NOVEL BAYESIAN CLUSTER ENUMERATION CRITERION FOR CLUSTER ANALYSIS WITH FINITE SAMPLE PENALTY TERM

Freweyni K. Teklehaymanot[1,2], Michael Muma[1], Abdelhak M. Zoubir[1,2]

[1] Signal Processing Group, Technische Universität Darmstadt, {ftekle, muma, zoubir}@spg.tu-darmstadt.de [2] Graduate School CE, Technische Universität Darmstadt, teklehaymanot@gsc.tu-darmstadt.de

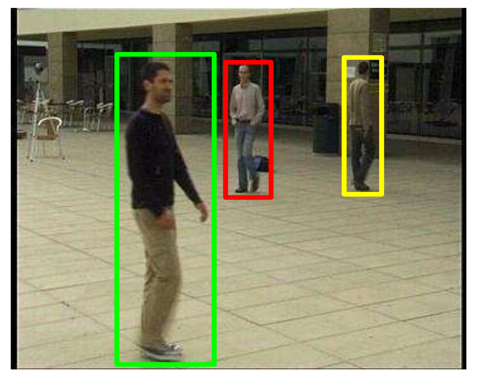
TECHNISCHE UNIVERSITÄT DARMSTADT

1 Motivation

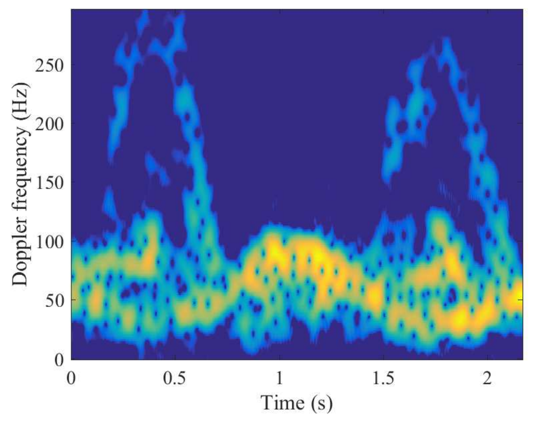
- Bayesian Information Criterion (BIC): extensively used in clustering.
- Prior to [1] BIC, for clustering, has not been derived from first principles.
- Finite-sample performance of asymptotic criterion is not satisfactory.

2 Contributions

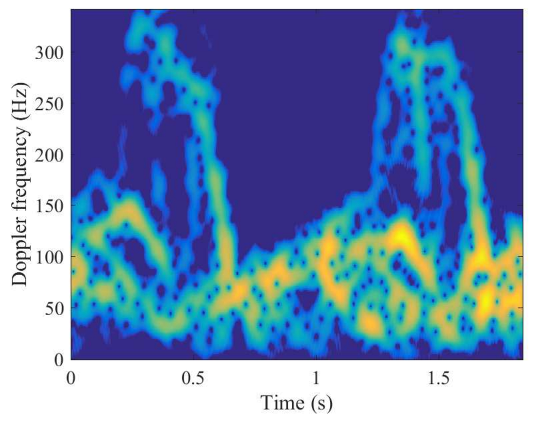
- Finite-sample BIC for clustering derived from first principles.
- Proposed criterion ($\mathrm{BIC}_{\mathrm{NF}}$) incorporates structure of the clustering problem and chooses model with maximum posterior probability.
- $\mathrm{BIC}_{\mathrm{NF}}$ successfully applied to camera networks [2] and radar [3].



