Contributions to Statistics

Matteo Grigoletto
Francesco Lisi
Sonia Petrone *Editors*

Complex Models and Computational Methods in Statistics

Physica-Verlag

Springer

Contributions to Statistics

For further volumes:
http://www.springer.com/series/2912

Matteo Grigoletto • Francesco Lisi
Sonia Petrone
Editors

Complex Models and Computational Methods in Statistics

Physica-Verlag
A Springer Company

Editors
Matteo Grigoletto
Francesco Lisi
Department of Statistical Sciences
University of Padua
Padua
Italy

Sonia Petrone
Department of Decision Sciences
Bocconi University
Milan
Italy

In collaboration with Department of Statistical Sciences, University of Padua, Italy

ISSN 1431-1968
ISBN 978-88-470-2870-8 ISBN 978-88-470-2871-5 (eBook)
DOI 10.1007/978-88-470-2871-5
Springer Milan Heidelberg New York Dordrecht London

Library of Congress Control Number: 2012956315

Jointly published with Physica-Verlag Heidelberg, Germany

Printed on acid-free paper

Physica is a brand of Springer
Springer is a part of Springer Science+Business Media (www.springer.com)

Preface

The use of computational methods in statistics to face complex problems and highly dimensional data, as well as the widespread availability of computer technology, is no news. The range of applications, instead, is unprecedented. As often occurs, new and complex data types require new strategies, demanding for the development of novel statistical methods and suggesting stimulating mathematical problems.

This volume presents the revised version of a selection of the papers given at S.Co. 2011, the *7th Conference on Statistical Computation and Complex Systems*, held in Padua, Italy, September 19–21, 2011. The S.Co. conference is a forum for the discussion of new developments and applications of statistical methods and computational techniques for complex and high-dimensional datasets.

Although the topics covered in this volume are diverse, the same themes recur, as research is mostly fueled by the need to analyse complicated data sets, for which traditional methods do not provide viable solutions. Among the topics presented we have estimation of traffic matrices in a communications network, in the presence of long-range dependence; nonparametric mixed-effects models for epidemiology; advanced methods for neuroimaging; efficient computations and inference in environmental studies; hierarchical and nonparametric Bayesian methods with applications in genomic studies; Markov switching models to explain regime changes in the evolution of realized volatility for financial returns; joint modelling of financial returns and multiple daily realized measures; classification of multivariate linear–circular data, with applications to marine monitoring networks; forecasting of electricity supply functions, using principal component analysis and reduced rank regression; clustering based on nonparametric density estimation; surface estimation and spatial smoothing, with applications to the estimation of the blood-flow velocity field. Whilst not exhaustive, this list should give a feel of the range of issues discussed at the conference.

This book is addressed to researchers working at the forefront of the statistical analysis of complex systems and using computationally intensive statistical methods.

We wish to thank all contributors who made this volume possible. Finally, thanks must go to the reviewers, who responded rapidly when put under pressure and helped improve the papers with their valuable comments and suggestions.

Padua, Italy — Matteo Grigoletto, Francesco Lisi
Milan, Italy — Sonia Petrone

Contents

A New Unsupervised Classification Technique Through Nonlinear Non Parametric Mixed-Effects Models

Laura Azzimonti, Francesca Ieva, and Anna Maria Paganoni

Abstract In this work we propose a novel unsupervised classification technique based on the estimation of nonlinear nonparametric mixed-effects models. The proposed method is an iterative algorithm that alternates a nonparametric EM step and a nonlinear Maximum Likelihood step. We apply this new procedure to perform an unsupervised clustering of longitudinal data in two different case studies.

1 Introduction

Unsupervised clustering is one of the main topics in data mining, i.e., the process of finding useful information from data [4]. We focus our attention on highly overdispersed longitudinal and repeated data, which are naturally described through mixed-effects models. Nonlinear mixed-effects models (NLME models) are mixed-effects models in which at least one of the fixed or random effects appears nonlinearly in the model function. They are increasingly used in several biomedical and ecological applications, especially in population pharmacokinetics, pharmacodynamic, immune cells reconstruction and epidemiological studies (see [6, 7, 13, 21]). In these fields, statistical modeling based on NLME models takes advantage of tools that allow to distinguish overall population effects from drugs effects or unit specific influence. In general, mixed-effects models include parameters associated with the entire population (fixed effects) and subject/group specific parameters (random effects). For this reason, mixed-effects models are able to describe the dynamics of the phenomenon under investigation, even in the presence of high between subjects variability. When the random effects represent a deviation from the common dynamics of the population, mixed-effects models

L. Azzimonti · F. Ieva (✉) · A.M. Paganoni
MOX - Department of Mathematics, Politecnico di Milano, Milan, Italy
e-mail: laura.azzimonti@mail.polimi.it; francesca.ieva@mail.polimi.it; anna.paganoni@polimi.it

M. Grigoletto et al. (eds.), *Complex Models and Computational Methods in Statistics*, Contributions to Statistics, DOI 10.1007/978-88-470-2871-5_1,

provide estimates both for the entire population's model and for each subject's one. In this work random effects have a different meaning, in fact they describe the common dynamics of different groups of subjects. In this framework, mixed-effects models provide only estimates for each group-specific model. Thanks to this property, it will be possible to consider mixed-effects models as an unsupervised clustering tool for longitudinal data and repeated measures. For this reason we focus our attention on the estimation of the distribution of the random effects $\mathcal{P}^*$.

