Latent Composite Likelihood Learning for the Structured Canonical Correlation Model: Supplementary Material

Ricardo Silva
Department of Statistical Science and CSML
University College London
RICARDO@STATS.UCL.AC.UK

Abstract

We present four pieces of supplementary material: first, an approach for the CDN inference problem of computing likelihood functions, which for our purposes we believe it is simpler to implement than other approaches presented in the literature; second, a discussion of the convergence of LEARNSTRUCTUREDCCA-II; third, brief comments on identification and initialization; fourth, details on the preprocessing of the NHS data.

1 SIMPLER CDN INFERENCE

An efficient procedure for transforming CDFs into PMFs is given in detail by Huang et al. (2010), which is particularly sophisticated and seemingly hard to implement. However, one can reduce the problem of computing PMFs from CDFs following the structure of Equation (5) – itself just a rearrangement of the general formulation (Joe, 1997) for binary variables: just introduce "pseudo" random variables corresponding to the difference indicators $\mathbf{Z}$ and construct the corresponding factor graph. Notice that the term $(-1)^{\sum_{i=1}^{p} z_i}$ is itself a product of univariate factors over the pseudo set $\mathbf{Z}$. Equation (5) is the "marginal" of a pseudo distribution $\mathcal{P}(\mathbf{Z}, \mathbf{Y})$ and can be found by any standard exact method of inference. We used junction trees. Figure 1 shows an example of reducing the problem of computing the PMF of graph $Y_1 \leftrightarrow Y_2 \leftrightarrow Y_3$. The result is analogous in the continuous case: one just have to create indicator variables that pick which factors are being derived and which are not.

This simple link is not mentioned in previous papers, to the best of our knowledge. In any case, the customized method described by Huang et al. (2010) readily includes details on how to generate parameter gradients, and it is useful as a framework for developing approximate algorithms (as already hinted by Huang and Frey, 2008): in our case, the pseudo "distribution" $\mathcal{P}(\mathbf{Z}, \mathbf{Y})$ can take negative values, therefore some standard variational methods cannot be directly applied to this representation (e.g., mean-field EM requires the logarithm of the joint).

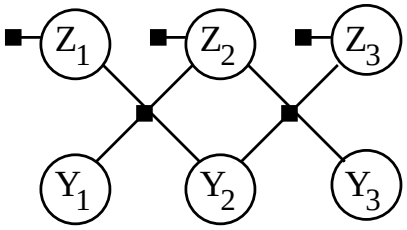

Figure 1: A factor graph representation for the "pseudo-distribution" of $Y_1 \leftrightarrow Y_2 \leftrightarrow Y_3$ induced by Equation (5).

2 CONVERGENCE

It is not obvious that LEARNSTRUCTUREDCCA-II converges, since it alternates between the optimization of two objective functions with respect to two parameter sets (i.e., we optimize $\mathcal{Q}_{\mathcal{G}_m}^{(\beta,\Sigma,\{q_{mn}(\cdot)\})}$ with respect to $\{\beta, \Sigma\}$, and we optimize $\mathcal{F}_{\mathcal{G}_m}^{(\beta,\Sigma)}$ with respect to $\mathcal{G}_m$).

Consider the two functions in Table 1. It is clear that

$$\mathcal{B}_{\mathcal{G}_m}^{(\beta,\Sigma,\{q_{mn}(\cdot)\})} \leq \chi_{\mathcal{G}_m}^{(\beta,\Sigma,\{q_{mn}(\cdot)\})}$$

since $\log \mathcal{P}(A, B) \leq \log \mathcal{P}(A)$ for any pair of events A and B. It is also the case that

$$\chi_{\mathcal{G}_m}^{(\beta,\Sigma,\{q_{mn}(\cdot)\})} \leq \mathcal{F}_{\mathcal{G}_m}^{(\beta,\Sigma)}$$

by Jensen's inequality.

