

---

# Robust Variational Autoencoders for Outlier Detection and Repair of Mixed-Type Data

---

Simão Eduardo\*<sup>1</sup>

Alfredo Nazábal\*<sup>2</sup> Christopher K. I. Williams<sup>12</sup> Charles Sutton<sup>123</sup>

<sup>1</sup>School of Informatics, University of Edinburgh, UK

<sup>2</sup>The Alan Turing Institute, UK; <sup>3</sup>Google Research

## Abstract

We focus on the problem of unsupervised cell outlier detection and repair in mixed-type tabular data. Traditional methods are concerned only on detecting which rows in the dataset are outliers. However, identifying which cells corrupt a specific row is an important problem in practice, and the very first step towards repairing them. We introduce the Robust Variational Autoencoder (RVAE), a deep generative model that learns the joint distribution of the clean data while identifying the outlier cells, allowing their imputation (*repair*). RVAE explicitly learns the probability of each cell being an outlier, balancing different likelihood models in the row outlier score, making the method suitable for OD in mixed-type datasets. We show experimentally that not only RVAE performs better than several state-of-the-art methods in cell OD and repair for tabular data, but also that it is robust against the initial hyper-parameter selection.

## 1 Introduction

The existence of outliers in real world data is a problem data scientists face daily, so outlier detection (OD) has been extensively studied in the literature (Chandola et al., 2009; Emmott et al., 2015; Hodge and Austin, 2004). The task is often *unsupervised*, meaning that we do not have annotations indicating whether individual cells in the data table are clean or anomalous. Although supervised OD algorithms have been proposed (Lee et al., 2018; An and Cho, 2015; Schlegl

et al., 2017), annotations of anomalous cells are often not readily available in practice. Instead, unsupervised OD attempts to infer the underlying clean distribution, and explains outliers as instances that deviate from that distribution. It is important to focus on the joint distribution over features, because although some outliers can be easily identified as anomalous by considering only the marginal distribution of the feature, many others are only detectable within the context of the other features (section 2.2 of (Chandola et al., 2009)). Recently deep models outperformed traditional ones for tabular data tasks (Klambauer et al., 2017), capturing their underlying structure better. They are an attractive choice for OD, since they have the flexibility to model a wide variety of clean distributions. However OD work has mostly focused on image datasets, repairing dirty pixels instead of cells in tabular data, e.g. (Wang et al., 2017b; Zhou and Paffenroth, 2017; Akrami et al., 2019).

Outliers present unique challenges to deep generative models. First, most work focuses on detecting anomalous data rows, without detecting which specific cells in a row are problematic (Redyuk et al., 2019; Schelter et al., 2018). Work on cell-level detection and repair often focuses on real-valued features, e.g. images (Zhou and Paffenroth, 2017; Wang et al., 2017b; Schlegl et al., 2017), or does not provide a principled way to detect anomalous cells (Nguyen and Vien, 2018). Since focus is on row OD, not enough care is given to cell granularity, which means it is often difficult to properly *repair* the dirty cells, e.g. large number of columns exist or when the data scientist is not a domain expert. Second, tabular data is often *mixed-type*, including both continuous and categorical columns. Although modelling mixed-type data has been explored (Nazabal et al., 2018; Vergari et al., 2019), the difficulty arises when handling outliers. Standard outlier scores are based on the probability that the model assigns to a cell, but these values are not comparable between likelihood models, performing poorly for mixed-type data. Finally, the effect of outliers in unsupervised learning can be insidious. Since deep generative models are highly flexible,

---

\* Joint first authorship.

they are not always robust against outliers (Hendrycks and Dietterich, 2019), overfitting to anomalous cells. When the model overfits, it cannot identify these cells as outliers, because it has modelled them as part of the clean distribution, and consequently, most repair proposals are skewed towards the dirty values, and not the underlying clean ones.

Our main contributions are: (i) Our *Robust Variational Autoencoder (RVAE)*, a novel fully unsupervised deep generative model for cell-level OD and repair for mixed-type tabular data. It uses a two-component mixture model for each feature, with one component for clean data, and the other component that robustifies the model by isolating outliers. (ii) *RVAE* models the underlying clean data distribution by down-weighting the impact of anomalous cells, providing a competitive outlier score for cells and a superior estimate of cell repairs. (iii) We present a hybrid inference scheme for optimizing the model parameters, combining amortized and exact variational updates, which proves superior standard amortized inference. (iv) *RVAE* allows us to present an outlier score that is commensurate across mixed-type data. (v) *RVAE* is robust against the selection of its hyper-parameters, while other OD methods suffer from fine tuning of their parameters to each specific dataset.

## 2 Variational Autoencoders

We consider a tabular dataset  $X$  with  $n \in \{1, \dots, N\}$  instances and  $d \in \{1, \dots, D\}$  features, where each cell  $x_{nd}$  in the dataset can be real (continuous),  $x_{nd} \in \mathbb{R}$ , or categorical,  $x_{nd} \in \{1, \dots, C_d\}$  with  $C_d$  the number of unique categories of feature  $d$ . Cells in the dataset are potentially corrupted with an unknown noising process appropriate for the feature type. The objective in this work is not only detecting the anomalous instances in the dataset, termed *row outliers*, but also determining the specific subset of cells that are anomalous, termed *cell outliers*, proposing potential *repair* values for them.

A common approach to unsupervised OD is to build a generative model  $p(X)$  that models the distribution of clean data. A powerful class of deep generative models are variational autoencoders (VAEs) (Kingma and Welling, 2014), which model  $p(X)$  as

$$p(X) = \prod_{n=1}^N \int d\mathbf{z}_n p(\mathbf{z}_n) p_\theta(\mathbf{x}_n | \mathbf{z}_n) \quad (1)$$

where  $p_\theta(\mathbf{x}_n | \mathbf{z}_n) = \prod_{d=1}^D p_\theta(x_{nd} | \mathbf{z}_n)$  and  $p_\theta(x_{nd} | \mathbf{z}_n)$  is the conditional likelihood of feature  $d$ ,  $\mathbf{z}_n \in \mathbb{R}^K$  is the latent representation of instance  $\mathbf{x}_n$ , and  $p(\mathbf{z}_n) = \mathcal{N}(\mathbf{0}, \mathbf{I})$  is an isotropic multivariate Gaussian prior. To handle mixed-type data, we choose the conditional like-

lihood  $p_\theta(x_{nd} | \mathbf{z}_n)$  differently for each feature type. For real features  $p_\theta(x_{nd} | \mathbf{z}_n) = \mathcal{N}(x_{nd} | m_d(\mathbf{z}_n), \sigma_d)$ , where  $\sigma_d$  is a global parameter. For categorical features  $p_\theta(x_{nd} | \mathbf{z}_n) = f(\mathbf{a}_d(\mathbf{z}_n))$ , where  $\mathbf{a}_d(\mathbf{z}_n)$  is an unnormalized vector of probabilities for each category and  $f$  is the softmax function. All  $m_d(\mathbf{z}_n)$  and  $\mathbf{a}_d(\mathbf{z}_n)$  are parameterized by feed-forward networks.

