Goodness-of-fit testing of the distribution of posterior classification probabilities for validating model-based clustering

Salima El Kolei[1] and Matthieu Marbac[2]

[1]: Univ.  , Ensai, CNRS, CREST-UMR 9194, 35000 , Rennes, France
[2]: Univ. Bretagne Sud, UMR CNRS 6205, LMBA, F-56000 Vannes, France.

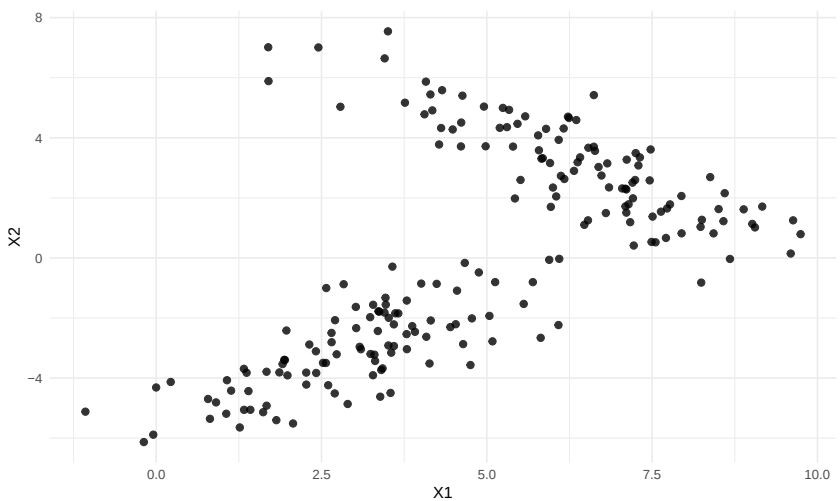
Rochebrune 2026

Geometric clustering and its limits

Data:

- Data to cluster: n independent realizations of $\mathbf{X} \in \mathbb{R}^J$.
- Target: estimating the partition among the individuals based on the observed data.

Goal of clustering:
Find a partition $\mathbf{z} = (z_1, \ldots, z_n)$ of n individuals into K groups.



K-means approach:
Minimize the intra-cluster inertia

$$\mathcal{C}(z, \boldsymbol{\mu}) = \sum_{k=1}^{K} \sum_{i:z_i=k} d(\boldsymbol{x}_i, \boldsymbol{\mu}_k).$$

Main limitations:

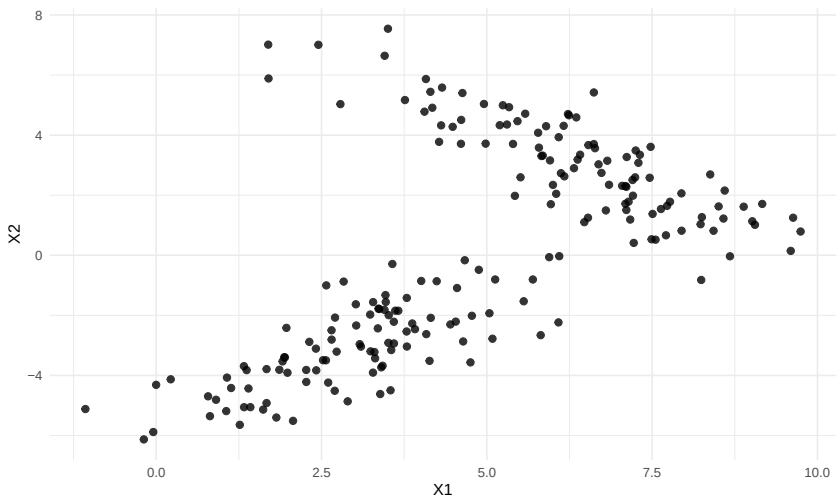
- **Distance choice:**
  Results are sensistive to the choice of d.
- **Hard partition:**
  Each point is strictly assigned to one cluster.
- **Uncertainty:**
  It does not account for classification uncertainty.

Model-based clustering:

$$f_{\boldsymbol{m},\boldsymbol{\theta}}(\boldsymbol{x}) = \sum_{k=1}^{K} \pi_k \eta_k(\boldsymbol{x}),$$

where $\boldsymbol{m}$ denotes the model, $\boldsymbol{\theta} \in \Theta_{\boldsymbol{m}}$ groups all the model parameters, the proportions $\pi_k > 0$ satisfy $\sum_{k=1}^{K} \pi_k = 1$ and η_k denotes the density of component k.

- Parametric mixture ($\boldsymbol{\theta}$ is finite-dimensional): $f_{\boldsymbol{m},\boldsymbol{\theta}}(\boldsymbol{x}) = \sum_{k=1}^{K} \pi_k \phi(\boldsymbol{x}; \boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k)$.
- Non-parametric mixture (η_{kj} is infinite-dimensional): $f_{\boldsymbol{m},\boldsymbol{\theta}}(\boldsymbol{x}) = \sum_{k=1}^{K} \pi_k \prod_{j=1}^{J} \eta_{kj}(x_j)$.
- Semi-parametric mixture (η is infinite-dimensional part): $f_{\boldsymbol{m},\boldsymbol{\theta}}(\boldsymbol{x}) = \sum_{k=1}^{K} \pi_k \eta(\boldsymbol{x} - \boldsymbol{\mu}_k)$.



Definition of a cluster:

Subjects arisen from the same mixture component, leading

$$\widehat{z}_{\boldsymbol{m},\boldsymbol{\theta}}(\boldsymbol{x}) = \underset{k=1,\ldots,K}{\arg\max}\, c_{\boldsymbol{m},\boldsymbol{\theta},k}(\boldsymbol{x})$$

with

$$c_{\boldsymbol{m},\boldsymbol{\theta},k}(\boldsymbol{x}) = \frac{\pi_k \eta_k(\boldsymbol{x})}{\sum_{\ell=1}^{K} \pi_\ell \eta_\ell(\boldsymbol{x})}.$$

In practice $(\boldsymbol{m}, \boldsymbol{\theta})$ are unknown.
We consider $\widehat{z}_{\widehat{\boldsymbol{m}},\widehat{\boldsymbol{\theta}}}$ and $c_{\widehat{\boldsymbol{m}},\widehat{\boldsymbol{\theta}},k}$.

Key advantages:

- **Classification uncertainty**
- **Model selection** $\widehat{\boldsymbol{m}} \in \mathcal{M}$ using information criteria (e.g., BIC).

The need for model validation

How to validate the clustering?

