Report for Robust Mean and Covariance Estimate by GAN

CAO, Yang
Department of Mathematics
ycaoau@ust.hk

ZENG, Wenqi
Department of Mathematics
wzengad@ust.hk

Abstract

In this project we performed robust mean and covariance matrix estimation using TV-GAN and JS-GAN on financial dataset with different assumptions on distribution. Considering the optimization difficulties of TV-GAN, we mainly conduct subsequent task of outlier detecting on JS-GAN and verified the impact of hidden layers. Also, we utilized the covariance matrix estimated by JS-GAN and combined it with Robust PCA. The results are consistent with theoretical derivations and meaningful in practical application.

1 Introduction

Robust statistics performs well for data drawn from a wide range of distributions especially for those who are not normal. Robustness means these statistics are not unduly affected by outliers and still produce good results when there are departures from parametric distribution. However, robust statistics like Tukey's median [6] and other depth-based estimators can be computational expensive and thus pose challenge to reach.

The initially proposed GAN [3] is based on game theory and includes one discriminator and one generator. The discriminator will output the probability indicating whether the input comes from training dataset while the generator will try to generate fake samples to deceive the discriminator. More generally, we understand GAN from the perspective of sample probability distribution. The generator of GAN implicitly defines a probability distribution and generates samples based on this distribution. The goal of GAN is to drive the implicit probability distribution defined by the generator to be close to that of the training sample set.

f-GAN [5] is a variant of GAN with benefits of various choices of divergence functions. Under the framework of f-GAN, these robust estimators who are maximizers of depth functions can be derived as the minimizers of variational lower bounds between the empirical distribution and generated distribution. This observation builds a bridge between developing variational lower bounds by neural network approximations and robust estimation, and provides a novel perspective on obtaining robust statistics.

2 Background

Robust estimation has become an important topic in statistics where presence of an unknown contamination poses both statistical and computational challenges. We consider the setting of Huber's model [4], the data are generated by

$$X_1, \cdots, X_n \sim (1-\epsilon)P_\theta + \epsilon Q \tag{1}$$

and our goal is to estimate the parameter θ. Specially, consider $P_\theta = N(\mu, I_p)$, the traditional robust maximum likelihood with coordinate wise median can not get a optimum estimation for μ. Therefore, we would need Tukey's median [6] to get the optimum value

$$\hat{\theta} = \arg\max_{\eta \in \mathbb{R}^p} \inf_{||u||=1} \frac{1}{n} \Sigma_{i=1}^n \mathbb{I}\{u^T X_i > u^T \eta\} \tag{2}$$

However, this method would be extremely computational expensive and would be unacceptable when the dimension is high. Hence, we would need some framework that have good performance in practice to approximate the Tukey's median.

2.1 Generative Adversarial Networks

The Generative Adversarial Networks (GANs) are a class of unsupervised machine learning algorithms. The generator (G) of GAN learns to map from random noise to some data distribution, and the discriminator (D) of GAN is a classifier which distinguish the generated data and the real data. The traning objective is to confuse the discriminator, which means increase its error rate. Then for value function $V(G, D)$, the target is actually solving

$$\min_G \max_D V(G, D) \tag{3}$$

Denote the distribution of real data as $P(x)$ with density $p(x)$, and the distribution of generated data as $Q(x)$ with density $q(x)$, our goal can be rewritten as

$$\min_T D(P||Q) \tag{4}$$

where D is some divergence function.

2.2 f-GAN

For any f where $f(1) = 0$ and f is convex, f-GAN [1] define the divergence function as

$$D_f(P||Q) = \int_x q(x) f(\frac{p(x)}{q(x)}) dx \tag{5}$$

As f is convex

$$D_f(P||Q) = \int_x q(x) f(\frac{p(x)}{q(x)}) dx \geq f(\int_x q(x) \frac{p(x)}{q(x)} dx) = f(1) = 0 \tag{6}$$

Let f^* be the convex conjugate of f, defined as

$$f^*(x) = \sup_{t \in \text{dom}_f} (xt - f(t)) \tag{7}$$

we have the property that f^* is convex, and

$$f(x) = \sup_{t \in \text{dom}_{f^*}} (xt - f^*(t)) \tag{8}$$

Consider the expression of $D_f(P||Q)$ and substitute t with $t = D(x)$, we can have

$$D_f(P||Q) = \int_x q(x) f(\frac{p(x)}{q(x)}) dx \tag{9}$$

$$= \int_x q(x) \sup_{t \in \mathrm{dom}_{f^*}} (\frac{p(x)}{q(x)} t - f^*(t)) dx \tag{10}$$

$$\geq \int_x p(x) D(x) dx - \int_x q(x) f^*(D(x)) dx \tag{11}$$

therefore

$$D_f(P||Q) \geq \sup_D \{\mathbb{E}_{x \sim P}[D(x)] - \mathbb{E}_{x \sim Q} f^*(D(x))\} \tag{12}$$

