Estimating the Contamination Factor's Distribution in Unsupervised Anomaly Detection

Lorenzo Perini*, Paul Bürkner, Arto Klami*

lorenzo.perini@kuleuven.be
paul-christian.buerkner@simtech.uni-stuttgart.de
arto.klami@helsinki.fi

https://people.cs.kuleuven.be/~lorenzo.perini/

@LorenzoPerini95

Anomaly Detection is the Task of Detecting the Instances that Deviate from a Normal Behaviour

Fraudulent transactions

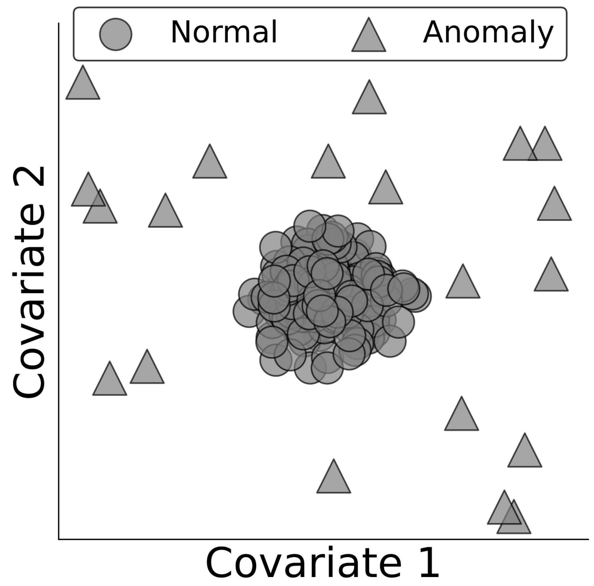
Machine breakdowns

Cyberattacks

University of Stuttgart
HELSINGIN YLIOPISTO
HELSINGFORS UNIVERSITET
UNIVERSITY OF HELSINKI
fwo
DTAI
KU LEUVEN

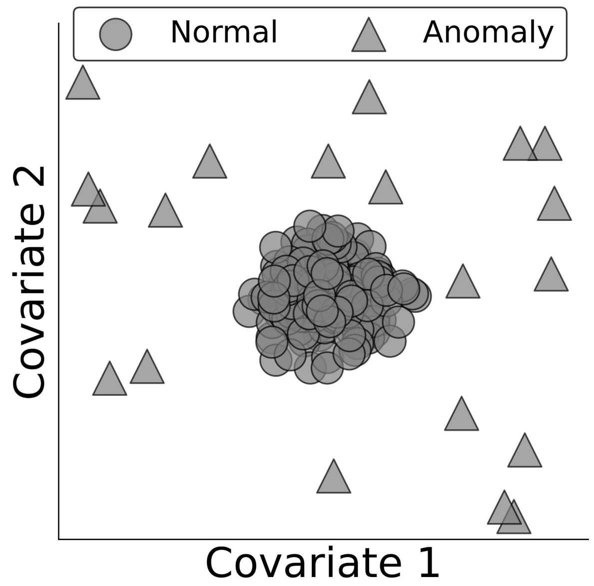
Because Labels Are Hard to Collect, Anomaly Detection Is Usually Tackled from an Unsupervised Perspective

Unlabeled data



Unsupervised Anomaly Detectors Exploit Data-Driven Intuitions to Assign Anomaly Scores

Unlabeled data

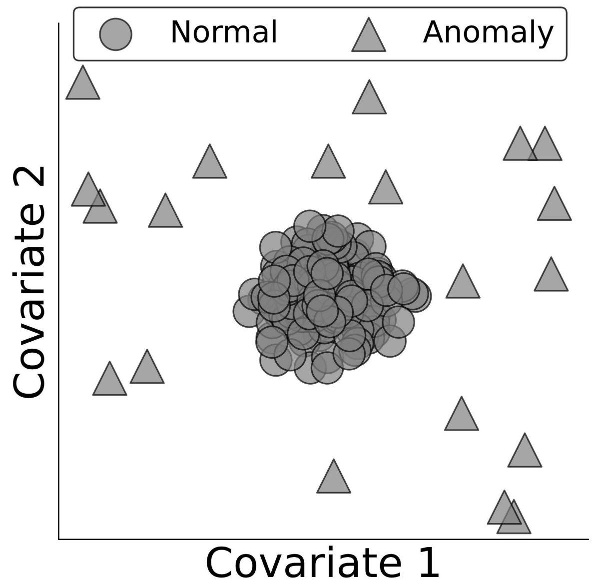


Anomaly Detector

Anomaly Scores
s_1
s_2
s_3
s_4
s_5
s_6
...
s_n

Transforming the Anomaly Scores into Hard Predictions Requires Setting a Decision Threshold

