Supplementary Material for Learning the Dependency Structure of Latent Factors

Yunlong He
Georgia Institute of Technology
heyunlong@gatech.edu

Yanjun Qi
NEC Labs America
yanjun@nec-labs.com

Koray Kavukcuoglu
NEC Labs America
koray@nec-labs.com

Haesun Park
Georgia Institute of Technology
hpark@cc.gatech.edu

1 More on Breast Cancer Data

In Section 4.4, SLFA achieves state-of-the-art result on the classification of microarray data, without using extra biological information. It outperforms Lasso-overlapped-group Jacob et al. (2009), a logistic regression approach with the graph-guided regularization, which utilizes a known biological network. This result indicates that SLFA might automatically explore and discover the deep information not covered by previous linear models. Therefore, we perform biological function-enrichment analysis on the learned latent bases from the Breast-Cancer experiments to confirm this indication and show the potential of SLFA for relational analysis of latent factors in biological data set.

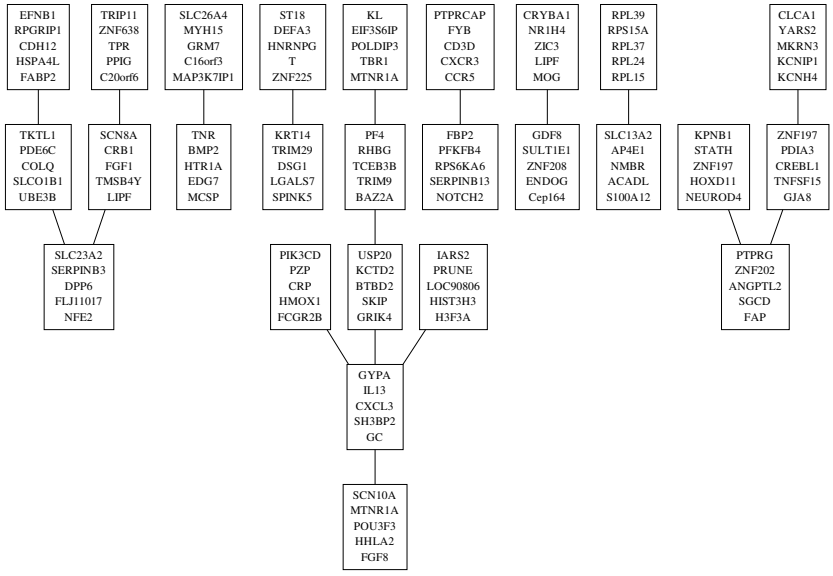


Figure 1: Positive interactions between gene groups. The top 5 genes associated with each latent factor are presented in the graph. Each edge corresponds to a negative element in the sparse precision matrix $\mathbf{\Phi}$.

In our analysis, we fix number of latent gene groups as 50 and plot and strongest positive and negative interactions in Figure 1 and Figure 2. The values associating each gene to a certain latent bases are sorted and we use the top 20 genes to represent and analyze each latent group and the top 5

genes are included in the graph. Using the functional annotation tool from Da Wei Huang & Lempicki (2008), we verify that the recovered latent groups appear reasonable based on the biological validations. The learned pairwise relationships between them also make good senses. For example, the 3^{rd} latent basis is enriched with proteins from "extracellular region", which is a region external to the outermost structure of a cell. For cells without external protective structures it refers to space outside of the plasma membrane Da Wei Huang & Lempicki (2008). Similarly, the 12^{th} latent basis is enriched with proteins from "plasma membrane", which is the membrane region surrounding a cell that separates the cell from its external environment. In the learned Φ matrix, the 12^{th} factor is partially correlated (positively) to the 3^{rd} latent factor, which is strongly verified by the closeness of their enriched cellular locations. Almost all top 20 genes in the 44^{th} latent bases are associated to the biological process "translational elongation". This important process carries the job of successive addition of amino acid residues to a nascent polypeptide chain during protein biosynthesis. This factor is strongly correlated (negatively) with the 3^{rd} latent factor based on the learned precision matrix. This also makes a lot of sense, since essentially the 44^{th} protein group locates at the intracellular which is not bounded by a lipid bilayer membrane and occurring within the cell.

