Modelling Public Procurement with Gaussian Mixture VGAE

Daniel Tanis[1] and Edoardo Pona[2]

[1] University of Cambridge, UK,
[2] Kings College London, UK

Abstract. We develop β–Graph Variational Autoencoder with a Gaussian mixture distribution and apply it to a network of public procurement records. We show that this model is able to accurately reconstruct the edges between buyers and sellers observed during training, learn meaningful and partially disentangled representations of this transaction network and generalise predictions to an out-of-sample set of unobserved edges. We analyse the model outputs to illustrate how the learned representations can be used by domain experts to learn additional properties of procurement networks and to inform other downstream tasks.

Keywords: graph variational autoencoder, public procurement, interpretability, disentangled representations

1 Modelling Transaction Networks

1.1 Research problem

In this research we develop a generative model capable of learning latent properties of a transaction network. By observing one year of public procurement records, we are able to learn relevant and interpretable properties of how buyers and sellers are connected which is highly informative of how firms will bid in the future.

From a policy perspective, there are two interesting ways of motivating why this problem is relevant. First, consider the problem of anticipating what a procurement market would look like. For instance, setting up an the purchase of an IT system in a region that has no experience with that sector yet. Researchers and policymakers might be interested in simulating how this scenario would look like; what is the level of competition that could be expected in the long run or how concentrated the market might be. Generating alternative markets by combining information from previously observed sectors could give good insights to that end.

Alternatively, consider the case of trying to anticipate the effects of an intervention on a market, such as the debarment or other strategies of collusion deterrence. If intervention affects how procurement markets operate, it is desirable to have a setting in which the data generating model of procurement market can be simulated, as opposed to only observing a single snapshot of this system.

That is, some model that captures the underlying processes that produce the observed data and that can generate alternative and equally likely configuration of the same data distribution potentially even creating counterfactual versions of that distribution.

With this in mind, we propose a model that combines strong reconstruction performance and generalisation to an out-of-sample set of edges, with somewhat interpretable latent features. The interpretability of the model inputs and outputs is intended to illustrate potential uses of this model for policy makers and domain experts, which might demand a less "black-box" modelling approach.

In the following sections we present the data used in this analysis, the model derivation and following empirical results. Taken together, this paper is intended to illustrate a potentially powerful application of recent machine learning tools to a novel applied problem. In the conclusion, we explore downstream applications of this model with the objective of contributing to public policy and domain researchers.

1.2 Related Works

This research relates to two main bodies of literature. In social science and economics, there is substantial work on empirical models of public procurement which are generally focused on identifying the effects of specific interventions or detecting patterns of collusion [3, 6, 7]. On a different note, scholars have also focused on game-theoretical models of procurement [5, 13]. In the machine learning literature, there are increasing efforts in using VAE models for learning interpretable representations in different domains [9]. In this research, we extend the efforts from the machine learning community and try to learn meaningful features of a network of public procurement that both highlights relevant underlying features driving market dynamics and are useful for potential downstream prediction tasks.

1.3 Public procurement

Public procurement is one of the main instruments for policy delivery and amounts to 8-20% of countries' GDP [15]. Despite abundant regulations and efforts to increase transparency, public procurement is often subject to strategic gaming on behalf of both firms and public agencies. This threat requires policymakers to be careful when procuring goods and services, and further efforts to increase transparency and accountability are typically needed to improve the quality of these transactions.

In this work, we use data from the Brazilian Federal government. The data is freely available on a public application programming interface (API) created by the Brazilian government in 2015 [4]. This API extracts the information from the official procurement website where tender announcement and bidding happens, to create a systematic platform in which civil society can access and download this information.

We focus on the subsets of the data for which we have complete information on tenders, auction and following contractual data between 2015 and 2016. Selecting only these tenders has two clear benefits: first, it provides detailed and complete information about most stages of the procurement process, and second, it ensures that we only observe successful transactions that actually incurred as a federal resources spent and where companies actually delivered goods or services. One important limitation of this selection is that failed tenders or tenders with missing data could have a meaningful impact on the general bidding environment which we cannot directly observe.

In an average year, there are about 20,000 different companies, 12,000 unique auctions and approximately 100,000 bids, which generates an average of 10 bidders per tender. Most tenders have a generally low value and involve activities such as purchasing office supplies or providing small service contracts, such as consulting contracts or cleaning services. This, in fact, is a property of the how the data is collected: the online bidding platform is designed to cater to standard goods and services, while custom purchases, which may also include more expensive construction and professional services, are allowed to use a different bidding regime.