Labeling humans in camera networks



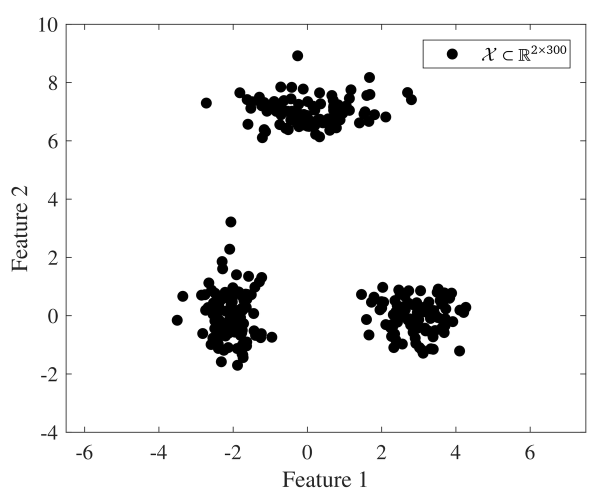
Doppler signature of Person A

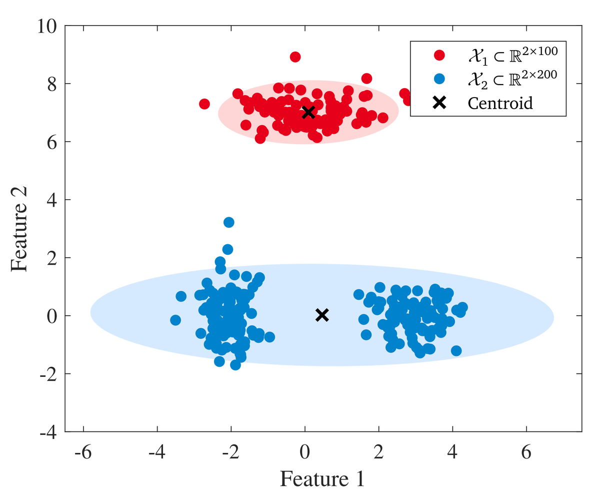


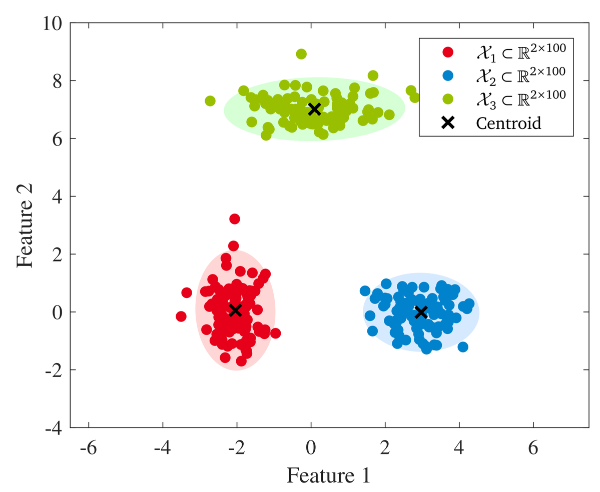
Doppler signature of Person B

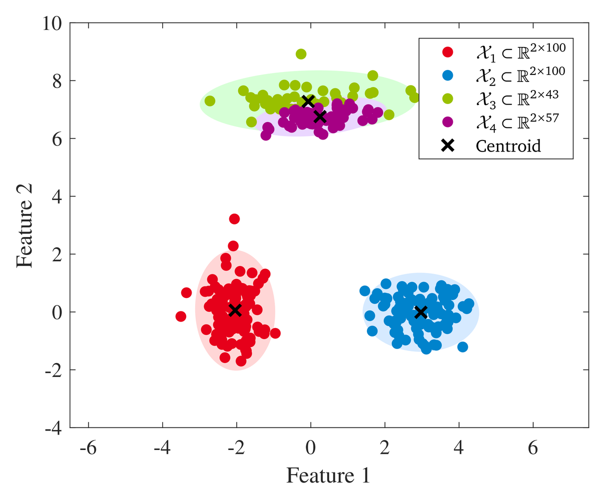
3 Problem Formulation

- $\boldsymbol{x}_k \sim \mathcal{N}(\boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k)$, $k \in \mathcal{K} \triangleq \{1,\ldots,K\}$: i.i.d Gaussian random variables; K: number of clusters; $\boldsymbol{\mu}_k \in \mathbb{R}^{r\times 1}$ and $\boldsymbol{\Sigma}_k \in \mathbb{R}^{r\times r}$: centroid and covariance matrix of the kth cluster
- $\mathcal{X}_k \subset \mathbb{R}^{r\times N_k}, k \in \mathcal{K}$: cluster containing realizations of $\boldsymbol{x}_k$
- $\mathcal{X} \triangleq \{\mathcal{X}_1,\ldots,\mathcal{X}_K\} \subset \mathbb{R}^{r\times N}$: observed data set; $N=\sum_{k=1}^{K} N_k$
- $\mathcal{M} \triangleq \{M_{L_{\min}},\ldots,M_{L_{\max}}\}$: family of candidate models
- $\mathcal{X}$ partitioned into $l = L_{\min},\ldots,L_{\max}$ clusters, using model M_l with parameters $\boldsymbol{\Theta}_l = [\boldsymbol{\theta}_1,\ldots,\boldsymbol{\theta}_l]$
- **Goal:** estimate the number of clusters in $\mathcal{X}$ given $\mathcal{M}$









