The Latent Block Model: a useful model for high dimensional data

Keribin Christine*, Celeux Gilles, Robert Valérie
INRIA Saclay and Laboratoire de Mathématiques d'Orsay, Université Paris Sud, Université Paris Saclay,
F-91405 Orsay, France - christine.keribin@math.u-psud.fr

Abstract

The Latent Block Model (LBM) designs in a same exercise a clustering of the rows and the columns of a data array. Typically the LBM is expected to be useful to analyze huge data sets with many observations and many variables. But it encounters several numerical issues with big data set: maximum likelihood is jeopardized by spurious maxima and selecting a proper model is challenging since there are a lot of models are in competition. In this communication, we analyze these numerical issues. In particular, we make use of Bayesian inference to avoid spurious solutions and propose an efficient way to scan the model set. Moreover, we advocate the exact Integrated Completed Likelihood (ICL) criterion to select a proper and consistent LBM. The methods and algorithms will be ilustrated with pharmacovigilance data involving large arrays of data.

Keywords: mixture; model selection; ICL; pharmacovigilance.

1. Introduction

Clustering is an essential tool to discover hidden structure and knowledge from data by detecting groups of observations that are similar within a group and dissimilar from one group to another one. This is an unsupervised learning method, as the group membership is not known and has to be jointly discovered with the group characteristics. Mixture models is a flexible probabilistic approach to deal with clustering. A finite parametric mixture model assumes that the observations come from K several distinct populations called components, each one having is own distribution, but the group membership (which unit belongs to which component) is unknown. A frequent assumption is that the component distributions belongs to the same parametric family $\varphi(x;\alpha)$, so that the components only differ by a parameter value. Hence, the mixing density is

$$f(x) = \sum_{k=1}^{K} \pi_k \varphi(x;\alpha_k)$$

where $\pi_1, \ldots, \pi_K$ are the mixing weights. When $x \in R^d$ is multidimensional, the observations form a matrix with n rows (the observations) and d columns (the variables). The challenge of modern data is now to learn from data with large n and d, and the question is not only to cluster the observations, but also to cluster simultaneously the observations and the variables, leading to a tremendous parsimonious data representation. This is called co-clustering and has many applications in many fields such as recommendation systems (where the rows are the customers and the columns the goods, and we want to cross groups of customers and groups of goods), text mining (to co-cluster words and documents), genomics (to co-cluster genes and experimental conditions) for example. As for clustering there are many ways to perform co-clustering, and we will focus here on the latent block model (LBM), as it provides some challenging features on theoretical as well as methodological aspects. We first present the model and describe keys features (estimation, consistency and model selection). LBM is applied to deal with pharmacovigilance data. We propose a way to scan the model set and extend MLBM to cluster simultaneously two distinct sets of variables observed on the same statistical units

2. Latent Block Model

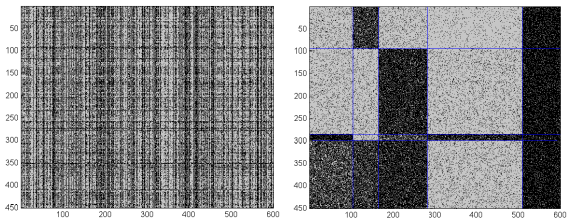

Figure 1: $n \times d = 450 \times 600$ observations (left) and their reorganization according to the underlying structure in 4×5 blocks (right)

The latent block model is a probabilistic model for co-clustering. It defines on a data matrix $X = (x_{ij})$ of n rows and d columns a block clustering structure as the Cartesian product of a row partition $\mathbf{z}$ by a column partition $\mathbf{w}$ with three main assumptions:

- row assignments (or labels) $\mathbf{z}_i$, $i = 1, \ldots, n$, are independent from column assignments (or labels) $\mathbf{w}_j$, $j = 1, \ldots, d$: $p(\mathbf{z}, \mathbf{w}) = p(\mathbf{z})p(\mathbf{w})$;
- row labels are independent, with a common multinomial distribution: $\mathbf{z}_i \sim \mathcal{M}(1, \boldsymbol{\pi} = (\pi_1, \ldots, \pi_g))$; in the same way, column labels are i.i.d. multinomial variables: $\mathbf{w}_j \sim \mathcal{M}(1, \boldsymbol{\rho} = (\rho_1, \ldots, \rho_m))$.
- conditionally to row and column assignments $(\mathbf{z}_1, \ldots, \mathbf{z}_n) \times (\mathbf{w}_1, \ldots, \mathbf{w}_d)$, the observed data X_{ij} are independent, and their (conditional) distribution $\varphi(., \alpha)$ belongs to the same parametric family, which parameter α only depends on the given block:

$$X_{ij} | \{z_{ik} w_{j\ell} = 1\} \sim \varphi(., \alpha_{k\ell})$$

  where z_{ik} is the indicator membership variable of whether row i belongs to row-group k and $w_{j\ell}$ is the indicator variable of whether column j belongs to column-group ℓ.

Hence, the complete parameter set is $\boldsymbol{\theta} = (\boldsymbol{\pi}, \boldsymbol{\rho}, \boldsymbol{\alpha})$, with $\boldsymbol{\alpha} = (\alpha_{11}, \ldots, \alpha_{gm})$. With these assumptions, the likelihood of the *complete data* is

$$p(\mathbf{x}, \mathbf{z}, \mathbf{w}; \boldsymbol{\theta}) = p(\mathbf{z}; \boldsymbol{\theta}) p(\mathbf{w}, \boldsymbol{\theta}) p(\mathbf{x} | \mathbf{z}, \mathbf{w}; \boldsymbol{\theta}) = \prod_{i,k} \pi_k^{z_{ik}} \prod_{j,\ell} \rho_\ell^{w_{j\ell}} \prod_{i,j,k,\ell} \varphi\left(x_{ij}; \alpha_{k\ell}\right)^{z_{ik} w_{j\ell}}$$

The labels are usually unobserved, and the *observed likelihood* is obtained by marginalization over all the label configurations:

$$p(\mathbf{x}; \boldsymbol{\theta}) = \sum_{\mathbf{z} \in \mathcal{Z}, \mathbf{w} \in \mathcal{W}} \left(\prod_{i,k} \pi_k^{z_{ik}} \prod_{j,\ell} \rho_\ell^{w_{j\ell}} \prod_{i,j,k,\ell} \varphi\left(x_{ij}; \alpha_{k\ell}\right)^{z_{ik} w_{j\ell}} \right)$$

LBM deals with matrix of homogeneous data, such as binary (Govaert and Nadif, 2008), Gaussian (Lomet, 2012), categorical (Keribin et al., 2015) or count (Govaert and Nadif, 2010) data. It involves a double missing data structure $\mathbf{z}$ for rows and $\mathbf{w}$ for columns, and the observed likelihood can not factorize as a product of the mixing density as for simple mixture models. This implies that the likelihood (and its logarithm) is rapidly not tractable event for few observations and few blocks, as the marginalization involves $k^n \times d^m$ terms. Due to this particularity, its study is very interesting, on a theoretical as well as a methodological point of view.

Estimation Recall first that the model is generically identifiable (identifiability up to a set a null measure): this was proved the for binary and categorical data (Keribin et al., 2015), and can be easily extended for distribution $\varphi(x; \alpha)$ whose simple mixtures are identifiable. Even if the observed likelihood is intractable for LBM, the maximum likelihood estimation with the EM algorithm could be thought to be feasible, as it does not involve the observed likelihood, but the complete which does not suffer the same difficulty. The E-step

requires to compute the expectation of the complete log-likelihood, conditionally to the observations and a given current parameter value $\theta^{(c)}$:

$$Q(\theta|\theta^{(c)}) = \sum_{i,k} s_{ik}^{(c)} \log \pi_k + \sum_{j,\ell} t_{j\ell}^{(c)} \log \rho_\ell + \sum_{i,j,k,\ell} e_{i,j,k,\ell}^{(c)} \log \varphi(x_{ij}; \alpha_{k\ell})$$

where

$$s_{ik}^{(c)} = P(z_{ik} = 1|\theta^{(c)}, \mathbf{x}), \quad t_{j\ell}^{(c)} = P(w_{j\ell} = 1|\mathbf{x}; \theta^{(c)}), \quad e_{i,j,k,\ell}^{(c)} = P(z_{ik}w_{j\ell} = 1|\mathbf{x}; \theta^{(c)}).$$