2 Gaussian Mixture Variational Graph Autoencoder

The Graph The procurement records are treated as the bipartite, unweighted and undirected graph $G = (S, T, E)$, where S and T are disjoint sets of nodes corresponding to suppliers and tenders, respectively, and E a set of edges connecting suppliers to tenders. The graph $G^S_{proj}(S_{proj}, E_{proj})$ is created by projecting the graph G onto the supplier's space. Nodes $S_{proj} \subseteq S$ and the set of edges E_{proj} are constructed based on the original edges of the bipartite graph. For the rest of this report, I will assume that $S_{proj} = S$, which means that all suppliers in the bipartite graph are also in the projected graph.

More formally, the projection of $G = (S, T, E)$ in S, where $S \cap T = \varnothing$, is the $G^S_{proj}(S, E_{proj})$. For every $u, v \in S$, edge $(u, v) \in E_{proj}$ if and only if $\exists i \in T$ where $(u, i) \in S$ and $(v, i) \in S$. The weights of edges $(u, v) \in E_{proj}$ correspond to $|edges(u \in S) \cap edges(v \in S)|$.

The graph G, containing both suppliers and buyers, provides a compact representation of this public procurement market in a given period. So, for instance, the number of bidders for a given tender t is immediately represented by $degree(t)$ and, similarly, the number of tenders supplier s participates in a period of time is represented by $degree(s)$. The projected graph, G^S_{proj}, also indicates the level of overlapping bids (direct or indirect) that the two companies share and the neighborhood of nodes, $N(v) = \{u|(u, v) \in E_{proj}\}$, provides a useful representation of markets.

2.1 Model Derivation

Background For the task of building a generative model of such process, consider the data that is actually available in graph G(S,T,E). We observe the edges

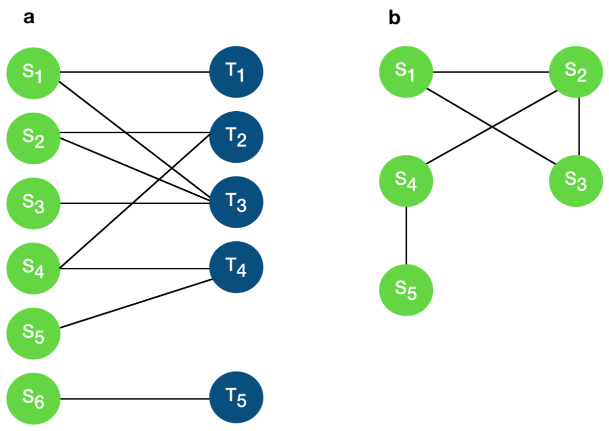

Fig. 1. Bipartite graph (a) and corresponding S projection (b).

connecting suppliers and tenders in a given period of time. We do not observe, however, underlying factors that have created these edges, such as the ones related to the company's decision to bid on a tender or the time of the year in which a tender happened that made a willing company otherwise unable to bid. A collection of non-observable factors together may determine the probability of edges connecting suppliers and companies to be observed in the database. The same set of factors could potentially create very similar graphs or, similarly, small variations of such factors could create different yet reasonable graphs. Uncovering these latent factors is crucial to achieve any sort of generative modelling of this process. To solve this problem, we turn towards the field of variational inference.

In more general terms, the objective is to learn the latent variables z that help in predicting the probability of any X features:

$$P(\mathbf{X}) = \int P(\mathbf{X}|\mathbf{Z})P(\mathbf{Z})dz. \qquad (1)$$

Because variables z are not observable, the model does not require detailed instructions and observations to model features z. Estimating parameters of a distribution instead of a deterministic function is known as variational inference and, in this particular setting, the problem of learning these properties of a latent space distribution in an unsupervised manner can be formulated as a *variational autoencoder* (VAE) [10].

Gaussian Mixture Variational Graph Encoder Thanks to the flexibility of the VAE model, it is possible to design an encoder to directly take graph structured data as input, as seen in the Variational Graph Autoencoder (VGAE) [12]. Without loss of generality, the encoder's task is to transform the input graph into parameters of a probability distribution. In VGAE this is achieved with a graph convolution layer [12] taking the graph in input, followed by two parallel graph convolutional layers computing mean and standard deviations for the latent space distributions. Each graph convolutional layer is defined as $GCN(\mathbf{A}, \mathbf{X}) = ReLU(\tilde{\mathbf{A}}\mathbf{X}\mathbf{W})$ where $\tilde{\mathbf{A}}$ is the symmetrically normalized adjacency matrix $\tilde{\mathbf{A}} = \mathbf{D}^{\mathbf{-1/2}}\mathbf{A}\mathbf{D}^{\mathbf{-1/2}}$ and $\mathbf{X}$ is a matrix of features corresponding to each node, or an identity matrix in the featureless case.

