Lecture 23: Introduction to Sparse Recovery

Up to this point in the course we have seen many problems & models, all of which roughly are of the form

$$\text{(*)} \quad \min_{\underline{\beta}} \frac{1}{2}\|A\underline{\beta} - \underline{Y}\|_2^2 + \underbrace{\frac{\lambda}{2}\|\beta\|_2^2}_{\text{Regularization}},$$

ex supervised learning, Ridge regression, graphical SSL, classification, kernel regression broadly.

We also saw, through Cross-Validation, that the choice of $\lambda \geq 0$ has a profound impact on the quality of our solutions. This is super important when A has fewer rows than columns! ie, the system $A\underline{\beta} = \underline{Y}$ is underdetermined.

Observe that if we differentiate (*) wrt $\underline{\beta}$ we get.

$$A^T(A\hat{\beta} - Y) + \lambda\hat{\beta}$$

$$\Rightarrow (A^TA + \lambda I)\hat{\beta} = A^TY.$$

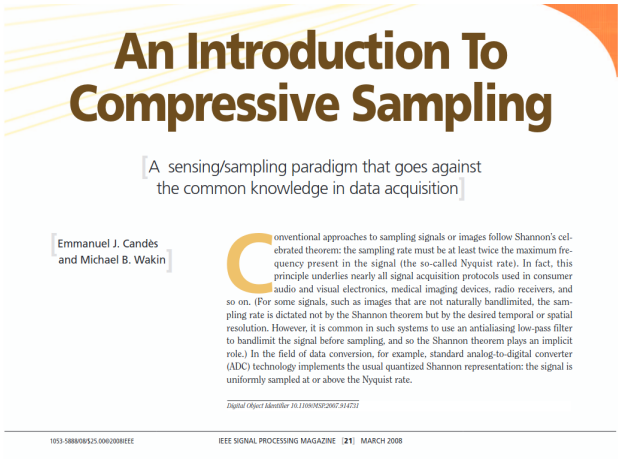
if $A \in \mathbb{R}^{N\times M}$ with $N < M$ then $A^TA \in \mathbb{R}^{M\times M}$ is rank deficient & so not invertible! Then adding λI to A^TA results in an invertible matrix & stabelizes the solution $\hat{\beta}$.

Hence, the term "regularization" since λI makes the problem more regular.

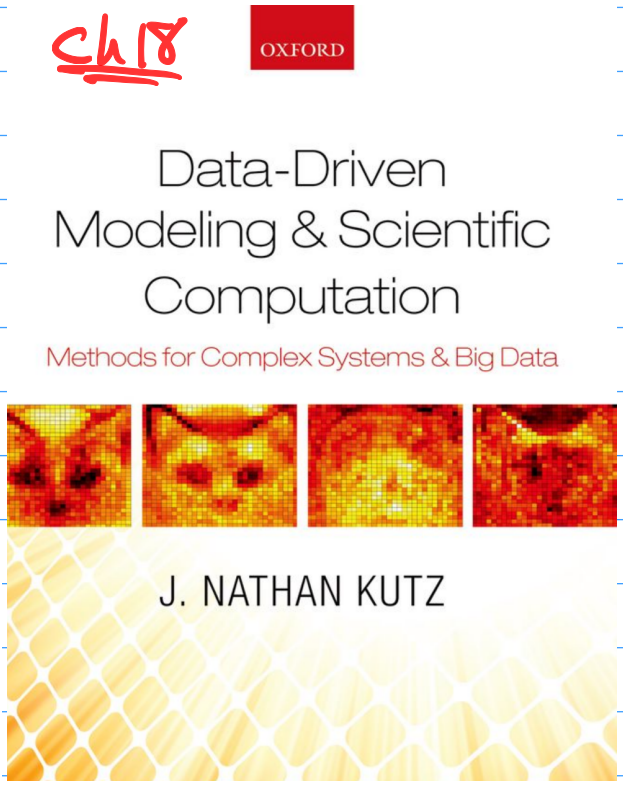
But one might wonder, what is special about $\|\beta\|_2$? why not another norm?