As exact inference for  $p_\theta(\mathbf{z}_n | \mathbf{x}_n)$  is generally intractable, a variational posterior  $q_\phi(\mathbf{z}_n | \mathbf{x}_n)$  is used; in VAEs this is also known as the encoder. It is modelled by a Gaussian distribution with parameters  $\mu(\mathbf{x}_n)$  and  $\Sigma(\mathbf{x}_n)$

$$q_\phi(\mathbf{z}_n | \mathbf{x}_n) = \mathcal{N}(\mathbf{z}_n | \mu(\mathbf{x}_n), \Sigma(\mathbf{x}_n)) \quad (2)$$

where  $\phi = \{\mu(\mathbf{x}_n), \Sigma(\mathbf{x}_n)\}$  are feed-forward neural networks, and  $\Sigma(\mathbf{x}_n)$  is a diagonal covariance matrix. VAEs are trained by maximizing the lower bound on the marginal log-likelihood called the *evidence lower bound (ELBO)*, given by

$$\mathcal{L} = \frac{1}{N} \sum_{n=1}^N \sum_{d=1}^D \mathbb{E}_{q_\phi(\mathbf{z}_n | \mathbf{x}_n)} [\log p_\theta(x_{nd} | \mathbf{z}_n)] - D_{KL}(q_\phi(\mathbf{z}_n | \mathbf{x}_n) || p(\mathbf{z}_n)), \quad (3)$$

where the neural network parameters of the decoder  $\theta$  and encoder  $\phi$  are learnt with a gradient-based optimizer. When VAEs are used for OD, typically an instance in a tabular dataset is an outlier if the expected likelihood  $\mathbb{E}_{q_\phi(\mathbf{z}_n | \mathbf{x}_n)} [\log p_\theta(\mathbf{x}_n | \mathbf{z}_n)]$  is small (An and Cho, 2015; Wang et al., 2017b).

## 3 Robust Variational Autoencoder (RVAE)

To improve VAEs for OD and repair, we want to make them more robust, by automatically identifying potential outliers during training, so they are downweighted when training the generative model. We also want a cell-level outlier score which is comparable across continuous and categorical attributes. We can achieve both goals by modifying the generative model.

We define here our robust variational autoencoder (RVAE), a deep generative model based on a two-component mixture model likelihood (decoder) per feature, which isolates the outliers during training. RVAE is composed of a clean component  $p_\theta(x_{nd} | \mathbf{z}_n)$  for each dimension  $d$ , explaining the clean cells, and an outlier component  $p_0(x_{nd})$ , explaining the outlier cells. The mixing variable  $w_{nd} \in \{0, 1\}$  acts as a gate to determine whether cell  $x_{nd}$  should be modelled by the clean component ( $w_{nd} = 1$ ) or the outlier component ( $w_{nd} = 0$ ). We define the marginal likelihood of the mixture model

model under dataset  $X$  as<sup>1</sup>

$$p(X) = \prod_{n=1}^N \sum_{\mathbf{w}_n} \int d\mathbf{z} p(\mathbf{z}_n) p(\mathbf{w}_n) p(\mathbf{x}_n | \mathbf{z}_n, \mathbf{w}_n), \quad (4)$$

$$p(\mathbf{x}_n | \mathbf{z}_n, \mathbf{w}_n) = \prod_{d=1}^D p_\theta(x_{nd} | \mathbf{z}_n)^{w_{nd}} p_0(x_{nd})^{1-w_{nd}}, \quad (5)$$

where  $\mathbf{w}_n \in \{0, 1\}^D$  is modelled by a Bernoulli distribution  $p(\mathbf{w}_n) = \prod_{d=1}^D \text{Bernoulli}(w_{nd} | \alpha)$ , and  $\alpha \in [0, 1]$  is a parameter that reflects our belief about the cleanliness of the data. To approximate the posterior distribution  $p(\mathbf{z}, \mathbf{w} | \mathbf{x})$ , we introduce the variational distribution

$$q_{\phi, \pi}(\mathbf{w}, \mathbf{z} | \mathbf{x}) = \prod_{n=1}^N q_\phi(\mathbf{z}_n | \mathbf{x}_n) \prod_{d=1}^D q_\pi(w_{nd} | \mathbf{x}_n), \quad (6)$$

with  $q_\phi(\mathbf{z}_n | \mathbf{x}_n)$  defined in (2) and  $q_\pi(w_{nd} | \mathbf{x}_n) = \text{Bernoulli}(w_{nd} | \pi_{nd}(\mathbf{x}_n))$ . The probability  $\pi_{nd}(\mathbf{x}_n)$  can be interpreted as the predicted probability of cell  $x_{nd}$  being clean. This approximation uses the mean-field assumption that  $\mathbf{w}$  and  $\mathbf{z}$  are conditionally independent given  $\mathbf{x}$ . Finally, the ELBO for the RVAE model can be written as

$$\begin{aligned} \mathcal{L} &= \frac{1}{N} \sum_{n=1}^N \sum_{d=1}^D \mathbb{E}_{q_\phi(\mathbf{z}_n | \mathbf{x}_n)} [\pi_{nd}(\mathbf{x}_n) \log p_\theta(x_{nd} | \mathbf{z}_n)] \\ &+ \frac{1}{N} \sum_{n=1}^N \sum_{d=1}^D \mathbb{E}_{q_\phi(\mathbf{z}_n | \mathbf{x}_n)} [(1 - \pi_{nd}(\mathbf{x}_n)) \log p_0(x_{nd})] \\ &\quad - \frac{1}{N} \sum_{n=1}^N D_{KL}(q_\phi(\mathbf{z}_n | \mathbf{x}_n) || p(\mathbf{z}_n)) \\ &\quad - \frac{1}{N} \sum_{n=1}^N D_{KL}(q_\pi(\mathbf{w}_n | \mathbf{x}_n) || p(\mathbf{w}_n)). \end{aligned} \quad (7)$$

Examining the gradients of (7) helps to understand the robustness property of the RVAE. The gradient of  $\mathcal{L}$  with respect to the model parameters  $\theta$  is given by

$$\nabla_\theta \mathcal{L} = \frac{1}{N} \sum_{n=1}^N \sum_{d=1}^D \pi_{nd}(\mathbf{x}_n) \mathbb{E}_{q_\phi(\mathbf{z}_n | \mathbf{x}_n)} [\nabla_\theta \log p_\theta(x_{nd} | \mathbf{z}_n)]. \quad (8)$$

We see that  $\pi_{nd}(\mathbf{x}_n)$  acts as a weight on the gradient. Cells that are predicted as clean will have higher values of  $\pi_{nd}(\mathbf{x}_n)$ , and so their gradients are weighted more highly, and have more impact on the model parameters. Conversely, cell outliers with low values of  $\pi_{nd}(\mathbf{x}_n)$  will have their gradient contribution down-weighted. A similar formulation can be obtained for the encoder parameters  $\phi$ .

