Out-of-Distribution Detection using Maximum Entropy Coding

Mojtaba Abolfazli, Mohammad Zaeri Amirani, Anders Høst-Madsen, June Zhang, Andras Bratincsak

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 [21]. 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 [18].

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 [21] 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.

A. Related Work

There have been a number of works on OOD or GAD, and we will just discuss a few. In statistics there are for example the Kolmogorov-Smirnoff (KS) test [7] and the Pearson χ^2 test [20].

M. Abolfazli, M. Zaeri Amirani, A. Høst-Madsen, and J. Zhang are with the Department of Electrical and Computer Engineering, University of Hawaii at Manoa, 2540 Dole Street, Honolulu, HI 96822; e-mail: {mojtaba,ahm,zjz}@hawaii.edu.

A. Bratincsak is with the Department of Pediatrics, John A. Burns School of Medicine, University of Hawaii, Honolulu, HI 96813; e-mail: andrasb@hphmg.org.

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

In machine learning one-class learning has been used [6, 23]. Another approach is using probabilistic generative models and considering the OOD problem in the latent domain [25]. Reference [24] proposed a method based on the empirical entropy. [5, 37] 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 [7], the empirical CDF of x^M is compared with the CDF of P. In the Pearson χ^2 test [20], 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. Some methods were proposed in [12, 15]. 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) [27, 14]. We have used this to find transients in [30] and it was used for change point detection in [35, 36]. 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 [18] 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 [29] 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 [21]. 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 which has the form [8]

$$P_T(\mathbf{x}) = \exp\left(\boldsymbol{\lambda}^T \mathbf{T}(\mathbf{x}) - A(\boldsymbol{\lambda})\right) \tag{1}$$

Here $A(\boldsymbol{\lambda})$ is the log-partition functions ensuring P_T integrates to one. The statistic therefore has to be one that has a valid maximum entropy distribution. While the maximum entropy distribution seems reasonable, it also has the following straighforward optimality property

Proposition 1. *Consider the set of distributions P' that satisfy $E_{P'}[\mathbf{T}(\mathbf{x})] = \mathbf{t}$. Among those the maximum entropy distribution P_T is the minimax coding distribution, i.e., it achieves*

$$\inf_{P:E_P[\mathbf{T}(x)]=\mathbf{t}} \sup_{P':E_{P'}[\mathbf{T}(x)]=\mathbf{t}} E_{P'}[-\log P(\mathbf{x})]$$

Proof. It is clear from (1) that $E_{P'}[-\log P(\mathbf{x})]$ is independent of P'. If there were some other distribution $\tilde{P}$ that had lower minimax coding length than P_T for all P', it would therefore have to be strictly smaller for all P'. But P_T is optimum if the data is actually generated by P_T, so this is not possible. □

Maximum entropy is actually commonly used in universal source coding. One method for universal source coding (of discrete sources) is to first transmit the type of the sequence (i.e., the histogram) and then the specific sequence with that type (e.g., [8, Section 13.2] – assuming a uniform distribution over sequences, which is precisely maximum entropy.

The importance of this property can be seen as follows. The log-likelihood ratio test with P_T can be written as

$$L_P(\mathbf{x}^M) - L_{P_T}(\mathbf{x}^M) \geq \tau$$

To the first order, the detection performance depends on $E_{P'}[L_P(\mathbf{x})] - E_{P'}[L_{P_T}(\mathbf{x})]$. According to Proposition 1 the second term is independent of P'. In many cases the first term is also independent of P'. For example if P is uniform over $[0, 1]$ or if P is itself a maximum entropy distribution for a subset of the statistic $\mathbf{T}$. In those cases, using maximum entropy results in a detection performance independent of P' and only dependent on $\mathbf{T}$ – to the first order.

It is natural to think that a more complex statistic will be able to capture more types of deviation from the default distribution. But 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 $M \to \infty$, To prove it, we need the following Lemma

Lemma 2. *Let $\mathbf{T}$ be a statistic and P_T the corresponding maximum entropy distribution. Suppose that data is generated according to $P_T(\mathbf{x}; \mathbf{T} = \mathbf{t})$ and let $\hat{\mathbf{t}} = \frac{1}{M}\sum_{i=1}^{M} T(\mathbf{x}_i)$. Then*

$$2 \ln 2(-\log P(\mathbf{x}^M; \mathbf{T} = \mathbf{t}) + \log P(\mathbf{x}^M; \mathbf{T} = \hat{\mathbf{t}}) \xrightarrow{D} \chi^2_m$$

Proof. Maximum entropy distributions (1) are in the exponential family. Notice that $\hat{\mathbf{t}} = \frac{1}{M}\sum_{i=1}^{M} \mathbf{T}(\mathbf{x}_i)$ is the maximum likelihood estimator for this model through some transformation $\hat{\mathbf{t}} \mapsto \hat{\boldsymbol{\lambda}}$ [22]. We also have [22]

$$\sqrt{M}\left(\hat{\boldsymbol{\lambda}} - \boldsymbol{\lambda}\right) \xrightarrow{D} N(0, \mathbf{J}^{-1})$$

where $\mathbf{J} = \mathbf{J}(\boldsymbol{\lambda})$ is the Fisher information matrix. Let us use the notation $\mathbf{J}(\hat{\boldsymbol{\lambda}}) = \nabla^2_{\boldsymbol{\lambda}} \left(A(\boldsymbol{\lambda})\right)|_{\boldsymbol{\lambda}=\hat{\boldsymbol{\lambda}}}$, so:

$$E\left[\mathbf{J}(\hat{\boldsymbol{\lambda}})\right] = -\nabla^2_{\boldsymbol{\lambda}} \ln P_T(x; \boldsymbol{\lambda}) = \nabla^2_{\boldsymbol{\lambda}} A(\boldsymbol{\lambda}) = \mathbf{J}$$

Now

$$\begin{aligned} R &= -\ln P(\mathbf{x}^M; \mathbf{T} = \mathbf{t}) + \ln P(\mathbf{x}^M; \mathbf{T} = \hat{\mathbf{t}}) \\ &= \boldsymbol{\lambda}^T T(\mathbf{x}^M) - MA(\boldsymbol{\lambda}) - \hat{\boldsymbol{\lambda}}^T T(\mathbf{x}^M) + MA(\hat{\boldsymbol{\lambda}}) \end{aligned}$$

