Mixture Density Conditional Generative Adversarial Network Models (MD-CGAN)

Jaleh Zand [1 2] **Stephen Roberts** [1 2]

Abstract

Generative Adversarial Networks (GANs) have gained significant attention in recent years, with particularly impressive applications highlighted in computer vision. In this work, we present a Mixture Density Conditional Generative Adversarial Model (MD-CGAN), where the generator is a Gaussian Mixture model, with a focus on time series forecasting. Compared to examples in vision, there have been more limited applications of GAN models to time series. We show that our model is capable to estimate a probabilistic posterior distribution over forecasts and that, in comparison to a set of benchmark methods, the MD-CGAN model performs well, particularly in situations where noise is significant in the time series.

1. Introduction

Generative Adversarial Networks have been one of the raft of breakthroughs in Deep Learning methods in recent years. Several different variations of the model have been introduced since the method was first introduced by (Goodfellow et al., 2014). One of the most popular variations of the work is the Conditional Generative Adversarial Networks (CGAN), (Mirza & Osindero, 2014), in which the generator and discriminator are both conditioned on some observed information. Within time series forecasting, future values are conditioned on information observed from the past - either from the time series itself, exogenous data or a combination of the two. This makes the CGAN approach particularly useful for time series prediction.

Most applications of (C)GAN have been within computer vision and, to a lesser extent, in natural language processing.

[1]Machine Learning Research Group, University of Oxford, Oxford, UK [2]Oxford-Man Institute of Quantitative Finance, University of Oxford, Oxford, UK. Correspondence to: Jaleh Zand <jz@robots.ox.ac.uk>, Stephen Roberts <sjrob@robots.ox.ac.uk>.

The 36th International Conference on Machine Learning, Long Beach, California, 2019.

The literature on the application of GAN models to problems associated with time series is so far limited. However, some work shows the potential usefulness of the method. For example, (Esteban et al., 2017) apply a recurrent GAN to generate realistic synthetic medical data series. (Zhou et al., 2018) apply the GAN model to forecast high-frequency stock datasets. In recent work, (Luo et al., 2018) use the GAN models to generate missing values for incomplete time series.

In this work, we present a method that expands on the CGAN algorithm. In our model, the generator estimates a multimodal posterior distribution, via a finite mixture of Gaussians. Unlike most variations of GAN models, whereby the generator makes a point estimation, MD-CGAN is capable of estimating a flexible probability distribution. This paper is set out as follows. In Section 2 we present the structure of the MD-CGAN model. In section 3 we test the model on various datasets and discuss the results. Finally in Section 4, we conclude.

2. The MD-CGAN Model Framework

We consider a time series, y_t. Our aim is to infer the posterior over some $y_{t'>t}$, conditioned on a set of observations which we denote $\mathbf{x}_t$. In order to make the posterior estimation we model the conditional density $p(y_{t'}|\mathbf{x}_t)$ as an adversarial network. To achieve this we use a Mixture Density Network (MDN) model similar to the one presented by Bishop (2006) for the generator G. The inputs to the generator network are $\mathbf{x}_t$ and $\mathbf{z}_g$, where $\mathbf{z}_g$ is a vector of samples randomly selected from a normal distribution. The output of the $G_{t'}(\mathbf{x}_t, \mathbf{z}_g)$ are the parameters of the Gaussian mixture model, $\alpha_{t'}$, $\sigma_{t'}$, and $\mu_{t'}$. As first proposed in Bishop (2006), we achieve this by using latent variables $\mathbf{s} = \{\mathbf{s}_\alpha, \mathbf{s}_\sigma, \mathbf{s}_\mu\}$ that are conditioned on the inputs, and where the mapping from $[\mathbf{x}_t, \mathbf{z}_g] \mapsto \mathbf{s} \mapsto [\alpha_{t'}, \sigma_{t'}, \mu_{t'}]$ is modelled via our network. As the mixture coefficients, α_i, must satisfy $\sum_{i=1}^{m} \alpha_i(\mathbf{x}_t) = 1$, hence we map $\mathbf{s}_\alpha$ to α via the *softmax* function. As the elements of σ are strictly positive so we adopt, $\sigma = \exp(\mathbf{s}_\sigma)$. Finally the means can be mapped directly from the latent variables, hence $\mu = \mathbf{s}_\mu$. This formalism allows us to directly model the predictive likelihood conditioned on an input, and the likelihood of G, indicated as $\mathcal{L}(G)$ is as follows:

$$\mathcal{L}(G_{t'}(\mathbf{x}_t, \mathbf{z}_g)) = \sum_{i=1}^{m} \alpha_{t'}(\mathbf{x}_t, \mathbf{z}_g) \mathcal{N}_i(y_{t'} | \mu_{t'}(\mathbf{x}_t, \mathbf{z}_g), \sigma_{t'}(\mathbf{x}_t, \mathbf{z}_g)), \quad (1)$$

where m is the number of mixing coefficients. We note that the generator outputs the parameters of a Gaussian mixture model, and therefore it offers extreme flexibility in modelling the posterior distribution for y_t'.

As in the CGAN model, the discriminator, D, is also conditioned on $\mathbf{x}_t$. The input to the discriminator model is, by design, $\mathbf{x}_t\mathcal{L}(y_{t'})$ and the output $\mathbf{x}_t$. For true values of $y_{t'}$, the likelihood has the maximum value of 1. The generator tries to 'fool' the discriminator by generating $G_{t'}$ where its likelihood is close to 1. The loss function for the generator, L_G in Equation 2, reflects this concept. The discriminator network, on the other hand, tries to differentiate between true $y_{t'}$ values and the pseudo-values created by the generator. The loss function for the discriminator, L_D in Equation 3, reflects this, where the lowest value is achieved when $\mathcal{L}(y_{t'})$ is 1 and $\mathcal{L}(G_{t'}(\mathbf{x}_t, \mathbf{z}_g))$ is 0.

$$L_G = \mathbb{E}_{z \sim P_z(z)}[-\mathcal{L}(G_{t'}(\mathbf{x}_t, \mathbf{z}_g))] \quad (2)$$

$$L_D = \mathbb{E}_{y \sim P_{data}(y)}[\|\mathbf{x}_t\mathcal{L}(y_{t'}) - \mathbf{x}_t\|^2] + \mathbb{E}_{z \sim P_z(z)}[\|\mathbf{x}_t\mathcal{L}(G_{t'}(\mathbf{x}_t, \mathbf{z}_g))\|^2] \quad (3)$$

The algorithm follows the steps in Algorithm 1. Figure 1 illustrates the structure and the interaction between the generator and the discriminator for the MD-CGAN model.

Algorithm 1 MD-CGAN Algorithm

for number of training iterations **do**
  for k steps do **do**
    Sample m noise samples, $\{z^1,...,z^m\}$ from $p_g(z)$
    Sample m data points, $\{x^1,...,x^m\}$ from $p_{data}(x)$
    Update the discriminator by descending its stochastic gradient:
    $$\nabla_{\theta_{\mathcal{L}}} \sum_{i=1}^{m} [\|\mathbf{x}\mathcal{L}(y) - x\|^2 + \|\mathbf{x}\mathcal{L}(G(z^i, x^i))\|^2]$$
  end for
  Sample m noise samples, $\{z^1,...,z^m\}$ from $p_g(z)$
  Update the generator by descending its stochastic gradient:
  $$\nabla_{\theta_g} \sum_{i=1}^{m} -\mathcal{L}(G(z^i, x^i))$$
end for

3. Experimental Results

3.1. Comparison with other Deep Learning Models

We compare the MD-CGAN model to the Mixture Density Network model (MDN) introduced by (Bishop, 2006), the CGAN model and a standard neural network (SNN). We run experiments on various time series datasets. Our results illustrate that the posterior estimated by the MD-CGAN has lower mean square error (MSE) compared to other models, for longer horizons and in the presence of noise or unforeseen shifts in noise levels.

The MD-CGAN model has two major advantages over the SNN. The first is the capability to estimate the posterior distribution instead of a point estimation. We note that for the purpose of our experiments, the mixture coefficient m is set to 1, and therefore the estimated posterior is effectively a Gaussian.