<sup>1</sup>Recall from (Bishop, 2006) (Section 9, page 431) mixture models can also be written in product form using mixing variables  $w_{nd}$ , as we adopt here.

### 3.1 Outlier Model

The purpose of the outlier distribution  $p_0(x_{nd})$  is to explain the outlier cells in the dataset, completely removing their effect in the optimization of the parameters of clean component  $p_\theta$ . For categorical features, we propose using the uniform distribution  $p_0(x_{nd}) = C_d^{-1}$ . Such a uniform distribution assumption has been used in multiple object modelling (Williams and Titsias, 2003) as a way to factor in pixel occlusion. In (Chemudugunta et al., 2006) a similar approach for background words is proposed. For real features, we standardize the features to have mean 0 and standard deviation 1. We use an outlier model based on a broad Gaussian distribution<sup>2</sup>  $p_0(x_{nd}) = \mathcal{N}(x_{nd} | 0, S)$ , with  $S > 1$ . Anomalous cells modelled by the outlier component will be further apart from  $m_d(\mathbf{z}_n)$  relative to clean ones.

Although more complex distributions can be used for  $p_0(x_{nd})$ , we show empirically that these simple distributions are enough to detect outliers from a range of noise levels (Section 4). Furthermore, RVAE can easily be extended to handle other types of features (Nazabal et al., 2018), e.g.: for count features we can use a Poisson likelihood, where the outlier component  $p_0$  would be a Poisson distribution with a large rate; for ordinal features we could have an ordinal logit likelihood, where  $p_0$  can be a uniform categorical distribution.

### 3.2 Inference

We use a hybrid procedure to train the parameters of RVAE that alternates amortized variational inference using stochastic gradient for  $\phi$  and  $\theta$ , and coordinate ascent over  $\pi$ . When we do not amortize  $\pi$ , but rather treat each  $\pi_{nd}(\mathbf{x}_n) \in [0, 1]$  as an independent parameter of the optimization problem, then an exact solution for  $\pi_{nd}(\mathbf{x}_n)$  is possible when  $\phi$  and  $\theta$  are fixed. Optimizing the ELBO (7) w.r.t.  $\pi_{nd}(\mathbf{x}_n)$ , we obtain an exact expression for the optimum<sup>3</sup>:

$$\begin{aligned} \hat{\pi}_{nd}(\mathbf{x}_n) &= g \left( r + \log \frac{\alpha}{1 - \alpha} \right), \quad (9) \\ r &= \mathbb{E}_{q_\phi(\mathbf{z}_n | \mathbf{x}_n)} \left[ \log \frac{p_\theta(x_{nd} | \mathbf{z}_n)}{p_0(x_{nd})} \right] \end{aligned}$$

where  $g$  is the sigmoid function. The first term in (9) represents the density ratio  $r$  between the clean component  $p_\theta(x_{nd} | \mathbf{z}_n)$  and the outlier component  $p_0(x_{nd})$ . When  $r > 1$  it will bias the decision towards assuming the cell being clean, conversely  $r < 1$  it will bias the decision towards the cell being dirty. Such a ratio  $r$  has

<sup>2</sup>This is standard, see (Quinn et al., 2009; Gales and Olsen, 1999)

<sup>3</sup>The derivation of equation (9) is provided in the Supplementary Material (Section 2)

arisen in the literature (Hido et al., 2011; Yamada et al., 2017). The second term in (9) represents our prior belief about cell cleanliness, defined by  $\alpha \in [0, 1]$ . Higher values of  $\alpha$  will skew the decision boundary towards a higher  $\hat{\pi}_{nd}(\mathbf{x}_n)$ , and vice-versa. This coordinate ascent strategy is common in variational inference for conjugate exponential family distributions (Jordan et al., 1999). We term this model RVAE-CVI (Coordinate ascent Variational Inference) below.

Alternatively  $\pi_{nd}(\mathbf{x}_n)$  can be obtained using amortized variational inference. However two problems arise in the process. First an inference gap is introduced by amortization, leading to slower convergence to the optimal solution. Secondly, there might not be enough outliers in the data to properly train a neural network to recognize the decision boundary between clean and dirty cells. We term this model RVAE-AVI (Amortized Variational Inference). RVAE inference is summarized in Algorithm 1, for both the coordinate ascent version (RVAE-CVI) and the amortized version (RVAE-AVI).

---

**Algorithm 1** RVAE Inference
 

---

```

1: procedure RVAE( $\eta$  learning rate,  $M$  batch size,
    $T$  number epochs,  $\alpha$  prior value)
2:   if RVAE-AVI = True then
3:     Define NN parameters:  $\Psi = \{\phi, \theta, \tau\}$ ;  $\triangleright \tau$ 
       is NN params of  $\pi_\tau(\mathbf{x}_n)$  encoder (AVI)
4:   else if RVAE-CVI = True then
5:     Define NN parameters:  $\Psi = \{\phi, \theta\}$ ;
6:     initialize  $\Psi$ ;
7:   for  $1, \dots, T$  do
8:     Sample mini-batches  $\{X_m\}_{m=1}^M \sim p(X)$ ;
9:     Evaluate  $p_\theta(x_{md}|\mathbf{z}_m)$  and  $p_0(x_{md}) \forall m, d$ ;  $\triangleright$ 
       Forward-Pass of Deep AE
10:    if RVAE-CVI = True then
11:      Infer  $\hat{\pi}_{md}, \forall m, d$  using eq. (9);
12:    else if RVAE-AVI = True then
13:      Evaluate encoder  $\pi_\tau(\mathbf{x}_n)$ ;  $\triangleright$ 
       Forward-Pass of NN
14:     $g_\Psi \leftarrow \nabla_\Psi \mathcal{L}(\Psi, \pi(\mathbf{x}_n), \alpha)$  using eq. (7);
15:     $\Psi \leftarrow \Psi + \eta \cdot \text{Adam}(\Psi, g_\Psi)$ 
    
```