A wide literature exists for parametric modeling of random effects distribution in linear and non linear mixed-effects models. In this framework, Maximum Likelihood (ML) estimators are generally preferred because of their consistency and efficiency. However, due to the non linearity of the likelihood, we are not always able to provide explicitly the parameter estimators. A general and complete overview of linear multilevel models is given in [12]. An analogous overview for nonlinear case is given in [10]. In [9] it is shown how `R` and `S-plus` tools estimate linear and generalized linear mixed-effects models with parametric, in particular Gaussian, random effects. Concerning nonlinear models, in [11] a ML estimation of Gaussian random effect is provided for peculiar nonlinear forms. A stochastic approximation of traditional EM algorithm (SAEM) for estimating Gaussian random effects is suggested in [14], whereas an exact EM algorithm is described in [24]. Finally, [25] introduces a Laplace approximation for nonlinear random effects marginal distributions. However, parametric assumptions may sometimes result too restrictive to describe very heterogeneous or grouped populations. Moreover, when the number of measurements for unit is small, predictions for random effects are strongly influenced by the parametric assumptions. For these reasons nonparametric (NP) framework, which allow $\mathcal{P}^*$ to live in an infinite dimensional space, is attractive. Moreover, it provides in a very natural way a clustering tool, as we will highlight later.

Methods for the estimation of linear nonparametric random effects distribution in linear and generalized linear mixed-effects models have been proposed in [1, 2], whereas [3, 6, 15, 23], among others, deal with nonparametric nonlinear models.

In this work we propose a novel estimation method for nonlinear nonparametric mixed-effects models, aimed at unsupervised clustering. The proposed method is an iterative algorithm that alternates a nonparametric EM step and a nonlinear Maximum Likelihood step. The present algorithm is implemented in `R` program (version 2.13.0, R Development Core Team [20]) and the `R` source code is available upon request. To the best of our knowledge, this is the first example of free software for the estimation of nonlinear nonparametric mixed-effects models.

In Sect. 2 the general framework of the work is sketched out, and in Sect. 3 the algorithm for the estimation of nonlinear nonparametric random effect (NLNPEM) is described. Section 4 contains applications to case studies. Concluding remarks and further developments of this work are finally discussed in Sect. 5. Technical details in the estimation algorithm are discussed in Appendix.

2 Model and Framework

We consider the following NLME model for longitudinal data:

$$\begin{aligned} \mathbf{y}_i &= f(\boldsymbol{\beta}, \mathbf{b}_i, \mathbf{t}) + \boldsymbol{\epsilon}_i \;\; i = 1, \ldots, N \\ \boldsymbol{\epsilon}_i &\sim \mathcal{N}(\mathbf{0}, \sigma^2 \mathbb{I}_n) \qquad \text{i.i.d.} \end{aligned} \tag{1}$$

where $\mathbf{y}_i \in \mathbb{R}^n$ is the response variable evaluated at times $\mathbf{t} \in \mathbb{R}^n$ and f is a general, real-valued and differentiable function with $p + q$ parameters. Each parameter of f is treated either as fixed or as random. Fixed effects are parameters associated with the entire population whereas random effects are subject-specific parameters that allow to identify clusters of subjects. $\boldsymbol{\beta} \in \mathbb{R}^p$ is a vector that contain all fixed effects and $\mathbf{b}_i \in \mathbb{R}^q$ is the vector for the i-th subject random effects.

The function f is nonlinear at least in one component of the fixed or random effects. The errors ϵ_{ij} are associated with the j-th measurement of the i-th longitudinal data. They are normally distributed, independent between different subjects and independent within the same subject. In general, the proposed method could also take into account of a different number of observations, located at different times, for different subjects. In (1) we chose not to consider this case in order to ease the notation, but the generalization is straightforward.

Usually random effects are assumed to be Normal distributed, $\mathbf{b}_i \sim \mathcal{N}_q(\mathbf{0}, \Sigma)$, with unknown parameters that, together with $\boldsymbol{\beta}$ and σ, can be estimated through methods based on the likelihood function (see [18]). In this parametric framework the maximum likelihood estimators are generally favored by their statistical properties, i.e., consistency and efficiency. Nevertheless the parametric assumptions could be too restrictive to describe highly heterogeneous or grouped data, so it might be necessary to move to a nonparametric approach. In our case, we assume $\mathbf{b}_i$, for $i = 1, \ldots, N$, independent and identically distributed according to a probability measure $\mathcal{P}^*$ that belongs to the class of all probability measures on $\mathbb{R}^q$. $\mathcal{P}^*$ can then be interpreted as the mixing distribution that generates the density of the stochastic model in (1). Looking for the ML estimator $\hat{\mathcal{P}}^*$ of $\mathcal{P}^*$ in the space of all probability measures on $\mathbb{R}^q$, the discreteness theorem proved in [16] states that $\hat{\mathcal{P}}^*$ is a discrete measure with at most N support points. Moreover under suitable hypotheses on the distribution of the response variable, satisfied, for example, by densities in the exponential family, the ML estimator is also unique as proved in [17]. Therefore the ML estimator of the random effects distribution can be expressed as a set of points $(\mathbf{c}_1, \ldots, \mathbf{c}_M)$, where $M \leq N$ and $\mathbf{c}_l \in \mathbb{R}^q$, and a set of weights $(\omega_1, \ldots, \omega_M)$, where $\omega_l \geq 0$ and $\sum_{l=1}^{M} \omega_l = 1$.

