Population Synthesis Meets Deep Generative Modelling

Stanislav S. Borysov, Jeppe Rich, Francisco C. Pereira

Department of Management Engineering, Technical University of Denmark, DTU, 2800 Kgs. Lyngby, Denmark

Abstract

Agent-based transport models depend to a high degree on the formation of the underlying population. Models and methods for generating such populations, possibly under constraints to reflect future margins, are commonly referred to as “population synthesis” models. Historically, different approaches have been proposed, ranging from deterministic approaches, such as the Iterative Proportional Fitting (IPF) algorithm (Deming and Stephan, 1940; Rich and Mulalic, 2012), to simulation based approaches of the Markov Chain Monte Carlo (MCMC) type (Farooq et al., 2013). Often, matrix fitting methods such as the IPF has been used in combination with post-simulation based methods in order to translate prototypical individuals into true micro agents. While the existing models are capable of producing acceptable results for agents with relative few socio-economic and spatial characteristics, these methods do not scale well when the dimensionality of the underlying distribution becomes large. As a result, in many cases, these methods are not able to accommodate the increasing need for more dimensions that result from, e.g. smaller zones, the combination of household-based and individual-based synthesis and more detailed variables in general. In this paper, we propose a different approach to population synthesis based on generative models from the deep learning framework. Contrarily to existing methods, these new methods are scalable and can handle a very large number of both numerical and categorical attributes at the same time.

Methodology

Modern generative models, such as Variational Autoencoders (VAEs) (Kingma and Welling, 2013) and Generative Adversarial Networks (GANs) (Goodfellow et al., 2014), have primarily been used for image processing. For example, they are capable of generating plausible (photo-realistic) pictures, effectively learning the complex underlying joint probability distribution of training data. In this work, we use a Variational Autoencoder, which is essentially a latent variable model (Fig. 1). The main purpose of a VAE is to estimate the joint probability distribution of the data through a sequence of nonlinear transformations, which are usually represented as a deep neural network, applied to a low-dimensional latent space with simple Gaussian properties (Fig. 1).

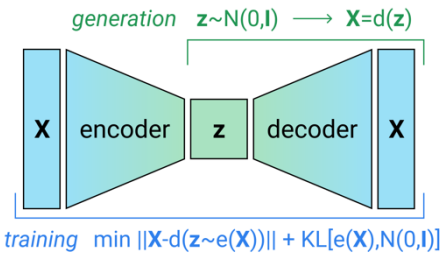

Figure 1. *Variational Autoencoder. During the training step, it learns to encode the training data* ***X*** *into the latent space* ***z*** *and reconstruct them back, enforcing normal distribution of the latent variable. To generate new samples, the encoder is discarded and the decoder network is used.*

During the training phase, the original data go through an encoding network, which maps data to the latent space. In the second stage, the data are reconstructed back to the original form using a decoding network. Although it is a probabilistic graphical model, it is possible to train the model by using the standard backpropagation algorithm, which opens a path towards large-scale problems unlike MCMC-based approaches. Once a VAE has been trained, samples that mimic the distribution of the original data can be generated by doing Gaussian random sampling in the latent space and transforming these samples back to the original data space using the decoder network. Statistical properties of the generated samples naturally follow statistical properties of the original data. Moreover, being a fully probabilistic model, a VAE is capable of estimating the log-likelihood of the data, which is of practical use for model comparison and outlier detection.

From a population synthesis perspective, there are several benefits related to the use of generative models. First, a full joint probability distribution for the data may effectively solve a range of privacy issues. By using a generative model, it is possible generate population samples with the same statistical properties as the real population but avoiding any references to real persons. As a result, public authorities could make publicly available generative models trained on the whole dataset as an alternative to micro-sampling. Second, generative models can be used for data compression, where the related space requirements grow only with the number of the model's parameters. Third, generative models can impute missing values in a natural way. Finally, examination of the data in the latent space can reveal new data properties such as clustering behaviour.

At a more practical level, the method could also replace current population synthesis practise by combining the VAE model, through which it would be possible to draw detailed population samples, with re-sampling schemes or quota-based sampling schemes to create sub-samples of the data which accommodate certain known future margins. In this way, it would be possible to create future populations which, on the one hand, reflect the correlation structure of the original population but, on the other hand, are restricted according to future population targets.

