Using Biclustering Methods to Classify High-dimensional Data

Biclustering via Sparse SVD

Introduction

- The objective of this project was to understand and implement the SSVD algorithm introduced by paper "Biclustering via Sparse Singular Value Decomposition" by Mihee Lee, Haipeng Shen, Jianhua Z. Huang and J. S. Marron.
- Motivation: Biclustering methods play important roles in analyzing high-dimensional low sample size datasets.
- Sparse singular value decomposition was proposed based on SVD. Sparsity was obtained by adding sparsity-inducing penalties

Method

- When using SVD to seek low rank approximation to a matrix X, the best rank-one approximation $(s_1, \boldsymbol{u}_1, \boldsymbol{v}_1) = \text{argmin}_{s,\boldsymbol{u},\boldsymbol{v}} ||\boldsymbol{X} - s\boldsymbol{u}\boldsymbol{v}^T||_F^2$
- In order to achieve sparsity in u and v, sparsity-inducing penalty terms were added to above function. The best rank-one approximation matrix using SSVD was found by minimizing $||\boldsymbol{X} - s\boldsymbol{u}\boldsymbol{v}^T||_F^2 + \lambda_u P_1(s\boldsymbol{u}) + \lambda_v P_2(s\boldsymbol{v})$
- The paper used adaptive lasso penalty: $P_1(s\boldsymbol{u}) = s\sum_{i=1}^{n} \omega_{1,i}|u_i|$, $P_2(s\boldsymbol{v}) = s\sum_{i=1}^{d} \omega_{2,j}|v_j|$ $\omega_{1,i}$s and $\omega_{2,j}$s are data-driven weights.
- $\boldsymbol{u}$ and $\boldsymbol{v}$ were solved using iterative algorithm.

Application to lung cancer data

- The data consist of expression levels of 12,625 genes, measured from 56 subjects. (56×12625 matrix)
- Algorithm converges within 5-10 iterations.
- The number of genes selected in each layer is much less than 12,625: there are, respectively, 3205, 2511, and 1221 genes involved in the three layers, corresponding to the nonzero entries of the $\boldsymbol{v}_k$ vectors.

Biclustering via Sparse SVD

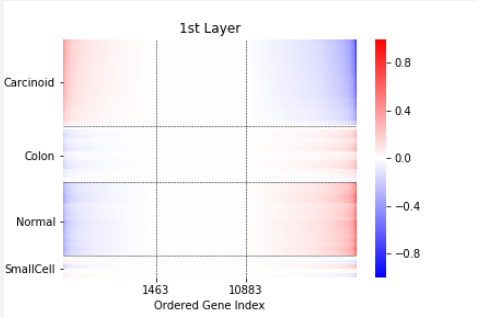

Figure: Lung cancer data: plot of first SSVD layer. Genes are rearranged according to an increasing order of the entries of the entries of $\boldsymbol{v}_1$, subjects are rearranged according to the values of $\boldsymbol{u}_k$ within each subject group.

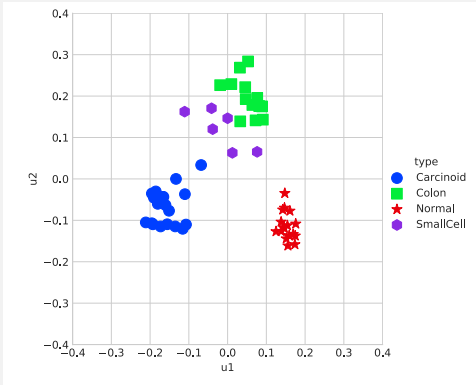

Figure: Lung cancer data: scatterplots of the entries of $\boldsymbol{u}_1$, $\boldsymbol{u}_2$,first two sparse singular vectors. These two vectors reveal three subject clusters.

Conclusion

- SSVD can automatically perform gene selection.
- Selected genes correspond to informative grouping of the subjects.
- The selection is performed both on the rows and columns, SSVD procedure take into account potential row-column interactions and thus suitable for biclustering.

Fitting spatial models to gun violence data

Introduction

- The objective of this project was to explore spatial patterns of gun incidents happened in the United States, 2017 and fit different spatial models to the data.
- The data was obtained from gunviolencearchive.org. It contained 59511 records, including the number of injuries, number of death and location information of each gun incident.
- A general view of where each gun related violence incident took place and the number of victims involved in 2017.

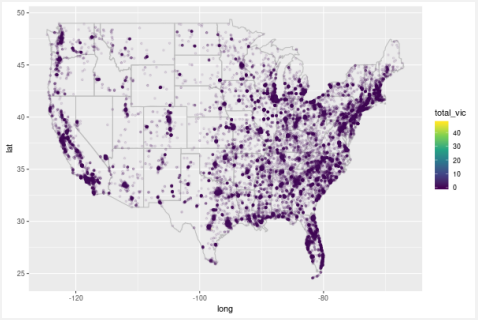

Figure: 2017 Gun shot violence distribution. A majority of the incidents happened in the Eastern United States as well as the west coast.

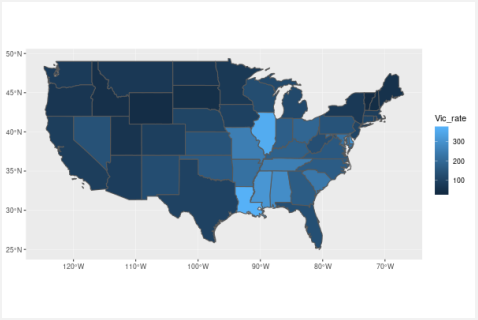

Figure: Victims per Million People 2017: spatial maps of the number of gun related violence victims, of which the second one is scaled by state populations.

Fitting spatial models to gun violence data

Method

- Simultaneous Autoregressive(SAR) Model: $y(s) = \phi \sum_{s'} \frac{W_{s,s'}}{W_{s,\cdot}} y(s') + \epsilon$
- Conditional Autoregressive(CAR) Model: $y(s)|y(-s) \sim N\left(\sum_{s'} \frac{W_{s,s'}}{W_{s,\cdot}} y(s'),\ \sigma^2\right)$
- Spatial GLM and Intrinsic Autoregressive Model.

Results

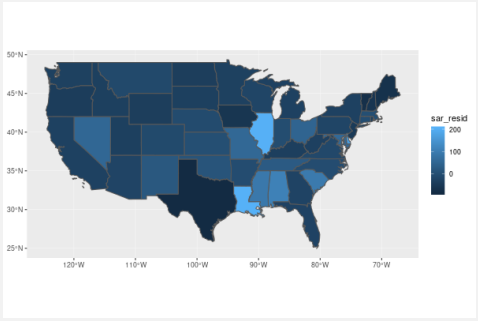

Figure: Residual plot of SAR model. There still exists spatial autocorrelation between states.

Model	RMSE	Moran'I	Geary'C
CAR	62.7792	-0.0979	0.9419
SAR	64.7380	0.0147	0.8272
GLM	683.062	0.2095	0.7314
IAR	3.1488	-0.1537	1.2163

Table: Measures of goodness of the four models we fitted including RMSE, Moran'I of residual and Geary'C of residual

- All the models perform well with relatively small RMSE values except the GLM model.
- Among all the models, IAR model has the lowest RMSE and its Moran'I of residual is the closest to zero. Therefore, IAR model performs the best in our analysis.

Conclusion

- To improve the results of our model, we could search for different datasets taken from other years and use them to fit the spatial models to see if the findings remain valid.