As mentioned above, in this paper we propose an algorithm for the joint estimation of $\boldsymbol{\beta}$, $(\mathbf{c}_1, \ldots, \mathbf{c}_M)$, $(\omega_1, \ldots, \omega_M)$ and σ^2 in the nonlinear framework of model (1). The proposed method maximizes the following likelihood

$$L(\boldsymbol{\beta}, \sigma^2 \big| \, \mathbf{y}) = p(\mathbf{y} | \, \boldsymbol{\beta}, \sigma^2) = \sum_{l=1}^{M} \omega_l \frac{1}{(2\pi\sigma^2)^{(nN)/2}} \mathrm{e}^{-\frac{1}{2\sigma^2} \sum_{i=1}^{N} \sum_{j=1}^{n} \left(y_{ij} - f(\boldsymbol{\beta}, \mathbf{c}_l, t_j)\right)^2} \tag{2}$$

with respect to fixed effects $\boldsymbol{\beta}$, error variance σ^2, and the random effects distribution $(\mathbf{c}_l, \omega_l)$, $l = 1, \ldots, M$. Each iteration of the algorithm described in Sect. 3 increases the likelihood in (2).

Concerning the distribution of random effects, for each $l = 1, \ldots, M$, $\mathbf{c}_l$ and ω_l represent the group-specific parameters and the corresponding weights in the mixture (2), respectively. Notice that we do not have to fix a priori the number M of support points, but it is computed by the algorithm. Since we don't have to specify a priori the number of support points and in consequence the number of groups, the nonparametric mixed-effects model could be interpreted as an unsupervised clustering tool for longitudinal data. This tool could be very useful in order to identify groups of subjects to be used in the analysis and to cluster observations.

3 NLNPEM Algorithm

The algorithm proposed for the estimation of the parameters of model (1) arises from the framework described in [22], and it increases at each iteration the likelihood (2). The algorithm alternates two steps: the first one is a nonparametric EM step whereas the second one is a nonlinear maximum-likelihood step. The nonparametric EM step estimates the discrete q-dimensional distribution $(\mathbf{c}_l, \omega_l), l = 1, \ldots, M$ of the random effects $\mathbf{b}_i$. The non linear maximum likelihood step provides an estimation of the fixed effects β and the variance σ^2, given $\mathbf{b}_i$.

The nonparametric EM step consists in an update of the parameters of the discrete distribution $(\mathbf{c}_l, \omega_l), l = 1, \ldots, M$ that increases the likelihood function (2). The property of increasing the likelihood was proved in [22]. The update is the following:

$$\begin{cases} \omega_l^{\text{up}} = \dfrac{1}{N} \displaystyle\sum_{i=1}^{N} W_{il} \\ \mathbf{c}_l^{\text{up}} = \arg\max\limits_{\mathbf{c}} \left[\displaystyle\sum_{i=1}^{N} W_{il} \ln p(\mathbf{y}_i \,|\, \boldsymbol{\beta}, \sigma^2, \mathbf{c}) \right] \end{cases} \tag{3}$$

where

$$W_{il} = \frac{\omega_l \, p(\mathbf{y}_i \,|\, \boldsymbol{\beta}, \sigma^2, \mathbf{c}_l)}{\sum_{k=1}^{M} \omega_k \, p(\mathbf{y}_i \,|\, \boldsymbol{\beta}, \sigma^2, \mathbf{c}_k)}$$

and

$$p(\mathbf{y}_i \,|\, \boldsymbol{\beta}, \sigma^2, \mathbf{c}) = \frac{1}{(2\pi\sigma^2)^{n/2}} \mathrm{e}^{-\frac{1}{2\sigma^2} \sum_{j=1}^{n} \left(y_{ij} - f(\boldsymbol{\beta}, \mathbf{c}, t_j)\right)^2}.$$

The coefficients W_{il} represent the probability of $\mathbf{b}_i$ being equal to $\mathbf{c}_l$ conditionally to the observation $\mathbf{y}_i$ and given the fixed effects $\boldsymbol{\beta}$ and the variance σ^2, that is

$$W_{il} = p(\mathbf{b}_i = \mathbf{c}_l \,|\, \mathbf{y}_i, \boldsymbol{\beta}, \sigma^2)$$

in fact,

$$W_{il} = \frac{p(\mathbf{b}_i = \mathbf{c}_l)\, p(\mathbf{y}_i \,|\, \boldsymbol{\beta}, \sigma^2, \mathbf{c}_l)}{p(\mathbf{y}_i \,|\, \boldsymbol{\beta}, \sigma^2)} = \frac{p(\mathbf{y}_i, \mathbf{b}_i = \mathbf{c}_l \,|\, \boldsymbol{\beta}, \sigma^2)}{p(\mathbf{y}_i \,|\, \boldsymbol{\beta}, \sigma^2)} = p(\mathbf{b}_i = \mathbf{c}_l | \mathbf{y}_i, \boldsymbol{\beta}, \sigma^2).$$

In order to estimate $\mathbf{b}_i$ for $i = 1, \ldots, N$, we want to maximize the probability of $\mathbf{b}_i$ conditionally to the observations $\mathbf{y}_i$ and given the fixed effects $\boldsymbol{\beta}$ and the error variance σ^2. For this reason the estimation of the random effects, $\hat{\mathbf{b}}_i$, is obtained maximizing W_{il} over l, that is

$$\hat{\mathbf{b}}_i = \mathbf{c}_{\tilde{l}} \text{ if } \tilde{l} = \arg\max_l W_{il}.$$

