Cheap robust learning of data anomalies with analytically solvable entropic outlier sparsification

Illia Horenko[a,1]

[a]Faculty of Informatics, Institute of Computing, Universitá della Svizzera Italiana, TI-6900 Lugano, Switzerland

Edited by David Weitz, Harvard University, Cambridge, MA; received November 2, 2021; accepted January 30, 2022

Entropic outlier sparsification (EOS) is proposed as a cheap and robust computational strategy for learning in the presence of data anomalies and outliers. EOS dwells on the derived analytic solution of the (weighted) expected loss minimization problem subject to Shannon entropy regularization. An identified closed-form solution is proven to impose additional costs that depend linearly on statistics size and are independent of data dimension. Obtained analytic results also explain why the mixtures of spherically symmetric Gaussians—used heuristically in many popular data analysis algorithms—represent an optimal and least-biased choice for the nonparametric probability distributions when working with squared Euclidean distances. The performance of EOS is compared to a range of commonly used tools on synthetic problems and on partially mislabeled supervised classification problems from biomedicine. Applying EOS for coinference of data anomalies during learning is shown to allow reaching an accuracy of 97% ± 2% when predicting patient mortality after heart failure, statistically significantly outperforming predictive performance of common learning tools for the same data.

sparsification | outlier detection | mislabeling | regularization | entropy

Detection of data anomalies, outliers, and mislabeling is a long-standing problem in statistics, machine learning (ML), and artificial intelligence (1–4). Let $\{x_1, x_2, \ldots, x_T\}$ be a fixed dataset (where data instances x_t are possibly augmented with labels), let θ be a set of ML model parameters, and let $g(x_t, \theta)$ be a scalar-valued loss function measuring a misfit of the data instance x_t. Then, a wide class of learning methods and anomaly detection algorithms can be formulated as numerical procedures for a minimization of the following functional:

$$\left\{\hat{w}, \hat{\theta}\right\} = \arg\min_{w_\cdot, \theta} \sum_{t=1}^{T} w_t g(x_t, \theta), \qquad [1]$$

where $0 \leq w_t \leq 1$ is the outlyingness, taking the values close to zero if the data point x_t is an anomaly (1, 5–7). If w and θ are both unknown, then the above problem [1] for simultaneous estimation of model parameters and loss weights becomes ill posed. Common approaches deal with this ill-posedness problem imposing additional parametric assumptions on w, for example, based on parametric thresholding of one-dimensional linear projections in Stahel–Donoho estimators or deploying other parametric tools [like $\chi(D)$-distribution quantiles to determine outliers of a D-dimensional normal distribution] (1, 5, 7, 8). An appealing idea would be to make this ill-posed problem well posed in a nonparametric way, by regularizing it with one of the common regularization approaches. For example, applying $l1$ regularization could result in a sparsification of w and zeroing out the outlying data points from the estimation (9). However, applying $l1$ and other sparsification methods results in a polynomial cost scaling required for a numerical solution of the resulting optimization problems—and would limit the solution of [1] to relatively small problems (10).

The key message of this brief report is in showing that the simultaneous well-posed detection of anomalies and learning of parameters θ in [1] can be achieved computationally very efficiently by means of the minimization of expected loss from the right-hand side of [1]—performed simultaneously to the regularized Shannon entropy maximization of the loss weight distribution w,

$$\left\{w^{(\alpha)}, \theta^{(\alpha)}\right\} = \arg\min_{w \in \mathbb{P}^{(T)}} L(w, \theta, \alpha),$$

$$\text{where } L(w, \theta, \alpha) = \sum_{t=1}^{T} w_t g(x_t, \theta) + \alpha \sum_{t=1}^{T} w_t \log w_t,$$

$$\text{such that } \; w \in \mathbb{P}^{(T)},$$

$$\mathbb{P}^{(T)} := \left\{ w \in \mathbb{R}^T \,\middle|\, w \geq 0 \;\wedge\; \sum_{t=1}^{T} w_t = 1 \right\}. \qquad [2]$$

The following *Theorem* summarizes the properties of this problem's solutions.

