Long Regression Applied to Short Term Traffic Prediction discussion_N250410

April 16, 2025

1 Introduction

In this document I am following the thread opened by `dicussion_N20241023.pdf` trying to show that for this specific application I would expect to have greater performance when the model abides by some specific constraints. These constraints amount to the intuition that higher traffic or a higher number of offenders during the sensing time-window is likely to be followed by a higher number of offenders during the predicting time-window. The conclusion of the previous document suggests that this increase in performance was not to be seen because of a lack of flexibility in the chosen models. I am here considering a much more flexible model, its representation is simple because feature values are here integers and actually with not too high values. Now the proposed model could easily be transformed into first-ordre splines when the breaking points are regularly padded.

[1] promotes the use of statistical tests in regression. I think that this addresses the overfitting issue. In the considered application, I do not see that we have clear information regarding the model itself or regarding the probability distributions of the errors. Instead I am using a validation dataset to address this overfitting issue, the dataset being large enough.

2 Problem Statement

The model proposed is linear with respect to the learned parameters. It is one-dimensional and it takes advantage of the fact that the feature D is an integer. $\widehat{Y}$ denotes the predicted value of c_n when the sensor captures a density of D at timestamp $t_n = nT_e$.

$$\widehat{Y} = \sum_{k=0}^{+\infty} a_k \mathbf{1}\left(D = k\right) + a_\infty \tag{1}$$

The parameters a_k are stacked in a column vector, whose size depends on the training set (i.e. its length is $1 + K = 1 + \max_n d_n$).

$$A = \left[a_0 \, \dots \, a_{K-1} \, a_\infty\right]^T \tag{2}$$

The input values (read from the feature values) are a $N \times (K+1)$-matrix

$$\mathbf{X} = \begin{bmatrix} \mathbf{1}(d_0 = 0) & \mathbf{1}(d_0 = 1) & \dots & \mathbf{1}(d_0 = K-1) & 1 \\ \mathbf{1}(d_1 = 0) & \mathbf{1}(d_1 = 1) & \dots & \mathbf{1}(d_1 = K-1) & 1 \\ & & \vdots & & \\ \mathbf{1}(d_{N-1} = 0) & \mathbf{1}(d_{N-1} = 1) & \dots & \mathbf{1}(d_{N-1} = K-1) & 1 \end{bmatrix} \tag{3}$$

where N is the number of samples in the dataset considered. The output values (i.e. predicted values claimed by the training set) are

$$\mathbf{Y} = \left[c_0 \dots c_{N-1}\right]^T \tag{4}$$

The model-predicted values are

$$\widehat{\mathbf{Y}} = \mathbf{X}A = \left[\widehat{c}_0 \dots \widehat{c}_{N-1}\right]^T \tag{5}$$

The objective is to minimize on the testing set

$$\sum_{n=0}^{N-1} |\widehat{y}_n - y_n| \tag{6}$$

In the following sections, I suggest some modifications too improve the objective.

3 Description of the Proposed Models

In the implementation `method` is an integer indicating the number of the model.

1. OLS with Numerical Regularization
2. OLS with Ridge Regression
3. OLS constrained to non-negative predictions
4. OLS constrained to non-decreasing D-dependency
5. LAD with Numerical Regularization
6. LAD with Ridge Regression
7. LAD constrained to non-decreasing D-dependency

3.1 Model 1: OLS with Numerical Regularization

The Ordinary Least Square (OLS) technique finds A by minimizing

$$\sum_{n=0}^{N-1} (\widehat{y}_n - y_n)^2 = (\mathbf{Y} - \mathbf{X}A)^T(\mathbf{Y} - \mathbf{X}A) \tag{7}$$

Equation (1) considers an infinite number of parameters, some could be used in the testing set without being considered during training. To solve this numerical issue, it is needed to set values to such parameters and thanks to a_∞, it makes sense to zero these parameters. An other way of expressing this idea is too search A by minimizing

$$\sum_{n=0}^{N-1} (\widehat{y}_n - y_n)^2 + \lambda \left(\sum_{k=0}^{+\infty} a_k^2 + a_\infty^2 \right) = (\mathbf{Y} - \mathbf{X}A)^T(\mathbf{Y} - \mathbf{X}A) + \lambda A^T A \tag{8}$$

The size of A depends on the training set.

$$K = \max_n d_n \text{ and } A \in \mathbb{R}^{K+1} \tag{9}$$

During prediction, the size of A is increased with zeros just before its last component a_∞, it thereby adapts to the possible higher values of d_n in the testing set.

