Regularization, Sparsity and Energy Minimization

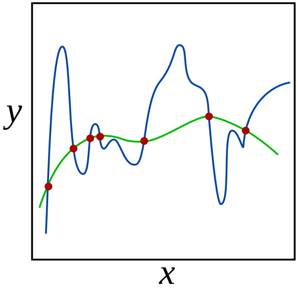
Source: Boyd and Vandenberghe, Convex Optimization (2014).

Regularization

Sparsity

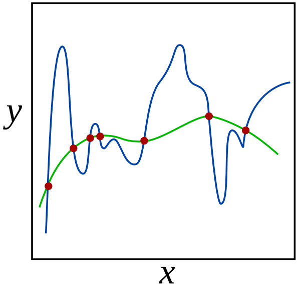
Energy minimization

Introduction



Suppose that we want to fit a regression model: $Ax \approx b$

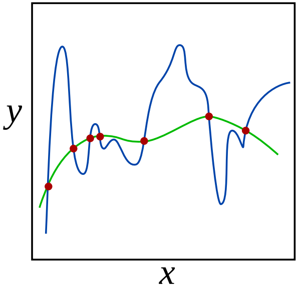
Introduction



Suppose that we want to fit a regression model: $Ax \approx b$

But model with large coefficients (large x) can lead to spurious predictions

Introduction



Suppose that we want to fit a regression model: $Ax \approx b$

But model with large coefficients (large x) can lead to spurious predictions

Our goal is to find a vector x that is small, and also makes the residual $\|Ax - b\|$ small

$$\text{minimize } (\|Ax - b\|, \|x\|)$$

How can we control both norms at the same time?

Regularization

We can formulate our problem as minimizing the weighted sum of the objectives:

$$\text{minimize } \|Ax - b\| + \gamma\|x\|$$

where $\gamma > 0$ is a parameter. Such method is called **regularization** parameter

- $\gamma = 0$: we try to find x that solves $Ax \approx b$

Regularization

We can formulate our problem as minimizing the weighted sum of the objectives:

$$\text{minimize } \|Ax - b\| + \gamma\|x\|$$

where $\gamma > 0$ is a parameter. Such method is called **regularization** parameter

- $\gamma = 0$: we try to find x that solves $Ax \approx b$
- γ large: we try to minimize $\|x\|$

Regularization

We can formulate our problem as minimizing the weighted sum of the objectives:

$$\text{minimize } \|Ax - b\| + \gamma\|x\|$$

where $\gamma > 0$ is a parameter. Such method is called **regularization** parameter

- $\gamma = 0$: we try to find x that solves $Ax \approx b$
- γ large: we try to minimize $\|x\|$

Alternatively, we can minimize the weighted sum of squared norms:

$$\text{minimize } \|Ax - b\|^2 + \gamma\|x\|^2$$

Tikhonov regularization

The most common form of regularization is with the Euclidean norms

$$\text{minimize } \|Ax - b\|_2^2 + \lambda \|x\|_2^2 \tag{1}$$

This is called **Tikhonov regularization problem**

Tikhonov regularization

The most common form of regularization is with the Euclidean norms

$$\text{minimize } \|Ax - b\|_2^2 + \lambda \|x\|_2^2 \tag{1}$$

This is called **Tikhonov regularization problem**

The regression problem $Ax \approx b$ that solves for x via (1) is called **ridge regression**.

Tikhonov regularization

The most common form of regularization is with the Euclidean norms

$$\text{minimize } \|Ax - b\|_2^2 + \lambda \|x\|_2^2 \tag{1}$$

This is called **Tikhonov regularization problem**

The regression problem $Ax \approx b$ that solves for x via (1) is called **ridge regression**.The solution of (1) is:

$$x = (A^T A + \lambda I_d)^{-1} A^T b$$

Since $A^T A + \lambda I_d$ is invertible for any $\lambda > 0$, the Tikhonov regularized least-squares solution requires no invertibility assumptions on the matrix $A^T A$

