MATH 680: Project Proposal

David Fleischer, Annik Gougeon

Last Update: 29 March, 2018

Introduction

We wish to study the method of nonparametric regression with hierarchical penalization as outlined by Haris, Shojaie, and Simon (2016b). In particular, wish to investigate its properties, applications, as well as outline suitable classes of optimization techniques for such problems.

Background

Consider the problem of estimating the relationship between responses $Y = [y_1, ..., y_n]^T \in \mathbb{R}^n$ and predictors $\mathbb{X} = [\boldsymbol{x}_1, ..., \boldsymbol{x}_n]^T \in \mathbb{R}^{n \times p}$, $\boldsymbol{x}_i = [x_{i1}, ..., x_{ip}] \in \mathbb{R}^p$. Suppose that Y and $\mathbb{X}$ are relation through an additive relationship

$$y_i = \sum_{j=1}^p f_j\left(x_{ij}\right) + \epsilon_i.$$

The method of estimation outlined herein focuses on a nonparametric method of basis expansion/projection estimators of Y. Specifically, for each predictor $X_j = [x_{1j}, ..., x_{nj}]^T \in \mathbb{R}^n$, we generate the expansion of X_j according to a finite set of basis functions $\left\{\psi_k\left(z\right)\right\}_{k=1}^K$, where K is the truncation level that will be determined data-adaptively (discussed shortly). Let $\Psi_K^{(j)} \in \mathbb{R}^{n \times K}$ be the set of basis functions correspond to predictor X_j, with entry $(i,k)^{\text{th}}$ given by

$$\Psi^{(j)}_{K,(i,k)} = \psi_k(x_{ij}), \quad 1 \leq k \leq K,\, 1 \leq i \leq n.$$

Of present interest is the case of a *polynomial basis expansion* $\psi_k^{(j)}(z) = z^k$ so that the j^{th} predictor undergoes the expansion

$$X_j = \begin{bmatrix} x_{1j} \\ \vdots \\ x_{nj} \end{bmatrix} \mapsto \Psi_K^{(j)} = \begin{bmatrix} \psi_1(x_{1j}) & \psi_2(x_{1j}) & \cdots & \psi_K(x_{1j}) \\ \vdots & \vdots & \ddots & \vdots \\ \psi_1(x_{nj}) & \psi_2(x_{nj}) & \cdots & \psi_K(x_{nj}) \end{bmatrix} = \begin{bmatrix} x_{1j} & x_{1j}^2 & \cdots & x_{1j}^K \\ \vdots & \vdots & \ddots & \vdots \\ x_{nj} & x_{nj}^2 & \cdots & x_{nj}^K \end{bmatrix}.$$

We may estimate the additive functions $f_j(x_{ij})$ by the <u>sparse additive `hierbasis`</u> estimator $\widehat{f}_j\left(x_{ij}\right) = \sum_{k \leq K} \widehat{\beta}_{j,k}^{\texttt{S-hier}} \psi_k\left(x_{ij}\right)$, $j = 1, ..., p$, such that each $\widehat{\boldsymbol{\beta}}_j^{\texttt{S-hier}} \in \mathbb{R}^K$ is simultaneously estimated the penalized minimization problem

$$\left[\widehat{\boldsymbol{\beta}}_1^{\texttt{S-hier}}, ..., \widehat{\boldsymbol{\beta}}_p^{\texttt{S-hier}}\right] = \underset{\boldsymbol{\beta}_j \in \mathbb{R}^K}{\arg\min} \left\{ \frac{1}{2n} \left\| Y - \sum_{j=1}^p \Psi_K^{(j)} \boldsymbol{\beta}_j \right\|_2^2 + \lambda \sum_{j=1}^p \Omega_j\left(\boldsymbol{\beta}_j\right) + \frac{\lambda^2}{\sqrt{n}} \sum_{j=1}^p \left\| \Psi_K^{(j)} \boldsymbol{\beta}_j \right\|_2 \right\}, \qquad (1)$$

where

$$\Omega_j\left(\boldsymbol{\beta}_j\right) = \frac{1}{\sqrt{n}} \sum_{k=1}^K w_k \left\| \Psi_{k:K}^{(j)} \boldsymbol{\beta}_{j,\,k:K} \right\|_2$$

for $w_k = k^m - (k-1)^m$, $\Psi^{(j)}_{k:K}$ denotes the submatrix of columns $k, k+1, ..., K$, and $\boldsymbol{\beta}_{k:K}$ the corresponding subvector of $\boldsymbol{\beta}$.