---

### 3.3 Outlier Scores

A natural approach to determine which cells are outliers in the data is computing the likelihood of the cells under the trained model. In a VAE, the scores for row and cell outliers would be

$$\begin{aligned}
 \text{Cell: } & -\mathbb{E}_{q_\phi(\mathbf{z}_n|\mathbf{x}_n)} [\log p_\theta(x_{nd}|\mathbf{z}_n)] \\
 \text{Row: } & -\sum_{d=1}^D \mathbb{E}_{q_\phi(\mathbf{z}_n|\mathbf{x}_n)} [\log p_\theta(x_{nd}|\mathbf{z}_n)], \quad (10)
 \end{aligned}$$

where a higher score means a higher outlier probability. However, likelihood-based outlier scores present several problems, specifically for row scores. In mixed type datasets categorical features and real features are modelled by probability and densities distributions respectively, which have different ranges. Often this leads to continuous features dominating over categorical ones. With the RVAE we propose an alternative outlier score based on the mixture probabilities  $\hat{\pi}_{nd}(\mathbf{x}_n)$

$$\text{Cell: } -\log \hat{\pi}_{nd}(\mathbf{x}_n) \quad \text{Row: } -\sum_{d=1}^D \log \hat{\pi}_{nd}(\mathbf{x}_n), \quad (11)$$

where again a higher score means a higher outlier probability. Notice that the row score is just the negative log-probability of the row being clean, given by  $\hat{\pi}_n = \prod_{d=1}^D \pi_{nd}(\mathbf{x}_n)$ . These mixture-based scores are more robust against some features or likelihood models dominating the row outlier score, making them more suitable for mixed type datasets.

### 3.4 Repairing Dirty Cells

Cell repair is related to missing data imputation, however this is a much **harder** task, since positions of anomalous cells are not given, and need to be inferred. After the anomalous cells are identified, a robust generative model allows to impute them given the dirty row directly. In general, repair under VAE-like models can be obtained via maximum a posteriori (MAP)

$$\hat{x}_{nd}^i = \arg \max_{x_{nd}} p_\theta(x_{nd}|\mathbf{z}_n) \quad \mathbf{z}_n \sim q_\phi(\mathbf{z}_n|\mathbf{x}_n^o), \quad (12)$$

where superscript  $i$  denotes imputed or clean cells (depending on context), and  $o$  corresponds to observed or dirty cells. In the case of RVAE,  $p_\theta(x_{nd}|\mathbf{z}_n)$  is the clean component responsible for modelling the underlying clean data, see (5). This reconstruction is akin to robust PCA's clean component. In practice, for real features  $\hat{x}_{nd}^i = m_d(\mathbf{z}_n)$ , the mean of the Gaussian likelihood, and for categorical features  $\hat{x}_{nd}^i = \arg \max_c f(a_{dc}(\mathbf{z}_n))$ , the highest probability category. Other repair strategies are discussed in the Supplementary Material (Section 10).

## 4 Experiments

We showcase the performance of RVAE<sup>4</sup> and baseline methods, for both the task of identifying row and cell outliers and repairing the corrupted cells in the data. Four different datasets from the UCI repository (Lichman, 2013), with a mix of real and categorical features, were selected for the evaluation (see Supplementary Material, Section 1). We compare RVAE with

<sup>4</sup>[https://github.com/sfme/RVAE\\_MixedTypes/](https://github.com/sfme/RVAE_MixedTypes/)

ABDA (Vergari et al., 2019) in a different OD task in the Supplementary material (Section 9).

#### 4.1 Corruption Process

All datasets were artificially corrupted in both training and validation sets. This is a standard practice in OD (Futami et al., 2018; Redyuk et al., 2019; Krishnan et al., 2016; Natarajan et al., 2013), and a necessity in our setting, due to the scarcity of available datasets with labelled cell outliers. No previous knowledge about corrupted cell position, or dataset corruption proportion is assumed. For each dataset, a subset of cells are randomly selected for corruption, following a two step procedure: a) a percentage of rows in the data are selected at random to be corrupted; b) for each of those selected rows, 20% of features are corrupted at random, with different sets of features being corrupted in each select row. For instance, a 5%-20% scenario means that 5% of the rows in the data are randomly selected to contain outliers, and for each of these rows, 20% of the features are randomly corrupted, leading to 1% of cells corrupted overall in the dataset. We will consider for the experiments five different levels of row corruption, {1%, 5%, 10%, 20%, 50%}, leading to five different levels of cells corrupted across the data, {0.2%, 1%, 2%, 4%, 10%}. This corruption process is repeated 5 times (instances). Here, we show results for the aggregate of all datasets for one of those instances, leaving the full disclosure of the results to the Supplementary Material (Section 8).

**Real features:** Additive noise is used as a noising process, with dirty cell values obtained as  $x_{nd}^o \sim x_{nd}^i + \zeta$ , with  $\zeta \sim p_{noise}(\mu, \eta)$ . Four different noise distributions  $p_{noise}$  are explored: *Gaussian noise* ( $\mu = 0, \eta = 5\sigma_d$ ), with  $\sigma_d$  the standard deviation of feature  $d$ ; *Laplace noise* ( $\mu = 0, \eta = \{4\sigma_d, 8\sigma_d\}$ ); *Log-Normal noise* ( $\mu = 0, \eta = 0.75\sigma_d$ ); and a *Mixture of two Gaussian noise components* ( $\mu_1 = -0.5, \eta_1 = 3\sigma_d$ , with probability 0.6 and  $\mu_2 = 0.5, \eta_2 = 3\sigma_d$  with probability 0.4).

**Categorical features:** The noising process is based on the underlying marginal (discrete) distribution. We replace the cell value by a dirty one by sampling from a *tempered categorical distribution*<sup>5</sup> (and excluding the current clean category):

$$x_{ndc}^o \sim \frac{p_c(x_{nd}^i)^\beta}{\sum_{c=1}^{C_d} p_c(x_{nd}^i)^\beta}, \quad (13)$$

with the range  $\beta = [0, 0.5, 0.8]$ . Notice that, when  $\beta = 0$ , the noise process reduces to the uniform distribution, while when  $\beta = 1$ , the noising process follows the marginal distribution.

<sup>5</sup>Also known as *power heuristic* in importance sampling.