In the implementation, I've chosen a small value for $\lambda = 10^{-3}$.

3.2 Model 2: OLS with Ridge Regression

This model is similar to model 1, except that now the value λ is chosen so as to best perform on a validation dataset. This changes the model, I call it a ridge regression because it gives a prior on the values of a_k, here it is zero. And the importance of the prior is based on performance on the validation set.

3.3 Model 3: OLS constrained to non-negative predictions

This model is similar to model 1, except that non-negative inputs are constrained to yield non-negative outputs.

$$\left[D \geq 0 \Rightarrow \widehat{Y} \geq 0\right] \quad \Leftrightarrow \quad [\forall k,\ a_k + a_\infty \geq 0 \text{ and } a_\infty \geq 0] \quad \Leftrightarrow \quad M_3 A \geq 0 \tag{10}$$

where 0 is a null column vector and M_3 is defined as a $(K+1)\times(K+1)$-matrix

$$M_3 = \begin{bmatrix} 1 & & & & & 1 \\ & 1 & & 0 & & 1 \\ & & \ddots & & & 1 \\ & 0 & & 1 & & 1 \\ & & & & 1 & 1 \\ & & & & & 1 \end{bmatrix} \tag{11}$$

This model has not been implemented, as both model 1 and model 2 when trained on the dataset, they both have this desired property. Equation 10 is tested in `experiment.m` at `number=5`.

3.4 Model 4: OLS constrained to non-decreasing D-dependency

Model 4 is similar to model 3, it adds a new constraint, which is that higher values of D should yield higher values of C.

$$\left[D \geq 0 \Rightarrow \widehat{Y} \geq 0 \text{ and } D' \geq D \Rightarrow \widehat{Y}' \geq \widehat{Y}\right] \quad \Leftrightarrow \quad [-a_\infty \leq a_0 \leq a_1 \leq \ldots \leq a_k \leq \ldots a_{K-1} \leq 0 \leq a_\infty] \quad \Leftrightarrow \quad M_4 A \geq 0 \tag{12}$$

where M_4 is a $(K+2)\times(K+1)$-matrix defined as

$$M_4 = \begin{bmatrix} 1 & 0 & 0 & 0 & & 0 & 1 \\ -1 & 1 & & & & & 0 \\ & -1 & 1 & & 0 & & \\ & & \ddots & \ddots & & & \\ & & & \ddots & \ddots & & \\ & 0 & & & -1 & 1 & \\ & & & & & -1 & 0 \\ & & & & & & 1 \end{bmatrix} \tag{13}$$

Equation 10 is tested in `experiment.m` at `number=5`.

3.5 LAD with Numerical Regularization

The Least Absolute Deviation (LAD) technique finds A by minimizing

$$\sum_{n=0}^{N-1} |\widehat{y}_n - y_n| = \|\mathbf{Y} - \mathbf{X}A\|_1 \tag{14}$$

where $\|\ \|_1$ is the l_1-norm (i.e. sum of the absolute values of all components).

As in model 1, we use a numerical regularization and minimize

$$\|\mathbf{Y} - \mathbf{X}A\|_1 + \lambda\|A\|_1 \tag{15}$$

3.6 LAD with Ridge Regression

As in model 2, we find λ for which A that minimizes equation (15) on the training dataset yields actually the smallest deviation on the validation set (i.e. $.\|Y - AX\|_1$).

3.7 LAD constrained to non-decreasing D-dependency

As in model 4, the optimization is

$$\text{minimize } \|\mathbf{Y} - \mathbf{X}A\|_1 + \lambda\|A\|_1 \text{ subjected to } M_4 A \geq 0 \tag{16}$$

4 Results and Discussion

4.1 First Experiment

I followed the process described in `discussionN_20240422.pdf` to extract data sets from the csv file containing speeds, speed limits and timestamps. The training set include the first 15 days. The validation set include the following 15 days and the testing set include the following 15 days. Samples are not distinguished with respect to speed limits. The sampling period $(\frac{1}{f_e})$ is 5 minutes that is an increased duration that increases the values of D collected. Only one feature is considered, that of D. $\widehat{Y}$ is the predicted value of c_n at the corresponding time stamp. This is implemented in experiments number 1 and 2.

Table 1

Model_Num	Dev_Tr	Dev_Va	Dev_Te	λ	`exp`
1	8.20	8.15	10.47	0.001	4
2	8.20	8.14	10.54	0.002	4
4	10.67	9.65	12.70	0.001	7
5	8.03	7.90	10.27	0.001	8
6	8.00	7.82	10.38	0.002	9
7	10.48	8.61	13.81	0.001	10

Note the last column indicates where in the file `experiment.m`, this is computed.