Note that the equation holds when $D = f'(\frac{p(x)}{q(x)})$. Then the training goal for GAN can be written as

$$G^* = \arg\min_G \max_D \{\mathbb{E}_{x \sim P}[D(x)] - \mathbb{E}_{x \sim Q} f^*(D(x))\} \tag{13}$$

2.3 f-Learning

The sample version of variational lower bounds in above derivation of f-GANs leads to learning method [1] [2] below

$$\hat{P} = \arg \inf_{Q \in \mathbb{Q}} \sup_{p \in \mathbb{P}} [\frac{1}{n} \Sigma_{i=1}^n f'(\frac{p(X_i)}{q(X_i)}) - \mathbb{E}_Q f^*(f'(\frac{p(X)}{q(X)}))] \tag{14}$$

Specially, if we choose $f = (x-1)_+$, $\mathbb{Q} = \{N(\eta, I_p) : \eta \in \mathbb{R}^p\}$ and $\mathbb{P} = \{N(\tilde{\eta}, I_p) : ||\tilde{\eta} - \eta|| \leq r\}$, we would get TV-GAN. And if we choose $f = x \log x - (x+1) \log \frac{x+1}{2}$ while preserving $\mathbb{P}$ and $\mathbb{Q}$, we would get JS-GAN.

2.4 Robust PCA

Then we would like to approximate the covariance matrix, a classical way to do that is to do Principle Component Analysis, which look for a matrix decomposition

$$X = L + E \tag{15}$$

Here, $X \in \mathbb{R}^{p \times n}$ is the data matrix, L is a low-rank matrix, and E with a small Frobenius norm is the error matrix. However, PCA is very sensitive to outliers, that is why we need some more robust algorithm.

Robust PCA aims to find the following decomposition

$$X = L + S \tag{16}$$

where L is matrix with low rank, and S is sparse. That is to solve

$$\min ||X - L||_0 \tag{17}$$

$$s.t. \mathrm{rank}(L) \leq k \tag{18}$$

However, directly solving this decomposition would be NP-hard. In practice, we usually relax the restriction a little bit. The simplest convex relaxation is to relax $||X - L||_0$ to $||X - L||_1$, and $\mathrm{rank}(L)$ to $\Sigma_i \sigma_i(L)$, where σ_i are the singular values of L.

3 Experiment

Our dataset consists of stock price from 01/01/2015 to 01/04/2020 on selected 50 companies from representative industries like energy, health care, utility, etc. The first experiment on financial data is estimation of robust mean and covariance matrix. In order to facilitate visualization, we have drawn the distribution of the real data and the generator for comparison. We also testified the hidden layer issue of JS-GAN referred in lecture with experimental proof. Next, we used the robust estimator by JS-GAN under student-t assumption into the detection of outliers, of which values are outside 99.75% of the distribution are considered as outliers, as well as robust PCA.

3.1 Robust Mean and Covariance Matrix Estimation

One interesting thing to notice is the importance of hidden layers in JS-GAN. If we drop hidden layers in discriminator, then robust estimators will no longer be robust resulting from failure of distinguishing distribution P and Q by JS-GAN. More details about relevant mathematical derivation can be found in paper [1]. We use a simple simulation to verify this phenomenon where Huber's model is set to $0.8 * N(1, 1) + 0.2 * N(t, 1)$. We can observe the estimator under zero hidden layers in discriminator comes close to $0.8 + 0.2t$ (grey line) while the estimator with two hidden layers turns out to be more robust to true mean 1.