Among the discovered latent factors, some are of great significance to breast cancer. For example, all top five proteins in the 2^{nd} latent group involve with the biological process "response to hormone stimulus", which performs a functional change in the activity of a cell as a result of a hormone stimulus. Other top ranked proteins in this group are mostly "glycoproteins" which are often important integral membrane proteins and play key roles in cell-cell interactions. This latent factor involves the critical gene "BRCA1" (BreastCance-1, early onset), whose defects are a cause of genetic susceptibility to breast cancer. Another three top genes "ITGA2", "PIAS3" and "RNASEL" in this group are have all been studied in cancer susceptibility.

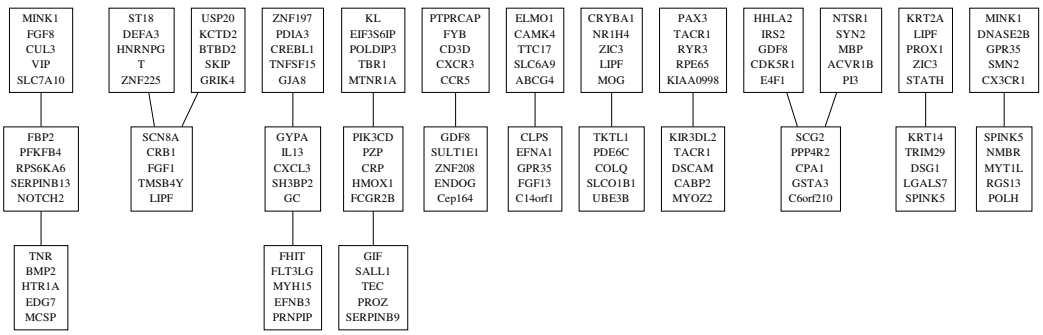
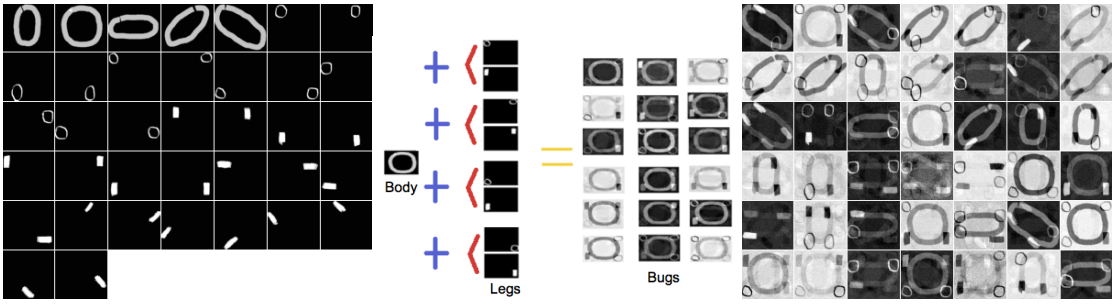
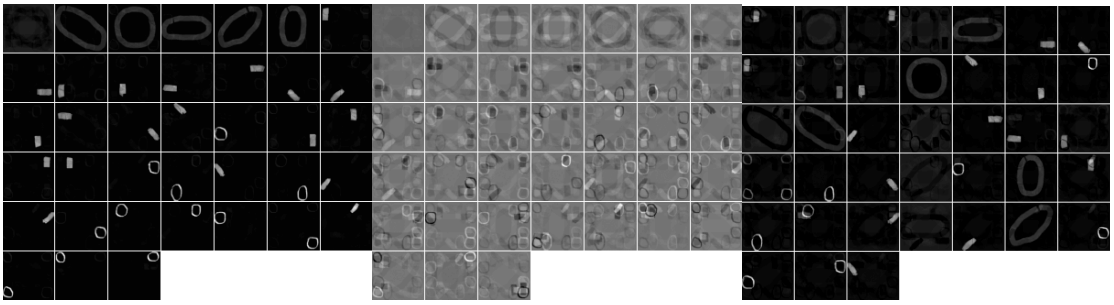


Figure 2: Negative interactions between gene groups.The top 5 genes associated with each latent factor are presented in the graph. Each edge corresponds to a positive element in the sparse precision matrix $\mathbf{\Phi}$.



