Out-of-Distribution Detection using Maximum Entropy Coding and Generative Networks

Mojtaba Abolfazli*, Mohammad Zaeri Amirani*, Anders Høst-Madsen*, June Zhang*, Andras Bratincsak†
*Department of Electrical and Computer Engineering, †Department of Pediatrics
University of Hawaii
e-mail: {mojtaba,zaeri,ahm,zjz}@hawaii.edu, andrasb@hphmg.org

***Abstract*—Given a default distribution P and a set of test data $x^M = \{x_1, x_2, \ldots, x_M\}$ this paper seeks to answer the question if it was likely that x^M was generated by P. For discrete distributions, the definitive answer is in principle given by Kolmogorov-Martin-Löf randomness. In this paper we seek to generalize this to continuous distributions. We consider a set of statistics $T_1(x^M), T_2(x^M), \ldots$. To each statistic we associate its maximum entropy distribution and with this a universal source coder. The maximum entropy distributions are subsequently combined to give a total codelength, which is compared with $-\log P(x^M)$. We show that this approach satisfied a number of theoretical properties.**

For real world data P usually is unknown. We transform data into a standard distribution in the latent space using a bidirectional generate network and use maximum entropy coding there. We compare the resulting method to other methods that also used generative neural networks to detect anomalies. In most cases, our results show better performance.

I. Introduction

We consider the following problem. Given a *default* distribution P (which could be continuous or discrete), and a set of test data $x^M = \{x_1, x_2, \ldots, x_M\}$ (which for now does not have to be IID), we would like to determine if it is likely that x^M was generated by P or by another distribution. For (binary) discrete data, this problem was solved theoretically by Kolmogorov and Martin-Löf through Kolmogorov complexity [1]. The starting point is what is called a P-test, which can be thought of as testing a specific data statistic (e.g., is the mean the correct one according to P). There are many such statistics, and a universal test is one that includes all statistics. Martin-Löf showed that a universal (sum) P-test is given by $K(x^M|M) < M$ when P is uniform, where K is Kolmogorov complexity. Replacing Kolmogorov complexity (which is uncomputable) with a universal source coder, this was used to develop Atypicality in [2].

In this paper we consider the following specific problem. We start with a continuous distribution P over $\mathbb{R}^n$ and IID test data $\mathbf{x}^M = \{\mathbf{x}_i \in \mathbb{R}^n\}, i = 1, \ldots, M > 1$. The question is if $\mathbf{x}^M$ is likely to have been generated by P. This problem is known as out-of-distribution (OOD) detection and is also called group anomaly detection (GAD). For this setup, Kolmogorov complexity cannot be directly applied. Our aim in this paper is to still use the principles of Martin-Löf randomness [1] to develop principled methods. We base this on statistics of the data, to which we associate maximum entropy distributions, which can in turn be used for coding.

The proofs of theorems can be found in the extended version of the paper [3].

The research was funded in part by the NSF grant CCF-1908957 and NIH grant 1202518S0.

A. Related Work

There have been several works on OOD or GAD, and we will just discuss a few. In statistics, there are for example the Kolmogorov-Smirnoff (KS) test [4] and the Pearson χ^2 test [5].

In machine learning, one-class learning has been used [6, 7]. Another approach is using probabilistic generative models and considering the OOD problem in the latent domain [8]. Reference [9] proposed a method based on the empirical entropy. [10, 11] used AAE and VAE neural networks to transform the data.

Many of the ML methods simply directly or indirectly use likelihood for OOD, i.e., how likely was it that data came from P? However, likelihood is not a good measure. As an example, suppose that P is the uniform IID distribution over $[0, 1]$. Then any sequence x^M is equally likely, i.e., nothing is OOD according to likelihood. The statistical tests for OOD therefore use the samples x^M to generate an alternative distribution. In the KS test [4], the empirical CDF of x^M is compared with the CDF of P. In the Pearson χ^2 test [5], the empirical PMF of x^M is compared with the PMF of P. This shows that OOD detection has to be generative, in the sense of coming up with an alternative distribution for the OOD data, and this is also consistent with Kolmogorov Martin-Löf randomness.

It is difficult to extend KS test to higher dimensions since computing the empirical CDFs depends on the arrangements of dimensions. A method was proposed in [12] and later improved in [13]. As opposed to that, our method can be used for high-dimensional data.

The method in this paper is related to Rissanen's minimum description length (MDL) [14, 15]. We have used this to find transients in [16] and it was used for change point detection in [17, 18]. The problem and methodology in this paper are different, so these papers are not directly applicable.