Theorem. *For any fixed* $\{x_1, x_2, \ldots, x_T\}$ *and* θ *such that* $\sup_t |g(x_t, \theta)| < \infty$ *and* $\alpha > 0$, *constrained minimization problem [2] admits a unique closed-form solution* $w^{(\alpha)}$,

$$w_t^{(\alpha)} = \frac{\exp\left(-\alpha^{-1} g(x_t, \theta)\right)}{\sum_{t=1}^{T} \exp\left(-\alpha^{-1} g(x_t, \theta)\right)}. \qquad [3]$$

The proof of the *Theorem* is provided in *SI Appendix*. It is straightforward to validate that the numerical cost of computing [3] scales linearly in statistics size T and is independent of the data dimension D—in contrast to common regularization techniques that require polynomial cost scaling in the data dimension D (10).

If the loss function $g(x_t, \theta)$ is a squared Euclidean distance (as in the least-squares methods), then, according to the above *Theorem*, the unique probability distributions w minimizing [2] are from the α-parametric family of spherically symmetric Gaussians, with the dimension-wise variance σ^2 being $\sigma^2 = 0.5\alpha$. This result provides an interesting insight into the density-based methods, for example, in t-stochastic neighbor embedding (t-SNE) (11)—one of the most popular nonlinear dimensions reduction approaches in the area of biomedicine (with over 20,000 citations according to Google Scholar). This method searches for the optimal low-dimensional approximations of the high-dimensional densities defined in a heuristic way as mixtures of spherically symmetric Gaussians,

$$w_t = \frac{\exp\left(-|||x_i - x_j|||^2 / 2\sigma^2\right)}{\sum_k \exp\left(-|||x_i - x_k|||^2 / 2\sigma^2\right)}, \qquad [4]$$

Author contributions: I.H. designed research, performed research, contributed new reagents/analytic tools, analyzed data, and wrote the paper.

The author declares no competing interest.

[1]Email: horenkoi@usi.ch.

This article contains supporting information online at https://www.pnas.org/lookup/suppl/doi:10.1073/pnas.2119659119/-/DCSupplemental.

Published February 23, 2022.

with a multiindex $t = (i, j)$. According to the above *Theorem*, this heuristics—building a computational foundation of tSNE—is actually equivalent to the optimal nonparametric density estimate [**3**], in the sense that it is simultaneously minimizing the expectation of the pairwise squared Euclidean distances between the data points (when considering loss function $g(x_t, \theta) = \|\theta - x_t\|^2$ with $t \equiv j$ and $\theta \equiv x_i$ in [**2**]) and simultaneously maximizing the entropy of w (i.e., providing the least-biased estimation) and is obtained with an explicitly computable closed-form expression. Furthermore, solution [**3**] also provides a recipe for computing such t-SNE density estimates in the cases with non-Euclidean loss functions g.

It is straightforward to verify that the simultaneous learning of the parameters θ and probability densities w can be performed with the monotonically convergent entropic outlier sparsification (EOS) algorithm (see *Algorithm 1*).

Eq. **4** establishes a relation between the Gaussian variance parameter σ^2 and the entropic sparsification parameter α in [**3**], indicating a possibility of inferring the optimal sparsification parameter α^* for the given data. For example, optimal σ^2 in [**4**]—and hence the optimal sparsification parameter value α^*—can be obtained by maximizing the log-likelihood of the distribution w_t^α with respect to the parameter α; that is, $\alpha^* = \arg\max_{\alpha>0} \sum_t \log(w_t^\alpha)$. In the practical examples of EOS below, we will follow a simpler grid search approach for selecting the optimal sparsification parameter α^*—deploying the same multiple cross-validation procedure that is commonly used for determining metaparameter values in AI and ML. On a predefined grid of α values, we will select those values that show the best overall model performance on the validation data that were not used in the model training.