- No ground truth (external labels) is available in practice.
- Information criteria select the best model among candidates, but this model is not necessarily "good" or well-specified. We need to validate it.

Objective

Investigate if $c_{\widehat{\boldsymbol{m}},\widehat{\boldsymbol{\theta}}_{\boldsymbol{m},n}}(\boldsymbol{x})$ is well fitted.

Our requirements for a validation procedure:

1. **No re-estimation:** Avoid the computational cost of fitting new parameters.
2. **Generality:** Based only on the estimated classification matrix and the ability to sample from the model.

Well-specified model

Well-specified model:
Let f_0 denote the true density function, $f_{\boldsymbol{m},\boldsymbol{\theta}}$ is said to be *well-specified to fit the data distribution* if

$$\forall \boldsymbol{x},\ f_{\boldsymbol{m},\boldsymbol{\theta}}(\boldsymbol{x}) = f_0(\boldsymbol{x}).$$

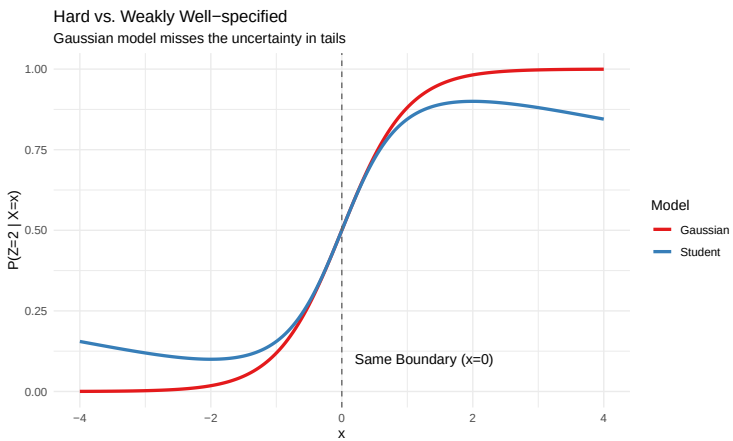
GOF on the distribution of X with estimator $\widehat{\boldsymbol{\theta}}_{\boldsymbol{m},n}$[1]

- Depends on the specific mixture.
- **X** can be an high-dimensional vector.
- Modelling the distribution of **X** is not our aim, it is a tool!

Well-specified model for clustering

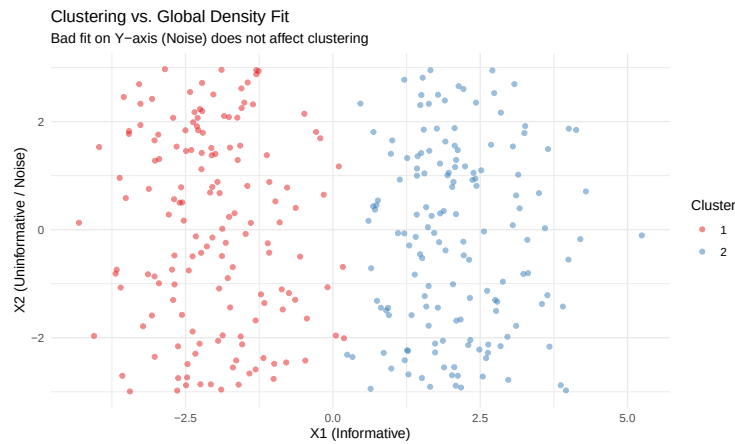
Well-specified for hard clustering	$\forall \boldsymbol{x},\ \arg\max_k c_{\boldsymbol{m},\boldsymbol{\theta},k}(\boldsymbol{x}) = \arg\max_k c_{0,k}(\boldsymbol{x}).$
(weakly) Well-specified for soft clustering	$\mathcal{L}_{f_0}(c_{\boldsymbol{m},\boldsymbol{\theta}}(\mathbf{X})) = \mathcal{L}_{f_{\boldsymbol{m},\boldsymbol{\theta}}}(c_{\boldsymbol{m},\boldsymbol{\theta}}(\mathbf{X})).$

Hard vs. Uncertainty



Same boundary, but Gaussian model underestimates uncertainty in tails vs. Student.

Clustering vs. Density Fit



Global density is misspecified (Unif vs Gauss on X_2), but classification is valid.

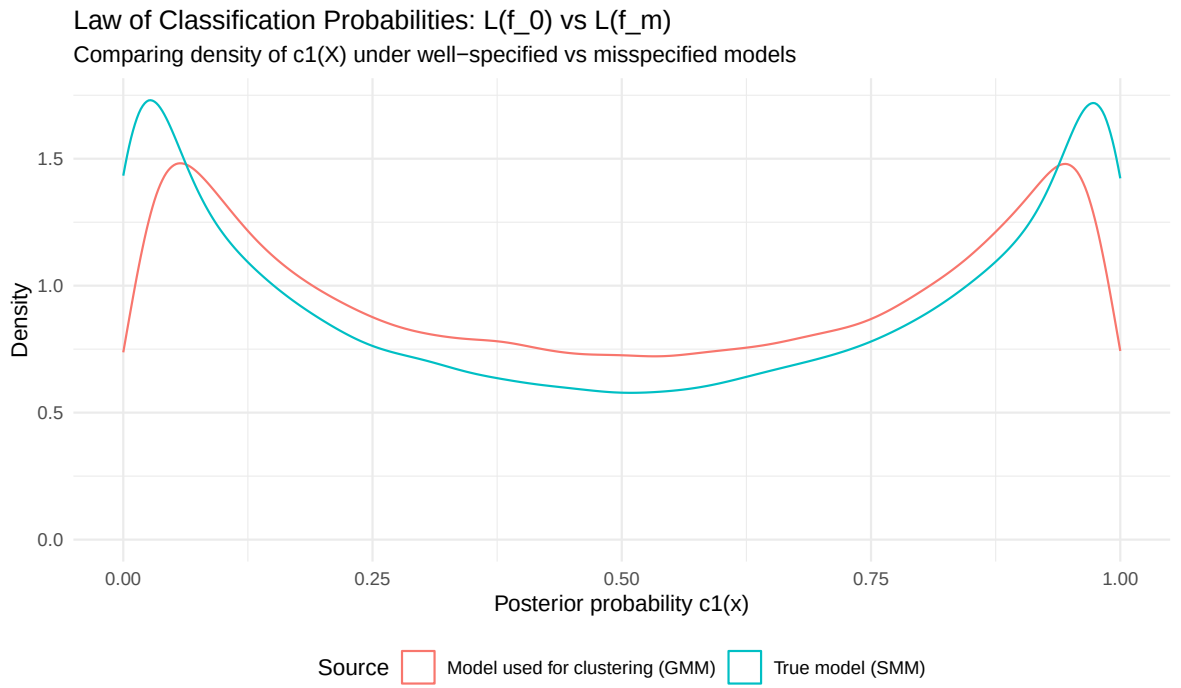
[1] Henry Braun. "A simple method for testing goodness of fit in the presence of nuisance parameters". In: *Journal of the Royal Statistical Society Series B: Statistical Methodology* 42.1 (1980), pp. 53–63.

