Bayesian Approach For Model Selection In Tensor Decompositions

M. S. Bennet Praba(A.P)[1], Tharun. J[2], S.B. Rishik[3], Akilan Ganesan[4], M.Mageswaran[5]
[1, 2, 3, 4, 5] SRMIST

Abstract- *Various tensor decompositions use different arrangements of factors to explain multi-way data. Components from different decompositions can vary in the number of parameters. Allowing a model to contain components from different decompositions results in a combinatoric number of possible models. The correct use of model evaluation and model selection techniques is vital in academic machine learning research. We consider model selection to balance approximation error with the number of parameters, but due to the number of possibilities post-hoc model selection is infeasible. Instead, we incrementally build a model. Estimating the adequate number of components is an important yet difficult problem in multi-way modelling. We demonstrate how a Bayesian framework for model selection based on automatic relevance determination (ARD) can be adapted to the Tucker and CandeComp/PARAFAC (CP) models.*

Keywords- tensor decompositions, multi-way arrays, model selection, bayesian information criterion.

I. INTRODUCTION

Tensor decompositions are in frequent use today in a variety of fields including psychometrics, chemometrics, image analysis, web data mining, bio-informatics, neuroimaging, and signal processing. Tensors are also known as multi-way arrays, multidimensional matrices or hypermatrices are generalizations of vectors (first order tensors) and matrices (second order tensors). The two most commonly used decompositions of tensors are the Tucker model and the more restricted canonical decomposition (CandeComp) and Parallel Factor Analysis (PARAFAC) model. We will presently denote the CandeComp/PARAFAC model as CP.

The Tucker model represents the data spanning the nth modality by the vectors (loadings) given by the J n columns of A (n) such that the vectors of each modality interact with the vectors of all remaining modalities with strengths given by a so-called core tensor G. As a result, the Tucker model encompasses all possible linear interactions between vectors pertaining to the various modalities of the data. The CP model is a special case of the Tucker model where the size of each modality of the core array G is the same while interaction is only between columns of same indices such that the only non zero elements are found along the diagonal of the core. Thus, the CP model can be expressed as a Tucker model with diagonal core.

In particular, by appropriate scaling of each component the CP model can be expressed as a Tucker model with unit diagonal core. The Tucker model can in turn be expressed as the CP model by duplicating components of different indices to form additional CP components. Notice, in the Tucker model a rotation of a given loading matrix A(n) can be compensated by a counter rotation of the core G, . For the CP model it is not possible in general to rotate the loading and still keep the core diagonal. Thus, the CP model is in general unique up to scale and permutation.As the CP model corresponds to the Tucker model with diagonal core Tucker decompositions in which only some off diagonal elements are non-zero can be considered a representational interpolation between the Tucker and CP decomposition. Hence, whereas the Tucker model encompass all potential interaction between the components of each modality through the core array G, the CP model only allow for interactions between columns of A(n) with same indices. The sparse Tucker model can be considered a model between the Tucker and CP model where interactions are present within a few of the components across the various modalities. Several strategies exist for simplifying the Tucker core.

The Tucker solution was rotated such that the Tucker core would have as many small loadings as possible. Thus, by regularizing the Tucker model excess components can be turned off and the Tucker core be simplified. By assigning priors for the model parameters and learning the hyper parameters of these priors the method is able to turn off excess components and simplify the core structure at a computational cost of fitting the conventional Tucker/CP model. To investigate the impact of the choice of priors we based the ARD on both Laplace and Gaussian priors corresponding to regularization by the sparsity promoting l_1 norm and the conventional l_2 norm, respectively. While the form of the priors had limited effect on the results obtained the ARD approach turned out to form a useful, simple, and efficient tool

for selecting the adequate number of components of data within the Tucker and CP structure.

For the Tucker and CP model the approach performs better than heuristics such as the Bayesian information criterion (BIC), Akaikes information criterion (AIC), DIFFIT and the numerical convex hull (NumConvHull) while operating only at the cost of estimating an ordinary CP/Tucker model. For the CP model the ARD approach performs almost as well as the core consistency diagnostic (CorConDiag). We will presently estimate the adequate degree of regularization by a Bayesian approach named automatic relevance determination (ARD). Two types of regularization will be considered; the sparsity promoting l_1 regularization as well as the more conventionally applied l_2 ridge regression regularization. The approach readily generalize to the CP model and will also here be used to estimate the number of components. Choosing the right model is in particular challenging in the Tucker model as the number of components is specified for each modality separately. This renders heuristics such as the DIFFIT, numerical convex hull (NumConvHull), Bayesian information criterion (BIC) and Akaikes information criterion (AIC) as well as cross-validation approaches computationally expensive as n models have to be evaluated. Furthermore, while model selection for the CP model has been guided by heuristics based on the core consistency diagnostic (CorConDiag), no such heuristics exist for the Tucker model. In two-way analysis it is common to evaluate the eigenvalue spectrum and truncate the singular value decomposition (SVD). Although this approach does not have a straightforward multi-linear counterpart approximate approaches have been given forming the fast DIFFIT. However, this approach can not account for additional constraints such as non-negativity.