Given the hierarchical nature of our data generating process, we extend the VGAE with a more expressive latent distribution: a mixture of gaussians [1]. Such a distribution keeps many of the advantages of the single Gaussian while allowing to overcome some of its limitations. A simple formulation for a mixture of Gaussian distributions is the following [1]

$$p(\mathbf{X}) = \sum_k \pi_k \mathcal{N}(\mathbf{X}|\boldsymbol{\mu_k}, \boldsymbol{\Sigma_k}) \tag{2}$$

where the weights π_k represent a valid categorical probability distribution.

In general, the Gaussian mixture distribution allows the model to capture two relevant properties of the data; multimodaility and a multi-layered data generation process. The former is a particularly important problem to tackle dealing with inference in a setting where there are important levels of cluster correlation which may or may not be observed.

To achieve this we deviate from the design of the encoder in [12]. The input convolutional layer is maintained unchanged as $\mathbf{h} = GCN(\tilde{\mathbf{A}}, \mathbf{X})$. This then feeds into an array of graph convolutional layers, two for each gaussian in the latent mixture modelling the mean and the standard deviation: $GCN_{i\mu}, GCN_{i\sigma}$. Parallel to that, a classifier is implemented as yet another graph convolutional layer fed with h, outputting the mixture distribution probabilities, which can be optionally interpreted as logits for an auxiliary classification task $\mathbf{c} = GCN(\tilde{\mathbf{A}}, \mathbf{h})$. Samples from this latent distribution are then drawn, and the decoder architecture is unchanged from [12]. The graph is reconstructed from the inner product of the samples $\mathbf{Z}$ from the latent space.

The model is summarised by the following equations:

$$\boldsymbol{\mu_k} = \tilde{\mathbf{A}} ReLU(\tilde{\mathbf{A}}\mathbf{X}\mathbf{W_0})\mathbf{W_{k\mu}} \tag{3}$$

$$\boldsymbol{\sigma_k} = \tilde{\mathbf{A}} ReLU(\tilde{\mathbf{A}}\mathbf{X}\mathbf{W_0})\mathbf{W_{k\sigma}} \tag{4}$$

$$\boldsymbol{\pi} = softmax(\tilde{\mathbf{A}} ReLU(\tilde{\mathbf{A}}\mathbf{X}\mathbf{W_0})\mathbf{W}_{\pi}) \tag{5}$$

where $softmax(x)_i = \frac{e^{x_i}}{\sum_{j=1}^{K} e^{x_j}}$ is the generalization of the logistic function to multiple dimensions [1] and $ReLU(\mathbf{x}) = max(0, \mathbf{x})$ is the rectified linear unit.

$$\mathbf{Z} \sim \sum_{k=1}^{K} \boldsymbol{\pi_k} \mathcal{N}(\mu_k, diag(\sigma_k)) \tag{6}$$

where i is the number of gaussians in our mixture, coinciding with the number of classes in our data when using preforming the auxiliary classification task.

$$\mathbf{\hat{A}} = \sigma(\mathbf{Z}\mathbf{Z}^{\mathbf{T}}) \tag{7}$$

where $\hat{A}$ is the reconstructed adjacency matrix.

Learning and disentanglement Following the traditional VAE literature, we optimise the variational evidence lower bound:

$$\mathcal{L} = \mathbb{E}\left[\log P(\mathbf{X}|\mathbf{Z}) - D_{KL}[Q(\mathbf{Z}|\mathbf{X})||P(\mathbf{Z})]\right]. \tag{8}$$

Furthermore, to disentangle the latent representations, we introduce a β parameter as in [8].

$$\mathcal{L} = \mathbb{E}\left[\log P(\mathbf{X}|\mathbf{Z}) - \beta D_{KL}[Q(\mathbf{Z}|\mathbf{X})||P(\mathbf{Z})]\right]. \tag{9}$$

This equation is a generalisation of the loss defined in (8). For values of $\beta > 1$ the relative importance of the KL entropy between prior and encoded space increases. Because the prior is a combination of a uniform distribution and independent gaussians, it is an orthogonal set of uncorrelated dimensions. As such, the increased cost for the posterior to diverge from the prior puts extra pressures on Z to be factorised while still being sufficient to reconstruct the data [2]. There is a lower cost for adding information in such a way that it (partially) maintains the orthogonality imposed by the prior while trying to learn the weights that reconstruct the original data.

