Orthogonal Approximation of Marginal Likelihood of Generative Models

Václav Šmídl
Institute of Information Theory and Automation
Czech Academy of Science, Prague
smidl@utia.cas.cz

Jan Bím
Dept. of Computer Science
Czech Technical University
bim.jan@fel.cvut.cz

Tomáš Pevný*
Dept. of Computer Science
Czech Technical University
pevnytom@fel.cvut.cz

Abstract

This paper presents a new approximation of the marginal likelihood of generative models which is used as a score for anomaly detection. The score is motivated by the shortcoming of the popular reconstruction error that it can behave arbitrarily outside the known samples. The proposed score corrects this by orthogonal combination of the reconstruction error and the likelihood in the latent space. As experimentally shown on benchmark problems from anomaly detection and illustrated on a toy problem, this combination lends the score robustness to outliers. Generative models evaluated with this score outperformed the competing methods especially in tasks of learning distribution from data corrupted by anomalies. Finally, the score is compatible with contemporary generative models, namely variational auto-encoders and generative adversarial networks.

1 Motivation

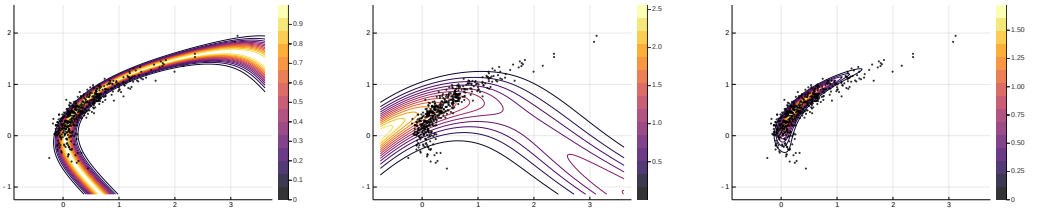
Generative models have made tremendous progress thanks to neural networks in past years with primary focus on generation of realistic samples. However, since some of them, such as variational autoencoder (VAE) [11], are based on the transformation of variables formula, the resulting model represents an estimate of probability distribution of training samples. Thus, the model can be used to evaluate probability of a new sample to be generated from the model. However, exact evaluation requires to integrate over the latent variable which is problematic. While methods for exact marginalization are available (HMC was proposed even in the first publication proposing VAE [11]), they are too expansive to run routinely for evaluation of large number of samples as is common in anomaly detection. Specifically in anomaly detection, it is not important to evaluate the exact likelihood but only to establish an relative probability between any two data points. Thus only a *score* function providing an order of the data is typically required. The score function can be transformed in any way that preserves order of the data, since the threshold for anomaly classification is determined in the subsequent step. Thus, both likelihood and log-likelihood function can be used as scores. Therefore, we will often use term "score" as shorter term for "approximation of the marginal likelihood".

The most popular score for generative models is the reconstruction error, which is used for the autoencodes [1, 23, 22] (although some works utilized it with restricted Boltzmann machines [7]).

*Tomáš Pevný is also with Avast Software s.r.o.

4th workshop on Bayesian Deep Learning (NeurIPS 2019), Vancouver, Canada.

This score corresponds to log-likelihood of the observed data with Dirac approximation of the prior on the latent variable. An alternative to the reconstruction error score is a score based on probability of the data projected on the latent variable with respect to the prior distribution [25]. We conjecture that such approximation is too coarse and its impact on accuracy is so severe that vanilla k-nearest neighbor with basic L_2 distance is frequently superior to approaches based on Variational Auto-Encoders in [18]. This problem is demonstrated in Figure 1a, where the reconstruction error of the autoencoder network defines the manifold well but also assigns high probability in areas where no data have been observed. The distribution in the latent space (image of the encoding function), assigns correct distribution on the manifold, but due to the many-to-one relation of the encoder, the areas outside the manifold received probability that is too high Figure 1b .



(a) Reconstruction error (b) Likelihood in latent space $p(z)$ (c) The proposed score

Figure 1: Anomaly scores of three different approaches for a toy problem, $x = [z^2, z]^\top + e$, where $p(z) = \mathcal{N}(0.5, 0.15)$ and $p(e) = \mathcal{N}([0, 0]^\top, 0.01\mathbf{I})$. (a) reconstruction error $p_{RE}(x) = \exp\left(-\frac{1}{2}||x - f(g(x))||_2^2/\sigma_{RE}\right), \sigma_{RE} = \text{var}(x - f(g(x)))$, (b) $p(z)|_{z=g(x)}$, and (c) the proposed score (4).

