for $i$-th data point from $j$-th source (say $F_{ji}$). In slice sampling, $F_{ji}$ grows gradually based on the slice variable and requires sampling of $F_{ji}$ new $\beta_k$ variables using ARS. On the other hand, the approximate Gibbs sampler, to sample $F_{ji}$, approximates an intractable integral (Eq (3.22) in [8]) using Monte Carlo samples. For the synthetic experiments described above, slice sampler takes about 100 iterations to converge and runs in 6.67 minutes. On the other hand, the approximate Gibbs sampler [8] requires about 1000 iterations to converge to the true number of factors taking totally 52.3 minutes. This timing analysis is performed using a Windows PC with Intel i7@3.4 GHz and 8 GB RAM.

5.2 Experiments-II: Real Data

5.2.1 Results using NIPS 0-12 Dataset

Our first real-world dataset is the NIPS 0-12 dataset, which contains the articles from the proceedings of Neural Information Processing Systems (NIPS) conference published between 1988 and 1999. In this dataset\(^1\), text articles are divided into nine different sections/tracks plus one miscellaneous section/track. We work with nine sections which are Cognitive Science (CS), Neuroscience (NS), Learning Theory (LT), Algorithms and Architecture (AA), Implementations (IM), Speech and Signal Processing (SSP), Visual Processing (VP), Applications (AP) and Control, Navigation and Planning (CNP). We treat each section as one data source, each section has its own focus and differs in distribution. NHFA exploits the underlying sharing across different sections while still retaining the focus of individual section by maintaining a hierarchy. Second, this allows us to compare our results with two baseline methods (i.e. Gupta et al [8] and Teh et al [18]) as they use the same dataset with similar settings.

Out of 1564 articles in total across nine sections, we randomly select 80 articles from each section and use them for training. Similar to [8, 18], the test set is chosen from VP section (consisting of 44 articles) and kept fixed throughout the experiment with NIPS 0-12 dataset. On average, the number of words per article are approximately 1000. We compute the perplexity on the test set and report the performance in terms of perplexity per document. Given the training data $X_{1,J}$ and a test set $X_j$ from $j$-th data source, perplexity per document (PPD) is defined as

$$PPD (X_j) = \exp \left(- \frac{1}{\bar{N}} \log p (X_j | X_{1,J}) \right) \quad (33)$$

where $\bar{N}$ denotes the number of documents in the test set.

Baseline Methods To compare the proposed model with other related works, we choose three baselines.

- **Baseline-1a ["No auxiliary"]:** This baseline is a linear Poisson-gamma based factor analysis model which totally relies on the target data and does not use any auxiliary data for training.