SEMI-PARAMETRIC ESTIMATION OF SHIFTS AND APPLICATION TO ROAD TRAFFIC FORECASTING

F. GAMBOA*, J-M. LOUBES** AND E. MAZA*

Description of the institution : MIST-R Project

The framework of the following work is a national research project MIST-R (http ://www.math.u-psud.fr/ ACI.html) whose purpose is roadtrafficking travel time forecasting. This project, granted by the french national research center, with head director Jean-Michel Loubes (CNRS and University Montpellier 2) and local directors Mehdi Danech-Pajouh (INRETS) Fabrice Gamboa (university Toulouse 3) and Michele Sebag (CNRS and university Paris Sud) is composed of 4 research teams made of statisticians, informaticians, and professionals of road traffic, involving 16 people. They aim first at understanding the different phenomena appearing in road traffic, mixing both periodic and random effects. Then building models of roadtrafficking enables to forecast time travel at short range. Hence, our method relies on a probabilistic study of real data, which enables to predict with more accuracy than standard methods, which are based on pure modeling without taking into account the probabilistic framework. The strength of this project is to gather the knowledge of both high level researchers and ingeneers. Starting from a pratical issue, we use statistical theory to improve the results, keeping in mind the pratical goal of our project, which is the main purpose of applied mathematics.

1. Methodology for time travel forecasting

The purpose of our study is short term travel time forecasting (up to 4 hours) on the Parisian highway network. Here, road traffic is described by the velocities of the vehicles. So, we aim at estimating road traffic velocities at all the points of the observation grid. The data we use are the following : the Parisian road network infrastructure is equipped with measurement stations, located approximately every 500 meters on the road network. These stations measure road traffic evolution by calculating, for every fixed period and all the day, the mean velocity of the vehicles flow. Using such large amount of data, our methodology is the following.

- Data clustering : this step is based on a classification method. Indeed, once a distance adapted to speed curves is properly chosen, we gather together similar functions in a small number of cluster, where the meaning of similar is given by the distance. Thus we obtain subgroups representing an archetype of the evolution of daily vehicle velocities.
- Structural estimation : here, our goal is to extract the feature from the whole curves contained in the group. This is the point we want to stress in this paper. In this case, all the curves seem to be deduced one from another by a single shift, see for example Figure 1(a). We observe speed curves with a same traffic jam or speed reduction but with different starting times of the phenomenon for each curve. So, for each cluster we need

to find the best representative profile, i.e the structural estimator, which can be done by estimating the shift parameters. Indeed, because of these shifts the mean curve is not representative enough and does not convey the true information. In our example, see the mean curve with dotted line on Figure 1(d).
- Real time classification for forecasting : this final step consists in comparing real time data to the different archetypes selected in the previous step and using the closest model to forecast the close evolution of the velocities of the vehicles.

Combining these different strategies leads to a complete methodology for short time travel time forecasting. This work is still under process but the first results we obtained are very encouraging, improving the standard forecasting procedures. See for example Figure **??**.

2. Structural estimation with shifts estimation

Shift estimation - the framework:

More and more often, the outcome of an experiment is not a random variable, but a noisy sample of curves. Examples of such data might be growth curves, longitudinal data in medicine, speech signals, traffic data or expenditure curves for some goods in the econometric domain. The individuals usually experience similar events which are explained by a pattern but the starting time of the event occurs sooner or later. Hence, computing a classical representative curve for this sample severely distorts the analysis of the data. Indeed, the average curve (usually the mean or the median) oversmooths the studied phenomenon and distorts the reality. Hence the solution we propose is a two step strategy
- First estimate deformation between the curves,
- Then apply the inverse of the deformations to align the data and be able to estimate the feature of the observed phenomenum.

Since many years, some work has been done to find a representative of a large sample of close enough functions.

Matematical modeling:

A sample of functions may be modeled as follows. We observe, for each curve j, $j = 1, \dots, J$, at consecutive times t_{ij}, $i = 1, \dots, n_j$, noisy data y_{ij}. We assume that there exists functions $j = 1, \dots, J$ (f_j) such that the measures y_{ij}, $j = 1, \dots, J$, $i = 1, \dots, n$ are

$$Y_{ij} = f_j(t_{ij}) + \varepsilon_{ij},\ j = 1, \dots, J,\ i = 1, \dots, n,$$

where $(\varepsilon_{ij})_{j=1,\dots,J,\, i=1,\dots,n}$ are i.i.d. random variables, representing the observation noise and n stands for the number of observations for each curve. So, we assume that the functions (f_j) are close from each other in the sense that there exists an unknown **archetype** f and unknown **warping functions** $(h_j)_{j=1,\dots,J}$ such that

$$\forall j \in \{1, \dots, J\},\ \forall t \in [0, T],\ f_j(t) = f \circ h_j(t).$$