During the nonparametric EM step, we could also reduce the support of the discrete distribution. The reduction of the support is performed in order to cluster the random effects. This support reduction consists in both making points very close to each other collapse and removing points with very low weight and not associated with any subject. In particular if two points are too close, that is $\|\mathbf{c}_l - \mathbf{c}_k\| < D$, where D is a tuning tolerance parameter, then we replace $\mathbf{c}_l$ and $\mathbf{c}_k$ with a new point $\mathbf{c}_{\min\{l,k\}} = (\mathbf{c}_l + \mathbf{c}_k)/2$ with weight $\omega_{\min\{l,k\}} = \omega_l + \omega_k$. Otherwise, if $\omega_l < \tilde{\omega}$, where $\tilde{\omega}$ is another tuning tolerance parameter, and the subset $\left\{i : \hat{\mathbf{b}}_i = \mathbf{c}_l\right\}$ is empty, we remove the point $\mathbf{c}_l$. The thresholds D and $\tilde{\omega}$ are two complexity parameters that affect the estimation of the nonparametric distribution; $\tilde{\omega}$ is linked to the size of the smallest group that we want to detect, while D represents the minimum allowed distance between different points of the discrete random effects distribution; the higher D is set, the lower is the number of groups. For this reason the two complexity parameters define a trade-off between bias and high number of groups. In this work we prefer setting D low in order to obtain a higher number of groups and, in case, cluster them later. A rule of thumb for setting these threshold parameters is the following: D may be much smaller than the standard deviation within groups, on the other hand, $\tilde{\omega}$ may be set of the same order of the inverse of the total number of observations in the dataset.

The nonlinear maximum likelihood step provides the estimation of the fixed effects $\boldsymbol{\beta}$ and the errors variance σ^2, given $\mathbf{b}_i = \hat{\mathbf{b}}_i$. In this step we maximize the nonlinear log-likelihood:

$$\ell(\boldsymbol{\beta}, \sigma^2 \big| \mathbf{y}, \hat{\mathbf{b}}) = -\frac{nN}{2} \ln(2\pi\sigma^2) - \frac{1}{2\sigma^2} \sum_{i=1}^{N} \sum_{j=1}^{n} \left(y_{ij} - f(\boldsymbol{\beta}, \hat{\mathbf{b}}_i, t_j) \right)^2 \qquad (4)$$

where $\hat{\mathbf{b}}_i$ is the estimation of random effects for the i-th subject provided in the nonparametric EM step.

The algorithm, given a starting discrete distribution with N support points for the random effects and a starting estimate for the fixed effects, alternates the nonparametric EM step and the nonlinear maximum likelihood step until convergence. More details, together with the sketch of the algorithm, are reported in Appendix.

In order to validate the proposed estimation algorithm and to compare it with already existing procedures for the linear framework, an intensive simulation study has been performed and detailed in [5]. In the first simulation study (see [5], Sect. 3.2), we compared the results obtained in a linear framework with those obtained with the algorithm introduced in [1] and implemented in the `npmlreg` `R`-package (see [8]). In the second one (see [5], Sect. 3.3) we considered two classic nonlinear functions f in (1): the exponential and the logistic growth curves. For each case a test set of simulated curves has been designed and the algorithm performance in the estimation of the random effects has been evaluated computing the Wasserstein distance between the true and the estimated distribution of the random effects.

In the linear framework NLNPEM method performs very well and its results are comparable with those obtained with the already existing npmlreg method; for a large number of groups, npmlreg method doesn't detect some points of the nonparametric distribution or even doesn't reach convergence, whereas NLNPEM performs well, even ignoring the true number of groups. Both in the linear and nonlinear framework we obtain a very high level of agreement, measured in term of Wasserstein distance between the true distribution generating data and the one estimated by NLNPEM algorithm. The NLNPEM method is also able to capture correctly outlier groups even in highly unbalanced situations.

4 Application to Case Studies

In this section we apply the proposed method to two different datasets: the first one contains the carbon dioxide uptake photosynthetic response curves in a sample of 12 different plants. It is a classical dataset for the study of longitudinal curves presented in [19] in a study of the cold tolerance of a C_4 grass species, *Echinochloa crus-galli* and analyzed also in [18]. The second one describes the number of Hospital Discharges of patients affected by Acute Myocardial Infarction (AMI) without ST-segment Elevation (NON-STEMI) along the time period 2000–2007, grouped by hospital and relative to the 30 largest clinical institutions of Regione Lombardia. The explorative analysis of these data is aimed at detecting groups with similar behaviours.

4.1 Carbon Dioxide Uptake

In the first case we consider the carbon dioxide ($C0_2$) uptake [$\mu mol \cdot m^{-2} \cdot s^{-1}$] of 12 plants, measured at several levels of ambient $C0_2$ concentration [$\mu L/L$], see Fig. 1. In [19] an exponential growth model is proposed to capture the common shape of the curves. In this case the nonlinear function to be used in the model (1) is:

$$f(t) = \alpha \left(1 - e^{-\lambda t}\right)$$

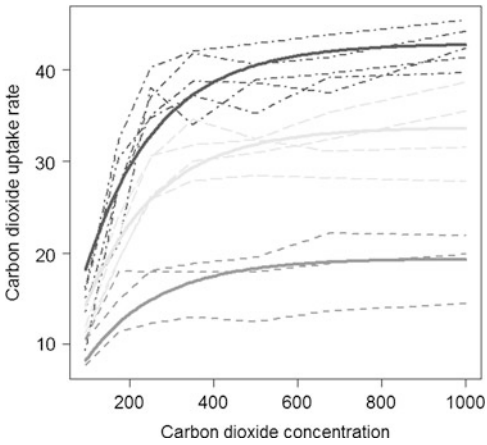

Fig. 1 Carbon dioxide uptake photosynthetic curves for 12 plants. Real data are colored according to the NLNPEM clusters and NLNPEM fitted models are superimposed

which is nonlinear in λ. The two parameters α and λ represent, respectively, the asymptote and the growth rate.

