1 2/7/2014

1.1 Best Shape-Constrained Projection

Even if a function f_0 is non-zero, it may be possible that the best convex approximation is zero. But, it cannot be that the best convex approximation and the best concave approximation are both zero.

Theorem 1.1. *Let $f_0 : C \to \mathbb{R}$ be a function with a bounded Hessian; let p be some distribution on C. Suppose*

$$\operatorname*{argmin}_{f \in \mathcal{C}_x \cup \mathcal{C}_v} \mathbb{E}(f_0(x) - f(x))^2 = 0$$

then $f_0 = 0$ necessarily.

Proof. Suppose that the $\operatorname{argmin}_{f \in \mathcal{C}_x \cup \mathcal{C}_v} \mathbb{E}(f_0(x) - f(x))^2 = 0$, then it must be that

$$\operatorname*{argmin}_{c \in \mathbb{R}} \mathbb{E}\Big(f_0(x) - cx^\mathsf{T} x\Big)^2 = 0$$

Since the optimal c in the above optimization is $c^* = \frac{\mathbb{E}[f_0(x)x^\mathsf{T}x]}{\mathbb{E}(x^\mathsf{T}x)^2}$, we have that $\mathbb{E}[f_0(x)x^\mathsf{T}x] = 0$.

We know that there exists a convex function f' such that $f' = f_0 + rx^\mathsf{T}c$ for some large $r > 0$.

By assumption and convexity of f', it must be that

$$\operatorname*{argmin}_{c \in \mathbb{R}} \mathbb{E}\Big(f_0(x) - cf'(x)\Big)^2 = 0$$

We know by similar argument then that $\mathbb{E}[f_0(x)f'(x)] = 0$.

However, $\mathbb{E}[f_0(x)f'(x)] = \mathbb{E}[f_0(x)^2 + rf_0(x)x^\mathsf{T}x] = \mathbb{E}[f_0(x)^2]$. This is a contradiction.

□

Thus, if we tried both convex and concave projection and the solution is zero, then we can be sure that the original function is identically zero.

The same argument applies if we consider the best single component approximation:

$$\operatorname*{argmin}_{f_k \in \mathcal{C}_x \cup \mathcal{C}_v} \mathbb{E}\Big(f_0(x) - f_k(x_k)\Big)^2$$

Because $\mathbb{E}[f_0(x)x_k^2] = \mathbb{E}\Big[\mathbb{E}[f_0(x) \mid x_k]x_k^2\Big] = \mathbb{E}[f_k^*(x_k)x_k^2]$. Thus, we can apply the same argument and use $f_k^*(x_k)$ where needed.

2 1/23/2014

2.1 Additive Faithfulness Case Study with Quadratic Function and Gaussian Distribution

We consider a quadratic function $f(x) = x^\mathsf{T} Hx + c^\mathsf{T} x$ and a Gaussian distribution $X \sim N(0, \Sigma)$.

We then have a closed form for the additive approximation.

- If f does not depend on x_j, then $f_j^*(x_j) = 0$.
- If f depends on x_j, then, letting H_j be the j-th row of H and Σ_j be the j-th row of Σ:

$$f_j^*(x_j) = H_j^\mathsf{T} \Sigma_j \frac{1}{\Sigma_{jj}} x_j^2 + c_j x_j$$

Let us assume $\Sigma_{jj} = 1$ for all j and that $c = 0$ for convenience. We then have two direct corollaries:

Corollary: We have additive convexity if and only if $\text{diag}(H\Sigma) \geq 0$.

Corollary: We have additive faithfulness if and only if $\text{diag}(H\Sigma) \neq 0$.

As an example where additive convexity and additive faithfulness are violated. Let $H = [1, 2; 2, 5]$ and $\Sigma = [1, -c; -c, 1]$. For $c = 0.5$, additive faithfulness is violated; for $c > 0.5$, additive convexity is violated.

Proof. We will show that the $f_j^*(x_j)$'s, so described, satisfy the KKT stationarity equations.

$$f_j^*(x_j) = \mathbb{E}[f(x) - \sum_{k \neq j} f_k^*(x_k) \,|\, x_j] \qquad \text{for all } j$$

To prove this, we use the following conditional mean and conditional covariance property of the multivariate Gaussian distribution.

$$\mathbb{E}[x_k \,|\, x_j] = \Sigma_{jk} \Sigma_{jj}^{-1} x_j$$
$$\mathbb{E}[x_k x_{k'} \,|\, x_j] \text{ is some constant for all } x_j$$
$$\mathbb{E}[x_k^2 \,|\, x_j] \text{ is some constant for all } x_j$$

□

Why can Gaussian distribution violate additive faithfulness? Because $\frac{\partial p(x_{-j} \,|\, x_j)}{\partial x_j}$ is always large for some values of x_{-j}.

3 1/2/2014

3.1 Convex-minus-Quadratic Estimation

Instead of estimating a convex-plus-concave function, it is theoretically sufficient to estimate a convex-minus-quadratic function.

