Investigation of Sparse Hierarchical Regularization for Basis Expansion Methods

Exploration and Expansion Regression via `HierBasis`

David Fleischer Annik Gougeon

Last Update: 03 May, 2018

Contents

1 Introduction

The method of nonparametric regression regularization described in Haris, Shojaie, and Simon (2016b) provides a flexible framework and implementation of a sparse hierarchical penalty via the `R` package `HierBasis`. The proposal offered by Haris, Shojaie, and Simon (2016b) outlines a convex penaltization and estimation technique that is suggested to be well-suited to high-dimensional problems. In particular, we wish to verify and expand upon the `HierBasis` framework in the context of sparse additive modelling, focusing on the problem of prediction of a continuous response and variable selection.

1.1 Problem Description

We restrict the attention of this project to focus on the problem of regression of a continuous response $y = [y_n, ..., y_n] \in \mathbb{R}^n$ on a high-dimensional design matrix $\mathbb{X} = [\boldsymbol{x}_1, ..., \boldsymbol{x}_n]^T = [X_1, ..., X_p] \in \mathbb{R}^{n \times p}$, such that

$$\begin{aligned} \boldsymbol{x}_i &= [x_{i1}, ..., x_{ip}] \quad \text{(observation } i\text{)} \\ X_j &= [x_{1j}, ..., x_{nj}]^T \quad \text{(predictor } j\text{)}. \end{aligned}$$

We consider the problem of estimating additive components $\{f_j\}_{j=1}^p$ of the additive model

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

for a sparse set of active features embedded within the design matrix $\mathbb{X}$. The proposal offered by Haris, Shojaie, and Simon (2016b) considers the class of basis expansion estimators (Cencov (1962)) defined by a finite set of basis functions $\{\psi_k(z)\}_{k=1}^K$, with some notion of increasing complexity (in k) and for a truncation level K to be adaptively selected. Let $\Psi_K^{(j)} \in \mathbb{R}^{n\times K}$ be the basis expansion corresponding to the j^{th} predictor X_j, with $(i,k)^{th}$ entry associated with observation x_{ij} and basis function ψ_k,

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

Then, through the basis expansion functions, the design matrix $\mathbb{X} \in \mathbb{R}^{n\times p}$ maps to a set of p $(n \times K)$ matrices

$$\mathbb{X} \overset{\psi}{\mapsto} \left\{\Psi_K^{(j)} \in \mathbb{R}^{n\times K}\right\}_{j=1}^p.$$

Of present interest is the set of polynomial basis functions $\{\psi_k(z)\}_{k=1}^K = \left\{z^k\right\}_{k=1}^K$ so that

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

We estimate the additive components f_j by the sparse additive `HierBasis` estimator $\widehat{f_j}$ given by

$$\widehat{f_j}(x_{ij}) = \sum_{k=1}^K \widehat{\beta}_{j,k}^{\texttt{SA-hier}} \psi_k(x_{ij}), \quad j = 1, ..., p,$$

such that the $j = 1, ..., p$ coefficient vectors $\widehat{\boldsymbol{\beta}}_j^{\texttt{SA-hier}} = \left[\widehat{\beta}_{j,1}^{\texttt{SA-hier}}, ..., \widehat{\beta}_{j,K}^{\texttt{SA-hier}}\right] \in \mathbb{R}^K$ are simultaneously estimated by the solution of the minimization problem

$$\left[\widehat{\boldsymbol{\beta}}_1^{\texttt{SA-hier}}, ..., \widehat{\boldsymbol{\beta}}_p^{\texttt{SA-hier}}\right] = \underset{\boldsymbol{\beta}_1,...,\boldsymbol{\beta}_p}{\arg\min} \left\{ \frac{1}{2} \left\| yg - \sum_{j=1}^p \Psi_K^{(j)} \boldsymbol{\beta}_j \right\|_2^2 + \lambda \sum_{j=1}^p \Omega_j\left(\beta_j\right) + \frac{\lambda^2}{\sqrt{n}} \sum_{j=1}^p \left\| \Psi_K^{(j)} \boldsymbol{\beta}_j \right\|_2 \right\}, \quad (1)$$

where

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

for weights $w_k = k^m = (k-1)^m$ penalization weights for the k^{th}-order basis estimator, $\Psi_{k:K}^{(j)}$ denotes the submatrix of columns $k, k+1, ..., K$ of $\Psi_K^{(j)}$, and $\boldsymbol{\beta}_{j,k:K}$ denotes the corresponding subvector of $\boldsymbol{\beta}_j$.