Case study

As a case study, we use the Danish National Travel Survey (TU) data, which is known to be one of the largest coherent trip diaries worldwide. It contains 24-hour trip diaries for more than 330.000 Danish residents and provides detailed information for approximately 1 million trips in the period from 1992 to present. To test the methodology, we develop a VAE model for the TU sample from 2010, consisting of 23.754 records, and test how well synthetically generated samples are able to reproduce the properties of the original sample. For each person in the survey, we select a total of 34 numerical attributes (e.g. income, age, average travel time) and 3 categorical attributes (sex, education and occupation). The resulting distribution is 60-dimensional and a combination of categorical variables and purely numerical variables.

In the test, a VAE model with a fully connected neural network with 2 hidden layers of 64 neurons each for both encoding and decoding parts has been applied. The VAE is trained through a standard back-propagation algorithm on a train set and validated on a test set. Synthetic data is generated by drawing samples from the 16-dimensional Gaussian latent space and by passing these samples through the decoding network. It is found that the synthetic distributions provide a good approximation to the marginal distributions of the original data (Fig. 2). We also compared joint distribution of the synthetic data and a simple marginal baseline model, which assumes independence of all features, $p(\mathbf{x}) = \prod p(x_i)$. The VAE (corr = 0.90, SRMSE = 2.18, R^2 = 0.78) significantly outperformed this baseline (corr = 0.74, SRMSE = 2.84, R^2 = 0.47). It is worth noting

that even a slightly more advanced conditional baseline for MCMC sampling would suffer from the curse of dimensionality and require an exponential number of conditionals.

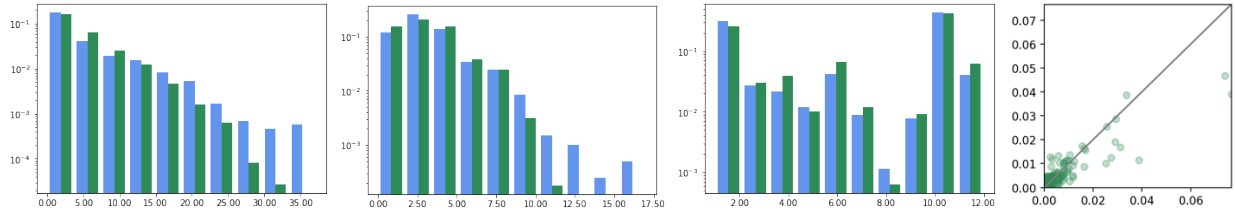

Figure 2. *Examples of marginal distributions from the true population (blue) and generated samples from the VAE (green): distance to the nearest station from home, number of trips, occupation by category (from left to right); and frequency comparison of the true and synthetic populations.*

So far, we have showed the possibility of generating synthetic population for a given year. Another important challenge is to generate samples for future years in order to facilitate forecasts of the underlying transport model. Typically, such future populations are constructed by imposing marginal constraints that reflect known margins for the future population, e.g. age, gender or income constraints. Incorporating such constraints directly into the VAE is an ongoing research problem. However, alternative sample-based solutions, such as importance sampling or quota-based sampling, can be used in order to fulfil this purpose.

As a future research direction, we also consider similar approaches applied to multi-agent problems (e.g. generation of households) and daily activities patterns.

Literature

Deming, W. E. and Stephan, F. F. (1940) On the least squares adjustment of a sampled frequency table when the expected marginal totals are known. Annals of Mathematical Statistics, 11 (4), 427–444.

Farooq, Bilal, Michel Bierlaire, Ricardo Hurtubia and Gunnar Flötteröd. Simulation Based Population Synthesis. Transportation Research Part B: Methodological 58, (2013): pp 243-263.

Rich, J., Mulalic, I., 2012. Generating synthetic baseline populations from register data. Transportation Research Part A 46, 467–479.

Kingma, D.P. and Welling, M., 2013. Auto-encoding variational bayes. arXiv preprint arXiv:1312.6114.

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