In our study, we assume that the observations can be written as a regression model, where we observe for each individual $j \in \{1, \dots, J\}$ an unknown function f translated by parameters $\theta_j^* \in \mathbb{R}$, which are to be estimated in order to align the shifted curves and then build an estimator of the function f. Hence we can write

$$Y_{ij} = f\left(t_{ij} - \theta_j^*\right) + \varepsilon_{ij}.$$

Methodology:

The difficulty of the work is that the estimation of the shift parameters can not rely on the pattern f which is unknown, but these quantities are deeply linked. That is the reason why we will use an M-estimator built on the Fourier series of the data given by the coefficients d_{jl}, $j = 1, \ldots, J$, $l = -(n-1)/2, \ldots, (n-1)/2$. We have consider the estimation problem in the frequency domain. Under identifiability assumptions, we provide a consistent method to estimate $\left(\theta_j^*\right)_{j=1,\ldots,J}$ when f is unknown. The estimator we construct relies on **semiparametric estimation theory** and is defined as the solution of the following minimization problem

$$M_n((\hat{\theta}_j)_{j=1,\ldots,J}) = \min_{\theta} (\ldots) .$$

We prove that these estimators are close to the real shift and that fluctuations of our estimates are asymptotically Gaussian. We alos provide an efficient algorithm to compute the sifts and finally build the structural estimator of the archetype.

Numerical results for roadtrafficking data:

For our traffic data example, the results are the followings. Figure 1(a) represent a particular cluster on a particular counting station. Figure 1(b) shows estimated shifts. Shifted curves are plotted on figure 1(c). So, in this homogeneous cluster, where only a shift phenomenon appears, difference are obvious between the mean curves in figure 1(d) of shifted curves (solid line) and of primary curves (dotted line). Hence, the shift estimated mean is clearly more representative of the individual behaviour.

** CNRS - Laboratoire de Mathématiques, Equipe de Probabilités, Statistique et Modélisation, UMR 8628 - bâtiment 425, Université Paris Sud, 91405 Orsay cedex

* Laboratoire de Statistique et Probabilités, UMR C5583 - Université Paul Sabatier, 118, route de Narbonne, F-31062 Toulouse cedex 4

E-mail address: Fabrice.Gamboa@math.ups-tlse.fr
E-mail address: Jean-Michel.Loubes@math.u-psud.fr
E-mail address: Elie.Maza@math.ups-tlse.fr

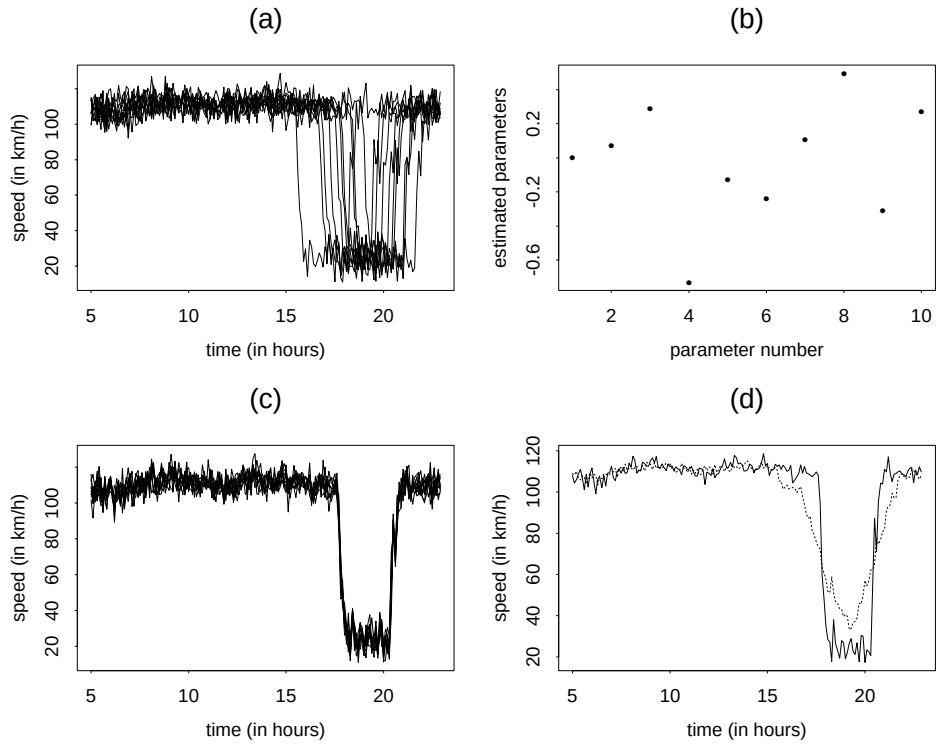

FIG. 1. On Figure (a), we see the observed data in an homogeneous claster. Figure (b) shows estimated shifts and shifted curves are plotted on Figure 1(c). Figure 1(d) shows the mean of shifted curves (solid line) and of primary curves (dotted line).