As a remedy to above problems, we propose an orthogonal approximation of the marginal likelihood that: (i) is theoretically justified, (ii) combines advantages of reconstruction error and probability of the projection to the latent space, and (iii) it has reasonable computational complexity. This work therefore does not propose a new model/architecture for detecting anomalies, but it proposes *a new score* compatible with existing VAE and GAN.

After acceptance of this work, we have found that the same score has been proposed independently to us in [16]. Since the score is the same, we focus on studying properties of the score with different generative models (such as vanilla Variational Autoencoder) on a very different suite of problems [15], and comparing the proposed score to legacy anomaly detectors, namely Isolation Forests [12] and k-Nearest Neighbors [8]. Although these legacy detectors lacks the fancy "deep" sticker, they perform very well on problems with hand-designed features, as has been shown in [18]. Thus, this study nicely complements the work of [16], showing the strength (robustness to outliers in the training set) and the weakness of the method (sensitivity to the quality of fit of the latent space).

2 Marginal Likelihood of Generative Models

Consider generative distribution of d-dimensional data samples x defined by

$$p(x) = \int p_\theta(x|z)p(z)dz \qquad\qquad p_\theta(x|z) = \mathcal{N}(x; f_\theta(z), \sigma^2\mathbf{I}), \tag{1}$$

and a chosen prior on the latent k-dimensional variable z which is either fixed $p(z) = \mathcal{N}(0, I)$, or with additional parameters $p_\theta(z)$ [20]. Distribution $p_\theta(x|z)$ is known as the *decoder* with $f_\theta(z)$ being a neural network parametrizing the mean of the Normal distribution. The aim is to estimate all parameters from the set of observations, $\{x^{(i)}\}_{i=1}^n$. Various model introduce auxiliary objects, for example VAE introduces *encoder*, which is a conditional probability distribution $q_\phi(z|x)$ parametrized similarly to the decoder as $q_\phi(z|x) = \mathcal{N}(z; g_\phi(x), \text{diag}(\sigma(x)))$. The aim is to estimate parameters (θ, ϕ) from the data.

The estimated parameters uniquely define marginal likelihood in (1), but since it is a complex integral, it is for VAE often approximated by the reconstructions error

$$p(x) \approx \int p_\theta(x|z)\delta(z - g_\phi(x))dz \propto \exp\left(-\|x - f(g(x))\|^2\right), \tag{2}$$

where $\propto$ denotes equality up to a multiplicative constant. Note that in this score, the integration over z is replaced by an evaluation of the likelihood at the "most probable" point given by the encoder $q_\phi(z|x)$. A different approach is used with the flow-based models, such as [14], where the the probability is evaluated by the change of variables formula (see Appendix B for computational details):

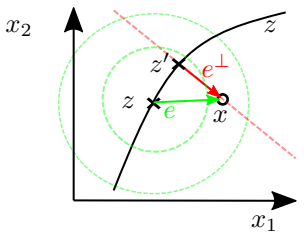
$$p(x) = p_z(f^{-1}(x)) \left| \frac{\partial f^{-1}(x)}{\partial x} \right|. \tag{3}$$

However, there are some unexplained phenomena reported in [13] when this formula is used. Moreover, auto-regressive flows can be used only when the dimension of $\mathcal{Z}$ is equal to that of $\mathcal{X}$, i.e. $d = k$.

We propose an alternative to generative model with isotropic noise ($\mathbf{M}_1$), which uses orthogonal decomposition of the noise ($\mathbf{M}_2$):

$$\mathbf{M_1} : x = f(z) + e, \qquad \mathbf{M_2} : x = f(z') + e^\perp,$$

where z' is a point in the latent space, and $e^\perp$ is the observation noise which lies in the normal space of the manifold defined by $f_\theta(z)$. This idea is visualized in the figure right, where the conventional noise is denoted by green and the proposed orthogonal model by red color.



Denoting $x' = f(z')$ we can decompose any point x into

$$x = x' + e^\perp,$$

where x' lies on a k-dimensional manifold, and $e^\perp$ in its normal space (of $d - k$ dimensions). Due to the (local) orthogonality, we consider x' and $e^\perp$ to be independent and the probability distribution $p(x)$ to be well approximated $p(x')p(e^\perp)$. The probability of $p(x')$ is given by the change of coordinate formula from $p_z(z)$, and the $p(e^\perp) \approx p(e)$. This reasoning yields the proposed approximation of the marginal likelihood:

$$p(x) \approx p(x')p(e) = p_z(f^{-1}(x')) \left| \frac{\partial f^{-1}(x')}{\partial x} \right| \mathcal{N}(x - x', \sigma^2 \mathbf{I}). \tag{4}$$

Note that the assignment $p(e^\perp) = p(e)$ is correct up to a normalizing constant if the z' point is correctly estimated. This can be achieved e.g. by optimization, see Appendix A.