We expand R as a Taylor series around $\hat{\boldsymbol{\lambda}}$. Notice that $\frac{\partial R}{\partial \boldsymbol{\lambda}}\big|_{\boldsymbol{\lambda}=\hat{\boldsymbol{\lambda}}} = 0$ because $\hat{\boldsymbol{\lambda}}$ is the MLE. Thus,

$$R = -\frac{M}{2}\left(\hat{\boldsymbol{\lambda}} - \boldsymbol{\lambda}\right)^T \mathbf{J}(\hat{\boldsymbol{\lambda}}) \left(\hat{\boldsymbol{\lambda}} - \boldsymbol{\lambda}\right) + o(\boldsymbol{\lambda}^2)$$

According to [22, p. 271] $\mathbf{J}(\boldsymbol{\lambda})$ is a continuous function of $\boldsymbol{\lambda}$. As $\hat{\boldsymbol{\lambda}} \xrightarrow{p} \boldsymbol{\lambda}$ then $\mathbf{J}(\hat{\boldsymbol{\lambda}}) \xrightarrow{p} \mathbf{J}$ by Slutsky's theorem [31]. Therefore, also by Slutsky's theorem, the first term in R is χ^2. Finally, applying the univariate delta method on the function:

$$g(\hat{\boldsymbol{\lambda}}) = R + \frac{M}{2}\left(\hat{\boldsymbol{\lambda}} - \boldsymbol{\lambda}\right)^T \mathbf{J}(\hat{\boldsymbol{\lambda}}) \left(\hat{\boldsymbol{\lambda}} - \boldsymbol{\lambda}\right)$$

results in $g(\hat{\boldsymbol{\lambda}}) \xrightarrow{p} 0$, which proves the Lemma.

□

Theorem 3. *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.

Proof. Let

$$R = 2\ln 2(-\log P(\mathbf{x}^M; \mathbf{T} = \mathbf{t}) + \log P(\mathbf{x}^M; \mathbf{T} = \hat{\mathbf{t}}))$$

Then according to Lemma 2 the false alarm probability satisfies

$$\lim_{M\to\infty} P_{FA} = \lim_{M\to\infty} P(R > \tau) = 1 - F_{\chi^2_k}(\tau) \tag{2}$$

Consider a sequence of distributions $P_{t_M}(\mathbf{x}) = P_T(\mathbf{x}; \mathbf{T} = t_M)$ that satisfy

$$\lim_{M\to\infty} M D(P_{t_M} \| P_{t_0}) = K \tag{3}$$

for some constant K. For a maximum entropy distribution, we have

$$M D(P_{t_M} \| P_{t_0}) = M(\boldsymbol{\lambda} - \boldsymbol{\lambda}_0)^T E_{\boldsymbol{\lambda}}[\mathbf{T}(\mathbf{x}] + M(A(\boldsymbol{\lambda}_0) - A(\boldsymbol{\lambda}))$$

In order for (3) to be satisfied, we must then have

$$\lim_{M\to\infty} M(\boldsymbol{\lambda} - \boldsymbol{\lambda}_0) = \mathbf{v} \tag{4}$$

for some vector $\mathbf{v}$. Now write

$$\begin{aligned} R = 2\ln 2 \big(&-\log P_{t_0}(\mathbf{x}^M) + \log P_{t_M}(\mathbf{x}^M) \\ &- \log P_{t_M}(\mathbf{x}^M) + \log P_{\hat{\mathbf{t}}}(\mathbf{x}^M)\big) \end{aligned}$$

The first term can be written as

$$\begin{aligned} &- \ln P_{t_0}(\mathbf{x}^M) + \ln P_{t_M}(\mathbf{x}^M) \\ &\quad = (\boldsymbol{\lambda} - \boldsymbol{\lambda}_0)^T \sum_{i=1}^{M} T(\mathbf{x}_i) + M(A(\boldsymbol{\lambda}_0) - A(\boldsymbol{\lambda})) \\ &\quad = M(\boldsymbol{\lambda} - \boldsymbol{\lambda}_0)^T \frac{1}{M} \sum_{i=1}^{M} T(\mathbf{x}_i) + M(A(\boldsymbol{\lambda}_0) - A(\boldsymbol{\lambda})) \\ &\quad \xrightarrow{P} K \end{aligned}$$

because of the law of large numbers and (3) and (4). At the same time

$$2\ln 2 \left(-\log P_{t_M}(\mathbf{x}^M) + \log P_{\hat{t}}(\mathbf{x}^M)\right) \xrightarrow{D} \chi^2_k$$

according to Lemma 2. There is the difference that the parameter is not fixed, but examining the proof of Lemma 2 shows that this makes no difference. Then from Slutsky's theorem

$$R \xrightarrow{D} \chi^2_k + K$$

Thus

$$\lim_{M\to\infty} P_D = F_{\chi^2_k}(\tau - K) \tag{5}$$

The theorem now follows by solving (2) and (5) with respect to K for given α and β.

□

Perhaps a more intuitive measure of distance between distributions is total variation distance, for continuous distributions

$$d_{\text{TV}}(P_{y_M}, P_{y_0}) = \frac{1}{2}\int |P_{y_M}(x) - P_{y_0}(x)|dx$$

Then Pinsker's inequality gives

$$\lim_{M\to\infty} \inf_{P_{y_M}} \sqrt{M} d_{\text{TV}}(P_{t_M}, P_{t_0}) \leq \sqrt{F^{-1}_{\chi^2_m}(\beta) - F^{-1}_{\chi^2_m}(\alpha)}$$

What this means is that the closest distribution that can be detected as OOD is

$$D(P_{t_M}\|P_{t_0}) \approx \frac{F^{-1}_{\chi^2_m}(\beta) - F^{-1}_{\chi^2_m}(\alpha)}{M}$$

This increases nearly linearly with m. 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 consider simple and complex statistics simultaneously, but with a penalty for more complex models in light of Theorem 3.