In conclusion, no efficient approach for the estimation of the number of components in the Tucker model is known. Thus, the aim of this paper is to use regularization to turn off excess components in the CP and Tucker model and thereby select the model order and simplify the core and to optimize the amount of regularization from data and to achieve these objective at the cost of estimating a conventional multi-way mode

II. RELATED WORKS

The greedy approach [1] can estimate a model consisting of combination of tensor decompositions. This approach is analogous to sparse coding with a union of dictionaries. Linear synthesis models are fundamental to multivariate signal processing tasks such as denoising, compression, and classification. In this work we explore an approach to approximate data arranged in a tensor, or multi-way array, via a combination of data-dependent bases. The bases are chosen from two or more sets each estimated by tensor decomposition yielding orthogonal components. Truncated singular value decomposition (SVD) finds the optimal reduced rank approximations of data stored in matrices. In multivariate signal processing, there may be multiple ways the signal can be arranged before approximation. For instance, if the signal is arranged as a tensor, then a large number of decomposition/approximation models have been proposed that exploit structure along different modes of the data. Tensor decompositions that can be written as a summation of component tensors each formed as tensor outer products. Decompositions of this nature are of interest because orthogonality can be enforced on any of the factors of the outer product and the resulting components tensors will be orthogonal.

Intristic Bayesian Factor's [3] is completely automatic Bayes factors, in that they are based only on the data and standard non informative priors. Note, however, that issues such as the "optimal" choice of training samples in dependent-data situations are yet to be resolved. It also seem to correspond to actual Bayes factors for reasonable "intrinsic priors," thus attaining a type of "second-order" Bayesian correspondence; in contrast, most other default methods achieve (at best) a first-order correspondence with Bayesian methods, with many having a systematic bias in favour of the more complex model. Compared with other "second-order" Bayesian methods, such as the method of Jeifreys, IBF's have the advantage of being very generally applicable. IBF's apply to non nested, as well as nested, model comparisons and can be applied to any distributions. IBF's can also be used for default Bayesian hypothesis testing. IBF's can be applied in situations in which even the usual Bayesian asymptotics (e.g., BIC) does not apply. IBF's can be used for default multiple model comparison and prediction. IBF's are invariant to univariate transformations of the data. If suitably invariant non informative prior distributions are used, they are also invariant to choice of the parametrizations of the models.

The PARAFAC decomposition is known as one of the most commonly used tools in tensor signal/data processing. Unfortunately, its classical algorithms barely take the potential statistical and/or deterministic prior information of the decomposed tensor into consideration, while the modern ones are usually problem oriented, which limits their applications [4].To fill in this gap, the PARAFAC decomposition of a tensor is brought into the framework of Bayesian inference.

By introducing transition models for the loading/factor matrices of a tensor, the PARAFAC decomposition can be formulated as an alternating Bayesian filter. By means of the flexibility of the Bayesian filter, the proposed filtering decomposition approach illustrated can cover two commonly used priors – parametric and time transition ones. Under the linear Gaussian assumption, the proposed in [2] filter can be implemented as an Alternating (Matrix) Kalman Filter. Analyses show that the performance of the proposed filter is similar to the reported ALS based algorithms when priors are unavailable. The results of numerical simulations show that our Bayesian approach outperforms the reported PARAFAC decomposition algorithms in literature, especially for the cases where the statistical and/or deterministic priors are offered such as the target tracking application of a bi static ULA MIMO radar system.

III. INFORMATION CRITERION

Information criteria offer a computationally appealing way of estimating the generalization performance of the model. In model selection AIC [5] and the BIC [6] have traditionally been used as simple approximations to the expectation of the negative log likelihood and the model evidence respectively . Here, the number of components are selected such that the following two quantities are minimized.

$$AIC = -2 \log L + K = S \log(SSE/S) + K$$
$$BIC = -2 \log L + K \log S = S \log(SSE/S) + K \log S$$

Where L is the likelihood of the model, K is the number of parameters in the model, and S is the number of datapoints. For least square estimation this reduces to the expressions to the right where SSE is the sum of squared error. Thus, the criteria defines a trade off between reduction in reconstruction error and complexity of the model. Notice that BIC tends to penalize model complexity more heavily than AIC, hence, gives a more conservative estimate of what is considered the best model.

