Sequential Parallel LASSO Models for eQTL Analysis

Anhong He
Syracuse University
900 South Crouse Ave
Syracuse, NY 13244, U.S.
anhonghe@gmail.com

Benika Hall
University of North Carolina at Charlotte
9201 University City Blvd.
Charlotte, NC 28223
bjohn157@uncc.edu

Jia Wen
University of North Carolina at Charlotte
9201 University City Blvd.
Charlotte, NC 28223
jwen6@uncc.edu

Yingbin Liang
Syracuse University
900 South Crouse Ave
Syracuse, NY 13244, U.S.
yliang06@syr.edu

Xinghua Shi
University of North Carolina at Charlotte
9201 University City Blvd.
Charlotte, NC 28223
x.shi@uncc.edu

1. ABSTRACT

The availability of large-scale genomic and transcriptomic data on populations makes it necessary to perform computationally intensive expression quantitative trait locus (eQTL) analysis. Modeling in a sparse learning framework, LASSO based tools are powerful for eQTL analysis. However, classical LASSO becomes limited for big genomic data. We thus propose two novel methods, namely sequential LASSO and parallel LASSO, to conduct eQTL analysis for datasets of ultra-high dimension. We theoretically prove the consistency of our methods under mild conditions and perform extensive simulations on synthetic data to validate our methods. We also apply our methods to a real human genomics database demonstrate the application of our method.

Categories and Subject Descriptors

J.3 [**Life and Medical Sciences**]: Biology and Genetics

General Terms

Algorithms

Keywords

LASSO models, eQTL analysis

2. MOTIVATION

Modern genetic variation databases for eQTL analysis are always of extremely high dimensionality. Given the availability of these massive data collections, dimensionality reduction is critical to improve the algorithm performance over these datasets. Least absolute shrinkage and selection operator (LASSO)[1] is one of many outstanding methods in dimensionality reduction for eQTL analysis. However, LASSO involves in the whole design matrix for computation and is thus limited by computing capacity (e.g. computer memory, time consumption) for big genomic data. In our study, we introduce two novel methods named sequential LASSO and parallel LASSO to allow efficient learning for datasets of ultra-high dimension. We theoretically prove the consistency under mild conditions and perform extensive simulations on synthetic data to validate our methods. We then apply our methods to a real human genomics database demonstrate the application of our method.

BCB'15, September 9–12, 2015, Atlanta, GA, USA.
ACM 978-1-4503-3853-0/15/09.
http://dx.doi.org/10.1145/2808719.2811449.

3. METHODS

In population based eQTL analysis, the genotypes of genetic variants are taken as the explanatory variant X and gene expression levels of the same individuals as the response variant Y. We employ a multi-variate linear regression model with sparsity penalty to capture the associations between genotypes and gene expression. In terms of a Gaussian graphical model[2], we assume $\{X, Y\}$ are a sparse graph and our objective is to retrieve all nodes X_i which directly links to Y in the graph.

$$\beta^* = \arg\min_{\beta} \frac{1}{n}\|y - X\beta\|_2^2 + \lambda\|\beta\|_1 \quad (1)$$

Equation-(1) shows the classical LASSO, where the whole design matrix $X_{n\times p}$ encoded by genotypes is involved. The intuition for Sequential/Parallel LASSO is based on the idea of "divide and conquer". Specifically, we break the problem into subproblems by splitting the whole feature set into smaller subsets, and applying classical LASSO on each subset. Then through interactions among results of subproblems, we expect to generate a consistent result to the original problem.

3.1 Sequential LASSO

For both Sequential and Parallel LASSO, we partition the whole feature set W into K non-intersected subsets, denote as $W = \{G_1, G_2, \cdots G_K\}$. Sequential LASSO performs feature selection multiply times with partial information $\{X_{G_k}, Y\}$ in each step, as shown in the structure chart in Figure-(1).

