1 Some Insight

Let us again consider this noiseless optimization:

$$\min_{B} \|B\|_{\infty,1}$$
$$\text{s.t. } y_j \geq y_i + B_i^{\mathsf{T}}(X^j - X^{i*}) \forall i, j$$

First, notice that, because (1) each constraint involves only one B^i and (2) the y_j's are not variables, the optimization can be decomposed into n sub-optimizations each done independently.

$$\min_{\beta \in \mathbb{R}^p} \|\beta\|_1$$
$$\text{s.t. } y_j \geq y_{i*} + \beta^{\mathsf{T}}(X^j - X^{i*}) \forall j$$

This optimization is analogous to finding a subgradient at just X^{i*}. This decomposition allows us to compare directly with the linear case because now, in both cases, we are solving for a single vector.

Let us simply the expression further. Define $\Delta \in \mathbb{R}^{p \times n}$ where the j-th row $\Delta_j = X^j - X^{i^*}$, and define $\mathbf{v} \in \mathbb{R}^n$ where $\mathbf{v}_j = y_j - y^{i*}$. Now the optimization takes on a familiar form.

$$\min_{\beta \in \mathbb{R}^p} \|\beta\|_1$$
$$\text{s.t. } \mathbf{v} \geq \beta^{\mathsf{T}} \Delta$$

The optimization is similar to the linear case, but we have an inequality instead of a strict equality here. The inequality is the culprit!

The solution to the optimization will lie on the intersection of the boundary of the constraint set and the boundary of a L_1 norm ball of some minimal radius.

Let $\widehat{\beta}$ be the solution to the above optimization. Because $\widehat{\beta}$ lies on the boundary of the constraint set, some inequalities must be equalities. Let I be the set of inequalities which became equalities when we plug in $\widehat{\beta}$. Then $\widehat{\beta}$ is equivalently the solution of the following optimization:

$$\min_{\beta \in \mathbb{R}^p} \|\beta\|_1$$
$$\text{s.t. } \mathbf{v}_I = \beta^{\mathsf{T}} \Delta_I$$

where Δ_I is a $p \times |I|$ sub-matrix and $\mathbf{v}$ is a $|I|$-dimensional sub-vector. Now we see the problem: if $|I|$ is very small, then it is as if we are solving lasso with

very few samples–$|I|$ number of samples instead of n!

How do we interpret I? Intuitively, I think of it as all j for which $y_j \approx y_{i*} = B_{i*}^{\mathsf{T}}(X^j - X^{i*})$. That is, it is the set of points for which the convex function is close to linear. For strongly convex functions like $\|x\|_2^2$, I would be very small. For linear functions like $\beta^{*\mathsf{T}}x$, $|I| = n$.

Therefore, this explains why our optimization gives sparse solution if the true function is linear and highly non-sparse solution if the true function is strongly convex.

2 Experiments

In the soft-recovery experiment, we measure $\frac{\|B_S\|_{\infty,1}}{\|B\|_{\infty,1}}$. The recovery curve plateaus. It should converge to 1 but appears to be doing so very slowly.

As a thought experiment, let's consider a noiseless support recovery optimization:

$$\begin{aligned} &\min_B \|B\|_{\infty,1} \\ &\text{s.t. } y_i \geq y_j + B_j^{\mathsf{T}}(X^i - X_j) \forall i, j \end{aligned}$$

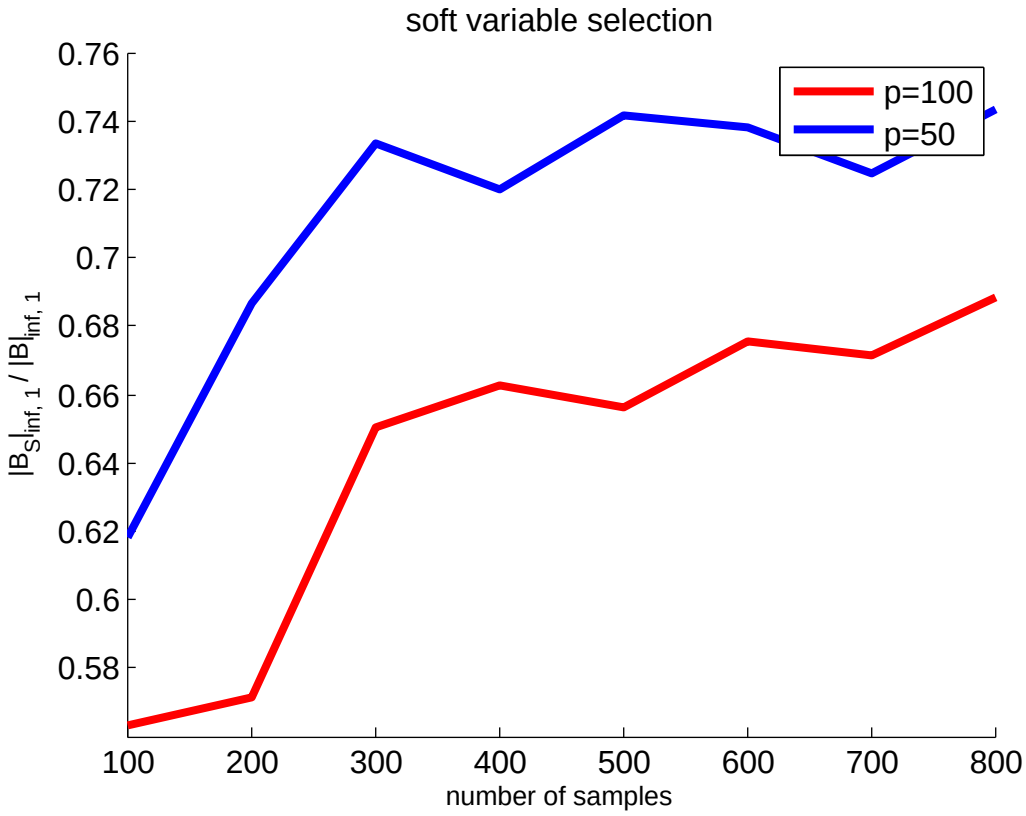
The optimum of this program should become exactly sparse as $n \to \infty$, but may become sparse at a very slow rate with respect to n. The penalized least-square solution then, like a heat-seeking missile thrown off target by a decoy, homes in to the non-sparse optimum rather than the actual sparse solution.

