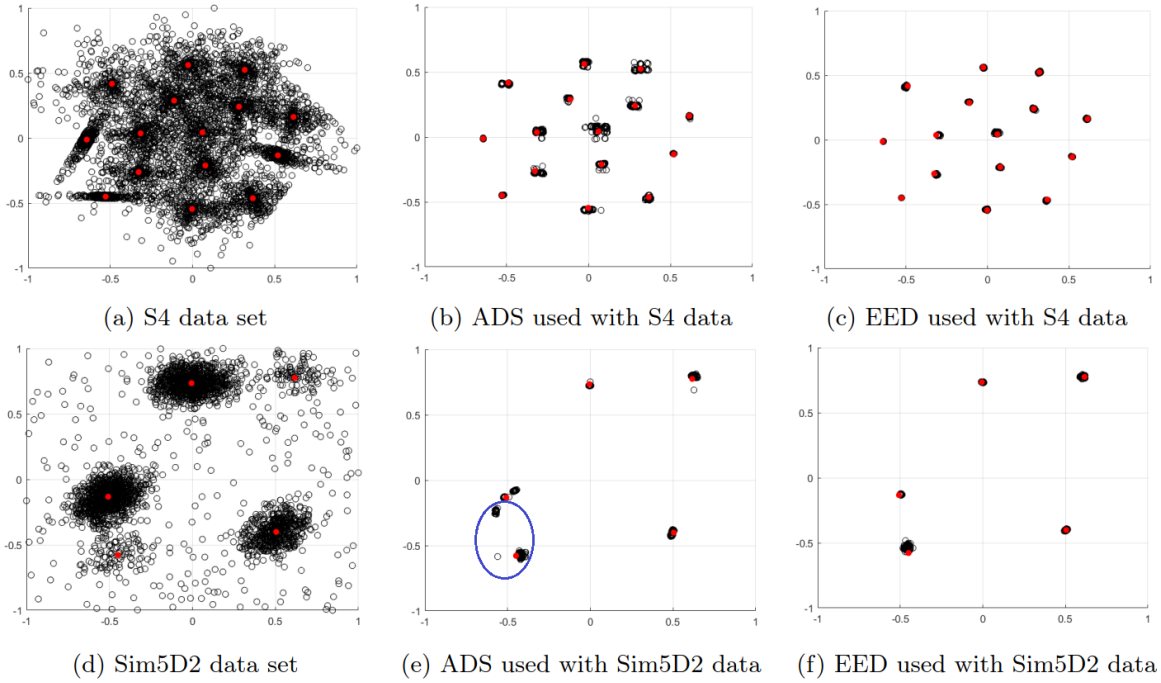
1. Select initial prototypes using the K-means++ algorithm.
2. Compute the mean vector and covariance matrix of the conditional multivariate distribution using the EM method with the maximum negative log-likelihood criterion.
3. Associate data vectors to the nearest prototypes using distance estimation
4. Use cluster centers as new prototypes.
5. Repeat steps 3 and 4 until the locations of prototypes do not change.

1 Results

Table 1 shows that the root mean square errors (RMSEs) to the real centroids are lower with the K-spatial-medians with EED-based distance estimation compared to the K-spatial-medians with ADS-based estimation. The best results are given in bold. Experiments were performed using 100 repetitions of the replicated clustering (with 100 replicates) and regeneration of missing values (in total 20 % missing values). In [4], cluster validation indices recommended more accurate results when the obtained prototypes based on distance estimation were used as starting points for the K-spatial-medians clustering based on ADS. Figure 1 shows how clustered prototypes, based on repeated clustering, are distributed around the real centroids in S4 and Sim5D2 data sets. Clearly, EED-based results are more compactly divided around the real centroids. Centroids which belong to bottom left side in Figure 1 (e) are illustrated using a blue ellipse.

Table 1: RMSEs between real centroids and clustered centroids

	S1	S2	S3	S4	Sim5D2	Sim2D2	O200	O2000
ADS	0.005	0.005	0.011	0.011	0.030	0.048	0.046	0.031
(std)	(0.006)	(0.001)	(0.002)	(0.002)	(0.026)	(0.004)	(0.010)	(0.004)
EED	**0.002**	**0.003**	**0.005**	**0.005**	**0.012**	**0.023**	**0.022**	**0.005**
(std)	**(0.000)**	**(0.000)**	**(0.001)**	**(0.001)**	**(0.002)**	**(0.004)**	**(0.010)**	**(0.001)**
EED-ADS	0.004	0.005	0.009	0.009	0.014	0.048	0.041	0.027
(std)	(0.000)	(0.000)	(0.001)	(0.001)	(0.002)	(0.004)	(0.009)	(0.003)



(a) S4 data set (b) ADS used with S4 data (c) EED used with S4 data

(d) Sim5D2 data set (e) ADS used with Sim5D2 data (f) EED used with Sim5D2 data

Figure 1: Clustering results of repeated clustering for two synthetic data sets using K-spatial-median clustering with and without distance estimation.

References

[1] E. Eirola, G. Doquire, M. Verleysen, and A. Lendasse. Distance estimation in numerical data sets with missing values. *Information Sciences*, 240:115–128, 2013.

[2] E. Eirola, A. Lendasse, V. Vandewalle, and C. Biernacki. Mixture of gaussians for distance estimation with missing data. *Neurocomputing*, 131:32–42, 2014.

[3] D. P. P. Mesquita, J. P. P. Gomes, A. H. Souza Junior, and J. S. Nobre. Euclidean distance estimation in incomplete datasets. *Neurocomputing*, 248:11–18, 2017.

[4] M. Niemelä and T. Kärkkäinen. Improving clustering and cluster validation with missing data using distance estimation methods. In T. Tuovinen, J. Periaux, and P. Neittaanmäki, editors, *Computational Sciences and Artificial Intelligence in Industry: New Digital Technologies for Solving Future Societal and Economical Challenges*, pages 123–133. Springer International Publishing, 2022.