Regularized Lasso Approach for Parameter Fusion in Data Harmonization

Lu Tang and Peter XK Song, PhD
Department of Biostatistics, University of Michigan, Ann Arbor, MI

BIOSTATISTICS

INTRODUCTION

- Data sets from multiple different sources are often combined to increase statistical power. Examples include
  - Combining genetic studies
  - Combining data from study sites in multi-center clinical trial
  - Combining survey data from different countries
- **Always gain more power with combined data sets? Not necessarily true**. Heterogeneity leads to extra complexity due to
  - different study goals
  - different study populations
  - different experimental protocols
  - different study coordination
- Consequences: possibly biased result and misleading conclusion.
- Propose a new approach to harmonize data sets from different studies via regularized lasso. It has shown to be very useful and appealing especially when data heterogeneity exists.

BACKGROUND

EXISTING METHODS (to handle heterogeneity)

- Interaction of covariate and study indicator.
- Random study effects.
- Covariate effects may be grouped in term of similar effect size among some homogeneous studies.

RESEARCH GOALS

Primary goals

Goal 1: To be parsimonious in the model specification but in the meantime allowing enough parameters to capture the inter-study heterogeneity.

Goal 2: Define a score to quantify the heterogeneity level.

Secondary goal

Accommodate linear, binomial, and Poisson responses.

FUSED LASSO[1] (when $p = 1$)

We adapt the idea of fused lasso to achieve parameter fusion. The original objective can be translated into an optimization problem.

$$\min_{\beta \in \mathbb{R}^{Kp}} -\log \mathcal{L}(\beta) + P_\lambda(\beta)$$

$$P_\lambda(\beta) = \lambda \sum_{k=1}^{K-1} \sum_{k'>k}^{K} \omega_{k,k'} |\beta_k - \beta_{k'}|$$

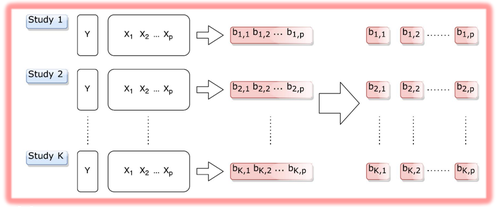
CHALLENGES

- Redundant penalty terms when considering all possible pairs.
- Numerical challenges of optimization under the generalized linear model (GLM) framework.

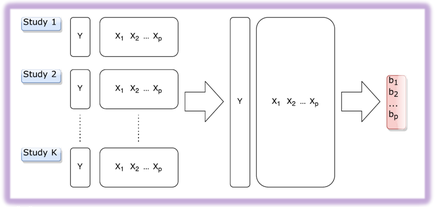
METHOD

THE IDEA

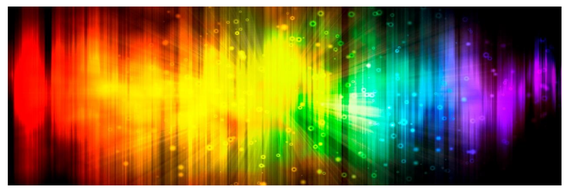
Consider the spectrum below where the two extremes represent the cases of heterogeneous (left) and homogeneous (right). Most existing methods focus on the two extremes. **We are interested in identifying a realistic model in between.**

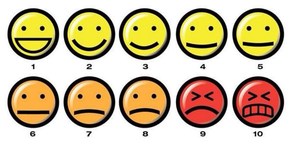


Heterogeneous



Homogeneous





"On a scale of 1 to 10"

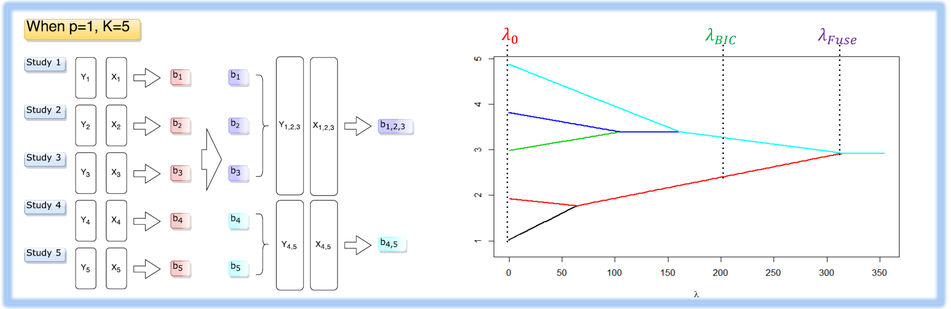
-Baymax

REDUCING TO LASSO (implemented using R package *penalized*)

If parameter ordering were known, the penalty term could be simplified as

$$P_\lambda(\beta) = \lambda \sum_{k=2}^{K} \omega_k |\beta_{(k)} - \beta_{(k-1)}|.$$

With linear reparametrization, the original penalty can be turned into a lasso penalty.



