KU LEUVEN

Generating individual claims using generative adversarial networks

Yves-Cédric Bauwelinckx with
E.J. Menvouta, T. Verdonck and J. Dhaene
KU Leuven

0 Outline

1 Outline

1 Introduction - Synthetic data

What?

- **Fake, generated data** made to **resemble** the **original, real data**

1 Introduction - Synthetic data

What?

- **Fake, generated data** made to **resemble** the **original, real data**

Why?

- Rise in data driven methods for modeling
- But: limited data available due to **privacy and ethics** concerns
- No private information in synthetic data $\Rightarrow$ **data can be shared**

1 Introduction - Synthetic data

What?

- **Fake, generated data** made to **resemble** the **original, real data**

Why?

- Rise in data driven methods for modeling
- But: limited data available due to **privacy and ethics** concerns
- No private information in synthetic data $\Rightarrow$ **data can be shared**

How?

- Traditionally: scenario generators with assumptions
- Recently: Machine learning, **generative models**
- Generative model **learns underlying distribution** from real data
- Sample from learned distribution to create synthetic data

2 Outline

2 Background - Synthetic data in insurance

- Synthetic data for various types of data in insurance
  - Simulation of driver telematics (So, Bouchez and Valdez, 2021)
  - Simulation of insurance fraud network data (Campo and Antonio, 2023)

2 Background - Synthetic data in insurance

- Synthetic data for various types of data in insurance
  - Simulation of driver telematics (So, Bouchez and Valdez, 2021)
  - Simulation of insurance fraud network data (Campo and Antonio, 2023)
- Generative AI in insurance
  - Generative Synthesis of Insurance Datasets (Kuo, 2020)
  - Synthesizing Property & Casualty Ratemaking Datasets using Generative Adversarial Networks (Côté et al. 2020)
  - Variational autoencoder for synthetic insurance data (Jamotton and Hainaut, 2023)

2 Background - Synthetic data in insurance

- Synthetic data for various types of data in insurance
  - Simulation of driver telematics (So, Bouchez and Valdez, 2021)
  - Simulation of insurance fraud network data (Campo and Antonio, 2023)
- Generative AI in insurance
  - Generative Synthesis of Insurance Datasets (Kuo, 2020)
  - Synthesizing Property & Casualty Ratemaking Datasets using Generative Adversarial Networks (Côté et al. 2020)
  - Variational autoencoder for synthetic insurance data (Jamotton and Hainaut, 2023)
- Individual claim generators
  - Individual Claims History Simulation Machine (Gabrielli and Wüthrich, 2018)
  - SynthETIC (Avanzi et al., 2021)

2 Background - Simulating claims reserving data

- Individual Claims History Simulation Machine (Gabrielli and Wüthrich, 2018)
  - Uses 35 neural networks, used over 8 sequential steps
  - Trained on data, but requires several assumptions
  - Performs well for Chain-ladder reserving method

2 Background - Simulating claims reserving data

- Individual Claims History Simulation Machine (Gabrielli and Wüthrich, 2018)
  - Uses 35 neural networks, used over 8 sequential steps
  - Trained on data, but requires several assumptions
  - Performs well for Chain-ladder reserving method
- SynthETIC (Avanzi et al., 2021)
  - Uses 8 modules
  - Offers flexibility to the user
  - Several distributional assumptions where the user can change parameters

2 Goal

- One model, trained in one go
- Data driven, make assumptions as lenient as possible
- Model should be easily adaptable to new datasets
- Quality of data should be close to original data

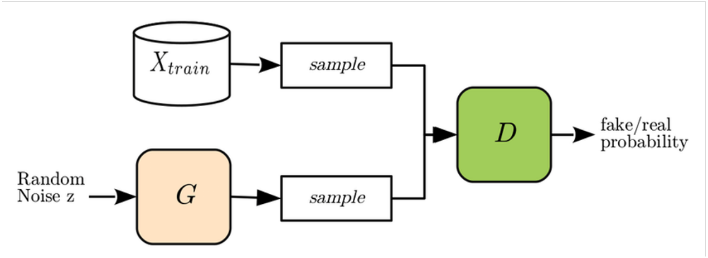
2 Goal

- One model, trained in one go
- Data driven, make assumptions as lenient as possible
- Model should be easily adaptable to new datasets
- Quality of data should be close to original data