#### 4.2 Evaluation metrics

In the OD experiments, we use Average Precision (AVPR) (Everingham et al., 2014; Salton and McGill, 1986), computed according to the outlier scores of each method. AVPR is a measure of area under the precision-recall curve, so higher is better. For cell outliers we report the macro average of the AVPR for each feature in the dataset.<sup>6</sup> In the repair experiments, different metrics are necessary depending on the feature types. For real features, we compute the Standardized Mean Square Error (SMSE) between the estimated values  $\hat{x}_{nd}^i$  and the original ground truth in the dirty cells  $x_{nd}^i$ , normalized by the empirical variance of the ground truth values:  $SMSE_d = \frac{\sum_{n=1}^{N_c^d} (x_{nd}^i - \hat{x}_{nd}^i)^2}{\sum_{n=1}^{N_c^d} (x_{nd}^i - \bar{x}_d)^2}$ , where  $\bar{x}_d$  is the statistical mean of feature  $d$  and  $N_c^d$  is the number of corrupted cells for that feature<sup>7</sup>. For categorical features, we compute the Brier Score between the one-hot representation of the ground truth  $x_{nd}^i$  and the probability simplex estimated for each category in the feature:  $Brier_d = \frac{1}{2N_c} \sum_{n=1}^{N_c^d} \sum_{c=1}^C (x_{ndc}^i - p_c(x_{nd}^o))^2$ , where  $p_c(x_{nd}^o)$  is the probability of category  $c$  for feature  $d$ ,  $x_{ndc}^i$  the one-hot true value for category  $c$ , and  $C$  the number of unique categories in the feature. We used the coefficient  $\frac{1}{2}$  in the Brier score to limit the range to  $[0, 1]$ . We name both metrics in the rest of the paper as SMSE for simplicity, but the correct metric is always used for each type.

#### 4.3 Competing Methods

We compare to several standard OD algorithms. Most methods are only concerned about row OD, whilst only a few can be used for cell OD. For more details on parameter selection and network settings for RVAE and competitor methods, see the Supplementary Material (Section 3).

**Exclusively row outlier detection.** We consider *Isolation Forest (IF)* (Liu et al., 2008), an OD algorithm based on decision trees, which performed quite well in the extensive comparison of (Emmott et al., 2015); and *One Class Support Vector Machines (OC-SVM)* (Chen et al., 2001) using a radial basis function kernel.

**Row and cell outlier detection.** We compare to (i) estimating the *Marginal Distribution* for each feature and using the negative log-likelihood as the outlier score. For real features we fit a Gaussian mixture model with the number of components chosen via Bayesian Information Criterion. The maximum number of compo-

<sup>6</sup>The AVPR macro average is defined as the average of the AVPR for all the features in a dataset.

<sup>7</sup>In our experiments  $\bar{x}_d = 0$  in practice, since the data has been standardized before using any method

nents is set at 40. For categorical features, the discrete distribution is given by the normalized category frequency; (ii) a combination of *OC-SVM* and *Marginal Distribution* for each feature. We use Platt scaling to transform the outlier score of OC-SVM for each row (to obtain log-probability), and then combine it with marginal log-likelihood of each feature. This score, a combined log-likelihood, is then used for cell OD; (iii) VAEs with  $\ell_2$  regularization and outlier scores given by (10); (iv) *DeepRPCA* (Zhou and Paffenroth, 2017), an unsupervised model inspired by robust PCA. The data  $X$  is divided in two parts  $X = R + S$ , where  $R$  is a deep autoencoder reconstruction of the clean data, and  $S$  is a sparse matrix containing the estimated outlier values (see Supplementary Material, Section 3, for further description). Outlier scores for rows are given by the Euclidean norm  $\sqrt{\sum_{d=1}^D |s_{nd}|^2}$ , whilst cell scores are given by  $|s_{nd}|^2$ , where  $s_{nd} \in S$ . (v) A set of *Conditional Predictors* (*CondPred*), where a neural network parametrizing  $p_\theta(\mathbf{x}_n)$  is employed for each feature in the data given the rest<sup>8</sup>. However,  $\ell_2$  regularization is necessary to prevent overfitting, and the model is overall much slower to train than VAE.

**Repair:** We compare to VAE, *DeepRPCA*, *Marginal Distribution* method and *Conditional Predictor* (*CondPred*) method for repairing dirty cells (same model parameters as in OD). We use (12) for all VAE-based methods. For *DeepRPCA* we use  $\hat{X}^i = R$ . For *CondPred* the estimate is  $\hat{x}_{nd}^i = \arg \max_{x_{nd}} p_\theta(x_{nd} | \mathbf{x}_n^o \setminus x_{nd}^o)$ , with  $\mathbf{x}_n^o \setminus x_{nd}^o$  meaning all features in  $\mathbf{x}_n^o$  except  $x_{nd}^o$ . The *Marginal Distribution* method takes  $x_{nd}^o$  and uses as estimate the mean of the closest GMM component in the real line. For RVAE, results using a different inference strategy are provided in Supplementary Material, using pseudo-Gibbs sampling (Rezende et al., 2014).

#### 4.4 Hyperparameter selection for competing methods

In order to tune the hyperparameters for the competing methods, we reserved a validation set with known inlier/outlier labels and ground truth values. This validation set was **not** used by the RVAE method. Thus the performance obtained by the competitor methods is an *optimistic* estimate of their performance in practice. Note also that RVAE-CVI is robust to the selection of its parameter  $\alpha$  in (9), as we will show in Section 4.8. In Figure 1 we compare the performance of the conditional predictor method and VAE, with respect to RVAE-CVI when  $\ell_2$  regularization is not used, and when the best  $\ell_2$  regularization value is used for each dataset. We can observe clearly that a significant gap

<sup>8</sup>It can be seen as a pseudo-likelihood model given by  $p_\theta(\mathbf{x}_n) \approx \prod_d p_\theta(x_{nd} | \mathbf{x}_n \setminus x_{nd})$

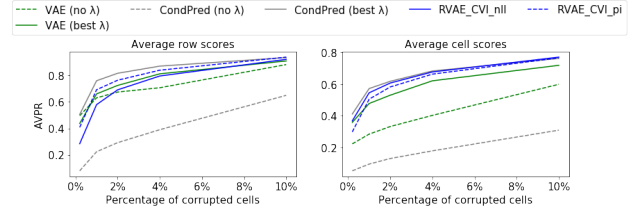


Figure 1: Row and cell OD performance (higher means better) of VAE and CondPred methods without L2 regularization and best choice of  $\lambda$ .

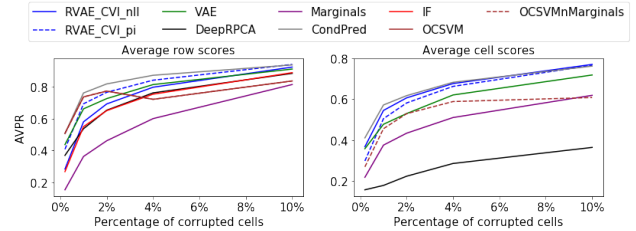


Figure 2: Row and cell OD scores for the average of the four datasets in 5 different cells corruption levels. Left: AVPR at row level. Right: AVPR at cell level.

exists in the performance of these competitor methods when not fine-tuned, making it explicit the dependence of these methods from a labelled validation set. In the rest of the experiments we will use the best possible version of each competitor method.