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 T_i so that $D(P_{T_i}\|P)$ can be made arbitrarily small – but with a complexity cost according to Theorem 3. 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 has 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 [21]. We use a coding approach, inspired by Kolmogorov complexity and universal source coding in Atypicality [18]. 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) [27, 14], 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 [28, 11], 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

$$\min_i -\log P'_{T_i}(\mathbf{x}^M) + L(\mathbf{T}_i) + \log^*(i) + \tau < -\log P(\mathbf{x}^M) \tag{6}$$

where $L(\mathbf{T}_i)$ is the length of the code to encode the (quantized) statistic $\mathbf{T}_i$ – we will later describe how precisely the coding is done. 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 [33]. We can then write

$$-\log\left(\sum_{i=1}^{\infty} P'_{T_i}(\mathbf{x}^M) 2^{-L(\mathbf{T}_i)-\log^*(i)}\right) + \tau < -\log P(\mathbf{x}^M) \tag{7}$$

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 (7) in our implementation.

While we know from Theorem 3 that we should consider statistics of varying complexity, it is not obvious that combining them using (6) or (7) 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.

We will show that in some cases, coding results in optimum performance in the sense of Theorem 3. Consider a sequence of statistics $\mathbf{T}_i$ such that $\mathbf{T}_i$ is a subset of the statistics in $\mathbf{T}_{i+1}$. Suppose that the default distribution is the maximum entropy distribution corresponding to $\mathbf{T}_1$, and suppose the OOD data is generated by the maximum entropy distribution corresponding to $\mathbf{T}_k$ for some $k > 1$. If k is known, Theorem 3 gives the peformance. As k is not known, we use (6) or (7).

Theorem 4. *Consider the same setup as in Theorem 3 and suppose (6) or (7) is used for detection. Assume* $L(T_i) = \frac{|\mathbf{T}_i|}{2}\log M$ *and ignore* $\log^* m$. *Then the detection performance is the same as if* k *were known and given by*

$$\lim_{M\to\infty} \inf_{P_t \in S_{\alpha,\beta}(M)} \frac{M}{\log M} \frac{1}{2\ln 2} D(P_t \| P_{t_0}) \leq \frac{|\mathbf{T}_k|}{2} \tag{8}$$

Proof. Let us assume we consider statistics $\mathbf{T}_i$ for $i \leq K(M)$ where $K(M)$ is an increasing function of M. We can bound the false alarm probability by the union bound

$$P_{FA} \leq \sum_{i=1}^{K(M)} P\left(-\log P(\mathbf{x}^{\mathbf{M}}) + \log P_{T_i}(\mathbf{x}^M) \geq \tau + \frac{|\mathbf{T}_i|}{2}\log M\right) \tag{9}$$

According to Lemma 2 $2\ln 2(-\log P(\mathbf{x}^{\mathbf{M}}) + \log P_{T_i}(\mathbf{x}^M))$ is approximately χ^2. We will first evaluate (9) under some simplifying assumptions: 1) $2\ln 2(-\log P(\mathbf{x}^{\mathbf{M}}) + \log P_{T_i}(\mathbf{x}^M))$ is exact χ^2, 2) $|\mathbf{T}_i| = i$, 3) $\tau = 0$. We can then use the Chernoff bound on (9) to get

$$\begin{aligned} P_{FA} &\leq \sum_{i=1}^{K(M)} \exp\left(-i\ln M\right)\left(\frac{i\ln M}{2i}\right)^{i/2} \\ &= \sum_{i=1}^{K(M)} \left(\frac{\sqrt{\ln M}}{\sqrt{2}M}\right)^i \\ &\leq \frac{\sqrt{\ln M}}{\sqrt{2}M}\frac{1}{1-\frac{\sqrt{\ln M}}{\sqrt{2}M}} \end{aligned}$$

We notice that $P_{FA} \to 0$ not matter what $K(M)$ is, that is, at some point it will be less than the given α. Of course the χ_2 approximation is not exact. However, we can always find some $K(M) \to \infty$ so that the

χ_2 approximation is good enough so that still $P_{FA} \to 0$. If $|\mathbf{T}_i| > i$ it can only lead to faster decrease towards zero. Finally, since we only require $P_{FA} \leq \alpha$, we could decrease τ below zero. The upside is that we can ensure $P_{FA} \leq \alpha$ for sufficiently large M, even if the assumptions are not satisfied.

Consider a sequence of distributions $P_{t_M}(\mathbf{x}) = P_{T_k}(\mathbf{x}; \mathbf{T}_k = \mathbf{t}_M)$ that satisfy

$$\lim_{M\to\infty} \frac{M}{\log M} D(P_{t_M} \| P_{t_0}) = K \tag{10}$$

for some constant K. For a maximum entropy distribution, we have

$$D(P_{t_M} \| P_{t_0}) = (\boldsymbol{\lambda} - \boldsymbol{\lambda}_0)^T E_{\boldsymbol{\lambda}}[\mathbf{T}(\mathbf{x}] + (A(\boldsymbol{\lambda}_0) - A(\boldsymbol{\lambda}))$$

In order for (10) to be satisfied, we must then have

$$\lim_{M\to\infty} \frac{M}{\log M} (\boldsymbol{\lambda} - \boldsymbol{\lambda}_0) = \mathbf{v} \tag{11}$$

for some vector $\mathbf{v}$. Now write

$$\begin{aligned} R = 2 \ln 2 \big(&- \log P_{t_0}(\mathbf{x}^M) + \log P_{t_M}(\mathbf{x}^M) \\ &- \log P_{t_M}(\mathbf{x}^M) + \log P_{\hat{\mathbf{t}}}(\mathbf{x}^M) \big) \end{aligned}$$