$\Rightarrow$ G(enerative) A(dversarial) N(etwork) with causal structure

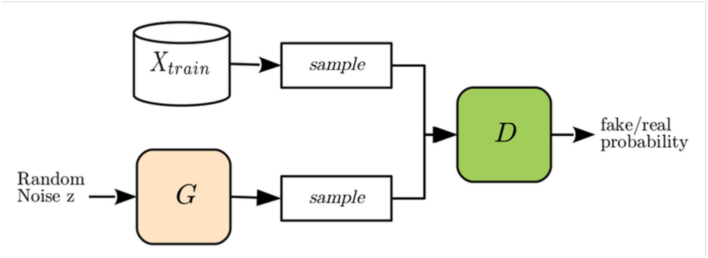
3 Outline

3 Model - GAN



- G(enerator) en D(iscriminator) $\Rightarrow$ 2 neural networks
- Generator and Discriminator compete against each other (adversarial)
- Goal: generator maps random noise to real data distribution

3 Model - GAN



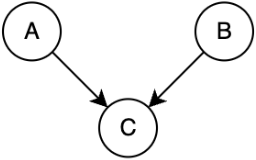
- **Generator generates** a random **sample** to "fool" Discriminator
- **Discriminator** tries to **distinguish real from generated samples**
- Discriminator gives feedback to Generator (Through a loss function)
- Generator performs better ⇒ Discriminator performs better ⇒ Generator performs better, etc.

3 Model - Causal framework

- The generator is typically one neural network
- Instead, we use a framework from Causal-TGAN (Wen et al., 2021)
  - Small neural network for each variable, following the causal structure
  - All small neural networks make up one big neural network

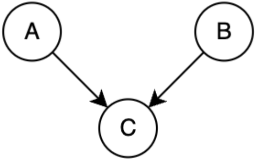
3 Model - Causal framework

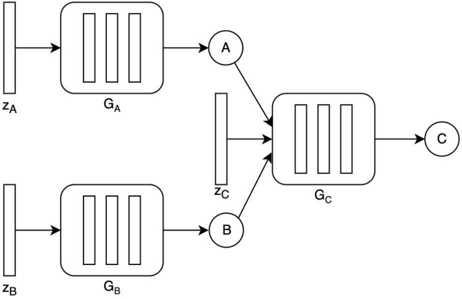
- The generator is typically one neural network
- Instead, we use a framework from Causal-TGAN (Wen et al., 2021)
  - Small neural network for each variable, following the causal structure
  - All small neural networks make up one big neural network



3 Model - Causal framework

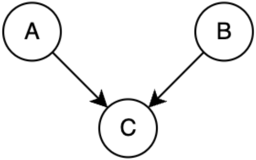
- The generator is typically one neural network
- Instead, we use a framework from Causal-TGAN (Wen et al., 2021)
  - Small neural network for each variable, following the causal structure
  - All small neural networks make up one big neural network

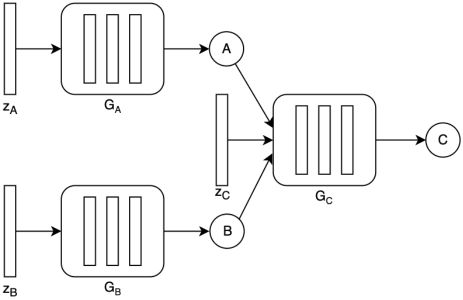




3 Model - Causal framework

- The generator is typically one neural network
- Instead, we use a framework from Causal-TGAN (Wen et al., 2021)
  - Small neural network for each variable, following the causal structure
  - All small neural networks make up one big neural network





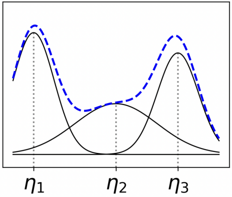
- Partial or no causal graph is also possible
- Shown to perform better than non-causal counterpart

3 Model - Distributions

- Data transformation from CTGAN (Xu et al., 2019)
- Discrete distributions
  - One-hot encoding
  - All categories are sampled evenly during training period

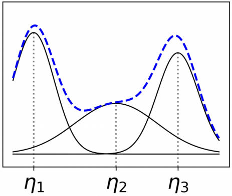
3 Model - Distributions

- Data transformation from CTGAN (Xu et al., 2019)
- Discrete distributions
  - One-hot encoding
  - All categories are sampled evenly during training period
- Continuous distributions
  - Variational Gaussian mixture (VGM) model for each column
  - Each marginal distribution gets represented as a mixture of Gaussian distributions



