JOBVITE

Lossy Compression of Covariance Matrices

Shiqi Zhang

PURDUE UNIVERSITY

Introduction

Motivation:

- Jobvite's candidate matching product employs large multivariate normal distributions to predict candidates' fitness for job requisitions as part of a proprietary algorithm called Fuzzy Tags
- The large covariance matrices defining these distributions require 10s of gigabytes of data incurring large costs for storage and transmission of these models
- Lossy compression techniques could significantly reduce those costs if model degradation is minimized

Goal:

- Develop a lossy compression method for covariance matrix from arbitrary multivariate normal distribution
- The method should monitor the performance gains as well as the information loss
- The method should maintain the following mathematical properties of the original covariance matrix
  - (1) positive definite matrix; (2) original values from diagonal of the covariance matrix
- Algorithms for compression must be reasonably efficient

Methodology

Idea:

- Inducing sparsity in the covariance matrices by forcing independence between groups of random variables
- Clustering variables and forcing independence between clusters induce sparsity suitable for compression
- Figure 1 shows one example of original matrix K and compressed matrix $\hat{K}$. By forcing independence between variables x_1, x_2, x_3 and x_4, only the off-diagonal elements a, b, c need to be stored, achieving a compression ratio of 50%

Demonstration Data:

- Current data set: a covariance matrix constructed from Stack Overflow questions and their tags
- Matrix size: 256 MB

$$\mathbf{K:}\ \begin{bmatrix} 1 & a & b & d \\ a & 1 & c & e \\ b & c & 1 & f \\ d & e & f & 1 \end{bmatrix} \longrightarrow \hat{\mathbf{K}}\mathbf{:}\ \begin{bmatrix} 1 & a & b & 0 \\ a & 1 & c & 0 \\ b & c & 1 & 0 \\ 0 & 0 & 0 & 1 \end{bmatrix}$$

Figure 1: original matrix and compressed matrix

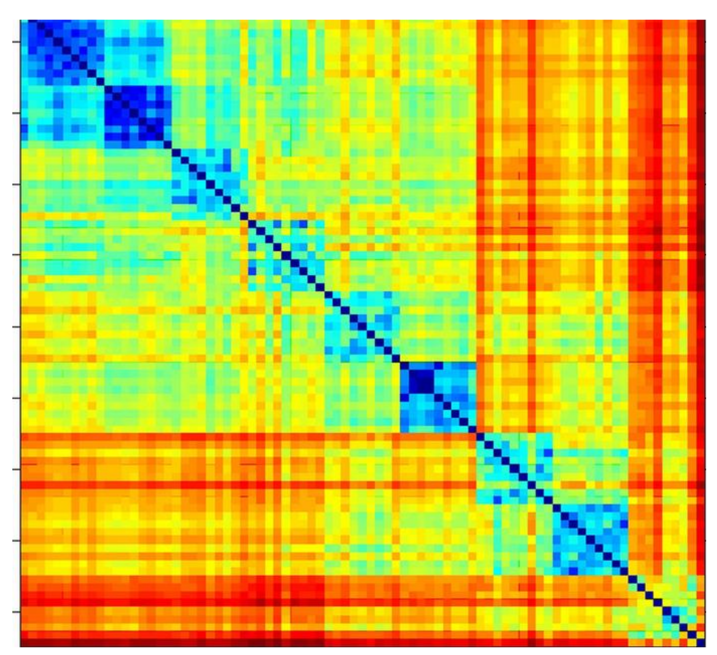

Figure 2: matrix structure after spectral clustering

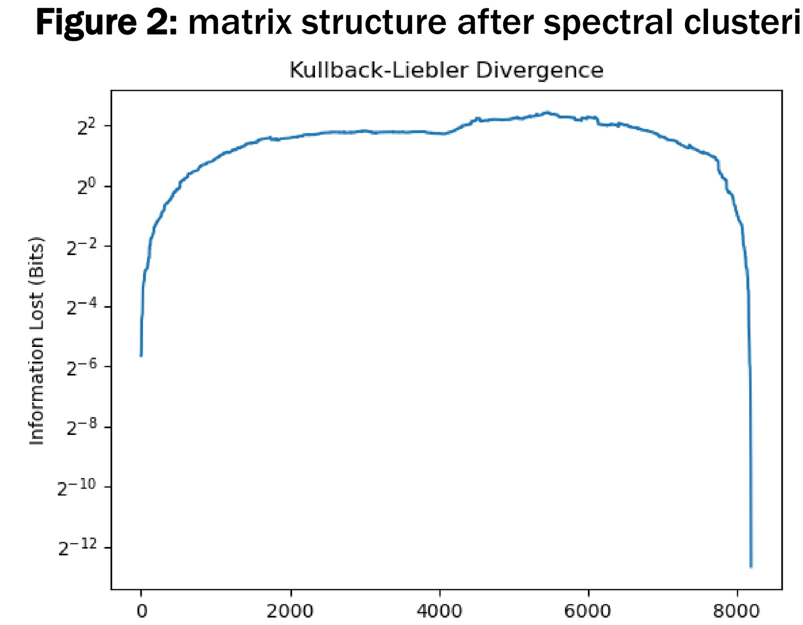

Figure 3: Kullback-Liebler divergence

Spectral Clustering

What it is:

- A clustering method commonly employed on graph data structures
- It uses the eigenvalues and eigenvectors of the graph Laplacian constructed from the covariance matrix

How to use:

- Calculate the Laplacian matrix from the original covariance matrix and compute its eigenvalues
- The second smallest eigenvalue is called the Fiedler value and the corresponding eigenvector is the Fiedler vector. Fiedler vector is used to sort the original covariance matrix

Our enhancements:

- The sorted covariance matrix is partitioned into two components by our objective function: Kullback-Leibler divergence per discarded element
- Hierarchical clustering is performed and the above process is applied recursively for the subclusters

Example:

- Figure 2 presents one example of the matrix structure after performing spectral clustering

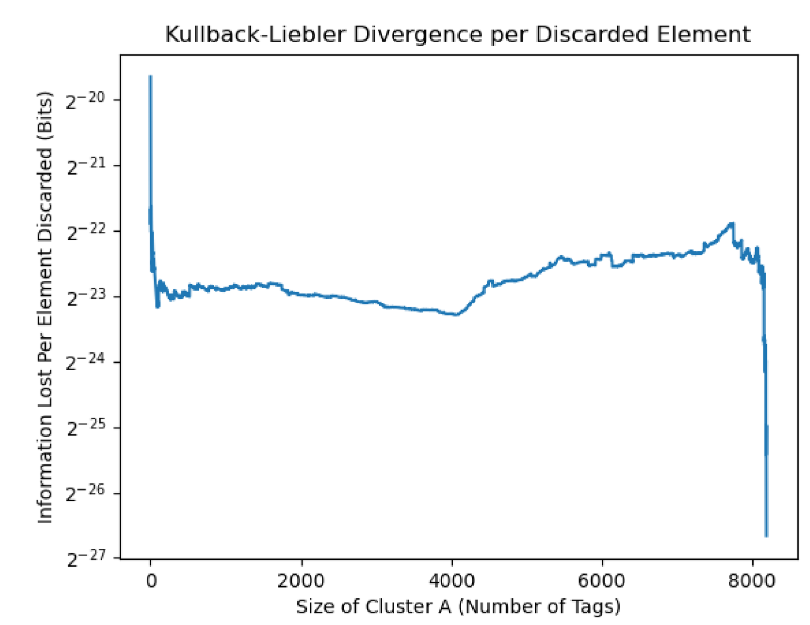

Figure 4: Kullback-Liebler divergence per discarded element

Acknowledgements

I would like to thank both the Data Mine and Jobvite for providing this opportunity to work on this project. Specifically, I would like to thank my mentor Jonathan Roose for patient instructions and continuous support. I would also like to thank to Margaret Ann Betz, James Joko and Dr. Mark Daniel Ward for helpful feedbacks.

Kullback-Leibler divergence

What it is:

- It is a measure of information loss between probability distributions

How to use:

- The original covariance matrix Σ_0 and the compressed one Σ_1 represent multivariate normal distributions with zero mean. Their rank is denoted by k. This formula describes their KL divergence:

$$D_{KL} = 0.5 \times (tr(\Sigma_1^{-1}\Sigma_0) - k + \ln(\det \Sigma_1) - \ln(\det \Sigma_0))$$

Efficient Computation:

- $tr(\Sigma_1^{-1}\Sigma_0) - k = 0$ for our compression scheme
- $\ln(\det \Sigma_0)$ is constant for all partitions
- Cholesky decomposition quickly computes $\ln(\det \Sigma_1)$ for partitions

Results:

- Figure 3 depicts the Kullback-Leibler divergence between the original covariance matrix and the compressed covariance matrix. It represents the information loss for every possible partitioning
- Figure 4 depicts the Kullback-Leibler divergence per discarded element. This is our objective function for partitioning, as it balances information loss with compression achieved

Conclusion

- Our work proves compression of covariance matrices via clustering retains key mathematical properties
- Kullback-Leibler divergence quantifies information loss due to compression within the multivariate normal distribution
- Our method quickly computes KL divergence of proposed clusters through algebraic simplifications and Cholesky decomposition

References

- Zare, Habil; P. Shooshtari; A. Gupta; R. Brinkman (2010). "Data reduction for spectral clustering to analyze high throughput flow cytometry data". BMC Bioinformatics. 11: 403. doi:10.1186/1471-2105-11-403. PMC 2923634. PMID 20667133
- Kullback, S. (1959), Information Theory and Statistics, John Wiley & Sons. Republished by Dover Publications in 1968; reprinted in 1978: ISBN 0-8446-5625-9