II. Methodology

There is no true generalization of universal source coding used for Atypicality in [2] in the real case, but we will still

maintain the idea of coding. For assuming that $\mathbf{x}^M$ comes from the default distribution, P, the codelength of the test set as argued by Rissanen [19] is

$$L_P(\mathbf{x}^M) = -\log P(\mathbf{x}^M) = -\log \prod_{i=1}^{M} P(\mathbf{x}_i).$$

We would like to compare this with a "universal" codelength.

Our starting point is Martin-Löf's idea of a P-test [1]. We consider a statistic $\mathbf{T} : \mathbb{R}^n \to \mathbb{R}^k$ with $\hat{\mathbf{t}} = \frac{1}{M}\sum_{i=1}^{M} \mathbf{T}(\mathbf{x}_i)$. If $\hat{\mathbf{t}} \neq E_P[\mathbf{T}(\mathbf{x})]$ one could consider the test data OOD. In order to put this both in a likelihood ratio test framework and a coding framework, we need to associate an alternative distribution with the statistic $\mathbf{T}$. The natural choice for such a distribution is the maximum entropy distribution [20], which we call P'_T.

In the maximum entropy distribution, the dimension of the statistic, k, corresponds to the number of free parameters.

It is natural to think that a more complex statistic will be able to capture more types of deviation from the default distribution. However, it might not be better for OOD detection, as the following theorem shows. Consider a maximum entropy distribution P_T and suppose that the default distribution $P = P_{t_0} = P_T(\mathbf{x}; \mathbf{T} = \mathbf{t}_0)$. Set a desired false alarm probability $\alpha = P_{FA}$ and detection probability $\beta = P_D$. Let $\mathcal{S}_{\alpha,\beta}(M)$ be the set of distributions $P_t = P_T(\mathbf{x}; \mathbf{T} = \mathbf{t})$ that can be detected with $P_{FA} \leq \alpha$ and $P_D \geq \beta$ with M samples. The question is how little deviation from the default distribution is needed for detection; we measure this by the radius $\inf_{P_t \in S_{\alpha,\beta}(M)} D(P_t \| P_{t_0})$, where $D(\cdot\|\cdot)$ is relative entropy. This radius can be calculated asymptotically as $M \to \infty$,

Theorem 1. *Let P_T be a maximum entropy distribution. Consider detection between $\mathbf{T} = \mathbf{t}_0$ and $\mathbf{T} \neq \mathbf{t}_0$. Fix the false alarm probability $\alpha = P_{FA}$ and the detection probability $\beta = P_D$ as $M \to \infty$. Let $\mathcal{S}_{\alpha,\beta}(M)$ be the set of distributions $P_t = P_T(\mathbf{x}; \mathbf{T} = \mathbf{t})$ that can be detected with $P_{FA} \leq \alpha$ and $P_D \geq \beta$ with M samples.. Then*

$$\lim_{M\to\infty} \inf_{P_t \in S_{\alpha,\beta}(M)} M \frac{1}{2\ln 2} D(P_t \| P_{t_0}) = F^{-1}_{\chi^2_k}(\beta) - F^{-1}_{\chi^2_k}(\alpha)$$

where $F_{\chi^2_k}$ is the CDF of the χ^2 distribution with k degrees of freedom.

This means that the closest (in the relative entropy sense) alternative distribution P' that can be detected as OOD depends on the choice of the statistic $\mathbf{T}$ since

$$\begin{aligned} D(P_{t_M} \| P_{t_0}) &\approx \frac{F^{-1}_{\chi^2_k}(\beta) - F^{-1}_{\chi^2_k}(\alpha)}{M} \\ &\approx \frac{\sqrt{2k}}{M}\left(\Phi^{-1}(\beta) - \Phi^{-1}(\alpha)\right) \end{aligned}$$

where Φ is the Normal CDF. This increases as $\sqrt{k}$, the size of the statistic. Thus, higher-complexity models are more difficult to detect, or more precisely, small deviations in high-complexity models are more difficult to detect.

The conclusion is that one should try to detect OOD with the simplest statistics possible. Yet, the statistic also has to be complex enough to capture deviations. The solution to this dilemma is to consider simple and complex statistics simultaneously, but with a penalty for more complex models in light of Theorem 1.