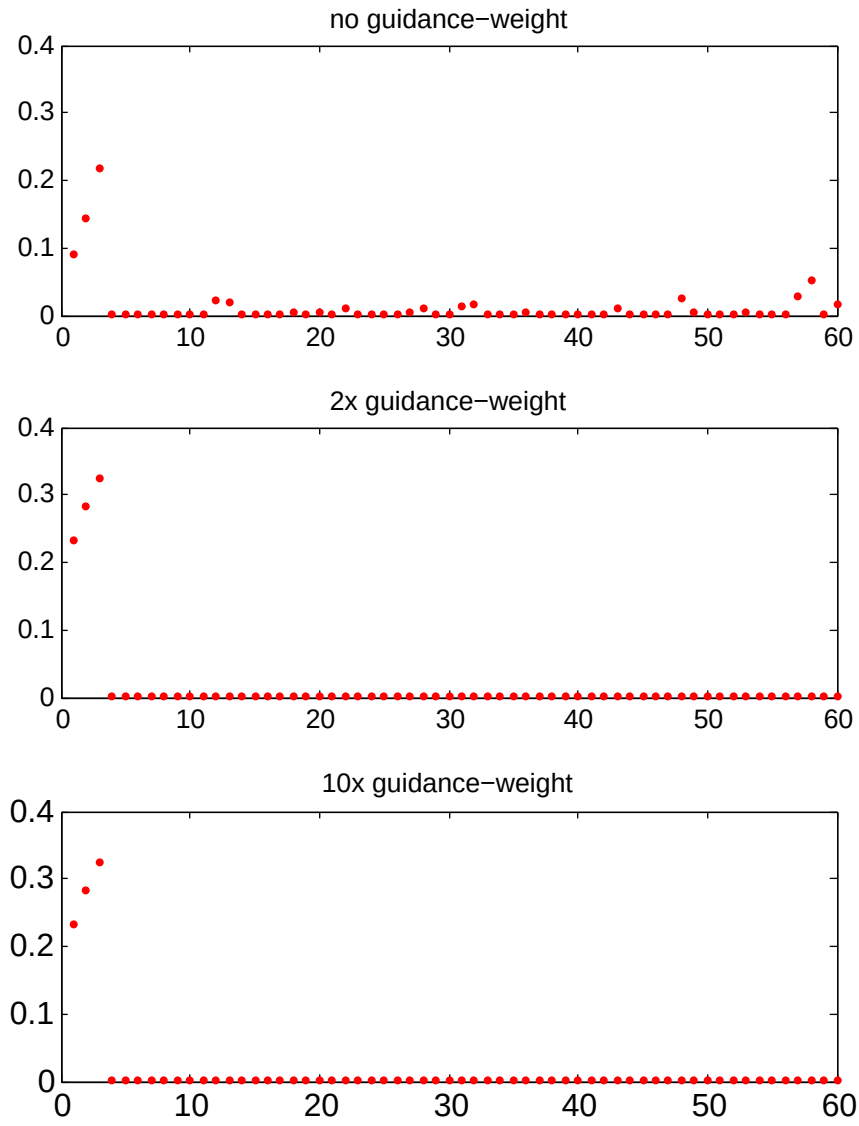
We have not implemented this optimization but can still approximate it with the penalized square error optimization by setting lambda very small.

What we see if that we must weigh the dimensions to exactly recover the support. We must use a weighed form of the objective.

$$\begin{aligned} &\min_B \sum_{s=1}^{p} \lambda_s \|B_{s\cdot}\|_\infty \\ &\text{s.t. } y_i \geq y_j + B_j^{\mathsf{T}}(X^i - X_j) \forall i, j \end{aligned}$$

Where λ_s is small for $s \in S$ and large for $s \notin S$.


soft variable selection
p=100
p=50
0.76
0.74
0.72
0.7
0.68
0.66
0.64
0.62
0.6
0.58
100
200
300
400
500
600
700
800
number of samples
$|B_S|_{inf,1} / |B|_{inf,1}$


no guidance−weight
0.4
0.3
0.2
0.1
0
0
10
20
30
40
50
60
2x guidance−weight
0.4
0.3
0.2
0.1
0
0
10
20
30
40
50
60
10x guidance−weight
0.4
0.3
0.2
0.1
0
0
10
20
30
40
50
60