Hybrid Bilinear and Trilinear Models for Exploratory Analysis of Three-Way Poisson Counts

Juha Raitio, Tapani Raiko, and Timo Honkela

Aalto University School of Science,
Department of Information and Computer Science,
P.O.Box 15400, FI-00076 AALTO, Finland
{juha.raitio,tapani.raiko,timo.honkela}@aalto.fi
http://ics.aalto.fi/

Abstract. We propose a probabilistic model class for the analysis of three-way count data, motivated by studying the subjectivity of language. Our models are applicable for instance to a data tensor of how many times each subject used each term in each context, thus revealing individual variation in natural language use. As our main goal is exploratory analysis, we propose hybrid bilinear and trilinear models with zero-mean constraints, separating modeling the simpler and more complex phenomena. While helping exploratory analysis, this approach leads into a more involved model selection problem. Our solution by forward selection guided by cross-validation likelihood is shown to work reliably on experiments with synthetic data.

Keywords: tensor factorization, multilinear model, unsupervised learning, exploratory data analysis, text analysis, Grounded Intersubjective Concept Analysis

1 Introduction

As a generic task, analysis of counts in relation to two categorical variables also known as factors, *ways* or *modes* is encountered in a vast number of scientific studies and engineering applications. In order to make the problem setting and evaluation of the present work more accessible, we concentrate on a concrete example from text analysis where the counts of selected words in a set of documents is represented as a *term-document matrix*, words indexed as rows and documents as columns. Common analyzes of this representation include relating the documents to each other by the counts of the word occurrences in documents, or studying the relation of words by their co-occurrences in documents. As such this comprises an example of *2-way data analysis*.

It may be of interest to additionally study how the counts of the term-document matrix vary according to some other factor such as the author of the document. In fact, if the variation between documents according to the author is included in analysis, one may be able to attribute some of the variation

in the word counts to the language use of the author, and consequently, give more accurate inference on relations between words (or documents) in general. The term-document data augmented with information about the authors has a representation as *3-way array*. 3-way data arrays, or in general multi-way data, can be studied for example using methods of *tensor data analysis*. For a generic introduction to the topic, see e.g. [13].

The present work is originally motivated by the finding that there can be substantial individual variation in how natural language expressions are used and interpreted. In [6], a method called Grounded Intersubjective Concept Analysis (GICA) has been introduced. The essence of the GICA method is to model individual variation in using natural language expressions and for this purpose, a 3-way analysis of Subject-Object-Context (SOC) tensors is needed. The analysis of such tensors may reveal individual differences in style but, more importantly, indicate subjectivity in modeling the relationship between language and the world. If this kind of subjectivity remains unrecognized, various kinds of problems related to communication may arise.

A more specific motivation for the present work stems from the fact that in [6] the analysis of Subject-Object-Context tensors was conducted by flattening the 3-way arrays to 2-way matrices. These matrices can then be straightforwardly analyzed using traditional data analysis methods such as PCA, SVD or ICA. Each direction of flattening introduces a point of view and may, as such, provide important insights into the data when analyzed. However, the flattening of the original data appears to be useful, but possibly inadequate approach. It appears necessary to devise a methodology that would make it possible to analyze all the relationships without first determining which modes of the array are in focus. As discussed above, traditional term-document matrices are formed by counting the number of instances of each term in each document and by storing this count in the element that corresponds to the row associated with the term in question and the column associated with the particular document. The GICA data is formed following the same basic principle, but adding a third mode which is used to include all the subjects being included in the analysis. Moreover, rather than considering frequency counts in whole documents, the counts concern typically some context window of a given length.