The second advantage is the adversarial structure of the MD-CGAN, which enables the model to navigate through unforeseen shifts in noise levels and have lower MSE in comparison to the non-adversarial models.

We perform experiments on four datasets, the Mackey-Glass chaotic dataset, sunspot dataset (Royal Observatory of Belgium, 2015), US initial jobless claims (U.S. Department of Labor, 2018) (USIJC), and the EURUSD foreign exchange daily rates (EURUSD FX rate). For consistency, the generator in the MD-CGAN, has the same structure as the MDN, the generator of the CGAN, and the SNN for our experiments. For each dataset, the series is split to the test and training sets. We assess the performance based on MSE, which is computed over the test dataset that consists of 400 data points post the training set. All algorithms have as input the last k data points, where k is set to 5 for the purpose of our experiments. We note that CGAN and SNN make point estimate predictions while MD-CGAN and MDN estimate posterior *distributions* at each time interval. The mixture coefficient m is set to one for all experiments to enable easier comparison and the mean of the distribution is considered the estimated value for the MDN and MD-CGAN models for the purpose of calculating the errors.

Mackey-Glass and Sunspot datasets: Error comparison across the models are indicated in Table 1. The GAN models (CGAN and MD-CGAN), do not outperform other models for these datasets.

Adding normally distributed noise to above test datasets: To evaluate the additional benefit of adversarial training, we add (30% by amplitude) normally distributed noise to the test data, noting that the noise is not added to the training dataset. This creates a shift in the level of noise between the training set and the test set. Mean-square errors are indicated in Table 1, referenced as "with noise". In this

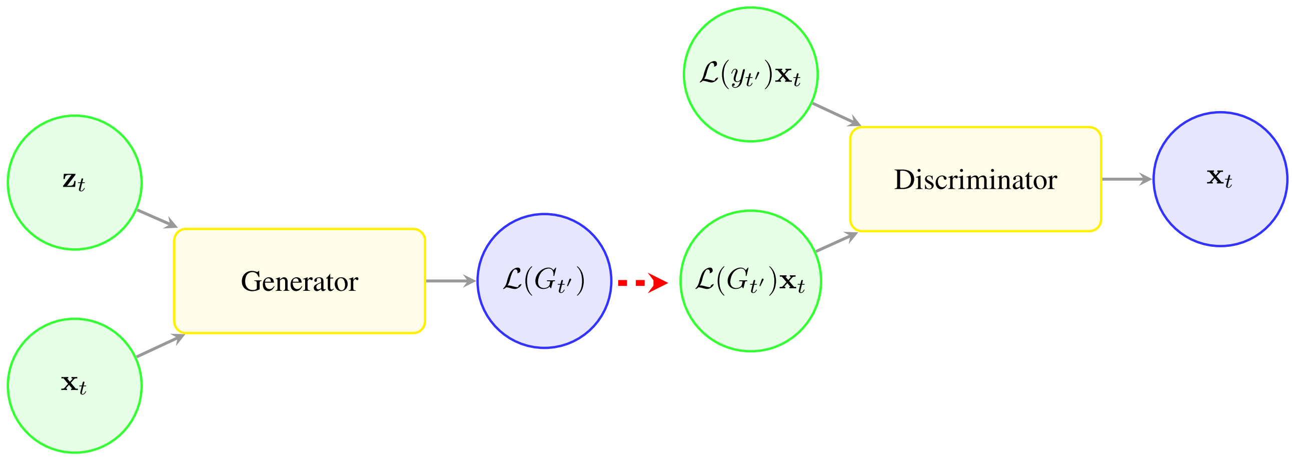

Figure 1. Schematic of our proposed MD-CGAN model showing Generator and Discriminator components.

scenario, the GAN models outperform the SNN and MDN., with the MD-CGAN having the best performance of all models. GAN approaches are thus able to more effectively deal with systematic shifts in additive noise.