It is clear that optimizing $\chi_{\mathcal{G}_m}^{(\beta,\Sigma,\{q_{mn}(\cdot)\})}$ with respect to $\{\beta, \Sigma\}$ is equivalent to optimizing $\mathcal{Q}_{\mathcal{G}_m}^{(\beta,\Sigma,\{q_{mn}(\cdot)\})}$. However, this does not guarantee that $\mathcal{B}_{\mathcal{G}_m}^{(\beta,\Sigma,\{q_{mn}(\cdot)\})}$ will also increase, or at least not decrease[1]. Similarly, the same applies for the relationship between optimizing $\mathcal{F}_{\mathcal{G}_m}^{(\beta,\Sigma)}$ with

[1]Consider the following simpler case: for a trivariate dis-

respect to $\mathcal{G}_m$, and again whether $\mathcal{B}_{\mathcal{G}_m}^{(\beta,\Sigma,\{q_{mn}(\cdot)\})}$ will not decrease.

Hence, although a clean proof of convergence seems elusive at this point, all the simulations showed not only convergence, but we observed monotone convergence with respect to $\mathcal{F}_{\mathcal{G}_m}^{(\beta,\Sigma)}$. Since it is true that updating $q_{mn}(\cdot)$ to $\mathcal{P}(\Theta_{mn} \mid \mathbf{Y}_{mn}^{1:N}, \beta, \Sigma, \mathcal{G}_m)$, for a fixed set of parameters and structure, will not decrease $\mathcal{B}_{\mathcal{G}_m}^{(\beta,\Sigma,\{q_{mn}(\cdot)\})}$ (Neal and Hinton, 1998), a sufficient condition for convergence is that the lowest bound $\mathcal{B}_{\mathcal{G}_m}^{(\beta,\Sigma,\{q_{mn}(\cdot)\})}$ does not decrease as parameters and structure are updated (assuming "convergence" here means find a local optimal of $\mathcal{B}_{\mathcal{G}_m}^{(\beta,\Sigma,\{q_{mn}(\cdot)\})}$).

3 IDENTIFIABILITY AND INITIALIZATION

Identifiability is relevant not only to validate the use of composite likelihood, but for the overall interpretation of the resulting structure. A full analysis of the identifiability of the model space is beyond of the scope of this paper. It is possible nevertheless to borrow the main results from Silva et al. (2006) to establish sufficient conditions: if X_m mutually d-separates three of its observed children Y_a, Y_b, Y_c plus a fourth observed variable (child of X_m or not), all of them being mutually dependent, then the coefficients of Y_a, Y_b, Y_c are identifiable. If every pair $\{X_m, X_n\}$ jointly d-separates two children of X_m and two children of X_n, then correlation σ_{mn} is identifiable. If the coefficients of a given Y_i and a given Y_j are identifiable, as well as the correlation of their latent parents in $\mathbf{X}_S$ (if different), then their copula coefficient is identifiable.

A more formal analysis and the relationship between parameter identifiability and structure search is left for future work. For now, we use these conditions to motivate a parameter initialization procedure. Given the recent success on carefully designed initialization methods for complex non-convex optimization problems in latent variable modeling (Hinton et al., 2006), it is also of interest to propose such methods for the structured CCA problem.

In this case, we suggest creating a tabu list of coefficients and triplets of vertices such that no bi-directed edges among these vertices are allowed at the beginning, and coefficients are also not allowed to change over iterations. This is done by considering, for each latent variable X_i, all subsets of size three among its children. The score of a triplet $\{Y_a, Y_b, Y_c\}$ is defined by the log-likelihood of the model that has $\{Y_a, Y_b, Y_c\}$ and a fourth observed variable as children of a single latent variable[2]. The triplet of the highest score is considered tabu, and the coefficients obtained by fitting this 4-variable model are used during the search procedure without modification. After the algorithm converges, the tabu list is removed and the procedure continues from that point until no new edge modifications are introduced. The motivation is assuming that the true model satisfies the basic identifiability conditions, and the goal is to identify one relevant triplet for each latent variable.