For the first term we have

$$\begin{aligned} &\frac{1}{\log M} \big(- \ln P_{t_0}(\mathbf{x}^M) + \ln P_{t_M}(\mathbf{x}^M) \big) \\ &\quad = \frac{1}{\log M} \left((\boldsymbol{\lambda} - \boldsymbol{\lambda}_0)^T \sum_{i=1}^{M} T(\mathbf{x}_i) + M(A(\boldsymbol{\lambda}_0) - A(\boldsymbol{\lambda})) \right) \\ &\quad = \frac{M}{\log M} (\boldsymbol{\lambda} - \boldsymbol{\lambda}_0)^T \frac{1}{M} \sum_{i=1}^{M} T(\mathbf{x}_i) + \frac{M}{\log M} (A(\boldsymbol{\lambda}_0) - A(\boldsymbol{\lambda})) \\ &\quad \xrightarrow{P} K \end{aligned}$$

because of the law of large numbers and (3) and (11). At the same time

$$2 \ln 2 \big(- \log P_{t_M}(\mathbf{x}^M) + \log P_{\hat{t}}(\mathbf{x}^M) \big) \xrightarrow{D} \chi^2_k$$

The detection probability is

$$P_D \geq P \left(R \geq \frac{|\mathbf{T}_k|}{2} \log M \right)$$

Then from Slutsky's theorem

$$R \xrightarrow{D} \chi^2_k + K$$

Thus

$$\lim_{M\to\infty} P_D = F_{\chi^2_k}(\tau - K) \tag{12}$$

The theorem now follows by solving (2) and (12) with respect to K for given α and β.

□

While the above gives some indication that coding works, 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. Histogram

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 [8]: 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 m 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 $m = 1$, so this is a good sequence of statistics for this problem according to the theory in Section II.

Let $M\hat{p}_j$ be the number of samples in the j-th interval. The maximum entropy distribution then is

$$f(x) = \hat{p}_{q(x)} m$$

where $q(x)$ is the interval that contains x. We need to use this distribution to code x^M. Let q^M be the quantization of x^M. The coding can then be done as follows: first transmit $\hat{p}_j$ (the type of q^M), and then which specific sequence q^M is in that type class; it is now known in which interval x_i is, and only the specific value has to be transmitted with $-\log m$ bits. We can then write the total codelength as

$$\hat{L} = \hat{L}_q - M \log m \tag{13}$$

where $\hat{L}_q$ is the codelength to encode q^M with a universal coder (e.g., as outlined above). If $m \ll M$ this is [32]

$$\hat{L}_q = MH(\hat{p}) + \frac{m-1}{2} \log M + O\left(\frac{1}{M}\right)$$

The question is how to choose m. It is clear that it is necessary to have unbounded m to catch arbitrary deviations from the uniform distribution, and the following theorem shows this is also sufficient to some extent

Theorem 5. *Suppose that the test data was generated by a continuous distribution. Then the histogram detector satisfies $P_{FA} \to 0$ and $P_D \to 1$ as $M \to \infty$ with suitable choice of $m \to \infty$.*

Proof. Let x be a point where $P(x) \neq \widehat{P}(x)$, lets say $P(x) < \widehat{P}(x)$. Since $P, \widehat{P}$ are continuous, there exists a neighborhood $N(x)$ so that $\forall x' \in N(x) : P(x') < \widehat{P}(x')$. It then follows that $\int_{N(x)} P(x')dx' < \int_{N(x)} \widehat{P}(x')dx'$. We now choose K so large that some histogram bin, call it i, is completely within $N(x)$. Then $p_i \neq \tilde{p}_i$, where $\tilde{p}_i$ is the probability of the i-th bin according to $\widehat{P}$. It follows that $D(p\|\tilde{p}) > 0$. Since $D(p\|\hat{p}) \xrightarrow{D} 0$ under H_0 and $D(p\|\hat{p}) \xrightarrow{D} D(p\|\tilde{p}) > 0$ under H_1, the theorem follows. □

Still, this leaves open how exactly to choose m. Without some type of model selection (e.g., MDL) the only choice is to let m be some function of M. The issue here is that there are contradictory requirements on m. First, as we have to let $m \to \infty$, for large M we get a large m, which is not always advantageous according to Theorem 3. To detect smooth distributions one requires $m \ll M$, while to detect concentrated distributions one would like $m > M$, as argued below. A traditional MDL approach for selecting m (e.g., [16, 26]) would choose m to get a good histogram approximation, e.g., it would not give $m > M$, which is desired for OOD. On the other hand, with (6) or (7) there are no restrictions on m – we will see that it is indeed possible to have an infinite sum in (7). One advantage of the coding approach is that it allows "degenerate" models with $m > M$.

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 small bin size, i.e., large m. 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 an 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 (i.e., uniform) and m is so large that all samples are in different bins. Then the negative log-likelihood according to (13) is

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

(when $m \gg M$ a more efficient coder is to transmit the sequence q^M uncoded) which is unbounded (negative) as $m \to \infty$, i.e., for some m, $\hat{L} + \tau < 0$ 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 6. *For a histogram detector with unbounded number of bins there is a universal coder so that*

1) *For the detector using (6)*
   - *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 (7)*
   - *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.*

Proof. The paper [32] has a coder for q^M for $m > M$, but for $m \gg M$ a more efficient scheme is possible. The encoding is done as follows

1) The encoder encodes the number b of bins that have at least one sample; this requires $\log M$ bits.
2) The encoder encodes which bins have at least one sample. This requires $\log \binom{m}{b}$ bits.
3) The encoder transmits which bin each sample is in; this is now transmission of M samples from a b alphabet source.

The only difference from [32] is that it makes it *explicit* that there might be less than M non-zero bins. Since the decoder then knows exactly the number of non-zero bins, if $b = M$ the encoder does not have to encode the distribution over bins, which is $\frac{1}{M}$ for all bins. But in the coding scheme of [32] the encoder always has to encode the distribution over bins since the decoder does not know the number of non-zero bins from the first step. The conclusion is that the coding schemes has the same redundancy and regret.

The total codelength is

$$\begin{aligned}
\hat{L} &= \log^* m + \log \binom{m}{b} - M \log m + \kappa(M) \\
&\geq \log^* m + mH\left(\frac{b}{m}\right) + \frac{1}{2} \log \frac{m}{8b(m-b)} - M \log m + \kappa(M) \\
&= \log^* m + b \log m - M \log m + \kappa(M)
\end{aligned}$$

Here $\kappa(M)$ denote terms that are bounded in m, e.g. $O(1)$. We see that if $b \leq M - 2$ this can be negative for sufficiently large m no matter what is $\kappa(M)$, but is always strictly positive for $b > M - 2$. We next calculate $P(b \leq M - 2)$, as follows

- We pick $M - 2$ bins. This can be done $\binom{m}{M-2}$ ways.
- We distribute the M samples in the $M - 2$ bins, where some are allowed to be empty. This can be done $\binom{k+k-2-1}{k-3}$ ways.