Fig. 1 summarizes numerical experiments comparing EOS to common data anomaly detection and learning tools on

Algorithm 1 Entropic outlier sparsification algorithm for the solution of optimization problem [**2**]

For a given $\{x_1, x_2, \ldots, x_T\}$, and $\alpha > 0$, randomly choose initial $w^{(1)}$
$I = 1;\ L^{(I)} = \infty;\ \Delta L^{(I)} = \infty\ I = 1;\ L^{(I)} = \infty;\ \Delta L^{(I)} = \infty$
while $\Delta\mathcal{L}^{(I)} > tol$ **do**
  $\theta^{(I)} \leftarrow$ solution of [**2**] for fixed $w^{(I)}$
  $w^{(I+1)} \leftarrow$ evaluating [**3**] for fixed $\theta^{(I)}$
  $L^{(I+1)} \leftarrow L(w^{(I+1)}, \theta^{(I)}, \alpha)$
  $I \leftarrow I + 1$
  $\Delta L^{(I)} \leftarrow L^{(I-1)} - L^{(I)}$.

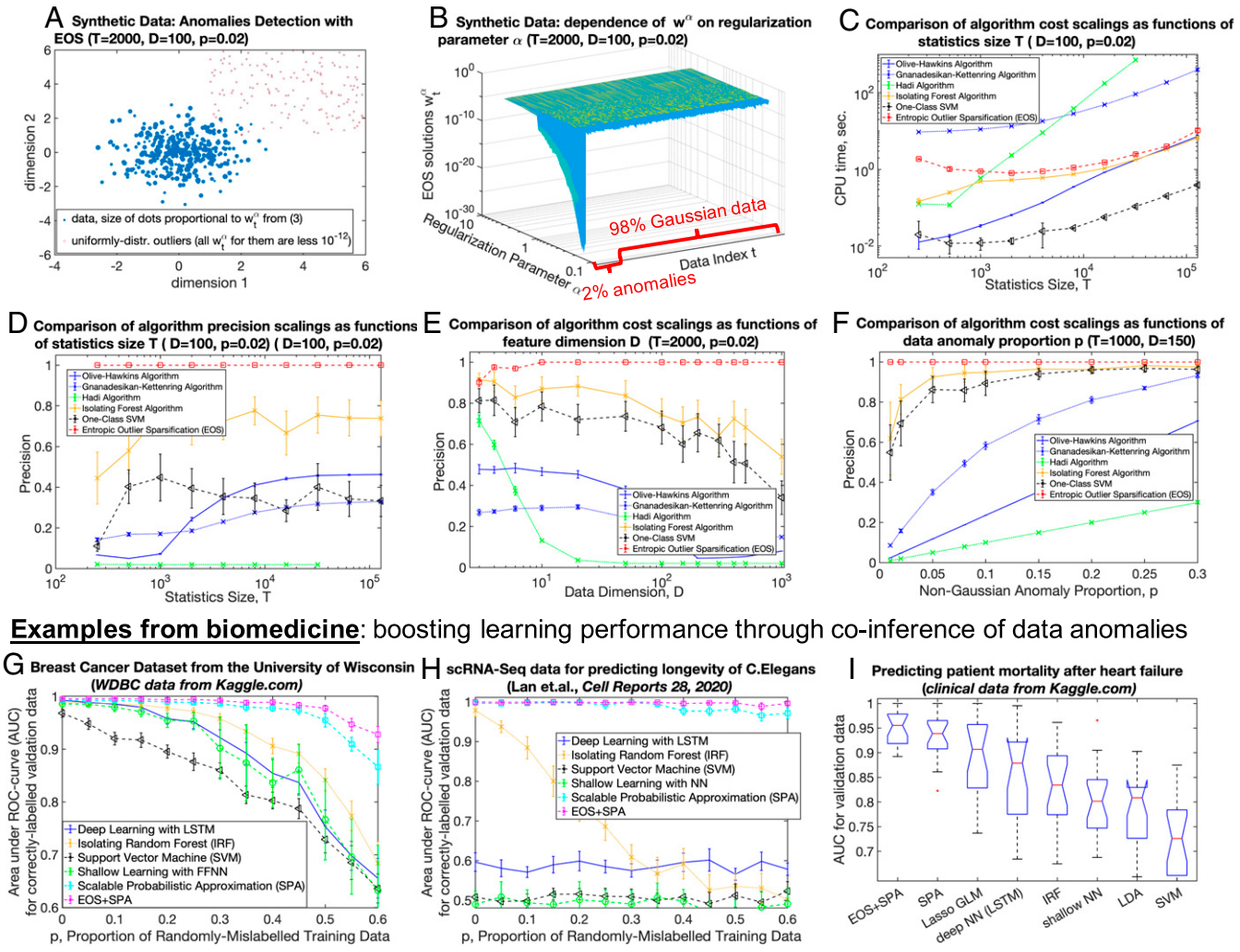

Fig. 1. Comparison of EOS algorithm for the solution of optimization problem [**2**] to common methods of data anomaly detection (*A*–*F*) and supervised classifier learning (*G*–*I*) on synthetic and real data examples from refs. 12–14.

randomly generated synthetic datasets (representing multivariate normal distributions with asymmetrically positioned uniformly distributed outliers; Fig. 1 *A*–*F*) and three biomedical datasets with various proportions of randomly mislabeled data instances in the training sets (Fig. 1 *G*–*I*). All of the compared algorithms are provided with the same information and run with the same hardware and software; 50 cross-validations were performed in every experiment to visualize the obtained 95% CIs. In numerical experiments with synthetic data (Fig. 1 *A*–*F*), the EOS algorithm is deployed, with g being the negative point-wise multivariate Gaussian loglikelihood, that is, with $g(x_t, \mu, \Sigma) = 1/D\left(0.5\log\det(\Sigma) + 0.5(x_t-\mu)^{\dagger}\Sigma^{-1}(x_t-\mu)\right)$, where μ and Σ are Gaussian mean and covariance, respectively. Iterative estimation of weighted mean and covariance in the θ-step of the EOS algorithm is performed using analytical estimates of the weighted Gaussian covariance and mean, and convergence tolerance *tol* is set to 10^{-12}. Total computational costs and statistical precisions—the latter are measured as the numbers of correctly identified points not belonging to the Gaussian distribution divided by the total number of identified outliers—are performed for various