Supplementary Materials

This online Supplement contains additional details about the experiments. Specifically, we provide additional details on the datasets used, and we extend the results of Q1,Q2 and Q3.

Table 1. Properties of the 22 datasets used. For each dataset, we report the number examples, the number of original covariates, and the ground-truth contamination factor.

Dataset	# Examples (N)	# Covariates	True γ^*
ALOI	12384	27	0.0304
Annthyroid	7129	21	0.0749
Arrhythmia	271	259	0.0996
Cardiotocography	1734	21	0.0496
Glass	214	7	0.0421
InternetAds	1682	1555	0.0499
KDDCup99	48113	40	0.0042
Lymphography	148	47	0.0405
PageBlocks	5473	10	0.1023
Parkinson	53	22	0.0943
PenDigits	9868	16	0.0020
Pima	526	8	0.0494
Shuttle	1013	9	0.0128
SpamBase	2661	57	0.0500
Stamps	340	9	0.0912
T15	42125	10	0.0668
T21	18509	10	0.0529
WBC	223	9	0.0448
WDBC	367	30	0.0272
WPBC	160	33	0.0562
Waveform	3443	21	0.0290
Wilt	4655	5	0.0200

Data. Table 1 shows the details of the used datasets. The datasets vary in terms of number of examples (from around 50 to more than 48000), number of covariates (from 5 to more than 1500) and the contamination factor (from around 0.004 to more than 0.10). Note that even the highest contamination factor is around 0.10, confirming the general assumption of anomalies being rare.

Q1-Q2. γGMM's estimated distribution. Table 2 shows the MAE between γGMM's sample mean and the true value γ^* on a per-dataset basis. With respect to the true value γ^* and out of 22 experiments, the sample mean is:

- a *good estimate* (i.e., MAE ≤ 0.01) for 7 datasets (Cardiotocography, InternetAds, Pima, SpamBase, T21, WPBC, Wilt);
- a *slightly imprecise estimate* (i.e., $0.01 <$ MAE ≤ 0.03) for 7 datasets (ALOI, Annthyroid, Glass, Lymphography, Parkinson, Stamps and T15);
- a *not-optimal estimate* (i.e., $0.03 <$ MAE ≤ 0.05) for 6 datasets (KDDCup99, PageBlocks, PenDigits, WBC, WDBC, and Waveform);
- a *bad estimate* (MAE > 0.05) for just two datasets (Arrhythmia, and Shuttle).

This shows, again, that the estimated distribution is well-calibrated.

Table 2. Mean Absolute Error (MAE) between the true contamination factor and γGMM's sample mean for the 22 datasets.

Dataset	γGMM's sample mean	True γ^*	MAE
ALOI	0.0596	0.0304	0.0292
Annthyroid	0.0499	0.0749	0.0250
Arrhythmia	0.0385	0.0996	0.0611
Cardiotocography	0.0446	0.0496	0.0050
Glass	0.0711	0.0421	0.0290
InternetAds	0.0487	0.0499	0.0012
KDDCup99	0.0502	0.0042	0.0460
Lymphography	0.0587	0.0405	0.0182
PageBlocks	0.0638	0.1023	0.0385
Parkinson	0.0711	0.0943	0.0232
PenDigits	0.0446	0.0020	0.0426
Pima	0.0390	0.0494	0.0104
Shuttle	0.0728	0.0128	0.0600
SpamBase	0.0580	0.0500	0.0080
Stamps	0.0627	0.0912	0.0285
T15	0.0417	0.0668	0.0251
T21	0.0490	0.0529	0.0039
WBC	0.0802	0.0448	0.0354
WDBC	0.0657	0.0272	0.0385
WPBC	0.0631	0.0562	0.0069
Waveform	0.0614	0.0290	0.0324
Wilt	0.0260	0.0200	0.0060

Q3. Selecting the anomaly detectors to compute the F_1 score. Because we aim at studying the effect of the contamination factor on the detectors' performance, we compare the F_1 scores only over the detectors that work well for each of the datasets. For each dataset D, we use as set of detectors those achieving the greatest F_1 score using the true contamination factor, i.e. $\arg\max_{f_m} \{F_1(f_m, D, \gamma^*)\}$. This means that, for each dataset, we (1) use each detector separately to make predictions using the true contamination factor γ^*, (2) measure their F_1 score, (3) keep those detectors that obtain the greatest F_1, and (4) use them to compute the F_1 deterioration using the point-estimates of the contamination factor. Table 3 lists the detectors used for each dataset to compute the F_1 deterioration. Observe that sometimes only a single detector obtains the greatest F_1 score, while sometimes several detectors get the same F_1 score.

Q3. False alarms and false negatives. Finally, Table 4 shows the false alarm (false positive) rate and the false negative rate. The majority of the threshold estimators provide extremely high estimates for the contamination factor, shown here as extremely low false negative rates, but they would be useless in practice because of their high false alarm rate. In fact, this metric is important as false alarms result in real costs for the company (e.g., turning the wind turbine off to wait until the ice on the blades melts down), while reducing trust in the detection system. Our method reduces the false alarm rate compared to most of the baselines, including QMCD and KARCH that achieve the second and third best F_1 scores on average. On the other hand, IQR and MTT have the lowest false positive rates, due to the fact that they often underestimate the contamination factor as supported by the false negative table.

Table 3. List of detectors with the greatest F_1 score when using the true contamination factor to set the threshold. For each dataset, we use such a subset of detectors to compute the deterioration.

Dataset	Anomaly Detectors
ALOI	KNN
Annthyroid	HBOS
Arrhythmia	IForest-HBOS-COPOD
Cardiotocography	KNN
Glass	LOF
InternetAds	LSCP
KDDCup99	COPOD
Lymphography	KNN-LOF-OCSVM-HBOS
PageBlocks	LOF
Parkinson	LSCP-HBOS-COPOD-LSCP-HBOS-COPOD
PenDigits	KNN-IForest-LOF-OCSVM-LSCP-Ae-VAE-HBOS-LODA-COPOD
Pima	IForest
Shuttle	KNN-OCSVM-Ae-VAE-HBOS-KNN-OCSVM-Ae-VAE-HBOS
SpamBase	LSCP
Stamps	LSCP
T15	OCSVM
T21	OCSVM
WBC	KNN-LOF-OCSVM-LODA-COPOD
WDBC	KNN-LOF-OCSVM-LSCP-Ae-VAE-LODA-COPOD
WPBC	OCSVM
Waveform	OCSVM
Wilt	LOF

Table 4. Mean and standard deviation of the false alarm rate (left) and false negative rate (right) obtained by using each method's γ estimate to set the threshold (the lower the better). Regarding the false alarms, γGMM has the third best mean and outperforms QMCD and KARCH, which are the second and third best baseline when measuring the F_1 score. On the other hand, γGMM obtains higher false negative rates than almost all the competitors, due to the fact that the threshold estimators overestimate the true contamination factor.

False Alarm Rate	
Method	Mean $\pm$ std.
IQR	0.009 ± 0.008
MTT	0.027 ± 0.024
γGMM	**0.042 ± 0.015**
QMCD	0.059 ± 0.018
KARCH	0.147 ± 0.047
CHAU	0.190 ± 0.035
ZSCORE	0.221 ± 0.050
YJ	0.390 ± 0.139
FILTER	0.454 ± 0.054
DSN	0.477 ± 0.134
HIST	0.513 ± 0.100
FGD	0.533 ± 0.183
AUCP	0.591 ± 0.088
MCST	0.611 ± 0.287
GESD	0.616 ± 0.106
REGR	0.643 ± 0.105
MAD	0.731 ± 0.083
CLF	0.757 ± 0.077
EB	0.785 ± 0.077
WIND	0.809 ± 0.076
MOLL	0.816 ± 0.082
BOOT	0.862 ± 0.079

False Negative Rate	
Method	Mean $\pm$ std
BOOT	0.001 ± 0.002
WIND	0.001 ± 0.002
MOLL	0.001 ± 0.002
EB	0.001 ± 0.003
MAD	0.002 ± 0.003
CLF	0.002 ± 0.003
GESD	0.003 ± 0.005
REGR	0.003 ± 0.005
AUCP	0.004 ± 0.006
HIST	0.006 ± 0.008
FGD	0.007 ± 0.011
DSN	0.007 ± 0.009
FILTER	0.007 ± 0.009
MCST	0.007 ± 0.012
YJ	0.009 ± 0.010
ZSCORE	0.017 ± 0.015
CHAU	0.019 ± 0.016
KARCH	0.021 ± 0.018
QMCD	0.034 ± 0.024
γGMM	**0.036 ± 0.025**
MTT	0.042 ± 0.028
IQR	0.044 ± 0.029