4 Proposed Bayesian Cluster Enumeration Criterion With Finite Sample Penalty Term

- **Main idea:** maximization of posterior probability of candidate models $M_l \in \mathcal{M}$ given $\mathcal{X}$

$$M_{\hat{K}} = \underset{\mathcal{M}}{\arg\max}\ \log p(M_l|\mathcal{X})$$

$p(M_l|\mathcal{X})$: posterior probability of M_l given $\mathcal{X}$; $\hat{K}$: estimated number of clusters in $\mathcal{X}$

- New BIC for clustering derived from first principles in [1]: applicable to broad class of data distributions

$$\mathrm{BIC}_{\mathrm{G}}(M_l) \triangleq \log p(M_l|\mathcal{X}) \approx \log \mathcal{L}(\hat{\boldsymbol{\Theta}}_l|\mathcal{X}) - \frac{1}{2}\sum_{m=1}^{l}\log\left|\hat{\boldsymbol{J}}_m\right|$$

$\mathcal{L}(\hat{\boldsymbol{\Theta}}_l|\mathcal{X})$: likelihood function; $\hat{\boldsymbol{J}}_m$: Fisher information matrix of observations from the mth cluster; $|\cdot|$: determinant

$$\hat{\boldsymbol{J}}_m \triangleq -\left.\frac{d^2 \log \mathcal{L}(\boldsymbol{\theta}_m|\mathcal{X}_m)}{d\boldsymbol{\theta}_m d\boldsymbol{\theta}_m^\top}\right|_{\boldsymbol{\theta}_m=\hat{\boldsymbol{\theta}}_m} \in \mathbb{R}^{q\times q}$$

- Special case of [1]: $\mathcal{X}$ is distributed as multivariate Gaussian

$$\mathrm{BIC}_{\mathrm{N}}(M_l) = \sum_{m=1}^{l} N_m \log N_m - \sum_{m=1}^{l}\frac{N_m}{2}\log\left|\hat{\boldsymbol{\Sigma}}_m\right| - \frac{q}{2}\sum_{m=1}^{l}\log N_m$$

$\hat{\boldsymbol{\Sigma}}_m$: estimate of the covariance matrix of the mth cluster; N_m: number of data points in the mth cluster

- **Contribution:** derivation of the penalty term for the finite-sample regime

$$\mathrm{BIC}_{\mathrm{NF}}(M_l) = \mathrm{BIC}_{\mathrm{N}}(M_l) + \frac{1}{4}r(r+1)l\log 2 + \frac{1}{2}\sum_{m=1}^{l}\log\left|\hat{\boldsymbol{\Sigma}}_m\right| - \frac{1}{2}\sum_{m=1}^{l}\log\left|\boldsymbol{D}^\top\hat{\boldsymbol{F}}_m\boldsymbol{D}\right|$$

$$\hat{\boldsymbol{J}}_m = \begin{bmatrix} N_m\hat{\boldsymbol{\Sigma}}_m^{-1} & \boldsymbol{0}_{r\times\frac{1}{2}r(r+1)} \\ \boldsymbol{0}_{\frac{1}{2}r(r+1)\times r} & \frac{N_m}{2}\boldsymbol{D}^\top\hat{\boldsymbol{F}}_m\boldsymbol{D} \end{bmatrix}$$

$\boldsymbol{D} \in \mathbb{R}^{r^2\times\frac{1}{2}r(r+1)}$: duplication matrix; $\hat{\boldsymbol{F}}_m \triangleq \hat{\boldsymbol{\Sigma}}_m^{-1}\otimes\hat{\boldsymbol{\Sigma}}_m^{-1} \in \mathbb{R}^{r^2\times r^2}$

- Estimated number of clusters

$$\hat{K}_{\mathrm{BIC}_{\mathrm{NF}}} = \underset{l=L_{\min},\ldots,L_{\max}}{\arg\max}\ \mathrm{BIC}_{\mathrm{NF}}(M_l)$$

5 Results

- Data-1: Gaussian data set with $K=5$ clusters; Data-2: Gaussian data set with $K=6$ clusters
- $L_{\min}=1$ and $L_{\max}=2K$
- $\mathrm{BIC}_{\mathrm{O}}$: the original BIC

$$\mathrm{BIC}_{\mathrm{O}}(M_l) = \sum_{m=1}^{l} N_m\log N_m - \sum_{m=1}^{l}\frac{N_m}{2}\log\left|\hat{\boldsymbol{\Sigma}}_m\right| - \frac{ql}{2}\log N$$

