# Least absolute deviations

Zé Vinícius

July 19, 2018

## 1 Least absolute deviations

$\ell_1$-norm optimization has been effectively used to estimate models on non-Gaussian noise, especially noise whose distributions possess heavy tails, such as Laplacian noise.

In this note, I followed the Majorization-Minimization (MM) framework in to derive the maximum a posteriori (MAP) affine model subject to Laplacian noise and with prior information that the model coefficients follow a joint Laplacian distribution.

This problem is classicly known as the Least Absolute Deviations and can be solved by a class of algorithms known as Iteratively Reweighted Least-Squares.

Mathematically, the cost function is given as follows

$$L(\boldsymbol{\beta}) = ||\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta}||_1 + \lambda||\boldsymbol{\beta}||_1, \tag{1}$$

in which $\boldsymbol{y}$ is the $n \times 1$ data vector, $\boldsymbol{X}$ is the $n \times m$ design matrix, $\beta$ is the $m \times 1$ vector of parameters, and $\lambda$ is a real-valued hyperparameter that controls the strength of the regularization. Our goal is to solve $\mathsf{argmin}_{\boldsymbol{\beta}}\, L(\boldsymbol{\beta})$.

Following the MM framework, we need to find an upper bound for $L$, let's call it $g$, i.e., $g(\boldsymbol{x}|\boldsymbol{z}) \geq L(\boldsymbol{x})$, equality holding at $\boldsymbol{x} = \boldsymbol{z}$. Additionally, we hope that $g$ will be "nicier" than $L$, meaning that we expect to be easier to find a point $\boldsymbol{z}_{t+1}$, such that $g(\boldsymbol{z}_{t+1}|\boldsymbol{z}_t) \leq g(\boldsymbol{z}_t|\boldsymbol{z}_t)$. The best scenario being that $\boldsymbol{z}_{t+1}$ is a minimizer of $g(\cdot|\boldsymbol{z}_t)$. If we can construct such $g$, we can generate a sequence of points that will lead to a feasible point of $L$ under some mild conditions[1].

[1] The interested reader is suggested to check out reference (Sun et. al. 2016) for details.

Using eq. (16) from (Sun et. al. 2016), it follows that one possible $g$ can be constructed as

$$g(\boldsymbol{\beta}_{k+1}|\boldsymbol{\beta}_k) = \frac{1}{2}\left\{||\boldsymbol{\Sigma}_k^{-1}(\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta}_{k+1})||_2^2 + ||\boldsymbol{\Sigma}_k^{-1}(\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta}_k)||_1 \right. \\ \left. + \lambda\left[\frac{||\boldsymbol{\beta}_{k+1}||_2^2}{||\boldsymbol{\beta}_k||_1} + ||\boldsymbol{\beta}_k||_1\right]\right\}, \qquad (2)$$

in which $\boldsymbol{\Sigma}_k = \mathrm{diag}(\boldsymbol{w})$, $w_i = |y_i - \boldsymbol{X}_i\boldsymbol{\beta}_k|$.

It can be noticed that

$$\mathsf{argmin}_{\boldsymbol{\beta}_{k+1}}\ g(\boldsymbol{\beta}_{k+1}|\boldsymbol{\beta}_k) = \mathsf{argmin}_{\boldsymbol{\beta}_{k+1}}\ ||\boldsymbol{\Sigma}_k^{-1}(\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta}_{k+1})||_2^2 + \lambda\frac{||\boldsymbol{\beta}_{k+1}||_2^2}{||\boldsymbol{\beta}_k||_1} \qquad (3)$$

which is the least squares cost function with a $\ell_2$ regularization component, which has an analytical minimum given as

$$\boldsymbol{\beta}_{k+1} = \left(\boldsymbol{X}^T\boldsymbol{\Sigma}_k^{-1}\boldsymbol{X} + \frac{\lambda}{||\boldsymbol{\beta}_k||_1}\boldsymbol{I}\right)^{-1}\boldsymbol{X}^T\boldsymbol{\Sigma}_k^{-1}\boldsymbol{y}. \qquad (4)$$

## References


1. Sun, Y *et. al.*, Majorization-Minimization Algorithms in Signal Processing, Communications, and Machine Learning. *IEEE Transactions on Signal Processing*, 2016.