In a preliminary study, we compared the effect of the aforementioned initialized procedure in LEARNSTRUCTUREDCCA-II against a random initialization by sampling coefficients from independent standard Gaussians. In 30 synthetic studies, the simpler initialization performed slightly worse on average in terms of edge omission and parameter fitting, and errors were more spread out. It also took longer to converge.

A suggestion for future work is that identifiability conditions can also be weakened, in the spirit of Hoyer et al. (2008).

4 THE NHS DATASET

The NHS survey contained 37 sections which in principle could be used as 37 latent variables. However, we filtered these variables according to the following criteria:

- some sections were applied only to a subset of the population (e.g., Mental Health staff received a different version of Section 6, concerning training). We removed those;
- questions with a very high empirical probability of 0 or 1 (> 0.97) were removed to speed the procedure up;
- some sections had conditional subquestions (e.g., Question 8, on appraisals, had three questions that depended on a positive answer to a fourth question). Those were removed too;
- sections with fewer than 4 items were also removed in order to make the problem harder for the initialization procedure;
- finally, we decided to use the 11 questions relating to quitting the position (Section 12) and overall job satisfaction (Section 13) in our testing stage. That is, we did not model them as part of our latent variable model.

The remaining sections, 9 in total, were used to derive the partition. Here we list the corresponding sections and questions derived from the preprocessing, as published by (Care Quality Comission and Aston University, 2010):

tribution $\log \mathcal{P}(Y_1, Y_2, Y_3;\ \theta)$, optimizing $\log \mathcal{P}(Y_1 \mid Y_2; \theta) + \log \mathcal{P}(Y_1 \mid Y_3; \theta)$ is not a guarantee that $\log \mathcal{P}(Y_1 \mid Y_2, Y_3;\ \theta)$ will increase.

[2]We do that in two stages: first, we find a latent variable X_j whose children have the highest canonical correlation score with the children on X_i. Then the fourth variable is any element in the union of the children of X_i and X_j.

Table 1: Components of a Pairwise Composite Likelihood Score Function

$$\mathcal{B}_{\mathcal{G}_m}^{(\beta,\Sigma,\{q_{mn}(\cdot)\})} = \sum_{m<n}\sum_{Y_i\in\mathcal{S}_m}\sum_{Y_j\in\mathcal{S}_n}\int q_{mn}(\Theta_{mn})\log\frac{\mathcal{P}(\mathbf{Y}_{mn}^{1:N}\mid\mathcal{G}_m,\beta,\Sigma,\theta_{ij})}{q_{mn}(\Theta_{mn})}d\,\Theta_{mn} + \frac{1}{|\mathcal{S}|-1}\sum_{m=1}^{|\mathcal{S}|}\sum_{n\neq m}\sum_{\{Y_i,Y_j\}\subset\mathcal{S}_m}\int q_{mn}(\Theta_{mn})\log\frac{\mathcal{P}(\mathbf{Y}_{mn}^{1:N}\mid\mathcal{G}_m,\beta,\Sigma,\Theta_{mn})}{q_{mn}(\Theta_{mn})}d\,\Theta_{mn}$$

$$\chi_{\mathcal{G}_m}^{(\beta,\Sigma,\{q_{mn}(\cdot)\})} = \sum_{m<n}\sum_{Y_i\in\mathcal{S}_m}\sum_{Y_j\in\mathcal{S}_n}\int q_{mn}(\Theta_{mn})\log\frac{\mathcal{P}(\mathbf{Y}_{i}^{1:N},\mathbf{Y}_{j}^{1:N}\mid\mathcal{G}_m,\beta,\Sigma,\theta_{ij})}{q_{mn}(\Theta_{mn})}d\,\Theta_{mn} + \frac{1}{|\mathcal{S}|-1}\sum_{m=1}^{|\mathcal{S}|}\sum_{n\neq m}\sum_{\{Y_i,Y_j\}\subset\mathcal{S}_m}\int q_{mn}(\Theta_{mn})\log\frac{\mathcal{P}(\mathbf{Y}_{i}^{1:N},\mathbf{Y}_{j}^{1:N}\mid\mathcal{G}_m,\beta,\Sigma,\Theta_{mn})}{q_{mn}(\Theta_{mn})}d\,\Theta_{mn}$$