For the Tucker model the DIFFIT procedure has been proposed to estimate the adequate number of components [7]. In the DIFFIT procedure, all potential models are evaluated and The DIFFIT for the m th model is then calculated as

$$DIF(m) = ExpVar(m) - ExpVar(m-1)$$
$$DIFFIT(m) = DIF(m)/DIF(m+1)$$

And the model with largest DIFFIT value taken to be the most adequate model when disregarding DIFFIT values based on too small values of DIF . Hence, the optimal model is given by the model that has the largest contribution to the explained variance relative to consecutive models corresponding to the region of maximal curvature in the graph of {m, ExpVar}. An approximate evaluation of DIFFIT forming the fastDIFFIT [9] is given by evaluating the eigenvectors of $X_{(n)}$ for all n-modes and take the best models R_m formed by the higher order singular value decomposition (HOSVD) .A refinement of the above approach correcting for the number of free (FP) for parameters the pth Tucker model form the NumConvHull approach [8] given by

$$FPDIF(p) = FP(p) - FP(p-1)$$
$$NumConvHull(p) = (DIF(p) / FPDIF(p))$$
$$(DIF(p+1) / FP(p+1))$$

For the CP model the core consistency has been used as a heuristic to access the adequate number of components [10]. The core consistency measures the degree of cross-talk between the components of the CP model by estimating the corresponding Tucker model core G given the CP loadings.

Since the Tucker model encompass all potential interactions between components of the various modes non-zero values in the off diagonal of the Tucker core indicate that structure in components of different indices over the modalities can combine resulting in so-called cross-talk. Too many components will result in a strong degree of cross-talk across the loadings of the modes thus will yield a low value of the CorConDiag. Too few components on the other hand will exhibit a low degree of cross-talk. Thus, a heuristic for the 'correct' number of components is taken to be just before a major drop-off in the graph of {d, CorConDiag} in [10].ARD is a hierarchical Bayesian approach widely used for model selection [11].

In ARD hyper parameters explicitly represents the relevance of different features by defining the range of variation for these features, usually by modelling the width of a zero-mean Gaussian prior imposed on the model parameters. If the width becomes zero, the corresponding feature cannot have any effect on the predictions. Hence, ARD optimizes these hyper parameters to discover which features are relevant. While ARD based on Gaussian priors can prune excess components Gaussian priors do not in general admit sparse representation within the active components hence does not necessarily favour simple parsimonious representations. The reason being that the l_2-regularization penalizes elements by their squares and as such penalizes large values relatively more than small values. The Laplace prior on the other hand is known to admit sparse representation as it corresponds to a l_1 regularization thus is the closest convex proxy to minimizing for the number of non-zero elements in the model .

IV. IMPLEMENTATION

Data Sets Used

1. Synthetic data

A data set with Tucker(3,4,5) structure was randomly generated with size 30 × 40 × 50. All the factors as well as the core array were drawn from a normal N(0,1)-distribution, i.e. with zero mean and variance is 1. Gaussian noise was added to the data such that SNR = 0 dB.

2. Flow injection analysis

This data set is given by the absorption spectra over time for three different chemical analytes measured in 12 samples with different concentrations, i.e. 12(samples) × 100(wavelengths) × 89(times), ideally this dataset form a Tucker(3,6,4) model.

3. Amino acid fluorescence

This data set contains the excitation and emission spectra of five samples of different amounts of tyrosine, tryptophane and phenylalanine forming a 5(samples) × 51(excitation) × 201(emission) array. Hence the data can be described by a three component CP model.

4. Sugar process data

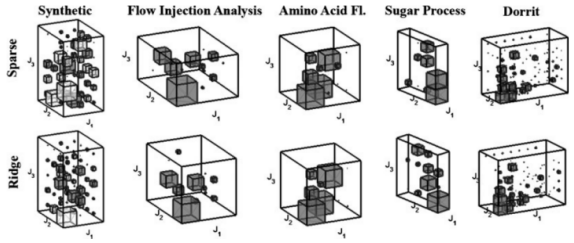
This data set contain emission and excitation spectra measurements in 265 samples forming a 265(samples) × 571(emissions) × 7(excitations) array. The data was modelled by a four component CP model where the number of components were estimated based on an extensive split half analysis.

