Graphical Models: Introducing BetaMixture Method for Correlation Detection

Lucy Liu

University of Connecticut

October 24, 2024

Overview

- Objective: Introduce the betaMixture method and its applications in high-dimensional data analysis.
- Topics covered:
  - High-dimensional challenges in regression and graphical models.
  - Convex geometry and correlation detection.
  - The betaMixture method.
  - Applications: Riboflavin gene expression analysis.

Problem Introduction: Linear regression

- Linear function:

$$y = \beta_0 + \sum_{j=1}^{p} \beta_j x_j + \epsilon$$

- where y is the outcome (response) variable is , p is the number of predictors, x_j , sample size is n, and random (Gaussian) noise is $\epsilon \sim N(0, \sigma^2)$

Problems appear when P is greater than n

- Using matrix notation, the parameter vector is estimated by the ordinary least squares formula $\hat{\beta} = (X'X)^{-1}X'Y$
- If $P > n$, routine estimation of regression parameters is not possible since the inverse of matrix $X'X$ does not exist. β is then unidentifiable.
- Even if $n > P$, inference about β may be impractical when P is sufficiently large because standard errors are often large and the width of the confidence interval grows with P.
- Standard errors inflate with more predictors:

$$\text{Width of Confidence Interval} \propto \sqrt{\frac{P(n-1)F_{P,n-P,\alpha}}{n(n-P)}}$$

The problem with LASSO

- Traditional linear regression methods face challenges when the number of predictors P exceeds the sample size n, leading to issues like unidentifiable parameters.
- In high-dimensional settings, typical methods like LASSO make strong assumptions about sparsity of the mean vector and rows of the covariance matrix, but correlation between columns of X is inevitable when P is large.
- Approaches to deal with correlations have relied on dimension reduction to restore validity of the requirement that $X'X$ is invertible.

Troublesome Assumptions

- Relationship between Y and X may not be linear.
  - Ex. a quantitative trait may depend on the expression of many genes so that a change in the expression of one gene may not occur without simultaneous change in many other genes.
- β sparsity assumption and covariance matrix sparsity may not be valid. It is possible that a trait is associated with hundreds or even thousands of genes.
  - Ex. If genes form a highly connected network, which may be necessary because the trait requires the production of many different proteins or it may be evolutionary beneficial as a way to protect against mutations
- Assumption of underlying low dimensionality may not be not valid

Introduction to Graphical Models

- **Graphical Models**: Graphical models are probabilistic models that represent dependencies between random variables as a graph.
- **Nodes** represent variables $X_1, X_2, ..., X_P$.
- **Edges** between nodes represent conditional dependencies between variables.
- **Goal**: Detect significant correlations between features in high-dimensional data.

Convex Geometry

- High-dimensional spaces behave counter-intuitively.
- Convex geometry is used to show that random pairs of uncorrelated vectors are almost orthogonal with high probability in high dimensions.
- The angle θ between two random vectors is governed by:

$$\sin^2\theta \sim \text{Beta}\left(\frac{n-1}{2}, \frac{1}{2}\right)$$

- This can be used to detect correlations by testing deviation from the null.

BetaMixture Method: Overview

- The betaMixture method models pairwise correlations using a mixture of beta distributions.
- Empirical Bayes two-group approach:
  - Null: $f_0(z) = \frac{z^{(n-1)/2-1}(1-z)^{-1/2}}{B\left(\frac{n-1}{2}, \frac{1}{2}\right)}$
  - Alternative: $f_{a,b}(z) = \frac{z^{a-1}(1-z)^{b-1}}{B(a,b)}$
- By treating the data as P points in an n-dimensional space, it leverages high-dimensional geometry to detect relationships between variables.

Mixture of Beta Distributions

- Frequentist Approach:
  - Null hypothesis: Predictors are uncorrelated.
  - Detect edges based on the distribution of angles.

$$\sin^2(\theta_{\text{threshold}}) \sim Q_\delta$$

  where Q_δ is a quantile of the Beta distribution.
- Bayesian Approach (betaMix):
  - Empirical Bayes method for improved power.
  - Mixture of beta distributions to identify correlated predictors.
  - an edge in the graph exists if the posterior null probability (under f_0) is smaller than some threshold,

$$m_0^{(t)} < \tau$$

  where τ controls the false discovery rate (FDR).

Mathematical Model of BetaMixture

- Using the properties of the beta distribution for angles between random vectors, the method identifies non-null correlations, helping to build graphical models without assuming sparse networks.
- The BetaMixture model:

$$p(z_j) = p_0 f_0(z_j) + (1 - p_0) f_{a,b}(z_j)$$

- Parameters a, b, and p_0 are estimated using the EM algorithm.

Estimating the BetaMixture Model

- E-step: Estimate posterior probabilities of null and alternative components:

$$\hat{m}_0^{(t)} = \frac{p_0^{(t-1)} f_0(z_j)}{p_0^{(t-1)} f_0(z_j) + (1 - p_0^{(t-1)}) f_{a^{(t-1)}, b^{(t-1)}}(z_j)}$$

- M-step: Maximize likelihood to update a, b, p_0.