S3: Flexibility of working, questions 5–9, 11 (e.g., "My employer offers working reduced hours")

S4: Types of training provided by the Trust, questions 12–16 (whether the staff member has taken "Any supervised on-the-job training")

S7: Statements about immediate manager, questions 36–40 ("My immediate manager... gives me clear feedback on my work")

S14: Statements about responsibilities and workload, questions 71–76 ("I do not have time to carry out all my work")

S15: Relationship to workmates, questions 77–82 ("The people I work with treat me with respect")

S16: Statements about the Trust where staff member works, questions 83–89 ("Care of patients / service users is my Trust's top priority")

S19: Opportunities at work, questions 106–110 ("There are opportunities for me to progress in my job")

S20: Statements about working in the NHS, questions 111–116 ("I understand the national vision for the NHS")

S22: statements about improving work practices, questions 120–124 ("I am able to make suggestions to improve the work of my team / department")

Figure 2 shows the resulting network over the final selected 45 variables. Light blue points do not represent connections: instead, they highlight the given partition. Yellow points represent connections between observed variables that measure different latent variables, and red points connect variables within the same partition set.

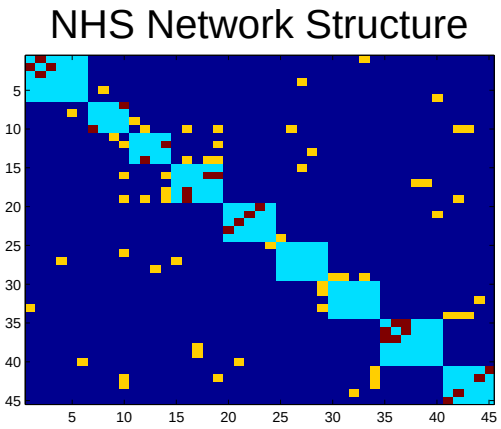

Figure 2: The learned NHS network. Figure best seem in color.

References

Care Quality Commission and Aston University. Aston Business School, National Health Service National Staff Survey, 2009 [computer file]. *Colchester, Essex: UK Data Archive [distributor], October 2010. Available at* HTTPS://WWW.ESDS.AC.UK, *SN: 6570*, 2010.

G. Hinton, S. Osindero, and Y. Teh. A fast learning algorithm for deep belief nets. *Neural Computation*, 18: 1527–1554, 2006.

P. Hoyer, S. Shimizu, A. Kerminen, and M. Palviainen. Estimation of causal effects using linear non-Gaussian causal models with hidden variables. *International Journal of Approximate Reasoning*, 49:362–378, 2008.

J. Huang and B. Frey. Cumulative distribution networks and the derivative-sum-product algorithm. *Proceedings of the 24th Conference on Uncertainty in Artificial Intelligence*, 2008.

J. Huang, N. Jojic, and C. Meek. Exact inference and

learning for cumulative distribution functions on loopy graphs. *Advances in Neural Information Processing Systems*, 23, 2010.

H. Joe. *Multivariate Models and Dependence Concepts*. Chapman-Hall, 1997.

R. Neal and G. Hinton. A view of the EM algorithm that justifies incremental, sparse and other variants. *In M. Jordan (Ed.), Learning in Graphical Models*, pages 355–368, 1998.

R. Silva, R. Scheines, C. Glymour, and P. Spirtes. Learning the structure of linear latent variable models. *Journal of Machine Learning Research*, 7:191–246, 2006.