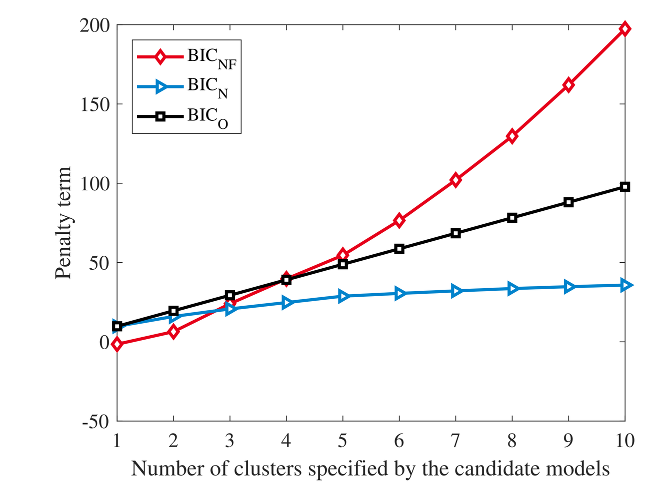
- $p_{\mathrm{det}} = \frac{1}{\mathrm{MC}}\sum_{s=1}^{\mathrm{MC}}\mathbb{1}_{\{\hat{K}_s=K\}}$, p_{det}: empirical probability of detection; MC: number of Monte Carlo experiments; $\mathbb{1}_{\{.\}}$: indicator function

Results for Data-1

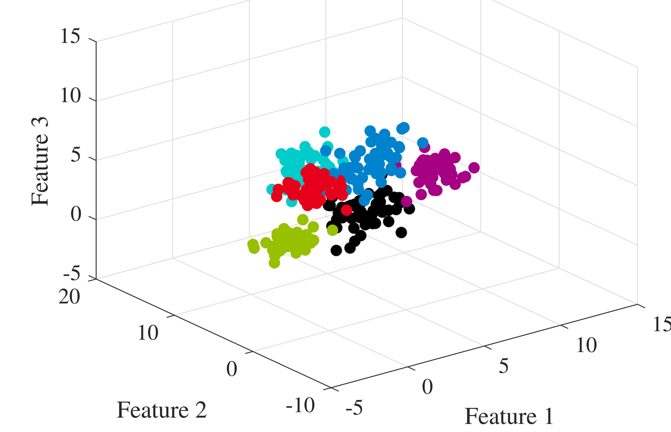
	N_k	10	50	100	1000
$p_{\mathrm{det}}(\%)$	$\mathrm{BIC}_{\mathrm{NF}}$	**77.6**	**100**	**100**	**100**
	$\mathrm{BIC}_{\mathrm{N}}$	0	77.8	96.2	**100**
	$\mathrm{BIC}_{\mathrm{O}}$	26.4	99.3	99.7	**100**

Results for Data-2

	N_k	50	100	250	1000
$p_{\mathrm{det}}(\%)$	$\mathrm{BIC}_{\mathrm{NF}}$	**82.1**	**96.7**	**98.7**	**99.3**
	$\mathrm{BIC}_{\mathrm{N}}$	64.7	92.9	98.1	**99.3**
	$\mathrm{BIC}_{\mathrm{O}}$	51.7	91.1	**98.7**	**99.3**



Penalty term for Data-1 when $N_k=10$



Data-2 when $N_k=50$

References

[1] F. K. Teklehaymanot, M. Muma, and A. M. Zoubir, "A novel Bayesian cluster enumeration criterion for unsupervised learning," *IEEE Trans. Signal Process. (under review)*, [Online-Edition: https://arxiv.org/abs/1710.07954v2], 2018.

[2] F. K. Teklehaymanot, M. Muma, and A. M. Zoubir, "Diffusion-Based Bayesian Cluster Enumeration in Distributed Sensor Networks," *IEEE Statist. Signal Process. Workshop (SSP) (accepted)*, 2018.

[3] F. K. Teklehaymanotand, A.-K Seifert, M. Muma, M. G. Amin, and A. M. Zoubir, "Bayesian Target Enumeration and Labeling Using Radar Data of Human Gait," *26th Eur. Signal Process. Conf. (EUSIPCO) (under review)*, 2018.

[4] F. Fleuret, J. Berclaz, R. Lengagne, and P. Fua, "Multicamera people tracking with a probabilistic occupancy map," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 30, pp. 267–282, 2018.

[5] F. K. Teklehaymanot, M. Muma, and A. M. Zoubir, "Robust Bayesian cluster enumeration criterion for unsupervised learning," under review, 2018.

download full paper here:

download Matlab codes here:

The work of F. K. Teklehaymanot is supported by the 'Excellence Initiative' of the German Federal and State Governments and the Graduate School of Computational Engineering at Technische Universität Darmstadt. The work of M. Muma is supported by the 'Athene Young Investigator Programme' of Technische Universität Darmstadt.