Unlabeled data



Anomaly Detector →

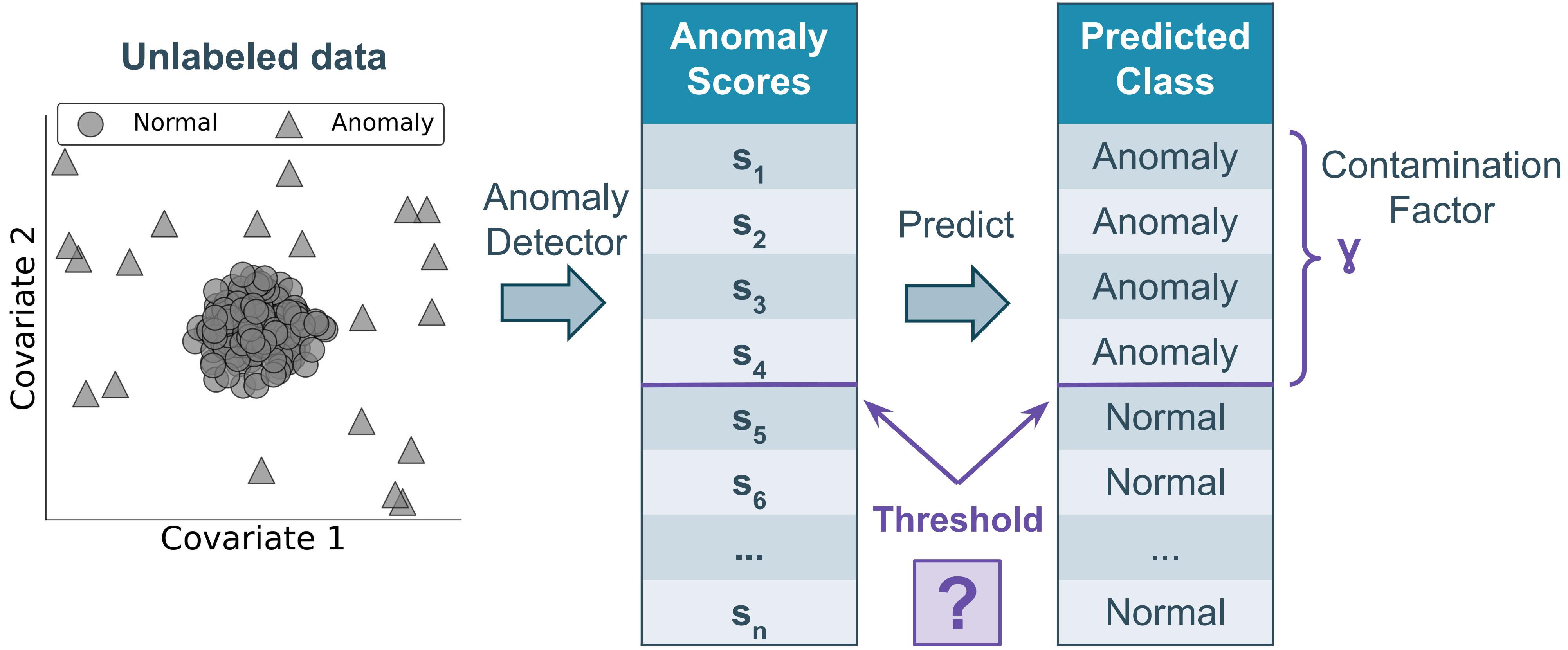
Anomaly Scores	Predicted Class
s_1	Anomaly
s_2	Anomaly
s_3	Anomaly
s_4	Anomaly
s_5	Normal
s_6	Normal
...	...
s_n	Normal

