Multivariate Triangular Quantiles for Novelty Detection

UNIVERSITY OF WATERLOO

Jingjing Wang, Sun Sun, Yaoliang Yu

{jingjing.wang, sun.sun, yaoliang.yu}@uwaterloo.ca

Summary

✓ **Novelty detection**: detects "novel" or "unusual" samples in data stream.

✓ We extend the univariate quantile function to the multivariate setting through triangular maps.

✓ We present a new framework for neural novelty detection, which recover, unifies and extends many existing approaches.

✓ We apply the multiple gradient descent algorithm to novelty detection and obtain an efficient end-to-end implementation of our framework.

Triangular Quantile Map

- The cumulative distribution function (CDF) F and the quantile function Q of a *univariate* random variable X:

$$F(x) = \Pr(X \le x), \qquad Q(u) = F^{-1}(u) := \inf\{x : F(x) \ge u\}.$$

Generalizing to the multivariate setting? Easy for F, Not obvious for Q.

- If U follows the uniform distribution over the interval $[0,1]$, then $Q(U)$ follows the distribution F.

Definition: Let $\mathbf{X}$ be a random vector in $\mathbb{R}^d$, and let $\mathbf{U}$ be uniform over the unit hypercube $[0,1]^d$. We call an **increasing triangular** map $\mathbf{Q} : [0,1]^d \to \mathbb{R}^d$ *the* triangular quantile map of $\mathbf{X}$ if $\mathbf{Q}(\mathbf{U}) \sim \mathbf{X}$.

General Framework for Novelty Detection

Let $\mathbf{f} : \mathbb{R}^d \to \mathbb{R}^m$ be a feature map and $\mathbf{X}$ a random sample from the unknown density p. We propose to learn the density $\mathbf{f}_{\#}p$ of the latent random vector $\mathbf{Z} = \mathbf{f}(\mathbf{X})$.

Problem Formulation

We learn the feature map $\mathbf{f}$ and the TQM $\mathbf{Q}$ *simultaneously* by minimizing the following objective:

$$\min_{\mathbf{f},\mathbf{Q}} \ \gamma \mathsf{KL}(\mathbf{f}_{\#}p \| \mathbf{Q}_{\#}q) + \lambda \ell(\mathbf{f}) + \zeta g(\mathbf{Q}), \tag{1}$$

- g: potential constraints on the increasing triangular map $\mathbf{Q}$
- ℓ: loss associated with learning the feature map $\mathbf{f}$
- q: a fixed reference density (e.g., the uniform density over $[0,1]^m$)
- $\zeta, \lambda, \gamma \ge 0$: regularization constants

W.l.o.g. we parameterize the TQM as

$$\mathbf{Q} = \mathbf{T} \circ \mathbf{\Phi}^{-1}, \quad \text{where} \tag{2}$$

- $\mathbf{\Phi} = (\Phi, \ldots, \Phi)$ with Φ the CDF of standard univariate Gaussian
- $\mathbf{T} : \mathbb{R}^d \to \mathbb{R}^d$ an increasing triangular map.

Thresholding Rules

Once $\mathbf{f}$ and $\mathbf{Q}$ are estimated, we can detect novel test samples by either thresholding the density function of the latent variable $\mathbf{Z}$ or thresholding its TQM.

- **by thresholding density**:

$$p_{\mathbf{Z}}(\mathbf{z}) = 1/|\mathbf{Q}'(\mathbf{Q}^{-1}(\mathbf{z}))| = \frac{1}{|\mathbf{T}'(\mathbf{T}^{-1}(\mathbf{z}))|} \cdot \prod_{j=1}^{m} \varphi(\Phi^{-1}(z_j)), \quad \text{where} \quad \varphi = \Phi'. \tag{3}$$

We declare a test sample $\tilde{\mathbf{X}}$ to be "novel" if

$$\log|\mathbf{T}'(\mathbf{T}^{-1}(\mathbf{f}(\tilde{\mathbf{X}})))| + \mathbf{1}^\top \Phi^{-2}(\mathbf{f}(\tilde{\mathbf{X}}))/2 \ge \tau, \tag{4}$$

Note: $\mathbf{T}$ is increasing triangular, efficient to compute $\mathbf{T}^{-1}$ and $|\mathbf{T}'|$

Downside: τ is usually difficult to guess.

