Discovering Sparse Covariance Structures with the Isomap

Bo Chang

Graphical Models Reading Group

July 3, 2015

Overview

- We talked about covariance estimation with Cholesky decomposition and GLM.
- A large class of methods with assumption: variables have a natural ordering, such as longitudinal data, time series, spatial data.
- We will first talk about some other methods in that class, then how to discover a structured ordering with Isomap, as proposed by Wagaman, A. S., & Levina, E. (2009).

Banding

- Given a $p \times p$ sample covariance matrix $\hat{\mathbf{\Sigma}} = [\hat{\sigma}_{ij}]$ and an integer k, $0 \leq k \leq p$, the k-banded estimator is defined by

$$B_k(\hat{\mathbf{\Sigma}}) = [\hat{\sigma}_{ij}\mathbf{1}\{|i-j| \leq k\}].$$

- Asymptotic results: the banded estimator and its inverse are consistent as long as

$$\frac{\log p}{n} \to 0.$$

- Not necessarily positive definite.

Tapering

- Given a tapering matrix $\mathbf{W} = [w_{ij}]$, a tapered estimator is defined by

$$B_{\mathbf{W}}(\hat{\boldsymbol{\Sigma}}) = \hat{\boldsymbol{\Sigma}} \circ \mathbf{W} = [\hat{\sigma}_{ij} w_{ij}].$$

- Schur product theorem: if $\mathbf{W}$ is positive definite, $B_{\mathbf{W}}(\hat{\boldsymbol{\Sigma}})$ is positive definite.
- Banding corresponds to

$$\mathbf{W} = [\mathbf{1}\{|i-j| \leq k\}]$$

Tapering

- For example, the trapezoidal tapering matrix

$$w_{ij} = \begin{cases} 1 & \text{if } |i-j| \leq l_h, \\ 2 - |i-j|/l_h & \text{if } l_h < |i-j| < l, \\ 0 & \text{otherwise,} \end{cases}$$

  for a given tapering parameter l and $l_h = l/2$.
- Asymptotic results: the tapered estimator and its inverse are consistent as long as

$$\frac{\log p}{n} \to 0.$$

Thresholding

- The thresholded estimator for a $\lambda \geq 0$ is defined by

$$T_\lambda(\hat{\boldsymbol{\Sigma}}) = [\hat{\sigma}_{ij}\mathbf{1}\{\hat{\sigma}_{ij} \geq \lambda\}].$$

- Different from banding and tapering, thresholding is permutation-invariant.
- Asymptotic results: the thresholded estimator and its inverse are consistent as long as

$$\frac{\log p}{n} \to 0.$$

- Wagaman, A. S., & Levina, E. (2009) claim that banding has a better convergence rate than thresholding. While Pourahmadi, M. (2013) does not completely agree.

Multidimensional Scaling (MDS)

- Mapping data in $\mathbb{R}^p$ to a lower-dimensional $\mathbb{R}^k$, preserving the pairwise dissimilarity as well as possible.
- Observations $x_1, x_2, \ldots, x_N \in \mathbb{R}^p$, and let d_{ij} be the dissimilarities between observations i and j. Often we choose Euclidean distance $d_{ij} = \|x_i - x_j\|$.
- **Metric multidimensional scaling**: find values $z_1, z_2, \ldots, z_N \in \mathbb{R}^k$ to minimize

$$S_M(z_1, z_2, \ldots, z_N) = \sum_{i \neq j} (d_{ij} - \|z_i - z_j\|)^2.$$

- A gradient descent algorithm is used to minimized S_M.

Multidimensional Scaling (MDS)

- **Classical multidimensional scaling**: Start with similarities $s_{ij} = \langle x_i - \bar{x}, x_j - \bar{x} \rangle$. Minimize

$$S_C(z_1, z_2, \ldots, z_N) = \sum_{i,j} (s_{ij} - \langle z_i - \bar{z}, z_j - \bar{z} \rangle)^2.$$

- Classical multidimensional scaling is equivalent to principal components analysis.

Multidimensional Scaling (MDS)

- Metric MDS and Classical MDS approximate the actual dissimilarity or similarities. **Nonmetric multidimensional scaling** effectively uses only ranks.
- Minimize

$$S_{NM}(z_1, z_2, \ldots, z_N, \theta) = \frac{\sum_{i \neq j}(\theta(d_{ij}) - \|z_i - z_j\|)^2}{\sum_{i \neq j} \|z_i - z_j\|^2},$$

  where θ is an increasing function.
- With θ fixed, minimize over z_i by gradient descent. With z_i fixed, use isotonic regression to find best monotonic approximation θ.