Weakly well-specified model for soft clustering

Definition:
For a model $\boldsymbol{m}$, let $\boldsymbol{\theta}^{\star}_{\boldsymbol{m}}$ be the parameter that minimizes the loss function considered during the clustering step with model $\boldsymbol{m}$. The model $\boldsymbol{m}$ is *well-specified for soft clustering* if:

$$\mathcal{L}_{f_0}(c_{\boldsymbol{m},\boldsymbol{\theta}^{\star}_{\boldsymbol{m}}}(\mathbf{X})) = \mathcal{L}_{f_{\boldsymbol{m},\boldsymbol{\theta}^{\star}_{\boldsymbol{m}}}}(c_{\boldsymbol{m},\boldsymbol{\theta}^{\star}_{\boldsymbol{m}}}(\mathbf{X})),$$

where $\mathcal{L}_f(\cdot)$ denotes the law (distribution) under density f.



Challenges (GOF)

- Not easy to generalize usual testing procedures for multivariate data.
- $\boldsymbol{\theta}^{\star}_{\boldsymbol{m}}$ is unknown and we only have access to its estimator $\widehat{\boldsymbol{\theta}}_{\boldsymbol{m},n}$.

Weakly well-specified model for soft clustering

The model $\boldsymbol{m}$ is *weakly well-specified for soft clustering* if:

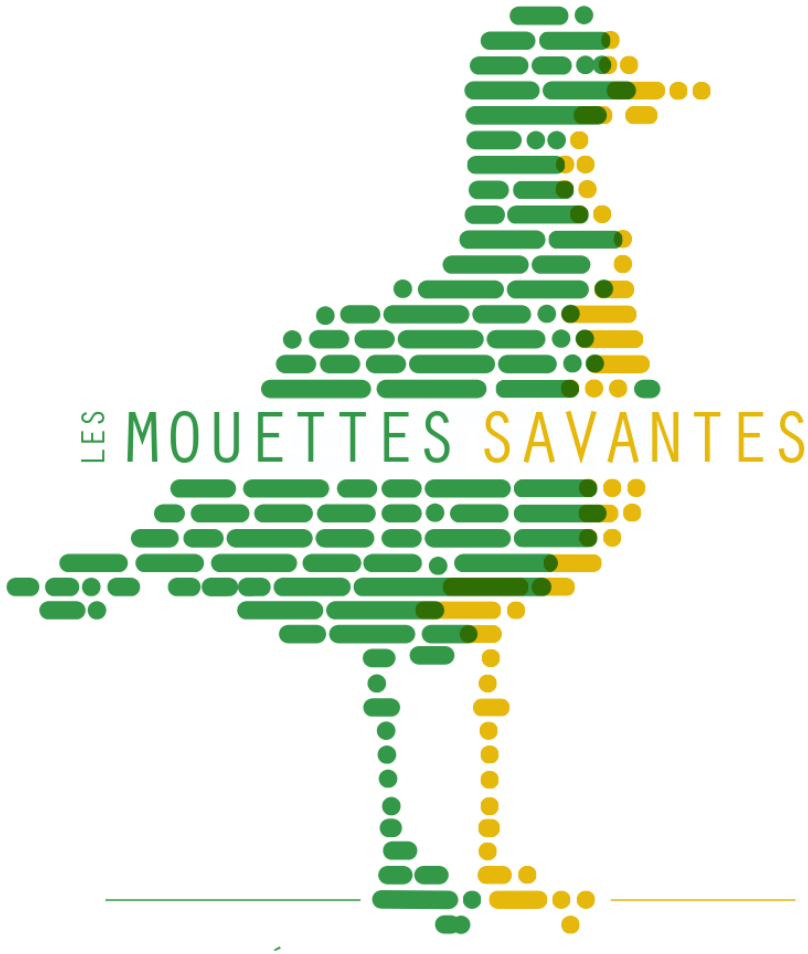
$$\mathcal{L}_{f_0}(c_{\boldsymbol{m},\boldsymbol{\theta}^\star_{\boldsymbol{m}}}(\mathbf{X})) = \mathcal{L}_{f_{\boldsymbol{m},\boldsymbol{\theta}^\star_{\boldsymbol{m}}}}(c_{\boldsymbol{m},\boldsymbol{\theta}^\star_{\boldsymbol{m}}}(\mathbf{X})).$$

Weakly well-specified model for soft clustering

The model $\boldsymbol{m}$ is *weakly well-specified for soft clustering* if:

$$\mathcal{L}_{f_0}(c_{\boldsymbol{m},\boldsymbol{\theta}^\star_{\boldsymbol{m}}}(\mathbf{X})) = \mathcal{L}_{f_{\boldsymbol{m},\boldsymbol{\theta}^\star_{\boldsymbol{m}}}}(c_{\boldsymbol{m},\boldsymbol{\theta}^\star_{\boldsymbol{m}}}(\mathbf{X})).$$

L'égalité en loi peut se tester par le théorème de la fonction m(o)uette



Testing procedure

$$\mathcal{H}_0 : \ \forall \varphi \in \mathcal{E},\ \mathbb{E}_{f_0}[\varphi(c_{\boldsymbol{m},\boldsymbol{\theta}^\star_{\boldsymbol{m}}}(\mathbf{X}))] = \mathbb{E}_{f_{\boldsymbol{m},\boldsymbol{\theta}^\star_{\boldsymbol{m}}}}[\varphi(c_{\boldsymbol{m},\boldsymbol{\theta}^\star_{\boldsymbol{m}}}(\mathbf{X}))].$$

Weakly well-specified model for soft clustering

The model $\boldsymbol{m}$ is *weakly well-specified for soft clustering* if:

$$\mathcal{L}_{f_0}(c_{\boldsymbol{m},\boldsymbol{\theta}^\star_{\boldsymbol{m}}}(\mathbf{X})) = \mathcal{L}_{f_{\boldsymbol{m},\boldsymbol{\theta}^\star_{\boldsymbol{m}}}}(c_{\boldsymbol{m},\boldsymbol{\theta}^\star_{\boldsymbol{m}}}(\mathbf{X})).$$