(a) True Bases (b) Creation (c) Samples



(d) SLFA Basis (e) PCA Basis (f) NMF Basis

Figure 3: Basis Recovered by Different Methods

2 More on Toy Example

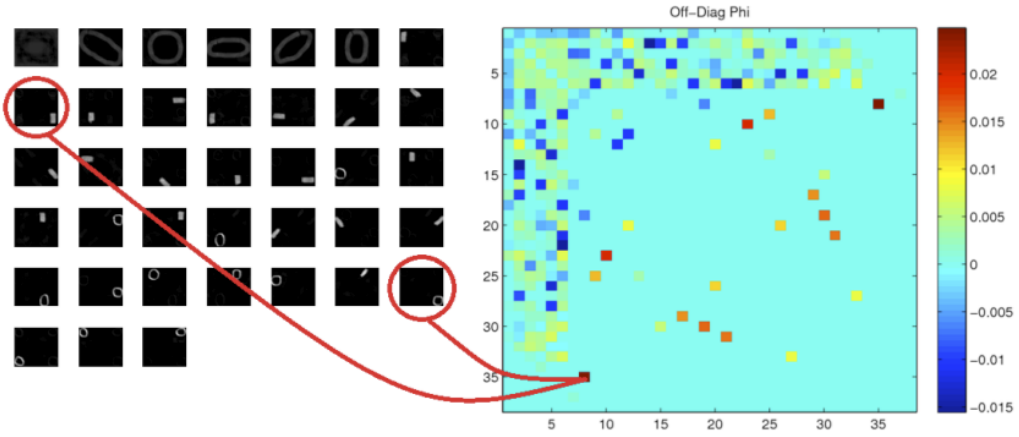


Figure 4: This figure shows how the sparse precision matrix (right heat graph) looks like, and how it indicates the relationship between latent bases. For instance, a positive value in $\mathbf{\Phi}$ indicates exclusion between a pair of square leg and circular leg.

3 More on Topic Visualization of NIPS Documents

SLFA explicit models the relationships between latent topics and therefore provides a convenient way to analyze and visualize document topics. We present a subgraph the positively related topics presented in Section 4.3. For completeness, we present the full graph of positively related topics in Figure 5. We can also use the precision matrix to find pairs of negatively related topics (see Figure 6 for a subgraph). It is worth noticing that the analysis methodology proposed in Section 3.1 can be used after Correlated Topic Model (CTM) Blei & Lafferty (2006) is trained, to generate a relational graph of topics (see Figure 7). We also refer readers to Jenatton et al. (2010) to see how our relational

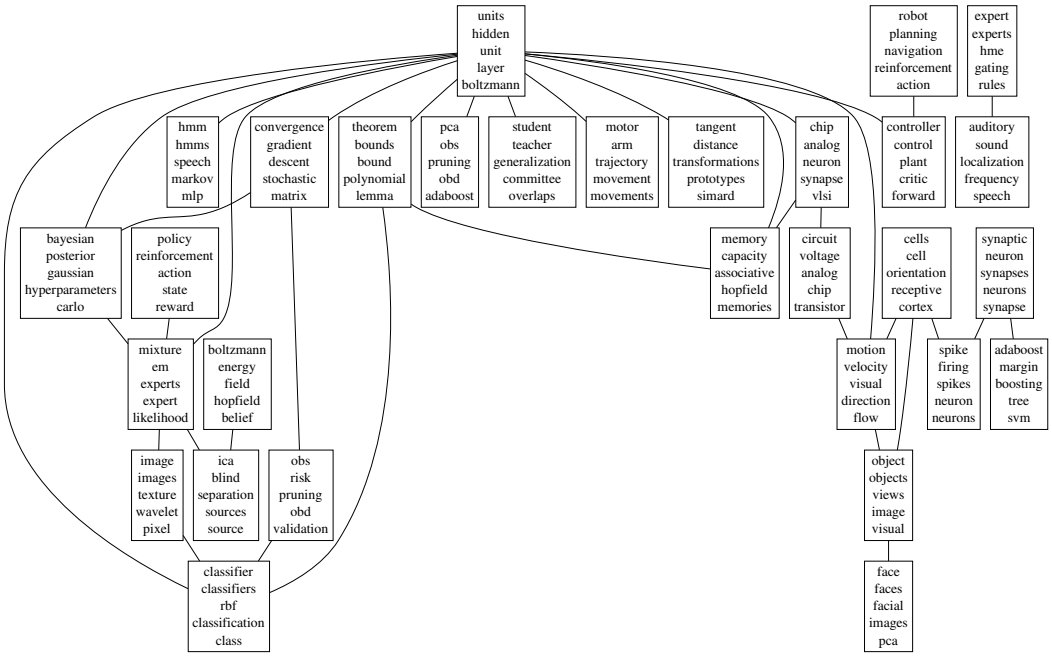


Figure 5: The full graph of positively related topics discovered from NIPS text corpus. Each edge corresponds to a negative element in the sparse precision matrix $\mathbf{\Phi}$.