In this analysis we consider only random effects for the asymptote, that means that the mixed-effects model becomes

$$\mathbf{y}_i = a_i \left(1 - \mathrm{e}^{-\lambda \mathbf{t}}\right) + \boldsymbol{\epsilon}_i$$

where $\boldsymbol{\epsilon}_i \sim \mathcal{N}(\mathbf{0}, \sigma^2 \mathbb{I}_n)$ are i.i.d. errors, a_i are the random effects for the asymptote ($b_i = a_i$), and λ is the fixed effect for the growth rate ($\beta = \lambda$).

The NLNPEM algorithm clusters the plants in $M = 3$ different groups, according to the estimated discrete distribution of the random effect for the asymptote (see Fig. 1). The estimated fixed effect is $\hat{\lambda} = 0.006$, the estimated discrete measure $\hat{\mathcal{P}}^*$ is concentrated on $(\hat{\mathbf{c}}_1, \hat{\mathbf{c}}_2, \hat{\mathbf{c}}_3) = (19.39, 33.71, 42.89)$ with weights $(\hat{\omega}_1, \hat{\omega}_2, \hat{\omega}_3) = (0.25, 0.33, 0.42)$ and the estimated variance is $\hat{\sigma}^2 = 8.94$. This analysis, performed with $D = 5$ and $\tilde{\omega} = 0.05$, backs up the presence of three groups of plants according to different asymptotes and automatically detects an unsupervised cluster structure. This result is in total agreement with a k-means clustering of the random asymptote point estimates computed following the traditional parametric approach [18] that assumes a Normal model for the random effect. Nevertheless, in that case, a critical point is the choice of k, the number of groups, which is set equal to three after maximizing the average silhouette width. On the contrary the number of groups is automatically computed in the NLNPEM method.

4.2 Acute Myocardial Infarction Without ST-Segment Elevation

The second example analyzed comes from epidemiological studies carried out using administrative databanks. In fact, Fig. 2 represents the normalized number of NON-STEMI diagnoses along the time period 2000–2007 grouped by hospital

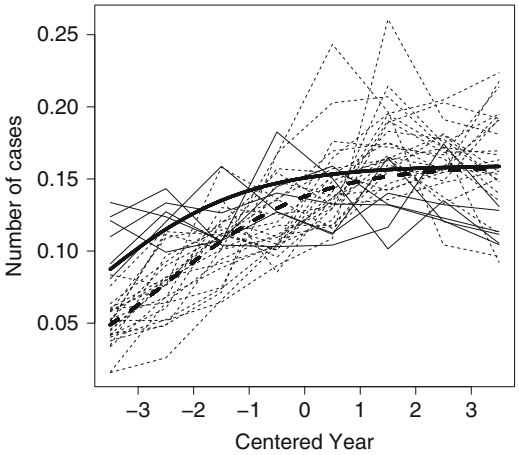

Fig. 2 Standardized number of AMI without ST-segment elevation diagnoses in the period 2000–2007 in the 30 largest clinical institutions of Lombardia Region. The year has been centered and normalization has been carried out standardizing the yearly number of diagnoses for each hospital by total number of diagnoses in the time window 2000–2007. Clusters pointed out by NLNPEM algorithm are highlighted, respectively, by *solid* and *dashed lines*. NLNPEM fitted models are superimposed

and relative to the 30 largest clinical institutions of Regione Lombardia. For each hospital the yearly number of diagnoses has been standardized by the hospital total number of diagnoses in the time period 2000–2007. As pointed out in [13] a logistic growth model with random inflection seems to capture the common "S-shaped" growing pattern; for this reason we consider a logistic growth model. In this case, the nonlinear function to be used in model (1) is:

$$f(t) = \frac{\alpha}{1 + \mathrm{e}^{-\frac{t-\delta}{\gamma}}}$$

where α represent the asymptote, δ is the inflection point, which correspond to the time at which the growth curve reaches the half of the asymptote, and γ is the time elapsed between δ and the time at which the growth curve reaches 3/4 of the asymptotic level. The parameters γ and α are treated as fixed effects while the inflection point is treated as random, as suggested in [13]. The model becomes:

$$\mathbf{y}_i = \frac{\alpha}{1 + \mathrm{e}^{-\frac{\mathbf{t}-d_i}{\gamma}}} + \boldsymbol{\epsilon}_i$$

where $\boldsymbol{\epsilon}_i \sim \mathcal{N}(\mathbf{0}, \sigma^2 \mathbb{I}_n)$ are i.i.d. errors, d_i represent the random effects for the inflection point, while α and γ represent the fixed effects. In particular $b_i = d_i$ and $\boldsymbol{\beta} = (\alpha, \gamma)$.