Testing procedure

$$\mathcal{H}_0 : \ \forall \varphi \in \mathcal{E},\ \mathbb{E}_{f_0}[\varphi(c_{\boldsymbol{m},\boldsymbol{\theta}^\star_{\boldsymbol{m}}}(\mathbf{X}))] = \mathbb{E}_{f_{\boldsymbol{m},\boldsymbol{\theta}^\star_{\boldsymbol{m}}}}[\varphi(c_{\boldsymbol{m},\boldsymbol{\theta}^\star_{\boldsymbol{m}}}(\mathbf{X}))].$$

Weakly well-specified model for soft clustering

The model $\boldsymbol{m}$ is *weakly well-specified for soft clustering* if:

$$\mathcal{L}_{f_0}(c_{\boldsymbol{m},\boldsymbol{\theta}^\star_{\boldsymbol{m}}}(\mathbf{X})) = \mathcal{L}_{f_{\boldsymbol{m},\boldsymbol{\theta}^\star_{\boldsymbol{m}}}}(c_{\boldsymbol{m},\boldsymbol{\theta}^\star_{\boldsymbol{m}}}(\mathbf{X})).$$

Testing procedure

$$\mathcal{H}_0 : \forall \varphi \in \mathcal{E},\ \mathbb{E}_{f_0}[\varphi(c_{\boldsymbol{m},\boldsymbol{\theta}^\star_{\boldsymbol{m}}}(\mathbf{X}))] = \mathbb{E}_{f_{\boldsymbol{m},\boldsymbol{\theta}^\star_{\boldsymbol{m}}}}[\varphi(c_{\boldsymbol{m},\boldsymbol{\theta}^\star_{\boldsymbol{m}}}(\mathbf{X}))].$$

Let $\psi_{\boldsymbol{m},\boldsymbol{\theta},\varphi}$ be the function defined for any $\varphi \in \mathcal{E}$ by

$$\psi_{\boldsymbol{m},\boldsymbol{\theta}^\star_{\boldsymbol{m}},\varphi}(\mathbf{X}) = \varphi(c_{\boldsymbol{m},\boldsymbol{\theta}^\star_{\boldsymbol{m}}}(\mathbf{X})) - \mathbb{E}_{f_{\boldsymbol{m},\boldsymbol{\theta}^\star_{\boldsymbol{m}}}}[\varphi(c_{\boldsymbol{m},\boldsymbol{\theta}^\star_{\boldsymbol{m}}}(\mathbf{X}))],$$

then the null hypothesis is defined as

$$\mathcal{H}_0 : \forall \varphi \in \mathcal{E}, \mathbb{E}_{f_0}[\psi_{\boldsymbol{m},\boldsymbol{\theta}^\star_{\boldsymbol{m}},\varphi}(\mathbf{X})] = 0,$$

and the alternative hypothesis

$$\mathcal{H}_1 : \exists \varphi \in \mathcal{E}, \mathbb{E}_{f_0}[\psi_{\boldsymbol{m},\boldsymbol{\theta}^\star_{\boldsymbol{m}},\varphi}(\mathbf{X})] \neq 0.$$

Three challenges

- the expectation defining the null hypothesis involves an infinite number of functions φ, whereas only a finite number of moment conditions can be tested in practice.
- $\boldsymbol{\theta}^\star_{\boldsymbol{m}}$ is unknown and we only have access to its estimator $\widehat{\boldsymbol{\theta}}_{\boldsymbol{m},n}$.
- $\mathbb{E}_{f_{\boldsymbol{m},\boldsymbol{\theta}}}[\varphi(c_{\boldsymbol{m},\boldsymbol{\theta}^\star_{\boldsymbol{m}}}(\mathbf{X}))]$ is generally not explicit.

Characterizing the Calibration

Moment conditions

$$\mathcal{H}_0 : \forall \varphi \in \mathcal{E}, \mathbb{E}_{f_0}[\psi_{\boldsymbol{m},\boldsymbol{\theta}^\star_{\boldsymbol{m}},\varphi}(\mathbf{X})] = 0.$$

We consider p functions $\varphi_1, \ldots, \varphi_p$, to construct the p-dimensional vector

$$\Psi_{\boldsymbol{m},p}(\mathbf{X};\boldsymbol{\theta}^\star_{\boldsymbol{m}}) = \left[\psi_{\boldsymbol{m},\boldsymbol{\theta}^\star_{\boldsymbol{m}},\varphi_1}(\mathbf{X}), \quad \ldots \quad , \psi_{\boldsymbol{m},\boldsymbol{\theta}^\star_{\boldsymbol{m}},\varphi_p}(\mathbf{X})\right]^\top,$$

where $\psi_{\boldsymbol{m},\boldsymbol{\theta}^\star_{\boldsymbol{m}},\varphi,j}(\mathbf{X}) = \varphi_j(c_{\boldsymbol{m},\boldsymbol{\theta}^\star_{\boldsymbol{m}}}(\mathbf{X})) - \mathbb{E}_{f_{\boldsymbol{m},\boldsymbol{\theta}^\star_{\boldsymbol{m}}}}[\varphi_j(c_{\boldsymbol{m},\boldsymbol{\theta}^\star_{\boldsymbol{m}}}(\mathbf{X}))]$.

The null hypothesis is replaced by

$$\mathcal{H}_0^{(p)} : \mathbb{E}_{f_0}[\Psi_{\boldsymbol{m},p}(\mathbf{X};\boldsymbol{\theta}^\star_{\boldsymbol{m}})] = \mathbf{0}_p.$$

Characterizing the Calibration

Moment conditions

$$\mathcal{H}_0 : \forall \varphi \in \mathcal{E}, \mathbb{E}_{f_0}[\psi_{\boldsymbol{m},\boldsymbol{\theta}^\star_{\boldsymbol{m}},\varphi}(\mathbf{X})] = 0.$$