- **by thresholding TQM**:

let $N \subseteq [0,1]^m$ be a subset whose (uniform) measure is $1-\alpha$ for some $\alpha \in (0,1)$, $\tilde{\mathbf{X}}$ is "novel" if

$$\mathbf{Q}^{-1}(\mathbf{f}(\tilde{\mathbf{X}})) \notin N. \tag{5}$$

For instance, we can choose N to be the cube centered at $(1/2, \ldots, 1/2)$ and with side length $(1-\alpha)^{1/m}$, in which case

$$\mathbf{Q}^{-1}(\mathbf{f}(\tilde{\mathbf{X}})) \notin N \iff \|\mathbf{Q}^{-1}(\mathbf{f}(\tilde{\mathbf{X}})) - \tfrac{1}{2}\|_\infty \ge (1-\alpha)^{1/m}/2. \tag{6}$$

Upside: Control Type-I error

Experiment

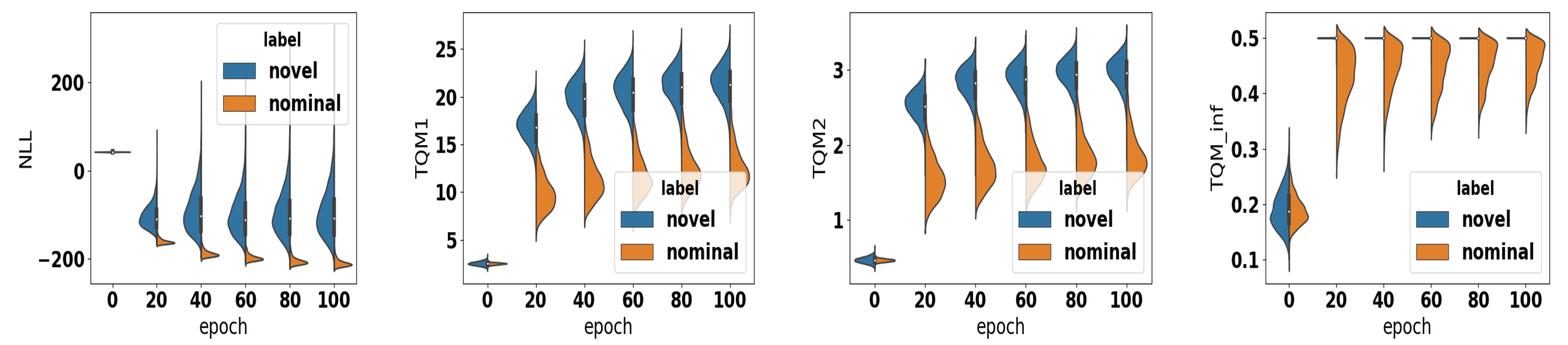

Figure: Violin plots: 1) DEN; 2) TQM1; 3) TQM2; and 4) TQM∞.