MAKING SENSE OF λ

Selection

Can be selected using AIC, BIC, or cross-validation. Here BIC is used.

Interpretation

If we view the solution path as the continuous spectrum above where λ_{BIC} captures the best compatibility of the data sets, then the relative location of $\boldsymbol{\lambda_{BIC}}$ with respect to $\boldsymbol{\lambda_{Fuse}}$ reflects the heterogeneity level of the collection of data sets.

EXTENSIONS

- Higher dimension (when p is large).
- Allow sparseness in estimation.
- Different λ for different parameter.
- Time-to-event response (survival).

SIMULATION

Simulation 1. Consider the mean response association $h(Ey_{ki}) = \beta_{k0} + \beta_{k1} x_{ki}$ where $K = 3;\ n_1 = n_2 = n_3 = 200$.

Matched pattern rate (MPR) is

$$MPR = \frac{\text{\# of replicates with matched grouping pattern}}{\text{\# of replicates}}.$$

True β pattern	Response type	Average $\hat{\beta}$ size	MPR of β_0	MPR of β_1	Overall MPR	Average λ_{BIC}	Average λ_{Fuse}	$\frac{\lambda_{BIC}}{\lambda_{Fuse}}$
Heterogeneous	Linear	6.00	1.00	1.00	1.00	0.00	250.63	0.000
β_0=(-1, 0, 1)	Binomial	5.99	0.99	1.00	0.99	0.03	64.54	0.005
β_1=(1, 0, -1)	Poisson	6.00	1.00	1.00	1.00	0.00	650.30	0.000
Mixed pattern	Linear	4.01	1.00	0.99	0.99	0.76	140.79	0.005
β_0=(1, 0, 1)	Binomial	4.05	0.93	0.92	0.86	1.13	29.65	0.038
β_1=(0.5, 1.5, 1.5)	Poisson	4.02	1.00	0.98	0.98	0.50	846.76	0.001
Homogeneous	Linear	2.05	0.97	0.98	0.95	0.68	0.86	0.791
β_0=(0.1, 0.1, 0.1)	Binomial	2.06	0.98	0.96	0.94	1.21	1.55	0.781
β_1=(0.5, 0.5, 0.5)	Poisson	2.04	0.98	0.98	0.96	0.70	0.81	0.864

* Number of simulation replicates = 100

Simulation 2. Consider linear response with a larger scale where $K = 50; p = 2;\ n_k = 50$ for $k = 1, \ldots, 50$.

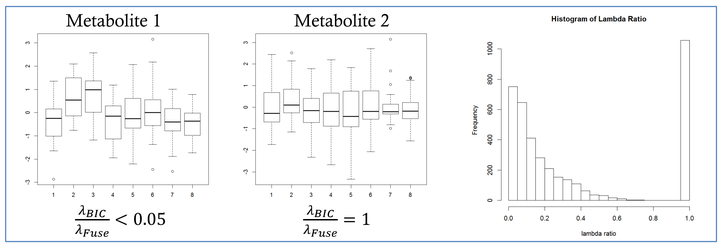
$$\beta_0 = \left(\underbrace{-1, \ldots, -1}_{15}, \underbrace{0, \ldots, 0}_{20}, \underbrace{1, \ldots, 1}_{15}\right);\quad \beta_1 = \left(\underbrace{1, \ldots, 1}_{15}, \underbrace{0, \ldots, 0}_{20}, \underbrace{-1, \ldots, -1}_{15}\right)$$

Average $\hat{\beta}$ size	MPR of β_0	MPR of β_1	Overall MPR	Average λ_{BIC}	Average λ_{Fuse}	$\frac{\lambda_{BIC}}{\lambda_{Fuse}}$
7.4	0.61	0.57	0.44	1.74	349	0.005

* Number of simulation replicates = 100

APPLICATION

- Data from the Early Life Exposure in Mexico to ENvironmental Toxicants (ELEMENT) Project, a study of prenatal exposure effect on adolescent growth.
- The 250 blood samples were split into 8 batches in assay, 3911 **metabolites** were measured from each individual.
- The above method is used for the fusion of mean parameter and to quantify heterogeneity level of each metabolite due to batch effect (for quality control).



[1] R. Tibshirani, M. Saunders, S. Rosset, J. Zhu, and K. Knight. Sparsity and smoothness via the fused lasso. Journal of the Royal Statistical Society: Series B (Statistical Methodology), 67(1):91-108, 2005.