#### 4.5 Outlier detection

We compare the performance of the difference methods in OD, both at row and cell levels. We focus on Gaussian noise ( $\mu = 0, \eta = 5\sigma_d$ ) for real features and uniform categorical noise, i.e.  $\beta = 0$  in (13), relegating results on other noise processes scenarios to Section 4.7. In Figure 2 we show the average OD performance across all datasets for all OD models in terms of both row (left figure) and cell OD (right figure). We term RVAE-CVI-nll our model with outlier score as defined in (10) and RVAE-CVI-pi our model with outlier score as defined in (11). We relegate RVAE-AVI results to the Supplementary Material (Section 6), since RVAE-AVI is worse than RVAE-CVI in general. Additional results on the OD for each dataset are also available in the Supplementary Material (section 4). In the right figure, We can observe that RVAE-CVI is performing similar to the conditional predictor method on cell OD while being consistently better than the other methods. Additionally, it performs comparatively well in row OD, being similar to the conditional predictor at higher noise levels. We would like to remind that RVAE-CVI does not need a validation set to select its parameters. This means that RVAE-CVI is directly applicable for datasets where no ground truth is available, providing a comparable performance to other methods where pa-

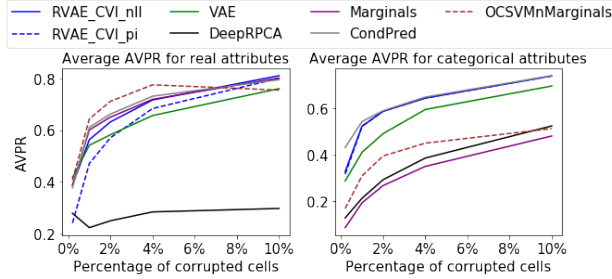


Figure 3: Average AVPR over all the features in the four datasets partitioned by type. Left: AVPR for real features. Right: AVPR for categorical features

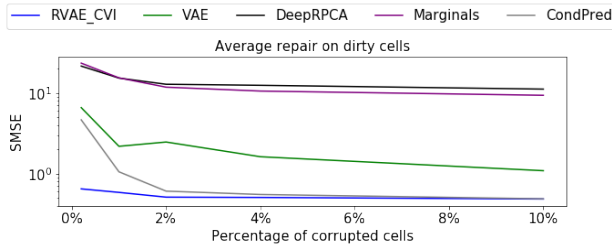


Figure 4: SMSE computed over the dirty cells in all datasets (lower means better). It shows the average over the four datasets for 5 different noised cells percentages. Y-axis is provided in log-scale.

parameter tuning for each dataset is necessary. Figure 2 (left figure) also confirms our hypothesis (Section 3.3) on the proper score to compute row outliers. We can see in the upper figure that RVAE-CVI using scores based on estimate  $\hat{\pi}_{nd}(\mathbf{x}_n)$ , as per score (11), are better for row OD compared to averaging different feature log-likelihoods (10). A further analysis of the OD performance of each model for the different feature types is shown in Figure 3. While the model based on estimating the marginal distribution works well for real features (left figure), it performs poorly on categorical features. Similarly, for the method combining OCSVM and the marginal estimator, it detects outliers better than the other methods in real features and low noise levels, but performs poorly for categorical features. In contrast, RVAE performs comparatively better across different types than the other models, with comparable performance to the conditional predictor.

#### 4.6 Repair

In this section, we compare the ability of the different models to repair the corrupted values in the data. We use the same noise injection process as Section 4.5. Figure 4 shows the average SMSE repair performance across datasets for all models when repairing the dirty cells in the data (more details in the Supplementary Material, Section 5). We can observe that RVAE-CVI outperforms the other models for all the different cell

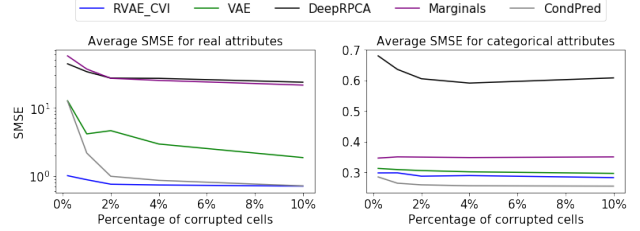


Figure 5: Average SMSE over all the features in the four datasets according to their type. Left: SMSE for real features. Right: SMSE for categorical features

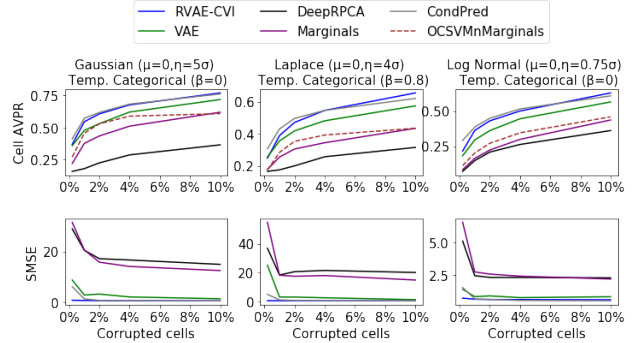


Figure 6: Effect of three different noising processes. Upper figures: average cell OD across datasets. Lower figures: average SMSE on the dirty cells

corruption scenarios, being of particularly significance in lower cell corruption regimes. This is significantly important since all the comparator methods required hyperparameter selection and still performed worse than RVAE-CVI. Also, in Figure 5 we can see the repair performance of different models according to the types of features in the data. Notice that RVAE-CVI is consistently better than the other models across real features while being slightly worse on categorical features.

#### 4.7 Robustness to Noising Processes

Figure 6 shows the performance of the different models across their different combinations of noise processes for all datasets and noise corruption levels (three other noise processes are covered in the Supplementary Material, Section 7). We notice that all the models perform consistently across different types of noise. RVAE-CVI is performing better in repair for low level noise corruption, while providing competitive performance in OD. Also, our choice of outlier models on Section 3.1 does not have a negative effect on the ability of RVAE to detect outliers and repair them. Different noise processes define what is feasible to detect and repair.