The statistics have to be chosen with the default distribution P in mind. Intuitively, the statistics have to indicate deviations from P well. To detect small deviations in distribution, one would like statistics so that $D(P_T\|P)$ can be made small. One way to obtain this is if P is itself a maximum entropy distribution for some value of $\mathbf{T}$ – then $D(P_T\|P)$ can be made arbitrarily small. Another possibility is to have a sequence of statistics $\mathbf{T}_i$ so that $D(P_{T_i}\|P)$ can be made arbitrarily small – but with a complexity cost according to Theorem 1. As an example, suppose that the default distribution is χ^2. If $\mathbf{T}(x) = (x, x^2)$ (mean and variance) the maximum entropy distribution is Gaussian, which is not close to χ^2; the consequence is that mean and variance have to change by large amounts for detection. But if $\mathbf{T}(x) = (x, \ln x)$, the maximum entropy distribution is Gamma, of which the χ^2 is a special case.

Using only maximum entropy distributions might seem limiting. However, the alternative distribution does not necessarily have to be modeled well. Suppose as an example the default distribution is $U[0,1]$, while data is generated according to $U[0,\theta]$, which is not maximum entropy. If $\theta > 1$, as soon as some $x > 1$ is seen, the default coder will give a codelength of infinity, and data will be declared OOD. If $\theta < 1$, the histogram distribution described below, Section II-B can be used, and this will detect $U[0,\theta]$.

A. Coding

Consider a sequence (finite or countable) of statistics $\mathbf{T}_i$ of varying complexity. We would like to combine all the statistics into a single test. This is similar to what is done for P-tests in Martin-Löf randomness [1]. We use a coding approach, inspired by Kolmogorov complexity and universal source coding in Atypicality [2]. The encoder and decoder both know the sequence of possible statistics $\mathbf{T}_i$. The idea is to use these statistics to encode the sequence $\mathbf{x}^M$ with the shortest codelength possible. Consider first a single statistic $\mathbf{T}$. One approach is that the encoder first calculates $\hat{\mathbf{t}} = \frac{1}{M}\sum_{i=1}^{M} \mathbf{T}(\mathbf{x}_i)$, conveys that to the decoder, and then encodes $\mathbf{x}^M$ with P_T. Since $\hat{\mathbf{t}}$ is real-valued, it has to be quantized to minimize total codelength. It will be noticed that is exactly as in Rissanen's minimum description length (MDL) [14, 15], and we can therefore use the rich theory from MDL. For example, one can use sequential coding instead of the two-step coding above. However, our aim is not to find a good model as in MDL.

Now consider the whole sequence of statistics $\mathbf{T}_i$. One approach to use the statistic $\mathbf{T}_i$ that results in the shortest codelength. From a coding point of view, the encoder needs to tell the decoder which statistic was used. This can be encoded using Elias code for the integers [21, 22], which uses $\log^* m + c$,

where $\log^* m = \log k + \log\log k + \cdots$, continuing until the argument to the log becomes negative, and c is a constant making Kraft's inequality satisfied with equality.

We can then write the resulting test explicitly as

$$\min_i -\log P'_{T_i}(\mathbf{x}^M) + L(\mathbf{T}_i) + \log^*(i) + \tau < -\log P(\mathbf{x}^M) \tag{1}$$

where $L(\mathbf{T}_i)$ is the length of the code to encode the (quantized) statistic $\mathbf{T}_i$. The threshold τ is chosen to achieve a desired false alarm probability. To recap, the coder on the left-hand side works by telling the decoder which encoder has been used, and then encoding according to it. However, a more efficient coder can be obtained by weighting the different coders, the principle in the Context Tree Weighting (CTW) coder, and other coders [23]. We can then write

$$-\log\left(\sum_{i=1}^{\infty} P'_{T_i}(\mathbf{x}^M) 2^{-L(\mathbf{T}_i)-\log^*(i)}\right) + \tau < -\log P(\mathbf{x}^M) \tag{2}$$

We call this *weighting*. Notice that this approach does not find a model with the shortest description length, and is therefore distinct from MDL. We will be using (2) in our implementation.

While we know from Theorem 1 that we should consider statistics of varying complexity, it is not obvious that combining them using (1) or (2) result in good OOD performance. Theoretical validation really is only possible asymptotically. The most meaningful limit is $k \to \infty$ while M is fixed or $M \ll k$ (if k is fixed and $M \to \infty$ one can just use the most complex model without much loss). Not all models allow $k \ll M$, and we will therefore limit the analysis to a specific case in the following section.