Then

$$\begin{aligned} P(b \leq M-2) &= \frac{m!}{(m-M+2)!(M-2)!} \cdot \frac{(2M-3)!}{M!(M-3)!} \cdot \frac{1}{m^M} \\ &= 1 \cdot \left(1-\frac{1}{m}\right) \cdots \left(1-\frac{M-2}{m}\right) \frac{1}{m^2} \cdot \frac{(2M-3)!}{M!(M-3)!} \\ &\leq \frac{1}{m^2} f(M) \end{aligned}$$

That means that the union bound over m is finite, which again means that the false alarm probability can be made arbitrary small for sufficiently large τ or M.

Consider instead the MDL-weighted coder. The criterion is

$$\sum_{m=m_0}^{\infty} 2^{-\hat{L}} = \sum_{m=m_0}^{\infty} m^{M-b(m)} 2^{-\log^* m} > C$$

where C is some constant the depends on τ and M. Under the default uniform distribution $b(m) = M$ for sufficiently large $m \geq m_0'$ with probability 1. But the sum $\sum_{m=m_0'}^{\infty} 2^{-\log^* m}$ is finite, and for C large enough the false alarm probability can be made arbitrarily small.

On the other hand, if at least one value is repeated, $b(m) < M$ for arbitrary large $m > m_0'$. But then the sum

$$\sum_{m=m_0}^{\infty} m^{M-b(m)} 2^{-\log^* m} \geq \sum_{m=m_0'}^{\infty} m 2^{-\log^* m} = \sum_{m=m_0'}^{\infty} 2^{-\log^* m + \log m} = \infty$$

Therefore, no matter how large C is, the data will be detected as anomalous.

□

The theorem also shows that the weighted detector (7) can be strictly better than the "model selection" detector (6); in terms of codelength (7) 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 [21]. Intuitively, this also makes sense: If P is very complicated, as universal coder without any knowledge of P would have to estimate P before it can start detecting deviations from P – requiring a large number of samples. On the other hand, it is difficult to use P in a universal source coder. Furthermore, in real-world implementation, P is not known, but found through machine learning. One option for using knowledge of P is to modify the ML distribution through some sort of online learning. However, this is not very feasible, and does not fit into the statistics framework we use. Therefore, the approach we take is to transform any distribution into a standard distribution, and then apply the coding approach. For known P this is alway possible, as follows

Lemma 7 ([30])**.** *For any continuous random variable* $\mathbf{X}$ *there exists an* n*-dimensional uniform random variable* $\mathbf{U}$*, so that* $\mathbf{X} = \check{\mathbf{F}}^{-1}(\mathbf{U})$.

For unknown P we use some type of bidirectional generative networks, described in more detail below. 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 3.

A simple example is given by the histogram approach in Section II-B. If P is very complex (e.g., with high frequency content) and the data is generated by a distribution very close to P, one would need very large m to detect this. On the other hand, if P is known, data can be transformed with the CDF, and even $m = 2$ 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 captures deviation from this model well. As mentioned in ??, the best statistics have the default model as maximum entropy distribution for some value of the statistic. 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 3. 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 $\mathbf{S}$ satysfying these constraints, and the maximum entropy distribution is the Gaussian distribution with covariance matrix $\mathbf{S}$ [9]. The method is called covariance selection [9].

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

$$f(\mathbf{x}) = \frac{\Gamma(n/2)}{\pi^{n/2} r^{n-2}} f_r(r^2) = \frac{\Gamma(n/2)}{\pi^{n/2} (r^2)^{n/2-1}} \exp\left(\lambda_0 - 1 + \sum_{i=1}^{m} \lambda_i t_i(r^2)\right)$$

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,

$$\begin{aligned} f(\mathbf{x}) &= \frac{\Gamma(n/2)}{\pi^{n/2} (r^2)^{n/2-1}} \frac{\beta^\alpha}{\Gamma(\alpha)} (r^2)^{\alpha-1} \exp(-\beta r^2) \\ &= \frac{\Gamma(n/2)}{\pi^{n/2}} \frac{\beta^\alpha}{\Gamma(\alpha)} (r^2)^{\alpha-n/2} \exp(-\beta r^2) \end{aligned} \tag{14}$$

The r^2 statistic detects deviations in the radial distribution of the test data. This can be complemented by detecting deviations in directional distribution. A natural statistic is $\mathbf{T}(\mathbf{x}) = \frac{\mathbf{x}}{\|\mathbf{x}\|}$. The corresponding maximum entropy distribution is the von Mises-Fisher distribution (for unit vectors $\mathbf{u}$)

$$f_d(\mathbf{u}) = C(\kappa) \exp(\kappa \boldsymbol{\mu}^T \mathbf{u})$$

where $C(\kappa)$ is a normalization constant, and

$$\begin{aligned} \boldsymbol{\mu} &= E[\|\mathbf{u}\|] \\ \frac{I_{m/2}(\kappa)}{I_{m/2-1}(\kappa)} &= \|E[\|\mathbf{u}\|]\| \end{aligned}$$

The Gaussian distribution with $\mathbf{\Sigma} = \mathbf{I}$ gives a von Mises-Fisher distribution with $\kappa = 0$. The total distribution for $\mathbf{x}$ is then

$$f(\mathbf{x}) = f_r(\|\mathbf{x}\|^2) f_d\left(\frac{\mathbf{x}}{\|\mathbf{x}\|}\right) \frac{1}{\|\mathbf{x}\|^{n-2}}$$

One can think of it as follows. The (sparse) covariance statistic can detect deviations in covariance, but not a deviation that has a covariance matrix $\mathbf{I}$, but a non-Gaussian distribution. The r^2 statistic and directional statistic can detect deviations from a Gaussian distribution. For example, if the components of $\mathbf{x}$ are IID, but not Gaussian, the covariance matrix is $\mathbf{I}$, but r^2 is not χ^2.