#### 4.8 Robustness to hyperparameter values

In this section, we examine the robustness of RVAE inference to the choice  $\alpha$ , and study its effect in both

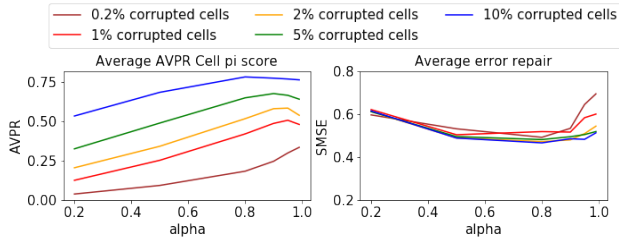


Figure 7: RVAE-CVI performance with different choices for  $\alpha$ . Left: average cell AVPR over the datasets. Right: average repair over dirty cells

OD and repair of dirty cells. We have analyzed values of  $\alpha$  in the set  $\{0.2, 0.5, 0.8, 0.9, 0.99\}$  and evaluated RVAE-CVI in all datasets under all levels of cell corruption and the noising process of Sections 4.5 and 4.6. Figure 7 shows the performance of RVAE-CVI in both OD (left figure) and repair (right figure) across different values of  $\alpha$ . Larger values of  $\alpha$  lead in general to a better OD performance, with a slight degradation when we approach  $\alpha = 1$ . Repair performance is consistent across different choices of  $\alpha$ , but values closer to 0 or 1 lead to a degradation when repairing dirty cells.

## 5 Related Work

There is relevant prior work in the field of OD and robust inference in the presence of outliers, a good meta-analysis study presented in (Emmott et al., 2015). Different deep models have been applied to this task, including autoencoders (Zong et al., 2018; Nguyen and Vien, 2018; Zhou and Paffenroth, 2017), (VAEs) (An and Cho, 2015; Wang et al., 2017b) and generative adversarial networks (GANs) (Schlegl et al., 2017; Lee et al., 2018). In (Nalisnick et al., 2018) the authors show that deep models trained on a dataset assign high likelihood to instances of a quite different dataset, which is problematic in OD. We identify outliers during training rather than from a fully-trained model, down-weighting their effect on parameter learning. Earlier in training, the model had less chance to overfit, so it should be easier to detect outliers.

Most closely related to our model are methods based on **robust PCA (RPCA) and autoencoders**. They focus on unsupervised learning in the presence of outliers, even though most methods need labelled data for hyper-parameter tuning (Candès et al., 2011; Zhou and Paffenroth, 2017; Zong et al., 2018; Nguyen and Vien, 2018; Xu et al., 2018; Akrami et al., 2019). RPCA-based alternatives, often assume that the features are real valued, and model the noise as additive with a Laplacian prior. A problem in RPCA-type models is that often the hyper-parameter that controls the outlier mechanism is dataset dependent and difficult to

interpret and tune. In (Wang et al., 2017b) the authors proposed using a VAE as a recurrent unit, iteratively denoising the images. This iterative approach is reminiscent of the solvers used for RPCA. However their work is not easily extended to mixed likelihood models and suffers from the same problems as VAEs when computing row scores (Section 3.3).

**Robust Variational Inference.** Several methods explore robust divergences for variational learning, under the presence of outliers (Regli and Silva, 2018; Futami et al., 2018), applied to supervised tasks. These divergences have hyper-parameters which are dataset dependent, and can be difficult to tune in unsupervised OD; in contrast, the  $\alpha$  hyperparameter used in RVAE is arguably more interpretable, and experimentally robust to misspecification. Recently a VAE model using one of these divergences at the decoder was proposed for down-weighting outliers (Akrami et al., 2019). However, in contrast to our model, they focused on image datasets and are not concerned with cell outliers. The same hyperparameter tuning problem arises, and it is not clear out to properly extend to categorical features.

**Bayesian Data Reweighting.** Wang et al (Wang et al., 2017a) proposes an approach that raises the likelihood of each observation by some weights and then infer both the latent variables and the weights from corrupted data. Unlike RVAE, these weights are only defined for each instance, so the method cannot detect cell-level outliers. Additionally, the parameters of the model are trained via MCMC instead of variational inference, making them more difficult to apply in the context of deep generative models.

**Classifier Confidence.** Several methods explore adding regularization to improve neural network classifier robustness to outliers (Lee et al., 2018; Hendrycks et al., 2019). However, the regularization hyper-parameters are not interpretable and often require a validation dataset to tune them. Other works like (Hendrycks and Gimpel, 2017), use the confidence of the predicted distribution as a measure of OD.

## 6 Conclusions

We have presented RVAE, a deep unsupervised model for cell OD and repair in mixed type tabular data. RVAE allows robust identification of outliers during training, reducing their contribution to parameter learning. Furthermore, a novel row outlier score for mixed type features was provided. RVAE performs similarly in OD and repair to other methods that heavily rely on fine tuning hyper-parameters with a trusted labelled set. Still, RVAE outperforms or matches the other models in dirty cell repair.



## Acknowledgements

This work was supported by The Alan Turing Institute under the EPSRC grant EP/N510129/1. This work was supported in part by the EPSRC Centre for Doctoral Training in Data Science, funded by the UK Engineering and Physical Sciences Research Council (grant EP/L016427/1) and the University of Edinburgh. We also thank reviewers for fruitful comments and corrections. The author Simão would like to thank helpful discussions with: Afonso Eduardo, Kai Xu and CUP group members in general.