5. Dorrit fluorescence data

This data set contains the emission and excitation spectra of 27 synthetic samples containing different concentrations of four chemical analytes forming a 27(samples) × 551(emissions) × 24(excitations) array . The data is adequately modelled by a four component CP model. Since the components of the four chemometrics data sets are non-negative the estimated models for these data were constrained to be non-negative.

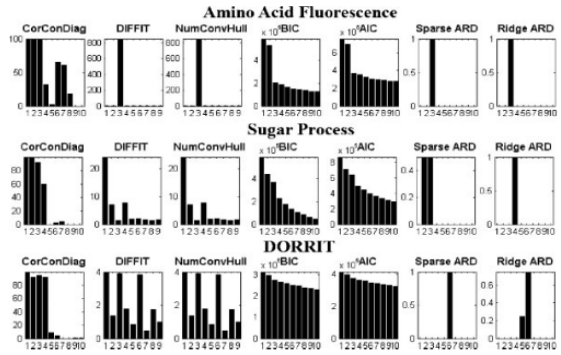
TUCKERS ANALYSIS

The impact of the choice of signal to noise ratio SNR is visible in the visualization of the models. For the synthetic data a clear breakpoint around SNR = 0 dB is found such that lower SNR values identify the correct model order for both sparse and ridge ARD Tucker whereas higher SNR values makes the ARD approach completely fail in identifying the correct number of components as the model fit noise. A similar behaviour is found for the remaining data sets. Namely that high SNR values tend to overestimate the number of components whereas low SNR values perform more stable. As such, the exact choice of SNR seem to have little impact on the model order found as long as SNR is not set too large. Thus, when there is no prior information as to the true SNR of the data it seems to be better to use low estimates of the SNR rather than large SNR values as large SNR values has a tendency to use too many components hence overfit the data. In the following analysis the SNR is set to SNR = 0 dB.



CP ANALYSIS

The estimated number of components for the three data sets can be found using the CorConDiag, DIFFIT/NumConvHull, BIC, AIC and sparse as well as ridge ARD CP. Notice how both BIC and AIC as for the Tucker analysis fail in estimating the adequate number of components. This is because the Tucker model and in particular the CP model are highly restricted models using only a few parameters to model a large amount of data. Thus, the complexity terms in BIC and AIC grows in general more slowly than the improvement in log(SSE) thus they tend to favour too complex models. The CorConDiag correctly identifies 3 components in the amino acid fluorescence data, 3–4 components in the sugar process data and 4 components in the Dorrit fluorescence data. The DIFFIT and NumConvHull correctly indicates a 3 component model for the amino acid fluorescence data but wrongfully a 1 component model for the Sugar process and a 1,3 or 6 component model for the Dorrit data. Both the sparse and ridge ARD methods correctly identified 3 components in the amino acid fluorescence data, for the sugar process the sparse ARD indicate a 2 or 3 component model whereas the ridge ARD correctly identifies a 4 component model. For the Dorrit data both sparse and ridge ARD indicate a 6 component model.



Thus, while the proposed ARD approach here perform better than heuristics such as DIFFIT/NumConvHull, BIC and AIC the CorConDiag seem to work somewhat better in estimating the number of components in the CP model. While the ARD approach outperforms heuristics such as DIFFIT, NumConvHull, AIC and BIC when estimating the number of components in the Tucker and CP model its found that the CorConDiag performed slightly better in estimating the adequate number of components in the CP model. The modelling inadequacies encountered for the ARD CP and Tucker is probably due to incorrect estimates of the SNR, deviation from Gaussianity in the noise, deviation from Gaussian and Laplace distributed components, the fact that the parameters were based on simple MAP estimates and finally due to limited amount of data for the identification of the model order. The reason why no approach correctly established the Tucker(3, 6, 4) structure of the flow injection analysis data and Tucker(4, 4, 4) of the sugar process data is because models with less components almost perfectly accounts for all data (VarExp > 0.99). On the other hand, for the Dorrit data the sparse ARD and ridge ARD failed in correctly identifying 4 components as excess components were able to model substantial parts of the data. Despite the different nature of the Gaussian and Laplace priors the results found based on the two priors were similar. This is because the ARD framework first and foremost turn off excess components while components that remain active are little influenced by the prior if their parameters are large.

Hence, if the d th component of the nth mode is important then its corresponding alpha will be small rendering the prior non informative and as a result give little effect in the estimation of that component. Thus, while the ARD framework effectively can turn off excess components the choice of prior seems to only have a limited effect on the components identified. Rather than estimating Sigma $_2$ from data where Sigma $_2$ is defined from a user given signal to noise ratio (SNR). The results obtained was only to a small degree sensitive to the defined SNR as long as the SNR was not set to high causing the model to over fit the data. Hence, although this parameter is user defined the actual choice of the parameter only has a limited impact on the models obtained.