Regularization

Sparsity

Energy minimization

1-norm regularization

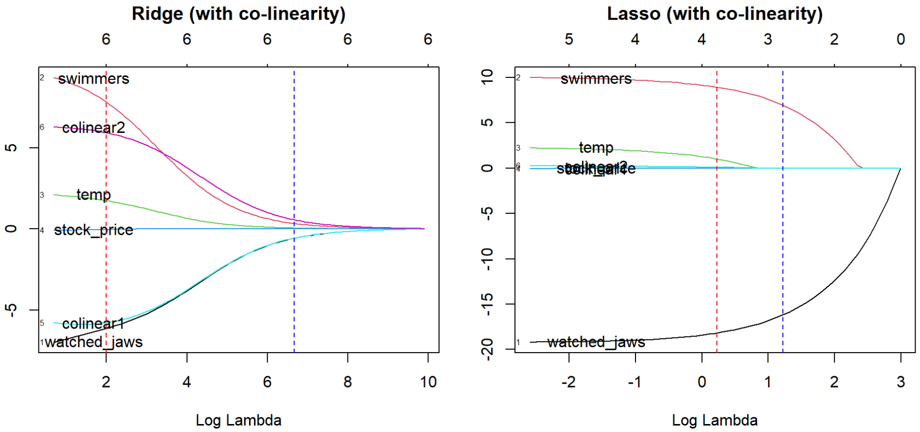
We can also regularized with the 1-norm:

$$\text{minimize } \|Ax - b\|_2 + \lambda \|x\|_1 \tag{2}$$

The regression problem $Ax \approx b$ that solves for x via (2) is called **LASSO**

Ridge vs LASSO

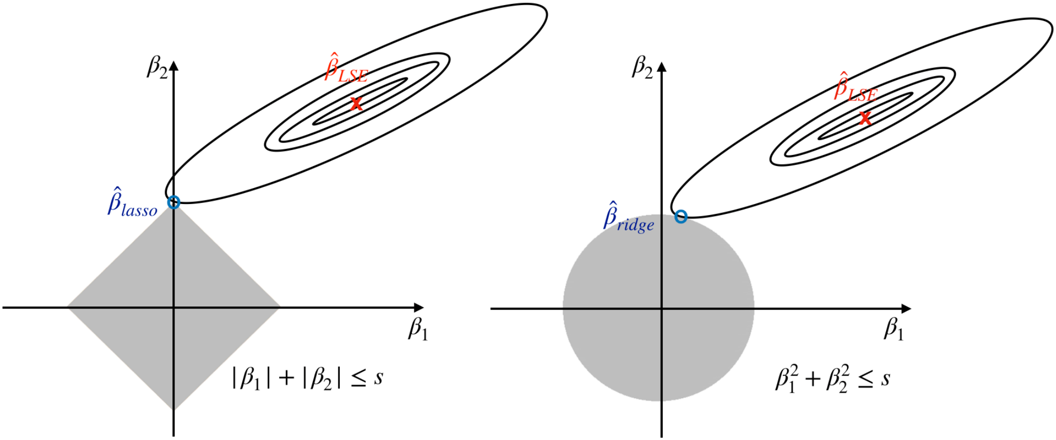
LASSO can help with **variable selection**



Plots of coefficients as functions of λ

Ridge vs LASSO

Graphical solutions to the LASSO (left) and Ridge (right) regression



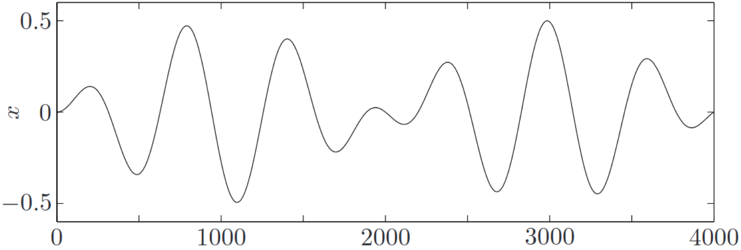
Feasible regions of 1- and 2-norm, and the level curve of $\|Ax - b\|^2$