One might think that subjectivity of language would be an exception rather than a rule, since semantics appear to be well defined through thesauri, ontologies, and other knowledge representations. However, as natural language is immersed with ambiguity, there is also a great amount of subjectivity and contextuality involved. A more detailed account on this matter is provided in [6]. Here it may be sufficient to refer to two examples. For the basic color terms, there seems to be a high degree of intersubjective agreement. Around the idea of prototypical red, green or blue there is not much subjective variation even though a particular context may shift the evaluation, like in the case of phrases "red skin" or "red wine" [3]. However, a lot more subjective variation is to be expected if less typical color names are considered, such as "purple", "khaki" or "orchid". An even more convincing example is when abstract words are consid-

ered. It is unlikely that all people mastering English would understand words like "democracy/-tic", "fair", "love", "wellbeing", or even "computation" in a mutually compatible manner. It should be obvious that there is variation in the interpretation in the use of these and many other words. Tools for the formal modeling and systematic analysis of this kind of semantic variation are not readily available and widely used, though. The Subject-Object-Context tensors [6] and the methodological development presented in this paper aim to alleviate the lack of tools and to provide a systematic framework for approaching this common but mostly unexplored phenomenon.

In this paper, we thus propose an unsupervised method for analyzing 3-way count data – including GICA data – where a 2-way analysis based on decomposition models is extended to allow 3-way analysis in the same framework in a probabilistic manner. The complexity of the original phenomenon is very high and the same concerns the data in question. With the methodology presented in this paper, it should become possible to explore the data so that useful conclusions can be made. In particular, findings that show that there is significant level of variation in the interpretation of some expression even if the context is the same.

2 Proposed Model

The count x_{ijk} indexed by the levels i, j and k in the ranges $\{1, 2, \ldots, I\}$, $\{1, 2, \ldots, J\}$ and $\{1, 2, \ldots, K\}$ of the three modes under consideration is modeled as Poisson distributed

$$P(x_{ijk}) = \text{Pois}(\exp(l_{ijk})), \tag{1}$$

where the *trilinear predictor*

$$\begin{aligned} l_{ijk} =& a_i^{(0)} + b_j^{(0)} + c_k^{(0)} \\ &+ \sum_{m_1=1}^{h_1} a_{im_1}^{(1)} b_{jm_1}^{(2)} + \sum_{m_2=1}^{h_2} b_{jm_2}^{(1)} c_{km_2}^{(2)} + \sum_{m_3=1}^{h_3} c_{km_3}^{(1)} a_{im_3}^{(2)} \\ &+ \sum_{m_4=1}^{h_4} a_{im_4}^{(3)} b_{jm_4}^{(3)} c_{km_4}^{(3)}. \end{aligned} \tag{2}$$

This specifies a model class that predicts the logarithm of the Poisson mean count by a specially structured trilinear model consisting of

- *bias parameters* $\mathbf{a}^{(0)}$, $\mathbf{b}^{(0)}$, and $\mathbf{c}^{(0)}$ for capturing the mean of each mode,
- all combinations of the *bilinear factorizations* with parameters $a_{i:}^{(q)}$, $b_{j:}^{(q)}$ and $c_{k:}^{(q)}$, $q = 1, 2$ for capturing interactions between modes,
- the *trilinear factorization* or the PARAFAC model [4] with parameters $a_{i:}^{(3)}$, $b_{j:}^{(3)}$ and $c_{k:}^{(3)}$ for capturing 3-way interactions between modes, and

– *hyperparameters* h_1, h_2, h_3 and h_4 for adjusting the model complexity,

where the subscript ":" is used to denote all values of the index of summation m within a factorization.

Without loss of generality, we assume that the vectors $\mathbf{a}^{(q)}$, $\mathbf{b}^{(q)}$, and $\mathbf{c}^{(q)}$, $q = 1, 2, 3$ are zero-mean in the sense that $\sum_i a_{im}^{(q)} = 0$, $\sum_j b_{jm}^{(q)} = 0$ and $\sum_k c_{km}^{(q)} = 0$ for all m (see Appendix for proof). These parameter vectors are also known as *loadings*.

The proposed model class can be interpreted as statistical multiple regression models, where a Poisson distributed count is regressed on three categorical (factorial) independents. The dimension of the parameter space is the number of parameters in a specific model, $I + J + K + h_1(I + J) + h_2(J + K) + h_3(I + K) + h_4(I + J + K)$. In the special case of $h_1 = h_2 = h_3 = h_4 = 0$ our specification is linear and equals that of a Generalized Linear Model [12] with logarithmic, canonical link function for Poisson distributed data. Our model is, however, nonlinear in parameters in its general form.