Predict →

Threshold

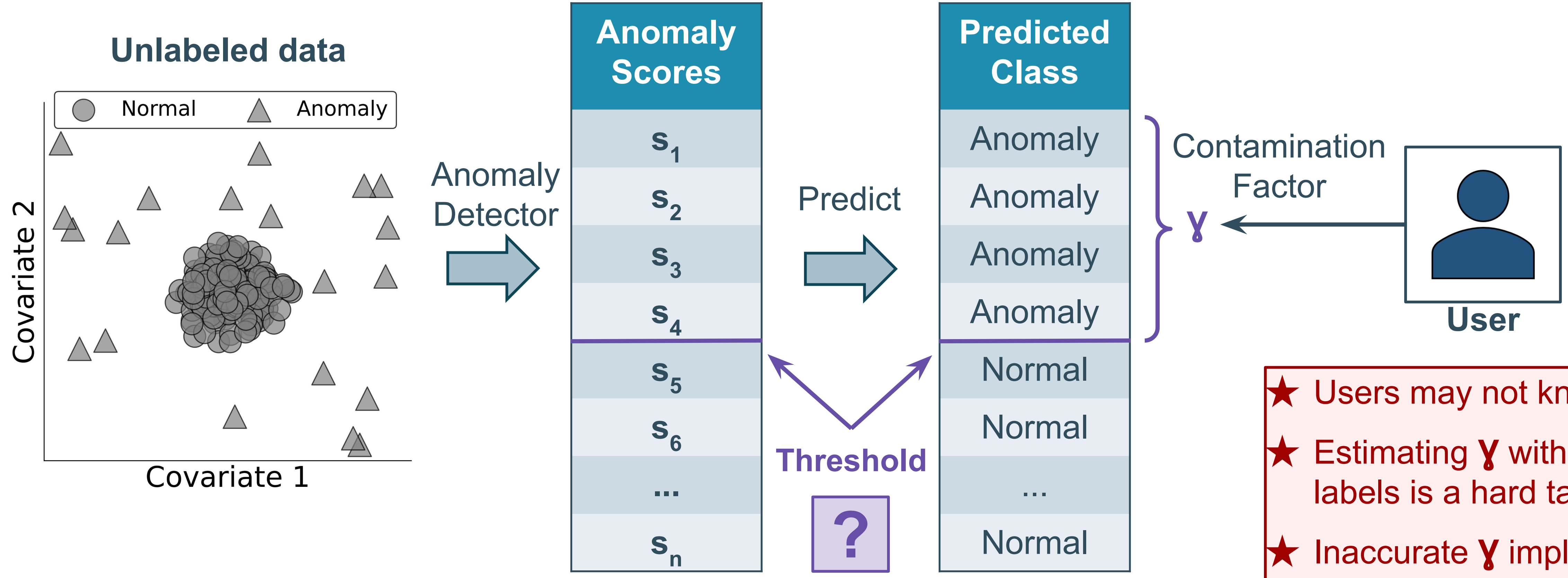
?

One Can Set the Threshold Using the Contamination Factor, i.e. the Proportion of Anomalies in the Data



University of Stuttgart · HELSINGIN YLIOPISTO HELSINGFORS UNIVERSITET UNIVERSITY OF HELSINKI · fwo · DTAI · KU LEUVEN

Setting a Correct Contamination Factor is <u>Essential</u> to Get Accurate Predictions



- ★ Users may not know γ
- ★ Estimating γ without labels is a hard task
- ★ Inaccurate γ implies wrong predictions

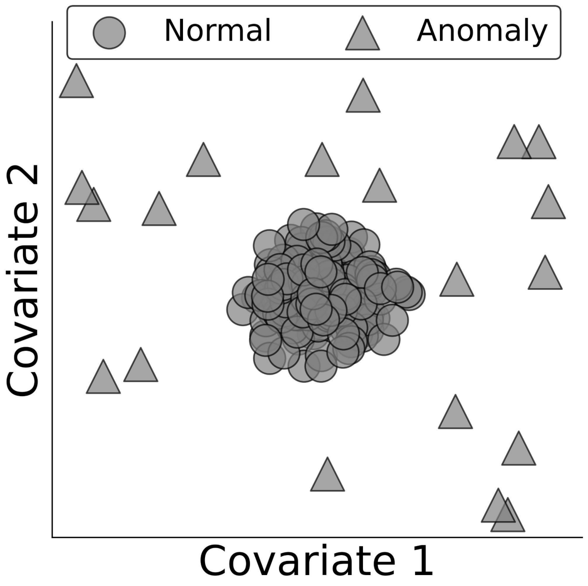
How Can We Estimate the Contamination Factor in an Unsupervised Setting?