4.1.1 Comparing Numerical Regularization and Ridge Regression

These experiments are conclusive that ridge regression is not doing better than using only the numerical regularization.

The third column of model 2 is below that of model 1 and the third column of model 6 is below that of model 5. This shows that on the validation set the choice of λ has reduced the deviation. This is of course no surprise as precisely the ridge regression here consists in choosing λ in such a way.

The fourth column of model 2 is higher that of model 1 and the fourth column of model 6 is higher that of model 5. This shows that this choice of λ proves to be less performing on the testing set. This means that ridge regression is here not a good idea, actually it means possibly giving a higher weight to the prior claiming that a_k should be zero that is claiming that the output is constant regardless of the value of X.

The low values of λ on the fifth column in models 2 and 6 show that giving a higher on this zero-priori is not efficient in increasing the performance on the validation set.

4.1.2 Comparing inducing models not constrained with models constrained to non-decreasing D-dependency

The fourth column of models 4 and 7 are greater than that of models 1 and 5. It shows that expectations have not been met, such constraints do not increase the performance.

4.1.3 Comparing inducing using the $L1$-norm with that of using the $L2$-norm

Because in our application, we are interested in the $L1$-norm, it makes sense to expect better performance when using the $L1$-norm. This expectation is supported when I compare for the validation dataset (column 3) for models 1 and 4 with 5 and 7. This expectation is also supported when I compare for that testing dataset for models 1 and 2 with 5 and 6. However this is not true considering model 4 and model 7 for the testing set, and the difference is important. I would say that this experiment is not conclusive on that matter.

Figure 1 shows on the D-dependency for models 1 and 4. For model 1, we see that the components $a_k + a_\infty$ are non-negative but not non-decreasing. These components predict y_n when $x_n = k$. For model 4, these components are non-decreasing which requested when solving the minimization problem. Those of models 2, 5 and 6 look like model 1 and those of model 7 look like model 4.

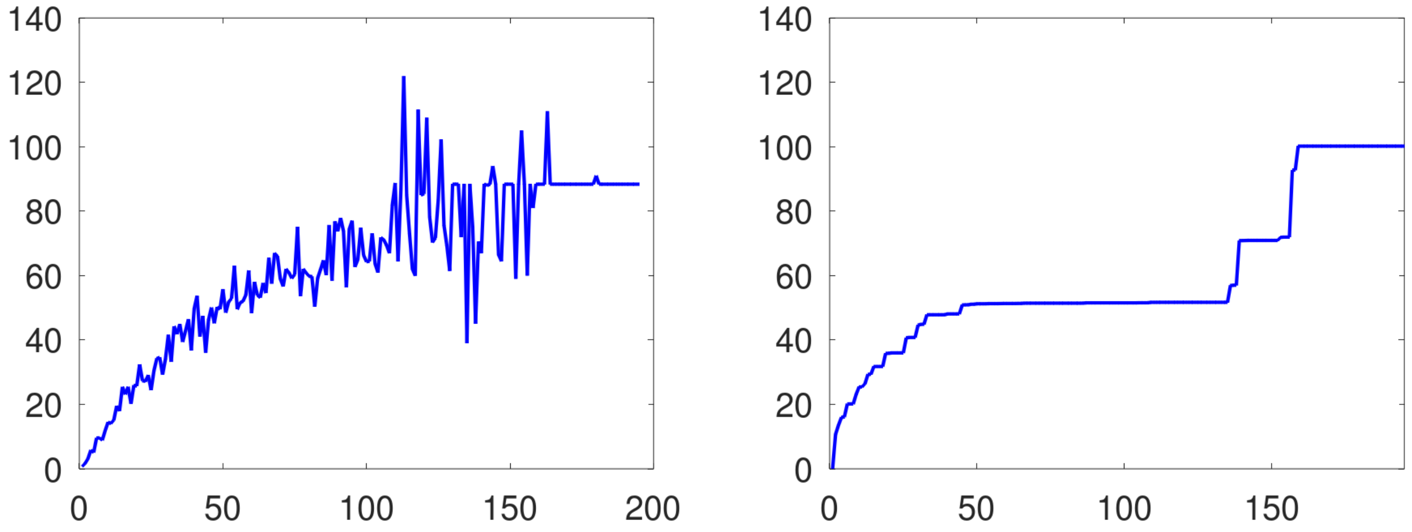

Figure 1: Components $a_k + a_\infty$ as a function of k the value of D for model 1 on the left and model 4 on the right.