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 unknown covariance matrix $\mathbf{\Sigma}$. Without loss of generality, we assume that everything is zero mean. We calculate CTW-based weighting criterion introduced in (7) to find if a batch of data $\mathbf{x}^M$ belongs to P or is OOD. In order to compute (7), 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. 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. This is described in Section IV-C.
4) Combine codelengths from step–2 and step–3 using CTW principle in (7) to get $\widehat{L}$.
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 [2]; 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 a number of structure learning methods such as graphical lasso (GLasso) [13]; 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 codelength of unique models and as a consequence, we consider unique conditional independence graphs.

The codelength of the universal multivariate Gaussian coder is the sum of two components: 1) $L(G)$, the number of bits needed to describe a conditional independence graph G which can be found using any graph coder described in [1]. 2) $L(\mathbf{x}^M|G)$, the number of bits to describe $\mathbf{x}^M$ using a multivariate Gaussian with conditional independence graph G. Any coding scheme that achieves the universal coding lower bound can be used. We chose predictive MDL [29] because it gives the actual codelength of data:

$$L(\mathbf{x}^M|G) = -\sum_{i=1}^{M-1} \log_2 P\left(\mathbf{x}_{i+1}; \hat{\theta}(\mathbf{x}_1, \ldots, \mathbf{x}_i)\right) \tag{15}$$

where $\hat{\theta}(\mathbf{x}_1, \ldots, \mathbf{x}_i)$ is the maximum likelihood estimate of the covariance matrix computed using the first i samples, $\{\mathbf{x}_1, \ldots, \mathbf{x}_i\}$. The solution of the estimate is constrained such that the corresponding conditional independence graph is G. In [17, Section 17.3], an iterative method was proposed to solve this constrained problem.

Algorithm 1 describes the implementation of the universal multivariate Gaussian coder.

C. Universal Gamma Coder

For $\mathbf{T}(r^2)$ statistics, we consider Gamma distribution as the maximum entropy distribution where χ^2 is a special case of it. In order to encode $\mathbf{x}^M$ using Gamma distribution, we use predictive MDL where we use maximum likelihood to estimate the parameters of Gamma distribution (shape and scale) sequentially.

Algorithm 1 UNIVERSAL MULTIVARIATE GAUSSIAN CODER

Input: IID zero-mean multivariate Gaussian data $\mathbf{x}^M$, GLasso regularization parameters $\{\lambda_1, \ldots, \lambda_K\}$
$\mathcal{G} = \{\}$ ▷ set of unique models
$\mathcal{L} = \{\}$ ▷ set of computed codelengths
for Each regularization parameter $\lambda \in \{\lambda_1, \ldots, \lambda_K\}$. **do**
    Use GLasso with λ to find G_λ
    if $G_\lambda \notin \mathcal{G}$ **then**
        $\mathcal{G} \leftarrow \mathcal{G} \cup G_\lambda$
        Compute $L(G_\lambda)$ using any of graph coders described in [2]
        Compute $L(\mathbf{x}^M|G_\lambda)$ using predictive MDL in (15)
        $L = L(G_\lambda) + L(\mathbf{x}^M|G_\lambda)$
        $\mathcal{L} \leftarrow \mathcal{L} \cup L$
Output: $\mathcal{L}$

D. Experiments on Synthetic Data

Although our approach, maximum entropy coding (MEC), is designed when data is normally distributed, the experiments show that it also performs well when the data comes from distributions that are marginally near-Gaussian. We compared our approach (MEC) to $d-$dimensional KS test (ddKS) method [15], a multi-dimensional two-sample KS test, for 6 different test scenarios of synthetically generated multivariate Gaussian and nearly-Gaussian data as 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, $\mathbf{Ax}$, 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. The test set, $\mathbf{x}^M$, is either generated from the default model, P or from the alternative model, $\widehat{P}$. We repeated the experiment 1000 times; in 500 experiments, data is drawn from default distribution and in 500 experiments, data is generated by alternative distribution. 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 test data do not exactly come from Gaussian distributions.

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	$\mathbf{A}_{ii} = 1, \mathbf{A}_{i,i-1} = \mathbf{A}_{i-1,i} = 0.5$ $\mathbf{A}_{i,i-2} = \mathbf{A}_{i-2,i} = 0.25$	$\mathbf{A}_{ii} = 1, \mathbf{A}_{i,i-1} = \mathbf{A}_{i-1,i} = 0.4$ $\mathbf{A}_{i,i-2} = \mathbf{A}_{i-2,i} = 0.2$ $\mathbf{A}_{i,i-3} = \mathbf{A}_{i-3,i} = 0.2$	$x_i \sim \text{Laplace}(0, i)$
CASE–4	same as CASE–3	same as CASE–3	$x_i \sim \text{Logistic}(0, i)$
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	$x_i \sim \text{StudentT}(i+4)$

V. UNKNOWN DEFAULT MODEL, P

In the previous section, we have shown that our coding-based OOD detection approach works well for multivariate Gaussian and near-Gaussian distributions. However, most real-world data are far from Gaussian. In fact, the default model 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. In this paper, we

TABLE II: AUROC comparing our approach (MEC) to the multi-dimensional KS test in [15] (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

used generative neural networks to transform arbitrary data to multivariate Gaussian. Our requirements are that 1) the transformation, $\mathbf{z} = f(\mathbf{x})$, from data the 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 considered Glow [19], a flow-based generative network. Glow is exactly invertible. Our method to solve OOD detection problem is described in Algorithm 2. Since the default distribution is unknown, training data is used to 1) train the Glow network (i.e., learn $g(\cdot)$ and $f(\cdot)$), 2) learn the multivariate distribution of the latent representation. Theoretically, this distribution can be specified a priori. However, in practice, we found that it is safer to estimate the distribution from data.

The disadvantage of Glow is that the latent space has to be the same dimension as the data space ($m = n$). This poses a limitation on our approach as it requires a number of matrix inversions of not necessarily well-conditioned matrices. We suggested downsampled the data before training Glow.