B. Asymptotic Analysis of Histogram Statistics

In this section we will show theoretically the advantage of the coding approach for combining statistics, limiting ourselves to scalar data for analytical tractability. We consider the case with the default distribution P uniform over $[0, 1]$. Any one-dimensional problem can be transformed into this by transformation with the CDF [20]: it is well known that for a continuous random variable X with CDF F_X, $U = F_X(X)$ has a uniform distribution on $(0, 1)$. We will see later that such transformations are essential for working with complex distributions. Thus, this is a general one-dimensional problem.

We use the following statistic: we divide the interval $[0, 1]$ into k equal-length subintervals and count the number of samples in each interval. The corresponding maximum entropy distribution is the uniform distribution over each subinterval. This is of course the histogram of the data. One notices that the default distribution is the histogram with $k = 1$, so this is a good sequence of statistics for this problem according to the theory in Section II.

This methodology is indeed a practical method in one dimension, competitive with KS. However, since we are mainly interested in high-dimensional data, we will not show any experimental results in this paper. We will use it to demonstrate theoretically that the code combining in (1) and (2) works.

In order to put this in a theoretical framework, we consider the case of a very concentrated alternative distribution. To detect this one does not need a large number of samples. If one has a few samples close together, this is a strong indication that the distribution is not uniform. This can be detected by a histogram with a small bin size, i.e., large k. For theoretical analysis, we will consider the extreme case of this, where the alternative distribution has a discontinuous CDF (i.e., discrete or mixed). There is then a good chance that two samples are identical, and that is a definite indication that the distribution is not uniform. We will show that the coding approach is able to detect this.

We will first argue that this is not possible without the coding combining in the sense that the false alarm probability is one no matter how large τ. Suppose that data is from the default distribution (uniform) and k is so large that all samples are in different bins. Then the negative log-likelihood is

$$\hat{L} = M \log M - M \log k$$

(when $k \gg M$ an efficient coder is to transmit the sequence q^M uncoded) which is unbounded (negative) as $k \to \infty$, i.e., for some k, $\hat{L} + \tau < 0$ no matter how large τ or M. Thus, without coding one has to limit $m < \infty$, and then one cannot detect identical samples for finite M.

On the other hand, with well-designed coding, we get the following result

Theorem 2. *For a histogram detector with an unbounded number of bins, there is a universal coder so that*

1) *For the detector using (1)*
   - *For sufficiently large τ and/or M, the false alarm probability can be made arbitrarily small.*
   - *If x^M has at least* three *non-unique samples (a sample repeated three times or two values repeated once), x^M will be classified as OOD with probability one.*
2) *For the weighted detector (2)*
   - *For sufficiently large τ and/or M, the false alarm probability can be made arbitrarily small.*
   - *If x^M has at least* two *non-unique samples x^M will be classified as OOD with probability one.*

The theorem also shows that the weighted detector (2) can be strictly better than the "model selection" detector (1); in terms of codelength (2) was already known to be better. At the same time, the proof of the theorem is based on a carefully designed coder, one that is optimum in the minimax coding sense. This indicates that using better coding in general – better coding meaning coders with shorter codelength – results in better detection.

III. Transformations

One important detail in Martin-Löf-Kolmogorov randomness detection is that the universal Turing machine implementing Kolmogorov complexity has as input also the default distribution P. In many cases, this disappears asymptotically, but not

in more complicated setups [1, Section 3.5]. The approach we take is to transform any distribution into a standard distribution, and then apply the coding approach.

This approach has several advantages

- Knowledge of P is utilized in the universal coder.
- Since the default distribution is always the same, a standard set of statistics can be used.
- Often small deviations from the default models can be captured by simple models, which is advantageous according to Theorem 1.

A simple example is given by the histogram approach in Section II-B. If P is very complex (e.g., with many extrema) and the data is generated by a distribution very close to P, one would need a very large k to detect this. On the other hand, if P is known, data can be transformed with the CDF, and even $k = 2$ might be able to detect the deviation.

IV. Multivariate Gaussian Default Model, P

We consider the case when the default model is Gaussian with zero mean and (known) covariance matrix $\mathbf{\Sigma}$; mostly we consider the case $\mathbf{\Sigma} = \mathbf{I}$. The aim is to find some statistics that capture deviation from this model well. The most obvious statistic is of course the mean and covariance, $\mathbf{T}(\mathbf{x}) = (\mathbf{x}, \mathbf{x}\mathbf{x}^T)$; the corresponding maximum entropy distribution is Gaussian $\mathcal{N}(\hat{\boldsymbol{\mu}}, \hat{\mathbf{\Sigma}})$. However, this is a high-complexity model, which should not be used alone according to Theorem 1. We therefore consider lower complexity models by specifying a sparse covariance matrix by requiring $\mathbf{\Sigma}^{-1}_{i,j} = 0$ for some coordinates $(i, j) \in J$ and putting $\mathbf{\Sigma}_{i,j} = \hat{\mathbf{\Sigma}}_{i,j}$ for $(i, j) \notin J$. There exists a unique positive definite matrix $\mathbf{S}$ satisfying these constraints, and the maximum entropy distribution is the Gaussian distribution with covariance matrix $\mathbf{S}$ [24]. The method is called covariance selection [24].