Unfortunately, computing $s_{ik}^{(c)}$, $t_{j\ell}^{(c)}$ and $e_{i,j,k,\ell}^{(c)}$ is again intractable, as it involves a marginalization that can not be factorized. However, the estimation can be performed either with numerical approximations (such as variational methods (VEM) imposing that the conditional joint distribution of the labels knowing the observations factorizes), or with a Bayesian approach (VBayes algorithm or Gibbs sampling).
VEM (Govaert and Nadif, 2008) presents several drawbacks such as its sensibility to starting values or its marked tendency to provide solutions with empty clusters, i.e. with fewer clusters than required after a maximum a posteriori classification rule. A possible way to attenuate the dependence of VEM to its initial values is to use a stochastic version of EM which incorporates a stochastic step between the E and M steps: the missing data are simulated according to their current conditional distribution. But a Gibbs sampling scheme is required to simulate the couple $(\mathbf{z}, \mathbf{w})$ as the joint conditional distribution cannot be exactly computed. Hence, SEM-Gibbs is not increasing the log-likelihood at each iteration, but it generates an irreducible Markov chain with a unique stationary distribution which is expected to be concentrated around the ML parameter estimate.
If SEM-Gibbs can bring an answer for the initialization, the problem of empty clusters still remains. Keribin et al. (2015) proposed to use Bayesian inference as a regularization tool for LBM and consider proper and independent non informative prior distributions for categorical LBM. However, as VEM, V-Bayes algorithm could be expected to be highly dependent on its initial values. Thus they recommended to initialize V-Bayes with the solution derived from the Gibbs sampler, which is easy to implement with conjugate priors.

Consistency Now, what are the asymptotic properties of the maximum likelihood estimator (MLE) and the variational estimator (VE)? This is a quite difficult theoretical question that has been recently solved. This was first studied on the Stochastic Block Model (SBM) which is a LBM with the same units in rows and columns, used to model graphs. In this case, there is only one set of latent variables $\mathbf{z}$. Celisse et al. (2012) proved in a seminal work that under the true parameter value, the conditional distribution of the assignments of a binary SBM converges to a Dirac of the real assignments. At the price of the strong assumption of the existence of an estimator of α converging at rate at least n^{-1}, they obtained the consistency of MLE and VE. Mariadassou and Matias (2015) presented a unified frame for LBM and SBM for observations coming from an exponential family, but cannot get rid off the previous assumption for consistency. Bickel et al. (2013) used a different approach: in the complete model where the labels are known, the consistency of the MLE is immediate. In case of a binary SBM, they showed that this property can be transferred to the MLE. Recently, Mariadassou et al. (2016) solved the consistency and the asymptotic normality of the MLE and VE for observations coming from an exponential family.

Model selection It remains to discuss model selection for LBM to choose a relevant number of blocks. This is an important challenge, especially for large data. BIC suffers of one drawback for LBM, as it needs the computation of the log-likelihood, which can only be approximated and whose quality cannot be assessed. On the other hand, the Integrated Completed Likelihood (ICL) criteria (Biernacki et al., 2000) defined as the logarithm of the integrated complete likelihood

$$ICL = \log p(\mathbf{x}, \mathbf{z}, \mathbf{w}|g, m) = \log \int p(\mathbf{x}, \mathbf{z}, \mathbf{w}|\theta; g, m) p(\theta; g, m) d\theta$$

is easily computed in exact and closed form using conjugate properties and is more appropriate for clustering as it favors well separated components. It is then well adapted for model selection in this case and the selected model has the maximum ICL. Notice that following Biernacki et al. (2000), the missing data $(\mathbf{z}, \mathbf{w})$

are replaced by their most probable inferred values $\widehat{\mathbf{z}}$, $\widehat{\mathbf{w}}$.

3. An application

The pharmacovigilance system aims at detecting as soon as possible potential associations between some drugs and adverse effects. An individual report consists in the list of prescribed drugs and observed effects. They are collected in a data matrix where rows represent the individual reports and columns are partitioned in two parts, one to report the drugs prescription and one to report the observed effects. The drug-columns are indicator memberships of whether this drug was prescribed and the effect-columns are indicator memberships of the presence of the effect. To give an idea, the World Health Organization data basis contained 3.7 millions of reports in 2004.