problem dimensions, statistic sizes, and outlier proportions. EOS was compared to all of the outlier detection methods available in the "Statistics" and "Machine Learning" toolboxes of MathWorks. Precision is chosen as the measure of performance here since it is more robust than the other common measures when the datasets are not balanced, for example, when the number of instances in one class (outliers) is much less than in the other class (nonoutliers). These results show that EOS allows a marked and robust improvement of outlier detection precision for all of the considered comparison cases. Data and MATLAB code are provided at https://github.com/horenkoi/EOS.

Next, real labeled datasets from biomedicine are considered, including two popular datasets—the University of Wisconsin Database for Breast Cancer diagnostics data (12) (Fig. 1 *G*) and the clinical dataset for predicting mortality after heart failure (13) (Fig. 1 *I*)—as well as a single-cell messenger RNA gene expression dataset from longevity research (14) (Fig. 1 *H*). The main focus here is on comparing the robustness of learning methods to outliers and mislabeled data instances in the training set, for common binary classifiers and for EOS that is equipped with loss function g from the scalable probabilistic approximation (SPA) classifier algorithm (15, 16). SPA is selected since it shows the highest robustness to mislabeling for all of the considered datasets (Fig. 1 *G* and *I*). As can be seen from Fig. 1 *G* and *H*, EOS with $g(x_t, \theta)$ from SPA (EOS+SPA, red dashed lines), allows a statistically significant improvement of prediction performance (measured with the common performance measure area under curve [AUC]) for all of the tested mislabeling proportions p for all of the considered biomedical examples. As was shown recently, coinference of data mislabelings can significantly improve predictive performance of supervised classifiers (17). Application of the EOS algorithm with model loss function $g(x_t, \theta)$ from SPA (EOS+SPA) allows achieving AUC of 0.96 and accuracy of $97\% \pm 2\%$ (*SI Appendix*, Fig. S1) when predicting patients' mortality after heart failure from clinical patients' data, statistically significantly outperforming common learning tools that do not deploy outlier coinference (Fig. 1*I*).