The minimum size of the covariance statistic that gives a valid maximum entropy distribution is the dimension of $\mathbf{x}$ (by only estimating the diagonal elements), which can still be high. We can also consider simpler statistics. One can start with $r^2 = \mathbf{x}^T\mathbf{x}$, and then consider statistics $\mathbf{T}(r^2)$. It is clear that the maximum entropy distribution corresponding to $\mathbf{T}(r^2)$ is uniformly distributed over an n-ball. Thus the maximum entropy distribution is $f(\mathbf{x}) = \frac{\Gamma(n/2)}{\pi^{n/2} r^{n-2}} f_r(r^2)$. For the default distribution, r^2 is χ^2 distributed. As discussed previously, it is advantageous to have the default distribution to be a special case of the maximum entropy distribution. We therefore use $\mathbf{T}(r^2) = (r^2, \ln(r^2))$ giving a Gamma maximum entropy distribution,

$$f(\mathbf{x}) = \frac{\Gamma(n/2)}{\pi^{n/2}(r^2)^{n/2-1}} \frac{\beta^\alpha}{\Gamma(\alpha)} (r^2)^{\alpha-1} \exp(-\beta r^2) \qquad (3)$$

Notice that these are just examples of possible statistics. One could also use higher-order moments and von Mises-Fisher statistics.

A. OOD under Multivariate Gaussian Default Model

As outlined, the default model P is assumed to be a known multivariate Gaussian distribution. The model for out-of-distribution data is also assumed to be Gaussian but with an unknown covariance matrix $\mathbf{\Sigma}$. Without loss of generality, we assume that everything is zero mean. We calculate the weighting criterion introduced in (2) to find if a batch of data $\mathbf{x}^M$ belongs to P or is OOD. In order to compute (2), we follow these steps:

1) Encode the data $\mathbf{x}^M$ with the known default model P. Therefore $L = -\log P(\mathbf{x}^M)$.
2) Encode the data $\mathbf{x}^M$ with universal multivariate Gaussian coder for all unique sparsity patterns obtained from covariance matrix estimation, giving a set of total codelengths $L_i, i = 1, \ldots, N$, where N is the number of distinct graphs considered. This is described in Section IV-B.
3) Encode the data $\mathbf{x}^M$ with universal Gamma distribution to account for $r^2 = \mathbf{x}^T\mathbf{x}$ statistics, giving L_g.
4) Combine codelengths from step–2 and step–3 using weighting in (2) to get $\widehat{L} = -\log\left(2^{-L_g} + \sum_{i=1}^{N} 2^{-L_i}\right)$ (the $\log^* k$ only matters for infinitely many models).
5) Given a threshold τ, the data $\mathbf{x}^M$ is OOD if $\widehat{L} + \tau < L$.

B. Universal Multivariate Gaussian Coder

A universal multivariate Gaussian coder was proposed in [25]; it is universal in the sense that it can be used to encode any multivariate Gaussian data. Our approach to finding the description length of a multivariate Gaussian model is based on characterizing the distribution by the sparsity pattern of the inverse covariance matrix, $\mathbf{\Sigma}^{-1}$. This sparsity pattern is known as the conditional independence graph, G, of the Gaussian. It can be found by using several structure learning methods such as graphical lasso (GLasso) [26]; these methods often use a regularization parameter, λ, to control for the sparsity of the solution. Each value of λ is associated with a conditional independence graph G. Here, we want to combine the codelength of unique models and as a consequence, we consider unique conditional independence graphs.

C. Experiments on Synthetic Data

We compared our approach, maximum entropy coding (MEC), to $d-$dimensional KS test (ddKS) method [13], a multi-dimensional two-sample KS test, for the test cases described in Table I. For Case-1&2, the distribution of both the default model, P and the alternative model, $\widehat{P}$ are multivariate Gaussian. For Case-3 to 6, the test data is generated from linear transformation of nearly-Gaussian distributions, $\mathbf{A}\mathbf{x}$, where $[\mathbf{A}_{ij}]$ is the transformation matrix.