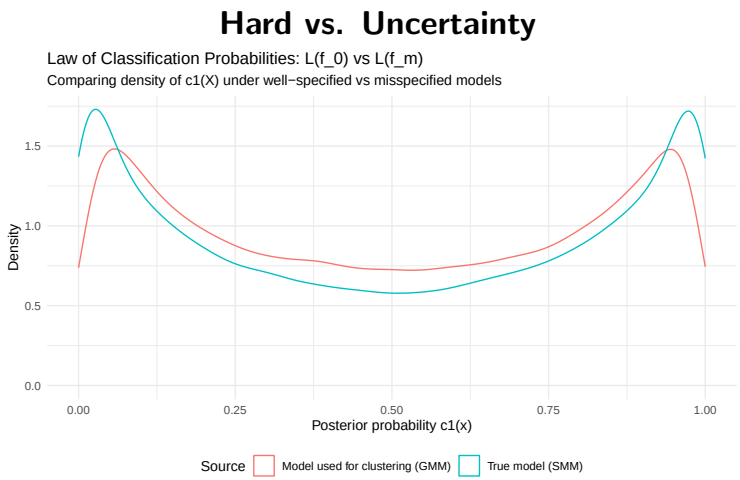
We consider p functions $\varphi_1, \ldots, \varphi_p$, to construct the p-dimensional vector

$$\Psi_{\boldsymbol{m},p}(\mathbf{X};\boldsymbol{\theta}^\star_{\boldsymbol{m}}) = \left[\psi_{\boldsymbol{m},\boldsymbol{\theta}^\star_{\boldsymbol{m}},\varphi_1}(\mathbf{X}), \quad \ldots \quad , \psi_{\boldsymbol{m},\boldsymbol{\theta}^\star_{\boldsymbol{m}},\varphi_p}(\mathbf{X})\right]^\top,$$

where $\psi_{\boldsymbol{m},\boldsymbol{\theta}^\star_{\boldsymbol{m}},\varphi,j}(\mathbf{X}) = \varphi_j(c_{\boldsymbol{m},\boldsymbol{\theta}^\star_{\boldsymbol{m}}}(\mathbf{X})) - \mathbb{E}_{f_{\boldsymbol{m},\boldsymbol{\theta}^\star_{\boldsymbol{m}}}}[\varphi_j(c_{\boldsymbol{m},\boldsymbol{\theta}^\star_{\boldsymbol{m}}}(\mathbf{X}))]$.

The null hypothesis is replaced by

$$\mathcal{H}_0^{(p)} : \mathbb{E}_{f_0}[\Psi_{\boldsymbol{m},p}(\mathbf{X};\boldsymbol{\theta}^\star_{\boldsymbol{m}})] = \mathbf{0}_p.$$



Basis: indicator functions (K=2)

φ_j	$\mathbb{E}_{f_0}$	$\mathbb{E}_{f_{\boldsymbol{m},\boldsymbol{\theta}^\star_{\boldsymbol{m}}}}$
$\mathbb{1}_{c_1 \leq 1/2}$	1/2	1/2
$\mathbb{1}_{c_1 \geq 1/2}$	1/2	1/2
$\mathbb{1}_{c_1 \leq 1/4}$	0.34	0.31

Basis: Bernstein functions (K=2)

φ_j	$\mathbb{E}_{f_0}$	$\mathbb{E}_{f_{\boldsymbol{m},\boldsymbol{\theta}^\star_{\boldsymbol{m}}}}$
$(1-c_1)^2$	0.37	0.35
$2c_1(1-c_1)$	0.34	0.29
c_1^2	0.37	0.35

Empirical likelihood

$$\mathcal{H}_0^{(p)} \; : \; \mathbb{E}_{f_0}[\Psi_{\boldsymbol{m},p}(\mathbf{X};\boldsymbol{\theta}_{\boldsymbol{m}}^\star)] = \mathbf{0}_p.$$

Empirical likelihood[2]: main ideas

- It is a nonparametric method of inference based on a data-driven likelihood ratio function.
- It does not require to specify the distributions for the data (relies only on moment constraints).

Empirical likelihood for mean testing

From an iid sample $\overline{\mathbf{X}} = (\mathbf{X}_1, \ldots, \mathbf{X}_n)$, EL builds a likelihood ratio in the spirit of

$$\ln \frac{\sup_\xi L_n(\boldsymbol{\theta}_{\boldsymbol{m}}^\star, \xi)}{\sup_{(\boldsymbol{\theta},\xi)} L_n(\boldsymbol{\theta}, \xi)}$$

without parametric assumptions

$$\mathcal{R}_{\boldsymbol{m},p}(\boldsymbol{\theta}_{\boldsymbol{m}}^\star; \overline{\mathbf{X}}) = \sup_{\omega_1(\boldsymbol{\theta}_{\boldsymbol{m}}^\star),\ldots,\omega_n(\boldsymbol{\theta}_{\boldsymbol{m}}^\star)} \left\{ \sum_{i=1}^n \ln(n\omega_i(\boldsymbol{\theta}_{\boldsymbol{m}}^\star)) : \omega_i(\boldsymbol{\theta}_{\boldsymbol{m}}^\star) \geq 0, \sum_{i=1}^n \omega_i(\boldsymbol{\theta}_{\boldsymbol{m}}^\star) = 1, \; \sum_{i=1}^n \omega_i(\boldsymbol{\theta}_{\boldsymbol{m}}^\star)\Psi_p(\mathbf{X}_i;\boldsymbol{\theta}_{\boldsymbol{m}}^\star) = \mathbf{0}_p \right\}$$

Properties

- **Wilks-type theorem:** under $\mathcal{H}_0$, $-2\mathcal{R}_{\boldsymbol{m},p}(\boldsymbol{\theta}_{\boldsymbol{m}}^\star; \overline{\mathbf{X}}) \; \rightarrow \; \chi^2_p$.
- **Confidence region:** Set of $\boldsymbol{\theta}_{\boldsymbol{m}}^\star$ that yield high values of $-2\mathcal{R}_{\boldsymbol{m},p}(\boldsymbol{\theta}_{\boldsymbol{m}}^\star; \overline{\mathbf{X}})$.

Comparison

- vs. LRT: keeps likelihood ratio structure without distributional assumptions.
- vs. CLT-based test: avoids plugging estimator of covariance and faster convergence to χ^2 limit.
- **Limitations:** computationally more demanding than CLT-based tests.

---

[2] Art B Owen. *Empirical likelihood*. Chapman and Hall/CRC, 2001.