## References

- Haleh Akrami, Anand A. Joshi, Jian Li, and Richard M. Leahy. Robust variational autoencoder, 2019.
- Jinwon An and Sungzoon Cho. Variational autoencoder based anomaly detection using reconstruction probability. *Special Lecture on IE*, 2:1–18, 2015.
- Christopher M Bishop. *Pattern Recognition and Machine Learning*. Springer, 2006.
- Emmanuel J Candès, Xiaodong Li, Yi Ma, and John Wright. Robust principal component analysis? *Journal of the ACM (JACM)*, 58(3):11, 2011.
- Varun Chandola, Arindam Banerjee, and Vipin Kumar. Anomaly detection: A survey. *ACM Computing Surveys (CSUR)*, 41(3):15, 2009.
- Chaitanya Chemudugunta, Padhraic Smyth, and Mark Steyvers. Modeling general and specific aspects of documents with a probabilistic topic model. In *NIPS*, 2006.
- Yunqiang Chen, Xiang Sean Zhou, and Thomas S Huang. One-class SVM for learning in image retrieval. In *ICIP (1)*, pages 34–37. Citeseer, 2001.
- Andrew Emmott, Shubhomoy Das, Thomas Dietterich, Alan Fern, and Weng-Keen Wong. A meta-analysis of the anomaly detection problem. *arXiv preprint arXiv:1503.01158*, 2015.
- Mark Everingham, S. M. Ali Eslami, Luc Van Gool, Christopher K. I. Williams, John M. Winn, and Andrew Zisserman. The pascal visual object classes challenge: A retrospective. *International Journal of Computer Vision*, 111:98–136, 2014.
- Futoshi Futami, Issei Sato, and Masashi Sugiyama. Variational inference based on robust divergences. In *AISTATS*, 2018.
- Mark John Francis Gales and Peder A. Olsen. Tail distribution modelling using the richter and power exponential distributions. In *EUROSPEECH*, 1999.
- Dan Hendrycks and Thomas G. Dietterich. Benchmarking neural network robustness to common corruptions and perturbations. *International Conference on Learning Representations (ICLR)*, abs/1807.01697, 2019.
- Dan Hendrycks and Kevin Gimpel. A baseline for detecting misclassified and out-of-distribution examples in neural networks. *International Conference on Learning Representations (ICLR)*, abs/1610.02136, 2017.
- Dan Hendrycks, Mantas Mazeika, and Thomas G. Dietterich. Deep anomaly detection with outlier exposure. *International Conference on Learning Representations (ICLR)*, abs/1812.04606, 2019.
- Shohei Hido, Yuta Tsuboi, Hisashi Kashima, Masashi Sugiyama, and Takafumi Kanamori. Statistical outlier detection using direct density ratio estimation. *Knowledge and Information Systems*, 26(2):309–336, 2011.
- Victoria Hodge and Jim Austin. A survey of outlier detection methodologies. *Artificial Intelligence Review*, 22(2):85–126, 2004.
- Michael I. Jordan, Zoubin Ghahramani, Tommi S. Jaakkola, and Lawrence K. Saul. An introduction to variational methods for graphical models. *Machine Learning*, 37:183–233, 1999.
- Diederik P. Kingma and Max Welling. Auto-encoding variational bayes. In *International Conference on Learning Representations (ICLR)*. 2014.
- Guenter Klambauer, Thomas Unterthiner, Andreas Mayr, and Sepp Hochreiter. Self-normalizing neural networks. In *NIPS*, 2017.
- Sanjay Krishnan, Jiannan Wang, Eugene Wu, Michael J. Franklin, and Kenneth Y. Goldberg. Activeclean: Interactive data cleaning for statistical modeling. *PVLDB*, 9:948–959, 2016.
- Kimin Lee, Honglak Lee, Kibok Lee, and Jinwoo Shin. Training confidence-calibrated classifiers for detecting out-of-distribution samples. *International Conference on Learning Representations (ICLR)*, abs/1711.09325, 2018.
- M. Lichman. UCI machine learning repository, 2013. URL <http://archive.ics.uci.edu/ml>.
- Fei Tony Liu, Kai Ming Ting, and Zhi-Hua Zhou. Isolation forest. In *2008 Eighth IEEE International Conference on Data Mining*, pages 413–422. IEEE, 2008.
- Eric Nalisnick, Akihiro Matsukawa, Yee Whye Teh, Dilan Gorur, and Balaji Lakshminarayanan. Do deep generative models know what they don’t know? *arXiv preprint arXiv:1810.09136*, 2018.
- Nagarajan Natarajan, Inderjit S. Dhillon, Pradeep Ravikumar, and Ambuj Tewari. Learning with noisy labels. In *NIPS*, 2013.

- Alfredo Nazabal, Pablo M. Olmos, Zoubin Ghahramani, and Isabel Valera. Handling Incomplete Heterogeneous Data using VAEs. *arXiv preprint arXiv:1807.03653*, 2018.
- Minh-Nghia Nguyen and Ngo Anh Vien. Scalable and interpretable one-class SVMs with deep learning and random fourier features. In *ECML/PKDD*, 2018.
- John A. Quinn, Christopher K. I. Williams, and Neil McIntosh. Factorial switching linear dynamical systems applied to physiological condition monitoring. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 31:1537–1551, 2009.
- Sergey Redyuk, Sebastian Schelter, Tammo Rukat, Volker Markl, and Felix Biessmann. Learning to validate the predictions of black box machine learning models on unseen data. In *Proceedings of the Workshop on Human-In-the-Loop Data Analytics*, pages 1–4, 2019.
- Jean-Baptiste Regli and Ricardo Silva. Alpha-beta divergence for variational inference. *CoRR*, abs/1805.01045, 2018.
- Danilo Jimenez Rezende, Shakir Mohamed, and Daan Wierstra. Stochastic backpropagation and approximate inference in deep generative models. *arXiv preprint arXiv:1401.4082*, 2014.
- Gerard Salton and Michael J McGill. Introduction to modern information retrieval. 1986.
- Sebastian Schelter, Dustin Lange, Philipp Schmidt, Meltem Celikel, Felix Biessmann, and Andreas Grafberger. Automating large-scale data quality verification. *Proceedings of the VLDB Endowment*, 11(12):1781–1794, 2018.
- Thomas Schlegl, Philipp Seeböck, Sebastian M Waldstein, Ursula Schmidt-Erfurth, and Georg Langs. Unsupervised anomaly detection with generative adversarial networks to guide marker discovery. In *International Conference on Information Processing in Medical Imaging*, pages 146–157. Springer, 2017.
- Antonio Vergari, Alejandro Molina, Robert Peharz, Zoubin Ghahramani, Kristian Kersting, and Isabel Valera. Automatic bayesian density analysis. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 33, pages 5207–5215, 2019.
- Yixin Wang, Alp Kucukelbir, and David M. Blei. Robust probabilistic modeling with bayesian data reweighting. In *ICML*, 2017a.
- Yu Wang, Bin Dai, Gang Hua, John Aston, and David P Wipf. Green generative modeling: Recycling dirty data using recurrent variational autoencoders. In *UAI*, 2017b.
- Christopher KI Williams and Michalis K Titsias. Learning about multiple objects in images: Factorial learning without factorial search. In *Advances in Neural Information Processing Systems*, pages 1415–1422, 2003.
- Haowen Xu, Wenxiao Chen, Nengwen Zhao, Zeyan Li, Jiahao Bu, Zhihan Li, Ying Liu, Youjian Zhao, Dan Pei, Yang Feng, Jie Chen, Zhaogang Wang, and Honglin Qiao. Unsupervised anomaly detection via variational auto-encoder for seasonal kpis in web applications. In *WWW*, 2018.
- Makoto Yamada, Song Liu, and Samuel Kaski. Interpreting outliers: Localized logistic regression for density ratio estimation. *CoRR*, abs/1702.06354, 2017.
- Chong Zhou and Randy C Paffenroth. Anomaly detection with robust deep autoencoders. In *Proceedings of the 23rd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pages 665–674. ACM, 2017.
- Bo Zong, Qiankun Song, Martin Renqiang Min, Wei Cheng, Cristian Lumezanu, Dae ki Cho, and Haifeng Chen. Deep autoencoding gaussian mixture model for unsupervised anomaly detection. In *International Conference on Learning Representations (ICLR)*, 2018.