For each test case, we performed OOD detection on a synthetically generated test dataset of size M. We repeated the experiment 1000 times. The AUROC for the six scenarios is shown in Table II. As it can be seen, our approach outperforms ddKS method in all cases, even for the cases where the OOD data do not come from Gaussian distributions.

V. Unknown Default Model, P

In the previous section, we have shown that our coding-based OOD detection approach works well for multivariate

TABLE I: Scenarios for generating synthetic test data.

	Parameters of Default model, P	Parameters of Anomalous model, $\widehat{P}$	Data generation
CASE–1	$\mathbf{\Omega}_{ii} = 1, \mathbf{\Omega}_{i,i-1} = \mathbf{\Omega}_{i-1,i} = 0.45$ $\mathbf{\Omega}_{16} = \mathbf{\Omega}_{61} = 0.45$	$\mathbf{\Omega}_{ii} = 1, \mathbf{\Omega}_{i,i-1} = \mathbf{\Omega}_{i-1,i} = 0.45$	$\mathbf{X} \sim \mathcal{N}(\mathbf{0}, \mathbf{\Omega}^{-1})$
CASE–2	$\mathbf{\Omega}_{ii} = 1, \mathbf{\Omega}_{i,i-1} = \mathbf{\Omega}_{i-1,i} = 0.45$ $\mathbf{\Omega}_{16} = \mathbf{\Omega}_{61} = 0.45$	$\mathbf{\Omega}_{ii} = 1, \mathbf{\Omega}_{i,i-1} = \mathbf{\Omega}_{i-1,i} = 0.5$ $\mathbf{\Omega}_{i,i-2} = \mathbf{\Omega}_{i-2,i} = 0.25$	$\mathbf{X} \sim \mathcal{N}(\mathbf{0}, \mathbf{\Omega}^{-1})$
CASE–3	$\mathbf{A}_{ii} = 1, \mathbf{A}_{i,i-1} = \mathbf{A}_{i-1,i} = 0.5$ $\mathbf{A}_{i,i-2} = \mathbf{A}_{i-2,i} = 0.25$	$\mathbf{A}_{ii} = 1, \mathbf{A}_{i,i-1} = \mathbf{A}_{i-1,i} = 0.4$ $\mathbf{A}_{i,i-2} = \mathbf{A}_{i-2,i} = 0.2$ $\mathbf{A}_{i,i-3} = \mathbf{A}_{i-3,i} = 0.2$	$x_i \sim$ Laplace$(0, i)$
CASE–4	same as CASE–3	same as CASE–3	$x_i \sim$ Logistic$(0, i)$
CASE–5	same as CASE–3	same as CASE–3	$x_i \sim \chi^2_{i+4}$
CASE–6	same as CASE–3	same as CASE–3	$x_i \sim$ StudentT$(i + 4)$

TABLE II: AUROC comparing our approach (MEC) to the multi-dimensional KS test in [13] (ddKS).

	$\mathbf{M = 25}$		$\mathbf{M = 50}$	
	MEC	ddKS	MEC	ddKS
CASE–1	**0.957**	0.824	**0.985**	0.939
CASE–2	**0.999**	0.987	**1.0**	**1.0**
CASE–3	**0.980**	0.553	**0.994**	0.582
CASE–4	**0.984**	0.564	**0.994**	0.594
CASE–5	**0.920**	0.563	**0.944**	0.591
CASE–6	**0.983**	0.707	**0.991**	0.888

Gaussian and near-Gaussian distributions. However, most real-world data are far from Gaussian. The default model usually is not known for real-world data. We overcome this by using a (non-linear) continuous transform so that the data is Gaussian in the transformed or the latent space, following the theory in Section III. In this paper, we used generative neural networks to transform arbitrary data to multivariate Gaussian. Our requirements are that 1) the transformation, $\mathbf{z} = f(\mathbf{x})$, from the data space, $\mathbf{x} \in \mathbb{R}^n$, to latent space, $\mathbf{z} \in \mathbb{R}^m$, is invertible. This means that there is a function g such that $\mathbf{x} = g(\mathbf{z}) = f^{-1}(\mathbf{z})$; 2) the distribution in the latent space can be specified (usually as a multivariate Gaussian).

For the transformation, we used Glow [27], a flow-based generative network. Glow is exactly invertible.