EOS and entropic sparsification Eq. **3** can be also applied to other types of leaning problems, for example, to feature selection and novelty detection problems.

Data Availability. Data and MATLAB code have been deposited in GitHub (https://github.com/horenkoi/EOS). Previously published data were used for this work (12–14).

ACKNOWLEDGMENTS. I.H. acknowledges funding from the Carl-Zeiss Foundation initiative "Emergent Algorithmic Intelligence."

1. D. L. Donoho, M. Gasko, Breakdown properties of location estimates based on halfspace depth and projected outlyingness. *Ann. Stat.* **20**, 1803–1827 (1992).
2. D. M. Rocke, D. L. Woodruff, Identification of outliers in multivariate data. *J. Am. Stat. Assoc.* **91**, 1047–1061 (1996).
3. P. Filzmoser, R. Maronna, M. Werner, Outlier identification in high dimensions. *Comput. Stat. Data Anal.* **52**, 1694–1711 (2008).
4. H. Wang, M. J. Bah, M. Hammad, Progress in outlier detection techniques: A survey. *IEEE Access* **7**, 107964–108000 (2019).
5. W. A. Stahel, "Robust estimation: Infinitesimal optimality and covariance matrix estimators," PhD thesis, Eidgenossische Technische Hochschule (ETH) Zurich, Zurich, Switzerland (1981).
6. P. J. Rousseeuw, B. C. Van Zomeren, Unmasking multivariate outliers and leverage points. *J. Am. Stat. Assoc.* **85**, 633–639 (1990).
7. R. A. Maronna, V. J. Yohai, The behavior of the Stahel-Donoho robust multivariate estimator. *J. Am. Stat. Assoc.* **90**, 330–341 (1995).
8. Y. Zuo, H. Cui, X. He, On the Stahel-Donoho estimator and depth-weighted means of multivariate data. *Ann. Stat.* **32**, 167–188 (2004).
9. D. L. Donoho, For most large underdetermined systems of equations, the minimal l1-norm near-solution approximates the sparsest near-solution. *Commun. Pure Appl. Math.* **59**, 907–934 (2006).
10. S. Huang, T. D. Tran, Sparse signal recovery via generalized entropy functions minimization. *IEEE Trans. Signal Process.* **67**, 1322–1337 (2019).
11. L. van der Maaten, G. Hinton, Visualizing data using t-SNE. *J. Mach. Learn. Res.* **9**, 2579–2605 (2008).
12. UCI Machine Learning, Data from "Breast cancer Wisconsin (diagnostic) data set." Kaggle. https://www.kaggle.com/uciml/breast-cancer-wisconsin-data. Accessed 1 October 2021.
13. D. Chicco, G. Jurman, Machine learning can predict survival of patients with heart failure from serum creatinine and ejection fraction alone. *BMC Med. Inform. Decis. Mak.* **20**, 16 (2020).
14. J. Lan *et al.*, Translational regulation of non-autonomous mitochondrial stress response promotes longevity. *Cell Rep.* **28**, 1050–1062.e6 (2019).
15. S. Gerber, L. Pospisil, M. Navandar, I. Horenko, Low-cost scalable discretization, prediction, and feature selection for complex systems. *Sci. Adv.* **6**, eaaw0961 (2020).
16. I. Horenko, On a scalable entropic breaching of the overfitting barrier for small data problems in machine learning. *Neural Comput.* **32**, 1563–1579 (2020).
17. S. Gerber *et al.*, Co-inference of data mislabelings reveals improved models in genomics and breast cancer diagnostics. *Front. Artif. Intell.* **4**, 739432 (2022).