Visualizing calibration: CDF comparisons

$$\hat{F}(x) = \frac{1}{n}\sum_{i=1}^{n} \mathbb{1}_{\{X_i \leq x\}} \text{ and } \hat{F}_{EL}(x; \boldsymbol{\theta}^{\star}_{\boldsymbol{m}}) = \sum_{i=1}^{n} \omega_i(\boldsymbol{\theta}^{\star}_{\boldsymbol{m}}) \mathbb{1}_{\{X_i \leq x\}}.$$

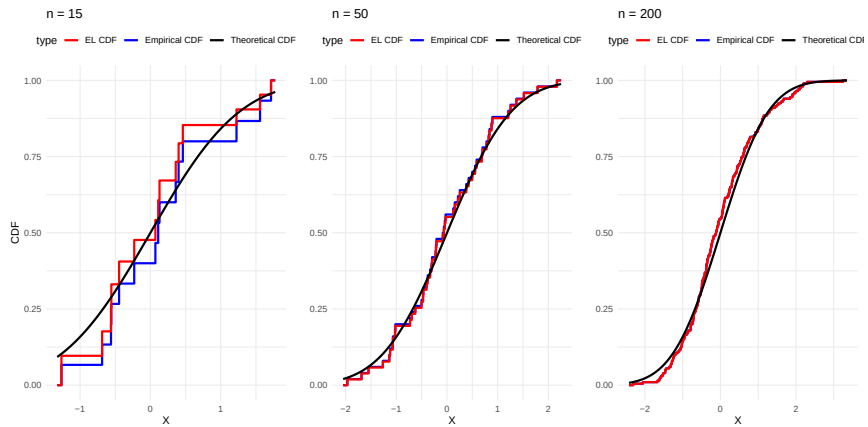


Figure: Comparison of CDFs in the standard Gaussian case with one well-specified constraint on the first moment $\mathbb{E}[X] = 0$.

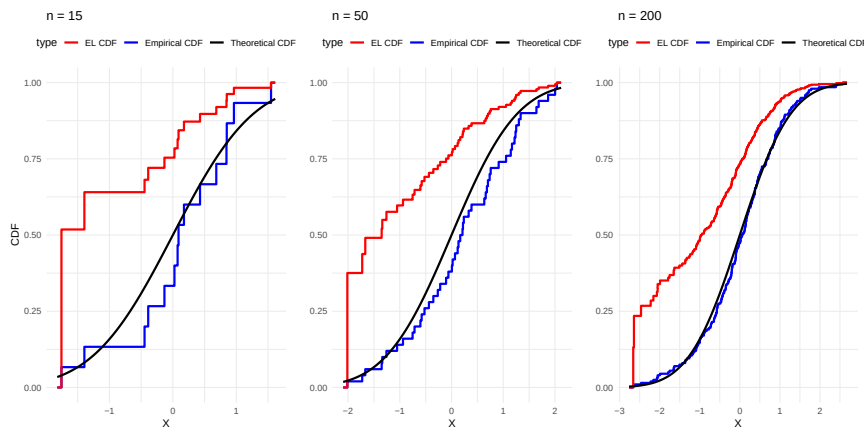


Figure: Comparison of CDFs in the standard Gaussian case with one miss-specified constraint on the first moment $\mathbb{E}[X + 1] = 0$.

Testing a value of the parameter

If $\Psi_p(\mathbf{X}; \boldsymbol{\theta}^\star_{\boldsymbol{m}}) \in \mathbb{R}^p$ has a full rank covariance matrix, then[3]

$$-2\mathcal{R}_{\boldsymbol{m},p}(\boldsymbol{\theta}^\star_{\boldsymbol{m}}; \overline{\mathbf{X}}) \xrightarrow{d} \mathcal{X}^2_p.$$

Steps of the proof

1. Controlling the Lagrange multipliers $\|\boldsymbol{\lambda}(\boldsymbol{\theta}^\star_{\boldsymbol{m}})\|_2 = O_{\mathbb{P}}(n^{-1/2})$, using that, under mild assumptions,
   - $\max_{1\leq i\leq n} \|\Psi_{\boldsymbol{m},p}(\mathbf{X}_i; \boldsymbol{\theta}^\star_{\boldsymbol{m}})\|_2 = o_{\mathbb{P}}(n^{1/2})$;
   - $\|n^{-1}\sum_{i=1}^n \Psi_{\boldsymbol{m},p}(\mathbf{X}_i; \boldsymbol{\theta}^\star_{\boldsymbol{m}})\|_2 = O_{\mathbb{P}}(n^{-1/2})$;
   - $S_n(\boldsymbol{\theta}^\star_{\boldsymbol{m}})$ converges in probability to a full rank matrix.
2. $$-2\mathcal{R}_{\boldsymbol{m},p}(\boldsymbol{\theta}^\star_{\boldsymbol{m}}; \overline{\mathbf{X}}) = \left(\frac{1}{\sqrt{n}}\sum_{i=1}^n \Psi_{\boldsymbol{m},p}(\mathbf{X}_i; \boldsymbol{\theta}^\star_{\boldsymbol{m}})\right)^\top S_n(\boldsymbol{\theta}^\star_{\boldsymbol{m}})^{-1} \left(\frac{1}{\sqrt{n}}\sum_{i=1}^n \Psi_{\boldsymbol{m},p}(\mathbf{X}_i; \boldsymbol{\theta}^\star_{\boldsymbol{m}})\right) + o_{\mathbb{P}}(1),$$
   where $S_n(\boldsymbol{\theta}^\star_{\boldsymbol{m}}) = n^{-1}\sum_{i=1}^n \Psi_{\boldsymbol{m},p}(\mathbf{X}_i; \boldsymbol{\theta}^\star_{\boldsymbol{m}})\Psi_{\boldsymbol{m},p}(\mathbf{X}_i; \boldsymbol{\theta}^\star_{\boldsymbol{m}})^\top$.