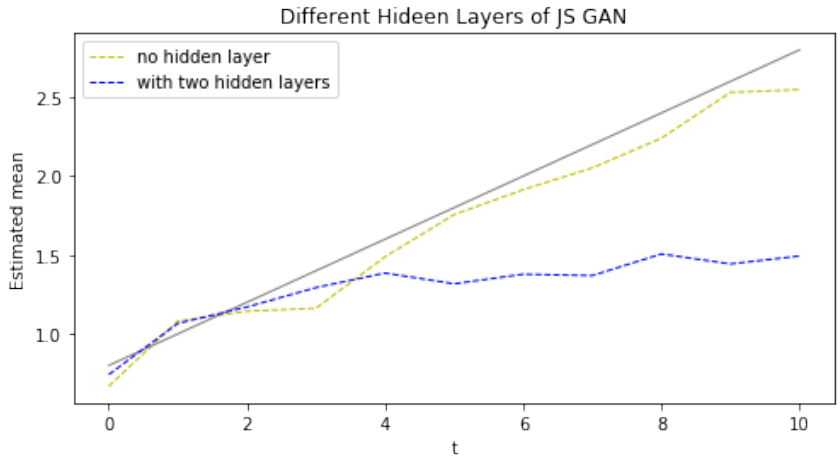

Figure 1: Comparison of different hidden layers in JS-GAN

Under assumptions on real data distribution in the above formula in several cases, namely Gaussian, Student-t, general Elliptical distribution, we calculate the robust estimators by TV-GAN and JS-GAN respectively and get the distribution as Figure 2 . For JS-GAN, Student-t seems to have better fitting performance compared with Gaussian and general Elliptical. However, TV-GAN performs not so well especially fitted by Elliptical distribution where it may suffer from optimization difficulties and fails to reach the saddle point because of far distance between true distribution and contamination distribution. Considering the unstable performance of TV-GAN under this situation, we turned to use JS-GAN with hidden layers in later experiments.

3.2 Outlier Detection by Discriminator Distribution Values

We chose JS-GAN fitted by Student-t distribution to find extreme samples whose discriminator values are larger than 99.75% percentile. In order to further understand the meaning of outliers, we found the stock price of the corresponding date shown in Figure 3 . The vast majority of outliers are concentrated in the stock market shocks at the end of 2019 when the spread of the coronavirus disease made global financial market greatly turbulent and experience the worst trading since 2008 financial crisis. It can be seen that the outliers detected by JS-GAN have been verified instructive in the actual stock market.

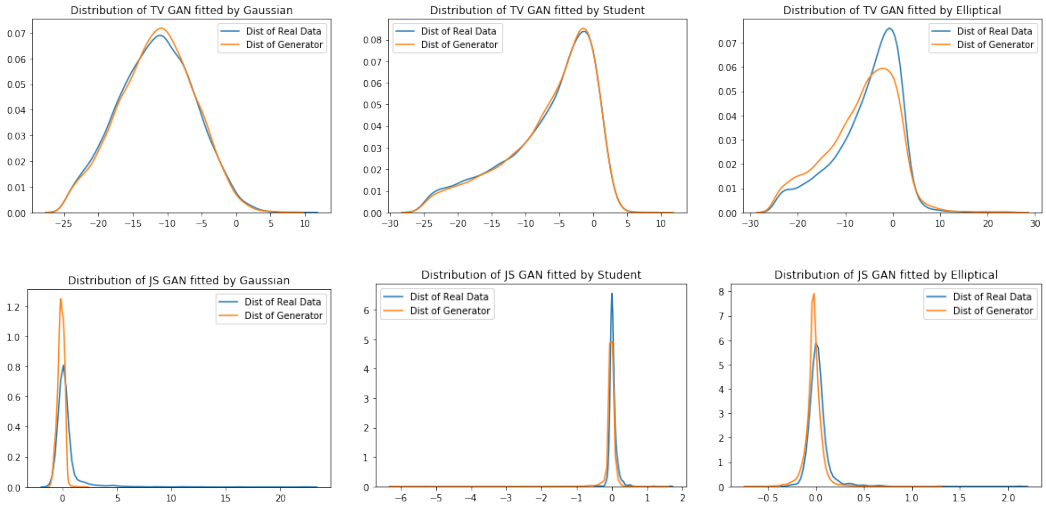

Figure 2: TV-GAN, JS-GAN fitted by Gaussian, Student-t, Elliptical distributions

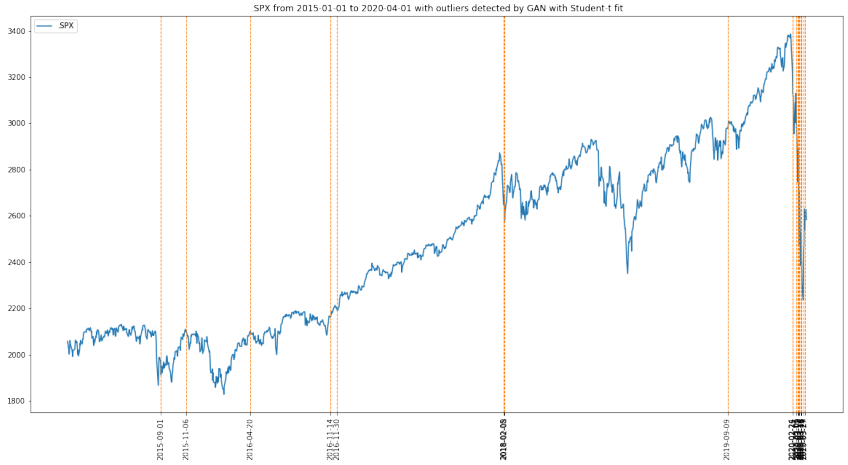

Figure 3: Outlier Detection by JS-GAN

3.3 Robust PCA

We fed standard covariance matrix to vanilla PCA and robust covariance matrix estimated by JS-GAN to Robust PCA and obtained visualization. Vanilla PCA has explaining ratio 0.481 and 0.068 for the first and second principle component respectively in Figure 4 while in Robust PCA case are 0.290 and 0.107 shown in Figure 5 . In top figure the original PCA is largely influenced by a few companies with instability, while the robust PCA in the bottom figure has a stable visualization with more spreads on the 50 companies by attenuating outliers. Since we selected representative companies in the ten industries, a decentralized distribution is more in line with our understanding.