2.1 Motivation: Exploratory Analysis

The reason we propose a combination of bilinear and trilinear terms instead of only the trilinear part, is the exploratory analysis of the results: We wish each phenomenon in the data to be modeled with as simple terms as possible. Since the trilinear part is often the most interesting but also the most difficult one for analysis, we hope to clarify it by separating the more trivial phenomena away. It is easy to see that the trilinear term could emulate the other terms by using constant loadings $\mathbf{1}$ for parameter vectors $\mathbf{a}$, $\mathbf{b}$ or $\mathbf{c}$. However, since we introduce the zero-mean constraint, we force the simpler terms to be used, too.

In the GICA context, the interpretation of the terms in Equation (2) is as follows. I is the number of people (or subjects), J is the number of terms (or objects) and K is the number of contexts. Biases describe how much text we have from each subject, and how common is each term and each context. The first bilinear term models how people prefer using some objects (or terms). This part is comparable to collaborative filtering. The second bilinear term is about how terms are used in contexts (or documents). This part is comparable to latent Dirichlet allocation. The third bilinear term models how common particular contexts are for different people, again comparable to collaborative filtering. The trilinear term can model the subjectivity of context to the use of terms.

2.2 Parameter Estimation

The parameter vectors **a**, **b**, and **c** are learned by fitting the model to i.i.d. data. The log-likelihood of the parameters is

$$\ln \prod_{i,j,k} P(x_{ijk} \mid l_{ijk}) = \sum_{i,j,k} \ln \mathrm{Pois}(x_{ijk} \mid \exp(l_{ijk})) \tag{3}$$

$$= \sum_{i,j,k} \ln \frac{\exp(l_{ijk})^{x_{ijk}} \exp(-\exp(l_{ijk}))}{x_{ijk}!} \tag{4}$$

$$= \sum_{i,j,k} [x_{ijk} l_{ijk} - \exp(l_{ijk}) - \ln(x_{ijk}!)] \tag{5}$$

and its partial derivative w.r.t. l_{ijk} is

$$\frac{\partial \ln \prod P}{\partial l_{ijk}} = x_{ijk} - \exp(l_{ijk}). \tag{6}$$

The gradient for fitting the parameters is further derived using the chain rule. Finding maximum likelihood estimates is subject to the zero-mean constraints of the parameter vectors.

2.3 Model Selection

The proposed algorithm for model selection is as follows. First, we set the hyperparameters $h_1 = h_2 = h_3 = h_4 = 0$ to estimate the biases $a_i^{(0)}$, $b_j^{(0)}$, $c_k^{(0)}$. Then model complexity is increased by incrementing the hyperparameters one at a time and thus introducing new components into the model. The new parameters are fitted while keeping the old ones fixed.

To avoid overfitting, proper hyperparameter values are determined by cross-validation [14], i.e., by splitting the tensor elements randomly into a number of equal sized partitions and then, in turn, holding out each partition from the parameter estimation as validation set. We stop increasing each hyperparameter whenever the probability of validation set, that is, its evidence for the model, stops increasing significantly. In cross-validation we compare the distribution of changes in the model evidences of the validation sets between before and after adding new parameters. We apply a non-parametric test (Wilcoxon signed-rank) to compare the significance level of the increase to a critical value.

After determining the hyperparameters, thus fixing the model complexity, the model parameters are estimated without holding out any data, and at the end, the whole model is fine-tuned by estimating all the parameters simultaneously.

3 Simulation Experiment

In order to assess our contribution, comprising of a statistical model class together with parameter estimation, model selection and data analysis procedures

Table 1. Summary of the experiment for identification of random models. Table A (*on the left*) displays the count for each value of the hyperparameters in the sample of 100 models. Table B (*on the right*) displays the accuracy of the method in the identification test in terms of the error rate for each hyperparameter and value, and in total over the types of factorizations and complexities.