VI. FUTURE SCOPE

Model selection is perhaps one of the most challenging problems in unsupervised learning. Its demonstrated how a simple Bayesian framework based on ARD could be adapted to multi-way models such as the Tucker and CP models.

Presently, each component of each mode was given its own prior and the priors were either solely Laplace or Gaussian, however, it's noted that other parametrizations of the priors are conceivable. Furthermore, only the most simple framework considered where loadings and hyper parameters were based on MAP estimation. Within the proposed Bayesian framework more involved methods based on sampling approaches to estimate model parameters as well as expectation propagation for the evaluation of predictive performance can be employed to further improve the model order estimation. This should be investigated in future work.
Finally, the ARD approach can only shrink models, i.e. remove components. Thus, once a component has been removed it can no longer be brought back. In particular this requires that J_n be chosen large enough to encompass all potential models. Future research should investigate methods that can adapt the ARD approach to grow if initialized by a model order that is too small.

VII. CONCLUSION

This Bayesian approach is computationally inexpensive as the method automatically removes excess components when estimating the model contrary to existing heuristics that requires the estimation and evaluation of all potential models. The proposed ARD framework forms an efficient tool forthe automatic estimation of components in the Tucker models and in the analysis of both synthetic and real data the method indeed effectively extracted reasonable number of components.

While the ARD approach outperforms heuristics such as DIFFIT, NumConvHull, AIC and BIC when estimating the number of components in the Tucker and CP model we found that the CorConDiag performed slightly better in estimating the adequate number of components in the CP model. The modeling inadequacies encountered for the ARD CP and Tucker is probably due to incorrect estimates of the SNR, deviation from Gaussianity in the noise, deviation from Gaussian and Laplace distributed components, the fact that the parameters were based on simple MAP estimates and finally

due to limited amount of data for the identification of the model order.

REFERENCES

[1] Austin J. Brockmeier, Jose C. Principe, Anh Huy Phan and Andrzej Cichocki - "A greedy algorithm for model selection of tensor decompositions." 2013 IEEE International Conference on Acoustics, Speech and Signal Processing

[2] Ming Shi, Dan Li and Jian - "Qiu Zhang An Alternating Bayesian Approach to PARAFAC Decomposition of Tensors' 2018 IEEE Access, Volume: 6 Pages: 36487 – 36499[3] James O. Berger and Luis R. Pericchi - "The Intrinsic Bayes Factor for Model Selection and PredictionJournal of the American Statistical Association" Vol. 91, No. 433 (Mar., 1996), pp. 109-122

[3] Qibin Zhao, Guoxu Zhou, Liqing Zhang, Andrzej Cichocki and Shun-Ichi Amari - "Bayesian Robust Tensor Factorization for Incomplete Multiway Data."IEEE Transactions on Neural Networks and Learning Systems (Volume: 27, Issue: 4 , April 2016)

[4] Thanh Huy Nguyen, Umut Şimşekli, Gaël Richard and Ali Taylan Cemgil - "Efficient Bayesian Model Selection in PARAFAC via Stochastic Thermodynamic Integration." IEEE Signal Processing Letters (Volume: 25 , Issue: 5 , May 2018)

[5] Akaike H. - "A new look at the statistical model identification." IEEE Trans. Automat. Contr. 1974; 19(6): 716–723.

[6] Bro R, Kjeldahl K, Smilde AK and Kiers HAL. - "Cross-validation of component models: a critical look at current methods." Anal. Bioanal. Chem. 2008; 390(5): 1241–1251.

[7] Timmerman ME and Kiers HAL. - "Three-mode principal components analysis: choosing the numbers of components and sensitivity to local optima." Br. J. Math. Stat. Psychol. 2000; 53: 1–16.

[8] Ceulemans E and Kiers HAL. Selecting among three-mode principal component models of different types and complexities: a numerical convex hull based method. Br. J. Math. Stat. Psychol. 2006; 59(1): 133 – 150.

[9] Kiers HAL and der Kinderen A. - "A fast method for choosing the numbers of components in tucker3 analysis." Br. J. Math. Stat. Psychol. 2003; 56: 119–125.

[10] Bro R and Kiers HAL. "A new efficient method for determining the number of components in parafac models." J. Chemom. 2003; 17(5): 274–286.

[11] (Alan) Qi Yuan, Minka TP, Picard RW, Ghahramani Z. - "Predictive automatic relevance determination by expectation propagation." ICML'04: Proceedings of the 21st International Conference on Machine Learning. ACM: New York, NY, USA, 2004; 85.