Several explanatory methods of automatic signal generation have been developed for over twenty years based on disproportionality measures see Marbac et al. (2016) for references. They are based on aggregated data (contingency table of drugs and effects), which suppose some homogeneity in the individuals. But it is reasonable to believe that the studied population is heterogeneous. Moreover, these matrices are so large that some drugs and adverse effects must be selected beforehand. LBM on contingency table can be considered in order to select subgroups of adverse effects and drugs with links.

The conditional density $\varphi(x, \alpha)$ for contingency table is Poisson, and following Govaert and Nadif (2010), the parameter $\alpha_{k\ell}$ for block (k, ℓ) is written as $\alpha_{k\ell} = \mu_i \nu_j \gamma_{k\ell}$ where μ_i denotes a row effect (two proportional rows will be in the same group), ν_j a column effect and $\gamma_{k\ell}$ an interaction for block (k, ℓ). Moreover, some identifiability conditions are needed:

$$\sum_i \mu_i = \sum_j \nu_j = \sum_{ij} x_{ij}.$$

This assures that $E(\sum_j x_{ij}) = \mu_i$ and $E(\sum_i x_{ij}) = \nu_j$ and naturally leads to estimate μ_i and ν_j beforehand by $\sum_i x_{ij}$ and $\sum_j x_{ij}$. Gibbs and V-Bayes algorithms are used with the following priors

$$\boldsymbol{\pi} \sim \mathcal{D}(a, \ldots, a), \quad \boldsymbol{\rho} \sim \mathcal{D}(a, \ldots, a), \quad \lambda_{k\ell} \sim \Gamma(\alpha, \beta), \tag{1}$$

$\mathcal{D}(a, \ldots, a)$ denoting a Dirichlet distribution and $\Gamma(\alpha, \beta)$ a Gamma distribution. The choice of hyperparameter values is important in Bayesian inference. Experiments on simulated data show that $a = 4$, as for categorical data, has a beneficial effect on spurious solutions, and then is hold in the following. The choice of α and β is more delicate. Simulations clearly advocate for $\delta = 1$. All $\beta \leq 1$ is also adequate. We then choose $\delta = 1$ and $\beta = 0.01$ to provide a distribution which is also weakly informative.

How to scan the model set ICL depends on a couple (g, m) and the set is much wider to explore than in the simple standard mixture model. We then propose a forward algorithm (Robert et al., 2016) that we call Bi-KM1. From a $g \times m$ partition, the $(g+1) \times m$ partition is analyzed. The algorithm is started from g different initializations, each being defined by a random split in two parts of one of the g components. The same is done for the $g \times (m+1)$ partition, and the model with the best ICL is kept. Although this search algorithm is sub-optimal, it is much more efficient that a random initialization on each model.

We compare it on simulated Poisson $\mathcal{P}(\lambda)$ count data with the greedy search (GS) of Wyse et al. (2014). In GS algorithm, labels are initialized with high values of g and m. Then, it randomly selects a row and allocates it to the row-class which improve ICL the most, taking into account the possible disappearance of a class. This process is also independently performed on the columns. Thus, the algorithm empties gradually the classes until convergence. Note that this procedure does not try to estimate the parameters of the underlying model as Bi-KM1 does, but only tries to define a partition. Both methods use the same priors and advocate the same hyperparameter values α and β, but have distinct preferences for a. For Bi-KM1, $a = 4$ avoids empty clusters while $a = 0.1$ for GS facilitates it. We use both algorithms with their predilection values on simulated data with different sizes (50×50, 100×100, 500×500) and three degrees of separation (easy, medium, difficult). In each cases, 100 matrices are generated. As GS is highly depending on the starting state, we take the best result after 10 random initializations. On the other hand, thanks to the Gibbs initialization, Bi-KM1 does not need to be initialized several times. For easy and medium cases, both algorithms are comparable, as they output same ICL values. But for the difficult case, Bi-KM1 outperforms GS for big sizes while GS has slightly better results for small data: this could be explained by the sensitivity to initialization of GS which is more pregnant for bigger sizes.