Auxiliary Prediction Task and Latent Distribution In order to improve control over the generative process of our model, we introduce an auxiliary classification layer. As mentioned above, this layer determines the mixture probabilities c which can be seen as weighting the samples from each gaussian. If we add an auxiliary cross entropy loss $H(\mathbf{y}, \mathbf{c}) = \sum_i \mathbf{y_i} log(\mathbf{p_i})$ where $\mathbf{p_i}$ is the class probability distribution $\mathbf{p} = softmax(\mathbf{c})$, we incentivise the model encode the prior that each individual class of nodes can be sampled from a single gaussian. This allows us to leverage additional information we have about classification of our data, and avoids the *mode collapse* [16] problem, where the gaussian mixture may collapse into a single gaussian. What we have described is a semi-supervised system, similar to [11].

3 Results

3.1 Node Latent Representations

The model is able to encode the companies in a rich embedding space that represents their properties. For example, the latent space is able to encode the companies' states as well as partially their regions. Regions are never shown to the model, meaning that they are being discovered from the companies' activity in the graph and their state. The same finding can also be seen when examining the correlation structures of the node latent embeddings. Once more, samples from the same state, as well as from the same region present a high correlation with each other.

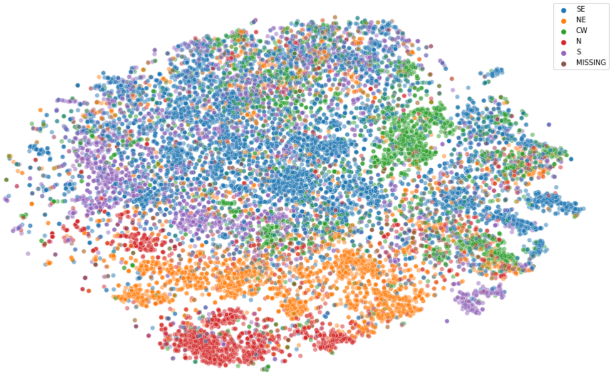

Fig. 2. t-SNE projection of the model embeddings coloured by region. The model is able to extrapolate regional clusters from state information - $\beta = 1$.

As in the original VGAE model, we can recover the graph adjacency matrix by taking the inner product of the latent embeddings. In evaluating performance on the reconstruction task, the trade-off between disentanglement and accuracy becomes evident from 5.

Finally, we also show that the different dimensions of the representation space are also activated according to properties of the procurement network that are not observed during training, namely sector of activity and node degree. Figure 6 illustrates how the produced values of dimensions correlate with the degree of the nodes, and Figure 7 illustrates how dimensions may capture the sector of activity. In this example, the sector used "building construction"-the most common one in the data- clearly activates some of the dimensions of the representation space while keeping others constant. As more than one dimension is activated when encoding these properties, it shows that there is no perfect disentanglement. However, since we do not have full control over the factors of variation, it is not possible to build a proper disentanglement metric [14].

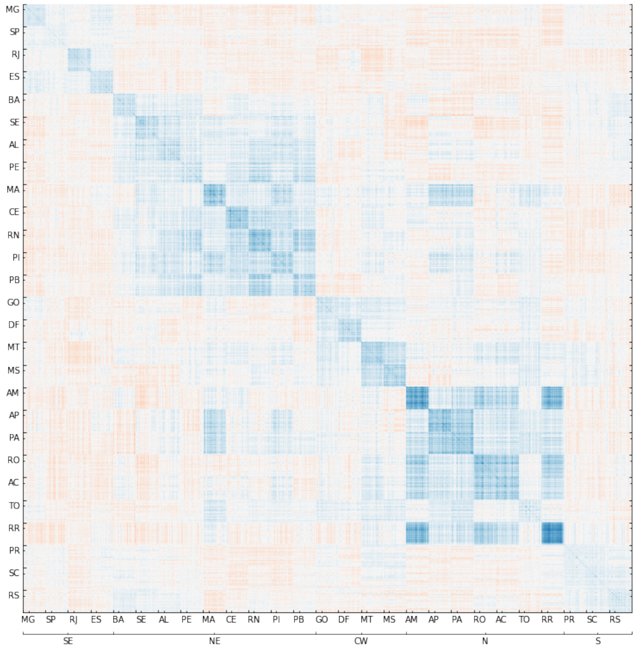

Fig. 3. Correlations between company latent embeddings, samples are grouped by state, and states are grouped into the 5 regions. We can see that the model is able to recover region-level information from the states information - $\beta = 1$

Fig. 4.