Notice that the proposed score uses reconstruction error (2) popular in the prior art together with the exact likelihood used in (3). It should, therefore, benefit from both scores and prevent pathological failures, as has been demonstrated on the motivational example in Figure 1. Moreover, unlike models based on auto-regressive flows, the proposed score is not restricted to cases when the input and the latent spaces have the same dimension.

3 Experiments: Anomaly Detection

The established definition of an anomaly is *an event occurring with a probability so low, that it raises suspicion of being generated by some other probability distribution*. This implies that a sample x is an anomaly if the probability density function $p(x)$ is very low.

The experimental comparison of the proposed score to the reconstruction error has been performed on nine problems adapted for use in anomaly detection by [6, 15] . These problems have been also used in study [18] comparing sophisticated methods based on neural networks to k-nearest neighbor [8] (kNN) and isolation forests (IF) methods [12] and it was found that kNN and IF dominated VAE and GAN. The below study uses only *easy* anomalies, as more difficult anomalous samples are located in areas of high densities of normal data, which raises doubts if they should be considered anomalous [6]. For each dataset, five distinct train/test splits were created with 80% of the dataset being used for training and the remaining 20% for evaluation. The training set was either clean without any outliers or contaminated with up to 10% anomalies. If there was less then 10% of anomalies, all available anomalous samples were added to the training dataset. Importantly, these anomalies are not labeled and the generative model is expected to learn the density describing legitimate as well as anomalous data. It is expected that the threshold for detection of the anomalies has to be set higher than in the case of clean data.

	no contamination				10% contamination			
	VAE				VAE			
Dataset	RE	Orth	kNN	IF	RE	Orth	kNN	IF
breast-cancer-wisconsin	0.87	0.95	**0.98**	0.94	0.77	**0.93**	0.86	0.86
cardiotocography	0.61	0.49	0.62	**0.64**	0.70	**0.82**	0.52	0.63
magic-telescope	0.76	0.91	0.86	**0.92**	0.85	**0.91**	0.9	0.81
pendigits	0.73	**0.97**	0.90	0.96	0.57	**0.69**	0.52	0.58
pima-indians	0.85	0.84	0.86	**0.88**	0.85	**0.92**	0.88	0.87
wall-following-robot	0.65	0.70	**0.72**	0.66	0.50	**0.59**	0.44	0.52
waveform-1	0.79	0.68	0.79	**0.82**	0.49	**0.70**	0.47	0.54
waveform-2	0.81	0.71	0.81	**0.83**	0.51	**0.72**	0.48	0.53
yeast	0.63	**0.81**	0.68	0.75	0.69	**0.70**	0.63	0.66

Table 1: Area under ROC curve of detectors based on variational autoencoder (VAE) with the proposed score based on orthogonal decomposition (Orth) and with the usual reconstruction error (RE), and also that of k-nearest neighbor (kNN) and Isolation forests (IF). The left / right part of the table show AUCs when the training set does not contain any outliers / is poluted with 10% outliers (or less if 10% is not present in the data set).

For each dataset and split of the data, a large number of models (280 to be exact) were trained differing by hidden layer dimensions $\in \{32, 64\}$, latent layer dimension (if smaller or equal than the data dimension) $\in \{2, 4, 9\}$. Both encoder and decoder contained three fully connected hidden layers of neurons with the "swish" activation function [17]. In order to represent the data space well, the decoder contained an extra output layer of neurons with linear activation function. All models were optimized using the ADAM optimizer [10] with default setting and with batch size 100 for 10000 steps. Finally, all experiments are implemented in the Julia programming language [3] with Flux.jl [9]. The code for the experiments is available at `https://github.com/anomaly-scores/Anomaly-scores`. Variance of the noise σ is treated as a hyper-parameter $\sigma^2 \in \{0.01, 0.1, 1, 10\}$.