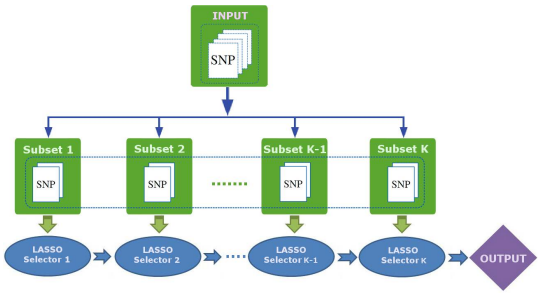

Figure 1: Structure Chart of Sequential LASSO

In k^{th} step LASSO works on a marginalized Gaussian sub-graph[2] $\{X_{S_{k-1}\cup G_k}, Y\}$. which includes edges inherent from the original graph and new edges induced by marginalization. For theoretical consistency analysis we first define three sets X_{T_k}, X_{I_k} and X_{O_k} for k^{th} step where $k = 1, \cdots, K$. X_{T_k} denotes nodes with direct links to Y in subgraph $\{X_{S_{k-1}\cup G_k}, Y\}$ inherited from the original graph $\{X, Y\}$. X_{I_k} denotes nodes with direct links to Y, but the links are induced by marginalization. We use $O_k = (S_{k-1} \cup G_k)\backslash(T_k \cup I_k)$ to denotes all (O)ther nodes in the subgraph.

Theoretical analysis shows Sequential LASSO conserves consistency if the selection is exact in each step, meaning that $X_{S_k} = X_{T_k \cup I_k}$ $\forall k = 1, \cdots, K$. This result is concise but the requirement is too strong. Our study shows that a relaxed requirement of X_{S_k} is sufficient to make Sequential LASSO stay consistency. In each step we alternatively require $T_k \subset S_k$ and $O_k \cap S_k = \emptyset$, $\forall k = 1, \ldots, K$. And we put no condition on I_k, which in practice extremely relaxes the strict requirement of LASSO. This is because that in practice the induced edges $X_{I_k} - Y$ tend to be of very small magnitude and easy to be missed.

3.2 Parallel LASSO

Parallel LASSO firstly performs selection independently on the partition $\{G_1, G_2, \cdots G_K\}$ to get feature set S_k and then performs selection on subgraph $\{X_{S_1 \cup S_2 \cdots S_K}, Y\}$ as a second stage. The structure chart in Figure-(2) show the details of Sequential LASSO.The performance guarantee for Parallel LASSO is similar to that of Sequential LASSO. We relax the requirement as $T_k \subset S_k$ and $O_k \cap S_k = \emptyset$, $\forall k = 1, \ldots, K+1$. Here, we take the second stage of selection as $(K+1)^{th}$ step.

4. RESULTS

We perform simulations on synthetic data to evaluate the performance of Sequential and Parallel LASSO. We demonstrate the stability of Sequential and Parallel LASSO for different feature set partitions. We show that the probability of exact recovery approaches 1 as the sample size becomes sufficient for Sequential and Parallel LASSO.

To demonstrate the application of sequential/parallel LASSO in practical scenarios, we apply our methods to genomic datasets available on human. Our results not only replicate some published eQTLs but also provide a new set of candidate eQTLs for further biological investigation.

5. CONCLUSION

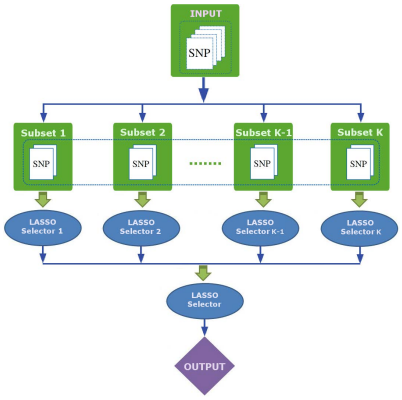

Figure 2: Structure Chart of Sequential LASSO

Sequential and Parallel LASSO are powerful tools to handle high-dimensional problems in a "divide and conquer" manner. The Parallel LASSO can be parallelized to alleviate computing time, and the Sequential LASSO is naturally extendable to a systematic data stream. Theoretical analysis based on Gaussian graphical models shows the consistency of our approaches and relaxed requirement over LASSO make our approach practically workable.

The essence of our methods is that we provide an extendable framework encoding LASSO models to find "direct" links between features and labels under the assumption of Gaussian distributions. Such a framework can be adapted for various applications, where we can encode many different machine learning models, not necessarily a LASSO variation or sparse learning models.

6. REFERENCES

[1] R. Tibshirani. Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society, Series B*, 58(1):267–288, 1996.

[2] D. Koller and N. Friedman. *Probabilistic graphical models: principles and techniques*. MIT press, 2009.