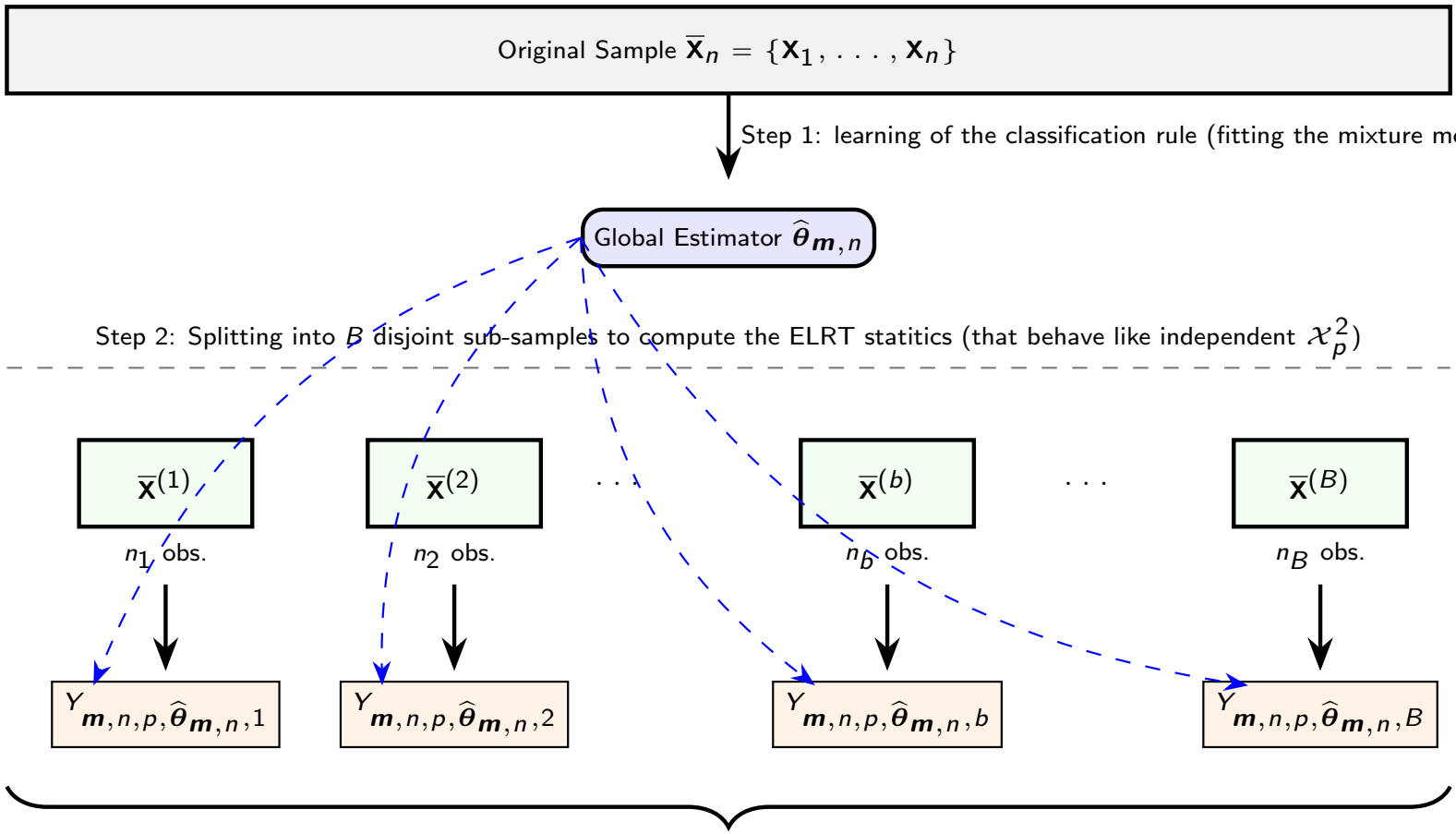
Impact of the estimator

$$-2\mathcal{R}_{\boldsymbol{m},p}(\widehat{\boldsymbol{\theta}}_{\boldsymbol{m},n}; \overline{\mathbf{X}}) \not\xrightarrow{d} \mathcal{X}^2_p,$$

since $\|\Psi_{\boldsymbol{m},p}(\cdot; \boldsymbol{\theta}^\star_{\boldsymbol{m}}) - \Psi_{\boldsymbol{m},p}(\cdot; \widehat{\boldsymbol{\theta}}_{\boldsymbol{m},n})\|_\infty \neq o_{\mathbb{P}}(n^{-1/2})$

[3] Jin Qin and Jerry Lawless. "Empirical likelihood and general estimating equations". In: *The Annals of Statistics* 22.1 (1994), pp. 300–325.

Splitting the data



Vector of statistics $\mathbf{Y}_{\boldsymbol{m},n,p,\widehat{\boldsymbol{\theta}}_{\boldsymbol{m},n}} = (Y_{\boldsymbol{m},n,p,\widehat{\boldsymbol{\theta}}_{\boldsymbol{m},n},1}, \ldots, Y_{\boldsymbol{m},n,p,\widehat{\boldsymbol{\theta}}_{\boldsymbol{m},n},B})^\top \in \mathbb{R}^B$

- $Y_{\boldsymbol{m},n,p,\widehat{\boldsymbol{\theta}}_{\boldsymbol{m},n},b} = -2\mathcal{R}_{\boldsymbol{m},p}(\widehat{\boldsymbol{\theta}}_{\boldsymbol{m},n}, \overline{\mathbf{X}}^{(b)})$.
- $n_b \to \infty$ to ensure the convergence of $Y_{\boldsymbol{m},n,p,\widehat{\boldsymbol{\theta}}_{\boldsymbol{m},n},b}$ to $\mathcal{X}_p$.
- $B \to \infty$ (imply n_b/n tends to zero) to have $\|\Psi_{\boldsymbol{m},p}(\cdot;\boldsymbol{\theta}^\star_{\boldsymbol{m}}) - \Psi_{\boldsymbol{m},p}(\cdot;\widehat{\boldsymbol{\theta}}_{\boldsymbol{m},n})\|_\infty = o_{\mathbb{P}}(n_b^{-1/2})$.
- p needs to grow as $n \to \infty$.

Distribution of the vector of LRT

Growing number of equations
The null hypothesis is

$$\mathcal{H}_0 : \forall \varphi \in \mathcal{E}, \mathbb{E}_{f_0}[\psi_{\boldsymbol{m},\boldsymbol{\theta}^\star_{\boldsymbol{m}},\varphi}(\mathbf{X})] = 0,$$

has been replaced by

$$\mathcal{H}_0^{(p)} : \mathbb{E}_{f_0}[\Psi_{\boldsymbol{m},p}(\mathbf{X};\boldsymbol{\theta}^\star_{\boldsymbol{m}})] = \mathbf{0}_p.$$

The number of equations p must increases with n.

Controlling the distribution of the vector of LRT
Under the null hypothesis and mild assumptions,

$$\lim_{n\to\infty} \left\| F_{Y_{\boldsymbol{m},n,p,\widehat{\boldsymbol{\theta}}_{\boldsymbol{m},n}}} - \prod_{b=1}^{B} F_{\mathcal{X}^2_p} \right\|_\infty = 0,$$

where $F_{Y_{\boldsymbol{m},n,p,\widehat{\boldsymbol{\theta}}_{\boldsymbol{m},n}}}$ is cdf of $Y_{\boldsymbol{m},n,p,\widehat{\boldsymbol{\theta}}_{\boldsymbol{m},n}}$ and $F_{\mathcal{X}^2_p}$ denotes the cdf of $\mathcal{X}^2_p$.

Main idea
Clustering validation is performed via goodness-of-fit testing based on the B test statistics $Y_{\boldsymbol{m},n,p,\widehat{\boldsymbol{\theta}}_{\boldsymbol{m},n},b}$.