The quality of detection is measured using the area under the ROC curve (AUC), which is considered as a standard in the field of anomaly detection. In total, 280 models differing by hyper-parameters were trained for each problem and data-split, the best combination of hyper-parameters was selected according to AUC on the training set, which simulates the scenario where some examples of anomalies are available for model selection.[2]

The results are summarized in Table 1. VAE with the proposed orthogonal score (denoted Orth) dominates other methods when training set is contaminated with outliers. On the contrary, for problems where the training data are clean, the reconstruction error is often better than the proposed score, and the prior art, k-NN and isolation forest, is dominating both of them. We suspect the poor detection on clean dataset is due to the mismatch between prior and posterior distributions on the latent space.

3.1 Robustness to outliers

To shed light on the robustness, we have performed a synthetic experiment on the toy problem from the motivation section, where the data were generated as $x = [z^2, z]^\top + e$, where $p(z) = \mathcal{N}(0.5, 0.15)$ and $p(e) = \mathcal{N}([0, 0]^\top, 0.01\mathbf{I})$. Furthermore, the training data were contaminated with 4 outliers sampled uniformly on the displayed grid. For this problem we have trained a vanilla variation autoencoder with 2 dense layers with a Normal distribution $\mathcal{N}(0, \mathbf{I})$ on the latent space.

Figure 2 shows the true and reconstructed data, samples from prior and projection of the data to the latent space, exact marginal likelihood, likelihood computed using the reconstruction error, likelihood using the transformation of variables, and the orthogonal score. We can see how the reconstruction error is distorted by the outliers, as the auto-encoder thrives to achieve perfect reconstruction. The posterior in the latent space is relatively well matched, with an occasional numerical outlier at the bottom. The proposed score is a product of the previous scores, and assigns high probability only to areas where both of these scores are high. We conjecture that the reason why the proposed orthogonal

[2]This scenario might not be as unrealistic as it sounds, as one typically has examples of a few anomalies.

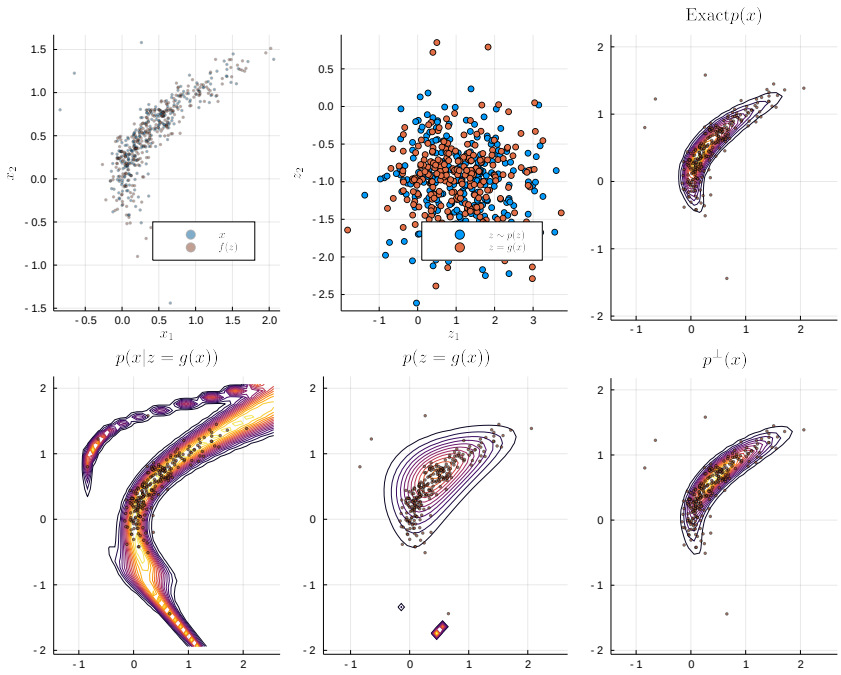


Figure 2: A toy example with data sampled from $x = [z^2, z]^\top + e$, $p(z) = \mathcal{N}(0.5, 0.15)$ and $p(e) = \mathcal{N}([0, 0]^\top, 0.01\mathbf{I})$, polluted by outliers. Top row in the left to right order shows: i) the original and reconstructed data, ii) samples from the prior and projection of the data via the encoder, iii) exact marginal likelihood computed numerically. Bottom row in the left to right order shows: i) the reconstruction error, ii) likelihood using the change of coordinates formula, and iii) the orthogonal score.