We Assume a Bayesian Perspective and Propose to Estimate the Contamination Factor's Distribution

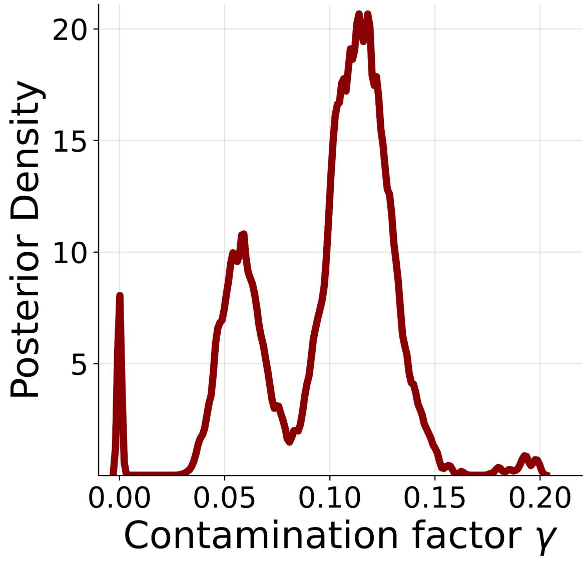
Task

Given:

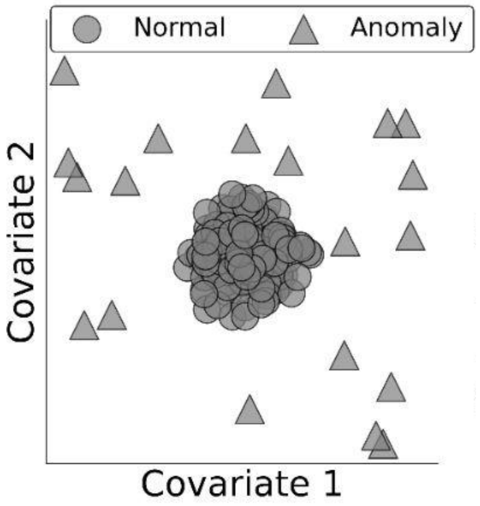
Unlabeled data



Estimate:



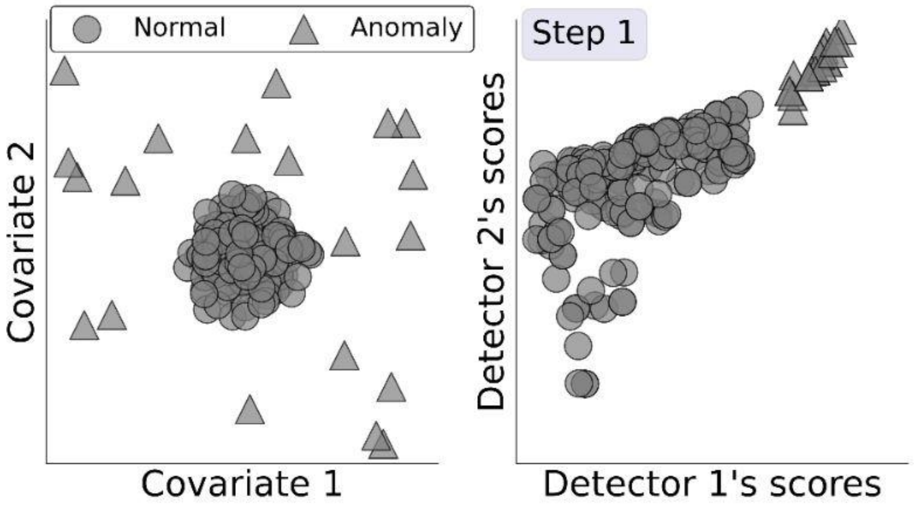
Our Method ɣGMM Estimates ɣ's Posterior in Four Steps