4.2 Second Experiment

I've chosen to make this second experiment to explore a possible explanation for this non-improving performances when D-dependency is constrained. Some specific values of D could be clues for machine learning that there is something special happening and in that situation the expectation of the number of offenders could be modified. This second experiment considers only samples for which the speed limit is 80km/h. Table 2 considers only models 1,4,5 and 7, so the validation set is here a second testing set.

Table 2

Model_Num	Dev_Tr	Dev_Va	Dev_Te	exp
1	0.732	0.637	0.539	13
4	3.365	2.059	1.25	14
5	0.572	0.422	0.304	15
7	3.450	2.109	1.28	17

The previous document `discussion_N20240422.pdf` had already shown that the number of offenders depends higly on the speed limit, there are much less when it is 80km/h. So the lower deviations do not show a higher accuracy.

This time both on the validation set and on the testing set, using $L1$-norm improves significantly the performance.

When considering the $L2$-norm, the constraint does not appear as improving. However when considering the $L1$-norm, a tiny improvement can be seen. At this point, I don't think we should be too conclusive.

4.3 Third Experiments

Based on the previous documents I have written, I am expecting that the density is sufficient to make predictions. However for a publication, it is maybe easier not to have to convince the reader with that information. So I am considering a new dataset where I am replacing D with DR, DR being the product of the density and the ratio that is the number of offenders, so it is also an integer. I am considering the same models and the same constraint which now means higher number of offenders are likely to be followed by a higher number of offenders.

Table 3

Model_Num	Dev_Tr	Dev_Va	Dev_Te	λ	exp
1	0.935	0.642	0.576	0.001	18
4	1.021	0.717	0.657	0.001	19
5	0.708	0.339	0.242	0.001	20
7	0.738	0.338	0.241	0.001	21

The same conclusions can be drawn.

This time both on the validation set and on the testing set, using $L1$-norm improves significantly the performance.

When considering the $L2$-norm, the constraint does not appear as improving. However when considering the $L1$-norm, a tiny improvement can be seen. At this point, I don't think we should be too conclusive, but the reader