	SVAE		VAE	
Dataset	RE	Orth	RE	Orth
breast cancer	0.70	0.92	0.77	**0.93**
cardiotocography	0.59	**0.88**	0.70	0.82
magic-telescope	0.86	**0.92**	0.85	0.91
pendigits	0.64	**0.69**	0.57	**0.69**
pima-indians	0.82	**0.94**	0.85	0.92
wall-following-robot	0.50	**0.62**	0.50	0.59
waveform-1	0.52	**0.82**	0.49	0.70
waveform-2	0.50	0.70	0.51	**0.72**
yeast	0.75	**0.83**	0.69	0.70

Table 2: Area under ROC curve of detectors based on spherical variational autoencoder (SVAE) and regular VAE with the proposed orthogonal score (Orth) and with the usual reconstruction error (RE).

likelihood is robust is that each of the previous scores is sensitive to different artifacts and their combination allows to suppress them.

The same toy example is used to illustrate the problem with mismatch between the prior and the encoded data in Figure 3. Note that the reconstruction error is not affected by the mismatch and identifies the latent space of the data well. However, the change of variables score is too concentrated on a small part of the data yielding underestimated marginal likelihood.

3.2 Richer prior model using VAMP prior

Below we study an impact of a richer family of prior distributions p_z on the latent layer in VAE. Specifically, we have used Spherical VAE (SVAE) with the VAMP prior [20] instead of a single-

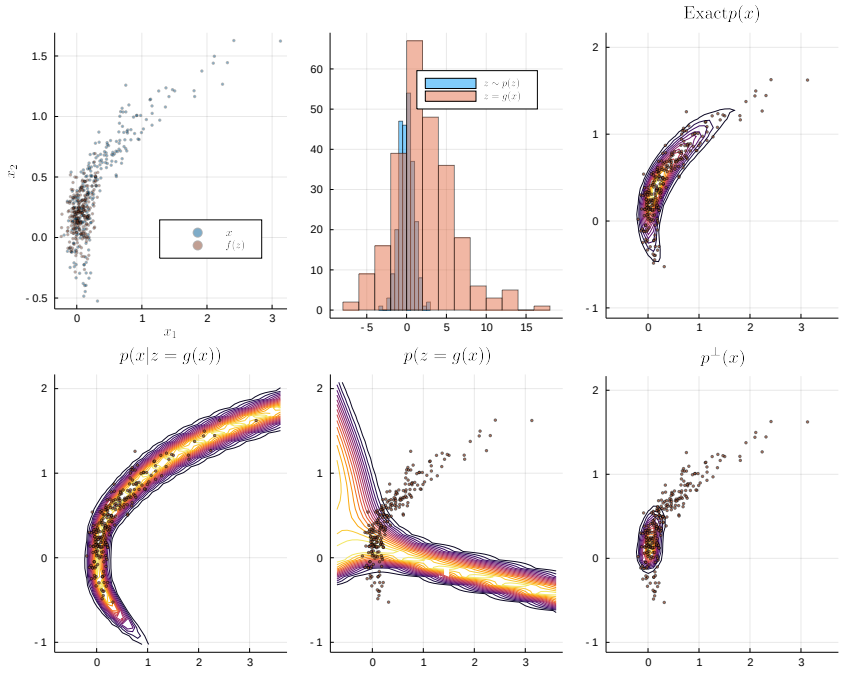


Figure 3: A toy example with data sampled from $x = [z^2, z]^\top + e$, $p(z) = \mathcal{N}(0.5, 0.15)$ and $p(e) = \mathcal{N}([0, 0]^\top, 0.01\mathbf{I})$, polluted by outliers. VAE with 1d latent variable. Top row in the left to right order shows: i) the original and reconstructed data, ii) samples from the prior and projection of the data via the encoder, iii) exact marginal likelihood computed numerically. Bottom row in the left to right order shows: i) the reconstruction error likelihood, ii) likelihood using the change of coordinates formula, and iii) the orthogonal score.

component Normal distribution. Spherical VAE has latent space restricted to unit sphere, hence the prior distribution in VAMP setting is a mixture of von Mises-Fisher distributions [5]. SVAE was previously shown to outperform the conventional choice in [2]. The SVAE is trained using Wasserstein divergence instead the usual KL divergence. The closeness of distributions $g(x)$ where $x \sim p(x)$ and $z \sim p(z)$ is measured using the Maximum Mean Discrepancy (MMD) with IMQ kernel [19], where its width c is treated as a hyper-parameter $c \in \{0.001, 0.01, 0.1, 1\}$. The number of components in the prior mixture was treated as a hyper-parameter with values $\in \{1, 4, 16\}$. Finally, the trade-off between enforcing reconstruction error and closeness of the distributions represented by $\beta \in \{0.01, 0.1, 1, 10\}$ which is also treated as a hyper-parameter.