The penalty described in (1) is defined by two terms. The first term containing Ω_j's is designed to provide a data-driven method of selecting the basis complexity/truncating the degree of the basis functions to some adaptively selected level $K_0 \leq K$. This term is derived from the hierarchical group lasso penalty (Zhao, Rocha, and Yu (2009)) and leads to hierarchical sparsity of the fitted parameters. That is, $\widehat{\beta}_{j,k} = 0 \implies \widehat{\beta}_{j,k'} = 0$ for all $k' \geq k$.

The second term in the sparse additive `HierBasis` penalty $\frac{\lambda^2}{\sqrt{n}} \sum_{j=1}^p \left\| \Psi_K^{(j)} \boldsymbol{\beta}_j \right\|_2$ imposes sparsity across the predictors $X_1, ..., X_p$ and induces additional sparsity across the solution space $\left\{\widehat{\boldsymbol{\beta}}_j^{\texttt{SA-hier}}\right\}_{j=1}^p$.

1.2 Problem Convexity

It is important to mention the convexity properties contained within this problem. First, we note that the `HierBasis` penalty, $\Omega(\beta)$ is of the hierarchical group lasso form (Zhao, Rocha, and Yu (2009)). That is, it belongs to the CAP (Composite Absolute Penalties) family of penalties, which enables the solution to achieve hierarchical sparsity. According to Zhao, Rocha, and Yu (2009), CAP estimators lead to stable estimates along with a more effective use of degrees of freedom. However, they do not generally result in sparser estimates than the lasso (Tibshirani (1996)).

We define the CAP penalty for hierarichal solution. Introduce a node that corresponds to some group of variables, say G_k, and let there be a total of n nodes. For every group G_k, define $G_{k:n}$ as the groups that should only be added to the model after G_k. For example, in a model with both main and interaction effects, the group of interaction effects can only be added after its main effects are included in the model. In the `HierBasis` case, this consists of the higher order terms of the set of basis functions, ψ_n.

Then, a hierarchical sparsity inducing CAP penalty can be defined as

$$T(\beta) = \sum_{i=1}^{n} \alpha_i \cdot \|(\beta_{G_i}, \beta_{G_{i:n}})\|_{\gamma_i},$$

where $\alpha_m > 0, \forall m$ and $1 \leq \gamma_i < \infty$. Note that α_m is a correction factor in the case that a coefficient appears in numerous groups. An important theorem from Zhao, Rocha, and Yu (2009) allows us to obtain convexity:

Theorem (Zhao, Rocha, and Yu (2009)) If $\gamma_i \geq 1, \forall i = 1, ...n$, then $T(\beta)$ is convex. Furthermore, if the loss function L is convex in β, then the objective function of the CAP optimization problem is convex.

It follows that our estimators $\widehat{\beta}_1^{\texttt{SA-hier}}, ..., \widehat{\beta}_p^{\texttt{SA-hier}}$ are indeed convex.

1.3 Solving the `HierBasis` Estimators

To solve the sparse additive problem Haris, Shojaie, and Simon (2016b) first solves the equivalent univariate problem by applying the results of Zhao, Rocha, and Yu (2009), Jenatton et al. (2010), Jenatton et al. (2011). By writing the problem in the form