Algorithm 2 GLOW ALGORITHM

Learn Default Model P
Input: IID training data $\mathbf{x}^N$
Learn functions $f(\cdot)$ and $g(\cdot)$ by training the neural network on $\mathbf{x}^N$. We used Glow; other invertible generative models can also be used
Estimate $\mathbf{\Sigma}_{train}$, the covariance matrix of $\mathbf{x}^N$ from the model giving the shortest codelength in Algorithm 1
Output: $f(\cdot)$ and $g(\cdot)$ of the neural network, learned default model $P = \mathcal{N}(\mathbf{0}, \mathbf{\Sigma}_{train})$

OOD Detection
Input: Outputs from Training, IID test data $\mathbf{x}^M, \tau$
Find the latent representation $\{\mathbf{z}_i = f(\mathbf{x}_i)\}$
Compute $L = -\sum_{i=1}^{M} \log(P(\mathbf{z}_i))$, where P is the learned default model
Compute $\widehat{L}$ by coding $\mathbf{z}^M$ with universal multivariate Gaussian coder and Gamma coder described in Section IV-B and IV-C
Output: Data $\mathbf{x}^M$ is OOD if $\widehat{L} + \tau < L$

VI. EXPERIMENTS ON REAL-WORLD DATA

We considered the digital image dataset MNIST [10] 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 [34].
- **Experiment 2:** Detect if a test set is from MNIST or non-MNIST [4].
- **Experiment 3:** Detect if a test set is from MNIST or synthetically-perturbed MNIST (see Table III and Figure 1 for the description and visualization of the different dataset).

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, $\mu = 0, \sigma = 0.05$

The training data consists of $60,000$ black and white images: $\{\mathbf{x} \in \mathbb{R}^{28\times 28}\}$. In order to use Algorithm 2 in the current implementation, we had to downsample the image from 28×28 to 8×8 pixels. This is because Glow can not do dimensionality reduction, and our approach needs to compute $\mathbf{\Sigma}^{-1}$. Inversion of high-dimensional matrices entails precision issues that can create invalid results; we are working on approximate matrix inversion to address this issue.

We solve OOD detection problem on test datasets of size M. We repeated the experiment 1000 times; in 500 experiments, test data is from the same dataset as the training data), and in 500 experiments, test data is from a different dataset than the training data. It should be mentioned that we can not compare to the ddKS method used in Section IV-D because the data is too high-dimensional for the KS test.

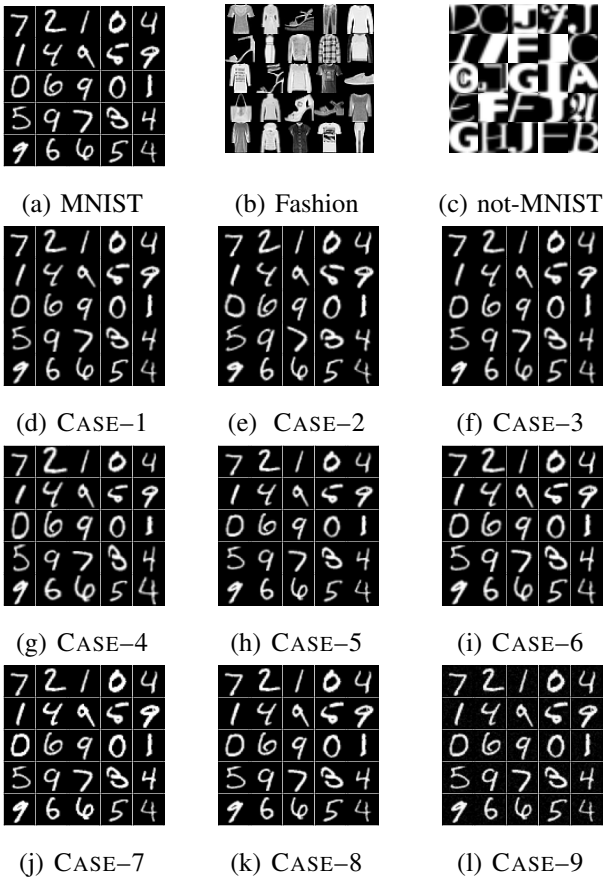
We compared our approach to another Glow-based method called Typicality [24]. We trained our model with the same hyperparameters and settings as [24]. For a fair comparison, we trained and tested Typicality with both downsampled 8×8 images. Table IV shows the AUROC for the experiments using different test set sizes M. Our method has higher performance than Typicality method in all cases except CASE–5. In particular, in CASE–8, Typicality gives an AUROC less than 0.5 (i.e., random guessing). This is because (as noted in the introduction) a likelihood-based approach, which does not learn alternative distribution, may encounter situations where the OOD test data has a high likelihood under the learned default model.

TABLE IV: AUROC for MNIST experiments comparing our MEC to Typicality [24] trained and tested on downsampled images. The best value for each case is boldfaced.

	$\mathbf{M = 50}$		$\mathbf{M = 100}$		$\mathbf{M = 200}$	
	MEC	Typicality	MEC	Typicality	MEC	Typicality
fashion MNIST	**1.000**	**1.000**	**1.000**	**1.000**	**1.000**	**1.000**
not-MNIST	**1.000**	**1.000**	**1.000**	**1.000**	**1.000**	**1.000**
CASE–1	**1.000**	0.995	**1.000**	**1.000**	**1.000**	**1.000**
CASE–2	**1.000**	0.998	**1.000**	**1.000**	**1.000**	**1.000**
CASE–3	**1.000**	**1.000**	**1.000**	**1.000**	**1.000**	**1.000**
CASE–4	**1.000**	0.502	**1.000**	0.505	**1.000**	0.510
CASE–5	0.788	**0.944**	0.855	**0.974**	0.911	**0.980**
CASE–6	**1.000**	**1.000**	**1.000**	**1.000**	**1.000**	**1.000**
CASE–7	**0.978**	0.788	**0.985**	0.849	**0.980**	0.911
CASE–8	**0.883**	0.430	**0.928**	0.380	**0.949**	0.315
CASE–9	**1.000**	**1.000**	**1.000**	**1.000**	**1.000**	**1.000**

VII. CONCLUSION, LIMITATIONS, AND FUTURE WORK

We developed a method for OOD detection based on MDL for the case when the default distribution P is known and when P is unknown. In terms of application, the latter case is more likely to occur. We showed with experiments that our approach outperforms KS test for multivariate Gaussian and near-Gaussian OOD detection problems. We tested our approach on MNIST dataset and compared with another OOD detection



(a) MNIST (b) Fashion (c) not-MNIST

(d) CASE–1 (e) CASE–2 (f) CASE–3

(g) CASE–4 (h) CASE–5 (i) CASE–6

(j) CASE–7 (k) CASE–8 (l) CASE–9

Fig. 1: Different datasets used in MNIST experiments. Note that the synthetically-perturbed images look very similar to the original.

method called Typicality. Our approach has higher performance in most of cases. It also does not give AUROC less than 0.5.