The NLNPEM algorithm clusters the hospitals in $M = 2$ different groups, according to the estimated discrete distribution of the random inflection point

(see Fig. 2). The estimated fixed effects are $\hat{\alpha} = 0.16$ and $\hat{\gamma} = 1.31$, the estimated discrete measure $\hat{\mathcal{P}}^*$ is concentrated on $(\hat{\mathbf{c}}_1, \hat{\mathbf{c}}_2) = (-3.76, -2.43)$ with weights $(\hat{\omega}_1, \hat{\omega}_2) = (0.2, 0.8)$ and the estimated variance is $\hat{\sigma}^2 = 7.7 \cdot 10^{-4}$. This analysis, performed with $D = 0.05$ and $\tilde{\omega} = 0.05$, backs up the presence of two groups of hospitals according to different inflection points.

Even if clinical best practice maintains that there is no evidence for a greater incidence of NON-STEMI in this period, it is known that since the early 2000s a new diagnostic procedure—the *troponin* exam—has been introduced and this could have produced an increased number of positive diagnoses by easing NON-STEMI detection. Hence, the presence of two clusters could be a consequence of the different hospital timings in the introduction and adoption of this practice. This hypothesis cannot be validated directly since the timings of adoption of the troponin exam by the 30 different hospitals included in the analysis are not available.

The good agreement with previous results detailed in [13] together with the great advantage of a nonparametric approach advocates the real profit in using this new estimation algorithm.

5 Conclusions

In this work, we proposed a new unsupervised clustering technique based on a new estimation method for nonlinear nonparametric mixed-effects models. The proposed method is based on an iterative algorithm (named NLNPEM) that alternates a nonparametric EM and an optimization step for the maximization of a nonlinear likelihood function. A simulation study both in linear and nonlinear setting of exponential and logistic growth has been carried out. Results show that NLNPEM performs well, even ignoring the real number of groups, in terms of Wasserstein distance between the true distribution generating data and the one estimated by NLNPEM algorithm, and that it always reaches convergence, even in those cases where several groups are present. We use this algorithm as an unsupervised clustering technique in the context of the explorative data mining. In particular two applications to real data of carbon dioxide uptake photosynthetic response curves and NON-STEMI number of diagnoses, respectively, are presented. In these two case studies the potential of our method in unsupervised clustering analysis is highlighted.

Appendix: Details on NLNPEM Algorithm

The NLNPEM is the following:

1. Define a starting discrete distribution for random effects with support on N points $(\mathbf{c}^{(0)}, \boldsymbol{\omega}^{(0)})$, a starting estimate for the fixed effects $\boldsymbol{\beta}^{(0)}$ and for $\sigma^{2(0)}$ and the tolerance parameters D and $\tilde{\omega}$.

2. Given $(\mathbf{c}^{(k-1)}, \boldsymbol{\omega}^{(k-1)})$, $\boldsymbol{\beta}^{(k-1)}$ and $\sigma^{2(k-1)}$, perform the EM step (without the support reduction) in order to update the support points $\mathbf{c}^{(k)}$ and the weights $\boldsymbol{\omega}^{(k)}$ of the random effect distribution, according to (3).
3. Given $(\mathbf{c}^{(k)}, \boldsymbol{\omega}^{(k)})$, perform the nonlinear maximum likelihood step in order to estimate the fixed effects $\boldsymbol{\beta}^{(k)}$ and the error variance $\sigma^{2(k)}$ maximizing (4).
4. Iterate Steps 2 and 3 until convergence.
5. Reduce the support of the discrete distribution, according to the tuning parameters D and $\tilde{\omega}$.
6. Given $(\mathbf{c}^{(k-1)}, \boldsymbol{\omega}^{(k-1)})$, $\boldsymbol{\beta}^{(k-1)}$, $\sigma^{2(k-1)}$, D and $\tilde{\omega}$, perform the EM step with the support reduction in order to update the support points $\mathbf{c}^{(k)}$ and the weights $\boldsymbol{\omega}^{(k)}$ of the random effect distribution, according to (3).
7. Given $(\mathbf{c}^{(k)}, \boldsymbol{\omega}^{(k)})$, perform the nonlinear maximum likelihood step in order to estimate the fixed effects $\boldsymbol{\beta}^{(k)}$ and the error variance $\sigma^{2(k)}$ maximizing (4).
8. Iterate Steps 6 and 7 until convergence.

The algorithm reaches convergence when parameters and discrete distribution stop changing or when there is no variation in the log-likelihood function.

Acknowledgments The case study in Sect. 4 is within the Strategic Program "Exploitation, integration and study of current and future health databases in Lombardia for Acute Myocardial Infarction" supported by "Ministero del Lavoro, della Salute e delle Politiche Sociali" and by "Direzione Generale Sanità—Regione Lombardia."

References

1. Aitkin, M.: A general maximum likelihood analysis of overdispersion in generalized linear models. Stat. Comput. **6**, 251–262 (1996)
2. Aitkin, M.: A general maximum likelihood analysis of variance components in generalized linear models. Biometrics **55**, 117–128 (1999)
3. Antic, J., Laffont, C.M., Chafaï, D., Concordet, D.: Comparison of nonparametric methods in nonlinear mixed effect models. Comput. Stat. Data Anal. **53**(3), 642–656 (2009)
4. Azzalini, A., Scarpa, B.: Data Analysis and Data Mining. Oxford Univeristy Press, Oxford (2012)
5. Azzimonti, L., Ieva F., Paganoni, A.M.: Nonlinear nonparametric mixed-effects models for unsupervised classification. Computational Statistics, Published online: 27 September 2012. DOI: 10.1007/s00180-012-0366-5
6. Davidian, M., Gallant, A.R.: The nonlinear mixed effects model with a smooth random effects density. Biometrika **80**(3), 475–488 (1993)
7. De Lalla, C., Rinaldi, A., Montagna, D., Azzimonti, L., Bernardo, M.E., Sangalli, L.M., Paganoni, A.M., Maccario, R., Di Cesare Merlone, A., Zecca, M., Locatelli, F., Dellabona, P., Casorati, G.: Invariant natural killer T-cell reconstitution in pediatric leukemia patients given HLA-haploidentical stem cell transplantation defines distinct CD4+ and CD4− subset dynamics and correlates with remission state. J. Immunol. **186**(7), 4490–4499 (2011)
8. Einbeck, J., Darnell, R., Hinde, J.: npmlreg: nonparametric maximum likelihood estimation for random effect models. [Online] http://CRAN.R-project.org/package=npmlreg (2009) (Accessed: 26 November 2012)