Unlabeled Data

Because Anomalies May Not Follow Patterns, We Map the Data into an <u>M-dimensional Anomaly Score Space</u>

Step 1

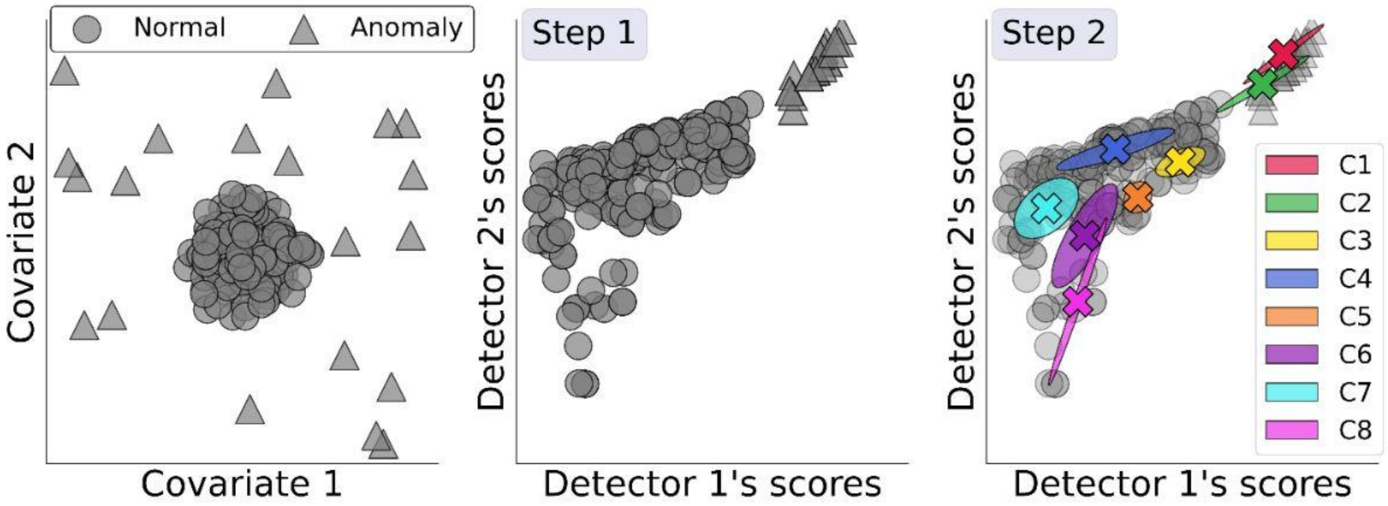


Unlabeled Data

1. Compute the anomaly scores using M unsupervised detectors

Linking Contamination and Data using a DPGMM: the Components' Mass Reflect the Contamination