Table 2 shows AUCs of anomaly detectors based on spherical VAE with a VAMP prior (denoted as SVAE) and with single Normal distribution (denoted as VAE). While the reconstruction error does not benefit from the richer prior, as can be expected since it is not part of the score, the proposed Orthogonal score is significantly better, as has been suggested in the previous section.

4 Conclusion

This work has proposed a new approximation of marginal likelihood of generative models combining reconstruction error and probability in the latent space. The approximation, which we prefer to call score rather than an approximate likelihood, is cheaper to compute than scores based on Monte Carlo marginalization. The experimental comparison to state of the art demonstrated that the new score is very robust to contamination of the training set by anomalies. In that case, generative models with the proposed score decisively outperform the conventional reconstruction error score in anomaly detection, as well as the k-NN and isolation forest methods. From an investigation on a toy problem, we suspect that the source of robustness is the combination of reconstruction error and likelihood in the latent space.

The experimental comparison also revealed that k-NN is frequently better when the training data are clean, i.e. they do not contain any outliers. We believe that this behavior is due to the mismatch between prior and posterior distributions in the latent space. In the future work, we would like to validate this suspicion using more sophisticated priors. Another direction of research would be to estimate, how the proposed approximation of the marginal likelihood differs from the true likelihood.

5 Acknowledgments

Research presented in this work has been supported by the Grant agency of Czech Republic no. 18-21409S. The authors also acknowledge the support of the OP VVV MEYS funded project CZ.02.1.01/0.0/0.0/16_019/0000765 "Research Center for Informatics".

References

[1] Jinwon An and Sungzoon Cho. Variational autoencoder based anomaly detection using reconstruction probability. *Special Lecture on IE*, 2:1–18, 2015. 1

[2] Caglar Aytekin, Xingyang Ni, Francesco Cricri, and Emre Aksu. Clustering and Unsupervised Anomaly Detection with l2 Normalized Deep Auto-Encoder Representations. In *Proceedings of the International Joint Conference on Neural Networks*, volume 2018-July, 2018. 3.2

[3] Jeff Bezanson, Alan Edelman, Stefan Karpinski, and Viral B. Shah. Julia: A fresh approach to numerical computing. `julialang.org/publications/julia-fresh-approach-BEKS.pdf`, 2017. 3

[4] Bin Dai and David Wipf. Diagnosing and enhancing vae models. *arXiv preprint arXiv:1903.05789*, 2019. A

[5] Tim R. Davidson, Luca Falorsi, Nicola De Cao, Thomas Kipf, and Jakub M. Tomczak. Hyperspherical variational auto-encoders. *arXiv:1804.00891 [cs, stat]*, 2018. 3.2

[6] Andrew F Emmott, Shubhomoy Das, Thomas Dietterich, Alan Fern, and Weng-Keen Wong. Systematic construction of anomaly detection benchmarks from real data. In *Proceedings of the ACM SIGKDD workshop on outlier detection and description*, pages 16–21. ACM, 2013. 3

[7] Ugo Fiore, Francesco Palmieri, Aniello Castiglione, and Alfredo De Santis. Network anomaly detection with the restricted boltzmann machine. *Neurocomputing*, 122:13–23, 2013. 1

[8] Stefan Harmeling, Guido Dornhege, David Tax, Frank Meinecke, and Klaus-Robert Müller. From outliers to prototypes: ordering data. *Neurocomputing*, 69(13-15):1608–1618, 2006. 1, 3

[9] Mike Innes. Flux: Elegant machine learning with julia. *Journal of Open Source Software*, 2018. 3

[10] Diederik P Kingma and Jimmy Ba. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*, 2014. 3

[11] Diederik P. Kingma and Max Welling. Auto-encoding variational bayes. *arXiv:1312.6114 [cs, stat]*, 2013. 1

[12] Fei Tony Liu, Kai Ming Ting, and Zhi-Hua Zhou. Isolation forest. In *2008 Eighth IEEE International Conference on Data Mining*, pages 413–422. IEEE, 2008. 1, 3

[13] Eric Nalisnick, Akihiro Matsukawa, Yee Whye Teh, Dilan Gorur, and Balaji Lakshminarayanan. Do deep generative models know what they don't know? *arXiv preprint arXiv:1810.09136*, 2018. 2, C

[14] George Papamakarios, Theo Pavlakou, and Iain Murray. Masked autoregressive flow for density estimation. In *Advances in Neural Information Processing Systems*, pages 2338–2347, 2017. 2