In this paper, we restricted the class of potential data models to only multivariate Gaussian distributions. Since the neural network transforms the default model into Gaussian, this is still powerful enough to detect subtle changes. However, if the test data is from distributions very different from the default distribution, it is unlikely that the transformed distribution will be close to Gaussian. The advantage of using MDL is that we can add any other model/coder into the framework as long as we account for complexity through MDL. In the future, we plan to expand our method with mixture of Gaussians and an extension of the histogram approach in one dimension to higher dimensions.

We also have numerical challenges with our current implementation because our method needs to compute covariance matrix Σ and precision matrix Σ^{-1}. High-dimensional data means that these matrices are large and precision issue means that the inversion results may not be valid covariance/precision matrices. We will investigate fast, approximate matrix inversion algorithms (e.g., [3]) to solve this problem.

REFERENCES

[1] Mojtaba Abolfazli, Anders Høst-Madsen, June Zhang, and Andras Bratincsak. Graph coding for model selection and anomaly detection in gaussian graphical models. In *ISIT'2021, Melbourne, Australia, July 12-20, 2021*, 2021.

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

[3] Michele Benzi and Miroslav Tuma. Orderings for factorized sparse approximate inverse preconditioners. *SIAM Journal on Scientific Computing*, 21(5):1851–1868, 2000.

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

[5] 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*, pages 173–189. Springer, 2018.

[6] Varun Chandola, Arindam Banerjee, and Vipin Kumar. Anomaly detection: A survey. *ACM computing surveys (CSUR)*, 41(3):1–58, 2009.

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

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

[9] A. P. Dempster. Covariance selection. *Biometrics*, 28(1):157–175, 1972. ISSN 0006341X, 15410420. URL http://www.jstor.org/stable/2528966.

[10] Li Deng. The mnist database of handwritten digit images for machine learning research. *IEEE Signal Processing Magazine*, 29(6):141–142, 2012.

[11] P. Elias. Universal codeword sets and representations of the integers. *Information Theory, IEEE Transactions on*, 21(2):194 – 203, mar 1975. ISSN 0018-9448. doi: 10.1109/TIT.1975.1055349.

[12] Giovanni Fasano and Alberto Franceschini. A multidimensional version of the kolmogorov–smirnov test. *Monthly Notices of the Royal Astronomical Society*, 225(1):155–170, 1987.

[13] Jerome Friedman, Trevor Hastie, and Robert Tibshirani. Sparse inverse covariance estimation with the graphical lasso. *Biostatistics*, 9(3):432–441, 2008.

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

[15] 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.

[16] PETER HALL and E. J. HANNAN. On stochastic complexity and nonparametric density estimation. *Biometrika*, 75(4):705–714, 12 1988. ISSN 0006-3444. doi: 10.1093/biomet/75.4.705. URL https://doi.org/10.1093/biomet/75.4.705.

[17] Trevor Hastie, Robert Tibshirani, and Jerome Friedman. The elements of statistical learning data mining, inference, and prediction, 2017.

[18] Anders Høst-Madsen, Elyas Sabeti, and Chad Walton. Data discovery and anomaly detection using atypicality. *IEEE Transactions on Information Theory*, 65(9), September 2019.

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

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

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

[22] Pierre Moulin and Venugopal V. Veeravalli. *Statistical Inference for Engineers and Data Scientists*. Cambridge University Press, 2019.

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

[24] Eric Nalisnick, Akihiro Matsukawa, Yee Whye Teh, and Balaji Lakshminarayanan. Detecting out-of-distribution inputs to deep generative models using typicality, 2019. URL https://arxiv.org/abs/1906.02994.

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

[26] J. Rissanen, T.P. Speed, and B. Yu. Density estimation by stochastic complexity. *IEEE Transactions on Information Theory*, 38(2):315–323, 1992. doi: 10.1109/18.119689.

[27] Jorma Rissanen. Modeling by shortest data description. *Automatica*, pages 465–471, 1978.

[28] Jorma Rissanen. A Universal Prior for Integers and Estimation by Minimum Description Length. *The Annals of Statistics*, 11(2):416 – 431, 1983. doi: 10.1214/aos/1176346150. URL https://doi.org/10.1214/aos/1176346150.

[29] Jorma Rissanen. Stochastic Complexity and Modeling. *The Annals of Statistics*, 14(3):1080 – 1100, 1986. doi: 10.1214/aos/1176350051. URL https://doi.org/10.1214/aos/1176350051.

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

[31] Robert J. Serfling. *Approximation Theorems of Mathematical Statistics*. Wiley-Interscience, 2001.

[32] G.I. Shamir. On the mdl principle for i.i.d. sources with large alphabets. *Information Theory, IEEE Transactions on*, 52(5):1939–1955, May 2006. ISSN 0018-9448. doi: 10.1109/TIT.2006.872846.

[33] F. M J Willems, Y.M. Shtarkov, and T.J. Tjalkens. The context-tree weighting method: basic properties. *Information Theory, IEEE Transactions on*, 41(3):653–664, 1995. ISSN 0018-9448. doi: 10.1109/18.382012.

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

[35] Kenji Yamanishi and Shintaro Fukushima. Model change detection with the mdl principle. *IEEE Transactions on Information Theory*, 64(9):6115–6126, 2018.

[36] 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*, 11(1):19795, 2021.

[37] 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.