Independent Component Analysis of Milan Mobile Network Data

Piercesare Secchi, Simone Vantini, Paolo Zanini

Abstract We analyze spatially dependent functional data observed on a dense lattice covering the metropolitan area of Milan; in each site of the lattice the average traffic carried by a major mobile phone provider is measured over periods of 15 minutes along a window of 14 days. These complex data are highly informative about population mobility and residence which is relevant for urban planning and the efficient design of service networks. The analysis is carried on by a suitable extension of the Independent Component Analysis for spatially dependent functional data.

Key words: Independent Component Analysis, Spatio-temporal Data, Erlang, Urban Mobility

1 Introduction and data description

Aim of this work is to understand if mobile traffic data can provide useful information for investigating population mobility in highly-populated areas. In particular we focus on the metropolitan area of Milan and center the analysis on Telecom Italia data. For each pixel of the covered area we observe the Erlang every 15 minutes for 14 days. The Erlang is a dimensionless unit calculated as the sum of the length of every call in a given time interval divided by the length of the interval (i.e., 15 minutes). For each pixel and for each quarter of an hour, this measure represents the average number of mobile phones simultaneously calling through the network, that, as a first approximation, can be considered proportional to the number of active people in that area at that time. The idea is to extract from this kind of data useful information for Green Move (GM), an interdisciplinary research project financed by Regione Lombardia involving different research groups at the Politecnico di Milano and regarding the development of a vehicle sharing system based on the concept of

Piercesare Secchi, Simone Vantini, Paolo Zanini
MOX - Dep. of Mathematics, Politecnico di Milano, p.za Leonardo da Vinci 32, 20133 Milan
e-mail: piercesare.secchi@polimi.it, simone.vantini@polimi.it, paolo.zanini@mail.polimi.it

"little, electric and shared vehicles". Our contribution to the project is to provide information about traffic flows and to find optimal places where to locate the docking stations of the system. The analyzed data describe a phenomenon in a 2D-space at different instants of time. This may be represented by a surface varying along time. To explore these spatial data we use an extension of Independent Component Analysis (ICA) - traditionally used to analyze time signals - in order to decompose the observed signal as a time-varying linear combination of a reduced number of time-invariant basis surfaces. The idea is to associate basis elements to population mobility and residence in order to provide relevant information for urban planning and for the efficient design of service networks.

The covered area is a rectangular lattice of 44 x 48 pixels (232m x 309m each) over the city of Milan. Temporal patterns associated to each pixel have been smoothed using a Fourier basis assuming a seven days periodicity. Hence we analyze a data matrix $X \in \mathbb{R}^{p \text{ x } n}$ where $p = 672$ is the number of instants (every 15 minutes for seven days) and $n = 2112$ is the number of pixels in the region. The purpose of the analysis is to represent X as the product of two matrices $C \in \mathbb{R}^{p \text{ x } k}$ and $S \in \mathbb{R}^{k \text{ x } n}$, where each row of S represents the evaluation in the n pixels of the corresponding basis surface and the element c_{ij} indicates the contribution of the j-th surface at time i.

2 Independent Component Analysis

Independent Component Analysis, also known as blind source separation, has received an increasing amount of attention from the signal-processing research community ([1]). We may define ICA by means of a statistical latent variables model. Consider p random variables $X_1, ..., X_p$ and assume that they are linear mixtures of p latent independent components:

$$X_i = c_{i1} S_1 + ... + c_{ip} S_p \tag{1}$$

for $i = 1, ..., p$. In matrix form: $\mathbf{X} = C\mathbf{S}$, where $\mathbf{X}$ and $\mathbf{S}$ are random vectors in $\mathbb{R}^p$ and C is the $\mathbb{R}^{p \text{ x } p}$ matrix of the coefficients.

The basic ICA assumption is that the components S_j are statistically independent. This represent a fine-tuning of the well-known Principal Component Analysis idea, that aims at decomposing the random variables $X_1, ..., X_p$ into linear mixtures of uncorrelated components. Indeed, PCA and ICA are meant to provide similar decompositions if data are normally distributed. The coefficients c_{ij} may be estimated by minimizing some quantities related to statistical independence, like entropy or mutual information. Entropy is a basic concept of information theory, usually interpreted as the degree of information carried by the distribution of a random element. The entropy H of a random vector $\mathbf{Y}$ with density f is defined as $H = -\int f(\mathbf{y}) \log f(\mathbf{y}) d\mathbf{y}$. Mutual information I between p scalar random variables $Y_1, ..., Y_p$ is in turn defined as $I = \sum_{i=1}^{p} H_i - H$ where H_i is the entropy of the random variable Y_i, while H is the entropy of the random vector $(Y_1, ..., Y_p)$. Mutual infor-

mation is a natural measure of the dependence between random variables ([2]). In fact, it is equivalent to the Kullback-Leibler divergence between the joint density f and the product of its marginal densities. It is always non-negative, and zero if and only if the variables are statistically independent. Mutual information is a common criterion for finding the ICA transform. Set $W = C^{-1}$ and define the ICA of a random vector $\mathbf{X}$ as $\mathbf{S} = W\mathbf{X}$, with W determined so that the mutual information of the transformed components S_j is minimized. The generalization to the case where the number of independent components is less than the number of mixtures (i.e., when C is rectangular) is simple to implement, for instance projecting the data into a subspace through PCA and then applying the ICA transform to the dimensionally-reduced data.