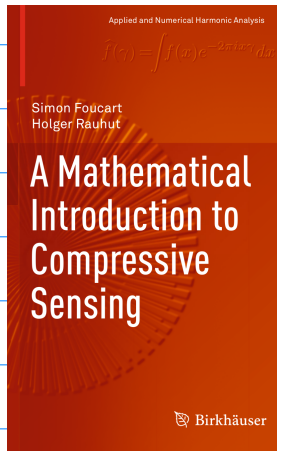
This is an excellent question & as it turns out it has profound implications. Particular choices of the norm, ex $\|\beta\|_p$ with $p \leq 1$ turns out to be very helpful. This is an advanced topic at the intersection of signal processing, convex analysis, applied math & statistics!

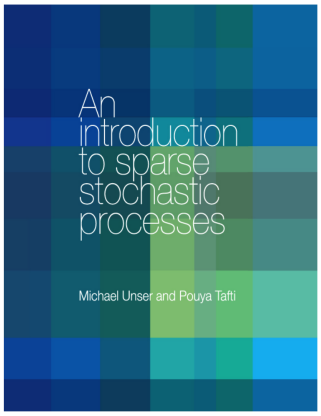
This line of thought also gave rise to the field of Compressed Sensing!



ch 18







23.1 The Role of Sparsity

The notion of "sparsity" is going to be key for us.

Consider a linear system

$A\beta = Y, \qquad \beta \in \mathbb{R}^N, \; Y \in \mathbb{R}^M, \; A \in \mathbb{R}^{M \times N}$

where $M < N$, so the system is underdetermined so it cannot be solved uniquely.

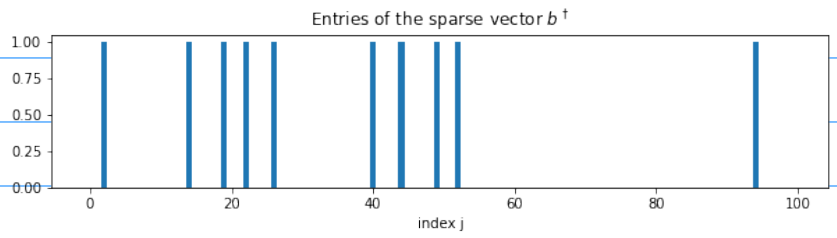
However, suppose that we have additional information that the vector $\underline{\beta}$ is sparse, more precisely it only has s-non-zero entries! (we say $\underline{\beta}$ is s-sparse)

Then, you might wonder whether you could find $\underline{\beta}$ if $s \ll M < N$?

Let us look at a quick example.

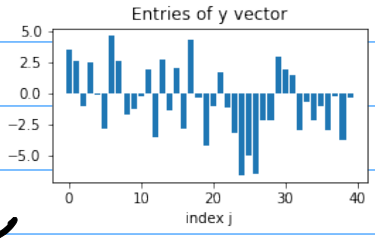
A vector $\underline{\beta}^t \in \mathbb{R}^{100}$ with only $s=10$ non-zero entries & a matrix $A \in \mathbb{R}^{40 \times 100}$ with random entries $A_{ij} \overset{iid}{\sim} N(0,1)$. Then set $\underline{y} = A\underline{\beta}^t \in \mathbb{R}^{40}$.

Our goal is then to infer $\underline{\beta}^t$ from $\underline{b}$, i.e., "solve" $A\underline{\beta} = \underline{y}$.



A

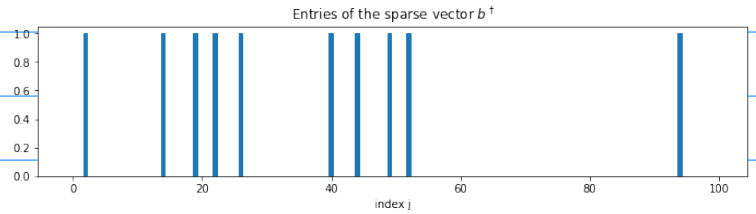
?