VI. EXPERIMENTS ON REAL-WORLD DATA

We considered the digital image dataset MNIST [28] where we do not know the default model distribution P for the data. Instead, we have a set of training data $\mathbf{x}^N$. We took the training data from the MNIST dataset and considered three sets of experiments for OOD detection:

- **Experiment 1:** Detect if a test set is from MNIST or fashion MNIST [29].
- **Experiment 2:** Detect if a test set is from MNIST or non-MNIST [30].
- **Experiment 3:** Detect if a test set is from MNIST or synthetically-perturbed MNIST (see Table III for the description of the different datasets).

The training data consists of $60,000$ black and white images: $\{\mathbf{x} \in \mathbb{R}^{28\times 28}\}$. In order to use our method in the current implementation, we had to downsample the image from 28×28 to 8×8 pixels. This is because currently inverting very large covariance matrices results in numerical instability. We are working on approximate matrix inversion to address this issue.

TABLE III: Scenarios for synthetically perturbing MNIST images. Rotation and shearing values are in degree, width and height shift in fraction, zoom, and brightness in range.

	Perturbation type, Value
CASE–1	Rotation, 5
CASE–2	Shearing, 20
CASE–3	[Width shift , Height shift], $[0.02, 0.02]$
CASE–4	Zooming, $[0.8, 1.2]$
CASE–5	Zooming, $[1, 1.1]$
CASE–6	Zooming, $[0.9, 1]$
CASE–7	Brightness, $[0.2, 2]$
CASE–8	Brightness, $[0.2, 1]$
CASE–9	Gaussian noise, $\mu = 0, \sigma = 0.05$

We solve OOD detection problem on test datasets of size M. We repeated the experiment 1000 times.

We compared our approach to another Glow-based method called Typicality [9]. We trained our model with the same hyperparameters and settings as [9]. We can not compare to the ddKS method used in Section IV-C because the data is too high-dimensional.

Table IV shows the AUROC for the experiments using different test set sizes M. Our method has higher performance than Typicality method in all cases except CASE–5.

TABLE IV: AUROC for MNIST experiments comparing our MEC to Typicality [9] trained and tested on downsampled images. The best value for each case is boldfaced.

	$\mathbf{M = 50}$		$\mathbf{M = 100}$	
	MEC	Typicality	MEC	Typicality
fashion MNIST	**1.000**	**1.000**	**1.000**	**1.000**
not-MNIST	**1.000**	**1.000**	**1.000**	**1.000**
CASE–1	**1.000**	0.995	**1.000**	**1.000**
CASE–2	**1.000**	0.998	**1.000**	**1.000**
CASE–3	**1.000**	**1.000**	**1.000**	**1.000**
CASE–4	**1.000**	0.502	**1.000**	0.505
CASE–5	0.788	**0.944**	0.855	**0.974**
CASE–6	**1.000**	**1.000**	**1.000**	**1.000**
CASE–7	**0.978**	0.788	**0.985**	0.849
CASE–8	**0.883**	0.430	**0.928**	0.380
CASE–9	**1.000**	**1.000**	**1.000**	**1.000**

VII. CONCLUSION

The paper has shown that maximum entropy coding can be used for OOD detection. It has a number of desirable theoretical properties and performs well on real world data.

REFERENCES

[1] Ming Li and Paul Vitányi, *An Introduction to Kolmogorov Complexity and Its Applications*, Springer, fourth edition, 2008.

[2] Anders Høst-Madsen, Elyas Sabeti, and Chad Walton, "Data discovery and anomaly detection using atypicality," *IEEE Transactions on Information Theory*, vol. 65, no. 9, September 2019.

[3] Mojtaba Abolfazli, Mohammad Zaeri Amirani, Anders Høst-Madsen, June Zhang, and Andras Bratincsak, "Out-of-distribution detection using maximum entropy coding," *ArXiv e-prints 2404.17023*, 2024.

[4] Gregory W Corder and Dale I Foreman, *Nonparametric Statistics: A Step-by-Step Approach*, John Wiley & Sons, 2014.

[5] Erich Leo Lehmann, Joseph P Romano, and George Casella, *Testing statistical hypotheses*, vol. 3, Springer, 1986.

[6] Varun Chandola, Arindam Banerjee, and Vipin Kumar, "Anomaly detection: A survey," *ACM computing surveys (CSUR)*, vol. 41, no. 3, pp. 1–58, 2009.

[7] Krikamol Muandet and Bernhard Schölkopf, "One-class support measure machines for group anomaly detection," *arXiv preprint arXiv:1303.0309*, 2013.

[8] Stephan Rabanser, Stephan Günnemann, and Zachary Lipton, "Failing loudly: An empirical study of methods for detecting dataset shift," *Advances in Neural Information Processing Systems*, vol. 32, 2019.