US initial jobless claims and EURUSD foreign exchange daily rate: As one of our primary interests is the forecasting in a finance setting, where stochastics are dominant, we look in detail at the performance of the GAN approaches to finance-related time series forecasting. We consider two candidate time series. Firstly, the US initial jobless claims over a 40 year period, sampled weekly and a foreign exchange (FX) rate between the EUR and the USD, over a period of 15 years sampled daily. Figure 2 highlights the relative performance of 1-step forecasts of the various approaches taken on these data sets. Out of sample mean-square error results are also provided in Table 1.

3.2. Longer-horizon forecasts

The forecasts in Subsection 3.1 are over 1-step. We ran the models over the longer-term forecast of ten weeks for both time series. We also perform comparisons against standard econometric linear models, namely an AR(5) model and the martingale model, where the expectation of the forecast is merely the present datum. Taking the martingale model as a baseline, we present in Table 2 the mean-square errors as the ratio to the martingale model error. We note that the MD-CGAN approach has ratios below one and provides the lowest error of all models in this scenario.

4. Conclusion

We present the MD-CGAN model, with enhanced performance in the presence of stochastic noise and for longer horizons. In the experiments presented, the MD-CGAN outperforms all non-linear models tested on the noisy Mackey Glass and Sunspot datasets as well as the financial time series, for both short and long term forecast horizons. As the forecasting horizon is extended we find it outperforms baseline linear models as well as other nonlinear approaches. As a GAN model, our approach retains adversarial robustness, most notable when noise is present in data. Furthermore, our MD-CGAN model can effectively estimate a flexible posterior distribution, in sharp contrast to standard GAN models. Exploiting the rich, multi-model, posterior distribution is not reported here but will feature in follow-up work. In summary, the MD-CGAN model combines the advantageous features of both mixture density and GAN methods. We see this as particularly useful in dealing with time series in which noise is significant and for providing robust long-term forecasts beyond simple point estimates.

References

Bishop, C. M. *Pattern Recognition and Machine Learning (Information Science and Statistics)*. Springer-Verlag, Berlin, Heidelberg, 2006. ISBN 0387310738.

Esteban, C., Hyland, S. L., and Rätsch, G. Real-valued (Medical) Time Series Generation with Recurrent Conditional GANs. *arXiv preprint arXiv:1706.02633*, 2017.

Goodfellow, I., Pouget-Abadie, J., Mirza, M., Xu, B., Warde-Farley, D., Ozair, S., Courville, A., and Bengio, Y. Generative Adversarial Nets. In *Advances in Neural Information Processing Systems*, pp. 2672–2680, 2014.

Luo, Y., Cai, X., Zhang, Y., Xu, J., et al. Multivariate time series imputation with generative adversarial networks. In *Advances in Neural Information Processing Systems*, pp. 1596–1607, 2018.