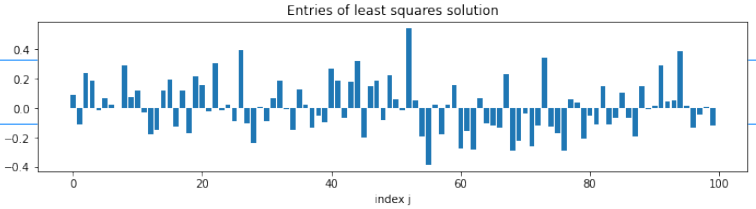


First thing we might try is least squares but this does not work well!

$$\underline{\beta}_{lsq} = \underset{\underline{\beta} \in \mathbb{R}^{100}}{\arg\min} \|A\underline{\beta} - \underline{y}\|_2^2$$

$\underline{\beta}_{lsq}$





We now try a diffent approach call Lasso which solves

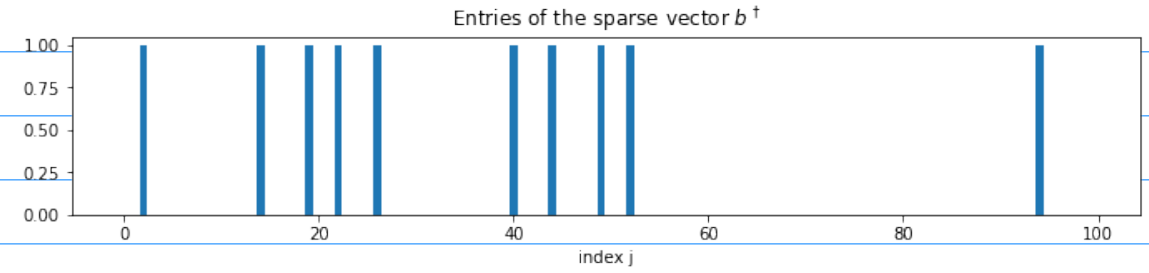
$$\beta_{Lasso} = \underset{\beta \in \mathbb{R}^{100}}{\arg\min} \frac{1}{2} \|A\beta - y\|_2^2 + \lambda \|\beta\|_1,$$

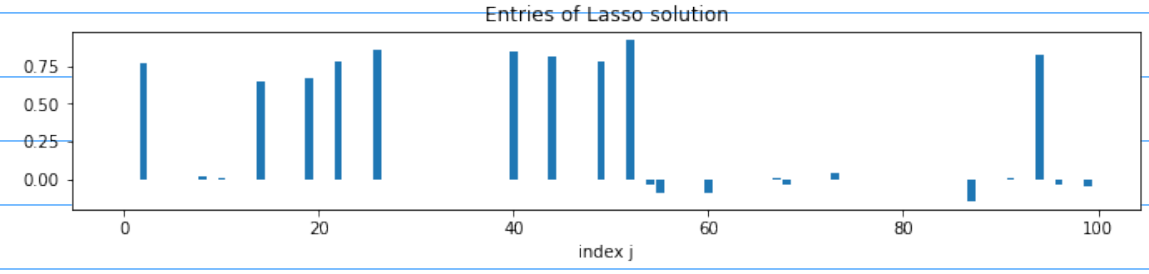
Compare to Ridge

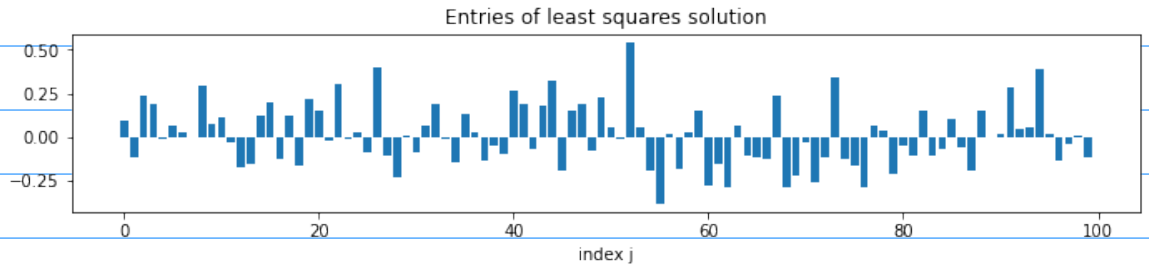
$$\beta_{Ridge} = \arg\min \frac{1}{2} \|A\beta - y\|_2^2 + \lambda \|\beta\|_2^2.$$