Isomap

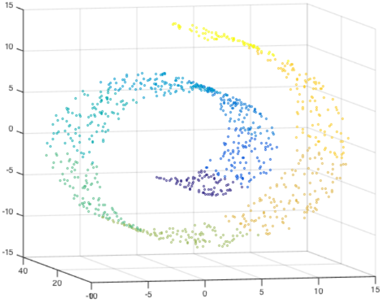
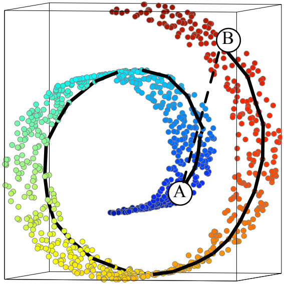


Figure: Swiss roll dataset.

- k-**nearest neighbors** (k-NN) adjacency graph: a sparse graph representing local structure.
- Distance or similarity can be defined based on the graph.

Isomap

- Euclidean distance between two neighboring nodes as weight of edge.
- Shortest paths: Djikstra's algorithm.
- Geodesic distance as dissimilarity, apply MDS.



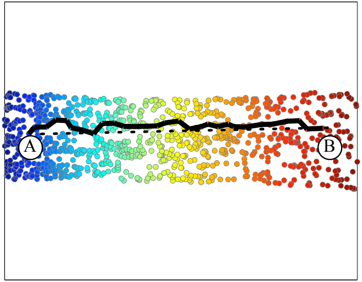


Figure: Euclidean distance vs geodesic distance.

The Isomap algorithm

1. For each point, find its r nearest neighbors using the dissimilarities $d(i,j)$. Construct a neighborhood graph by connecting each point to its r neighbors, with dissimilarities as the edge weights.

2. Estimate the geodesic distance $\tilde{d}_r(i,j)$ between each pair of points i, j by computing the shortest-path distance from i to j through the neighborhood graph.

3. Apply metric MDS to the matrix of pairwise shortest-path distances to obtain an embedding in $\mathbb{R}^d$. In our case, this means find $z_1, \ldots, z_p \in \mathbb{R}^1$ that minimize the stress function (known as stress 1 in the literature)

$$S(z_1, \ldots, z_p) = \frac{\sum_i \sum_j (|z_i - z_j| - \tilde{d}_r(i,j))^2}{\sum_i \sum_j |z_i - z_j|^2} .$$

This minimization reduces to computing the first eigenvector of a matrix derived from pairwise distances.

Isomap

There are many ways to define a dissimilarity measure.

- $d_{ij} = 1 - |\hat{\rho}_{ij}|$, where $\hat{\rho}_{ij}$ is the sample correlation coefficient between variable i and j.
- $d_{ij} = 1 - \hat{\rho}_{ij}$.
- $d_{ij} = C - |\hat{\sigma}_{ij}|$, where $\hat{\sigma}_{ij}$ is the sample covariance between variable i and j.

Isoband (Isomap+banding)

- From the ordering, we construct a $p \times p$ permutation matrix $\mathbf{P}$. The covariance matrix is reordered

$$\hat{\mathbf{\Sigma}}_0 = \mathbf{P}\hat{\mathbf{\Sigma}}\mathbf{P}^T.$$

- The banding operator B_k is applied to $\hat{\mathbf{\Sigma}}_0$, and then reorder back.

$$\hat{\mathbf{\Sigma}}_1 = \mathbf{P}^T B_k(\hat{\mathbf{\Sigma}}_0)\mathbf{P}.$$

Bootstrap

Bootstrapped isoband algorithm: modify Step 1 as follows:

1a Resample the observations with replacement T times and construct T bootstrap sample covariance matrices $\hat{\Sigma}_1^*, \ldots, \hat{\Sigma}_T^*$.

1b For each matrix $\hat{\Sigma}_t^*$, $t = 1, \ldots, T$, construct a neighborhood graph by connecting each variable to its r nearest neighbors using dissimilarities $d_t^*(i, j)$ based on $\hat{\Sigma}_t^*$.

1c Construct the final neighborhood graph by putting an edge between variables i and j if an edge is present between i and j in at least cT of the bootstrap graphs, where $c \in (0, 1)$ is a tuning parameter, and assign weight $d(i, j)$ (original dissimilarity) to the edge.

References

Wagaman, A. S., & Levina, E. (2009). Discovering sparse covariance structures with the isomap. Journal of Computational and Graphical Statistics, 18(3), 551-572.

Pourahmadi, M. (2013). High-Dimensional Covariance Estimation: With High-Dimensional Data. John Wiley & Sons.

Saul, L. K., Weinberger, K. Q., Ham, J. H., Sha, F., & Lee, D. D. (2006). Spectral methods for dimensionality reduction. Semisupervised learning, 293-308.

Trevor J.. Hastie, Tibshirani, R. J., & Friedman, J. H. (2009). The elements of statistical learning: data mining, inference, and prediction. Springer.

The End