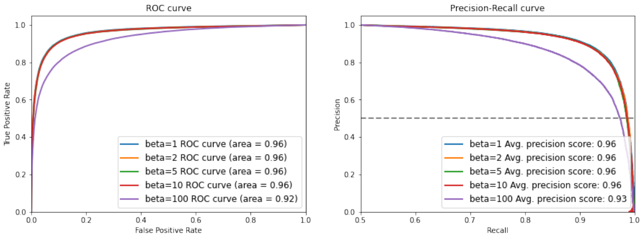

Fig. 5. Reconstruction performance of on a balanced test set

3.2 Downstream tasks

In this section we demonstrate how the latent embeddings generated by our model can be used for new downstream tasks. Using the representations produced with the 2015 data, we illustrate that it is possible to predict the probability of future edges on 2016 data with high accuracy and as well predicting the probability of a firm in the database being sanctioned in the future. Our procedure for both tasks consists in sampling a latent embedding from a trained

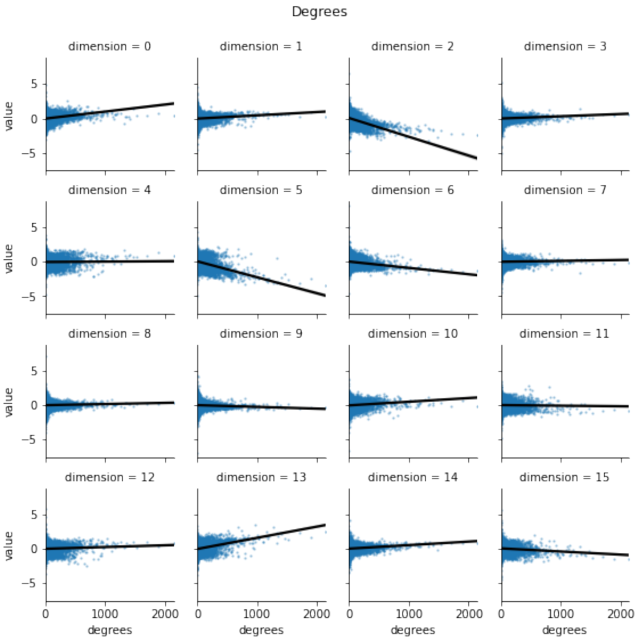

Fig. 6. Correlation between representation dimensions and node degree - $\beta = 100$

Gaussian Mixture VGAE model, with the auxiliary prediction task, on data from 2015. A logistic model with L1 regularisation [1] is then fitted to perform either downstream task. We repeat each experiment for each available β parameter.

Link Prediction For this task we sampled a balanced set of 20,000 edges from the graph generated on 2016 data. The edges are between companies that are present in both years 2015 and 2016. An edge is represented by the concatenation of the dot product between the edges of two companies and the element-wise multiplication of the two corresponding vectors. Performance is shown in Table 1 under **Link**.

Debarment Prediction A perhaps even more relevant task for our study is the prediction of whether a company will be debarred given its activity in the network. Unfortunately we are only able to match a modest amount of companies (85) in the network with the debarment data at our disposal. We use these and an equally sized sample of negative examples. The features used for this task are the latent representations produced for each company and the element-wise multiplication of these features with themselves. We achieve the results shown in Table 1 under **Debarment**.