4 Conclusion

In this project, we explored the connection of f-GAN and statistically optimal robust estimators under the framework of f-learning and revised how the depth function can be view as the variational lower bounds. As beginners to GAN, we tried to utilize tools in training GAN to obtain robust statistics and thus got enabled to see the wide adaption of GAN except for common fields such as image generation.

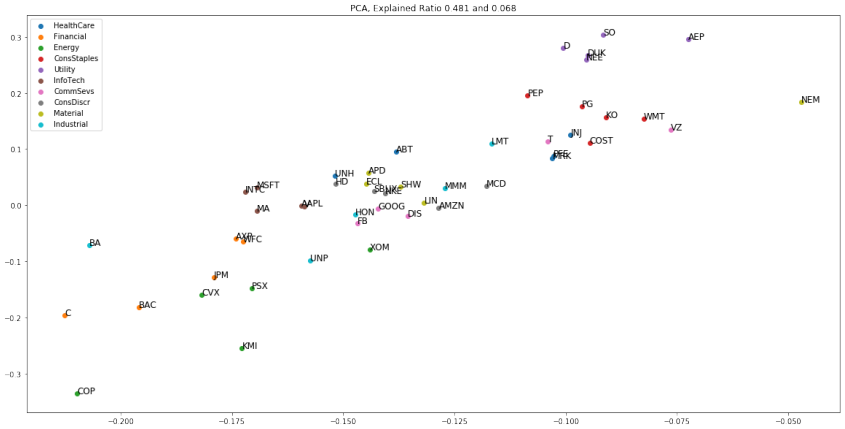

Figure 4: Visualization of PCA

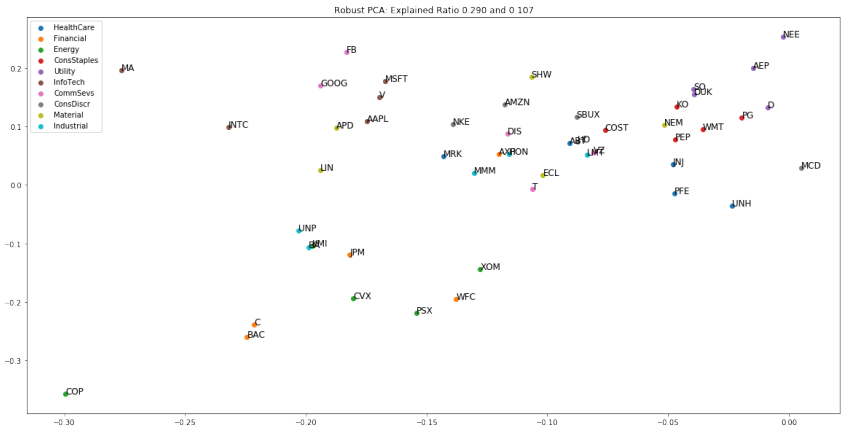

Figure 5: Visualization of Robust PCA

Although we did not encounter common problems such as mode collapse during training GAN, we sometimes have difficulty determining whether the model has converged. Also we should notice the necessity of hidden layers of neural network structures used in the GAN training. A neural network class without hidden layer is equivalent to matching linear features, and is thus not suitable for robust estimation.

5 Contribution

CAO, Yang: Discussion, Theoretic Derivation, Writing
ZENG, Wenqi: Discussion, Experiment, Writing

References

[1] C. Gao, J. Liu, Y. Yao, and W. Zhu. Robust estimation and generative adversarial nets. *arXiv preprint arXiv:1810.02030*, 2018.

[2] C. Gao, Y. Yao, and W. Zhu. Generative adversarial nets for robust scatter estimation: A proper scoring rule perspective. *CoRR*, abs/1903.01944, 2019.

[3] I. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, A. Courville, and Y. Bengio. Generative adversarial nets. In *Advances in neural information processing systems*, pages 2672–2680, 2014.

[4] P. J. Huber. Robust estimation of a location parameter. In *Breakthroughs in statistics*, pages 492–518. Springer, 1992.

[5] S. Nowozin, B. Cseke, and R. Tomioka. f-gan: Training generative neural samplers using variational divergence minimization. In *Advances in neural information processing systems*, pages 271–279, 2016.

[6] J. W. Tukey. Mathematics and the picturing of data. In *Proceedings of the International Congress of Mathematicians, Vancouver, 1975*, volume 2, pages 523–531, 1975.