

NIPS data analysis

Miaoyan Wang, Jan 14, 2021

The NIPS dataset [?] records the counts of words in 2,483 papers published from 1987 to 2003. We focus on the top 100 authors, 200 most frequent words, and normalize each word count by log transformation with pseudocount 1. The resulting dataset is an order-3 tensor $\mathcal{Y} \in \mathbb{R}^{100 \times 200 \times 17}$, where the entry $\mathcal{Y}(i, j, k)$ represents the log count of word i by author j in year k .

We first investigate the prediction accuracy of tensor completion. Table 2 compares the performance of our method, low-rank CP decomposition, and naive imputation by sample average. The reported MAE is averaged over five runs of cross-validation, with 20% entries for testing and 80% for training. We find that our method substantially outperforms the classical low-rank method for every r in the considered range. Further increment of rank appears to have little effect on the performance; doubling the rank to $r = 20$ yields only 5% relative deduction in MAD (see Appendix). In contrast, the low-rank CPT exhibits poor prediction even at a relatively high rank. The comparison highlights the advantage of our method in achieving accuracy while maintaining low complexity. In addition, the two tensor methods outperform the naive imputation, implying the necessity of incorporating tensor structure in the analysis.

Method	$r = 3$	$r = 6$	$r = 9$	$r = 12$	$r = 15$
NonparaT (Ours)	0.18 (0.002)	0.16 (0.002)	0.15 (0.001)	0.14 (0.001)	0.13 (0.001)
Low-rank CPT	0.22(0.004)	0.20(0.007)	0.19(0.007)	0.17(0.007)	0.17(0.007)
Naive imputation (Baseline)	0.32(.001)				

Table 2: Prediction accuracy measured in MAE in the NIPS data analysis. The reported MAEs are averaged over five runs of cross-validation, with standard errors in parentheses. For low-rank CPT, we use R function `rTensor` with default hyperparameters, and for our method, we set $H = 20$.

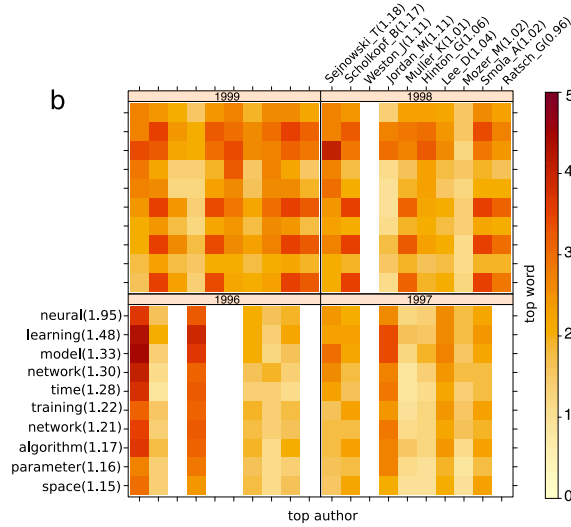


Figure 1: Estimated signal tensor with top authors and words in the years 1996-1999. Gray regions indicate no publication for the author in the specific year. Authors and words are ranked by marginal averages based on $\hat{\Theta}$, where the marginal average is denoted in the parentheses.

Next, we examine the estimated signal tensor $\hat{\Theta}$ from our method with $r = 10$ and the resolution parameter $H = 20$. The result changes little with higher values of r and H . We define $\hat{\Theta}(i, j, k)$ by our nonparametric estimate (6) if author i publishes at least one paper in year k , and zero otherwise. Figure 1 illustrates the subtensor of $\hat{\Theta}$ consisting of top authors and words (after excluding generic words such as *figure*, *current*, etc). The detected pattern is consistent with the active topics in the NIPS publication. For example, among the top words are *neural* (marginal mean = 1.95), *learning* (1.48), and *network* (1.21), whereas top authors are *T. Sejnowski* (1.18), *B. Scholkopf* (1.17), *M. Jordan* (1.11), and *G. Hinton*. We also find strong heterogeneity among word occurrence across authors and years. For example, *training* and *algorithm* are popular words for *B. Scholkopf*, and A. Smola in 1998-1999, whereas *model* occurs more often in *M. Jordan* in 1996. The achieved denoising accuracy and pattern detection demonstrates the applicability of our method. (would be interesting to verify the specific papers in that year.)