MATH 680: Project Proposal

David Fleischer, Annik Gougeon

Last Update: 12 March, 2018

Introduction

We wish to investigate the novel nonparameteric regression techniques outlined by Haris, Shojaie, and Simon (2016). In particular, we wish to study the method of adaptive truncation through (convex) hierarchical penalization in order to understand its properties, as well as outline suitable classes of optimization techniques for such problems.

Problem Outline

Univariate Case

Consider the univariate problem of nonparametric estimation from covariate-response pairs $\{(x_i, y_i) \mid x_i, y_i \in \mathbb{R}\}_{i=1}^n$. We assume that each y_i are related to the corresponding x_i through the functional relationship $y_i = f(x_i) + \epsilon_i$, where ϵ_i are i.i.d. with zero mean and constant variance.

Let $Y = [y_1, ..., y_n]^T$ and $X = [x_1, ..., x_n]^T$, and let $\Psi_K \in \mathbb{R}^{n \times K}$ be the basis expansion of X such that the $(i,k)^{\text{th}}$ element of Ψ_K is given by

$$\Psi_{K(i,k)} = \psi_k(x_i), \quad 1 \le k \le K, 1 \le i \le n$$

where ψ_k is the k^{th} basis function. For the purposes of hierarchical penalization we assume the ψ_k are ordered by some measure of complexity (i.e., $\psi_1(x) = x$, $\psi_2(x) = x^2$, ...). Then, we model responses Y via the basis expansion/projection estimator $\boldsymbol{\beta}$

$$Y = \Psi_K \boldsymbol{\beta} + \epsilon.$$

The unpenalized projection estimator is the solution to the minimization problem

$$\widehat{\boldsymbol{\beta}}^{\text{proj}} = \underset{\boldsymbol{\beta} \in \mathbb{R}^K}{\arg\min} \frac{1}{2n} \|Y - \Psi_K \boldsymbol{\beta}\|_2^2. \tag{1}$$

We should note that the choice of a truncation level K is not immediately obvious. Setting K too large leads to high variance estimates as the dimension of the basis expansion increases, while setting K too small leads to high bias estimates as the basis expansion is unable to capture additional complexity between Y and X. The `hierbasis` proposal for this problem (Haris, Shojaie, and Simon (2016)) is to consider a saturated basis ($K = n$) and apply a hierarchical penalization on (1) in order to simultaneously select K and estimates of $\boldsymbol{\beta}$. In particular, the problem seeks to solve the problem

$$\hat{\beta}^{\texttt{hier}} = \underset{\beta \in \mathbb{R}^n}{\arg\min} \left\{ \frac{1}{2n} \|y - \Psi\beta\|_2^2 + \lambda \Omega(\beta) \right\} \tag{2}$$

where

$$\Omega(\beta) = \frac{1}{\sqrt{n}} \sum_{k=1}^{n} w_k \|\Psi_{k:n} \boldsymbol{\beta}_{k:n}\|_2,$$

such that $\Psi_{k:n}$ denotes the final $k, k+1, ..., n$ columns of Ψ_n, $\boldsymbol{\beta}_{k:n}$ the final $k, k+1, ..., n$ entries of $\boldsymbol{\beta}$, $w_k = k^m - (k-1)^m$, and m, λ are tuning parameters.

is a heirarchical penalty of the class introduced by Zhao, Rocha, and Yu (2009). (to do... F. Bach (2009), Kim and Xing (2010))

- Introduction/motivation.
- Background material (2-4 papers).
- Why the problem is important/interesting/challenging/worth our time.
- What we wish to accomplish through the scope of the project.

Proposal

Solving

To solve (2) we apply the resualts of Zhao, Rocha, and Yu (2009) and Jenatton et al. (2010) and (via Haris, Shojaie, and Simon (2016)). In particular, under the basis expansion $\Psi_n \in \mathbb{R}^{n \times n}$ of X, let $\ell = \{1, ..., n\}$ be the set of column indices of Ψ_n and consider K subsets of indices ℓ

$$\mathcal{G} = \{g_1, ..., g_K\}, \quad g_k \subset \ell.$$

Denote by $\boldsymbol{\beta}_{g_k} = (\beta_j)_{j \in g_k}$ to be the subset of $\boldsymbol{\beta}$ corresponding to indices g_k. Now, define our general hierarchical penalty $\Omega(\boldsymbol{\beta})$ by

$$\Omega(\boldsymbol{\beta}) \ = \sum_{g \in \mathcal{G}} w_g \|\boldsymbol{\beta}_{g,0}\|_2$$

where $\boldsymbol{\beta}_{g,0}$ denotes a vector of coefficients whose entries are identical to $\boldsymbol{\beta}$ for indices of β in g, and 0 for indices of β not contained within g, i.e.,

$$\begin{cases} \beta_i \in \boldsymbol{\beta}_{g,0} & \text{for all } i \in g \\ 0 \in \boldsymbol{\beta}_{g,0} & \text{for all } i \notin g \end{cases}$$

Consider the objective function

$$f(\boldsymbol{\beta}) + \lambda \Omega(\boldsymbol{\beta}),$$

where $f(\boldsymbol{\beta}) = \frac{1}{2}\|Y - \Psi\boldsymbol{\beta}\|_2^2$ is our least-squares loss function. In the proximal gradient descent algorithm we linearize L about a current estimate $\boldsymbol{\beta}'$ of $\boldsymbol{\beta}$, and update $\boldsymbol{\beta}$ as the solution to the proximal problem

$$\arg\min_{\boldsymbol{\beta}} \left\{ f(\boldsymbol{\beta}') + (\boldsymbol{\beta} - \boldsymbol{\beta}')^T \nabla f(\boldsymbol{\beta}') + \lambda \Omega(\boldsymbol{\beta}) + \frac{L}{2}\|\boldsymbol{\beta} - \boldsymbol{\beta}'\|_2^2 \right\}.$$