Pharmacovigilance data Data collected between 2000 and 2010 by AFFSAPS [1], represents 219 340 individual reports and involves 2142 drugs and 4216 adverse effects. In a first attempt to run standard explanatory methods, none of the reference couples (drugs, effects) already reported by the Observational Medical Outcomes Partnership (OMOP) are detected, neither does Poisson LBM. This could be explained by the fact that the reference set is not exhaustive, or by the weakness of the signal in the french database, or by some co-prescription artefact. In order to reinforce the signal, we construct a reduced contingency table with the individuals that only took one drug and have only one effect, discarding co-prescription artefact: this represents around 20% of the reports and concerns 1482 drugs and 2239 adverse effects. ICL on LBM Poisson selects 19×20 blocks. Signals detected by LBM on the complete contingency table are grouped in blocks of the reduced contingency table with highest intensity parameter. However, OMOP reference signals are not grouped in general in the same blocks.

4. Multiple Latent Block Model

The contingency table provides a summary of individual reports and a way to deal with large data, but with limitations. A new trend is to directly proceed the individual data. For example, Marbac et al. (2016) recently used logistic regression, but effect by effect. As so far, individual reports in pharmacovigilance are not directly analyzed on the whole, especially because of the amount of data. This direct analysis could get rid of the co-prescription artefact.

Individual data can be seen as a two binary matrices (one for the drugs, the other for the effects) sharing the same row units. We propose to use a new approach that adapts a LBM model to co-cluster rows and columns of two binary tables by imposing the same row clusters and call it Multiple Latent Block Model (MLBM). Let note $x = (x_{ij})_{n\times J}$ the matrix of drugs (J different drugs), and $y = (y_{ik})_{n\times K}$ the matrix of the adverse effects (K studied effects): $x_{ij} = 1$ whether individual i reports a prescribed drug j, 0 otherwise; $y_{ik} = 1$ whether individual i reports an adverse effect k, 0 otherwise. We define a latent structure of G row-groups, H drug-groups, and L effect-groups and assume that:

- row $\mathbf{z}_i$, $i = 1, \ldots, n$, drugs $\mathbf{v}_j$, $j = 1, \ldots, J$ and effects $\mathbf{w}_k$, $k = 1, \ldots, K$ assignments are independent
- row labels are i.i.d. multinomial variables $\mathbf{z}_i \sim \mathcal{M}(1, \boldsymbol{\pi})$; drug labels are i.i.d. multinomial variables $\mathbf{v}_j \sim \mathcal{M}(1, \boldsymbol{\rho})$ and effect labels are i.i.d. multinomial variables $\mathbf{w}_k \sim \mathcal{M}(1, \boldsymbol{\tau})$.
- conditionally to row, drug and effect assignments $(\mathbf{z}, \mathbf{v}, \mathbf{w})$, the observed data x are independent with a (conditional) binary distribution φ which parameter only depends on the given block.

The parameter to estimate is $\boldsymbol{\theta} = c(\boldsymbol{\pi}, \boldsymbol{\rho}, \boldsymbol{\tau}, \boldsymbol{\alpha}, \boldsymbol{\beta})$ and MLBM has the following density:

$$p(\mathbf{x}, \mathbf{y}; \boldsymbol{\theta}) = \sum_{\mathbf{z}\in\mathcal{Z}, \mathbf{v}\in\mathcal{V}, \mathbf{w}\in\mathcal{W}} \left(\prod_{i,g} \pi_g^{z_{ig}} \prod_{j,h} \rho_h^{v_{k\ell}} \prod_{k,\ell} \tau_\ell^{w_{k\ell}} \prod_{i,j,g,h} \varphi\left(x_{ij}; \alpha_{gh}\right)^{z_{ig}v_{jh}} \prod_{i,k,g,\ell} \varphi\left(y_{ik}; \beta_{g\ell}\right)^{z_{ig}w_{k\ell}} \right)$$

MLBM is generically identifiable and all the methodological principles (algorithms, ICL for model selection, scan of the model set) can be easily extended to MLBM (Robert et al., 2015). We test on simulated data with $n = 20\,000, J = 200, K = 400$ and 300 reference signals (drug, advers effect). Several difficulties arise: (i) the data are now sparse (98% of 0 in the AFFSAPS data) and the $\mathcal{B}(1, 1)$ prior for the binary parameter proposed by Keribin et al. (2015) is not pertinent anymore. In fact, a $\mathcal{B}(2, 100)$ prior, which density mode is around 0.02, is much more adapted. (ii) the performances are affected by the data size and the three dimensions of the model set. To make up for this problem, we propose a two steps analysis: (S1) we first use simple Poisson LBM on the contingency table to extract a $(H_{\min}, L_{\min})$ couple for the number of drug-clusters and effect-clusters that we can be thought to have also meaning in the individual representation. (S2) We then use MLBM with starting model size $(2, H_{\min}, L_{\min})$. Assignments for the columns are initialized with results of (S1) and a kmeans is performed to cluster the rows. Then, we run the extension of Bi-KM1 adapted for three dimensions.