A	h_1	h_2	h_3	h_4	tot.
$h_: \equiv 0$	35	33	35	36	139
$h_: \equiv 1$	31	33	32	39	135
$h_: \equiv 2$	34	34	33	25	126
tot.	100	100	100	100	400

B	h_1	h_2	h_3	h_4	tot.
$h_: \equiv 0$	0	0	0	0	0
$h_: \equiv 1$	0	0	0.03	0.10	0.04
$h_: \equiv 2$	0	0.03	0	0.12	0.03
tot.	0	0.01	0.01	0.07	0.02

proposed in Sect. 2, we apply them to synthetic data generated using 100 random models in the proposed model class. For this purpose we first draw each hyperparameter for a model uniformly from $\{0, 1, 2\}$, and then draw values for the respective parameter vectors uniformly from $[-1, 1]$ and remove their mean. The drawn models are summarized in Table 1.A. Finally, we sample size $40 \times 25 \times 15$ ($I \times J \times K$) data tensors from these models.

In the simulation we identified each of the models independently, applying the proposed model selection procedure using 10-folds in cross-validation and critical value of 0.15 for entering new components. In the models that generated the 100 data tensors, we had in total 400 hyperparameter values to identify. The method failed in 7 of the hyperparameter values for the 3-way components (h_4) and in 2 for the 2-way components (h_2 and h_3). In all but one case out of the 9, the error was that one true generating component was excluded from the identified model. Once, for h_2, one extra component was included. In total 91 out of the 100 generating models were identified correctly. Table 1.B summarizes the results in identification accuracy.

In overall, according to this experiment, the model selection procedure is feasible. It seems that model estimation works surprisingly well despite the lack of guarantees for finding the global optimum. Failures in the identification may be due to suboptimal parameter values or to the sampling of cross-validation data. Consequently, the improvement in the model evidence by introduction of some components have not been considered significant in our conservative model selection procedure. It is interesting to note that off-by-one errors in the identification do not seem to induce further errors in subsequently identified components and hence the estimation procedure can be considered robust in this respect.

4 Discussion

Our method estimates the trilinear predictor tensor as a sum of finite number of constraint rank-one tensors, i.e., as constraint CANDECOMP[1]/PARAFAC[4] trilinear decomposition. Once this representation has been found, it follows from

the properties of the decomposition that the trilinear components are unique up to permutation and scaling of the parameter vectors under certain sufficient conditions [11] that are true in most real world data analyses.

It is known that in general the approximation of a tensor by the trilinear decomposition is an ill-posed problem [15] that does not have a bounded solution for some degenerate tensors. Also the greedy approach we apply (but not depend on) for fitting the decomposition incrementally, does not result in best fit in the sense that the optimal parameters of a less complex model are not generally optimal in a more complex model [9]. Both of these results are derived for approximations based on the Frobenius norm. We are not aware of results that are valid for our probabilistic metric. Additionally, there are results (e.g. [10]) that in real world applications the trilinear composition, fitted in the greedy manner, gives comparable performance besides its lower computational cost than the fitting of all of the parameters simultaneously.

Our model is a probabilistic generative model as opposed to traditional tensor factorization models. One benefit from this is the well-founded handling of missing values. We used missing values for the model selection by holding out validation elements in the tensor, but in general, the original data might contain missing elements, too. As the proportion of the missing values increases, modelling the posterior uncertainty of the parameters starts to become important. Our approach resembles the basic 2-way model in [7] and could be extended to the more advanced treatment such as variational Bayes.

A tensor of size $I \times J \times K$ can be factorized in many different ways, see [8] for a review on tensor factorization. We build upon the CANDECOMP/PARAFAC model in the multilinear predictor, that is, using factors $I \times h$ and $J \times h$ and $K \times h$, except that we include the simpler factorizations for explaining the other phenomena. Another difference is of course that we use it hierarchically as a parameter for the Poisson distribution. Tucker decomposition [5] is the oldest tensor factorization method, which uses $h_1 \times h_2 \times h_3$ and $I \times h_1$ and $J \times h_2$ and $K \times h_3$. It can be used also in computing nonnegative tensor factorizations [2]. Recently, an algorithm for solving a set of factorization problems with possibly coupled factors was given in [16].