Estimating the BetaMixture Model

E-step (Expectation Step):

- In this step, we compute the expected value of the latent variable m_0, which represents whether a pair of predictors follows the null distribution (uncorrelated).
- The posterior probability that the j-th pair is from the null distribution is:

$$\hat{m}_0^{(t)} = \frac{p_0^{(t-1)} f_0(z_j)}{p_0^{(t-1)} f_0(z_j) + (1 - p_0^{(t-1)}) f_{a^{(t-1)}, b^{(t-1)}}(z_j)}$$

where:

- $f_0(z_j)$ is the null beta distribution, Beta $\left(\frac{n-1}{2}, \frac{1}{2}\right)$.
- $f_{a,b}(z_j)$ is the alternative beta distribution, Beta(a, b).
- $p_0^{(t-1)}$ is the prior probability of the null hypothesis at iteration $t-1$.

Estimating the BetaMixture Model

M-step (Maximization Step):

- In this step, we maximize the expected complete data log-likelihood from the E-step by updating the parameters.
- Update the parameters a, b by solving:

$$\text{argmax} \sum_j \left[\hat{m}_0 j^{(t)} \log f_{a,b}(z_j) \right]$$

- This step is repeated iteratively until convergence.

Error Rate Control

- **False Discovery Rate (FDR)**: The betaMixture method controls FDR by setting thresholds for significance testing.
- For m pairwise comparisons, FDR control ensures that the proportion of false positives among significant results is bounded by a target level q.
- The betaMixture method controls the false discovery rate by setting thresholds based on beta distributions. $m_0^{(t)} < \tau$
- Since the null distribution is determined by the sample size, we can set τ so that $Q_\tau((n-1)/2, 0.5) = q$.
- High-dimensional spaces allow precise detection of correlations with minimal false positives.

Application: Riboflavin dataset

- Dataset:
  - Contains normalized expression data of 4088 genes and 71 samples (assumed to be independent).
  - Goal: identifying which genes are predictors of riboflavin production rate in *Bacillus subtilis*.
- High-dimensional data where $P > n$

Riboflavin Analysis Using BetaMixture

- Treat each gene as a point in $\mathbb{R}^{71}$.
- Calculate angles between pairs of gene expression vectors.
- Variable selection: detect significant correlations (edges) between the $P = 4088 + 1$ variables (genes and riboflavin production)
- Reporting the nodes which are found to be adjacent to the response variable's node.
- Threshold: $\sin^2\theta > 0.815$ ($|r| > 0.43$).

Visualizing the Riboflavin Network

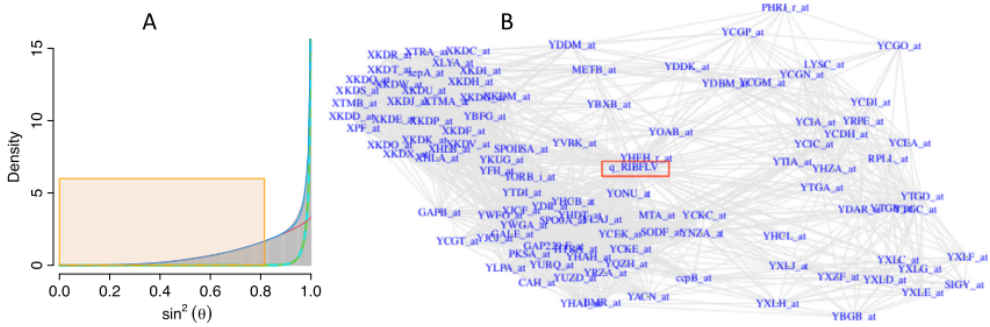

Fig. 5. A. The riboflavin data - fitted beta mixture model. B. 106 genes are selected as strong predictors for the production rate of riboflavin data.

- Fig. 5A shows the distribution of the z_j 's and the fitted mixture model.
- For variable selection, we're only interested in edges which connect to the riboflavin production rate variable (the highlighted node, q_RIBFLV in Fig. 5B)

Riboflavin Results

- 106 genes were identified as significant predictors of riboflavin production.
- Formed two large interconnected clusters.
- The large number of selected predictors and the strong dependence among them suggests that riboflavin production is an intricate process which cannot be explained well by a sparse, linear model.
- A change in one gene may cause a chain reaction in many other genes, possibly involving non-linear effects, making it complicated to predict the ultimate effect on the response variable.

Conclusion

- Summary:
  - BetaMixture method provides a powerful alternative to traditional sparse models.
  - Based on solid convex geometry principles for high-dimensional data.
  - Effective in complex, non-sparse networks like gene expression.
- Other Applications:
  - Spatial data (e.g., estimating spatial covariance matrices).
  - Classification problems (e.g., radar signal classification).
  - Complex high-dimensional datasets in various domains.