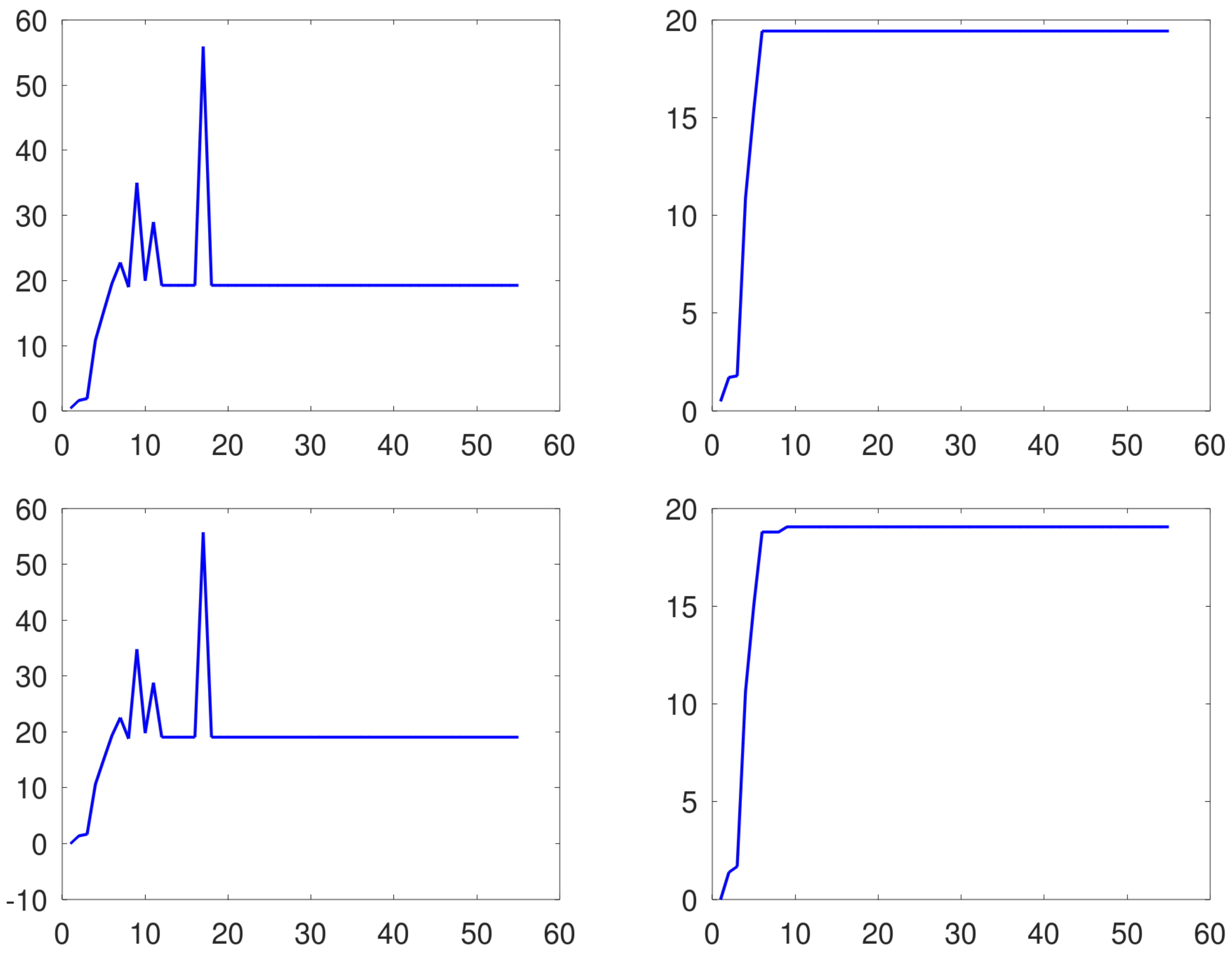

Figure 2: Components $a_k + a_\infty$ as a function of k the value of D for model 1, 4, 5 and 7 when we look at the figures using the classical reading order.

may think it is conclusive.

Figure 2 show the D-dependency for models 1, 4, 5 and 7. The reduced values of the horizontal axis indicates that the datasets 2 and 3 have much less samples. This is because I do not consider all samples for which the speed limits is being changed between the sensing time-window and the predicting time window.

A Proof of equation (10)

A.1 $\left[D \geq 0 \Rightarrow \widehat{Y} \geq 0\right] \quad \Rightarrow \quad \forall k,\ a_k + a_\infty \geq 0$ and $a_\infty \geq 0$

- Let $k \in \{0 \ldots K-1\}$ and $D = [k]$ a scalar equal to k. Then $\widehat{Y} = [a_k + a_\infty]$ is also a non-negative scalar and hence $a_k + a_\infty \geq 0$.

- Let D have a greater value than the size of A, then A has to be increased with zeros padded just above its last component. X is then defined with one row composed of two non-zero components, the first one corresponding to a null component of A. Then $0 \leq \widehat{Y} = a_\infty$.

A.2 $[\forall k,\ a_k + a_\infty \geq 0] \quad \Rightarrow \quad D \geq 0 \Rightarrow \widehat{Y} \geq 0$

Let $D = k \geq 0$, an integer. We get $\widehat{y} = a_k + a_\infty \geq 0$.

B Proof of equation (12)

B.1 Proving the implication

- The previous proof has already shown that $a_k + a_\infty \geq 0$ and $a_\infty \geq 0$. Applying this claim to $k = 0$ yields $a_0 = a_0 + a_\infty - a_\infty \geq -a_\infty$ and also $a_\infty \geq 0$.
- Let $k \in \{0 \dots K-2\}$, and $D = k$ and $D' = k+1$, this shows $\widehat{Y} = a_k + a_\infty \leq \widehat{Y}' = a_{k+1} + a_\infty$ and hence $a_k \leq a_{k+1}$.
- Let $D = K-1$ and $D' = K$ then $\widehat{Y} = a_{K-1} + a_\infty \leq \widehat{Y}' = a_\infty$. So $a_{K-1} \leq 0$.

B.2 Proving the converse

- Because the right statement includes that $a_k + a_\infty \geq 0$ and $a_\infty \geq 0$, the previous proof has already shown that $D \geq 0 \Rightarrow \widehat{Y} \geq 0$.
- Let $k' = D' \geq D = k$. Then $\widehat{Y}' - a_\infty = a_{k'} \geq \ldots \geq a_k = \widehat{Y} - a_\infty$, so we get $\widehat{Y}' \geq \widehat{Y}$.

References

[1] Norman R. Draper and Harry Smith. *Applied Regression Analysis.* John Wiley & Sons, Inc., 1998.