Regularization

Sparsity

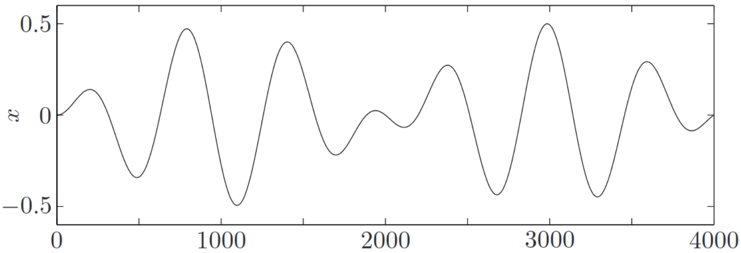
Energy minimization

Signal reconstruction



We consider signals in one dimension, e.g., audio signals, represented by a vector $x \in \mathbb{R}^n$

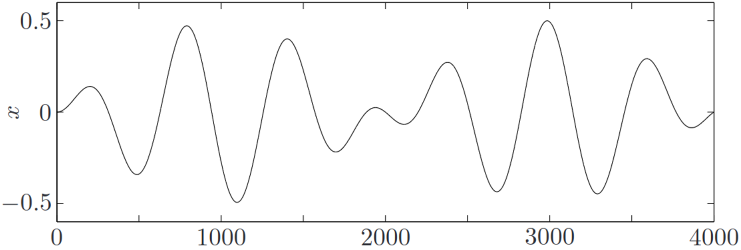
Signal reconstruction



We consider signals in one dimension, e.g., audio signals, represented by a vector $x \in \mathbb{R}^n$

The coefficients x_i correspond to the signal value at time i, evaluated (or sampled, in the language of signal processing)

Signal reconstruction



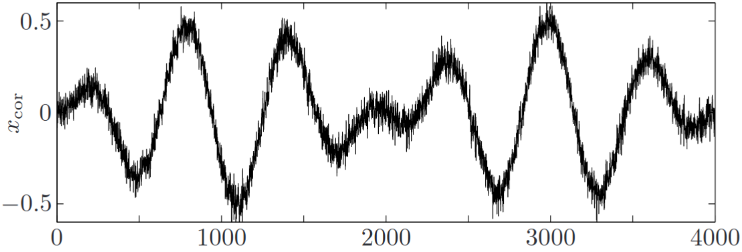
We consider signals in one dimension, e.g., audio signals, represented by a vector $x \in \mathbb{R}^n$

The coefficients x_i correspond to the signal value at time i, evaluated (or sampled, in the language of signal processing)

It is usually assumed that the signal does not vary too rapidly:

$$x_i \approx x_i + 1$$

Signal reconstruction



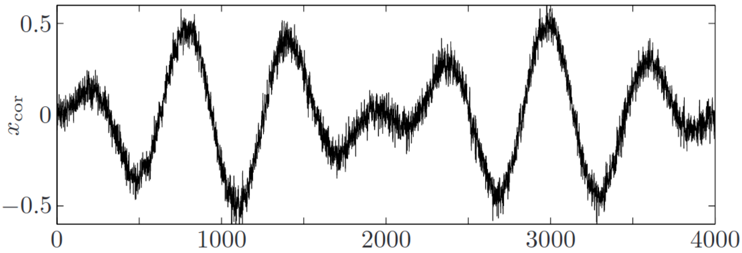
The signal x is corrupted by an additive noise v:

$$x_{\mathrm{cor}} = x + v$$

The goal is to form an estimate $\hat{x}$ of the original signal x, given the corrupted signal x_{cor}

This process is called **signal reconstruction**, **de-noising** or **smoothing**