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

where Ω is a hierarchical penalty of the form described above. We may apply the following proximal gradient descent algorithm (with complexity $O(p)$) (Jenatton et al. (2011)). Let $\Psi = UV$ such that $U \in \mathbb{R}^{n \times K}$, $U^T U/n = \mathbb{I}_K$, can be obtained via a QR decomposition on the basis expansion matrix of Ψ. Then, the univariate `HierBasis` problem

$$\widehat{\boldsymbol{\beta}}^{\texttt{hier}(K)} = \arg\min_{\boldsymbol{\beta} \in \mathbb{R}^K} \left\{ \frac{1}{2} \|y - \Psi_K \boldsymbol{\beta}\|_2^2 + \frac{\lambda}{\sqrt{n}} \sum_{k=1}^{K} (k^m - (k-1)^m) \|\Psi_{k:K} \boldsymbol{\beta}_{k:K}\|_2 \right\}$$

is solved by reformulating it in a proximal-gradient-descent-friendly format

$$\min_{\boldsymbol{\beta} \in \mathbb{R}^K} \left\{ \frac{1}{2} \left\|U^T y/n - \boldsymbol{\beta}\right\|_2^2 + \lambda \sum_{k=1}^{K} w_k \|\boldsymbol{\beta}_{k:K}\|_2 \right\}$$

which itself can be solved via a coordinate descent algorithm (Haris, Shojaie, and Simon (2016b))

With the univariate `HierBasis` estimators solved, we may now introduce the solution to the sparse additive `HierBasis` estimators using a block coordinate descent algorithm (Haris, Shojaie, and Simon (2016b))

Algorithm 1 Solving the Univariate `HierBasis` Problem

procedure HierBasis($y, U, \lambda, \{w_k\}_{k=1}^K$)
  Initialize $\boldsymbol{\beta}^{(1)} = \cdots = \boldsymbol{\beta}^K \leftarrow U^T y/n$
  for k = K, ..., 1 **do**
    Update $\boldsymbol{\beta}_{k:K}^{k-1} \leftarrow \left(1 - \frac{w_k \lambda}{\|\boldsymbol{\beta}_{k:K}^k\|_2}\right)_+ \boldsymbol{\beta}_{k:K}^k$
  return $\boldsymbol{\beta}^1$

Algorithm 2 Solving the Sparse Additive `HierBasis` Problem

procedure AdditiveHierBasis($y, \left\{\Psi_K^{(j)}\right\}_{j=1}^p, \lambda, \{w_k\}_{k=1}^K, \texttt{maxiter}$)
  Initialize $\beta_j \leftarrow 0$ for $j = 1, ..., p$
  while $l \leq$ `maxiter` **and** not converged **do**
    for $j = 1, ..., p$ **do**
      Set $r_{-j} \leftarrow y - \sum_{j' \neq j} \Psi_K^{(j')} \boldsymbol{\beta}_{j'}$
      Set $\tilde{w}_1 = w_1 + \lambda$, $\tilde{w}_k = w_k$, for $k = 2, ..., K$
      Update $\boldsymbol{\beta}_j \leftarrow \arg\min \left\{ \frac{1}{2n} \left\| r_{-j} - \Psi_K^{(j)} \boldsymbol{\beta} \right\|_2^2 + \frac{\lambda}{\sqrt{n}} \sum_{k=1}^K \tilde{w}_k \left\| \Psi_{k:K}^{(j)} \boldsymbol{\beta}_{j,k:K} \right\|_2 \right\}$
  return $\boldsymbol{\beta}_1, ..., \boldsymbol{\beta}_p$

1.4 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) 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?

2 Methods

2.1 The `hierbasis2` Package

We have create a companion package to `HierBasis`, named `hierbasis2`, in order to implement the above tests and features of the above proposal. This new package retains all of the user-facing functionality of the original `HierBasis` package, but now permits the user to manipulate some additional parameters, as well as introducing some new functions. That is, the new library has been designed with this project in mind, allowing us to explore the properties of the original `HierBasis` package in a modular, readable, and concise format.

2.1.1 Installation

As is the case for `HierBasis`, installation of `hierbasis2` can be done via `devtools::install_github`

```
#install.packages(devtools)
library(devtools)
install_github("dfleis/hierbasis2")
library(hierbasis2)
```

3 Results

4 Discussion

References

Cencov, NN. 1962. "Estimation of an Unknown Density Function from Observations." In *Dokl. Akad. Nauk, Sssr*, 147:45–48.

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.