The above penalty Ω_j is designed to provide a data-driven method of truncating the basis complexity to some $K_0 \leq K$, derived from the hierarchical group lasso penalty (Zhao, Rocha, and Yu (2009)), leading to hierarchical sparsity of the fitted parameters $\widehat{\beta}_k = 0 \implies \widehat{\beta}_{k'} = 0$, for $k' > k$. The second penalization term $\frac{\lambda^2}{n} \sum \left\| \Psi^{(j)}_K \boldsymbol{\beta}_j \right\|_2$ imposes additional sparsity through a linked penalization constant λ^2.

Solving for `hierbasis` Estimators

To solve (1) Haris, Shojaie, and Simon (2016b) applies the results of Zhao, Rocha, and Yu (2009), Jenatton et al. (2010), Jenatton et al. (2011). By writing the problem in the form

$$\min_{v \in \mathbb{R}^p} \left\{ \|u - v\|_2^2 + \lambda \Omega(v) \right\}, \tag{2}$$

where Ω is a hierarchical penalty of the form described in Zhao, Rocha, and Yu (2009), we may apply an efficient proximal gradient descent algorithm with complexity $O(p)$ (Jenatton et al. (2011)).

Proposal

Of consideration for this project, we wish to tackle the following questions:

(1) Can the `hierbasis` estimator procedure offer a material gain over the lasso estimator (Tibshirani (1996))? Preliminary tests, as well as the `hierbasis` documentation (Haris, Shojaie, and Simon (2016a)), suggest a marginal sparsity improvement with no worse predictive power, but at the cost of computational complexity.

(2) The `hierbasis` documentation (Haris, Shojaie, and Simon (2016a)) references a mixing parameter α controlling the relative importance of the hierarchical and the sparsity-inducing penalities. How does the manipulation of this parameter affect its performance? Is it feasible to select α through cross-validation?

(3) What is the effect of changing the form of the weights $w_k = k^m - (k-1)^m$ in the hierarchical penalty Ω? The documentation suggests implementing $m = 2$ or $m = 3$. Why are these two values optimal, and how does the procedure perform when another m is selected?

(4) What other optimization methods be used to solve `hierbasis`? Some initial trials seem to indicate some parameter convergence issues. Can other methods address the convergence issues?

(5) Can the ℓ_1 norm be implemented in either/both penalities in order to induce stricter sparsity? What are the consequences of using the ℓ_1 norm?

(6) How does the `hierbasis` estimator procedure and R package perform on new datasets and simulations? Is it feasible to use this method for large datasets, considering the compuation time. How does it compare to the lasso estimator (Tibshirani (1996)) in this regard?

References

Haris, Asad, Ali Shojaie, and Noah Simon. 2016a. *HierBasis: Nonparametric Regression and Sparse Additive Modeling.*

———. 2016b. "Nonparametric Regression with Adaptive Truncation via a Convex Hierarchical Penalty." *arXiv Preprint arXiv:1611.09972.*

Jenatton, Rodolphe, Julien Mairal, Guillaume Obozinski, and Francis Bach. 2011. "Proximal Methods for Hierarchical Sparse Coding." *Journal of Machine Learning Research* 12 (Jul): 2297–2334.

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

Tibshirani, Robert. 1996. "Regression Shrinkage and Selection via the Lasso." *Journal of the Royal Statistical Society. Series B (Methodological).* JSTOR, 267–88.

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.