3 Model - Distributions

- Data transformation from CTGAN (Xu et al., 2019)
- Discrete distributions
  - One-hot encoding
  - All categories are sampled evenly during training period
- Continuous distributions
  - Variational Gaussian mixture (VGM) model for each column
  - Each marginal distribution gets represented as a mixture of Gaussian distributions



- A value gets transformed to an indicator and a normalised value

4 Outline

4 Data

- 1 million samples from simulator of Gabrielli and Wüthrich

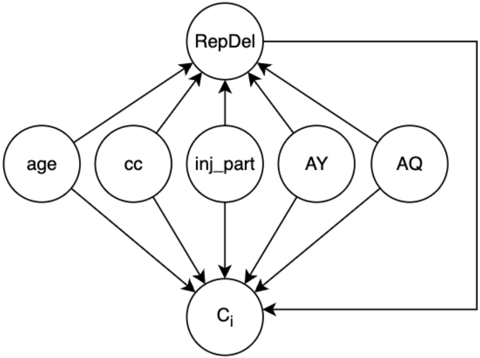
Name	Notation
Age	*age*
Claims code	*cc*
Accident year	*AY*
Accident quarter	*AQ*
Injured part	inj_part
Reporting delay	RepDel
Payment	P_i
Open status	O_i

- Payments and open status for i $\in$ 0, 1,..., 11 years

4 Data - Long tailed distributions

- Long tailed distributions are not captured as well with VGM
- Log-transformation of the columns with long-tails (Payments)
- Resulting distribution is more closely-packed together
- $\Rightarrow$ better representation by VGM

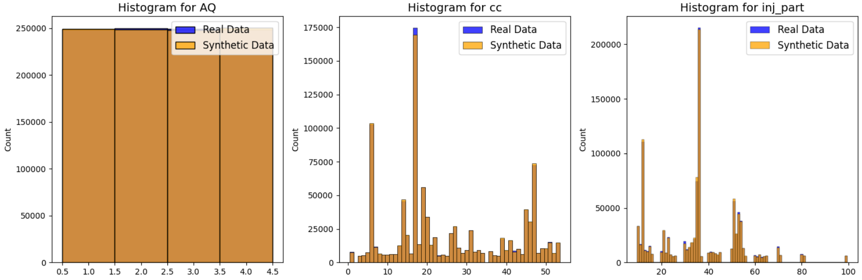
4 Data - Causal structure



- C_i represents the claim payment and open status in year i

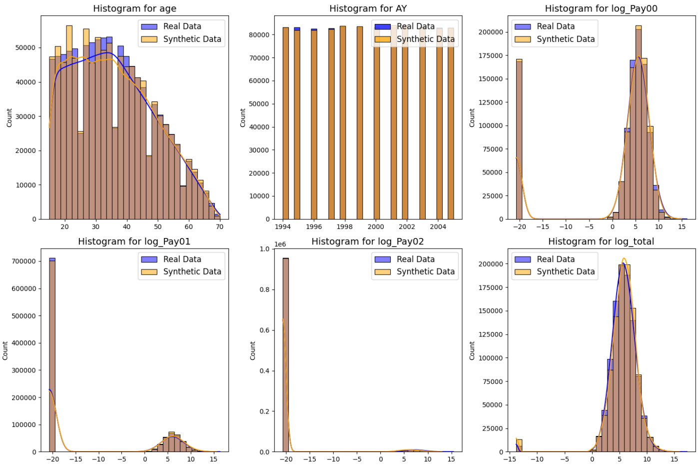
5 Outline

5 Results - Discrete



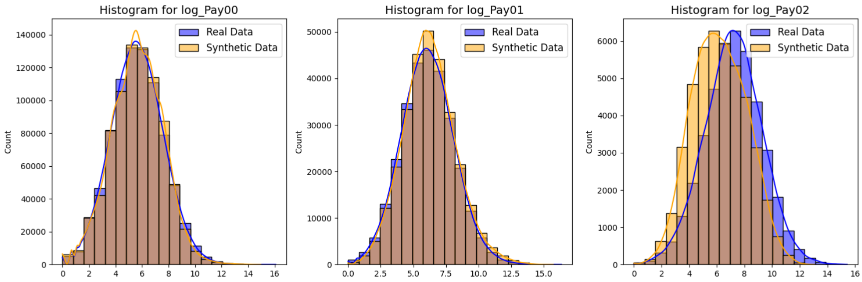
- Discrete distributions are reconstructed very well