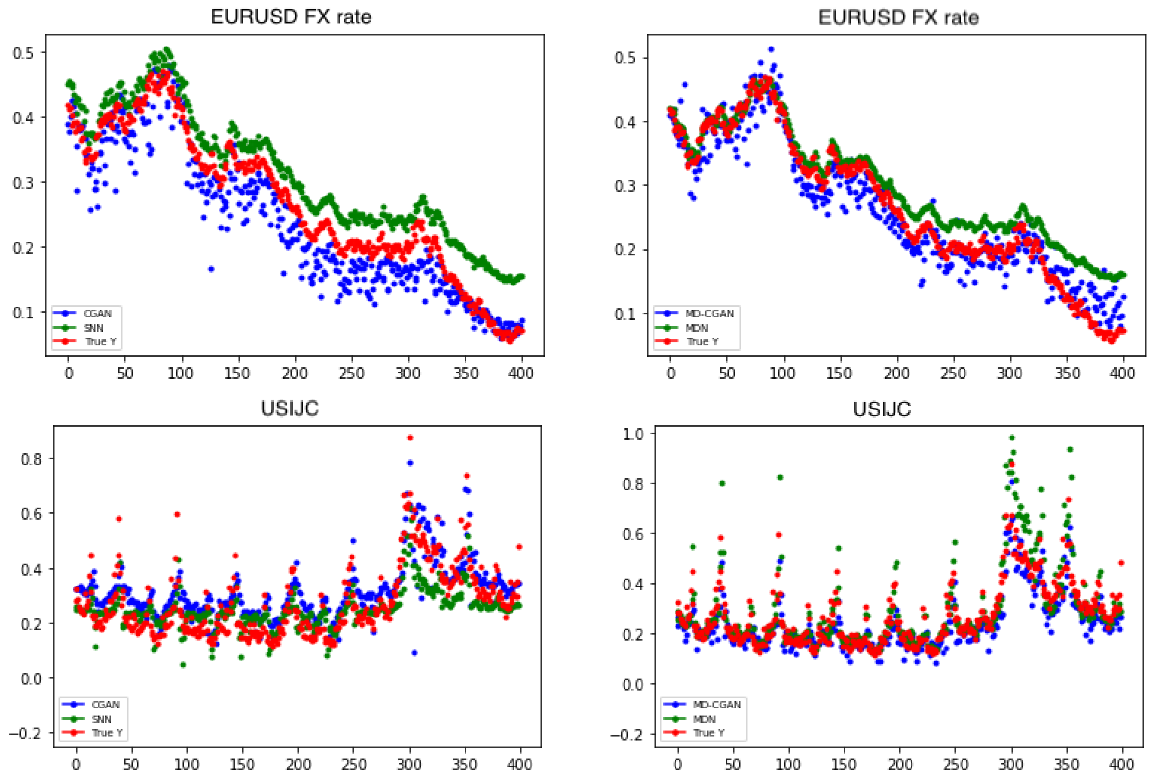

Figure 2. Results for the foreign exchange (FX) and US jobless claims (USIJC) time series. The left-hand plots show comparison of CGAN (blue), SNN (green) and ground truth (red). The right-hand plots show forecasts of our model, MD-CGAN (blue), the MDN approach (green) and true values (red). The y-axis is arbitrary after data normalization, and the x-axis is in samples. All data constitutes out of sample test data with a forecast horizon of one day for FX and one week for USIJC.

	Mackey-Glass	**Sunspot**	**Mackey-Glass with Noise**	**Sunspot with Noise**	**USIJC**	**EURUSD FX rate**
SNN	0.0014 *(0.0020)*	0.0154 *(0.0260)*	0.1640 *(0.1939)*	0.0536 *(0.0777)*	0.0070 *(0.0154)*	0.0022 *(0.0017)*
CGAN	0.0036 *(0.0048)*	0.0196 *(0.0358)*	0.0360 *(0.0525)*	0.0266 *(0.0380)*	0.0074 *(0.0157)*	0.0018 *(0.0026)*
MDN	**0.0002** ***(0.0003)***	**0.0105** ***(0.0239)***	0.1402 *(0.1597)*	0.0758 *(0.1144)*	0.0080 *(0.0238)*	0.0016 *(0.0020)*
MD-CGAN	0.0026 *(0.0030)*	0.0172 *(0.0293)*	**0.0264** ***(0.0392)***	**0.0203** ***(0.0359)***	**0.0041** ***(0.0089)***	**0.0008** ***(0.0011)***

Table 1. MSEs (standard deviations) for all experiments. All data was pre-normalized to $[0, 1]$ range.

	USIJC	**EURUSD FX rate**
AR(5)	0.78	1.91
SNN	0.79	1.25
CGAN	0.77	0.85
MDN	0.84	3.48
MD-CGAN	**0.73**	**0.76**

Table 2. Ratio of model MSE to martingale baseline model for a forecast horizon of 10 weeks.

Mirza, M. and Osindero, S. Conditional Generative Adversarial Nets. *arXiv preprint arXiv:1411.1784*, 2014.

Royal Observatory of Belgium. SILSO Data, 2015.

U.S. Department of Labor. The Unemployment Insurance Weekly Claims. Bureau of Labor Statistics, 2018.

Zhou, X., Pan, Z., Hu, G., Tang, S., and Zhao, C. Stock Market Prediction on High-Frequency Data Using Generative Adversarial Nets. *Mathematical Problems in Engineering*, 2018, 2018.