graph of NIPS topics is different from the tree structured topical graph learned by sparse hierarchical dictionary learning.

References

Blei, D.M. and Lafferty, J.D. Correlated topic models. *Advances in Neural Information Processing Systems*, 2006.

Da Wei Huang, B.T.S. and Lempicki, R.A. Systematic and integrative analysis of large gene lists using david bioinformatics resources. *Nature protocols*, 2008.

Jacob, L., Obozinski, G., and Vert, J.P. Group lasso with overlap and graph lasso. *Proceedings of the 26th Annual International Conference on Machine Learning*, 2009.

Jenatton, R., Mairal, J., Obozinski, G., and Bach, F. Proximal methods for sparse hierarchical dictionary learning. *Proceedings of the International Conference on Machine Learning*, 2010.

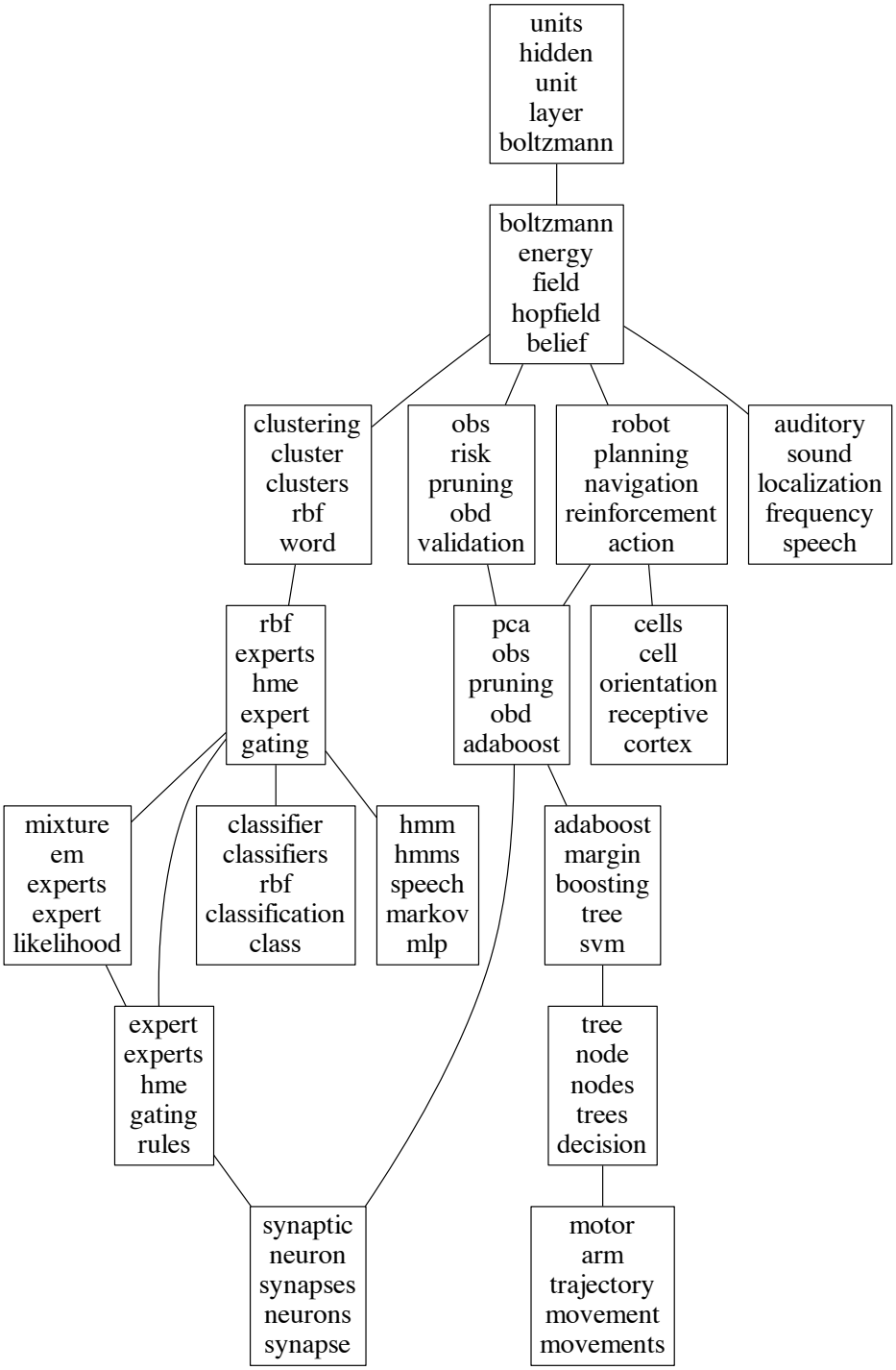


Figure 6: Negatively related topics discovered from NIPS text corpus. Each edge corresponds to a positive element in the sparse precision matrix $\mathbf{\Phi}$.

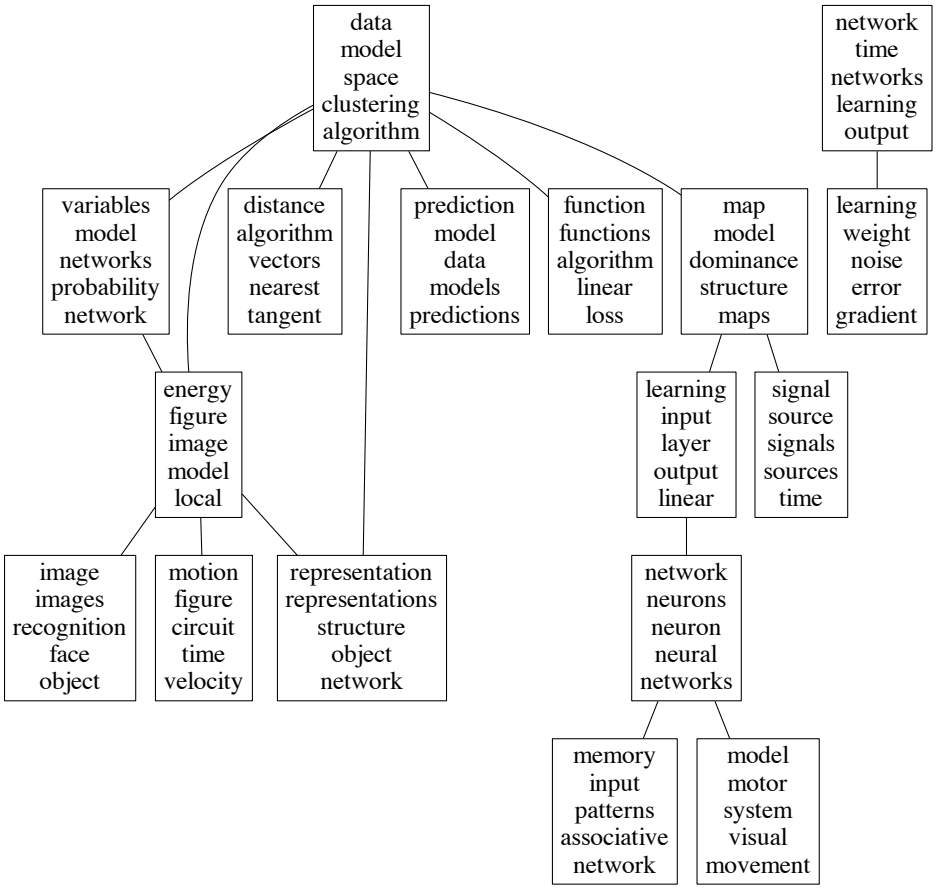


Figure 7: We use the Correlated Topic Model (CTM) to learn 40 topics from NIPS text corpus and then run sparse Gaussian graphical model to obtain the sparse precision matrix of the embedding vectors. Finally use the analysis methodology proposed in Section 3.1, we are able to generate a graph of positively related topics.