Step 2



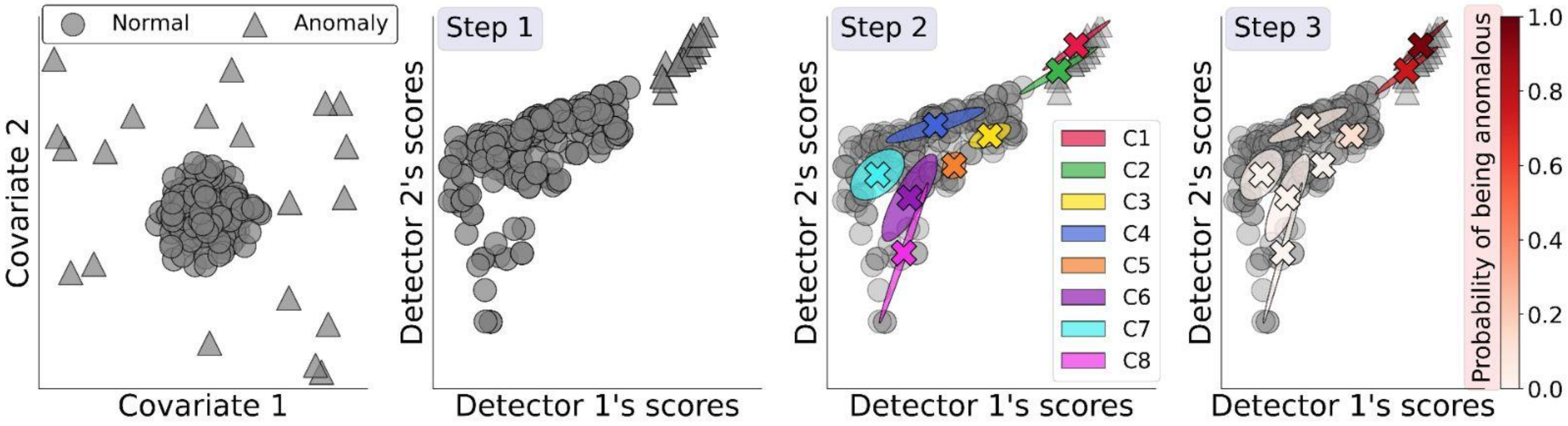
Unlabeled Data

1. Compute the anomaly scores using M unsupervised detectors

2. Model the family of scores as a DPGMM (using V.I.)

Because We Do Not Know Which Component Is Anomalous, We Estimate Their (Joint) Probability

Step 3



Unlabeled Data

1. Compute the anomaly scores using M unsupervised detectors

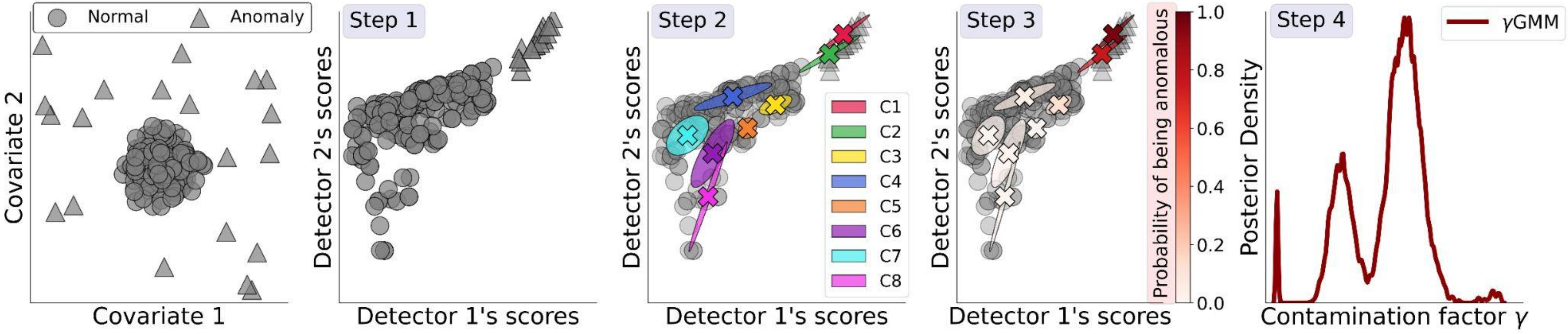
2. Model the family of scores as a DPGMM (using V.I.)

3. Compute the components' probability of being anomalous

University of Stuttgart
HELSINGIN YLIOPISTO HELSINGFORS UNIVERSITET UNIVERSITY OF HELSINKI
fwo
DTAI
KU LEUVEN

We Derive the Posterior by Combining Step 3's Probabilities with Step 2 Components' Posterior

Step 4



Unlabeled Data

1. Compute the anomaly scores using M unsupervised detectors

2. Model the family of scores as a DPGMM (using V.I.)

3. Compute the components' probability of being anomalous

4. Derive **γ**'s posterior distribution

$$p(\gamma) = \sum_{k=1}^{K} \mathbb{P}(\text{exact } c_1, \ldots, c_k)\mathcal{B}(a_k, b_k)$$

University of Stuttgart
HELSINGIN YLIOPISTO
HELSINGFORS UNIVERSITET
UNIVERSITY OF HELSINKI
fwo
DTAI
KU LEUVEN

How Does ɣGMM Compare to Existing Approaches?

Estimating the Contamination Factor's Distribution in Unsupervised Anomaly Detection

For further details:

★ Check out the paper online

★ Reach out to us via email

Lorenzo Perini, *Paul Bürkner, Arto Klami*

lorenzo.perini@kuleuven.be
paul-christian.buerkner@simtech.uni-stuttgart.de
arto.klami@helsinki.fi

https://people.cs.kuleuven.be/~lorenzo.perini/

@LorenzoPerini95