Result of (S1) gives $(H_{\min} = 27, L_{\min} = 29)$ and we note that although the model generating the data is not LBM, the algorithm tends to group signals in the same clusters. The highest intensity clusters own

[1] We warmly thank research team B3PHI from Inserm UMR 1181, Villejuif, to give us access to this data

signal proportion varying from 13% to 50%, while these proportions would be of 0.3% at random, so that the concentration of signal is real in these particular blocks. (S2) Bi-KM1 selects the $(G = 50, H = 32, L = 40)$ model, with more groups for drugs and effects. We focus on couple of blocks of high intensity (high values for α_{gh} and $\beta_{g\ell}$). They own generally only few drugs and effects in the corresponding groups, and are quasi-systematically associated with signals. Moreover, the process tend to split individuals that took a drug with effect from those who took the drug without effect. In case of real data, it could helps to define risk profile.

5. Conclusions

LBM uses a probabilistic approach to perform co-clustering, with well known properties and efficient algorithms. We propose a new algorithm Bi-KM1 to scan for model selection, hence reducing time processing. Dealing with french pharmacovigilance data, we test LBM on the summarized contingency table, then extend LBM to cluster two distinct sets of variables to directly treat the individual reports. On simulated data of 20 000 reports, 200 drugs and 400 effects, MLBM give promising results. A challenging problem arises with real individual data, as their size does not allow to read drug and effect matrices simultaneously. Methods and algorithms must then be adapted to continue to take advantage of this model.

References

Bickel, P., Choi, D., Chang, X., Zhang, H., et al. (2013). Asymptotic normality of maximum likelihood and its variational approximation for stochastic blockmodels. *The Annals of Statistics*, 41(4):1922–1943.

Biernacki, C., Celeux, G., and Govaert, G. (2000). Assessing a mixture model for clustering with the integrated completed likelihood. *IEEE transactions on pattern analysis and machine intelligence*, 22(7):719–725.

Celisse, A., Daudin, J.-J., Pierre, L., et al. (2012). Consistency of maximum-likelihood and variational estimators in the stochastic block model. *Electronic Journal of Statistics*, 6:1847–1899.

Govaert, G. and Nadif, M. (2008). Block clustering with bernoulli mixture models: Comparison of different approaches. *Computational Statistics & Data Analysis*, 52(6):3233–3245.

Govaert, G. and Nadif, M. (2010). Latent block model for contingency table. *Communications in Statistics-Theory and Methods*, 39(3):416–425.

Keribin, C., Brault, V., Celeux, G., and Govaert, G. (2015). Estimation and selection for the latent block model on categorical data. *Statistics and Computing*, 25(6):1201–1216.

Lomet, A. (2012). *Sélection de modèles pour la classification de données continues*. PhD thesis, Université Technologique de Compiègne.

Marbac, M., Tubert-Bitter, P., and Sedki, M. (2016). Bayesian model selection in logistic regression for the detection of adverse drug reactions. *Biometrical Journal*, 58(6):1376–1389.

Mariadassou, M., Brault, V., and Keribin, C. (2016). Normalité asymptotique de lestimateur du maximum de vraisemblance dans le modele de blocs latents. *48èmes journés de Statistique de la SFdS*.

Mariadassou, M. and Matias, C. (2015). Convergence of the groups posterior distribution in latent or stochastic block models. *Bernoulli*, 21(1):537–573.

Robert, V., Celeux, G., and Keribin, C. (2015). Un nouveau modèle pour la pharmacovigilance. *47èmes journés de Statistique de la SFdS*.

Robert, V., Celeux, G., and Keribin, C. (2016). Modèle des blocs latents et sĺection de modèles en pharmacovigilance. *48èmes journés de Statistique de la SFdS*.

Wyse, J., Friel, N., and Latouche, P. (2014). Inferring structure in bipartite networks using the latent block model and exact icl. arXiv preprint arXiv:1404.2911.