9. Fox, J.: Linear mixed models. Appendix to An R and S-PLUS Companion to Applied Regression. Sage Publications Inc. California (2002) http://cran.r-project.org/doc/contrib/Fox-Companion/appendix.html (Accessed: 26 November 2012)
10. Gallant, A.R.: Nonlinear Statistical Models. Wiley, New York (1987)
11. Goldstein, H.: Nonlinear multilevel models, with an application to discrete response data. Biometrika **78**(1), 45–51 (1991)
12. Hox, J.J.: Applied Multilevel Analysis. TT-Publikaties, Amsterdam (1995)
13. Ieva, F., Paganoni, A.M., Secchi, P.: Mining administrative health databases for epidemiological purposes: a case study on Acute Myocardial Infarctions diagnoses. In: Pesarin, F., Torelli, S. (eds.) Accepted for Publication in Advances in Theoretical and Applied Statistics. Springer, Berlin (2012) http://mox.polimi.it/it/progetti/pubblicazioni/quaderni/45-2010.pdf
14. Kuhn, E., Lavielle, M.: Maximum likelihood estimation in nonlinear mixed effect models. Comput. Stat. Data Anal. **49**(4), 1020–1038 (2005)
15. Lai, T.L., Shih, M.C.: Nonparametric estimation in nonlinear mixed-effects models. Biometrika **90**(1), 1–13 (2003)
16. Lindsay, B.G.: The geometry of mixture likelihoods: a general theory. Ann. Stat. **11**(1), 86–94 (1983a)
17. Lindsay, B.G.: The geometry of mixture likelihoods, Part II: the exponential family. Ann. Stat. **11**(3), 783–792 (1983b)
18. Pinheiro, J.C., Bates, D.M.: Mixed-Effects Models in S and S-Plus. Springer, Berlin (2000)
19. Potvinm, C., Lechowicz, M.J., Tardif, S.: The statistical analysis of ecophysiological response curves obtained form experiments involving repeated measures. Ecology **71**(4), 1389–1400 (1990)
20. R Development Core Team: R: a language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria [Online] http://www.R-project.org (2009) (Accessed: 26 November 2012)
21. Sheiner, L.B., Beal, S.L.: Evaluation of methods for estimating population pharmacokinetic parameters. III. Monoexponential model: routine clinical pharmacokinetic data. J. Pharmacokinet. Pharmacodyn. **11**(3), 303–319 (1980)
22. Schumitzky, A.: Nonparametric EM algorithms for estimating prior distributions. Appl. Math. Comput. **45**(2), 143–157 (1991)
23. Vermunt, J.K.: An EM algorithm for the estimation of parametric and nonparametric hierarchical models. Statistica Neerlandica **58**, 220–233 (2004)
24. Walker, S.: An EM algorithm for nonlinear random effects models. Biometrics **52**(3), 934–944 (1996)
25. Wolfinger, R.: Laplace's approximation for nonlinear mixed models. Biometrika **80**, 791–795 (1993)

Estimation Approaches for the Apparent Diffusion Coefficient in Rice-Distributed MR Signals

Stefano Baraldo, Francesca Ieva, Luca Mainardi, and Anna Maria Paganoni

Abstract The Apparent Diffusion Coefficient (ADC) is often considered in the differential diagnosis of tumors, since the analysis of a field of ADCs on a particular region of the body allows to identify regional necrosis. This quantity can be estimated from magnitude signals obtained in diffusion Magnetic Resonance (MR), but in some situations, like total body MRs, it is possible to repeat only few measurements on the same patient, thus providing a limited amount of data for the estimation of ADCs. In this work we consider a Rician distributed magnitude signal with an exponential dependence on the so-called b-value. Different pixelwise estimators for the ADC, both frequentist and Bayesian, are proposed and compared by a simulation study, focusing on issues caused by low signal-to-noise ratios and small sample sizes.

1 Introduction

Diffusion magnetic resonance (MR) is as an important tool in clinical research, as it allows to characterize some properties of biological tissues. When tumor areas are analyzed using this technique, it can be observed that the diffusion tensor, estimated from the magnetic MR magnitude signal, has reduced values in lesions with respect to surrounding physiological tissues, allowing to identify pathological areas or necrosis. When the tissue region of interest can be considered as isotropic the *Apparent Diffusion Coefficient* (ADC) is sufficient to characterize the diffusion

S. Baraldo (✉) · F. Ieva · A.M. Paganoni
MOX - Department of Mathematics, Politecnico di Milano, Milan, Italy
e-mail: stefano1.baraldo@mail.polimi.it; francesca.ieva@mail.polimi.it; anna.paganoni@polimi.it

L. Mainardi
Department of Bioengineering, Politecnico di Milano, Milan, Italy
e-mail: luca.mainardi@polimi.it

M. Grigoletto et al. (eds.), *Complex Models and Computational Methods in Statistics*, Contributions to Statistics, DOI 10.1007/978-88-470-2871-5_2,

properties of the tissue, and it is usually estimated from the exponential decay of the signal with respect to the *b-value*, the MR acquisition parameter. The assumption of isotropy is common and reasonable in various cases, like breast and prostate cancer (see, for example, [7, 10]).