Test statistics

Sum	$\sum_{b=1}^{B} Y_{\boldsymbol{m},n,p,\widehat{\boldsymbol{\theta}}_{\boldsymbol{m},n},b}$
Maximum	$\max_{1\leq b\leq B} Y_{\boldsymbol{m},n,p,\widehat{\boldsymbol{\theta}}_{\boldsymbol{m},n},b}$
Kolmogorov-Smirnov	$\sup_y \lvert\hat{F}_B(y) - F_{\chi^2_p}(y)\rvert$
Cramér-von Mises	$B\int_{-\infty}^{\infty}(\hat{F}_B(y) - F_{\chi^2_p}(y))^2 dF_{\chi^2_p}(y)$
The Anderson-Darling	$B\int_{-\infty}^{\infty}\frac{(\hat{F}_B(y)-F_{\chi^2_p}(y))^2}{F_{\chi^2_p}(y)(1-F_{\chi^2_p}(y))} dF_{\chi^2_p}(y)$

Practical Implementation: statistics and basis

Choice of the test statistics:
While several options exist (*max*, *sum*), we recommend **Kolmogorov-Smirnov (KS)**

- Better numerical performance in small samples ($n = 200$).
- More robust diagnostic by considering the full distribution of the moment vector.

Choice of the functional basis φ:
We recommend a **Bernstein basis** with a data-driven orthogonalization:

1. Generate K-variate Bernstein polynomials of degree s: $\varpi_{\boldsymbol{j}}(\boldsymbol{a}) = \frac{s!}{\prod j_k!} \prod a_k^{j_k}$.
2. Perform **PCA** on these polynomials to handle multicollinearity and the simplex constraint.
3. Use the first p principal components as the final functions.

Choice of the growing rates

1. Obtain the rate of convergence of $\widehat{\boldsymbol{\theta}}_{\boldsymbol{m},n}$ to $\boldsymbol{\theta}^{\star}_{\boldsymbol{m}}$ (at least $n^{-1/3}$)
2. Select a rate of growing for B based on $\|\widehat{\boldsymbol{\theta}}_{\boldsymbol{m},n} - \boldsymbol{\theta}^{\star}_{\boldsymbol{m}}\|$
3. Select a rate of growing for p based on $\|\widehat{\boldsymbol{\theta}}_{\boldsymbol{m},n} - \boldsymbol{\theta}^{\star}_{\boldsymbol{m}}\|$ and B.

Recommended default values (Parametric case):
Based on our numerical findings, we suggest:

Number of blocks	$B = \lfloor 2n^{1/4} \rceil$
Number of moments	$p = \lfloor 3n^{1/9} \rceil$

Some numerical experiments: parametric mixture models

$$f_{\boldsymbol{m},\boldsymbol{\theta}}(\boldsymbol{x}) = \sum_{k=1}^{K} \pi_k \prod_{j=1}^{J} \eta_{k,j}(x_j).$$

			Full Gaussian mixture model				Student mixture model			
n	p	B	$c = 0$	$c = 0.25$	$c = 0.50$	$c = 0.75$	$df = 6$	$df = 5$	$df = 4$	$df = 3$
200	5	6	0.336	0.580	0.508	0.628	0.669	0.692	0.742	0.798
400	6	7	0.061	0.098	0.110	0.406	0.255	0.344	0.477	0.611
800	6	8	0.049	0.054	0.081	0.621	0.129	0.197	0.334	0.565
1600	7	9	0.047	0.054	0.135	0.860	0.086	0.118	0.295	0.684
3200	7	10	0.044	0.064	0.362	0.973	0.088	0.155	0.378	0.913
6400	8	12	0.052	0.112	0.809	1.000	0.107	0.280	0.588	0.996

Table: Parametric case - Empirical power of the test under two alternative hypotheses: a GMM with correlated variables (c) and a Student's t-mixture model (df). The case $c = 0$ represents the null hypothesis (H_0). Rejection rates are computed over 5000 replicates ($\alpha = 0.05$).

n	p	B	$c = 0$	$c = 0.25$	$c = 0.50$	$c = 0.75$
200	5	6	0.274	0.354	0.384	0.522
400	6	7	0.060	0.082	0.085	0.330
800	6	8	0.053	0.056	0.062	0.585
1600	7	9	0.050	0.054	0.098	0.877
3200	7	10	0.049	0.064	0.276	0.958
6400	8	12	0.052	0.079	0.702	0.999

Table: Nonparametric case - Empirical level of the testing the conditional independence between variables within components. The case $c = 0$ represents the null hypothesis (H_0). Rejection rates are calculated over 5000 replicates at a nominal level $\alpha = 0.05$ obtained with the KS statistic.

Take home message

- Internal diagnostic for clustering:
  A novel validation method that directly tests the specification of posterior probabilities on the unit simplex, regardless of the data dimension.

- No parametric mixture required:
  The procedure is flexible and does not rely on a global parametric assumption. It is valid for both parametric and non-parametric frameworks.

- Key requirement:
  You only need a consistent estimator $\widehat{\boldsymbol{\theta}}_{\boldsymbol{m},n}$ and its rate of convergence to $\boldsymbol{\theta}^{\star}_{\boldsymbol{m}}$ to properly calibrate: B and p

- Numerical efficiency:
  No additional model fitting is required. The test is computationally light and handles high-dimensional data.

- Reference:
  El Kolei, Salima and Matthieu Marbac. "Goodness-of-fit testing of the distribution of posterior classification probabilities for validating model-based clustering." arXiv preprint arXiv:2511.04206 (2025).

- R package:
  GOFclustering available on CRAN.

Take home message

- Internal diagnostic for clustering:
  A novel validation method that directly tests the specification of posterior probabilities on the unit simplex, regardless of the data dimension.

- No parametric mixture required:
  The procedure is flexible and does not rely on a global parametric assumption. It is valid for both parametric and non-parametric frameworks.

- Key requirement:
  You only need a consistent estimator $\widehat{\boldsymbol{\theta}}_{\boldsymbol{m},n}$ and its rate of convergence to $\boldsymbol{\theta}^{\star}_{\boldsymbol{m}}$ to properly calibrate: B and p

- Numerical efficiency:
  No additional model fitting is required. The test is computationally light and handles high-dimensional data.

- Reference:
  El Kolei, Salima and Matthieu Marbac. "Goodness-of-fit testing of the distribution of posterior classification probabilities for validating model-based clustering." arXiv preprint arXiv:2511.04206 (2025).

- R package: