# ROBUST FILTERING AND CORRELATION-ALIGNED DIFFUSION FOR UNSUPERVISED TABULAR ANOMALY DETECTION

**Anonymous authors**
Paper under double-blind review

## ABSTRACT

Unsupervised anomaly detection trains on unlabeled data that contain anomalies. We address this contamination with two data-centric steps. First, a FastMCD-based pre-training filter removes rows outside the robust covariance support; it improves most detectors we compare. Second, we replace only the isotropic forward corruption of Diffusion Time Estimation (DTE), a diffusion-based detector, with correlation-aligned noise. Estimating this correlation from the retained rows, anomalies included, beats every cleaner set of rows we tested. Anisotropy-aware DTE (ADTE) combines these two steps and, in its categorical variant ADTE-C, a regression head that stabilizes results across seeds. On the 57 ADBench groups under the unsupervised protocol, ADTE-C reaches 77.0 AUCROC, to our knowledge the highest reported. Removing any one of the three components significantly lowers ADTE-C's mean AUCROC.

## 1 INTRODUCTION

In unsupervised anomaly detection, the training data are unlabeled and contain an unknown fraction of anomalies. Under the unsupervised protocol used in our evaluation, a detector is trained on the same unlabeled dataset it must rank, so the anomalies to be found are in its training data. ADBench's authors describe this setting as the one unsupervised detectors are primarily designed for (Han et al., 2022), and ODIM and ALTBI report under it (Kim et al., 2024; Cho et al., 2025). Our benchmark is the 57 groups of ADBench, 47 tabular and 10 of image and text embeddings (Section 4.1).

Diffusion Time Estimation (DTE; Livernoche et al., 2024) scores a row by the diffusion timestep a network predicts for it (Section 3). DTE also has a non-parametric estimator based on a nearest-neighbor approximation to the posterior over diffusion time, yielding a kNN-style distance score (Livernoche et al., 2024). Under isotropic corruption, the distance in this approximation is Euclidean; once the corruption follows the data's correlation it is Mahalanobis. Both kNN variants are therefore among the detectors we compare: Mahalanobis kNN's inverse-correlation metric downweights the leading directions (Shyu et al., 2003), the correlation's eigenvectors with the largest eigenvalues, whereas corruption aligned with the correlation places more noise variance along them (Sections 4.4 and 4.6). Two properties of DTE matter here. First, its score reflects how far a row lies from the rows the network was trained on, so under the unsupervised protocol the anomalies are part of the network's own reference set. Second, its isotropic corruption ignores the correlation structure of real tabular features.

We introduce Anisotropy-aware DTE (ADTE), which changes DTE in three places: (i) a pre-training robust-support filter, the minimum covariance determinant (MCD) filter, which removes rows that lie far from a robust estimate of the data's center and covariance; (ii) correlation-aligned corruption, which changes only DTE's forward noise, from isotropic to aligned with the correlation of the rows the filter retains; and (iii) a regression head that predicts the timestep itself, added only to the categorical variant ADTE-C, which predicts the timestep as one of $K$ bins. ADTE-C reaches a mean AUCROC of 77.0 on the 57 groups, the highest among the detectors we compare under the unsupervised protocol (Table 1).

We make three contributions.



Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

- **Robust filtering improves detectors of many kinds, not only ours.** One fixed filter, applied before training, improves detectors in each of Table 1's three blocks (classical, closed-form and neighbor-based detectors; recent learned detectors; the DTE family), with the largest gains for the Euclidean neighbor-based detectors (Section 4.2).
- **Aligning DTE's corruption with the data's correlation adds a gain on top of filtering, and this gain depends on the orientation of the correlation.** The gain is significant on the 57 groups and on the 47 tabular groups alone (Sections 4.3 and 4.6). Estimating the correlation from the rows the filter retains, anomalies included, beats even the cleaner pools: the labeled normal rows, and the retained rows with their anomalies deleted or replaced. Whitening the inputs, which makes the same corruption isotropic, is not a substitute for aligning the corruption (Section 4.5).
- **We show which anomaly geometries alignment helps, and where learned scoring improves on classical scoring.** On the real tabular groups, the gain is larger where a larger share of the anomalies' squared deviation from the normal mean lies along the leading eigenvectors of the normals' correlation, an association computed with the labels; synthetic anomalies planted along the leading eigenvector are detected better with alignment and those planted in three other geometries, a minor eigenvector among them, worse (Section 4.4). On tabular data, after the same filter, the network with isotropic corruption is not significantly better than Euclidean kNN or the closed-form Mahalanobis distance $T^2$; the aligned model has the highest mean AUCROC of the detectors we compare (Section 4.6).

## 2 Related Work

Unsupervised detectors score a row by the distance to its nearest neighbors or its local density relative to its neighbors (Ramaswamy et al., 2000; Breunig et al., 2000), by a network trained on its nearest-neighbor distances, as in LUNAR (Goodge et al., 2022), or by how few random splits are needed to isolate it (Liu et al., 2008). Other detectors score by tail probabilities under the empirical marginals or an empirical copula, as ECOD and COPOD do (Li et al., 2023; 2020), by an estimated support of the normal data (Schölkopf et al., 2001; Ruff et al., 2018), by autoencoder reconstruction error or latent density (Sakurada & Yairi, 2014; Zong et al., 2018), or by the loss of a self-supervised task, which in SLAD predicts the scale of random feature subvectors (Xu et al., 2023). A Mahalanobis distance under a robust covariance estimate is another such score (Rousseeuw & Van Driessen, 1999); we use the robust covariance not as the anomaly score but to select the training rows (Section 3.2).

Existing methods address training-data contamination by modifying the objective, refining the training rows before the final fit, or intervening during training. Latent Outlier Exposure modifies the objective, inferring a normal-or-anomalous label per row while updating the model (Qiu et al., 2022). SRR refines the training rows with an ensemble of one-class classifiers and repeats (Yoon et al., 2022); DSIL, demonstrated on DTE, progressively refines the unlabeled set by identifying and removing anomalies (Vincent et al., 2025); and SROC fits a detector on pre-trained image features, removes the highest-scoring training images and refits on the rest (Cordier et al., 2022). During training, RCA's collaborative autoencoders drop the rows with the largest reconstruction errors (Liu et al., 2021); ODIM and ALTBI exploit a network's tendency to fit normal rows first, ODIM by early stopping (Kim et al., 2024) and ALTBI by dropping high-loss rows from the objective (Cho et al., 2025); and K-DSM, a score-matching detector, removes low-density training points before each gradient step (Livernoche et al., 2026). The MCD fits location and scatter to the subset of rows of a given size whose covariance has the lowest determinant; FastMCD approximates it (Rousseeuw & Van Driessen, 1999). A recent tabular anomaly detection study discusses robust methods such as the MCD for handling training data with an unknown contamination level (Wei & Armanfard, 2026). Our filter leaves every detector's objective and training procedure unchanged: it is one fixed, detector-independent rule, the reweighted support of the MCD estimate, applied once before training, and it keeps a data-dependent share of the rows rather than a preset one. Section 4.2 measures its effect on every detector in Table 1, including ODIM, ALTBI-I and K-DSM.

Some diffusion-based detectors score by reconstruction error or denoising loss: DDAE scores by denoising reconstruction error (Sattarov et al., 2025), and TabADM, trained on contaminated tabular data, scores a row by its denoising loss summed over the diffusion timesteps and drops each minibatch's highest-loss rows during training (Zamberg et al., 2023). DTE's target is instead the

scalar timestep, whose definition does not involve the noise covariance, so replacing isotropic noise with correlated Gaussian noise leaves the target and the loss forms unchanged (Section 3.2).

Non-isotropic Gaussian corruption has been studied in generative diffusion: score-based diffusion has been trained with a fixed non-isotropic Gaussian noise model (Voleti et al., 2022) and with arbitrary Gaussian forward noise (Alido et al., 2025). More recently, the forward noise has been shaped by the data or chosen as an inductive bias: scaled anisotropically along the training distribution's principal component in image generation (Singh et al., 2025), controlled by frequency (Jiralerspong et al., 2025), or given a covariance that is diagonal in the frequency basis (Scimeca et al., 2025). K-DSM also sets its perturbation noise from the data, but per feature, with kurtosis-guided noise scales under a diagonal covariance, and it scores a row by the norm of the noise-scaled score its network estimates (Livernoche et al., 2026). Our corruption instead follows the data's correlation matrix, not one principal component or a per-feature scale; we estimate the correlation from the rows the MCD filter retains, anomalies included, and the detector scores by the predicted timestep. Sections 4.3 and 4.6 measure the gain. Sections 4.4 to 4.6 show how it depends on the anomalies' geometry, on the rows that estimate the correlation, and on whether the correlation enters as corruption or as a metric.

Joint classification–regression training has precedent in a multi-task age estimator that adds classification as an auxiliary task to regression (Liu et al., 2020). DTE itself compares an $\ell_2$ regression of the timestep with its inverse Gamma and categorical variants (Livernoche et al., 2024), and MoreRed regresses a normalized diffusion time (Kahouli et al., 2024). In multi-task learning, the training signals of related tasks can improve generalization (Caruana, 1997). Our regression head is trained jointly with the categorical objective on the same timestep, and its output enters the anomaly score at equal weight with the categorical one.

# 3 METHOD

## 3.1 BACKGROUND

Diffusion Time Estimation (DTE; Livernoche et al., 2024) is an anomaly detection method based on a forward diffusion process. Given a data point $x \in \mathbb{R}^D$, DTE constructs noisy samples by injecting isotropic Gaussian noise with a variance that grows according to a timestep $t$:

$$x_t = x + \sqrt{1 - \bar{\alpha}_t}\,\epsilon, \quad \epsilon \sim \mathcal{N}(0, I_D), \tag{1}$$

where $\bar{\alpha}_t = \prod_{j=0}^{t}(1 - \beta_j)$ with $\beta_j$ following a linear schedule from $\beta_0$ to $\beta_{T-1}$ over $T$ timesteps, and $t$ is drawn uniformly from $\{0, \ldots, T-1\}$ during training. DTE omits the mean-shrinking term of the standard denoising diffusion probabilistic model (DDPM; Ho et al., 2020). A neural network $f_\theta$ is trained to predict the diffusion timestep from the corrupted sample $x_t$. At inference, an unmodified data point is scored by the network's estimated diffusion timestep; anomalous points tend to lie farther from the normal data manifold and receive higher estimates.

DTE admits two parametric variants. The *categorical variant* (DTE-C) discretizes the diffusion timesteps into $K$ bins via $b(t) = \min(\lfloor t \cdot K/T \rfloor, K-1)$ and trains a network $f_\theta$ whose softmax output $f_\theta(x)$ is a probability vector over the $K$ bins. The released implementation passes this probability vector to a cross-entropy loss, which applies a further log-softmax, so the objective minimized is[1]

$$\mathcal{L}_{\text{cat}} = \mathbb{E}_{x,t}\left[\log \sum_{k=0}^{K-1} \exp\left(f_\theta(x_t)_k\right) - f_\theta(x_t)_{b(t)}\right]. \tag{2}$$

The anomaly score is the expected bin index $\hat{t}_{\text{cat}}(x) = \sum_{k=0}^{K-1} k \cdot f_\theta(x)_k \in [0, K-1]$. The *inverse Gamma variant* (DTE-IG) instead models the diffusion variance with an IG posterior, fits its scale by negative log-likelihood, and uses a scale-based score (Livernoche et al., 2024; Appendix H).

## 3.2 DATA-CENTRIC CORRECTIONS

Under the unsupervised protocol (Section 1), training on contaminated data can degrade neighbor-based detectors (Gu et al., 2019), which motivates filtering for DTE given the connection of its

[1]The original DTE paper writes the ordinary cross-entropy $-\log f_\theta(x_t)_{b(t)}$; its released code, provided to reproduce its results and used for our DTE results, optimizes the objective above.

non-parametric estimator to kNN (Livernoche et al., 2024, Section 3.2); Section 4.2 quantifies the filtering gains across detector families. ADTE combines the MCD filter and correlation-aligned corruption (this section) with, for the categorical variant, a regression head (Section 3.3); ADTE-C and ADTE-IG build on DTE-C and DTE-IG, respectively.

**The MCD filter.** To address contamination, we apply a fast preprocessing step based on the FastMCD estimator (Rousseeuw & Van Driessen, 1999). All rows of the unlabeled dataset are z-score standardized, projected to a lower-dimensional principal component analysis (PCA) subspace when dimensionality exceeds a threshold (Section 4.1), and filtered to remove suspected anomalies falling outside the robust covariance support.

We retain the estimator's reweighted support: rows whose squared Mahalanobis distance under the consistency-corrected raw estimate is below the 0.975 quantile of a $\chi^2_d$ distribution, where $d$ is the number of dimensions used to fit the estimator. The support fraction sets the size of the raw subset, not the fraction of rows kept; if the estimator fails or fewer than ten rows remain, we use all rows. A DTE model is then trained on the retained rows in the original standardized feature space; these rows form $X_{\text{filtered}}$. At inference every row of the dataset, retained or not, is scored.

**Correlation-aligned corruption.** To align the diffusion process with the underlying data geometry, we replace the isotropic noise with correlation-aligned noise $\epsilon \sim \mathcal{N}(0, \Sigma)$, where

$$\Sigma = (1 - s) \cdot \mathrm{Corr}(X_{\text{filtered}}) + s \cdot I \tag{3}$$

interpolates between the original isotropic noise ($s$=1, $\Sigma = I$) and fully correlation-aligned noise ($s$=0, $\Sigma = \mathrm{Corr}(X_{\text{filtered}})$); the correlation is computed over all retained rows, including any anomalies the filter retained; a feature constant on the retained rows enters $\Sigma$ with unit variance and zero correlation. We use $s$=0 (full alignment) for ADTE-C and $s$=0.75 for ADTE-IG; the shrinkage ablation is in Table 4 and Appendix F. In practice, if the Cholesky factorization of $\Sigma$ fails (near-singular at $s$=0), diagonal jitter is added at escalating levels ($10^{-6}$, $10^{-5}$, $10^{-4}$, $10^{-3}$); if all fail, the model falls back to isotropic noise (in our runs, $10^{-6}$ always sufficed and the fallback never triggered). For $s > 0$, the shrinkage term $s \cdot I$ ensures positive definiteness directly.

In the following equations, $\Sigma$ denotes the covariance used after any diagonal jitter or isotropic fallback. The corruption distribution becomes

$$q_\Sigma(x_t \mid x, t) = \mathcal{N}\left(x, (1 - \bar{\alpha}_t)\Sigma\right). \tag{4}$$

Thus, in the corruption density, the Euclidean deviation $\|x_t - x\|_2^2$ of isotropic DTE is replaced by the Mahalanobis deviation $(x_t - x)^\top \Sigma^{-1} (x_t - x)$. Equivalently, aligned corruption of the standardized rows is isotropic corruption of the whitened rows $\Sigma^{-1/2} x$. Since $\Sigma$ is fixed with respect to $t$ and the training targets of DTE-C and DTE-IG depend on the timestep alone, changing the corruption covariance modifies only the noise sampler; the network architecture, output format, and loss forms are unchanged. Appendix H extends DTE's inverse Gamma approximation (Livernoche et al., 2024, Sections 3.1 and 3.3) to the covariance $\Sigma$ and records the inherited implementation, including where its scoring differs from that paper's text.

### 3.3 A Jointly Trained Regression Head

We add a second head that regresses the timestep itself; in the full pipeline it raises mean AUCROC and narrows the spread of AUCROC over training seeds (Section 4.3).

The head predicts the normalized timestep $t/T \in \{0, 1/T, \ldots, (T-1)/T\}$, trained jointly with the categorical objective with head weight $\lambda$:

$$\mathcal{L} = \mathcal{L}_{\text{cat}} + \lambda \, \mathcal{L}_{\text{reg}}, \quad \mathcal{L}_{\text{reg}} = \mathbb{E}_{x,t}\left[\left(g_\phi(h_\theta(x_t)) - t/T\right)^2\right], \tag{5}$$

where $h_\theta(x) \in \mathbb{R}^{d_h}$ is the shared representation from the penultimate layer of $f_\theta$, and $g_\phi : \mathbb{R}^{d_h} \to (0, 1)$ is a lightweight network (Linear $\to$ ReLU $\to$ Linear $\to$ Sigmoid).

At inference, the two scores are placed on a common $[0, 1]$ scale by rescaling the expected bin index $\hat{t}_{\text{cat}}(x)$ (Section 3.1) to $\bar{t}_{\text{cat}}(x) = \hat{t}_{\text{cat}}(x)/(K-1) \in [0, 1]$, and the final anomaly score combines them:

$$\hat{t}(x) = \omega \cdot g_\phi(h_\theta(x)) + (1 - \omega) \cdot \bar{t}_{\text{cat}}(x), \tag{6}$$

with a fixed $\omega = 0.5$, giving equal weight to both outputs. The inference weight $\omega$ is separate from the training weight $\lambda$, which ADTE-C sets to $0.1$.

# 4 EXPERIMENTS

## 4.1 SETUP

We evaluate on the 57 groups of ADBench (Han et al., 2022): 47 tabular groups and 10 image and text groups (5 of each) given as fixed embeddings, each of the 10 scored as the mean over its member datasets. Every detector is trained on the whole unlabeled dataset, anomalies included, or, with the filter, on the rows the filter retains, and scores every row of the dataset.

Unless otherwise stated, numbers are means over five seeds and over the 57 groups, in percentage points of AUCROC and AUCPR (the areas under the ROC and the precision–recall curves; AUCPR is computed as average precision); differences and sums, in the tables and in the text, are computed from unrounded values. A "$\pm$" is the sample standard deviation (SD) of the five seed-level means, one 57-group mean per seed. Intervals are 95% bootstrap intervals, obtained by resampling the groups with replacement; a difference is significant when its interval excludes zero, and a non-significant difference is not evidence of no difference. Paired tests, where quoted, are two-sided Wilcoxon signed-rank tests (Wilcoxon, 1945) over group-level differences, except where a bootstrap $p$-value is named, with the $p$-value reported alongside rather than as a second criterion. Detectors run with the filter unless a filtering control is specified. The appendix gives the detector comparison split by modality, per-group results for ADTE-C, isotropic DTE-C (ADTE-C without alignment, with $K = 8$ and the head), ADTE-IG, both kNNs and $T^2$, the exclusion rules behind any reduced group count, and sensitivity checks on $K$, $s$, $\lambda$, $\omega$ and the filter's settings; per-group results for the remaining experiments are provided with the code.

The configuration of Section 3 is fixed for every dataset: for ADTE-C, $K = 8$ bins (DTE's published default is 7), $\lambda = 0.1$, $\omega = 0.5$ and $s = 0$; for ADTE-IG, which has no bins and no head, a shrinkage of $s = 0.75$; and, wherever the filter is used, an MCD support fraction of 0.75, with the estimator fitted on a 32-dimensional PCA projection of the rows when the data have more than 32 dimensions. DTE-NP, DTE's non-parametric estimator, runs at its release defaults (Appendix A). The filter keeps a data-dependent share of the rows, a median of 73% across the 57 groups (range 60–97%). We keep the diffusion schedule, multilayer perceptron (MLP) backbone and optimization settings of DTE's released implementation unchanged (Livernoche et al., 2024); Appendix A lists them. All GPU experiments ran on NVIDIA Ada Lovelace hardware: the comparison and ablation results (Tables 1–6, Section 4.6 and Appendix E) on an RTX 4000 Ada, and the analysis experiments (the controls of Section 4.2, Sections 4.4 and 4.5, and the eigenvalue cap of Section 4.6) on L4 GPUs; ADTE-C's 57-group mean differs by 0.13 AUCROC across the two. The classical, closed-form and neighbor-based detectors of Table 1, LUNAR excepted, and DTE-NP run on CPU; Table 28 in Appendix L lists every experiment's hardware. Baselines are re-run under the same pipeline, with and without the filter, each with the same configuration for every dataset; ALTBI-I (ALTBI with an importance-weighted autoencoder) and ODIM alone use per-dataset network sizes, the ones in their released implementations. Published numbers for the baselines follow their own protocols, metric conventions and dataset lists (ODIM and ALTBI-I report under the unsupervised protocol) and are compared with ours in Appendix A. Where the protocol matches, our numbers agree with the published ones (Appendix A, Table 11).

## 4.2 WHICH DETECTORS BENEFIT FROM FILTERING?

With the filter, ADTE-C has the highest mean AUCROC across the 57 groups among the detectors of Table 1, and ADTE-IG the highest AUCPR; we use ADTE-C for the alignment analyses that follow. The filter's benefit is not specific to our detector: it raises the mean AUCROC of 18 of the 19 detectors of Table 1, with the largest gains, 9.3 to 13.7 points, for the Euclidean neighbor-based ones. After the filter, DTE-NP ranks rows almost exactly as a nearest-neighbor detector (Appendix A), and LUNAR and both kNNs have mean AUCROC within 0.2 points of isotropic DTE-C.

Two comparisons show what the gain depends on. Which rows are removed matters, not only how many: each of eight filtering policies gives ADTE-C a higher mean AUCROC than no filter, the MCD filter the highest, ahead of ALTBI-I and ODIM used as filters (Appendix C, Table 18). Removing

Table 1: Every detector without and with the MCD filter under the unsupervised protocol (57 groups, %; mean ± SD; bold: the highest filtered mean of each metric). Rows form three blocks, each ordered by filtered AUCROC with the highest last: classical, closed-form and neighbor-based detectors; recent learned detectors; and the DTE family. Δ is filtered minus unfiltered AUCROC (ROC) and AUCPR (PR). Both kNN detectors use $k = 5$, as in DTE-NP's released implementation, and score each row by its distance to the fifth-nearest neighbor; Mahalanobis kNN measures distances with the inverse correlation matrix. $T^2$ is the squared Mahalanobis distance (Mahalanobis, 1936) to the mean under the covariance, and $T^2$ ($\Sigma{=}I$) the same with the identity covariance. The MCD detector is PyOD's robust-covariance Mahalanobis scorer (Zhao et al., 2019). ‡ODIM's 20news group averages the same five of its six members in both columns; the sixth could not be evaluated without the filter.

| | no filter | | + MCD filter | | Δ (filter) | |
|---|---|---|---|---|---|---|
| Method | AUCROC | AUCPR | AUCROC | AUCPR | ROC | PR |
| ECOD | 72.0±0.1 | 31.3±0.2 | 72.9±0.1 | 32.7±0.3 | +0.8 | +1.4 |
| $T^2$ ($\Sigma{=}I$) | 72.0±0.1 | 32.9±0.2 | 72.9±0.1 | 33.3±0.2 | +0.9 | +0.4 |
| COPOD | 72.3±0.1 | 30.4±0.2 | 73.1±0.1 | 30.7±0.3 | +0.8 | +0.3 |
| IForest | 74.0±0.1 | 33.4±0.4 | 74.1±0.2 | 29.0±0.4 | +0.1 | −4.4 |
| MCD detector | 75.2±0.3 | 34.2±0.8 | 74.4±0.3 | 33.7±0.7 | −0.8 | −0.5 |
| LOF | 61.4±0.4 | 18.5±0.2 | 75.1±0.2 | 33.5±0.6 | +13.7 | +14.9 |
| Mahalanobis kNN | 67.9±0.5 | 25.2±0.5 | 76.1±0.5 | 34.1±0.5 | +8.3 | +8.9 |
| LUNAR | 63.9±0.4 | 20.4±0.3 | 76.2±0.4 | 35.5±0.3 | +12.3 | +15.0 |
| Euclidean kNN | 67.2±0.4 | 24.5±0.5 | 76.5±0.4 | 36.6±0.4 | +9.3 | +12.1 |
| $T^2$ | 74.5±0.1 | 30.2±0.2 | 76.8±0.1 | 35.8±0.6 | +2.2 | +5.6 |
| ODIM‡ | 74.2±0.1 | 33.3±0.1 | 74.6±0.3 | 34.4±0.3 | +0.4 | +1.2 |
| SLAD | 66.7±0.4 | 24.0±0.4 | 75.0±0.2 | 36.5±0.7 | +8.3 | +12.4 |
| DDAE | 74.2±0.1 | 33.2±0.2 | 75.2±0.1 | 35.0±0.2 | +1.0 | +1.8 |
| ALTBI-I | 75.4±0.2 | 34.2±0.7 | 75.5±0.1 | 34.2±0.7 | +0.1 | −0.1 |
| K-DSM | 70.0±0.5 | 34.1±0.9 | 75.8±0.2 | 35.9±0.2 | +5.9 | +1.8 |
| ADTE-IG (ours) | 67.5±1.0 | 21.3±0.4 | 75.8±0.3 | **36.9**±0.4 | +8.3 | +15.7 |
| isotropic DTE-C | 74.2±0.5 | 30.8±0.5 | 76.3±0.2 | 35.9±0.7 | +2.1 | +5.2 |
| DTE-NP | 65.7±0.3 | 22.9±0.2 | 76.4±0.3 | 35.7±0.3 | +10.7 | +12.7 |
| ADTE-C (ours) | 74.2±0.9 | 33.6±0.8 | **77.0**±0.2 | 36.7±0.7 | +2.8 | +3.1 |

Table 2: Cumulative build-up from DTE-C at $K = 8$ (57 groups, %; AUCROC ± SD over seeds, and AUCPR). The last two rows are the isotropic-with-head and aligned-with-head cells of Table 3; the last row is ADTE-C.

| Configuration | AUCROC | AUCPR |
|---|---|---|
| DTE-C | 73.5±0.83 | 29.1 |
| + regression head | 74.2±0.47 | 30.8 |
| + MCD filter | 76.3±0.15 | 35.9 |
| + aligned corruption | **77.0**±0.19 | 36.7 |

Table 3: Corruption × head at $K = 8$ with the filter fixed, for DTE-C with and without the regression head (57 groups; AUCROC ± SD over seeds; effects with their intervals). Effects are differences of unrounded cell means; the interaction $p$ is from a two-sided Wilcoxon test over per-group interaction contrasts, the alignment effect with the head minus the alignment effect without it. AUCPR cells are in Appendix B, Table 15.

| | isotropic | aligned ($s{=}0$) |
|---|---|---|
| no head | 76.22±0.37 | 76.70±0.28 |
| with head | 76.30±0.15 | **77.02**±0.19 |
| alignment | +0.48 [−0.22, +1.39] (no head) | |
| | +0.73 [+0.14, +1.44] (head) | |
| head | +0.08 [−0.33, +0.49] (iso.) | |
| | +0.33 [+0.03, +0.68] (aligned) | |
| interaction | +0.25, $p = .07$ | |

Table 4: Shrinkage $s$, ADTE-C with the filter (57 groups).

| $s$ | 0 | 0.1 | 0.25 | 0.5 | 0.7 |
|---|---|---|---|---|---|
| AUCROC | 77.02 | 76.70 | 76.53 | 76.41 | 76.36 |
| AUCPR | 36.74 | 35.99 | 35.80 | 35.74 | 35.76 |

Table 5: Bin count $K$, ADTE-C with the filter (57 groups; AUCROC ± SD over seeds, and AUCPR).

| $K$ | 6 | 7 | 8 | 9 | 10 |
|---|---|---|---|---|---|
| AUCROC | 77.00±0.23 | 77.06±0.12 | 77.02±0.19 | 77.08±0.16 | 77.09±0.29 |
| AUCPR | 36.46 | 36.55 | 36.74 | 36.64 | 36.52 |

Table 6: Shrinkage $s$, ADTE-IG with the filter (57 groups).

| $s$ | 0 | 0.25 | 0.5 | 0.75 | 1 |
|---|---|---|---|---|---|
| AUCROC | 75.89 | 75.99 | 75.92 | 75.81 | 75.59 |
| AUCPR | 36.41 | 36.71 | 36.92 | 36.94 | 37.01 |

the same number of randomly chosen rows instead leaves the MCD filter significantly ahead on all four detectors tested, Euclidean kNN, LOF, isotropic DTE-C and ADTE-C (bootstrap $p$-values; Appendix C, Table 19). The training regime matters as well: when ADTE-C and isotropic DTE-C are trained on labeled normal rows only and score held-out rows, they do better without the filter on the 47 tabular groups (Appendix C).

### 4.3 What Does Each Component Contribute?

Table 2 builds the pipeline up from DTE-C at $K = 8$, ADTE-C's bin count, so that each step adds one component and none changes $K$: the regression head adds 0.7 AUCROC points before filtering, the filter 2.1, and aligned corruption 0.7, or 3.5 AUCROC points end to end. Removing any one of the three from ADTE-C costs mean AUCROC on the 57 groups: the filter 2.79 points [0.97, 4.87], alignment 0.73 [0.14, 1.44] and the head 0.33 [0.03, 0.68] (Tables 1 and 3). Averaged over all member datasets, a training run of unfiltered isotropic DTE-C at DTE's published $K = 7$ without the head takes 219 seconds; adding the filter, alignment and the head brings the full pipeline to 215 seconds, plus the filter's own time of about 2.5 seconds per dataset, measured separately on a CPU (Appendix C, Table 20). In operation counts, for $n$ retained rows in $D$ dimensions, alignment adds an $O(nD^2)$ correlation estimate and an $O(D^3)$ factorization per fit, and $O(D^2)$ work per corruption draw.

In the corruption-by-head factorial of Table 3, which fixes the filter on, alignment adds 0.5 AUCROC points without the head and 0.7 with it.

The gain from the regression head (Section 3.3) comes mostly through the lowest seed outcomes within groups: with the filter under aligned corruption it narrows the seed spread, the SD of a group's AUCROC over its five seeds, by 21% and raises the mean of the groups' lowest seed outcomes by 0.6 points [0.2, 1.2], while the highest does not change significantly. Without the filter under isotropic corruption, it narrows the spread by 34% and raises the lowest outcomes by 1.9 points [0.7, 3.2] (Appendix E). The regression output is an anomaly score in its own right: ADTE-C using it alone at inference comes within 0.07 points of the combined score in mean AUCROC over the 57 groups, and narrows the seed spread by 20%, against 21% for the combined score (Appendices E and F).

### 4.4 Which Anomaly Geometries Benefit from Alignment?

Alignment helps when the anomalies' displacements lie along the correlation's leading directions. We test this directly with planted anomalies (Table 7). For each of the 47 tabular groups we fit a Gaussian mixture to its labeled normal rows and draw a synthetic copy of the group from it, with as many normal rows and as many anomalies as the group has; each anomaly is a further draw displaced in one of four ways at Mahalanobis scale 5, calibrated from the normals' correlation (in root mean square for the sparse perturbations; construction details in Appendix D). The synthetic copy then goes through the same protocol as the real data, filter included, and we compare aligned with isotropic training of DTE-C and, as a second comparison, Mahalanobis kNN with Euclidean kNN. Table 7 reports these paired AUCROC differences for each planted geometry; it does not rank the geometries by detection difficulty. Alignment helps when the displacement lies along the correlation's leading eigenvector and hurts when it lies along a minor eigenvector, on a few sparse coordinates or in a random direction; the kNN contrast has the opposite sign in every row. Isotropic corruption and the Euclidean metric both ignore the data's correlation, and they do so in opposite ways: isotropic noise is smallest relative to the spread along the leading directions, while the Euclidean distance is dominated by them (the Gaussian case is worked out in Appendix D).

Table 7 does not establish whether alignment's benefit is specific to leading-direction anomalies or extends to other directions when the corruption favors them. A second experiment on the same synthetic normals tests this distinction: two anomaly sets are displaced along two orthogonal directions of similar variance, neither of them the leading one. We train two aligned models, each with the corruption variance increased along a different one of these directions. Each anomaly set is scored by both models, by the one inflated along its own direction (matched) and by the one inflated along the other (crossed). Mean AUCROC is 4.1 points [2.4, 6.0] higher for matched than for crossed pairings (43 of the 47 groups, one to five seeds; 32 wins and 11 losses; Appendix D). Averaged over both anomaly sets, the two models do not differ significantly: their mean AUCROC difference is $-0.2$ points [$-0.7$, +0.2]. The gain comes from the match: under this construction the trained detector detects better along the direction its corruption was inflated.

On the real tabular datasets, a label-based measure of anomaly geometry is associated with the alignment gain. We measure, for each of the 47 groups, the share of its anomalies' squared deviation from the normal mean that lies along the leading directions of the labeled normals' correlation (its top eigenvectors, as many as needed to carry a tenth of the total variance; on 42 of the 47 groups that is the single leading eigenvector), pooled over the group's anomalies and computed with the labels. The alignment gain correlates with that share (Spearman $\rho = .31$, $p = .03$, permutation test). In repeated random splits of the groups, the median share from one half is used to divide the other half; on average, held-out groups above this threshold gain 1.7 AUCROC points more than those below it (permutation $p = .01$, one-sided; Appendix F). Within datasets, a separate measure uses the retained rows' mean and correlation: anomalies that rise in rank under aligned training tend to be those whose deviation lies more within a wider leading eigenspace of the corruption's own covariance, the retained rows' correlation. This subspace is spanned by the leading eigenvectors that carry 90% of that matrix's total variance (45 groups; Appendix F).

### 4.5 Which Correlation Pools and Coordinate Systems Carry the Gain?

ADTE-C uses the correlation to shape its training corruption, while $T^2$ (with the correlation in place of its covariance) and Mahalanobis kNN use it as a distance metric. Alignment changes only the

Table 7: Planted anomalies at Mahalanobis scale 5 on synthetic copies of the 47 tabular groups (mean AUCROC difference in points, with interval); each column is a contrast within one pair of detectors, aligned minus isotropic DTE-C or Mahalanobis minus Euclidean kNN, all with the filter. Leading and minor directions are the leading and the smallest non-degenerate eigenvectors of the labeled normals' correlation; construction in the text and Appendix D.

| Planted geometry | DTE-C aligned − isotropic | kNN Mahalanobis − Euclidean |
|---|---|---|
| leading direction | +5.60 [+3.69, +7.68] | −4.55 [−7.14, −2.23] |
| minor direction | −3.80 [−6.01, −1.85] | +5.95 [+2.83, +9.04] |
| sparse coordinates | −3.32 [−4.44, −2.26] | +9.10 [+6.01, +12.24] |
| random direction | −2.60 [−3.73, −1.57] | +10.22 [+6.43, +14.03] |

Table 8: Correlation pools and orientations for ADTE-C's corruption (mean AUCROC difference from the reference row in points on the 47 tabular groups; interval; training rows and scored rows fixed). The labeled-normal row uses the labels. Rotation rows turn the correlation's eigenvectors toward a random basis, eigenvalues kept; the stated angles are the mean principal angles between the original and the rotated leading eigenspaces (dimension defined in Appendix F); the sign-flip row reverses the sign of the correlations between a random half of the coordinates and the rest, eigenvalues kept. Deletion and replacement pools, and the constructions, in Appendices G and F.

| Pool or orientation | AUCROC minus ours |
|---|---|
| retained rows (ours) | 0 (reference) |
| all rows | −0.54 [−1.35, +0.04] |
| labeled normal rows | −0.94 [−1.83, −0.26] |
| none (isotropic) | −0.83 [−1.71, −0.14] |
| rotated 12° | −0.53 [−1.05, −0.07] |
| rotated 23° | −0.83 [−1.48, −0.24] |
| rotated 38° | −1.15 [−1.99, −0.44] |
| rotated 49° | −1.15 [−2.15, −0.31] |
| random basis, same eigenvalues | −1.11 [−2.23, −0.19] |
| signs flipped on half the coordinates | −0.72 [−1.68, +0.06] |

Table 9: Five detectors after the same filter (%): AUCROC means on the 47 tabular and the 10 embedding groups, and the paired contrasts on the 47 (mean, interval, wins/losses).

| Detector | 47 tabular | 10 embedding | 47 ADTE-C − detector |
|---|---|---|---|
| ADTE-C | **79.64** | 64.72 | 0 (reference) |
| $T^2$ | 78.92 | 66.68 | +0.72 [−0.18, +1.83], 26/21 |
| isotropic DTE-C | 78.72 | 64.91 | +0.92 [+0.23, +1.73], 25/22 |
| Euclidean kNN | 78.51 | 66.89 | +1.13 [−0.09, +2.43], 22/25 |
| Mahalanobis kNN | 77.71 | 68.62 | +1.93 [+0.69, +3.28], 26/21 |

noise sampler (Section 3.2), so its effect depends on which rows estimate the sampler's correlation, the correlation pool; in the pipeline the pool is the retained rows, anomalies included. The upper rows of Table 8 vary that pool on the 47 tabular groups. Relative to the retained rows we use, estimating the correlation from all rows of the dataset, anomalies included, gives a lower mean AUCROC, and estimating it from the labeled normal rows alone costs 0.9 points, about as much as removing alignment altogether (0.8). Even the all-rows pool, which differs from the labeled normals only by the dataset's anomalies, is ahead of them by 0.39 points [0.11, 0.72]. When each group is split in half, with the filter and the pools built on one half and the other half scored, the labeled-normal pool costs 0.67 points [0.10, 1.39] relative to the retained pool on the 47 groups (Appendix G).

The filter retains anomalies in 35 of the 47 groups; across these 35 groups, the median fraction of retained rows that are anomalous is 4%. Part of the retained pool's advantage comes from those anomalies: deleting them from the pool costs 0.42 AUCROC points [0.08, 0.83] on 34 of those groups (ten seeds; Appendix G), and replacing them with an equal number of filter-rejected normal rows drawn at random costs 0.52 points [0.10, 0.99] on the 25 where the exchange is possible (Appendix G). On the same 35 groups, including the retained anomalies in the pool raises the correlation's variance along the direction from the retained normals' mean to the retained anomalies' mean by 0.37 [0.17, 0.59] in correlation units (every coordinate has variance one; it rises in 30 of the 35 groups; Appendix I), so the retained anomalies tilt the corruption toward the directions they occupy.

While ADTE-C prefers the retained pool, $T^2$ and Mahalanobis kNN prefer the labeled-normal one: the labeled-normal correlation that costs ADTE-C 0.9 points improves $T^2$ by 4.6 [2.1, 7.3] and Mahalanobis kNN by 1.6 [0.6, 2.7] (Appendix G). For those two detectors, which use the correlation as a metric, the preference for the labeled-normal pool is consistent with masking (Rousseeuw & van Zomeren, 1990): anomalies that enter the correlation estimate make themselves look less anomalous under the resulting metric.

The gain also depends on the coordinate system the network trains in. Aligned noise on the standardized data and isotropic noise on whitened data are the same corruption in two coordinate systems (Section 3.2), and the distribution of the timestep given the corrupted point is the same in both, so the network faces the same prediction problem (Appendix H). Yet, averaged over 43 of the 47 groups (three of them with fewer than five seeds; Appendix G), the isotropic model trained in whitened coordinates scores 4.2 AUCROC points [1.0, 8.3] below the aligned model on the standardized data; the difference must come from how the network learns in each coordinate system, so whitening the inputs is not a substitute for aligning the corruption.

The alignment gain depends on the orientation of the correlation beyond its eigenvalues (Table 8): the mean AUCROC cost grows with the rotation of its eigenvectors up to about 40° and then levels off at the cost of a random basis with the same eigenvalues. Even flipping the signs of the correlations between a random half of the coordinates and the rest, which keeps every variance, correlation magnitude and eigenvalue and changes only the eigenvectors, costs 0.72 points [−0.06, 1.68]. Alignment is not a change of noise scale either: we test five multipliers of the default per-timestep noise standard deviation, $m \in \{0.5, 1, 2, 3, 5\}$. The isotropic model's mean AUCROC is highest at $m = 1$, and each of the five isotropic means is significantly below the aligned model's mean at $m = 1$ (Appendix F).

### 4.6 When Does Learned Scoring Improve on Classical Scoring?

Alignment lifts the network to the highest tabular mean AUCROC among five detectors: Table 9 compares Euclidean kNN, Mahalanobis kNN and $T^2$, three classical detectors that need no network training, with isotropic DTE-C and ADTE-C, on the 47 tabular groups. Isotropic DTE-C is not significantly better than Euclidean kNN or $T^2$ (intervals in Appendix B). ADTE-C is significantly above its isotropic version (+0.92 AUCROC points [+0.23, +1.73]) and above Mahalanobis kNN (+1.93 [+0.69, +3.28]). ADTE-C uses the correlation only in training and inverts no covariance at inference; $T^2$ and Mahalanobis kNN score every row through an inverted covariance or correlation matrix. As in Section 4.4, the same correlation moves the two scorers in opposite directions: as kNN's distance metric it changes Mahalanobis kNN by −0.80 points [−1.83, +0.20] relative to Euclidean kNN, against the +0.92 it adds to the network, a gap of 1.72 points [0.45, 3.27]. On the 47 tabular groups alone, as on the 57, ADTE-C has the highest mean AUCROC of the detectors in Table 1 (per-detector tabular means in Appendix B, Table 13).

The 10 image and text groups reverse this: there the three classical detectors are all above both DTE-C rows; Euclidean kNN is 2.2 AUCROC points [1.0, 3.5] above ADTE-C and 2.0 [0.5, 3.5] above isotropic DTE-C, so most of the deficit is there before alignment (Table 9; per-group results in Appendix B, Table 16). Capping the five largest eigenvalues of the corruption's correlation at the sixth, with the matrix rescaled so that its trace stays $D$, raises ADTE-C on these 10 groups by 0.80 AUCROC points [0.23, 1.45], nearly all of it on the five text groups (1.58 points [0.86, 2.30], all five up; Appendix K).

## 5 Conclusion

We address contaminated training data in unsupervised anomaly detection with two data-centric steps. The MCD filter raises the mean AUCROC of most detectors we compare, the Euclidean neighbor-based ones most of all. Aligning DTE's corruption with the correlation of the retained rows, anomalies included, adds a significant gain and gives ADTE-C the highest mean AUCROC in Table 1. A jointly trained regression head adds a smaller, significant gain and stabilizes the categorical variant across seeds.

On planted anomalies, alignment is an inductive bias relative to isotropic corruption: it favors anomalies displaced along the leading directions of the normal rows' correlation, at the expense of those in the other tested geometries. Mahalanobis kNN uses the correlation as a distance metric and shows the reverse pattern relative to Euclidean kNN on the same planted anomalies (Table 7). On the real tabular groups, the alignment gain is larger where a larger share of the anomalies' squared deviation lies along the leading directions. Beyond shaping the corruption, the rows' second-order structure helps as a metric: $T^2$ has a higher mean AUCROC under its own rows' covariance than under the identity covariance, with and without the filter (Table 1). While $T^2$ and Mahalanobis kNN have a higher mean AUCROC with the correlation estimated from the labeled normal rows than from the retained rows, ADTE-C shows the reverse (Section 4.5). The retained anomalies tilt ADTE-C's corruption toward the directions they occupy. Our alignment suits the tabular groups, where its gain arises; on the 10 embedding groups the three classical detectors of Table 9 remain ahead of both DTE-C rows, and a differently shaped corruption, with the leading eigenvalues capped, recovers part of the gap (Section 4.6; Appendix M states the other bounds of our evidence). Two questions reach beyond unsupervised anomaly detection: which examples a model should learn from when its training data may be contaminated, and which directions its training noise should emphasize.

### AI USE STATEMENT

The research questions, the method and the conclusions of this work are the authors'. Under our direction, generative AI tools assisted with formulating and checking mathematical statements and writing out the derivations of Appendix H, proposing and refining hypotheses, designing experiments, implementing and testing code, and interpreting results. They also assisted with writing and editing the text and with searching the literature. No reported number or table entry was produced by a generative AI tool; all were computed by the released code. We reviewed all AI-assisted work: every statement, number and citation was checked by the authors against the experimental records and the cited works. We take responsibility for the final content of this work, including text, statements and artifacts produced with the aid of generative AI.

### REPRODUCIBILITY STATEMENT

Section 4.1 states the protocol, the configuration, the seeds and the definition of significance used throughout. Appendix A gives the data-loading rules, DTE's settings and the checks of our baseline numbers against the published ones; Appendix B gives per-group results; Appendix C the filtering policies, the random-removal control and the filter's cost; Appendix D the planted constructions; Appendix F the sensitivity to $K$, $s$, $\lambda$, $\omega$ and the filter's settings; Appendix G the correlation pools; Appendix H the derivations and their assumptions; and Appendix L, with Table 28, the hardware, seeds and groups of every experiment and the software versions. The code of the pipeline and of every experiment reported here and the per-group results behind the result tables are provided in an anonymized repository at `https://anonymous.4open.science/r/code32e0ebd2`.

## REFERENCES

Jeffrey Alido, Tongyu Li, Yu Sun, and Lei Tian. Whitened score diffusion: A structured prior for imaging inverse problems. In *Advances in Neural Information Processing Systems (NeurIPS)*, volume 38, pp. 72619–72646, 2025. doi: 10.52202/085713-2183.

Kris Boudt, Peter J. Rousseeuw, Steven Vanduffel, and Tim Verdonck. The minimum regularized covariance determinant estimator. *Statistics and Computing*, 30(1):113–128, 2020. doi: 10.1007/s11222-019-09869-x.

Markus M. Breunig, Hans-Peter Kriegel, Raymond T. Ng, and Jörg Sander. LOF: Identifying density-based local outliers. In *ACM SIGMOD International Conference on Management of Data*, pp. 93–104, 2000. doi: 10.1145/342009.335388.

Rich Caruana. Multitask learning. *Machine Learning*, 28(1):41–75, 1997. doi: 10.1023/A:1007379606734.

Seoyoung Cho, Jaesung Hwang, Kwan-Young Bak, and Dongha Kim. ALTBI: Constructing improved outlier detection models via optimization of inlier-memorization effect. In *AAAI Conference on Artificial Intelligence*, volume 39, pp. 11544–11552, 2025. doi: 10.1609/aaai.v39i11.33256. arXiv:2408.09791v2.

Antoine Cordier, Benjamin Missaoui, and Pierre Gutierrez. Data refinement for fully unsupervised visual inspection using pre-trained networks. *arXiv preprint arXiv:2202.12759*, 2022.

Adam Goodge, Bryan Hooi, See-Kiong Ng, and Wee Siong Ng. LUNAR: Unifying local outlier detection methods via graph neural networks. In *AAAI Conference on Artificial Intelligence*, volume 36, pp. 6737–6745, 2022. doi: 10.1609/aaai.v36i6.20629.

Xiaoyi Gu, Leman Akoglu, and Alessandro Rinaldo. Statistical analysis of nearest neighbor methods for anomaly detection. In *Advances in Neural Information Processing Systems (NeurIPS)*, volume 32, pp. 10923–10933, 2019. arXiv:1907.03813.

Songqiao Han, Xiyang Hu, Hailiang Huang, Minqi Jiang, and Yue Zhao. ADBench: Anomaly detection benchmark. In *Advances in Neural Information Processing Systems (NeurIPS), Datasets and Benchmarks Track*, volume 35, pp. 32142–32159, 2022. doi: 10.52202/068431-2329.

Jonathan Ho, Ajay Jain, and Pieter Abbeel. Denoising diffusion probabilistic models. In *Advances in Neural Information Processing Systems (NeurIPS)*, volume 33, pp. 6840–6851, 2020.

Thomas Jiralerspong, Berton Earnshaw, Jason Hartford, Yoshua Bengio, and Luca Scimeca. Shaping inductive bias in diffusion models through frequency-based noise control. *arXiv preprint arXiv:2502.10236*, 2025. Workshop paper at the DeLTa and FPI workshops, ICLR 2025.

Khaled Kahouli, Stefaan Simon Pierre Hessmann, Klaus-Robert Müller, Shinichi Nakajima, Stefan Gugler, and Niklas Wolf Andreas Gebauer. Molecular relaxation by reverse diffusion with time step prediction. *Machine Learning: Science and Technology*, 5(3):035038, 2024. doi: 10.1088/2632-2153/ad652c.

Dongha Kim, Jaesung Hwang, Jongjin Lee, Kunwoong Kim, and Yongdai Kim. ODIM: Outlier detection via likelihood of under-fitted generative models. In *International Conference on Machine Learning (ICML)*, volume 235 of *PMLR*, pp. 23941–23971, 2024. arXiv:2301.04257v2.

Zheng Li, Yue Zhao, Nicola Botta, Cezar Ionescu, and Xiyang Hu. COPOD: Copula-based outlier detection. In *IEEE International Conference on Data Mining (ICDM)*, pp. 1118–1123, 2020. doi: 10.1109/ICDM50108.2020.00135.

Zheng Li, Yue Zhao, Xiyang Hu, Nicola Botta, Cezar Ionescu, and George H. Chen. ECOD: Unsupervised outlier detection using empirical cumulative distribution functions. *IEEE Transactions on Knowledge and Data Engineering*, 35(12):12181–12193, 2023. doi: 10.1109/TKDE.2022.3159580.

Boyang Liu, Ding Wang, Kaixiang Lin, Pang-Ning Tan, and Jiayu Zhou. RCA: A deep collaborative autoencoder approach for anomaly detection. In *International Joint Conference on Artificial Intelligence (IJCAI)*, pp. 1505–1511, 2021. doi: 10.24963/ijcai.2021/208.

Fei Tony Liu, Kai Ming Ting, and Zhi-Hua Zhou. Isolation forest. In *Eighth IEEE International Conference on Data Mining (ICDM)*, pp. 413–422, 2008. doi: 10.1109/ICDM.2008.17.

Na Liu, Fan Zhang, and Fuqing Duan. Facial age estimation using a multi-task network combining classification and regression. *IEEE Access*, 8:92441–92451, 2020. doi: 10.1109/ACCESS.2020.2994322.

Victor Livernoche, Vineet Jain, Yashar Hezaveh, and Siamak Ravanbakhsh. On diffusion modeling for anomaly detection. In *International Conference on Learning Representations (ICLR)*, 2024. arXiv:2305.18593.

Victor Livernoche, Jie Zan, and Reihaneh Rabbany. Kurtosis-guided denoising score matching for tabular anomaly detection. *arXiv preprint arXiv:2605.06955*, 2026.

P. C. Mahalanobis. On the generalised distance in statistics. *Proceedings of the National Institute of Sciences of India*, 2(1):49–55, 1936. Reprinted in Sankhyā A 80(S1):1–7, 2018, doi: 10.1007/s13171-019-00164-5.

Fabian Pedregosa, Gaël Varoquaux, Alexandre Gramfort, Vincent Michel, Bertrand Thirion, Olivier Grisel, Mathieu Blondel, Peter Prettenhofer, Ron Weiss, Vincent Dubourg, Jake Vanderplas, Alexandre Passos, David Cournapeau, Matthieu Brucher, Matthieu Perrot, and Édouard Duchesnay. Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research*, 12(85):2825–2830, 2011.

Chen Qiu, Aodong Li, Marius Kloft, Maja Rudolph, and Stephan Mandt. Latent outlier exposure for anomaly detection with contaminated data. In *International Conference on Machine Learning (ICML)*, volume 162 of *PMLR*, pp. 18153–18167, 2022.

Sridhar Ramaswamy, Rajeev Rastogi, and Kyuseok Shim. Efficient algorithms for mining outliers from large data sets. In *ACM SIGMOD International Conference on Management of Data*, pp. 427–438, 2000. doi: 10.1145/342009.335437.

Peter J. Rousseeuw and Katrien Van Driessen. A fast algorithm for the minimum covariance determinant estimator. *Technometrics*, 41(3):212–223, 1999. doi: 10.1080/00401706.1999.10485670.

Peter J. Rousseeuw and Bert C. van Zomeren. Unmasking multivariate outliers and leverage points. *Journal of the American Statistical Association*, 85(411):633–639, 1990. doi: 10.1080/01621459.1990.10474920.

Lukas Ruff, Robert Vandermeulen, Nico Goernitz, Lucas Deecke, Shoaib Ahmed Siddiqui, Alexander Binder, Emmanuel Müller, and Marius Kloft. Deep one-class classification. In *International Conference on Machine Learning (ICML)*, volume 80 of *PMLR*, pp. 4393–4402, 2018.

Mayu Sakurada and Takehisa Yairi. Anomaly detection using autoencoders with nonlinear dimensionality reduction. In *MLSDA Workshop on Machine Learning for Sensory Data Analysis*, pp. 4–11, 2014. doi: 10.1145/2689746.2689747.

Timur Sattarov, Marco Schreyer, and Damian Borth. Diffusion-scheduled denoising autoencoders for anomaly detection in tabular data. In *ACM SIGKDD Conference on Knowledge Discovery and Data Mining (KDD)*, pp. 2478–2489, 2025. doi: 10.1145/3711896.3736910.

Bernhard Schölkopf, John C. Platt, John Shawe-Taylor, Alex J. Smola, and Robert C. Williamson. Estimating the support of a high-dimensional distribution. *Neural Computation*, 13(7):1443–1471, 2001. doi: 10.1162/089976601750264965.

Luca Scimeca, Thomas Jiralerspong, Berton Earnshaw, Jason Hartford, and Yoshua Bengio. Learning what matters: Steering diffusion via spectrally anisotropic forward noise. *arXiv preprint arXiv:2510.09660*, 2025.

Mei-Ling Shyu, Shu-Ching Chen, Kanoksri Sarinnapakorn, and LiWu Chang. A novel anomaly detection scheme based on principal component classifier. In *IEEE Foundations and New Directions of Data Mining Workshop, in conjunction with ICDM*, pp. 172–179, 2003.

Jyotirmai Singh, Samar Khanna, and James Burgess. Squeezed diffusion models. *arXiv preprint arXiv:2508.14871*, 2025. NeurIPS 2025 Workshop on Machine Learning and the Physical Sciences.

Elouan Vincent, Alexandre Dréan, Julien Perez, Marc Plantevit, and Céline Robardet. Diffusion for explainable unsupervised anomaly detection. In *IEEE International Conference on Data Science and Advanced Analytics (DSAA)*, pp. 1–10, 2025. doi: 10.1109/DSAA65442.2025.11248026.

Vikram Voleti, Christopher Pal, and Adam Oberman. Score-based denoising diffusion with non-isotropic Gaussian noise models. In *NeurIPS Workshop on Score-Based Methods*, 2022. arXiv:2210.12254.

Jack Yi Wei and Narges Armanfard. ICLAD: In-context learning for unified tabular anomaly detection across supervision regimes. *arXiv preprint arXiv:2603.19497*, 2026.

Frank Wilcoxon. Individual comparisons by ranking methods. *Biometrics Bulletin*, 1(6):80–83, 1945. doi: 10.2307/3001968.

Hongzuo Xu, Yijie Wang, Juhui Wei, Songlei Jian, Yizhou Li, and Ning Liu. Fascinating supervisory signals and where to find them: Deep anomaly detection with scale learning. In *International Conference on Machine Learning (ICML)*, volume 202 of *PMLR*, pp. 38655–38673, 2023.

In-Kwon Yeo and Richard A. Johnson. A new family of power transformations to improve normality or symmetry. *Biometrika*, 87(4):954–959, 2000. doi: 10.1093/biomet/87.4.954.

Jinsung Yoon, Kihyuk Sohn, Chun-Liang Li, Sercan Ö. Arik, Chen-Yu Lee, and Tomas Pfister. Self-supervise, refine, repeat: Improving unsupervised anomaly detection. *Transactions on Machine Learning Research (TMLR)*, 2022. August 2022; arXiv:2106.06115.

Guy Zamberg, Moshe Salhov, Ofir Lindenbaum, and Amir Averbuch. TabADM: Unsupervised tabular anomaly detection with diffusion models. *arXiv preprint arXiv:2307.12336*, 2023.

Yue Zhao, Zain Nasrullah, and Zheng Li. PyOD: A Python toolbox for scalable outlier detection. *Journal of Machine Learning Research*, 20(96):1–7, 2019.

Bo Zong, Qi Song, Martin Renqiang Min, Wei Cheng, Cristian Lumezanu, Daeki Cho, and Haifeng Chen. Deep autoencoding Gaussian mixture model for unsupervised anomaly detection. In *International Conference on Learning Representations (ICLR)*, 2018.

# A BENCHMARK, PROTOCOL AND CONFIGURATION DETAILS

**Groups and member datasets.** The 57 groups of ADBench are 47 tabular datasets, each its own group, and 10 groups of fixed image and text embeddings. Seven of the ten collect the datasets ADBench derives from one multi-class source: CIFAR10, FashionMNIST and SVHN with 10 members each, MNIST-C with 16 and MVTec-AD with 15 for images, and 20news with 6 and agnews with 4 for text. The remaining three text groups, amazon, imdb and yelp, are single datasets. The image and text groups use ADBench's released embeddings (ResNet18 features for the image groups, BERT features for the text groups). Every AUCROC and AUCPR we report is computed per member dataset and averaged within its group, so the 121 member datasets enter the statistics through their 57 groups. AUCPR is the area under the precision–recall curve, computed as average precision.

**Loader rules.** Two rules of the data loader apply to every detector of Table 1 and are redrawn per seed. A dataset with fewer than 1,000 rows is resampled to 1,000 rows with replacement, ADBench's default, and a dataset with more than 50,000 rows is subsampled to 50,000 rows without replacement, DTE's cap (ADBench's own default is 10,000). Every detector of Table 1 is run with the 50,000 cap; the planted experiment of Section 4.4 uses ADBench's 10,000.

**Protocol.** As stated in Section 4.1, every detector is trained on the dataset as given, anomalies included and unlabeled, or on the rows the filter retains, and scores every row of the dataset. This differs from DTE's evaluation on one point: DTE trains on a bootstrap draw of the dataset and scores the dataset; we train on the dataset itself, without the draw. The published evaluations of the other baselines differ from ours in more respects, summarized in Table 10; their published numbers are shown beside ours in Table 11, and where a matched population exists the comparison is made on it. A feature is treated as constant on the retained rows when its standard deviation on the standardized retained rows is below $10^{-8}$; on the tabular groups this flags exactly the features that take one value on the retained rows, and the same rule is applied for Mahalanobis kNN.

Table 10: Evaluation protocols of the published baseline results. "Train = score" means the detector is fitted on the rows it ranks.

| Source | Training rows and scored rows | Datasets and row cap |
|---|---|---|
| This paper | trains on the dataset (on the retained rows with the filter) and scores every row; no bootstrap draw | 57 groups; 50,000 |
| DTE (Livernoche et al., 2024) | bootstrap draw of the dataset; scores the dataset | 57; 50,000 |
| K-DSM (Livernoche et al., 2026) | DTE's bootstrap draw (contaminated setting) | 57; 50,000 |
| DDAE (Sattarov et al., 2025) | whole dataset for training and evaluation, with bootstrapping | 57; none stated |
| ODIM (Kim et al., 2024) | train = score | 57 (46 tabular, 6 image, 5 text); none stated |
| ALTBI-I (Cho et al., 2025) | train = score ("training AUC") | 57 (46 tabular, 6 image, 5 text); its code supports the 46 tabular; none stated |
| SLAD (Xu et al., 2023) | half of the normal rows for training; the rest scored | clean-normal protocol |
| LUNAR (Goodge et al., 2022) | normal rows remaining after the test allocation, split 85 : 15 into training and validation | other collections |
| ECOD (Li et al., 2023) | 60 : 40 train/test split | 30 datasets of the ODDS and DAMI collections |
| ADBench (Han et al., 2022) | 70 : 30 stratified split | 57; 10,000 |

**Published values and matched-population checks.** The published uncertainty conventions also differ from ours: the baselines' "±" values are averages of per-dataset standard deviations over seeds, whereas ours is the standard deviation of the five seed-level means; the two are not mixed in any table. For ALTBI-I, on the 46 tabular datasets its released code supports (the 47 tabular datasets without `24_mnist`), its published per-dataset AUCROC averages 76.9, our unfiltered run 76.8 and our filtered run 76.9, and a run of its released pipeline end to end under its own protocol, over three

Table 11: Published results on ADBench's 57 groups beside ours without and with the filter (AUCROC / AUCPR, %). The published protocols are those of Table 10; "±" conventions differ between sources and are not shown. The DTE-paper values are means of the 57 per-dataset entries in its appendix tables and ALTBI-I's 46-tabular value from its per-dataset entries, computed by us.

| Detector | Published (source) | Ours, no filter | Ours, + filter | Matched-population check |
|---|---|---|---|---|
| ALTBI-I | 76.2 / 34.8 (Cho et al., 2025) | 75.4 / 34.2 | 75.5 / 34.2 | 46 tabular: 76.9 published, 76.8 ours (76.9 with the filter) |
| ODIM | 74.8 / 36.8 (as evaluated by Cho et al., 2025) | 74.2 / 33.3 | 74.6 / 34.4 | 46 tabular: 75.7 / 36.6 published (Kim et al., 2024), 75.6 / 36.6 ours; five datasets within 0.16 |
| DDAE | 74.1 / — (Sattarov et al., 2025) | 74.2 / 33.2 | 75.2 / 35.0 | five datasets within 0.35 |
| SLAD | 66.6 / 24.3 (as evaluated by Livernoche et al., 2024) | 66.7 / 24.0 | 75.0 / 36.5 | — |
| K-DSM | 70.1 / 34.3 (Livernoche et al., 2026) | 70.0 / 34.1 | 75.8 / 35.9 | same 57 groups: 70.1 published, 70.0 ours |
| DTE-C ($K = 7$, no head) | 72.8 / 28.8 (Livernoche et al., 2024) | 73.4 / 29.5 | — | — |
| DTE-IG (isotropic) | 65.9 / 21.6 (Livernoche et al., 2024) | 67.5 / 21.3 (ADTE-IG, $s = 0.75$) | 75.8 / 36.9 | — |
| MCD detector | 74.8 / 33.7 (as evaluated by Livernoche et al., 2024) | 75.2 / 34.2 | 74.4 / 33.7 | — |
| IForest | 73.7 / 33.6 (as evaluated by Livernoche et al., 2024) | 74.0 / 33.4 | 74.1 / 29.0 | — |
| kNN | 72.8 / 30.8 ($k = 32$; as evaluated by Livernoche et al., 2024) | 67.2 / 24.5 ($k = 5$) | 76.5 / 36.6 | — |
| ECOD | 68.9 / 29.6 (as evaluated by Livernoche et al., 2024) | 72.0 / 31.3 | 72.9 / 32.7 | — |
| COPOD (Li et al., 2020) | 69.2 / 28.8 (as evaluated by Livernoche et al., 2024) | 72.3 / 30.4 | 73.1 / 30.7 | — |
| LOF | 61.1 / 18.1 (as evaluated by Livernoche et al., 2024) | 61.4 / 18.5 | 75.1 / 33.5 | — |
| DTE-NP | 72.2 / 29.5 (as evaluated by Livernoche et al., 2024; DTE's protocol draws a bootstrap training sample, so a self-match is not guaranteed for every scored row, Appendix J) | 65.7 / 22.9 | 76.4 / 35.7 | — |

initializations, gives 76.7; Table 1's 75.4 differs from the published 76.2 through the population, the 10 embedding groups, not the integration. For ODIM, the value in Table 11 is the evaluation reported by Cho et al. (2025), whose image groups use ADBench's Vision Transformer (ViT) embeddings where ours are its ResNet18 ones; Kim et al. (2024) report 75.7 AUCROC and 36.6 AUCPR on 46 tabular datasets, our unfiltered run on the same 46 gives 75.6 and 36.6, and our run of the released ODIM under its own protocol on five datasets agrees with the published per-dataset values to within 0.16 AUCROC points. For DDAE the corresponding five-dataset check, at three seeds, agrees to within 0.35 points; its published PR-AUC is trapezoidal where ours is average precision, so only AUCROC is juxtaposed. Our DTE-C row is the released configuration ($K = 7$, no head, no filter) run under the unsupervised protocol, and the nearest configuration we run for DTE-IG is ADTE-IG without the filter, whose corruption is not isotropic ($s = 0.75$). For LUNAR, $T^2$, $T^2$ ($\Sigma{=}I$) and Mahalanobis kNN we found no ADBench value published under a comparable protocol, so Table 11 has no row for them.

**DTE's settings.** DTE's published architecture and optimizer (hidden widths 256, 512 and 256, ReLU, dropout 0.5, Adam with learning rate $10^{-4}$, 400 epochs, 64 rows per optimization step) are taken from its paper. The released diffusion schedule is kept unchanged: $T = 400$ steps with $\beta$ linear from $10^{-4}$ to $10^{-2}$. DTE's paper lists a maximum timestep of 300 and a linear schedule from 0 to

0.01 (its Table 2 and Appendix D.2); we keep the released implementation's values. Its released code fixes five further choices, which we keep: dropout follows the second and third hidden layers only; there are no normalization layers; Adam runs with weight decay $5 \cdot 10^{-4}$; the learning rate is constant; and training runs its full budget with no early stopping. The regression head we add has one hidden layer of width 128.

**DTE-NP.** DTE's non-parametric estimator runs at the release's defaults: five neighbors, 1,000 timesteps with the linear $\beta$ schedule from $10^{-4}$ to $10^{-2}$, Euclidean distance, no exclusion of a row's own copy, and the release's mapping from the mean distance to the five neighbors to a timestep (the timestep that maximizes the inverse Gamma log-density scaled by that distance). The release's default of five neighbors is used rather than the 32 stated in DTE's text; the estimator is taken unchanged from the released repository. After the filter, the group-level Spearman correlation between DTE-NP's score and the mean distance to the five nearest retained rows is at least 0.997 on every group (median 0.9999); without the filter the minimum is 0.9967 (median 0.9998). The filter's effect on the 57 groups is $+10.70$ AUCROC points $[+6.07, +16.01]$, 39 wins and 18 losses ($+12.75$ $[+7.33, +19.03]$ on the 47; $+1.08$ $[-0.90, +4.30]$ on the 10). Without the filter DTE-NP is 1.5 points $[0.4, 3.0]$ below Euclidean kNN on the 57 groups (65.7 against 67.2); DTE-NP averages the five neighbor distances where kNN takes the largest, and it quantizes the result to a timestep. With the filter the two differ by 0.09 points, so DTE-NP's larger gain in Table 1 ($+10.7$ against $+9.3$) is that unfiltered gap (paired difference of the two gains $+1.44$ $[+0.45, +2.90]$, 38 wins and 19 losses).

**Neighbor detectors at $k = 32$.** Table 1 runs both kNN detectors at $k = 5$, DTE-NP's released default, and LOF at $k = 20$. At $k = 32$, the value of DTE's own kNN baseline, the same protocol gives the means of Table 12. Without the filter, Euclidean kNN reaches 73.0 mean AUCROC on the 57 groups, close to the 72.8 mean of DTE's published $k = 32$ kNN results; with the filter it reaches 76.6, so the filter's gain for kNN is $+3.6$ $[+0.6, +7.9]$ points at $k = 32$ against $+9.3$ at $k = 5$. ADTE-C's margin over the filtered $k = 32$ detectors is $+0.4$ $[-0.9, +2.0]$ over Euclidean kNN, $+0.8$ $[-0.6, +2.3]$ over Mahalanobis kNN and $+1.9$ $[+0.2, +3.8]$ over LOF on the 57 groups.

Table 12: Neighbor detectors at $k = 32$ under the unsupervised protocol (57 groups, %; mean ± SD over seeds). $\Delta$ is the filter's effect, filtered minus unfiltered, from unrounded means; the last two columns give the filtered mean AUCROC on the 47 tabular and the 10 embedding groups.

| | no filter | | + MCD filter | | $\Delta$ (filter) | | + MCD filter, AUCROC | |
|---|---|---|---|---|---|---|---|---|
| Method | AUCROC | AUCPR | AUCROC | AUCPR | ROC | PR | 47 tabular | 10 embedding |
| LOF | 62.6±0.1 | 19.8±0.2 | 75.1±0.2 | 34.3±0.5 | +12.5 | +14.5 | 76.8 | 67.4 |
| Mahalanobis kNN | 72.6±0.2 | 29.3±0.3 | 76.2±0.3 | 34.0±0.3 | +3.6 | +4.7 | 77.9 | 68.4 |
| Euclidean kNN | 73.0±0.2 | 30.8±0.2 | 76.6±0.1 | 36.6±0.2 | +3.6 | +5.7 | 78.8 | 66.4 |

**K-DSM.** K-DSM (Livernoche et al., 2026) runs at its released settings: six residual blocks of width 512, per-feature noise scales set from each feature's kurtosis, and a per-minibatch teacher filter of its own, which drops the rows that the teacher, an exponential moving average of the network's weights, scores above the minibatch's 80th percentile, and which is active in both the unfiltered and the filtered column of Table 1; the MCD filter is added in front of it.

**ODIM's 20news group.** The released ODIM driver runs a two-component Gaussian-mixture check every ten optimization steps. On the member `20news_2`, which has $2{,}497 = 39 \cdot 64 + 1$ rows, the last optimization step of every epoch holds a single row, on which the check raises, and the unfiltered run records no result at any seed; with the filter the row count changes and the run completes. Table 1 therefore averages the same five members of 20news in both ODIM columns. With all six members in the filtered column ODIM's 57-group mean would be $74.585 \pm 0.298$ against $74.621 \pm 0.296$ with the same five; the 47 tabular groups are unaffected.

## B Full Result Tables

This section gives the per-modality means behind Table 1 (Tables 13 and 14), the AUCPR cells of the corruption-by-head factorial of Table 3 (Table 15), the two paired contrasts referred to in Section 4.6, and per-group results for the five detectors of Table 9 and ADTE-IG (Tables 16 and 17).

Table 13: Every detector of Table 1 on the 47 tabular groups, without and with the MCD filter (%; mean ± SD over seeds; Δ = filtered minus unfiltered, from unrounded means). Rows in the order of Table 1. ‡ODIM's 20news group is not part of this population.

| | no filter | | + MCD filter | | Δ (filter) | |
|---|---|---|---|---|---|---|
| Method | AUCROC | AUCPR | AUCROC | AUCPR | ROC | PR |
| ECOD | 74.3±0.1 | 34.8±0.2 | 75.2±0.1 | 36.5±0.3 | +0.9 | +1.6 |
| $T^2$ ($\Sigma{=}I$) | 73.9±0.1 | 36.7±0.3 | 75.0±0.1 | 37.2±0.3 | +1.0 | +0.5 |
| COPOD | 74.5±0.1 | 33.8±0.3 | 75.5±0.1 | 34.2±0.3 | +1.0 | +0.4 |
| IForest | 76.4±0.1 | 37.3±0.5 | 76.5±0.2 | 32.0±0.4 | +0.2 | −5.2 |
| MCD detector | 77.7±0.4 | 38.8±0.9 | 76.5±0.3 | 37.9±0.8 | −1.2 | −0.9 |
| LOF | 60.4±0.4 | 19.3±0.3 | 76.7±0.3 | 37.0±0.7 | +16.3 | +17.7 |
| Mahalanobis kNN | 68.0±0.6 | 27.4±0.6 | 77.7±0.6 | 37.5±0.5 | +9.7 | +10.1 |
| LUNAR | 64.0±0.5 | 21.9±0.3 | 78.4±0.5 | 39.7±0.4 | +14.4 | +17.8 |
| Euclidean kNN | 67.4±0.5 | 26.4±0.6 | 78.5±0.5 | 40.9±0.5 | +11.1 | +14.5 |
| $T^2$ | 76.1±0.1 | 33.4±0.2 | 78.9±0.2 | 40.1±0.7 | +2.8 | +6.7 |
| ODIM‡ | 75.8±0.1 | 36.6±0.1 | 76.5±0.4 | 38.3±0.4 | +0.7 | +1.7 |
| SLAD | 67.6±0.4 | 26.0±0.5 | 77.3±0.2 | 40.8±0.8 | +9.7 | +14.8 |
| DDAE | 75.9±0.2 | 36.8±0.3 | 77.2±0.1 | 39.0±0.2 | +1.2 | +2.1 |
| ALTBI-I | 76.9±0.2 | 37.1±0.8 | 77.2±0.2 | 37.8±0.8 | +0.3 | +0.7 |
| K-DSM | 71.0±0.7 | 38.1±1.1 | 77.5±0.3 | 39.5±0.2 | +6.5 | +1.5 |
| ADTE-IG (ours) | 69.3±1.3 | 23.2±0.5 | 78.6±0.3 | 41.6±0.5 | +9.3 | +18.4 |
| isotropic DTE-C | 76.2±0.6 | 34.1±0.6 | 78.7±0.2 | 40.2±0.8 | +2.5 | +6.2 |
| DTE-NP | 65.7±0.4 | 24.7±0.3 | 78.4±0.4 | 39.8±0.3 | +12.7 | +15.1 |
| **ADTE-C (ours)** | 76.5±1.0 | 37.6±0.9 | 79.6±0.2 | 41.3±0.9 | +3.1 | +3.7 |

Table 14: Every detector of Table 1 on the 10 image and text groups, without and with the MCD filter (%; mean ± SD over seeds; Δ = filtered minus unfiltered, from unrounded means). ‡ODIM's 20news group averages the same five of its six members in both columns.

| | no filter | | + MCD filter | | Δ (filter) | |
|---|---|---|---|---|---|---|
| Method | AUCROC | AUCPR | AUCROC | AUCPR | ROC | PR |
| ECOD | 61.6±0.1 | 14.6±0.1 | 61.9±0.1 | 14.8±0.1 | +0.3 | +0.2 |
| $T^2$ ($\Sigma{=}I$) | 63.1±0.1 | 15.1±0.0 | 63.5±0.1 | 15.2±0.1 | +0.4 | +0.1 |
| COPOD | 61.9±0.1 | 14.6±0.1 | 62.0±0.1 | 14.7±0.1 | +0.1 | +0.1 |
| IForest | 63.1±0.3 | 15.3±0.1 | 62.8±0.2 | 14.5±0.1 | −0.3 | −0.8 |
| MCD detector | 63.8±0.2 | 12.6±0.1 | 64.5±0.1 | 14.1±0.2 | +0.7 | +1.4 |
| LOF | 65.9±0.1 | 14.7±0.1 | 67.5±0.1 | 16.9±0.1 | +1.6 | +2.2 |
| Mahalanobis kNN | 67.1±0.1 | 15.1±0.1 | 68.6±0.1 | 18.3±0.2 | +1.5 | +3.2 |
| LUNAR | 63.2±1.3 | 13.6±0.3 | 65.9±0.2 | 15.7±0.1 | +2.7 | +2.1 |
| Euclidean kNN | 66.1±0.3 | 15.4±0.1 | 66.9±0.1 | 16.5±0.1 | +0.8 | +1.0 |
| $T^2$ | 67.4±0.1 | 15.2±0.1 | 66.7±0.1 | 15.6±0.1 | −0.7 | +0.5 |
| ODIM‡ | 66.9±0.1 | 17.6±0.1 | 65.8±0.1 | 16.5±0.1 | −1.2 | −1.1 |
| SLAD | 62.6±0.1 | 14.9±0.2 | 64.1±0.0 | 16.0±0.1 | +1.6 | +1.1 |
| DDAE | 66.1±0.1 | 15.9±0.0 | 65.9±0.1 | 16.1±0.1 | −0.2 | +0.2 |
| ALTBI-I | 68.7±0.1 | 20.8±0.2 | 67.9±0.1 | 17.3±0.2 | −0.8 | −3.5 |
| K-DSM | 65.1±0.4 | 15.5±0.6 | 68.0±0.1 | 18.9±0.1 | +2.9 | +3.4 |
| ADTE-IG (ours) | 58.9±1.2 | 11.9±0.3 | 62.8±0.5 | 15.0±0.1 | +3.9 | +3.1 |
| isotropic DTE-C | 64.9±0.5 | 15.3±0.2 | 64.9±0.3 | 15.7±0.1 | −0.0 | +0.4 |
| DTE-NP | 65.7±0.1 | 14.7±0.1 | 66.7±0.1 | 16.3±0.1 | +1.1 | +1.6 |
| **ADTE-C (ours)** | 63.5±0.3 | 15.2±0.2 | 64.7±0.4 | 15.5±0.0 | +1.2 | +0.3 |

**Two paired contrasts of Section 4.6.** On the 47 tabular groups, after the filter, isotropic DTE-C minus Euclidean kNN is $+0.21$ AUCROC points $[-0.93, +1.39]$ (18 wins, 29 losses) and minus $T^2$ is $-0.20$ $[-0.98, +0.60]$ (25 wins, 22 losses); neither interval excludes zero.

## C The Filter: Alternatives, Controls and Cost

**Eight filtering policies and a no-filter control.** Table 18 gives ADTE-C's mean AUCROC on the 57 groups and on the 47 tabular groups when the MCD filter is replaced by each of seven other ways of choosing the rows to remove, and with no filter. Each alternative removes exactly 25% of the rows

Table 15: Corruption × head at $K = 8$ with the filter fixed, for DTE-C with and without the regression head, AUCPR (57 groups, %; mean ± SD over seeds). Effects are differences of unrounded cell means; the interaction $p$ is from a two-sided Wilcoxon test over per-group interaction contrasts, the alignment effect with the head minus the alignment effect without it.

| | isotropic | aligned ($s$=0) |
|---|---|---|
| no head | 36.00 ± 0.80 | 36.59 ± 0.78 |
| with head | 35.94 ± 0.69 | 36.74 ± 0.74 |
| alignment | +0.59 (no head) / +0.81 (head) | |
| head | −0.07 (iso.) / +0.15 (aligned) | |
| interaction | +0.22, $p = .33$ | |

by its own score, close to the MCD filter's median removal of 27%; the MCD trim control ranks rows by the robust Mahalanobis distance under the filter's reweighted estimate, the mean and rescaled covariance of the rows the filter retains, but removes a fixed 25%, separating the robust covariance from the retention budget. The two learned detectors used as filters run at reduced settings: ODIM as a filter uses a three-member ensemble rather than its published ten, and ALTBI-I as a filter averages the scores of its three trials into one ranking. The DTE-score filter removes the 25% of rows that an isotropic seven-bin DTE-C without the head scores highest, in one round. Every policy has a higher mean AUCROC than the no-filter control and the MCD filter the highest, on both populations. The MCD filter is ahead of the DTE-score filter by $+1.74$ AUCROC points $[+0.14, +3.63]$ (ahead on 31 of the 57 groups; two-sided Wilcoxon $p = .19$); the other six intervals include zero, and the Wilcoxon $p$-values of the six alternatives other than the DTE-score filter, each against the MCD filter, all exceed .05. The eight filters lie within $1.74$ points of one another, while the least of them is $1.0$ points above no filter.

**Training on labeled normal rows only.** When ADTE-C and isotropic DTE-C are trained on half of each dataset's labeled normal rows and score the held-out normal rows together with all anomalies, both do better without the filter. On the 47 tabular groups the no-filter model is ahead of the filtered one by $+2.32$ AUCROC points $[+0.52, +4.43]$ for ADTE-C (32 wins and 15 losses) and by $+2.95$ $[+0.87, +5.64]$ for isotropic DTE-C (29 and 18); on the 57 groups the two differences are $+1.83$ $[+0.29, +3.70]$ and $+2.83$ $[+1.10, +5.16]$. The difference between the two contrasts ($-0.63$ $[-2.02, +0.56]$ on the 47) is not significant.

**Removing the same number of rows at random.** Table 19 compares the MCD filter with removing, for each dataset and seed, the same number of rows chosen uniformly at random, for Euclidean kNN, LOF, isotropic DTE-C and ADTE-C. On the 47 tabular groups the filter's advantage over random removal is $+9.09$ $[+4.83, +14.45]$ for Euclidean kNN (31 wins, 16 losses), $+15.24$ $[+10.37, +20.34]$ for LOF (36 and 11), $+2.53$ $[+0.72, +4.72]$ for isotropic DTE-C (31 and 16) and $+3.23$ $[+0.99, +6.02]$ for ADTE-C (26 and 21).

**Other preprocessing transforms.** Without the filter, changing the pipeline's preprocessing to another standard transform lowers ADTE-C's mean AUCROC on the 57 groups in every case (AUCROC only): winsorization at the 1st and 99th percentiles followed by the z-scoring $73.49$ ($-0.70$ against the z-scored control's $74.19$), the Yeo–Johnson power transform (Yeo & Johnson, 2000) with its own standardization in place of the z-scoring $72.66$ ($-1.54$), min–max scaling in its place $70.91$ ($-3.28$; three seeds), scikit-learn's RobustScaler in its place $70.44$ ($-3.76$), a quantile transform to the normal distribution in its place $67.62$ ($-6.57$) and per-row unit normalization after the z-scoring $63.81$ ($-10.38$); on the 47 tabular groups the same six are $-0.81$, $-1.75$, $-4.19$, $-4.45$, $-7.21$ and $-10.03$ points against a control of $76.48$. The quantile and RobustScaler transforms produce constant scores on some datasets in seven of the 57 groups, which enter these means at an AUCROC of 50; scoring those datasets at 100 instead still leaves both transforms below the control and below winsorization.

**Cost.** Table 20 gives training times per training run, averaged over the 121 member datasets and the seeds (hardware in Table 28), and the total per seed, which is 121 times the mean. The filter's own time is excluded from the training times and measured separately on a CPU: 2.50 seconds per dataset

Table 16: Per-group AUCROC for the five detectors of Table 9 and ADTE-IG (%; group means), without and with the filter; the 47 tabular groups, then the 10 image and text groups.

| | ADTE-C | | isotropic DTE-C | | ADTE-IG | | Euclidean kNN | | Mahalanobis kNN | | $T^2$ | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| Group | no filter | filter | no filter | filter | no filter | filter | no filter | filter | no filter | filter | no filter | filter |
| 1_ALOI | 52.0 | 53.9 | 52.3 | 54.0 | 52.5 | 52.9 | 70.6 | 63.4 | 64.2 | 55.7 | 55.9 | 53.5 |
| 2_annthyroid | 96.0 | 90.0 | 92.5 | 90.7 | 91.1 | 93.6 | 81.1 | 82.3 | 79.6 | 79.1 | 64.1 | 93.3 |
| 3_backdoor | 87.3 | 89.6 | 86.5 | 89.9 | 82.8 | 91.9 | 74.3 | 91.7 | 71.4 | 91.8 | 90.4 | 93.2 |
| 4_breastw | 92.3 | 97.5 | 93.9 | 95.9 | 84.6 | 94.5 | 93.3 | 98.2 | 93.6 | 98.3 | 97.2 | 97.7 |
| 5_campaign | 75.9 | 73.4 | 78.8 | 74.3 | 72.2 | 71.6 | 74.1 | 74.6 | 75.5 | 73.7 | 76.6 | 75.2 |
| 6_cardio | 80.4 | 81.7 | 76.1 | 78.8 | 67.9 | 82.8 | 68.6 | 78.7 | 69.4 | 68.2 | 89.6 | 83.0 |
| 7_Cardiotocography | 53.4 | 50.9 | 49.6 | 48.2 | 54.9 | 58.4 | 49.1 | 50.0 | 48.7 | 52.2 | 60.3 | 53.4 |
| 8_celeba | 66.0 | 82.6 | 81.0 | 84.4 | 76.2 | 82.8 | 65.1 | 81.2 | 64.8 | 81.4 | 79.4 | 84.1 |
| 9_census | 63.6 | 61.1 | 64.9 | 62.5 | 62.4 | 59.9 | 67.2 | 63.8 | 66.9 | 60.4 | 65.9 | 61.6 |
| 10_cover | 69.8 | 84.4 | 75.0 | 82.9 | 69.9 | 82.9 | 79.3 | 84.8 | 79.9 | 83.4 | 93.2 | 85.3 |
| 11_donors | 76.8 | 93.7 | 78.9 | 91.8 | 52.4 | 90.8 | 73.6 | 91.8 | 73.6 | 90.7 | 79.8 | 91.5 |
| 12_fault | 63.8 | 57.4 | 57.6 | 58.0 | 52.6 | 52.1 | 72.0 | 63.9 | 70.5 | 58.9 | 62.7 | 50.4 |
| 13_fraud | 93.8 | 93.6 | 93.6 | 92.7 | 90.6 | 93.3 | 95.1 | 95.0 | 95.0 | 91.4 | 95.2 | 93.6 |
| 14_glass | 87.1 | 81.2 | 86.8 | 80.3 | 75.5 | 73.0 | 53.3 | 72.5 | 52.6 | 75.4 | 76.2 | 78.9 |
| 15_Hepatitis | 67.1 | 72.5 | 72.5 | 70.8 | 60.5 | 68.9 | 46.2 | 65.3 | 46.6 | 64.7 | 68.6 | 77.3 |
| 16_http | 99.5 | 99.6 | 99.5 | 99.6 | 98.7 | 99.9 | 9.2 | 99.9 | 9.2 | 99.9 | 99.6 | 99.9 |
| 17_InternetAds | 65.0 | 76.5 | 66.6 | 73.3 | 60.0 | 75.4 | 65.2 | 69.8 | 77.2 | 70.8 | 76.6 | 76.3 |
| 18_Ionosphere | 92.3 | 92.5 | 93.7 | 95.4 | 80.4 | 93.2 | 67.5 | 92.5 | 68.3 | 93.3 | 92.2 | 95.0 |
| 19_landsat | 65.0 | 62.7 | 54.0 | 57.2 | 44.3 | 56.0 | 57.6 | 56.0 | 48.9 | 57.3 | 46.4 | 59.7 |
| 20_letter | 87.7 | 83.7 | 83.2 | 83.0 | 72.9 | 82.6 | 90.1 | 85.7 | 89.2 | 85.3 | 80.4 | 81.4 |
| 21_Lymphography | 96.2 | 99.7 | 91.9 | 99.6 | 63.6 | 99.4 | 46.1 | 99.7 | 53.8 | 99.5 | 98.5 | 96.0 |
| 22_magic.gamma | 75.4 | 75.0 | 77.7 | 77.0 | 74.8 | 77.2 | 81.1 | 79.3 | 83.5 | 81.7 | 73.7 | 73.0 |
| 23_mammography | 83.2 | 84.9 | 82.7 | 85.2 | 85.1 | 84.9 | 83.8 | 85.3 | 84.0 | 85.0 | 86.0 | 83.5 |
| 24_mnist | 84.6 | 89.9 | 86.3 | 90.3 | 73.4 | 88.2 | 83.7 | 91.2 | 83.5 | 90.8 | 84.6 | 89.1 |
| 25_musk | 100.0 | 100.0 | 99.8 | 100.0 | 83.5 | 100.0 | 61.6 | 100.0 | 97.3 | 97.1 | 99.7 | 95.9 |
| 26_optdigits | 44.2 | 40.2 | 48.4 | 39.3 | 52.1 | 44.7 | 39.5 | 35.8 | 42.1 | 39.2 | 50.5 | 38.6 |
| 27_PageBlocks | 91.7 | 93.5 | 90.7 | 91.6 | 87.5 | 92.3 | 83.4 | 90.6 | 81.0 | 87.5 | 91.6 | 92.9 |
| 28_pendigits | 79.7 | 90.2 | 79.5 | 86.8 | 69.8 | 88.5 | 74.3 | 88.4 | 70.1 | 83.8 | 76.3 | 86.0 |
| 29_Pima | 65.0 | 64.9 | 65.2 | 64.2 | 55.5 | 60.9 | 66.9 | 66.1 | 67.0 | 66.7 | 68.5 | 68.5 |
| 30_satellite | 80.8 | 82.1 | 74.3 | 77.9 | 59.1 | 77.0 | 66.5 | 77.0 | 62.2 | 77.7 | 63.6 | 79.5 |
| 31_satimage-2 | 99.8 | 99.9 | 96.8 | 99.8 | 90.9 | 98.9 | 93.2 | 99.9 | 94.8 | 99.4 | 98.4 | 99.5 |
| 32_shuttle | 97.6 | 99.0 | 97.4 | 99.3 | 83.8 | 99.6 | 63.4 | 99.6 | 63.3 | 99.6 | 98.2 | 99.2 |
| 33_skin | 76.0 | 91.2 | 78.7 | 89.9 | 70.7 | 80.7 | 67.7 | 79.1 | 67.5 | 80.0 | 71.4 | 86.8 |
| 34_smtp | 95.5 | 95.5 | 95.3 | 95.5 | 87.2 | 95.9 | 91.7 | 96.1 | 91.5 | 96.1 | 81.8 | 95.1 |
| 35_SpamBase | 55.9 | 49.6 | 51.8 | 49.8 | 52.0 | 49.3 | 48.9 | 51.2 | 47.6 | 51.5 | 52.9 | 48.7 |
| 36_speech | 51.6 | 49.3 | 53.7 | 49.2 | 51.1 | 45.1 | 49.1 | 48.5 | 51.5 | 50.2 | 50.0 | 49.2 |
| 37_Stamps | 79.5 | 82.5 | 81.3 | 85.1 | 68.6 | 80.3 | 60.6 | 83.8 | 58.5 | 82.4 | 86.0 | 84.7 |
| 38_thyroid | 99.5 | 98.4 | 98.9 | 98.6 | 87.4 | 98.8 | 95.9 | 97.0 | 95.7 | 95.8 | 93.4 | 98.7 |
| 39_vertebral | 39.9 | 40.2 | 35.1 | 38.5 | 46.7 | 47.7 | 49.2 | 42.6 | 50.5 | 44.3 | 40.7 | 38.5 |
| 40_vowels | 89.6 | 94.4 | 94.7 | 96.2 | 79.1 | 95.1 | 97.7 | 96.5 | 98.2 | 95.8 | 91.2 | 92.3 |
| 41_Waveform | 77.2 | 76.4 | 60.7 | 62.0 | 53.9 | 61.5 | 72.3 | 72.3 | 59.6 | 59.5 | 57.2 | 57.2 |
| 42_WBC | 93.6 | 98.7 | 88.0 | 98.6 | 78.2 | 99.1 | 55.6 | 99.6 | 55.7 | 99.5 | 97.3 | 99.1 |
| 43_WDBC | 98.3 | 97.8 | 79.2 | 95.6 | 68.2 | 96.9 | 81.2 | 97.9 | 82.0 | 96.4 | 95.3 | 97.5 |
| 44_Wilt | 58.4 | 84.1 | 78.7 | 79.1 | 85.0 | 84.0 | 61.3 | 68.2 | 70.2 | 79.7 | 66.6 | 85.5 |
| 45_wine | 59.8 | 96.5 | 66.0 | 96.7 | 55.4 | 96.7 | 49.8 | 93.5 | 49.8 | 94.0 | 62.7 | 97.0 |
| 46_WPBC | 48.6 | 50.0 | 51.1 | 50.1 | 49.2 | 48.5 | 52.3 | 48.7 | 50.1 | 46.6 | 48.1 | 51.6 |
| 47_yeast | 40.3 | 40.2 | 40.2 | 40.4 | 42.7 | 45.1 | 40.6 | 40.5 | 40.3 | 40.3 | 40.4 | 40.9 |
| CIFAR10 | 64.6 | 64.4 | 65.2 | 65.0 | 59.8 | 63.1 | 66.2 | 66.0 | 68.2 | 68.1 | 68.4 | 65.3 |
| FashionMNIST | 86.3 | 86.4 | 86.4 | 87.1 | 76.0 | 85.0 | 86.7 | 87.9 | 89.1 | 90.4 | 89.0 | 86.9 |
| MNIST-C | 75.5 | 74.4 | 76.3 | 75.3 | 70.6 | 73.3 | 79.2 | 78.5 | 81.2 | 80.5 | 80.2 | 76.2 |
| MVTec-AD | 73.2 | 75.3 | 74.6 | 75.4 | 60.4 | 73.5 | 64.0 | 75.3 | 58.6 | 74.9 | 63.4 | 73.0 |
| SVHN | 59.2 | 59.4 | 60.2 | 59.5 | 56.6 | 58.8 | 61.2 | 60.8 | 62.8 | 62.9 | 62.9 | 59.9 |
| 20news | 53.5 | 58.0 | 58.7 | 60.1 | 54.1 | 55.2 | 57.7 | 57.5 | 62.6 | 61.2 | 62.4 | 61.4 |
| agnews | 59.0 | 61.4 | 63.4 | 62.9 | 55.3 | 56.8 | 66.4 | 65.3 | 72.2 | 72.1 | 72.0 | 68.2 |
| amazon | 54.3 | 57.3 | 57.1 | 55.8 | 50.8 | 54.6 | 61.0 | 60.9 | 60.9 | 60.8 | 60.5 | 60.1 |
| imdb | 49.0 | 49.6 | 44.6 | 47.1 | 50.4 | 50.7 | 49.6 | 49.8 | 48.3 | 48.1 | 48.1 | 50.0 |
| yelp | 60.3 | 61.0 | 62.7 | 61.0 | 55.2 | 57.5 | 68.6 | 67.1 | 67.3 | 67.3 | 67.0 | 65.8 |

Table 17: Per-group AUCPR for the five detectors of Table 9 and ADTE-IG (%; group means), without and with the filter; the 47 tabular groups, then the 10 image and text groups.

| | ADTE-C | | isotropic DTE-C | | ADTE-IG | | Euclidean kNN | | Mahalanobis kNN | | $T^2$ | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| Group | no filter | filter | no filter | filter | no filter | filter | no filter | filter | no filter | filter | no filter | filter |
| 1_ALOI | 3.2 | 3.4 | 3.2 | 3.4 | 3.6 | 3.3 | 7.6 | 4.3 | 6.1 | 3.6 | 3.8 | 3.4 |
| 2_annthyroid | 65.9 | 42.5 | 61.2 | 46.4 | 37.2 | 50.4 | 23.0 | 28.2 | 21.5 | 21.4 | 15.7 | 54.6 |
| 3_backdoor | 55.2 | 62.2 | 52.4 | 58.2 | 48.7 | 59.6 | 37.5 | 55.2 | 36.1 | 18.3 | 54.3 | 62.3 |
| 4_breastw | 75.5 | 91.5 | 78.5 | 86.5 | 62.5 | 87.9 | 89.2 | 95.9 | 90.7 | 96.0 | 93.6 | 94.2 |
| 5_campaign | 28.3 | 28.0 | 32.1 | 27.6 | 27.9 | 28.2 | 27.4 | 29.0 | 30.3 | 31.1 | 31.1 | 27.7 |
| 6_cardio | 39.7 | 50.6 | 30.4 | 48.0 | 18.8 | 47.3 | 28.6 | 48.5 | 31.6 | 23.2 | 46.4 | 53.6 |
| 7_Cardiotocography | 32.7 | 36.0 | 27.9 | 34.1 | 27.0 | 36.0 | 28.6 | 35.1 | 29.2 | 29.8 | 35.3 | 36.6 |
| 8_celeba | 4.3 | 10.9 | 7.4 | 11.3 | 6.8 | 12.0 | 4.2 | 11.2 | 4.2 | 11.3 | 7.8 | 11.8 |
| 9_census | 8.1 | 7.4 | 8.2 | 7.5 | 8.2 | 7.2 | 9.1 | 7.9 | 9.1 | 7.2 | 8.5 | 7.3 |
| 10_cover | 2.2 | 4.4 | 3.1 | 4.4 | 3.8 | 4.3 | 3.9 | 4.1 | 4.0 | 4.5 | 8.2 | 3.7 |
| 11_donors | 13.9 | 36.5 | 14.7 | 26.8 | 8.1 | 24.9 | 15.9 | 27.5 | 15.8 | 25.2 | 14.1 | 26.5 |
| 12_fault | 46.5 | 37.0 | 41.9 | 37.4 | 38.4 | 34.7 | 53.0 | 41.8 | 50.4 | 37.8 | 44.6 | 33.1 |
| 13_fraud | 67.1 | 17.0 | 67.0 | 19.0 | 17.5 | 20.5 | 8.9 | 16.2 | 8.3 | 54.0 | 14.4 | 58.5 |
| 14_glass | 21.3 | 14.7 | 17.2 | 14.6 | 12.7 | 16.1 | 11.6 | 12.8 | 8.8 | 13.1 | 12.8 | 10.1 |
| 15_Hepatitis | 28.1 | 41.2 | 40.6 | 40.2 | 27.2 | 37.3 | 18.3 | 32.8 | 16.7 | 25.1 | 28.3 | 40.8 |
| 16_http | 41.4 | 55.9 | 41.4 | 57.1 | 23.6 | 86.4 | 2.3 | 86.0 | 2.3 | 86.0 | 47.6 | 86.1 |
| 17_InternetAds | 40.1 | 41.9 | 30.3 | 38.1 | 29.0 | 41.6 | 27.9 | 33.8 | 36.4 | 35.3 | 37.5 | 37.1 |
| 18_Ionosphere | 87.4 | 92.1 | 90.5 | 94.5 | 67.5 | 93.2 | 73.3 | 92.7 | 73.7 | 93.5 | 91.5 | 94.6 |
| 19_landsat | 31.1 | 25.8 | 22.3 | 23.8 | 18.2 | 22.8 | 24.8 | 22.2 | 20.0 | 22.7 | 18.9 | 24.5 |
| 20_letter | 43.8 | 24.5 | 31.3 | 21.8 | 21.8 | 22.7 | 35.2 | 21.3 | 38.5 | 21.0 | 22.7 | 18.0 |
| 21_Lymphography | 48.9 | 93.4 | 40.8 | 89.9 | 14.9 | 83.4 | 8.2 | 93.5 | 9.5 | 87.3 | 68.0 | 64.5 |
| 22_magic.gamma | 68.1 | 68.9 | 69.4 | 70.0 | 61.2 | 70.0 | 74.1 | 71.4 | 76.3 | 73.5 | 66.0 | 66.4 |
| 23_mammography | 18.2 | 16.2 | 17.8 | 16.3 | 10.4 | 15.0 | 15.8 | 15.8 | 16.3 | 15.5 | 17.8 | 10.6 |
| 24_mnist | 38.8 | 44.1 | 41.8 | 44.1 | 26.6 | 42.7 | 38.7 | 47.8 | 39.8 | 47.9 | 38.3 | 42.4 |
| 25_musk | 99.0 | 98.9 | 91.7 | 99.4 | 21.9 | 98.9 | 21.1 | 100.0 | 60.4 | 67.2 | 89.8 | 60.4 |
| 26_optdigits | 2.4 | 2.2 | 2.6 | 2.2 | 3.1 | 2.4 | 2.2 | 2.0 | 2.3 | 2.2 | 2.7 | 2.1 |
| 27_PageBlocks | 59.7 | 62.6 | 53.4 | 58.1 | 48.7 | 60.3 | 47.0 | 52.6 | 42.4 | 37.3 | 54.0 | 62.8 |
| 28_pendigits | 7.1 | 12.4 | 5.7 | 7.5 | 5.6 | 8.7 | 7.5 | 9.4 | 5.9 | 6.1 | 5.6 | 7.6 |
| 29_Pima | 45.0 | 45.6 | 45.4 | 45.2 | 37.6 | 44.1 | 51.8 | 49.5 | 50.9 | 49.9 | 50.6 | 50.2 |
| 30_satellite | 69.0 | 78.2 | 63.7 | 75.1 | 39.9 | 74.1 | 51.9 | 73.0 | 49.1 | 72.1 | 51.4 | 75.1 |
| 31_satimage-2 | 86.4 | 97.0 | 18.2 | 86.0 | 9.3 | 89.7 | 34.8 | 97.7 | 30.2 | 64.4 | 36.7 | 67.4 |
| 32_shuttle | 61.7 | 86.0 | 62.8 | 89.0 | 33.0 | 93.0 | 16.4 | 92.5 | 15.9 | 96.5 | 87.1 | 86.5 |
| 33_skin | 31.7 | 53.5 | 34.4 | 51.0 | 28.4 | 38.1 | 26.8 | 37.5 | 26.5 | 39.1 | 28.7 | 44.8 |
| 34_smtp | 41.6 | 12.6 | 41.8 | 12.7 | 6.5 | 38.9 | 25.3 | 36.4 | 25.6 | 36.4 | 27.7 | 0.6 |
| 35_SpamBase | 42.4 | 37.9 | 39.8 | 38.2 | 41.6 | 39.0 | 39.5 | 40.0 | 38.5 | 40.4 | 39.8 | 39.3 |
| 36_speech | 1.9 | 1.6 | 2.0 | 1.7 | 2.1 | 1.6 | 2.0 | 1.9 | 2.5 | 2.0 | 2.4 | 2.0 |
| 37_Stamps | 24.6 | 29.7 | 25.6 | 31.6 | 19.8 | 29.4 | 17.9 | 33.2 | 16.2 | 30.3 | 30.3 | 29.2 |
| 38_thyroid | 83.8 | 69.2 | 73.9 | 70.9 | 17.2 | 69.4 | 27.3 | 44.6 | 26.4 | 29.5 | 25.8 | 72.6 |
| 39_vertebral | 10.3 | 9.9 | 9.4 | 9.6 | 12.6 | 11.2 | 11.4 | 10.3 | 12.0 | 11.0 | 9.9 | 9.5 |
| 40_vowels | 24.7 | 37.6 | 38.9 | 54.1 | 15.4 | 46.1 | 55.1 | 47.2 | 69.0 | 45.5 | 36.1 | 35.2 |
| 41_Waveform | 12.2 | 10.3 | 4.0 | 4.4 | 3.4 | 4.8 | 10.5 | 9.9 | 4.1 | 4.1 | 3.9 | 3.9 |
| 42_WBC | 30.3 | 74.8 | 20.9 | 73.8 | 22.5 | 90.1 | 25.0 | 95.0 | 27.4 | 94.9 | 57.4 | 89.1 |
| 43_WDBC | 49.3 | 43.1 | 11.0 | 29.3 | 7.5 | 37.1 | 29.8 | 48.8 | 34.8 | 36.5 | 33.5 | 42.4 |
| 44_Wilt | 5.9 | 13.9 | 11.8 | 11.3 | 22.7 | 14.1 | 6.2 | 7.5 | 8.0 | 11.7 | 7.4 | 15.1 |
| 45_wine | 13.1 | 63.9 | 17.2 | 66.6 | 15.6 | 63.3 | 8.6 | 44.0 | 8.5 | 49.3 | 11.2 | 67.8 |
| 46_WPBC | 23.3 | 23.3 | 26.2 | 23.5 | 26.2 | 22.9 | 24.6 | 22.5 | 24.0 | 22.0 | 22.2 | 23.9 |
| 47_yeast | 30.1 | 29.3 | 29.5 | 29.2 | 30.6 | 30.9 | 30.1 | 29.4 | 29.6 | 29.2 | 28.9 | 29.2 |
| CIFAR10 | 8.9 | 9.2 | 9.3 | 9.5 | 8.0 | 9.4 | 10.5 | 10.2 | 11.1 | 10.8 | 11.1 | 9.3 |
| FashionMNIST | 33.0 | 32.6 | 30.4 | 33.1 | 22.2 | 29.8 | 32.8 | 35.2 | 35.2 | 43.5 | 34.9 | 32.1 |
| MNIST-C | 18.9 | 17.3 | 18.4 | 17.9 | 13.8 | 17.2 | 19.7 | 20.1 | 22.3 | 25.2 | 20.5 | 18.4 |
| MVTec-AD | 53.8 | 57.0 | 54.7 | 56.7 | 39.7 | 56.1 | 46.7 | 57.7 | 35.4 | 58.2 | 37.8 | 54.5 |
| SVHN | 7.5 | 7.5 | 7.7 | 7.6 | 6.8 | 7.6 | 8.1 | 7.9 | 8.4 | 8.3 | 8.5 | 7.7 |
| 20news | 6.0 | 6.6 | 7.2 | 7.4 | 6.1 | 6.5 | 7.6 | 6.9 | 8.5 | 7.6 | 8.6 | 7.5 |
| agnews | 6.5 | 7.0 | 7.4 | 7.4 | 6.0 | 6.3 | 9.0 | 7.9 | 10.8 | 10.5 | 10.9 | 8.5 |
| amazon | 5.6 | 5.9 | 5.9 | 5.7 | 5.3 | 5.7 | 6.3 | 6.3 | 6.3 | 6.3 | 6.3 | 6.2 |
| imdb | 4.7 | 4.7 | 4.4 | 4.5 | 4.9 | 5.0 | 4.7 | 4.7 | 4.5 | 4.5 | 4.5 | 4.8 |
| yelp | 6.8 | 6.7 | 7.4 | 7.1 | 6.4 | 6.2 | 9.0 | 7.9 | 8.6 | 8.3 | 8.6 | 7.5 |

Table 18: ADTE-C with the MCD filter, seven other filtering policies and no filter (%; mean AUCROC ± SD over seeds on the 57 groups), ordered by the 57-group mean. Each alternative removes 25% of the rows by its own score; the MCD filter removes a data-dependent fraction (median 27%).

| Rows removed by | AUCROC (57) | AUCROC (47 tabular) |
|---|---|---|
| MCD filter (the paper's) | $\mathbf{76.89} \pm 0.14$ | $\mathbf{79.55}$ |
| MCD distance, fixed 25% trim | $76.67 \pm 0.36$ | 79.24 |
| ALTBI-I score | $76.37 \pm 0.40$ | 78.72 |
| ECOD score | $76.34 \pm 0.39$ | 78.86 |
| IForest score | $75.90 \pm 0.34$ | 78.41 |
| kNN distance ($k = 5$) | $75.86 \pm 0.54$ | 78.22 |
| ODIM score | $75.51 \pm 0.40$ | 77.81 |
| DTE-C score (isotropic, seven bins, no head) | $75.16 \pm 0.73$ | 77.40 |
| no filter | $74.19 \pm 0.72$ | 76.48 |

Table 19: The MCD filter against removing the same number of rows at random (57 groups, AUCROC points; mean, 95% interval, wins/losses). The "$p$" of "filter − random" is a bootstrap $p$-value from the same group resamples; the last column is the sum of the two contrasts before it.

| Detector | filter − random | $p$ | random − no filter | filter − no filter |
|---|---|---|---|---|
| Euclidean kNN | $+7.55$ $[+3.93, +12.02]$, 36/21 | .001 | $+1.72$ $[+0.70, +2.91]$ | $+9.27$ |
| LOF | $+12.79$ $[+8.63, +17.26]$, 41/16 | .0001 | $+0.93$ $[+0.03, +2.00]$ | $+13.71$ |
| isotropic DTE-C | $+2.14$ $[+0.65, +3.90]$, 36/21 | .013 | $-0.13$ $[-0.56, +0.31]$ | $+2.00$ |
| ADTE-C | $+2.88$ $[+1.02, +5.17]$, 32/25 | .009 | $-0.19$ $[-0.71, +0.29]$ | $+2.69$ |

on average (median 1.68, maximum 19.5 on `13_fraud`), about one part in 86 of the full pipeline's training time.

Table 20: Training time per training run (seconds, mean over the 121 member datasets and the seeds) and the total per seed. The filter's own time (2.5 s per dataset on a CPU) is excluded.

| Configuration | per training run | per seed (121 datasets) |
|---|---|---|
| ADTE-C (filter, aligned corruption, head, $K = 8$) | 215 s | 7.2 h |
| DTE-C, no filter, no head, published $K = 7$ | 219 s | 7.4 h |
| the same with the MCD filter | 158 s | 5.3 h |

## D Planted Anomalies and the Crossed Test

**Construction.** For each of the 47 tabular groups we fit a Gaussian mixture with full covariances to the dataset's labeled normal rows, choosing one to nine components by the Bayesian information criterion (BIC) with a covariance regularization of $10^{-6}$, and draw a synthetic copy of the dataset from it, with as many normal rows and as many anomalies as the dataset has after this experiment's loader rules (resampling to 1,000 rows below 1,000; a cap of 10,000 rows above). Every anomaly is a further draw from the mixture displaced by a vector $\delta$. Directions and scales are defined in the coordinates standardized on the labeled normal rows, with $C_N$ their correlation matrix and $r = 5$ the Mahalanobis scale. The leading- and minor-direction anomalies are displaced rigidly along the leading eigenvector of $C_N$ and along its smallest non-degenerate eigenvector (the eigenvector of the smallest eigenvalue above $10^{-9}$ times the largest), by $r\sqrt{\lambda}$ times the eigenvector with $\lambda$ its eigenvalue, so that $\delta^\top C_N^{-1}\delta = r^2$ exactly. The sparse perturbation adds a unit displacement of random sign on up to three coordinates per anomaly and scales all anomalies of a dataset by one common constant, chosen so that the mean of $\delta^\top C_N^{-1}\delta$ over the anomalies equals $r^2$; a common scale rather than a per-anomaly one keeps the scaling from re-introducing the correlation structure this perturbation is meant to ignore. The random-direction perturbation draws a direction uniformly on the sphere for each anomaly and scales each anomaly so that its own $\delta^\top C_N^{-1}\delta$ equals $r^2$. For the sparse and random-direction perturbations the eigenvalues of $C_N$ are floored at $10^{-2}$ (the mean eigenvalue of a correlation matrix being 1) when the Mahalanobis scale is computed. The two DTE-C models of Table 7 are ADTE-C and isotropic DTE-C.

**Directions at a common Mahalanobis scale.** Take the normal rows to be $\mathcal{N}(0, C)$ in the standardized coordinates and the corruption's covariance to be this same $C$, $C = U\Lambda U^\top$ with eigenvalues $\lambda_1 \geq \cdots \geq \lambda_D$ and eigenvectors $u_1, \ldots, u_D$. A displacement $\delta$ at Mahalanobis scale $r$ satisfies $\lambda_D r^2 \leq \|\delta\|^2 \leq \lambda_1 r^2$, with equality along $u_D$ and $u_1$ respectively: at a common Mahalanobis scale the leading-direction shift is the longest in ordinary distance and the minor-direction shift the shortest, which is what a Euclidean neighbor distance measures. Corruption adds noise to the normal rows. Isotropic corruption with noise standard deviation $\sigma_t$ at timestep $t$ ($\sigma_t^2 = 1 - \bar{\alpha}_t$) gives $\mathcal{N}(0, C + \sigma_t^2 I)$, with variance $\lambda_i + \sigma_t^2$ along $u_i$; aligned corruption gives $\mathcal{N}(0, (1 + \sigma_t^2)\, C)$, with variance $(1 + \sigma_t^2)\lambda_i$. The ratio of noise to the data's spread along $u_i$ is therefore $\sigma_t^2/\lambda_i$ under isotropic corruption, smallest along the leading directions and largest along the minor ones, and $\sigma_t^2$ in every direction under aligned corruption. A displacement says less about the diffusion time in directions where the data's spread dominates the noise: under isotropic corruption a direction whose variance far exceeds the noise level tells the time estimate almost nothing, whereas aligned corruption scales the noise to the spread in every direction, in the data's own units, and so brings the leading directions into the estimate. In this Gaussian case the corrupted point's marginal under aligned corruption is $\mathcal{N}(0, (1 + \sigma_t^2)\, C)$ at every $t$, so the timestep posterior depends on $x_t$ only through $x_t^\top C^{-1} x_t$, the Mahalanobis form, whereas under isotropic corruption it depends on $x_t^\top (C + \sigma_t^2 I)^{-1} x_t$, whose relative weights across directions change with $t$.

**The crossed test.** For each of the 47 tabular groups and each seed we take two adjacent eigenvectors $a$ and $b$ of the planting correlation $C_N$ near its median eigenvalue and the pair $u = (a + b)/\sqrt{2}$, $v = (a - b)/\sqrt{2}$ rotated within their plane, with the rotation angle solved so that the two directions have the same variance under the retained rows' correlation $C_R$. The match has to be solved for rather than assumed, because $C_R$ is the correlation of the rows the filter keeps from a dataset that contains the planted anomalies themselves; the angle is found by a bracketed root-find that repeats the planting, the filtering and the measurement at every trial. Directions enter the corruption operator in the pipeline's standardized coordinates: $a_u$ and $a_v$ are $u$ and $v$ rescaled coordinate-wise by the standardization and normalized, a diagonal rescaling that preserves the Mahalanobis radius but not orthogonality, so $|\cos(a_u, a_v)|$ is measured (tolerance $0.15$), $a_u^\top C_R a_u$ and $a_v^\top C_R a_v$ are matched to a relative tolerance of $0.50$, and the planting variances of $u$ and $v$ under $C_N$ to $0.20$. Three models are then trained on the retained rows: aligned corruption with the correlation matrix of $C_R + \gamma_u a_u a_u^\top$, aligned corruption with that of $C_R + \gamma_v a_v a_v^\top$, and the isotropic reference; $\gamma_u$ and $\gamma_v$ are solved separately so that the two inflated operators raise the variance along their own direction by the same increment, which is checked on the constructed matrices; the common increment is the smaller of half the largest increment either direction can reach after the renormalization to a correlation matrix and four times the smaller of the two directions' variances, and a construction whose increment would fall below $0.25$ is refused. Two anomaly sets are planted, displaced along $u$ and along $v$. A matched score is the inflated-$u$ model on the $u$ anomalies or the inflated-$v$ model on the $v$ anomalies, a crossed score the other pairing. The orientation contrast, the inflated-$u$ model minus the inflated-$v$ model averaged over the two anomaly sets, is $-0.2$ $[-0.7, +0.2]$; because $u$ and $v$ are labels the construction assigns, this quantity changes sign under relabeling, and only its magnitude and interval carry information.

**Groups and seeds in the crossed test.** Of 235 attempted (dataset, seed) constructions, 186 were admitted: 28 groups at five seeds, 7 at four, 3 at three, 4 at two and 1 at one; four groups have none (`25_musk`, `31_satimage-2`, `33_skin` and `34_smtp`, the last two with three features). Of the 49 refusals, 41 are searches that found no admissible pair within their budget, most often because the signed variance difference between the two directions did not change sign over the half turn searched (the filter's retained set, and with it $C_R$, is not a continuous function of the planted rows, so the difference need not cross zero); 5 stopped with the variance match above its tolerance of 0.50 in relative terms; and 3 (`17_InternetAds`, three seeds) were refused by a check of the construction itself. The refusals concentrate on heavily contaminated datasets and on low-dimensional ones.

## E Head Effects and Stability Statistics

All statistics in this section are at $K = 8$, in the setting of Table 1, on the 57 groups unless stated. The seed spread of a group is the standard deviation of its AUCROC over the five seeds; ratios of spreads are geometric means over groups, that is, exponentials of the mean log ratio, and their intervals are

the exponentiated 95% group-bootstrap intervals of the mean log ratio; a reduction of the spread quoted in Section 4.3 is one minus this ratio.

**Mean effects of the head.** The head's effect on mean AUCROC is $+0.67$ points $[-0.15, +1.42]$ without the filter under isotropic corruption, $+0.08$ $[-0.33, +0.49]$ with the filter under isotropic corruption and $+0.33$ $[+0.03, +0.68]$ with the filter under aligned corruption; of the three intervals, only the last excludes zero.

**Seed spread by setting.** Table 21 gives, for the three settings, the seed spread without the head relative to the filtered isotropic setting, the head's reduction of the spread, and the head's effect on each group's lowest and highest seed outcome. The two multipliers without the head are $2.76$ $[2.00, 3.99]$ for removing the filter (wider on 49 of the 57 groups) and $1.19$ $[1.01, 1.42]$ for aligning the corruption (wider on 34); the two share the filtered isotropic reference and are not independent. With the head, aligning the corruption multiplies the spread by $0.94$ $[0.78, 1.14]$ (27 groups wider, 30 narrower). The head's reductions in log spread relative to the filtered isotropic setting are $+0.41$ $[+0.17, +0.69]$ without the filter (Wilcoxon $p = .002$) and $+0.23$ $[+0.04, +0.43]$ with the filter under aligned corruption ($p = .06$). The head's reduction of the spread is largest where the spread without the head is largest.

Table 21: What the regression head changes, by setting (57 groups). Spread multipliers are geometric means over groups of the ratio of seed spreads; "groups narrower" counts the groups whose spread is smaller with the head; the lowest- and highest-outcome rows are the head's mean effect on each group's lowest and highest AUCROC over the five seeds (points, 95% interval, wins/losses; ties are not counted).

| | no filter, isotropic | filter, aligned | filter, isotropic |
|---|---|---|---|
| spread without head, relative to filter + isotropic | 2.76 [2.00, 3.99] | 1.19 [1.01, 1.42] | 1 |
| spread with head / without head | 0.656 | 0.788 | 0.990 |
| groups narrower with the head (of 57) | 44 | 34 | 29 |
| lowest seed outcome | +1.91 [+0.68, +3.22], 41/16 | +0.63 [+0.20, +1.22], 34/23 | +0.48 [−0.10, +1.11], 27/30 |
| highest seed outcome | +0.03 [−0.95, +0.95], 32/25 | −0.09 [−0.48, +0.28], 25/30 | −0.15 [−0.66, +0.31], 30/25 |

**Rank decomposition.** Sorting each group's five seed outcomes separately with and without the head and differencing rank by rank gives Table 22; each rank contributes one fifth of its entry to the mean effect. Without the filter the two lowest ranks contribute $+0.661$ of the $+0.671$ mean effect, and with the filter under aligned corruption $+0.255$ of $+0.327$.

Table 22: The head's effect on each rank of a group's five seed outcomes (57 groups, AUCROC points, mean over groups), lowest rank first.

| Setting | lowest | second | third | fourth | highest | mean of five |
|---|---|---|---|---|---|---|
| no filter, isotropic | +1.91 | +1.40 | +0.05 | −0.03 | +0.03 | +0.67 |
| filter, isotropic | +0.48 | +0.28 | −0.09 | −0.11 | −0.15 | +0.08 |
| filter, aligned | +0.63 | +0.64 | +0.32 | +0.13 | −0.09 | +0.33 |

**One output alone at inference.** Under the filter and aligned corruption, the trained ADTE-C models were used at inference with one output alone, without retraining (Appendix F gives the means). With the regression output alone at inference, they have a narrower seed spread than models trained without the head: the reduction in log spread is $+0.23$ $[+0.08, +0.37]$ (bootstrap $p = .001$; $+0.22$ $[+0.06, +0.37]$ on the 47 tabular groups), a spread ratio of $0.798$ $[0.692, 0.919]$ against $0.788$ for the combined score (Table 21). The combined score is narrower than the regression output alone by only a further $+0.01$ $[-0.07, +0.09]$ in log spread ($p = .72$; $+0.01$ $[-0.10, +0.10]$ on the 47). With the categorical output alone at inference, they still have a narrower spread than models trained without the head, a reduction in log spread of $+0.14$ $[+0.02, +0.27]$ ($p = .017$; $+0.12$ $[-0.01, +0.27]$ on the 47), so part of the head's effect on the spread arises in training; adding the regression output to the score at $\omega = 0.5$ gives a further reduction in log spread of $+0.10$ $[+0.03, +0.17]$ ($p = .002$; $+0.10$ $[+0.04, +0.17]$ on the 47). The intervals and $p$-values in this paragraph come from a bootstrap that resamples groups, with `19_landsat` and `30_satellite` as one cluster.

# F Sensitivity Checks and Geometry Constructions

**Shrinkage $s$.** Table 23 gives ADTE-C at $s \in \{0, 0.1, 0.25, 0.5, 0.7\}$ and ADTE-IG at $s \in \{0, 0.25, 0.5, 0.75, 1\}$. ADTE-C's mean AUCROC declines from 77.0 at $s = 0$ to 76.4 at $s = 0.7$ on the 57 groups (79.6 to 78.9 on the 47 tabular groups), and its AUCPR is highest at $s = 0$ as well: 36.7 against 35.7 to 36.0 at $s > 0$ (41.3 against 40.1 to 40.4 on the 47). ADTE-IG uses corruption shrunk toward isotropy; in the sweep, mean AUCPR increases toward isotropy while mean AUCROC varies little across the tested settings (a span of 0.44 points on the 47 tabular groups and 0.40 on the 57). The value we use does not differ significantly from full alignment: at $s = 0.75$ against $s = 0$, AUCROC $-0.12$ $[-1.05, +0.60]$ and AUCPR $+0.64$ $[-0.62, +2.04]$ on the 47 tabular groups, and $-0.07$ $[-0.85, +0.53]$ and $+0.53$ $[-0.51, +1.72]$ on the 57 (no multiplicity correction).

Table 23: The shrinkage sweep with the filter (%). ADTE-C at $K = 8$ with the head; ADTE-IG without bins or head. AUCROC ± SD over seeds and AUCPR, on the 57 groups and on the 47 tabular groups.

| | ADTE-C | | | | | ADTE-IG | | | |
|---|---|---|---|---|---|---|---|---|---|
| $s$ | AUCROC (57) | AUCPR (57) | AUCROC (47) | AUCPR (47) | $s$ | AUCROC (57) | AUCPR (57) | AUCROC (47) | AUCPR (47) |
| 0 | $77.02 \pm 0.19$ | 36.74 | $79.64 \pm 0.15$ | 41.27 | 0 | $75.89 \pm 0.43$ | 36.41 | 78.70 | 40.97 |
| 0.1 | $76.70 \pm 0.16$ | 35.99 | $79.19 \pm 0.16$ | 40.36 | 0.25 | $75.99 \pm 0.30$ | 36.71 | 78.84 | 41.34 |
| 0.25 | $76.53 \pm 0.15$ | 35.80 | $79.06 \pm 0.17$ | 40.15 | 0.5 | $75.92 \pm 0.43$ | 36.92 | 78.70 | 41.58 |
| 0.5 | $76.41 \pm 0.22$ | 35.74 | $78.93 \pm 0.23$ | 40.06 | 0.75 | $75.81 \pm 0.35$ | 36.94 | 78.57 | 41.61 |
| 0.7 | $76.36 \pm 0.17$ | 35.76 | $78.86 \pm 0.16$ | 40.08 | 1 | $75.59 \pm 0.63$ | 37.01 | 78.40 | 41.71 |

**Filter settings.** Table 24 varies the two settings of the MCD filter that the rest of the paper holds fixed, the support fraction $h = 0.75$ and the PCA dimension of 32, for ADTE-C with everything else unchanged. The support fraction is nominal: on the 47 tabular groups the median fraction of rows kept is 0.54, 0.69, 0.73, 0.77, 0.80 and 0.83 for $h = 0.50$, 0.70, 0.75, 0.80, 0.85 and 0.90. One step below the default is not distinguishable from zero ($h = 0.70$: $-0.07$ $[-0.63, +0.47]$ on the 47); $h = 0.50$ loses on 32 of the 47 groups; at $h \geq 0.80$ the median group on the 47 moves by at most 0.1 points, and the mean loss there, 1.6 to 2.0 points, comes from three to eight groups on which the wider filter keeps more anomalies among the training rows. For the PCA dimension, 33 of the 47 tabular groups have at most 32 features, so a PCA dimension of 64 changes nothing for them; on the 14 groups with more than 32 features the difference is $-0.91$ $[-2.52, +0.62]$, with movements of both signs, and on the 47 the differences for PCA dimensions 16 and 64, $-0.05$ and $-0.27$, are not distinguishable from zero. The projection is our substitute for a robust estimate in the original coordinates when the features outnumber what FastMCD handles well; a regularized MCD (Boudt et al., 2020) is the alternative that avoids the projection. These are ADTE-C's sensitivities to the filter's settings, not the alignment gain's.

Table 24: Sensitivity of ADTE-C to the filter's settings: mean AUCROC minus the default ($h = 0.75$, PCA dimension 32), in points, with 95% intervals; 57 groups and the 47 tabular groups.

| Setting | 57 groups | 47 tabular |
|---|---|---|
| support fraction $h = 0.50$ | $-1.60$ [$-3.19$, $-0.30$] | $-1.92$ [$-3.81$, $-0.36$] |
| $h = 0.70$ | +0.01 [$-0.46$, +0.48] | $-0.07$ [$-0.63$, +0.47] |
| $h = 0.80$ | $-1.34$ [$-3.49$, +0.05] | $-1.67$ [$-4.12$, +0.00] |
| $h = 0.85$ | $-1.55$ [$-3.75$, $-0.01$] | $-2.01$ [$-4.53$, $-0.18$] |
| $h = 0.90$ | $-1.23$ [$-3.10$, +0.18] | $-1.62$ [$-3.78$, +0.07] |
| PCA dimension 16 | $-0.01$ [$-0.67$, +0.54] | $-0.05$ [$-0.83$, +0.59] |
| PCA dimension 64 | $-0.12$ [$-0.56$, +0.26] | $-0.27$ [$-0.79$, +0.19] |

**Bin count $K$.** ADTE-C at $K = 6, 7, 8, 9, 10$ has mean AUCROC $77.00 \pm 0.23$, $77.06 \pm 0.12$, $77.02 \pm 0.19$, $77.08 \pm 0.16$ and $77.09 \pm 0.29$ and mean AUCPR 36.46, 36.55, 36.74, 36.64 and 36.52 on the 57 groups (on the 47 tabular groups AUCROC 79.59, 79.60, 79.64, 79.75 and 79.76): the tested values span 0.1 AUCROC point on the 57 and 0.2 on the 47, and no per-group test separates them. The alignment gain itself is not specific to $K = 8$: in one run that trains aligned and isotropic DTE-C at $K = 7$, 8 and 9, aligned minus isotropic is $+0.76 \pm 0.33$, $+0.60 \pm 0.18$ and $+0.89 \pm 0.16$ AUCROC points on the 57 groups and $+0.79 \pm 0.37$, $+0.83 \pm 0.18$ and $+1.02 \pm 0.20$ on the 47 tabular groups (mean ± SD over seeds of the seed-paired difference).

**Head weight $\lambda$.** The $\lambda$ ladder was run without the filter under isotropic corruption at $K = 9$ (57 groups): mean AUCROC / AUCPR $74.1 \pm 0.2$ / 30.5 at $\lambda = 0.1$, $73.5 \pm 0.3$ / 29.5 at $\lambda = 1$ and $71.6 \pm 0.5$ / 26.2 at $\lambda = 10$, against $73.4 \pm 0.6$ / 29.5 for DTE-C without the head at its published $K = 7$ (on the 47 tabular groups, AUCROC 76.2, 75.3 and 73.4 against 75.3). The no-head baseline is at $K = 7$, so the first step changes $K$ as well as adding the head.

**Inference weight $\omega$.** The trained ADTE-C models were used at inference, without retraining, with $\omega \in \{0, 0.1, 0.5, 0.9, 1\}$, from the categorical output alone to the regression output alone. Mean AUCROC on the 57 groups is 76.80, 76.99, 77.02, 76.98 and 76.96, and mean AUCPR 36.56, 36.81, 36.74, 36.32 and 36.28. The AUCROC range across the five settings is 0.22 points on the 57 groups (0.09 on the 47 tabular groups); the categorical output alone is 0.22 points below $\omega = 0.5$ $[-0.03, +0.48]$ and the regression output alone 0.07 below.

**The held-out median split.** The result of Section 4.4 on the energy share, the share of the anomalies' squared deviation from the normal mean, in coordinates standardized on the labeled normal rows, that lies along the leading directions, is tested without choosing a threshold on the data it is applied to: the 47 tabular groups are split at random into 24 and 23, the median energy share of the 24 is applied to the 23, and the alignment gain of the held-out groups above the median minus that of those below is averaged over 1,000 random splits, giving $+1.74$ AUCROC points. Its one-sided $p$-value, .010, comes from permuting the energy shares across groups and repeating the procedure inside every permutation with 200 splits in place of 1,000, with `19_landsat` and `30_satellite`, which share one feature matrix, kept as one block whose two shares exchange only with each other; the same block permutation gives the $p$-value of the Spearman correlation reported in the main text. When the split is repeated with the aligned and isotropic models behind Table 1, held-out groups above the threshold gain 1.66 AUCROC points more from alignment than those below it on the 47 tabular groups ($p = .010$), and 1.49 more on the 57 ($p = .009$); on the same models a single split at the median of all 47 shares puts 16 of the 23 groups above the median among those that gain from alignment, against 9 of the 24 below it (Fisher $p = .041$; mean gains $+1.87$ and $+0.00$); at $K = 9$, with the aligned and isotropic models of the bin-count sweep, the same held-out comparison gives 2.20 points on the 47 ($p = .002$).

**Within-dataset ranks.** The within-dataset result of Section 4.4 is computed on 45 of the 47 tabular groups (`16_http` and `34_smtp` are excluded because their anomalies' within-dataset energy share, defined below, is constant, so the correlation is undefined). In coordinates standardized on the retained rows (constant columns excluded), we take for each anomaly the share of its squared deviation from the retained rows' mean that falls in the leading eigenvectors of the retained rows' correlation, the eigenvectors that carry 90% of the trace (the anomaly's within-dataset energy share; up to a coordinate-wise rescaling, this correlation is the corruption's covariance), and the change in its percentile rank among the dataset's rows between aligned and isotropic training, positive when aligned training ranks it as more anomalous. Sorting each dataset's anomalies into quintiles of the share, the mean rank change rises monotonically from the lowest to the highest quintile; the top quintile exceeds the bottom one by $+0.027$ in percentile-rank units on average and on 37 of the 45 groups (sign test $p = 1.5 \cdot 10^{-5}$). The within-group Spearman correlation between an anomaly's share and its rank change has median $+0.19$ across the 45 groups and is positive on 35 of them (sign test $p = 2.5 \cdot 10^{-4}$; $p = .001$ under a permutation test that takes the most favorable of eight versions of the share, with eigensets carrying 10% to 95% of the trace and one version without a cutoff). Table 25 gives both statistics per group.

**Isotropic noise magnitudes.** Five isotropic magnitudes $m \in \{0.5, 1, 2, 3, 5\}$ times the default were trained on the 47 tabular groups, and aligned corruption at the same five magnitudes beside them. Isotropic DTE-C's mean AUCROC is 78.19, 78.72, 78.31, 77.83 and 77.29, so the default $m = 1$ is the best of the five; aligned DTE-C's is 78.65, 79.55, 79.38, 78.95 and 78.10, above the isotropic model at every magnitude, by $+1.07$ $[+0.27, +2.08]$ at $m = 2$. Here $m$ multiplies the per-timestep noise standard deviation $\sigma_t$, the matrix $\Sigma$ being unchanged, so the injected energy grows as $m^2$. The aligned default is above each isotropic magnitude: aligned DTE-C at $m = 1$ minus isotropic DTE-C at $m = 0.5$, 1, 2, 3 and 5 is $+1.36$ $[+0.34, +2.58]$, $+0.83$ $[+0.11, +1.71]$, $+1.23$ $[+0.05, +2.60]$, $+1.71$ $[+0.08, +3.78]$ and $+2.26$ $[+0.20, +4.81]$ (47 groups; group-bootstrap intervals).

Table 25: Within-dataset result per group (45 tabular groups): the mean change in percentile rank of the anomalies in the top quintile of the within-dataset energy share ($Q_5$) minus that of the bottom quintile ($Q_1$), and the within-group Spearman correlation $\rho$ between an anomaly's share and its rank change.

| Group | $Q_5 - Q_1$ | $\rho$ | Group | $Q_5 - Q_1$ | $\rho$ | Group | $Q_5 - Q_1$ | $\rho$ |
|---|---|---|---|---|---|---|---|---|
| 1_ALOI | +0.009 | +0.10 | 17_InternetAds | +0.052 | +0.16 | 32_shuttle | +0.000 | −0.04 |
| 2_annthyroid | −0.005 | −0.10 | 18_Ionosphere | −0.031 | +0.03 | 33_skin | +0.099 | +0.66 |
| 3_backdoor | +0.020 | +0.19 | 19_landsat | +0.042 | +0.19 | 35_SpamBase | +0.028 | +0.36 |
| 4_breastw | +0.087 | +0.63 | 20_letter | +0.095 | +0.46 | 36_speech | −0.011 | −0.17 |
| 5_campaign | +0.024 | +0.00 | 21_Lymphography | +0.004 | +0.70 | 37_Stamps | −0.038 | −0.33 |
| 6_cardio | +0.060 | +0.60 | 22_magic.gamma | −0.045 | −0.24 | 38_thyroid | −0.001 | −0.10 |
| 7_Cardiotocography | +0.043 | +0.30 | 23_mammography | −0.005 | −0.24 | 39_vertebral | +0.094 | +0.34 |
| 8_celeba | +0.013 | +0.16 | 24_mnist | +0.021 | +0.17 | 40_vowels | +0.002 | +0.13 |
| 9_census | +0.008 | +0.06 | 25_musk | +0.003 | +0.16 | 41_Waveform | −0.094 | −0.03 |
| 10_cover | +0.025 | +0.19 | 26_optdigits | +0.102 | +0.57 | 42_WBC | +0.003 | +0.22 |
| 11_donors | +0.011 | −0.06 | 27_PageBlocks | +0.017 | +0.20 | 43_WDBC | +0.026 | +0.52 |
| 12_fault | +0.093 | +0.31 | 28_pendigits | +0.020 | +0.21 | 44_Wilt | +0.031 | +0.13 |
| 13_fraud | +0.038 | +0.43 | 29_Pima | +0.092 | +0.57 | 45_wine | +0.024 | +0.85 |
| 14_glass | +0.019 | +0.28 | 30_satellite | +0.073 | +0.33 | 46_WPBC | +0.057 | +0.20 |
| 15_Hepatitis | +0.024 | −0.13 | 31_satimage-2 | +0.013 | +0.78 | 47_yeast | +0.076 | +0.71 |

**Conventions of Table 8's orientation rows.** Each of the six orientation rows is the paired per-group difference between the perturbed operator and the correlation on the 47 tabular groups, with a 95% percentile interval from 20,000 group-paired resamples drawn from one generator consumed in table order (a fresh generator moves an interval end by a few hundredths) and a two-sided Wilcoxon signed-rank test. The perturbed operators are built from the retained rows' correlation $C = U\Lambda U^\top$ as follows. The operator in a random basis with the same eigenvalues is $C_1 = (UQ)\Lambda(UQ)^\top$ with $Q$ a seeded random orthogonal matrix (uniformly distributed, from the QR factorization of a Gaussian matrix with the standard sign correction) whose determinant is made $+1$ by flipping the sign of one column, which leaves $C_1$ unchanged; the spectrum and trace are those of $C$ and only the eigen-directions change; the diagonal is not kept, so $C_1$ and the rotated operators $C_\xi$ below are not correlation matrices. The rotated operators are $C_\xi = (UQ_\xi)\Lambda(UQ_\xi)^\top$, where $Q_\xi$ interpolates from $I$ at $\xi = 0$ to $Q$ at $\xi = 1$ by rotating each $2 \times 2$ block of $Q$'s real Schur form by $\xi$ times its angle, at $\xi = 0.15, 0.30, 0.50, 0.70$; the angles given in Table 8, $12°$, $23°$, $38°$ and $49°$, are the mean principal angles between the leading eigenspaces of $C$ and of $C_\xi$, of dimension $k = \min(k_{90}, \max(1, \lfloor D/2 \rfloor))$ with $k_{90}$ the number of eigenvectors carrying 90% of the trace. The sign-flipped operator is $C_b = \mathrm{diag}(b)\, C\, \mathrm{diag}(b)$ with $b \in \{-1, +1\}^D$ having $b_1 = +1$ and exactly $\lfloor D/2 \rfloor$ negative entries among the other coordinates, chosen uniformly at random: the correlations between the two coordinate blocks change sign, the diagonal, the spectrum and the condition number are unchanged. The row averages three such patterns per dataset and seed (two where $D = 3$). The mean AUCROC of DTE-C on the 47 tabular groups is 79.55 with $C$ (this experiment's own run of the default), 79.03, 78.72, 78.40 and 78.40 along the path, 78.44 in the random basis and 78.72 with isotropic corruption.

**Two groups that share a feature matrix.** `19_landsat` and `30_satellite` are one feature matrix under two anomaly labelings. Without them, ADTE-C's margin over Mahalanobis kNN on the remaining 45 tabular groups is $+1.80$ AUCROC points $[+0.53, +3.20]$, 24 wins and 21 losses.

## G Correlation Pools, Whitening and Reduced Group Counts

**Reduced group counts.** Where an analysis uses fewer than the 47 tabular groups, its count follows an eligibility rule stated here with the excluded groups named; where fewer than five seeds were used, the seed counts are listed with them.

- **34** (deleting the retained anomalies from the pool; ten seeds): the 35 groups that retain at least one anomaly at some seed, without `17_InternetAds`, which the minimum pool size of $2D + 10$ rows excludes at $D = 1{,}555$. A group that retains no anomaly has nothing to delete; over all 46 eligible groups, those groups enter as zeros.
- **25** (replacing the retained anomalies by rejected normal rows): the 34 groups above without eight that lack enough filter-rejected normal rows for a count-matched replacement (`4_breastw`, `12_fault`, `22_magic.gamma`, `29_Pima`, `30_satellite`,

35_SpamBase, 41_Waveform, 47_yeast) and without 3_backdoor, whose replaced pool failed its reconstruction check at one seed; over the 37 groups these rules admit, the twelve that retain no anomaly enter as zeros.

- **35** (the direction of the retained anomalies): the 47 without the twelve groups that retain no anomalies at any seed, listed in Appendix I.
- **40** (the held-out pool): the 47 without seven groups whose labeled-normal pool is smaller than their retained pool on a training half (4_breastw, 12_fault, 22_magic.gamma, 29_Pima, 30_satellite, 35_SpamBase, 47_yeast).
- **43** (the crossed test; 28 groups at five seeds, 7 at four, 3 at three, 4 at two and 1 at one): the 47 without the four groups with no admitted construction, 25_musk, 31_satimage-2, 33_skin and 34_smtp (Appendix D).
- **14** (the PCA-dimension-64 contrast in the text of Appendix F): the tabular groups with more than 32 features, the only ones a PCA dimension of 64 affects.
- **46** (the ALTBI-I and ODIM checks of Table 11): the 47 tabular datasets without 24_mnist, the block both papers describe as tabular (Appendix A).
- **43** (whitening): the 47 without four groups whose whitened models return a constant score at every fitted seed, listed below.
- **45** (the within-dataset energy share): the 47 without 16_http and 34_smtp, whose within-dataset energy share is constant, so the per-anomaly correlation is undefined.
- ODIM's 20news group: both columns of Table 1 average the same five of its six members (Appendix A).

**Pool contrasts.** Relative to the retained pool, ADTE-C's mean AUCROC on the 47 tabular groups changes as follows: deleting the retained anomalies from the pool $-0.42$ $[-0.83, -0.08]$ on the 34 groups that retain anomalies (ten seeds; the retained pool ahead on 27, behind on 7), or $-0.33$ $[-0.64, -0.07]$ over the 46 eligible groups with the groups without retained anomalies entering as zeros; replacing them by rejected normal rows drawn uniformly $-0.52$ $[-0.99, -0.10]$ on the 25 groups where the exchange is possible (ahead on 16, behind on 9), or $-0.35$ $[-0.68, -0.06]$ over the 37 admitted groups; and by the least outlying rejected normal rows $-0.19$ $[-0.42, +0.01]$ over the same 37 (15 ahead, 10 behind, 12 zeros).

**Classical detectors with the labeled-normal pool.** For $T^2$ (in its correlation form) and Mahalanobis kNN ($k = 5$), replacing the correlation of the retained rows by that of the dataset's labeled normal rows, with the retained rows kept as the center and reference set, improves $T^2$ by $+4.57$ $[+2.14, +7.25]$ AUCROC points and Mahalanobis kNN by $+1.57$ $[+0.56, +2.73]$ on the 47 tabular groups, the two contrasts of Section 4.5.

**Whitening.** The whitened model of Section 4.5 transports the retained rows by the inverse of the Cholesky factor of the same $\Sigma$ that aligned corruption uses ($y = L^{-1}(x - \mu)$ with $\Sigma = LL^\top$ and $\mu$ the retained rows' mean) and trains DTE-C with isotropic corruption on the transported rows, so that the corruption law matches the aligned one exactly (Appendix H). On some (dataset, seed) fits the transported rows reach magnitudes at which the network's softmax output becomes constant in single precision and the model returns a constant score, which ranks nothing: all five seeds of 1_ALOI, 5_campaign and 9_census, the three fitted seeds of 3_backdoor, three seeds of 11_donors and of 7_Cardiotocography, two of 6_cardio among the tabular groups, and a few fits in two of the embedding groups. The contrast of Section 4.5 masks these fits and drops the four fully affected groups: $-4.21$ $[-8.27, -0.99]$ on 43 groups (40 at five seeds, 6_cardio at three, 7_Cardiotocography and 11_donors at two). Two other treatments give the same reading: keeping every fit at its returned value, $-4.74$ $[-8.48, -1.56]$ on the 46 groups with complete fits (31 groups lose), and excluding every group with an affected fit, $-2.56$ $[-5.90, -0.08]$ on 40. On the 57 groups the three treatments give $-4.13$ $[-7.28, -1.47]$ (all fits, 56 groups), $-3.59$ $[-6.89, -0.90]$ (53 groups) and $-2.03$ $[-4.89, +0.05]$ (48 groups). Across the treatments the whitening loss on the tabular groups is between three and six times the alignment gain of Table 8.

**The held-out pool.** To separate the pool from the rows it scores, each of the 47 tabular groups is split in half; the filter and both correlation pools (retained rows, labeled normal rows) are built on one

half, the model scores the other half, and the two folds are averaged before the seeds. On the 40 groups whose labeled-normal pool is at least as large as the retained pool on every seed and fold, the retained pool leads the labeled-normal pool by $+0.73$ AUCROC points $[+0.06, +1.57]$, 24 wins, 15 losses and one tie; on all 47 groups by $+0.67$ $[+0.10, +1.39]$, 30 wins, 16 losses and one tie; on the seven excluded groups by $+0.37$ $[-0.03, +0.69]$ (6 and 1), and, with `39_vertebral` counted among the excluded groups, on 39 and 8 groups by $+0.73$ $[+0.04, +1.59]$ and $+0.39$ $[+0.03, +0.67]$. On the 40 comparable groups, the same fitted models score their own training half at $+1.21$ $[+0.47, +2.11]$, so the retained pool's advantage is $+0.48$ $[+0.22, +0.80]$ larger on rows that were in the pool than on rows that were not (30 wins and 10 losses).

**Rotation ladder, random basis and sign flip.** Their construction is given with the conventions of Table 8's orientation rows in Appendix F.

## H Aligned Corruption, Whitening and the Inverse Gamma Variant

**Corruption densities.** Write $\sigma_t^2 = 1 - \bar{\alpha}_t$. Isotropic DTE corrupts a data point $x$ by $x_t = x + \sigma_t \epsilon$ with $\epsilon \sim \mathcal{N}(0, I_D)$, so $q(x_t \mid x, t) = \mathcal{N}(x, \sigma_t^2 I_D)$. The aligned variant draws $\epsilon \sim \mathcal{N}(0, \Sigma)$ with $\Sigma$ fixed during training, so $q_\Sigma(x_t \mid x, t) = \mathcal{N}(x, \sigma_t^2 \Sigma)$ with density

$$q_\Sigma(x_t \mid x, t) = (2\pi\sigma_t^2)^{-D/2} \, |\Sigma|^{-1/2} \exp\left( -\frac{(x_t - x)^\top \Sigma^{-1} (x_t - x)}{2\sigma_t^2} \right). \tag{7}$$

The only $t$-dependent change from the isotropic density is the quadratic form; $|\Sigma|^{-1/2}$ does not depend on $t$, so the bin labels $b(t)$, the regression target $t/T$ and the loss are unchanged.

**Whitening and the timestep posterior.** Let $\Sigma^{-1/2}$ be the symmetric inverse square root and $y = \Sigma^{-1/2} x$. Then $\Sigma^{-1/2} x_t = y + \sigma_t z$ with $z = \Sigma^{-1/2}\epsilon \sim \mathcal{N}(0, I_D)$: aligned corruption of $x$ is isotropic corruption of $y$. More generally, for any invertible linear map $A$ that does not depend on $t$, the distribution of the timestep given the corrupted point is the same in the two coordinate systems. Writing $p(t \mid x_t) \propto p(t) \int q_\Sigma(x_t \mid x, t)\, p(x)\, dx$, with $p(t)$ the uniform distribution of the timestep in training and $p(x)$ the data distribution, and changing variables to $Ax$ and $Ax_t$, every term acquires the same factor $|\det A|^{-1}$, which cancels in the normalization over $t$, so $p(t \mid Ax_t) = p(t \mid x_t)$. The quantity the network is trained to predict is therefore identical in the standardized and in the whitened coordinates. What is not invariant is the fitted network: the initialization, dropout, weight decay and Adam's coordinate-wise step sizes act in the parameter coordinates, and a linear change of the input coordinates changes them (the first layer's weights transform as $W \mapsto W A^{-1}$). The two also differ numerically: the aligned sampler draws noise through a Cholesky factor of $\Sigma$, which tolerates a near-singular $\Sigma$ with mild diagonal jitter (columns that are constant on the retained rows enter with zero correlation and unit variance), whereas whitening needs an inverse factor of $\Sigma$ ($L^{-1}$ in Appendix G), which amplifies estimation error along the directions of smallest variance. The difference measured in Section 4.5 between training in the two coordinate systems is thus a property of the learner, not of its target. $T^2$ floors the eigenvalues of its covariance at $10^{-6}$ of the largest, and Mahalanobis kNN inverts the same retained-row correlation as the corruption uses.

**Maximum-likelihood variance and the inverse Gamma form.** For a fixed clean point $x$ and fixed $\Sigma$, with $\tau = \sigma_t^2$,

$$\log q_\Sigma(x_t \mid x, \tau) = -\frac{D}{2} \log \tau - \frac{(x_t - x)^\top \Sigma^{-1} (x_t - x)}{2\tau} + \text{const}, \tag{8}$$

which is maximized at $\hat{\tau} = (x_t - x)^\top \Sigma^{-1} (x_t - x)/D$: the inferred variance is the Mahalanobis deviation under $\Sigma$ divided by $D$, and since $\sigma_t^2$ increases with $t$, a larger deviation maps to a larger timestep. With an inverse Gamma prior $\tau \sim \mathrm{IG}(a_0, b_0)$ of shape $a_0$ and scale $b_0$, the posterior is

$$\tau \mid x_t, x, \Sigma \;\sim\; \mathrm{IG}\left( a_0 + \frac{D}{2},\; b_0 + \tfrac{1}{2}(x_t - x)^\top \Sigma^{-1} (x_t - x) \right). \tag{9}$$

Dropping the prior (formally $a_0 = -1$, $b_0 = 0$, a constant kernel rather than a distribution) leaves the inverse Gamma kernel with shape $D/2 - 1$ and scale $\frac{1}{2}(x_t - x)^\top \Sigma^{-1} (x_t - x)$. DTE's approximation

(Livernoche et al., 2024, Sections 3.1–3.2) is this kernel with $\Sigma = I_D$ and with the squared distance to the closest point of the dataset in place of the distance to a given clean point, where the released estimator takes the mean distance to the five nearest rows (Appendix A); under aligned corruption the kernel's distance becomes the Mahalanobis form under $\Sigma$, with the same shape. The derivation above conditions on the clean point, and at inference DTE-IG has none: its network predicts the scale $b_\theta(x)$ from the uncorrupted input alone, trained by the log-likelihood of the timestep's variance $\sigma_t^2$ under the inverse Gamma with shape $D/2 - 1$ and scale $b_\theta(x_t)$, that is $(D/2 - 1)\log b_\theta(x_t) - b_\theta(x_t)/\sigma_t^2$ up to constants (the implementation adds a small constant to $b_\theta$ inside the logarithm), and the score is a fixed multiple of $b_\theta(x)$. The released implementation, which we use, divides $b_\theta(x)$ by $D/2 - 1$, whereas the paper's text divides by $D/2$, which gives the mode of that inverse Gamma; the two differ by a positive constant on a given dataset, so every ranking, and every AUCROC and AUCPR reported here, is the same under both.

**Practical condition.** The derivation assumes a fixed, positive definite $\Sigma$. In the implementation $\Sigma$ is the shrunk correlation of Section 3 after any diagonal jitter or isotropic fallback, symmetrized before its Cholesky factorization. The kernel's shape $D/2 - 1$ is positive only for $D > 2$; every ADBench dataset has at least three features. As in DTE, the closest-point approximation is coarsest at large timesteps, where the corrupted point lies far from every clean row and the kernel flattens; aligned corruption changes the distance in that step, not the step itself.

# I The Direction of the Retained Anomalies

This section defines the statistic behind the result of Section 4.5 that the anomalies the filter retains raise the correlation's variance along their own direction by $0.37$ $[+0.17, +0.59]$.

**Definitions.** For each dataset and seed, on the standardized rows, let $R$ be the rows the filter retains, $S$ the anomalies among them and $R \setminus S$ the retained normal rows. Let $C_R$ be the correlation matrix of the rows in $R$ and $C_{R\setminus S}$ that of the rows in $R \setminus S$, both computed on the columns that have non-zero spread in both row sets (a constant column has no correlation), and let $\Delta C = C_R - C_{R\setminus S}$, the change the retained anomalies make to the correlation the corruption is aligned with. Let $w$ be the unit vector, on the same columns, from the mean of the retained normal rows to the mean of the retained anomalies. The statistic is $w^\top \Delta C\, w$, the change in variance along $w$ in correlation units: both matrices have unit diagonal, so $\mathrm{tr}(\Delta C) = 0$ identically and the average of $w^\top \Delta C\, w$ over all directions is exactly zero. A positive value says that the retained anomalies raise the variance along their own direction at the expense of other directions. Per group the statistic is the mean over its member datasets and seeds; the interval is the 95% group bootstrap.

**Population and seeds.** Of the 47 tabular groups, 35 retain at least one anomaly after the filter; the other twelve (`11_donors`, `16_http`, `21_Lymphography`, `25_musk`, `28_pendigits`, `31_satimage-2`, `32_shuttle`, `34_smtp`, `38_thyroid`, `42_WBC`, `43_WDBC` and `45_wine`) retain none at any of the five seeds, so there $\Delta C = 0$ and $w$ is undefined, and the statistic is not defined for them. The statistic rests on five seeds for 34 of the 35 groups and on four for `37_Stamps`, whose filter retained no anomaly at one seed. Two of the groups, `19_landsat` and `30_satellite`, are one feature matrix under two labelings and return identical values.

**Related statistics on the same 35 groups.** The percentile of $w^\top \Delta C\, w$ among the values of $u^\top \Delta C\, u$ for 10,000 random unit directions $u$ averages $0.85$ $[0.73, 0.95]$ over groups (above $0.5$ on 30 of the 35); the acute angle between $w$ and the leading eigenvector of $\Delta C$ averages $33°$ $[28°, 39°]$ (below $45°$ on 27 of the 35); the largest principal angle between the leading eigenspaces of $C_{R\setminus S}$ and $C_R$ (of dimension $k = \min(k_{90}, \max(1, \lfloor D/2 \rfloor))$, with $k_{90}$ the number of eigenvectors carrying 90% of $C_R$'s trace and $D$ the number of live columns) averages $20.5°$ $[12.4°, 29.8°]$; and the ratio of Frobenius norms $\|\Delta C\|_F / \|C_R\|_F$ averages $0.084$ (median $0.047$), so the retained anomalies change the correlation little in norm and mostly along their own direction. This construction adds no anomalies and inflates nothing; it is distinct from the crossed test of Appendix D.

## J WHERE THE NEIGHBOR DETECTORS' FILTER GAIN COMES FROM

The filter's largest gains in Table 1 are those of the Euclidean neighbor-based detectors. Under the unsupervised protocol two bookkeeping effects, unrelated to which rows are anomalous, could contribute to them. First, a retained row scores against itself in the reference set while a removed row cannot (*self-match*). Second, the loader's resampling of datasets below 1,000 rows leaves exact copies, whose distance to their nearest neighbor is zero, and the filter removes some of them (*loader duplicates*). We decompose the filtered-minus-unfiltered AUCROC gain of Euclidean kNN, Mahalanobis kNN and LOF into these two components and the remainder, *reference-set cleaning*, the effect of removing rows from the reference set. Operationally, self-match is the gain under the unsupervised protocol minus the gain when each row is excluded from its own reference set, loader duplicates is the latter minus the gain when the rows are also taken at their native count, and cleaning is that last gain; the three add to the total exactly. Across the five neighbor-based detectors of Table 1 (LOF, LUNAR, both kNNs and DTE-NP), the filter improves 31 to 40 of the 57 groups each.

Table 26 gives the decomposition on the 57 groups and on the 47 tabular groups (one paired group bootstrap of 10,000 resamples per population). The cleaning component's interval excludes zero for all three detectors on both populations, and the two bookkeeping components together are $0.43$ of Euclidean kNN's total gain, $0.21$ of LOF's and $0.44$ of Mahalanobis kNN's on the 47 tabular groups, below half for all three. The loader-duplicate component is exactly zero on the groups with at least 1,000 native rows (43 of the 57, 35 of the 47), so its 57- and 47-group means are those of the 14 resampled groups (12 of them tabular) spread over each population.

Table 26: The filter's AUCROC gain for the neighbor-based detectors, split into self-match, loader duplicates and reference-set cleaning (points; mean and 95% interval over groups). Components are rounded independently and may differ from the total by $0.001$.

| Detector | total gain | self-match | loader duplicates | cleaning |
|---|---|---|---|---|
| *57 groups* | | | | |
| Euclidean kNN ($k = 5$) | $+9.272$ [$+5.183, +14.124$] | $+1.386$ [$+0.628, +2.282$] | $+2.771$ [$+0.741, +5.331$] | $+5.116$ [$+1.885, +9.272$] |
| Mahalanobis kNN ($k = 5$) | $+8.262$ [$+4.282, +13.097$] | $+1.091$ [$+0.500, +1.793$] | $+2.767$ [$+0.801, +5.141$] | $+4.404$ [$+1.121, +8.655$] |
| LOF ($k = 20$) | $+13.715$ [$+9.286, +18.362$] | $+0.253$ [$+0.043, +0.442$] | $+2.658$ [$+0.554, +5.350$] | $+10.805$ [$+6.881, +15.149$] |
| *47 tabular groups* | | | | |
| Euclidean kNN ($k = 5$) | $+11.069$ [$+6.134, +16.656$] | $+1.607$ [$+0.713, +2.663$] | $+3.187$ [$+0.779, +6.215$] | $+6.275$ [$+2.393, +11.185$] |
| Mahalanobis kNN ($k = 5$) | $+9.700$ [$+4.828, +15.314$] | $+1.291$ [$+0.568, +2.133$] | $+2.956$ [$+0.690, +5.718$] | $+5.452$ [$+1.467, +10.557$] |
| LOF ($k = 20$) | $+16.292$ [$+11.210, +21.505$] | $+0.301$ [$+0.054, +0.518$] | $+3.177$ [$+0.693, +6.315$] | $+12.814$ [$+8.302, +17.541$] |

## K THE TEN EMBEDDING GROUPS

The 10 image and text groups are where the energy share of Section 4.4 takes its extreme values (Table 27) and where Section 4.6 reports the reversed ordering of the detectors. The share uses the labels and no trained model; it depends on the loader's draw of the rows (seed 0) and on no training seed.

The text groups' shares, $0.34$ to $0.50$, are an order of magnitude above the image groups', $0.01$ to $0.07$. But in the text groups a single eigenvector carries 37 to 54% of the normal variance, so their shares are close to proportional (ratio $0.79$ to $0.99$): the text anomalies lie in the leading direction to about the extent that the normal rows' own variance does. In the image groups the ratio is $0.10$ to $0.55$: their anomalies lie relatively more outside the leading directions than the normal rows do.

**Damping the corruption's leading directions.** Capping the five largest eigenvalues of the corruption's correlation at the sixth and rescaling the matrix so that its trace stays $D$ raises ADTE-C's AUCROC on the ten groups by $+0.80$ points $[+0.23, +1.45]$ (8 of the 10 groups up; filter and head), nearly all of it in the five text groups ($+1.58$ $[+0.86, +2.30]$, all five up), with the five image groups at $+0.02$ $[-0.22, +0.22]$.

## L REPRODUCIBILITY AND COMPUTE

**Hardware by experiment.** Two GPU types were used, GPU A, an RTX 4000 Ada, and GPU B, an L4; CPU machines (x86 and arm64) ran the filter and the detectors and analyses that Table 28 marks

Table 27: The ten embedding groups: the energy share of Section 4.4 and the alignment contrast. The share is the fraction of the anomalies' squared deviation from the normal mean in the leading eigenvectors of the labeled normals' correlation (the fewest carrying a tenth of the variance); $k_{\text{top}}$ is their number; the mass is the variance they carry; share, $k_{\text{top}}$ and mass are averaged over the group's member datasets; the ratio is the mean over the group's member datasets of share over mass (1 when the anomalies' deviations are isotropic with respect to the normals' correlation). The last column is aligned minus isotropic DTE-C (AUCROC points; one run that trains both models, Appendix L).

| Group | Modality | Share | $k_{\text{top}}$ | Mass | Ratio | aligned − isotropic |
|---|---|---|---|---|---|---|
| CIFAR10 | image | 0.07 | 2.0 | 0.13 | 0.55 | −0.7 |
| FashionMNIST | image | 0.01 | 1.2 | 0.13 | 0.10 | −0.7 |
| SVHN | image | 0.06 | 1.3 | 0.13 | 0.45 | −0.1 |
| MNIST-C | image | 0.03 | 1.0 | 0.15 | 0.18 | −1.0 |
| MVTec-AD | image | 0.03 | 1.0 | 0.23 | 0.10 | −0.3 |
| 20news | text | 0.34 | 1.0 | 0.37 | 0.93 | −1.6 |
| agnews | text | 0.50 | 1.0 | 0.54 | 0.93 | −2.1 |
| amazon | text | 0.36 | 1.0 | 0.38 | 0.94 | −1.0 |
| imdb | text | 0.34 | 1.0 | 0.43 | 0.79 | +3.2 |
| yelp | text | 0.37 | 1.0 | 0.37 | 0.99 | −0.6 |

CPU. Section 4.1 states the assignment, and the table lists it per experiment with the seeds and groups of each. Every contrast compares two models run on the same datasets and seeds. Every contrast between two DTE-C models is computed in one environment; the comparisons between DTE-C and a classical detector pair GPU-trained models with CPU detectors, each in its own environment. ADTE-C's mean AUCROC differs across the two GPUs by $0.13$ points on the 57 groups ($77.02$ on GPU A against $76.89$ on GPU B) and by $0.090$ on the 47 tabular groups ($79.64$ against $79.55$); using GPU B's ADTE-C in place of GPU A's in Table 9 moves every contrast mean by at most $0.09$ points and flips no sign. The corruption-by-head factorial of Table 3 repeated on GPU B gives $76.27$, $76.83$, $76.29$ and $76.89$ for isotropic and aligned corruption without and with the head, alignment effects of $+0.56$ without and $+0.60$ with the head and an interaction of $+0.04$ (Wilcoxon $p = .39$), with the same signs as on GPU A. The rows the filter retains can differ between CPU architectures and thread counts on datasets whose correlation matrix is numerically singular, where FastMCD's subset search can settle in a different local optimum; the two $T^2$ rows and the Mahalanobis kNN row of Table 1 were therefore computed in one x86 environment, and no per-dataset comparison of the filter's retained rows across environments is made.

**Software.** The results were produced with PyTorch 2.4.1 (CUDA 12.4), scikit-learn 1.8.0 (Pedregosa et al., 2011), NumPy 2.4, SciPy 1.17.1, PyOD 2.0.7 (Zhao et al., 2019) and ADBench 0.1.11 (Han et al., 2022). The scikit-learn version determines the FastMCD implementation and with it the rows the filter retains. The code, the configurations of every experiment and the per-group results are in the anonymized repository named in the Reproducibility statement.

## M LIMITATIONS

Our alignment suits the tabular groups, where its gain arises: on the 10 embedding groups the three classical detectors of Table 9 remain ahead of both DTE-C rows. Capping the corruption's leading eigenvalues recovers part of the gap on the text groups (Section 4.6). The filter's benefit is measured under the unsupervised protocol, in which a detector trains on and ranks the same dataset (Section 4.2). The share of the anomalies' squared deviation along the leading eigenvectors (Section 4.4) uses the labels and describes where alignment helped. One global correlation carries second-order structure only, so multimodal or locally curved normal populations are a different regime.

Table 28: Hardware, seeds and groups of every experiment in the paper. GPU A: RTX 4000 Ada; GPU B: L4.

| Experiment | Hardware | Seeds | Groups |
|---|---|---|---|
| *Section 4.2* | | | |
| Table 1, every row, without and with the filter | CPU (classical detectors, kNNs, LOF, DTE-NP; $T^2$ and Mahalanobis kNN rows on x86); GPU A (LUNAR, ODIM, SLAD, DDAE, ALTBI-I, K-DSM, the DTE family) | 5 | 57 |
| DTE-NP against the mean distance to the five nearest rows | CPU (x86) | 5 | 57 |
| Eight filtering policies and the no-filter control | GPU B | 5 | 57 |
| Random-removal control | GPU B (DTE-C models); CPU (Euclidean kNN, LOF) | 5 | 57 |
| Training on labeled normal rows only | GPU B | 5 | 47 and 57 |
| *Section 4.3* | | | |
| Tables 2 and 3 | GPU A | 5 | 57 |
| Training time | GPU A | 5 | 121 datasets |
| Time of the filter | CPU (x86) | 5 | 121 datasets |
| *Section 4.4* | | | |
| Table 7 | GPU B (DTE-C); CPU (kNN) | 5 | 47 |
| Crossed test | GPU B | 1–5 per group (28 groups at five, 7 at four, 3 at three, 4 at two, 1 at one) | 43 |
| Energy share, median split | GPU B | 5 | 47 |
| Within-dataset anomaly ranks | GPU B (scores); CPU (analysis) | 5 | 45 |
| *Section 4.5* | | | |
| Table 8, pool rows, rotation ladder, random basis, sign flip; the five noise magnitudes | GPU B | 5 (4 on one group) | 47 |
| Held-out pool | GPU B | 5, two folds | 40 (and 47) |
| Deleting the retained anomalies from the pool | GPU B | 10 | 46 |
| Replacing them with rejected normal rows | GPU B | 5 | 37 |
| Direction of the retained anomalies | CPU | 5 (4 on one group) | 35 |
| Labeled-normal pool for $T^2$ and Mahalanobis kNN | CPU (arm64) | 5 | 47 |
| Whitened coordinates | GPU B | 5 (3 on one group, 2 on two) | 43 (variants on 40, 46, 48, 53 and 56) |
| *Section 4.6* | | | |
| Table 9 | GPU A (DTE-C rows); CPU (kNN rows; $T^2$ and Mahalanobis kNN on x86) | 5 | 47 and 10 |
| *Section 4.3 and Appendix E* | | | |
| Mean effects of the head, seed spread, lowest outcomes, rank decomposition, one output alone at inference, $\omega$ | GPU A | 5 | 57 |
| *Appendix* | | | |
| ALTBI-I's released pipeline end to end under its own protocol, three initializations (Appendix A) | GPU B | 3 | 46 |
| Released ODIM under its own protocol (Appendix A) | GPU B | 5 | 5 datasets |
| Released DDAE under its own protocol (Appendix A) | GPU B | 3 | 5 datasets |
| Neighbor detectors at $k = 32$ (Appendix A) | CPU (x86) | 5 | 57 |
| Other preprocessing transforms (Appendix C) | GPU B | 5 (3 for min–max) | 57 (and 47) |
| Sweeps of $K$, $s$ and $\lambda$ (Appendix F; Tables 4–6 and 23) | GPU A | 5 | 57 (and 47) |
| Aligned and isotropic DTE-C at $K = 7, 8, 9$ in one run (Appendix F) | GPU B | 5 | 57 and 47 |
| Filter-setting sweep, support fraction and PCA dimension (Appendix F, Table 24) | GPU B | 5 | 57 and 47 |
| Median split on Table 1's aligned and isotropic models and on the $K = 9$ pair of the bin-count sweep (Appendix F) | GPU A | 5 | 47 and 57 |
| Neighbor detectors' gain decomposition (Appendix J) | CPU (x86) | 5 | 57 and 47 |
| Energy share of the embedding groups (Appendix K) | CPU; labels only, no training | loader seed 0 | 10 |
| Embedding groups, aligned − isotropic (Appendix K) | GPU B | 5 | 10 |
| Eigenvalue cap (Section 4.6, Appendix K) | GPU B | 5 | 10 |
| Factorial of Table 3 repeated (Appendix L) | GPU B | 5 | 57 |