[9] Eric Nalisnick, Akihiro Matsukawa, Yee Whye Teh, and Balaji Lakshminarayanan, "Detecting out-of-distribution inputs to deep generative models using typicality," *ArXiv e-prints 1906.02994*, 2019.

[10] Raghavendra Chalapathy, Edward Toth, and Sanjay Chawla, "Group anomaly detection using deep generative models," in *Joint European Conference on Machine Learning and Knowledge Discovery in Databases*. Springer, 2018, pp. 173–189.

[11] Yufeng Zhang, Wanwei Liu, Zhenbang Chen, Ji Wang, Zhiming Liu, Kenli Li, and Hongmei Wei, "Towards out-of-distribution detection with divergence guarantee in deep generative models," *arXiv preprint arXiv:2002.03328*, 2020.

[12] Giovanni Fasano and Alberto Franceschini, "A multi-dimensional version of the kolmogorov–smirnov test," *Monthly Notices of the Royal Astronomical Society*, vol. 225, no. 1, pp. 155–170, 1987.

[13] Alex Hagen, Shane Jackson, James Kahn, Jan Strube, Isabel Haide, Karl Pazdernik, and Connor Hainje, "Accelerated computation of a high dimensional kolmogorov-smirnov distance," *arXiv preprint arXiv:2106.13706*, 2021.

[14] Jorma Rissanen, "Modeling by shortest data description," *Automatica*, pp. 465–471, 1978.

[15] Peter D. Grunwald, *The Minimum Description Length Principle*, MIT Press, 2007.

[16] Elyas Sabeti and Anders Høst-Madsen, "Data discovery and anomaly detection using atypicality for real-valued data," *Entropy*, p. 219, Feb. 2019, Available at https://doi.org/10.3390/e21030219.

[17] Kenji Yamanishi and Shintaro Fukushima, "Model change detection with the mdl principle," *IEEE Transactions on Information Theory*, vol. 64, no. 9, pp. 6115–6126, 2018.

[18] Kenji Yamanishi, Linchuan Xu, Ryo Yuki, Shintaro Fukushima, and Chuan-hao Lin, "Change sign detection with differential mdl change statistics and its applications to covid-19 pandemic analysis," *Scientific Reports*, vol. 11, no. 1, pp. 19795, 2021.

[19] Jorma Rissanen, "Stochastic complexity and modeling," *The Annals of Statistics*, , no. 3, pp. 1080–1100, Sep. 1986.

[20] Thomas.M. Cover and Joy.A. Thomas, *Elements of Information Theory, 2nd Edition*, John Wiley, 2006.

[21] Jorma Rissanen, "A universal prior for integers and estimation by minimum description length," *The Annals of Statistics*, , no. 2, pp. 416–431, 1983.

[22] P. Elias, "Universal codeword sets and representations of the integers," *Information Theory, IEEE Transactions on*, vol. 21, no. 2, pp. 194 – 203, mar 1975.

[23] F. M J Willems, Y.M. Shtarkov, and T.J. Tjalkens, "The context-tree weighting method: basic properties," *Information Theory, IEEE Transactions on*, vol. 41, no. 3, pp. 653–664, 1995.

[24] A. P. Dempster, "Covariance selection," *Biometrics*, vol. 28, no. 1, pp. 157–175, 1972.

[25] Mojtaba Abolfazli, Anders Host-Madsen, June Zhang, and Andras Bratincsak, "Graph compression with application to model selection," *arXiv preprint arXiv:2110.00701*, 2021.

[26] Jerome Friedman, Trevor Hastie, and Robert Tibshirani, "Sparse inverse covariance estimation with the graphical lasso," *Biostatistics*, vol. 9, no. 3, pp. 432–441, 2008.

[27] Durk P Kingma and Prafulla Dhariwal, "Glow: Generative flow with invertible 1x1 convolutions," *Advances in neural information processing systems*, vol. 31, 2018.

[28] Li Deng, "The mnist database of handwritten digit images for machine learning research," *IEEE Signal Processing Magazine*, vol. 29, no. 6, pp. 141–142, 2012.

[29] Han Xiao, Kashif Rasul, and Roland Vollgraf, "Fashion-mnist: a novel image dataset for benchmarking machine learning algorithms," *arXiv preprint arXiv:1708.07747*, 2017.

[30] Yaroslav Bulatov, "Notmnist dataset," *Google (Books/OCR), Tech. Rep.[Online]. Available: http://yaroslavvb. blogspot. it/2011/09/notmnist-dataset. html*, vol. 2, 2011.