Theorem 3.1. *Any function $h : \mathbb{R}^d \to \mathbb{R}$ with a bounded Hessian can be decomposed as $h(x) = f(x) - q(x)$ where $f(x)$ is convex and $q(x)$ is $cx^\mathsf{T}x$ for some $c \geq 0$.*

Proof. For large enough c, the Hessian of $h + c$ is a positive semidefinite. The function f is thus convex. □

Convex-minus-quadratic functions are faster to learn. The optimization program has about twice as few variables. It is also possibly easier to analyze.

3.1.1 Implementation

For additive modeling, I use the following optimization program. There are two parameters λ and L.

$$\begin{aligned} \min_{f_1,\ldots,f_p} \frac{1}{n} \sum_i \Big(y_i - \sum_j f_j(x_{ij}) + c_j x_{ij}^2 \Big)^2 + \lambda \sum_j \|\partial f_j - 2c_j x_j^2\|_\infty \\ \text{s.t. } f_j \text{ is convex} \\ c_j \geq 0 \quad \text{and} \quad c_j \leq L \end{aligned}$$

The second L parameter is necessary. A similar parameter is required in convex-plus-concave estimation as well, which I will explain in the next section. L can be interpreted as a lower bound on the second derivative of the estimated regression function. We set $L = 200$ in experiments.

Demonstration: We estimate a one-dimensional function so that we can visualize the behavior of convex-minus-quadratic functions. We set $lambda = 0$. The result is in figure 3.1.1.

Here, $n = 300$, the SNR is about 0.6 (high noise).

3.2 Tuning Parameters for Convex-plus-Concave Estimation

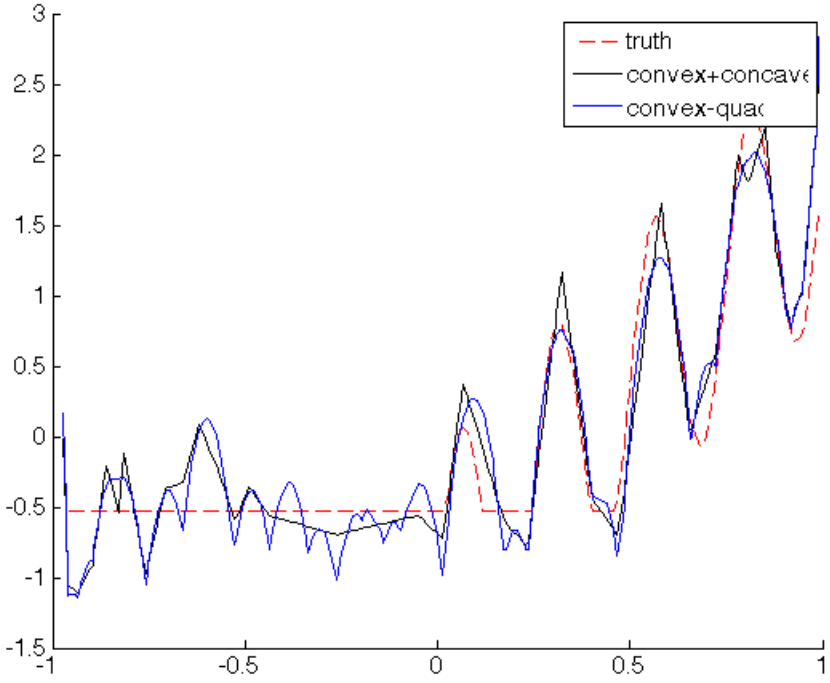
Convex-plus-concave estimation is not free of tuning parameters. Arbitrary sum of convex and concave functions can represent any function with a bounded Hessian.

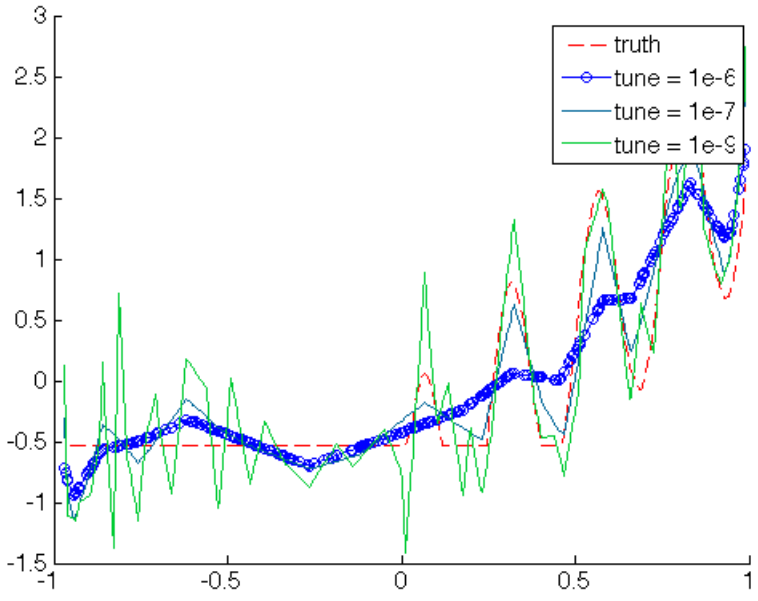
In Minhua's original SCCAM implementation, the objective is augmented with a square-penalty on the value of the gradients of f and g, the convex and the concave functions. The penalty looks like $10^{-6}\|\text{gradient}\|_2^2$. This penalty is not just for numerical stability; its magnitude affects the estimation.