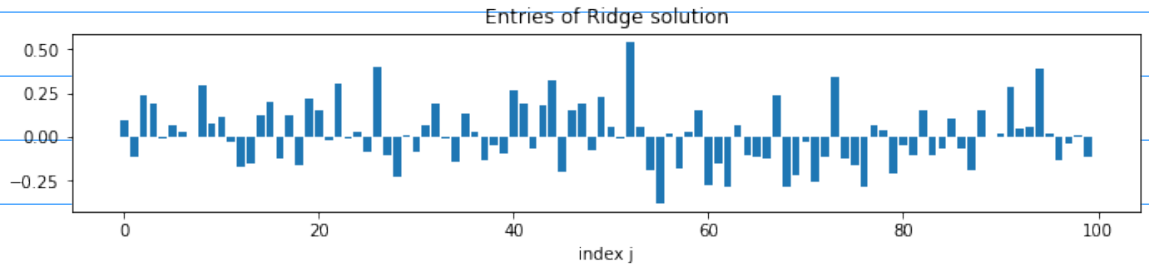
essentially we use a diffent regularization term.

But this ends up having a huge effect!









At an intuitive level the reason for the success of Lasso is that the $\|\cdot\|_1$-norm "promotes/prefers" solutions that are sparse while the $\|\cdot\|_2$-norm does not. Hence, Lasso/1-norm regularization are preferrable when we expect $\underline{x}$ to be sparse!

23.2 Intuition for Lasso

Solving a problem of the form

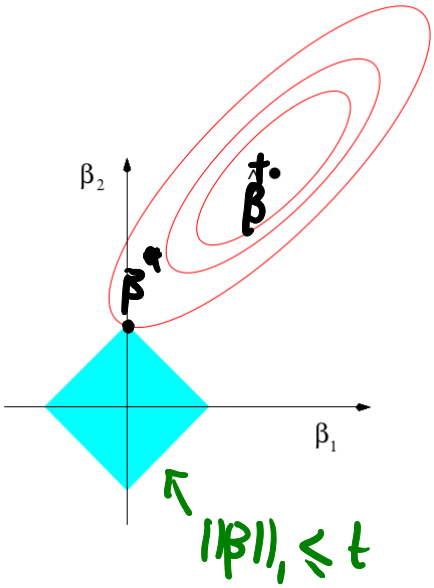
(I) $$\underline{\beta}^* = \underset{\underline{\beta}\in\mathbb{R}^N}{\operatorname{argmin}} \frac{1}{2}\|A\underline{\beta} - \underline{y}\|_2^2 + \lambda\|\underline{\beta}\|_p^p$$

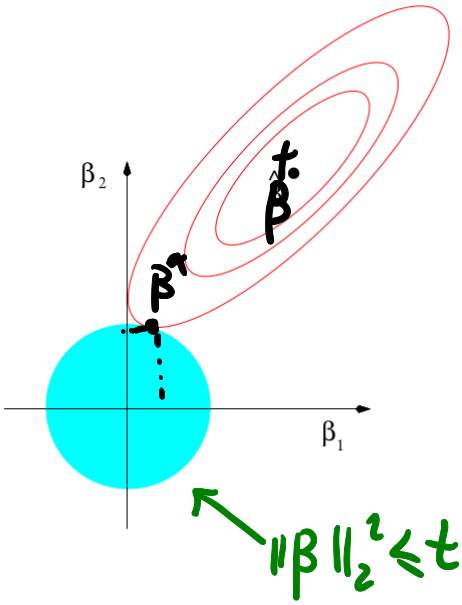
is equivalent to solving

(II) $$\underline{\beta}^* = \operatorname{argmin} \frac{1}{2}\|A\underline{\beta} - \underline{y}\|_2^2$$
$$\text{subject to } \|\underline{\beta}\|_p^p \leq t,$$

The value of t & λ are related to each other, indeed, (I) is the Lagrangian form of (II).

Here we can see the effect of the choice of the norm $\|\cdot\|_p$ in the regularization term.


β_2
β_1
$\hat{\beta}$
$\hat{\beta}^r$
$\|\beta\|_1 \leq t$


β_2
β_1
$\hat{\beta}$
$\hat{\beta}^r$
$\|\beta\|_2^2 \leq t$