References

1. Carroll, J.D., Chang, J.J.: Analysis of individual differences in multidimensional scaling via an n-way generalization of "eckart-young" decomposition. Psychometrika 35(3), 283–319 (1970), http://dx.doi.org/10.1007/BF02310791
2. Friedlander, M.P., Hatz, K.: Computing nonnegative tensor factorizations. Computational Optimization and Applications 23(4), 631–647 (March 2008)
3. Gärdenfors, P.: Conceptual Spaces. MIT Press (2000)
4. Harshman, R.: Foundations of the PARAFAC procedure: Model and conditions for an 'explanatory' multi-mode factor analysis. UCLA Working Papers in phonetics (16) (1970)
5. Hitchcock, F.L.: The expression of a tensor or a polyadic as a sum of products. Journal of Mathematics and Physics 6, 164–189 (1927)

6. Honkela, T., Raitio, J., Nieminen, I., Lagus, K., Honkela, N., Pantzar, M.: Using GICA method to quantify epistemological subjectivity. In: Proceedings of IJCNN 2012, International Joint Conference on Neural Networks (2012)
7. Ilin, A., Raiko, T.: Practical approaches to principal component analysis in the presence of missing values. Journal of Machine Learning Research (JMLR) 11, 1957–2000 (July 2010)
8. Kolda, T.G., Bader, B.W.: Tensor decompositions and applications. SIAM Review 51(3), 455–500 (2009)
9. Kolda, T.G.: Orthogonal tensor decompositions. SIAM Journal on Matrix Analysis and Applications 23(1), 243–255 (July 2001)
10. Kolda, T.G., Bader, B.W., Kenny, J.P.: Higher-order web link analysis using multilinear algebra. In: ICDM 2005: Proceedings of the 5th IEEE International Conference on Data Mining. pp. 242–249 (November 2005)
11. Kruskal, J.B.: Three-way arrays: Rank and uniqueness of trilinear decompositions, with application to arithmetic complexity and statistics. Linear Algebra and its Applications 18, 95–138 (1977)
12. McCullagh, P., Nelder, J.A.: Generalized linear models (Second edition). London: Chapman & Hall (1989)
13. Mørup, M.: Applications of tensor (multiway array) factorizations and decompositions in data mining (2011), http://onlinelibrary.wiley.com/doi/10.1002/widm.1/full
14. Picard, R.R., Cook, R.D.: Cross-validation of regression models. Journal of the American Statistical Association 79(387), pp. 575–583 (1984), http://www.jstor.org/stable/2288403
15. de Silva, V., Lim, L.H.: Tensor rank and the ill-posedness of the best low-rank approximation problem. SIAM J. Matrix Analysis Applications 30(3), 1084–1127 (2008), http://dblp.uni-trier.de/db/journals/siammax/siammax30.html#SilvaL08
16. Yilmaz, Y.K., Cemgil, A.T., Simsekli, U.: Generalised coupled tensor factorisation. In: Shawe-Taylor, J., Zemel, R.S., Bartlett, P.L., Pereira, F.C.N., Weinberger, K.Q. (eds.) NIPS. pp. 2151–2159 (2011), http://dblp.uni-trier.de/db/conf/nips/nips2011.html#YilmazCS11

Appendix: Removing Mean of Parameter Vectors

Without loss of generality, we can assume that $\sum_i a_{im}^{(q)} = 0$, $\sum_j b_{jm}^{(q)} = 0$ and $\sum_k c_{km}^{(q)} = 0$ for all m, $q = 1, 2, 3$ in Equation (2). This is because any non-zero mean in parameter vectors could be moved to a simpler term. For instance the mean $\mu_m^{(a1)}$ of $a_{im}^{(1)}$ could be moved to $b_j^{(0)}$ by noting that for all i and j

$$b_j^{(0)} + a_{im}^{(1)} b_{jm}^{(2)} = [b_j^{(0)} + \mu_m^{(a1)} b_{jm}^{(2)}] + [a_{im}^{(1)} - \mu_m^{(a1)}] b_{jm}^{(2)}. \tag{7}$$

For removing the mean $\mu_m^{(a3)}$ of $a_{im}^{(3)}$, we can increase h_2 by 1 and set the new part to

$$b_{jh_2}^{(1)} = \sqrt{\mu_m^{(a3)}} b_{jm}^{(3)}, \tag{8}$$

$$c_{kh_2}^{(2)} = \sqrt{\mu_m^{(a3)}} c_{km}^{(3)}. \tag{9}$$