Signal reconstruction

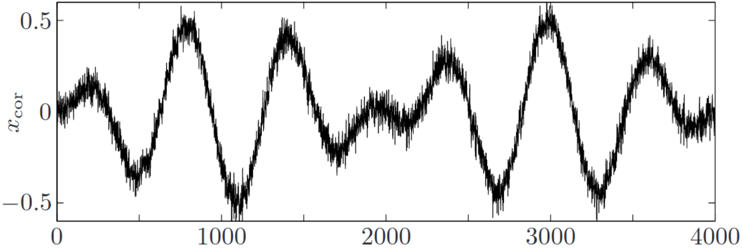


One simple formulation of the reconstruction problem is the following

$$\text{minimize}_{\hat{x}} \ \left(\|\hat{x} - x_{\mathrm{cor}}\|_2, \varphi(\hat{x})\right) \tag{3}$$

where the function $\varphi : \mathbb{R}^n \to \mathbb{R}$ is convex, and is called the **regularization function** or **smoothing objective**, which measures the roughness, or lack of smoothness, of the estimate $\hat{x}$

Signal reconstruction



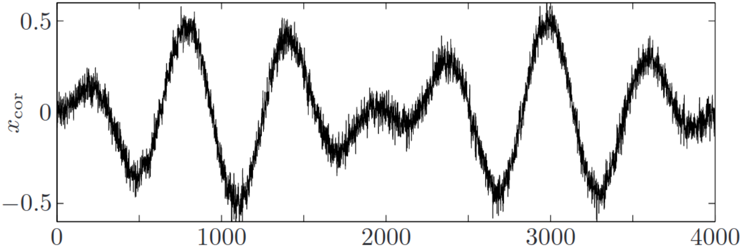
One simple formulation of the reconstruction problem is the following

$$\text{minimize}_{\hat{x}} \quad (\|\hat{x} - x_{\text{cor}}\|_2, \varphi(\hat{x})) \tag{3}$$

where the function $\varphi : \mathbb{R}^n \to \mathbb{R}$ is convex, and is called the **regularization function** or **smoothing objective**, which measures the roughness, or lack of smoothness, of the estimate $\hat{x}$

The reconstruction problem (3) seeks signals that are close to the corrupted signal (small $\|\hat{x} - x_{\text{cor}}\|_2$), and that are smooth (small $\varphi(\hat{x})$)

Signal reconstruction



One simple formulation of the reconstruction problem is the following

$$\mathrm{minimize}_{\hat{x}} \;\; (\|\hat{x} - x_{\mathrm{cor}}\|_2, \varphi(\hat{x})) \tag{3}$$

We can reformulate the signal reconstruction problem using regularization

$$\mathrm{minimize}_{\hat{x}} \; \|\hat{x} - x_{\mathrm{cor}}\|_2^2 + \lambda\varphi(\hat{x}) \tag{3'}$$

Measure of smoothness

The simplest reconstruction method uses the quadratic smoothing function

$$\phi_2(x) = \sum_{i=1}^{n-1} (x_{n+1} - x_n)^2$$

Measure of smoothness

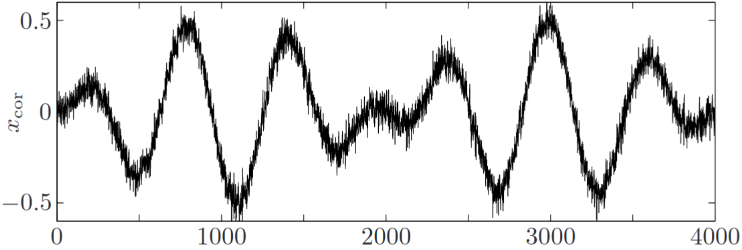
The simplest reconstruction method uses the quadratic smoothing function

$$\phi_2(x) = \sum_{i=1}^{n-1} (x_{n+1} - x_n)^2$$

This can be written as

$$\phi_2(x) = \left\| \begin{bmatrix} -1 & 1 & 0 & \dots & 0 & 0 & 0 \\ 0 & -1 & 1 & \dots & 0 & 0 & 0 \\ \vdots & \vdots & \vdots & & \vdots & \vdots & \vdots \\ 0 & 0 & 0 & \dots & -1 & 1 & 0 \\ 0 & 0 & 0 & \dots & 0 & -1 & 1 \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \\ \vdots \\ x_n \end{bmatrix} \right\|^2 = \|Dx\|_2^2$$