5 Results - Continuous



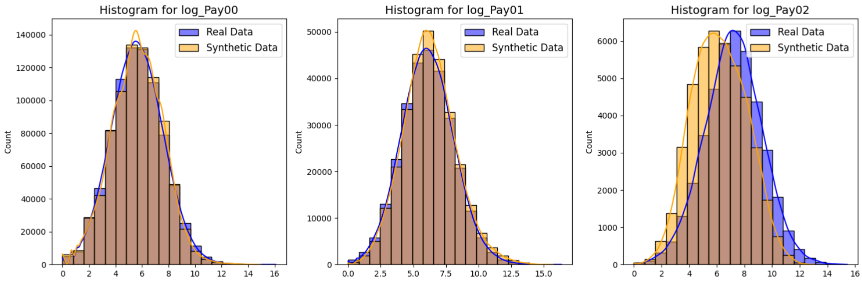
- Continuous distributions are reconstructed well

5 Results - Continuous



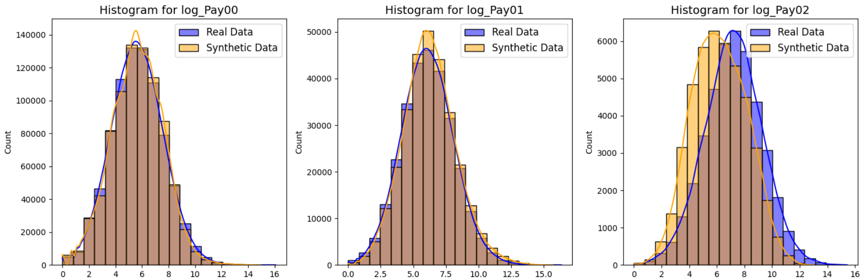
- Distribution of the non-zero claims

5 Results - Continuous



- Distribution of the non-zero claims
- Payments in the first and second year are well reconstructed
- Payments in third year are more sparse $\Rightarrow$ worse reconstruction

5 Results - Continuous



- Distribution of the non-zero claims
- Payments in the first and second year are well reconstructed
- Payments in third year are more sparse $\Rightarrow$ worse reconstruction
- Small difference in log transformed distribution can mean big difference when transformed back

5 Results - Payments

Statistic	Real data	Synthetic data
1^{st} year		
No claim payments	16.93%	17.20%
Average payment	2294.83	2437.02
Median payment	276.0	271.0
Largest payment (10^6)	9.60	3.14

5 Results - Payments

Statistic	Real data	Synthetic data
1^{st} year		
No claim payments	16.93%	17.20%
Average payment	2294.83	2437.02
Median payment	276.0	271.0
Largest payment (10^6)	9.60	3.14
2^{nd} year		
No claim payments	71.41%	72.32%
Average payment	4529.45	6377.70
Median payment	432.0	488.0
Largest payment (10^6)	12.83	11.33

5 Results - Payments

Statistic	Real data	Synthetic data
1^{st} year		
No claim payments	16.93%	17.20%
Average payment	2294.83	2437.02
Median payment	276.0	271.0
Largest payment (10^6)	9.60	3.14
2^{nd} year		
No claim payments	71.41%	72.32%
Average payment	4529.45	6377.70
Median payment	432.0	488.0
Largest payment (10^6)	12.83	11.33
Total (summed over 12 years)		
No claim payments	0.62%	1.29%
Average payment	4346.33	4652.30
Median payment	313.0	319.0
Largest payment (10^6)	31.7	38.9

6 Conclusion

- One model, trained in one go
  - Generator is one neural network

6 Conclusion

- One model, trained in one go
  - Generator is one neural network
- Data driven, make assumptions as lenient as possible
  - Assumption 1: distribution can be represented by a Gaussian Mixture
  - Assumption 2: causal graph (optional)

6 Conclusion

- One model, trained in one go
  - Generator is one neural network
- Data driven, make assumptions as lenient as possible
  - Assumption 1: distribution can be represented by a Gaussian Mixture
  - Assumption 2: causal graph (optional)
- Model should be easily adaptable to new datasets
  - User only needs to provide data and specify long-tailed variables
  - Causal graph is optional

6 Conclusion

- One model, trained in one go
  - Generator is one neural network
- Data driven, make assumptions as lenient as possible
  - Assumption 1: distribution can be represented by a Gaussian Mixture
  - Assumption 2: causal graph (optional)
- Model should be easily adaptable to new datasets
  - User only needs to provide data and specify long-tailed variables
  - Causal graph is optional
- Quality of data should be close to original data
  - Does generally well
  - Sparse data is generalised at a lower quality

Thank you for your attention!

Contact: yves-cedric.bauwelinckx@kuleuven.be