Inverse Extrapolation using Kernel Mixtures

Working Paper

Juste Raimbault

Date

Abstract

When using aggregated statistical data, one often observe an aggregated distribution over a classification. We propose a desaggregation method under the assumption of a kernel mixture.

1 Introduction

When studying geographical data, aggregation is generally done on different aspects. Spatial and temporal aggregations have been widely studied (MAUP, etc.). The aggregation over classes of a population may be problematic : for example, one can have the distribution of incomes for a population of a geographical area, but not the distributions conditioned to socio-economic classes (CSP) that are crucial for some statistical analyses. We describe here a method to reconstruct these, under the knowledge of population composition and simplifying assumptions on the shape of the conditional distributions. This problem is close to inverse problems in various fields of applied statistics, and using Gaussian mixtures is a well-known method [Yu et al., 2012]. It can also be understood as mode finding with a fixed number of modes [Carreira-Perpinan, 2000].

2 Method

2.1 Formalization

General case We assume a random variable W describing a population, for which we know the empirical distribution with density $f(w)$. The population is stratified into a finite number of categories $c \in C$. We assume that the shape of the distribution is known for each category and that it can be expressed with a kernel $k_c(w)$. With categories yielding a partition of the population, we have directly by independence, where w_c is the proportion of category c in the population such that $\sum_{c\in C} w_c = 1$,

$$f(w) = \sum_{c\in C} w_c \cdot k_c(w)$$

Parametrization Kernels are the unknown to be determined from the empirical distribution. In the general case, we want to minimize a cost function between the two distribution, i.e. solving

$$\min_{k_c} K(f, \sum_{c\in C} w_c k_c)$$

Let simplify the problem and take similar distributions for each category, such that $k_c = g(\vec{\alpha}_c)$. With a simple mean-square error cost function, our optimization problem becomes

$$\min_{\vec{\alpha}_c} \int_w \left(f(w) - \sum_c w_c \cdot \left[g(\vec{\alpha}_c) \right](w) \right)^2 dw$$

2.2 Implementation

Implement the optimization problem in R

In practice, we will have an histogram $(f(x_j))_j$ of the empirical distribution, yielding a discrete mean-square integration. The implemented function will be of the form

`extrapolate(distrib,weights,kernel,bounds)`

with `distrib` the histogram, `weights` composition of the population,`kernel` the kernel function (taking parameters and variable as arguments), `bounds` boundary for parameters (should be tunable for each category or general).

Remarks :

- kernel type can be left generic
- in what case do we have a convex problem (yielding unicity and speed of resolution) ?

3 Examples

try on data

References

[Carreira-Perpinan, 2000] Carreira-Perpinan, M. A. (2000). Mode-finding for mixtures of gaussian distributions. IEEE Transactions on Pattern Analysis and Machine Intelligence, 22(11):1318–1323.

[Yu et al., 2012] Yu, G., Sapiro, G., and Mallat, S. (2012). Solving inverse problems with piecewise linear estimators: From gaussian mixture models to structured sparsity. IEEE Transactions on Image Processing, 21(5):2481–2499.