Quadratic smoothing

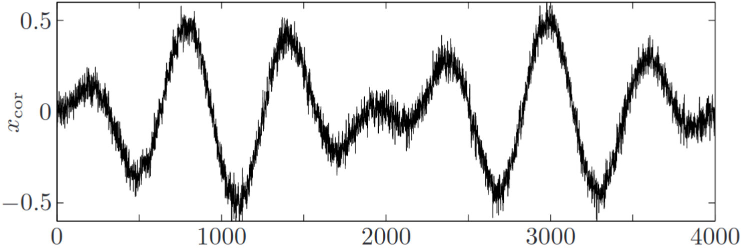


The optimization problem is

$$\text{minimize}_{\hat{x}} \ \|\hat{x} - x_{\mathrm{cor}}\|_2^2 + \lambda \|D\hat{x}\|_2^2$$

This problem is called **quadratic smoothing**

Quadratic smoothing



The optimization problem is

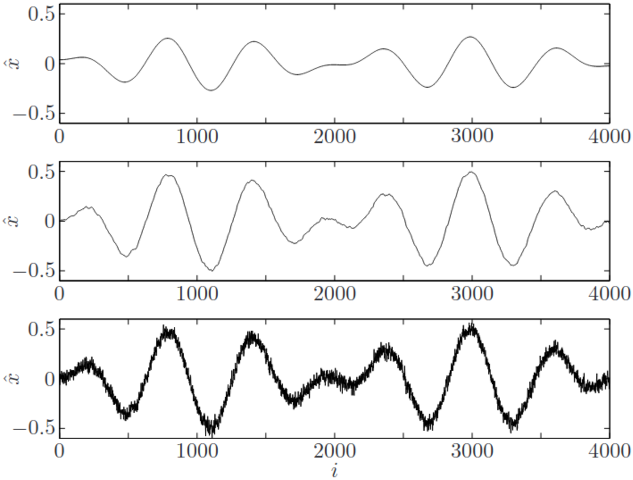
$$\text{minimize}_{\hat{x}} \ \|\hat{x} - x_{\text{cor}}\|_2^2 + \lambda \|D\hat{x}\|_2^2$$

This problem is called **quadratic smoothing**
The solution of this problem is:

$$\hat{x} = (1 + \lambda D^T D)^{-1} x_{\text{cor}}$$

Example

Result of reconstructing a signal $x \in \mathbb{R}^{4000}$



Top: $\|\hat{x} - x_{\mathrm{cor}}\| = 8$, Middle: $\|\hat{x} - x_{\mathrm{cor}}\| = 3$, Bottom: $\|\hat{x} - x_{\mathrm{cor}}\| = 1$

Total variation reconstruction

Simple quadratic smoothing works well as a reconstruction method when the original signal is very smooth

But any rapid variations in the original signal will be removed by quadratic smoothing

Total variation reconstruction

Simple quadratic smoothing works well as a reconstruction method when the original signal is very smooth

But any rapid variations in the original signal will be removed by quadratic smoothing

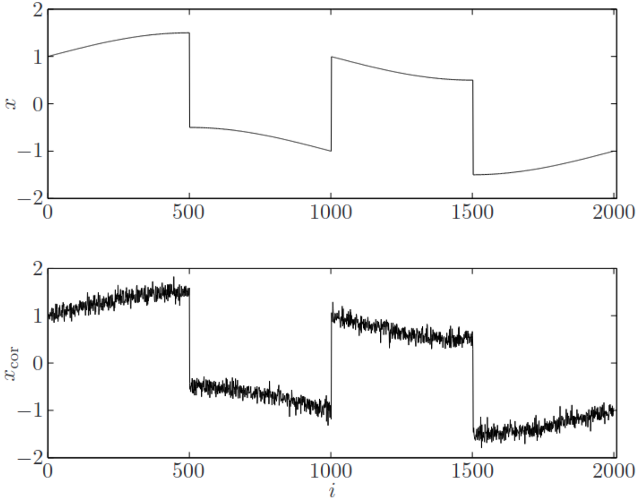
Alternatively, we can use the function:

$$\phi_{\mathrm{tv}}(x) = \sum_{i=1}^{n-1} |x_{n+1} - x_n| = \|Dx\|_1$$

which is called **total variation** of x

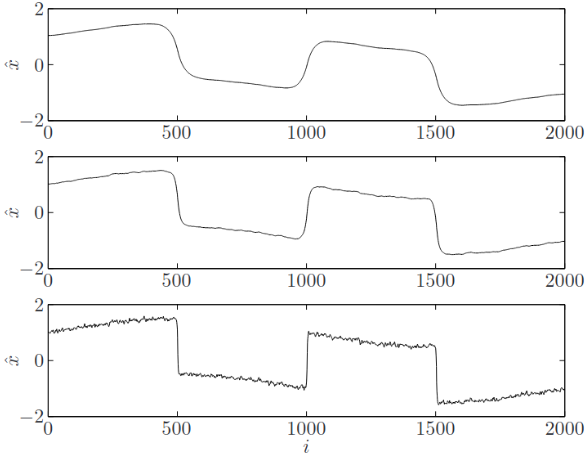
Example

Let's try to recover a signal $x \in \mathbb{R}^{2000}$



Example

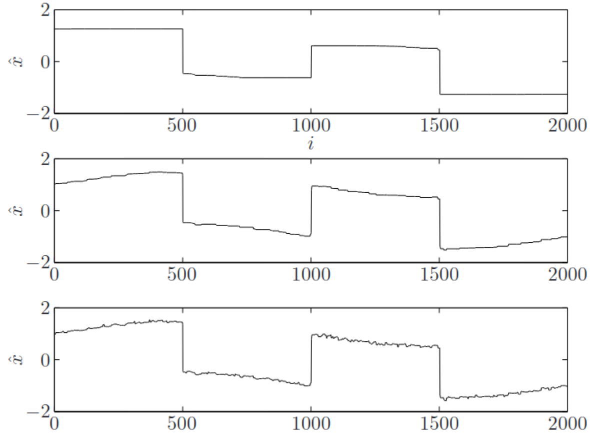
Reconstruction with quadratic smoothing



Top: $\|\hat{x} - x_{\text{cor}}\| = 10$, Middle: $\|\hat{x} - x_{\text{cor}}\| = 7$, Bottom: $\|\hat{x} - x_{\text{cor}}\| = 4$

Example

Reconstruction with TV reconstruction



Top: $\|Dx\|_1 = 5$, Middle: $\|Dx\|_1 = 8$, Bottom: $\|Dx\|_1 = 10$

Image denoising

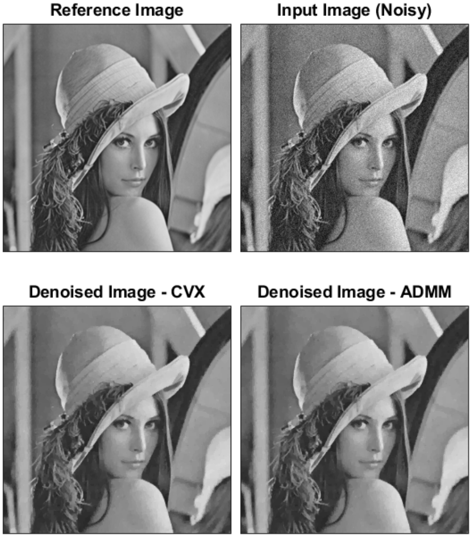
The TV reconstruction for images is:

$$\text{minimize}_{\hat{x}} \; \|\hat{x} - x_{\text{cor}}\|_2^2 + \lambda \text{TV}(\hat{x})$$

where

$$\text{TV}(x) = \sum_{i=1}^{m-1} \sum_{j=1}^{m} |x_{i,j} - x_{i+1,j}| + \sum_{i=1}^{m} \sum_{j=1}^{n-1} |x_{i,j} - x_{i,j+1}|$$

Example


Reference Image
Input Image (Noisy)
Denoised Image - CVX
Denoised Image - ADMM