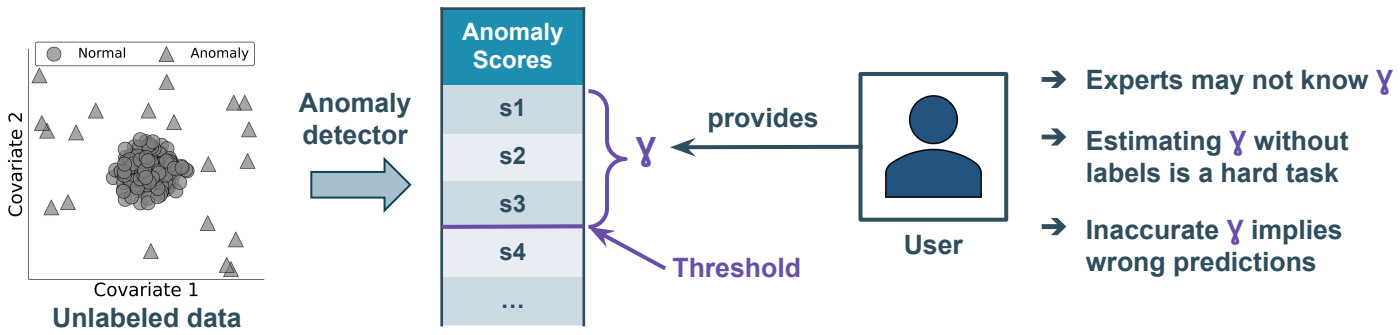
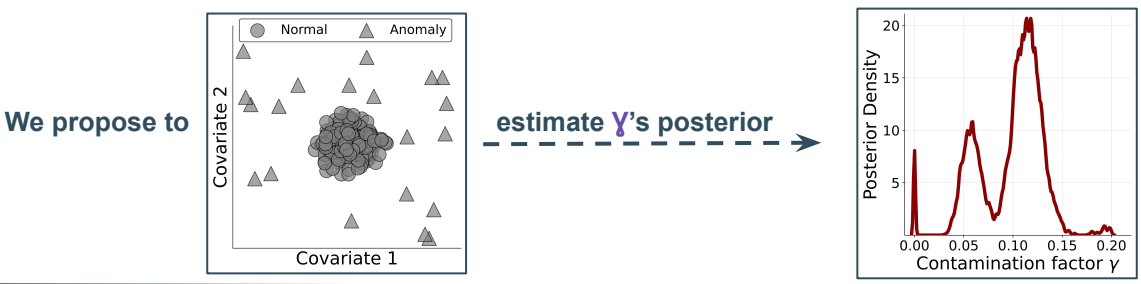
Estimating the Contamination Factor's Distribution in Unsupervised Anomaly Detection

Problem: thresholding the anomaly scores requires domain knowledge

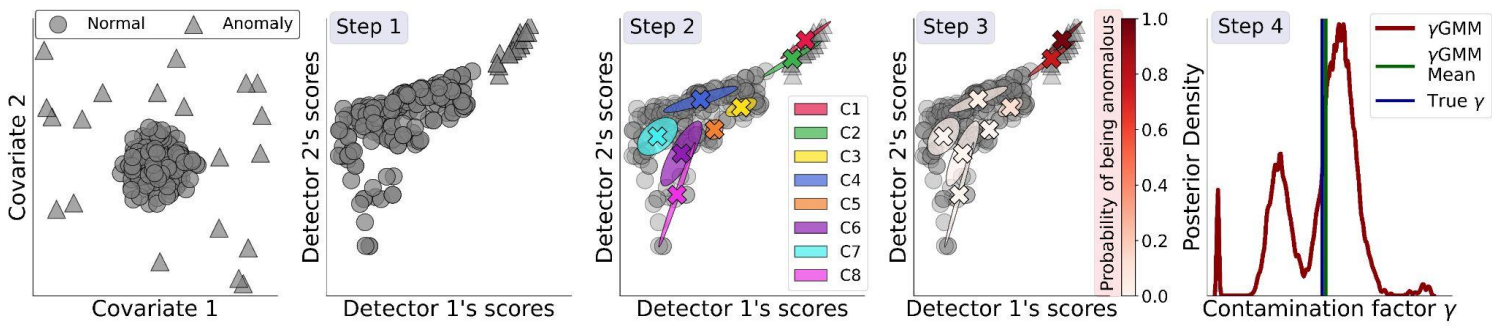


Task: estimate the contamination factor's posterior distribution using only unlabeled data



Insight: (a) Model the data in the anomaly score space, (b) identify the components flagged as anomalies by several detectors, and (c) estimate their mass as the contamination γ

We propose *γGMM*, a 4-steps approach:



Unlabeled Data

1. Compute the anomaly scores using M unsupervised detectors

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

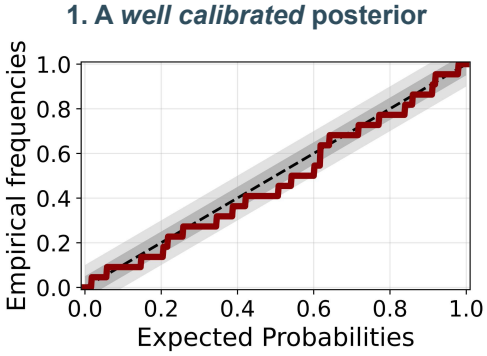
3. Compute the components' probability of being anomalous

4. Derive γ's posterior distribution

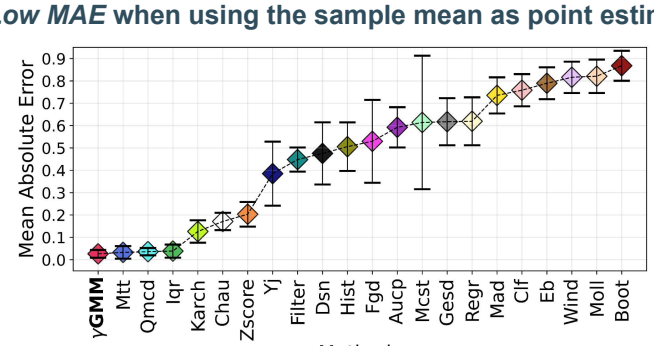
$$p(\gamma|S) = \sum_{k=1}^{K} p(c_1, \dots, c_k)\mathcal{B}(a_k, b_k)$$

Experiments on 20 datasets show that *γGMM* has

1. A *well calibrated* posterior



2. *Low MAE* when using the sample mean as point estimate



Lorenzo Perini, Paul Bürkner, Arto Klami

@LorenzoPerini95

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

SCAN HERE!

fwo DTAI KU LEUVEN

HELSINGIN YLIOPISTO HELSINGFORS UNIVERSITET UNIVERSITY OF HELSINKI

University of Stuttgart