Dividing through by the fixed L and removing the constant $f(\boldsymbol{\beta}')$ we rewrite the above problem as

$$\arg\min_{\boldsymbol{\beta}} \left\{ \frac{1}{L}(\boldsymbol{\beta} - \boldsymbol{\beta}')^T \nabla f(\boldsymbol{\beta}') + \frac{1}{2}\|\boldsymbol{\beta} - \boldsymbol{\beta}'\|_2^2 + \frac{\lambda}{L}\Omega(\boldsymbol{\beta}) \right\}$$

$$\arg\min_{\boldsymbol{\beta}} \left\{ \frac{1}{L}\boldsymbol{\beta}^T \nabla f(\boldsymbol{\beta}') - \frac{1}{L}\boldsymbol{\beta}'^T \nabla f(\boldsymbol{\beta}') + \frac{1}{2}\boldsymbol{\beta}^T\boldsymbol{\beta} - \boldsymbol{\beta}^T\boldsymbol{\beta}' + \frac{1}{2}\boldsymbol{\beta}'^T\boldsymbol{\beta}' + \frac{\lambda}{L}\Omega(\boldsymbol{\beta}) \right\}$$

$$\arg\min_{\boldsymbol{\beta}} \left\{ \frac{1}{L}\boldsymbol{\beta}^T \nabla f(\boldsymbol{\beta}') - \boldsymbol{\beta}^T\boldsymbol{\beta}' + \frac{1}{2}\boldsymbol{\beta}^T\boldsymbol{\beta} + \frac{\lambda}{L}\Omega(\boldsymbol{\beta}) \right\}$$

$$\arg\min_{\boldsymbol{\beta}} \left\{ \boldsymbol{\beta}^T \left[\frac{1}{L} \nabla f\left(\boldsymbol{\beta}'\right) - \boldsymbol{\beta}' + \frac{1}{2}\boldsymbol{\beta} \right] + \frac{\lambda}{L}\Omega(\boldsymbol{\beta}) \right\}$$

temp

$$\arg\min_{\boldsymbol{\beta}} \left\{ \frac{1}{2}\boldsymbol{\beta}^T\boldsymbol{\beta} - \boldsymbol{\beta}^T\left(\boldsymbol{\beta}' - \nabla f\left(\boldsymbol{\beta}'\right)\right) + \frac{1}{2}\left({\boldsymbol{\beta}'}^T\boldsymbol{\beta}' - 2{\boldsymbol{\beta}'}^T\nabla f\left(\boldsymbol{\beta}'\right) + \nabla f\left(\boldsymbol{\beta}'\right)^T \nabla f\left(\boldsymbol{\beta}'\right) \right) + \frac{\lambda}{L}\Omega(\boldsymbol{\beta}) \right\}$$

$$\arg\min_{\boldsymbol{\beta}} \left\{ \frac{1}{2}\boldsymbol{\beta}^T\boldsymbol{\beta} - \boldsymbol{\beta}^T\left(\boldsymbol{\beta}' - \nabla f\left(\boldsymbol{\beta}'\right)\right) + \frac{1}{2}\left(\boldsymbol{\beta}' - \nabla f\left(\boldsymbol{\beta}'\right)\right)^T \left(\boldsymbol{\beta}' - \nabla f\left(\boldsymbol{\beta}'\right)\right) + \frac{\lambda}{L}\Omega(\boldsymbol{\beta}) \right\}$$

$$\arg\min_{\boldsymbol{\beta}} \left\{ \frac{1}{2}\left(\boldsymbol{\beta} - \left(\boldsymbol{\beta}' - \nabla f\left(\boldsymbol{\beta}'\right)\right)\right)^T \left(\boldsymbol{\beta} - \left(\boldsymbol{\beta}' - \nabla f\left(\boldsymbol{\beta}'\right)\right)\right) \frac{\lambda}{L}\Omega(\boldsymbol{\beta}) \right\}$$

$$\arg\min_{\boldsymbol{\beta}} \left\{ \frac{1}{2}\left\|\boldsymbol{\beta} - \left(\boldsymbol{\beta}' - \nabla f\left(\boldsymbol{\beta}'\right)\right)\right\|_2^2 + \frac{\lambda}{L}\Omega(\boldsymbol{\beta}) \right\}$$

References

Bach, Francis. 2009. "High-Dimensional Non-Linear Variable Selection Through Hierarchical Kernel Learning." *arXiv Preprint arXiv:0909.0844*.

Haris, Asad, Ali Shojaie, and Noah Simon. 2016. "Nonparametric Regression with Adaptive Truncation via a Convex Hierarchical Penalty." *arXiv Preprint arXiv:1611.09972*.

Jenatton, Rodolphe, Julien Mairal, Guillaume Obozinski, and Francis R Bach. 2010. "Proximal Methods for Sparse Hierarchical Dictionary Learning." In *ICML*, 487–94. 2010. Citeseer.

Kim, Seyoung, and Eric P Xing. 2010. "Tree-Guided Group Lasso for Multi-Task Regression with Structured Sparsity."

Zhao, Peng, Guilherme Rocha, and Bin Yu. 2009. "The Composite Absolute Penalties Family for Grouped and Hierarchical Variable Selection." *The Annals of Statistics*. JSTOR, 3468–97.