[15] Tomáš Pevný. Loda: Lightweight on-line detector of anomalies. *Machine Learning*, 102(2):275–304, 2016. 1, 3

[16] Stanislav Pidhorskyi, Ranya Almohsen, and Gianfranco Doretto. Generative probabilistic novelty detection with adversarial autoencoders. In S. Bengio, H. Wallach, H. Larochelle, K. Grauman, N. Cesa-Bianchi, and R. Garnett, editors, *Advances in Neural Information Processing Systems 31*, pages 6822–6833. Curran Associates, Inc., 2018. 1

[17] Prajit Ramachandran, Barret Zoph, and Quoc V. Le. Searching for activation functions. *CoRR*, abs/1710.05941, 2017. 3

[18] Vít Škvára, Tomáš Pevný, and Václav Šmídl. Are generative deep models for novelty detection truly better? *arXiv preprint arXiv:1807.05027*, 2018. 1, 1, 3

[19] Ilya Tolstikhin, Olivier Bousquet, Sylvain Gelly, and Bernhard Schoelkopf. Wasserstein auto-encoders. *arXiv preprint arXiv:1711.01558*, 2017. 3.2

[20] Jakub M Tomczak and Max Welling. Vae with a vampprior. *arXiv preprint arXiv:1705.07120*, 2017. 2, 3.2

[21] Dmitry Ulyanov, Andrea Vedaldi, and Victor Lempitsky. It takes (only) two: Adversarial generator-encoder networks. In *Thirty-Second AAAI Conference on Artificial Intelligence*, 2018. A

[22] Dan Xu, Elisa Ricci, Yan Yan, Jingkuan Song, and Nicu Sebe. Learning deep representations of appearance and motion for anomalous event detection. *arXiv preprint arXiv:1510.01553*, 2015. 1

[23] Chong Zhou and Randy C Paffenroth. Anomaly detection with robust deep autoencoders. In *Proceedings of the 23rd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pages 665–674. ACM, 2017. 1

[24] Jun-Yan Zhu, Taesung Park, Phillip Isola, and Alexei A Efros. Unpaired image-to-image translation using cycle-consistent adversarial networks. In *Proceedings of the IEEE international conference on computer vision*, pages 2223–2232, 2017. A

[25] 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, 2018. 1

A Analysis of the noise model

Let's now briefly discuss relation of the isotropic model $\mathbf{M}_1$ and the orthogonal model $\mathbf{M}_2$. The assumptions of the latter might seem to be very restrictive at the first sight, as it might not model the reality accurately. Below it is argued that since in practice one almost always fails to identify the true f, the assumption does not decrease the expressiveness of the model, but it changes the structure of the noise. Note that the conventional isotropic residue e can be decomposed into part $e^{\|}$ lying in the space tangential to $f(z)$ and orthogonal: $e = e^{\|} + e^{\perp}$. Due to the bijection of f there exists z' such that $f(z') = f(z) + e^{\|}$. Consequently, x can be expressed as $x = f(z') + e^{\perp}$, which conforms with model $\mathbf{M}_2$. It effectively means that $e^{\|}$ is *absorbed* by the distribution on the latent space $p(z)$ and by the learned decoder f.

Recent analysis of [4] suggests that standard VAE with reduced dimension of the latent variable is learning well the manifold but no the distribution on it. This indicates that may be learning the proposed model $\mathbf{M}_2$ rather than the assumed model $\mathbf{M}_1$.

If we assume that the estimation procedure is learning the standard isotropic model, we can evaluate local correction of $p(z')$. From the above definitions it holds that $\epsilon^{\|} = f(z') - f(z)$. Using Taylor expansion at point $f(z')$ to approximate $f(z)$,

$$\epsilon^{\|} = f(z') - f(z) \approx \left.\frac{\partial f(z)}{\partial z}\right|_{z'} (z' - z),$$

where $J(z') = \left.\frac{\partial f(z)}{\partial z}\right|_{z'}$ denotes Jacobian of f at point z' and the error term has been omitted. Using this approximation, z' can be expressed as

$$z' = z + J(z')^{-1}\epsilon^{\|}.$$

Now recall that $z \sim \mathcal{N}(0, \mathbf{I})$ and $\epsilon^{\|} \sim \mathcal{N}(0, \sigma^2)$, which means that $z' \sim \mathcal{N}(0, \mathbf{I} + \sigma^2 J(z')J(z')^{\mathrm{T}})$. This correction removes the discrepancy between model assumed during training and during evaluation and the *corrected* score equals to

$$p(x) \approx \mathcal{N}\left(z'|0, \mathbf{I} + \sigma^2 J(z')J(z')^{\mathrm{T}}\right) \left|\frac{\partial f^{-1}(z')}{\partial z'}\right| \mathcal{N}(x - f(z'), \sigma^2\mathbf{I}). \tag{5}$$