Demonstration:

The experimental set-up is the same as before. The result is in figure 3.2

As one can see, when the tuning parameter (the coefficient for the $\|\text{gradient}\|_2^2$ penalty) is too small, the estimated function is fitting noise. When the tuning parameter is too large, the estimated function is too smooth.


truth
convex+concave
convex-quad
3
2.5
2
1.5
1
0.5
0
-0.5
-1
-1.5
-1
-0.5
0
0.5
1


truth
tune = 1e-6
tune = 1e-7
tune = 1e-9
3
2.5
2
1.5
1
0.5
0
-0.5
-1
-1.5
-1
-0.5
0
0.5
1

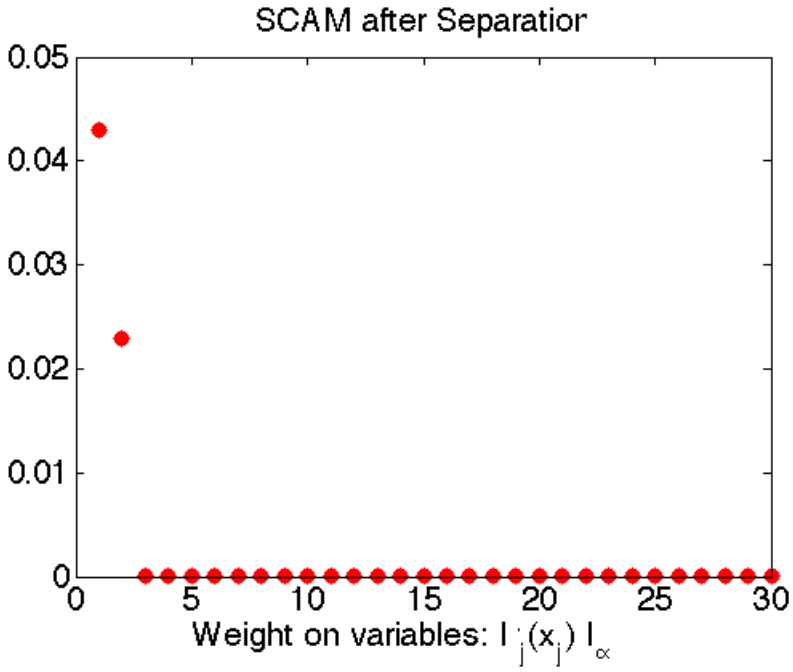
3.3 Convex-Concave Separation

We study the problem of variable selection on a convex-plus-concave function h. The approach has two steps:

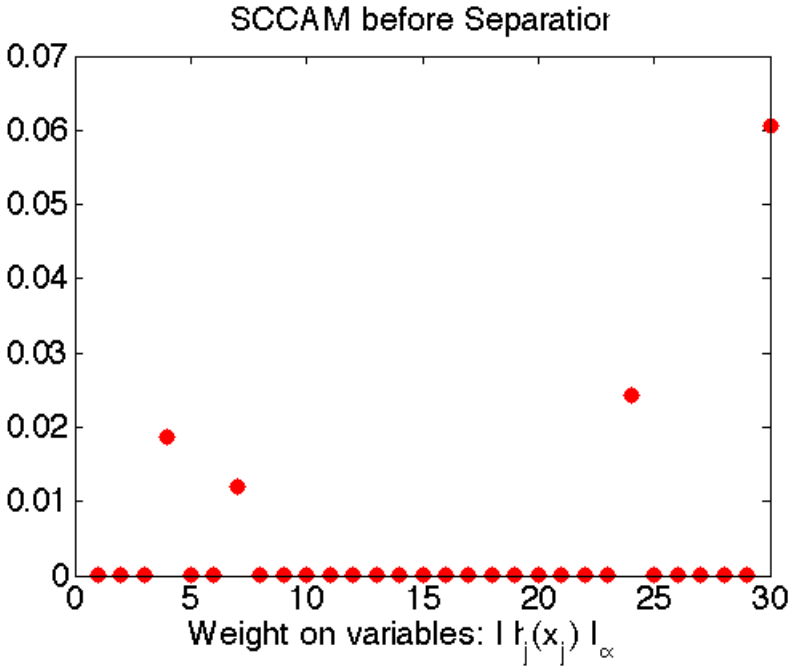
1. Learn **sparse non-additive** convex function f and concave function g such that $h = f + g$. This is the separation stage.
2. Apply sparse additive model on f and g.

Preliminary results are favorable. In our experiment, we let $p = 30, n = 300$. The true function is $h(x) = 2x_1x_2$ and uses only the first two variables. The SNR is 4. h is an example of a function that cannot be consistently estimated by additive modeling.

We can indeed identify the correct variables by applying additive model on f after separation.



Without separation however, the additive model indeed fails.


SCCAM before Separatior
0.07
0.06
0.05
0.04
0.03
0.02
0.01
0
0
5
10
15
20
25
30
Weight on variables: I h_j(x_j) I_∝