3 Data analysis and future developments

We apply ICA to our spatio-temporal mobile-network data. In $p = 672$ different time intervals we observe the surfaces $x_{1\cdot}, \ldots, x_{p\cdot} \in \mathbb{R}^n$, where n is the number of pixels covering the Milan area, each surface $x_{i\cdot}$ representing the spatial distribution of Erlang relative to time interval i. Then we look at these surfaces as linear combinations of a certain number, say $k \leq p$, of latent independent basis surfaces. Hence we represent:

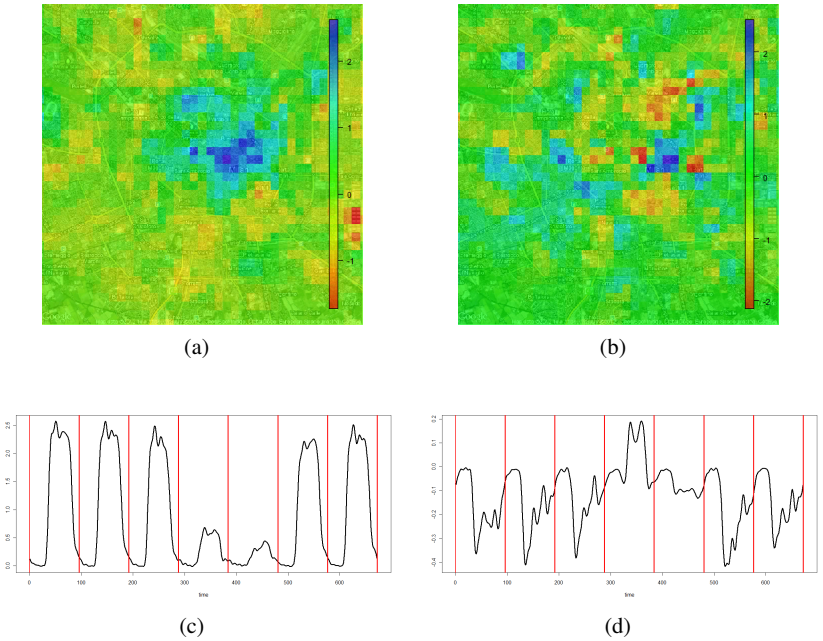
$$x_{iz} = c_{i1}s_{1z} + \ldots + c_{ik}s_{kz} \tag{2}$$

with $i = 1, \ldots, p$ and $z = 1, \ldots, n$. The k columns of the matrix $C \in \mathbb{R}^{p \text{ x } k}$ are the temporal weight profiles of the k basis surfaces.

It is interesting to compare different ICA transforms obtained for different values of k. Indeed, some latent surfaces appear for almost every choice of k. Figure (1) shows two estimated surfaces along with their temporal weight profiles, obtained with $k = 12$. The first surface is indeed robust with respect to the choice of the parameter k, and seems to represent the financial districts of Milan. Indeed, by inspecting its temporal weight profile, it appears that this surface weights more during the daily hours of working days than during the daily hours of weekend and it is turned off at nights. The second surface appears only for higher values of k and shows the high potential of the proposed exploratory method. Its temporal weight profile is active only on Saturday daily hours and the highest values on the surface correspond to the shopping districts of Milan. Hence the method is able to capture the temporal dynamics of some social urban features precisely localized in space. Within the GM project this information will help to localize specific urban areas that would profit from a third generation car sharing system, it will permit to identify locations for docking stations and it will be instrumental for the design of specific services offered to different communities (condo-sharing, firm-sharing, ...).

From the methodological point of view a deeper analysis is in order. One stimulating aspect of the problem is the choice of k that allows the best representation of the data. We are developing a hierarchical procedure where the choice of k is done along a tree. The output is a tree representation of nested basis capturing higher resolution details of urban dynamics. Moreover we want to compare our analysis with

those obtained through different approaches for similar problems, like Multivariate Curve Resolution ([3]) or decomposition through treelets ([4]). Finally, spatial dependence within each basis surface could be taken into account by introducing a penalization in the objective function. This last extension may well provide the main improvement in the proposed methodology.



(a) (b)

(c) (d)

Fig. 1: Two basis surfaces (top) with their temporal profiles (bottom) obtained applying ICA with k=12. The first day is Wednesday. The first surface (on the left) describes the financial districts of Milan. The second one (on the right) highlights shopping districts.

Acknowledgements Green Move project is supported by Regione Lombardia. Data are courtesy of Convenzione di ricerca DiAP - Telecom Italia, Politecnico di Milano (Italy).

References

1. Hyvarinen, A., Oja, E.: Independent Component Analysis: Algorithms and Applications. Neural Networks **13**, 411–430 (2000)
2. Cover, T. M., Thomas, J. A.: Elements of information theory. Wiley, New York (1991)
3. Jaumot, J., Gargallo, R., De Juan, A., Tauler, R.: A graphical user-friendly interface for MCR-ALS: a new tool for multivariate curve resolution in MATLAB. Chemometrics and Intelligent Laboratory System **76**, 101–110 (2005)
4. Secchi, P., Vantini, S., Vitelli, V.: Treelets analysis of spatially-dependent mobile network data. MOX-report, Departments of Mathematics, Politecnico di Milano (2012)