The above model assumes $J(z')$ to be constant, which is not technically correct. Alternatively, in the evaluation of auto-encoders, we know both z and z', the Taylor expansion can be made around z instead in z', which changes the criterion to

$$p(x) \approx \mathcal{N}\left(z'|0, \mathbf{I} + \sigma^2 J(z) J(z)^{\mathrm{T}}\right) \left|\frac{\partial f^{-1}(z')}{\partial z'}\right| \mathcal{N}(x - f(z'), \sigma^2 \mathbf{I}). \tag{6}$$

The formulation of the likelihood (6) can be used with generative models identified by GANs. The point z can be found by solving $\arg\min_z \|f(z) - x\|^2$, however in practice one would probably use variants with encoders satisfying cyclic properties [24, 21], as they might be faster to solve the optimization problem and also more stable.

B Determinant of the Jacobian

The calculation of the Jacobian in the evaluation of $p(x')$ in (4), where $x' = f(z')$ seems to be an ill-posed problem, because the Jacobian $\frac{\partial f(z)}{\partial z}$ is a rectangular matrix due to $f : \mathbb{R}^k \to \mathbb{R}^d$ with $d > k$. But recall that x' always lays on the manifold, and therefore the determinant should be calculated only with respect to the coordinate system on the manifold, which is of dimension k and therefore properly defined.

Let's align the coordinate system on x such that the last $d - k$ coordinates span the normal space of $f(z)$. Then due to the definition of the normal space it holds that the last $d-k$ columns of the Jacobian are all zeros, i.e. $(\forall j > k)\left(\frac{\partial f(z)_j}{z_i}\Big|_z = 0\right)$. Contrary, the first k components of the Jacobian define a local approximation of the manifold around the point $f(z)$. If f is a bijection, which is assumed here, they have a non-zero determinant.

In practice, the determinant can be easily calculated using Singular Value Decomposition of $J(f(z))$. Specifically, $\mathrm{svd}(J(f(z)) = \mathbf{U}\mathbf{\Sigma}\mathbf{V}^*$, where $\mathbf{U}$ and $\mathbf{V}$ are unitary matrices and Σ is a diagonal matrix with k non-zero singular values on the diagonal. The columns of V corresponding to zero singular values form a base of the normal space, and those corresponding to non-zero singular values form a base of the local approximation of the manifold. Since determinant is equal to the product of eigenvalues and singular values are their square roots, the product of squares of non-zero singular values is equal to the determinant of $J(f(z))$ with $f(z)$ determining the manifold.

C Is there a manifold in the data?

Since the data lie in the full space, it should be possible to find a mapping to the same dimensional latent space, in a similar manner as the flow based methods. The decomposition into a manifold and noise part may be seen as a simplification of the full model. We test if this modeling assumption is valid on an exhaustive search over all possible dimensions of the latent space on the "breast-cancer-wisconsin" dataset which has eight dimensions. The results of anomaly detection for all possible latent dimensions are displayed in Figure 4 with variability with respect to the splits of test/train data and hyper-parameters.

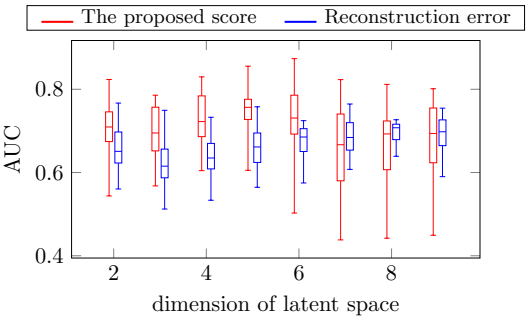


Figure 4: AUC of anomaly detection for a range of latent dimensions for the proposed criteria (red) and reconstruction error score (blue).

Note that while the conventional score based on reconstruction error is almost insensitive to the latent dimension, the performance of the anomaly detection based on the proposed score has a flat peak at 4–6 dimensions with decreasing performance for lower as well as higher dimensions. This suggests, that the modeling assumption of the low-dimensional latent space is beneficial. This may be relevant to the discussion on the performance of the full-dimensional latent space [13].