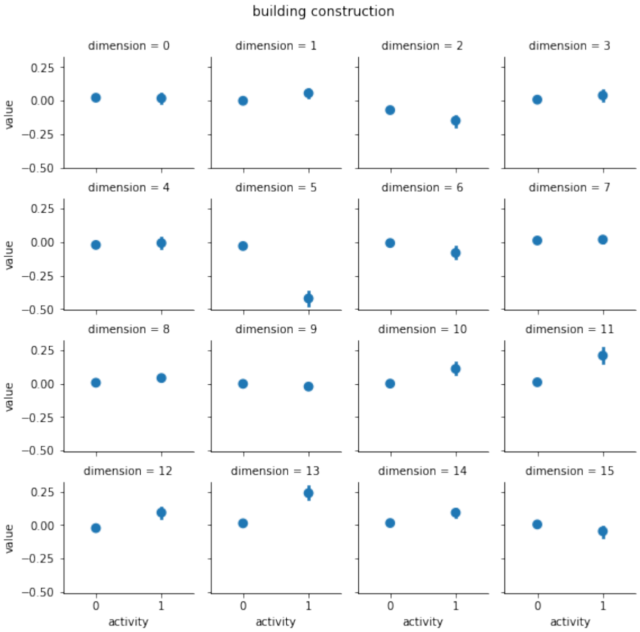

Fig. 7. Correlation between representation dimensions and sector of activity - $\beta = 100$

4 Conclusion

In this paper we propose a β–Graph Variational Autoencoder with a Gaussian mixture model and apply it to a novel data set, namely a network of buyers and sellers of public procurement in Brazil. We show that this model is capable of learning relevant properties of a procurement network and that the learned latent features not only capture feature relevant for reconstructing this network, but can also be used for potentially relevant downstream tasks. This model, and potential future developments, can be used by domain experts to better understand how complex transaction networks function and, more specifically, use the latent dimensions to simulate alternative scenarios of how procurement might

Table 1. Downstream task performance table

	Link		**Debarment**	
β	AUC	PR	AUC	PR
1	0.90	0.91	0.81	0.82
5	0.90	0.91	0.80	0.81
10	0.90	0.91	0.80	0.82
100	0.87	0.89	0.77	0.77

work. This might be particularly relevant to policy makers trying to anticipate how their actions might affect market dynamics.

References

1. Bishop, C.M.: Pattern recognition and machine learning. springer (2006)
2. Burgess, C.P., Higgins, I., Pal, A., Matthey, L., Watters, N., Desjardins, G., Lerchner, A.: Understanding disentangling in β-vae (2018)
3. Chassang, S., Kawai, K., Nakabayashi, J., Ortner, J.M.: Data driven regulation: Theory and application to missing bids. Tech. rep., National Bureau of Economic Research (2019)
4. ComprasNet: http://compras.dados.gov.br/
5. Compte, O., Lambert-Mogiliansky, A., Verdier, T.: Corruption and competition in procurement auctions. Rand Journal of Economics pp. 1–15 (2005)
6. Davis, P., Garcés, E.: Quantitative techniques for competition and antitrust analysis. Princeton University Press (2009)
7. Gerardino, M.P., Litschig, S., Pomeranz, D.: Can audits backfire? evidence from public procurement in chile. Tech. rep., National Bureau of Economic Research (2017)
8. Higgins, I., Matthey, L., Pal, A., Burgess, C., Glorot, X., Botvinick, M., Mohamed, S., Lerchner, A.: beta-vae: Learning basic visual concepts with a constrained variational framework. In: International Conference on Machine Learning (2017)
9. Honke, G., Higgins, I., Thigpen, N., Miskovic, V., Link, K., Duan, S., Gupta, P., Klawohn, J., Hajcak, G.: Representation learning for improved interpretability and classification accuracy of clinical factors from eeg. arXiv preprint arXiv:2010.15274 (2020)
10. Kingma, D.P., Welling, M.: Auto-encoding variational bayes (2014)
11. Kingma, D.P., Mohamed, S., Rezende, D.J., Welling, M.: Semi-supervised learning with deep generative models. In: Advances in neural information processing systems. pp. 3581–3589 (2014)
12. Kipf, T.N., Welling, M.: Variational graph auto-encoders (2016)
13. Laffont, J.J., Tirole, J.: A Theory of Incentives in Procurement and Regulation, vol. 1. The MIT Press, 1 edn. (1993), http://EconPapers.repec.org/RePEc:mtp:titles:0262121743
14. Locatello, F., Bauer, S., Lucic, M., Raetsch, G., Gelly, S., Schölkopf, B., Bachem, O.: Challenging common assumptions in the unsupervised learning of disentangled representations. In: international conference on machine learning. pp. 4114–4124. PMLR (2019)
15. OECD: https://www.oecd.org/gov/public-procurement/
16. Shi, W., Zhou, H., Miao, N., Zhao, S., Li, L.: Fixing gaussian mixture vaes for interpretable text generation. arXiv preprint arXiv:1906.06719 (2019)