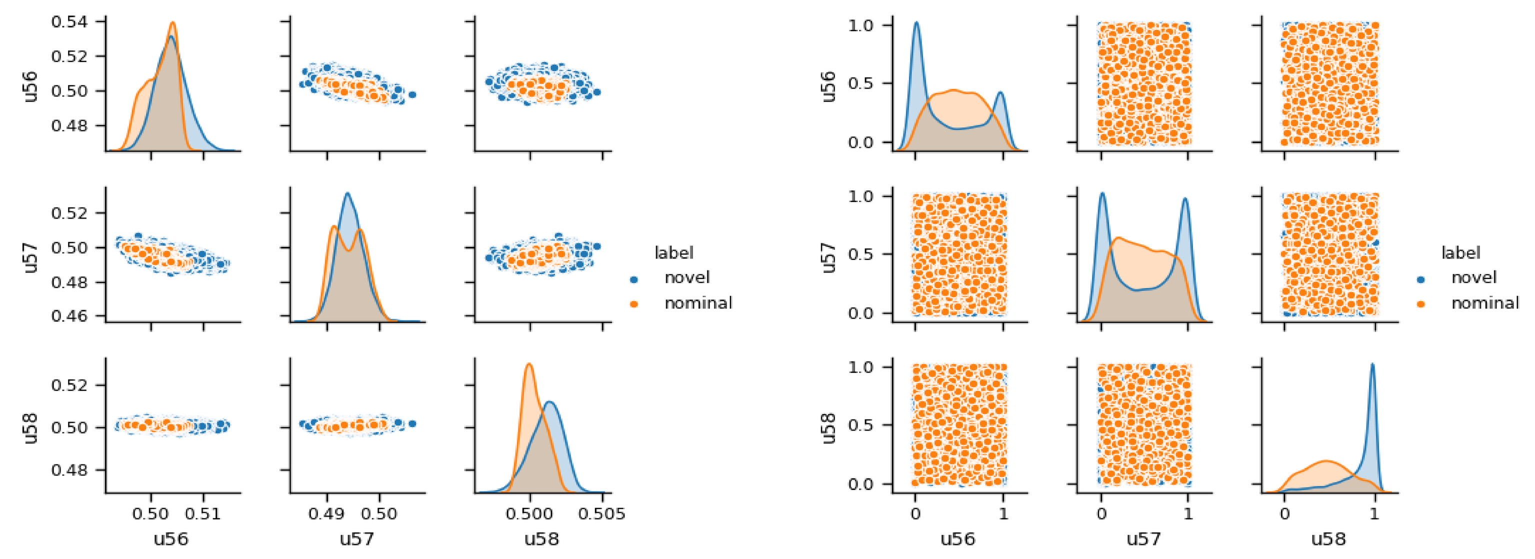

Figure: Marginal and joint distributions of pre-image in $[0,1]$ of test data (dimension: $56, 57$, and 58). 1) distributions at initialization; and 2) distributions at 1000 epochs of training.

Estimating TQM Using Deep Networks

Our framework (1) has three components which we implement as follows:

- **Feature Extractor for f**:

a deep autoencoder [2] composed by one encoder $\mathbf{Z} = \mathcal{E}(\mathbf{X}; \boldsymbol{\theta}_E)$ and one decoder $\hat{\mathbf{X}} = \mathcal{D}(\mathbf{Z}; \boldsymbol{\theta}_D)$. The Euclidean reconstruction loss:

$$\ell(\mathbf{f}) = \ell(\boldsymbol{\theta}_E, \boldsymbol{\theta}_D) = \sum_{i=1}^{n} \|\mathbf{X}_i - \hat{\mathbf{X}}_i\|^2. \tag{7}$$

- **Flow-based Neural Density Estimator for Q**:

the sum-of-squares (SOS) flow [1]

- **KL-divergence term**:

approximated empirically using the given sample $\{\mathbf{X}_1, \ldots, \mathbf{X}_n\}$. Upon dropping irrelevant constants we reduce the KL term in (1) to:

$$\min_{\boldsymbol{\theta}_Q} \sum_{i=1}^{n} \Big[\log|\mathbf{Q}'(\mathbf{Q}^{-1}(\mathbf{f}(\mathbf{X}_i)))| - \log q(\mathbf{Q}^{-1}(\mathbf{f}(\mathbf{X}_i)))\Big] \tag{8}$$

Objective Function

$$\min_{\boldsymbol{\theta}} \sum_{i=1}^{n} (1-\lambda) \underbrace{\Big[\log|\mathbf{T}'(\mathbf{T}^{-1}(\mathbf{Z}_i))| + \|\mathbf{T}^{-1}(\mathbf{Z}_i)\|^2/2\Big]}_{\text{negative log-likelihood } h(\mathbf{X}_i;\boldsymbol{\theta})} + \lambda \underbrace{\|\mathbf{X}_i - \mathcal{D}(\mathbf{Z}_i; \boldsymbol{\theta}_D)\|^2}_{\text{reconstruction loss } \ell(\mathbf{X}_i;\boldsymbol{\theta})},$$

$$\text{where} \quad \mathbf{Z}_i = \mathcal{E}(\mathbf{X}_i; \boldsymbol{\theta}_E) \tag{9}$$

Multiple Gradient Descent Algorithm

We cast the two competing objectives in (9) as multi-objective optimization. Using the multiple gradient descent algorithm (MGDA) [3]

Let gradient descent decide what λ to use in each iteration!

$$\lambda_t = \arg\min_{0 \le \lambda \le 1} \left\| \sum_{i \in I} (1-\lambda) \nabla h(\mathbf{X}_i; \boldsymbol{\theta}_t) + \lambda \nabla \ell(\mathbf{X}_i; \boldsymbol{\theta}_t) \right\|^2 \tag{10}$$

$$= \min\left\{1, \max\left\{0, \frac{\langle \nabla h_I - \nabla \ell_I, \nabla h_I \rangle}{\|\nabla h_I - \nabla \ell_I\|^2}\right\}\right\}, \tag{11}$$

where $I \subseteq \{1, \ldots, n\}$ is a minibatch of samples

References

[1] Jaini, P. and Yu, Y. Sum-of-Squares Polynomial Flow. *PMLR'19*

[2] Abati D, Porrello A, Calderara S, et al. Latent Space Autoregression for Novelty Detection. *CVPR'19*

[3] Désidéri, Jean-Antoine MGDA II: A direct method for calculating a descent direction common to several criteria. *Doctoral dissertation'12*