In many practical situations it may not be possible to collect more than few measures at different b-values, limiting the accuracy of the estimation. A reduction in the total number of measures necessary to achieve a certain accuracy is convenient in term of costs and allows to keep the patient involved in the MR procedure for a shorter amount of time (the experience may be unpleasant, especially when total body MR must be performed). The purpose of this work is to compare different frequentist and Bayesian approaches to the estimation of the ADC, underlining their statistical properties and computational issues.

2 Rice-Distributed Diffusion MR Signals

2.1 The Rice Distribution

The random variables we deal with derive from the complex signal $w = w_r + \mathrm{i}w_i$ measured in diffusion MR. It is usual to assume that both w_r and w_i are affected by a Gaussian noise with equal, constant variance, i.e. $w_r \sim \mathcal{N}(\nu\cos(\vartheta), \sigma^2)$ and $w_i \sim \mathcal{N}(\nu\sin(\vartheta), \sigma^2)$, with $\nu \in \mathbb{R}^+$ and $\vartheta \in [0, 2\pi)$. The quantity at hand is the modulus M of this signal, which has then a *Rice* (or *Rician*) distribution, that we will denote as $M \sim \mathrm{Rice}(\nu, \sigma^2)$. The density of this random variable has the form

$$f_M(m|\nu, \sigma^2) = \frac{m}{\sigma^2}\mathrm{e}^{-\frac{m^2+\nu^2}{2\sigma^2}} I_0\left(\frac{m\nu}{\sigma^2}\right)\mathbb{I}_{(0,+\infty)}(m), \tag{1}$$

where I_0 is the zeroth-order modified Bessel function of the first kind (see [1]). Using the series expression of I_0, it is possible to deduce a different, equivalent definition of a Rician random variable as $M = \sigma\sqrt{R}$, where R is a noncentral χ^2 variable that can be expressed as a mixture of $\chi^2(2P + 2)$ distributions with $P \sim Poisson(\nu^2/2\sigma^2)$. This formulation becomes particularly useful for sampling from a Rice distribution, as it allows an easy implementation of a Gibbs sampler.

2.2 Rice Exponential Regression

Diffusion MR aims at computing the diffusion tensor field on a portion of tissue, and this is achieved by analyzing the influence of water diffusion on the measured signal, under different experimental settings. In particular, the classical model for relating the magnitude signal to the acquisition parameters and the 3-dimensional diffusion tensor D is the *Stejskal–Tanner* equation

$$\nu_{\mathbf{g}} = \nu_0 \exp(-\mathbf{g}^{\mathrm{T}} D \mathbf{g} b), \tag{2}$$

where $\nu_{\mathbf{g}}$ is the "real" intensity signal we want to measure, ν_0 is the signal at $b = 0$ and the vector $\mathbf{g} \in \mathbb{R}^3$ is the applied magnetic gradient. The b-value is a function of other acquisition settings, which we will omit since their description and discussion is beyond the scope of this article. See, for example, [3] for an overview on MR techniques, including diffusion MR, and a discussion of various issues and recent advances in this field.

In general, even in the ideal noiseless case, at least six observations are needed to determine the components of the symmetric, positive definite diffusion tensor D, by varying the direction $\mathbf{g}$ of the magnetic field gradient. However, if the tissue under study can be considered as isotropic, the diffusion tensor has the simpler form $D = \alpha I$, where α is the ADC, a scalar parameter, and I is the identity matrix. This reduces model (2) to the following

$$\nu = \nu_0 \exp(-\alpha b) \tag{3}$$

for any vector $\mathbf{g}$ (in the following, we will omit it for ease of notation).

Equation (3) describes pointwise the phenomenon on the tissue region of interest. In this study we consider the pixels of a diffusion MR sequence of images as independent and focus on the estimation problem for a single point in space. We do not consider a spatial modeling for the ADC field: although it could be a useful way to filter noise and to capture underlying tissue structures; on the other hand, for diagnostic purposes it may be preferable to submit to the physician an estimate that has not been artificially smoothed.

3 Estimation Methods

In this section we present different methods for the estimation of α, the unknown parameter of interest. We consider a sample of signal intensities on a single pixel $M_i \sim \mathrm{Rice}(\nu_0 \mathrm{e}^{-\alpha b_i}, \sigma^2)$, $i = 1, \ldots, n$, and their respective realizations $\mathbf{m} = m_1, \ldots, m_n$ at b-values $\mathbf{b} = b_1, \ldots, b_n$.

The dispersion parameter σ^2 is usually measured over regions where almost pure noise is observed, and used as a known parameter in the subsequent estimates. This estimate of σ^2 is considered as reliable, since it can be based on a very large number of pixels, so we will consider the case of known dispersion parameter.

We consider nonlinear least squares, maximum likelihood and three Bayesian point estimators. In the case of a simple $\mathrm{Rice}(\nu, \sigma^2)$ random variable an iterative method of moments estimator has been proposed in [2], but this technique has no straightforward extension to the case of covariate-dependent ν, while moment equations would be difficult to invert in the considered case. Moreover, under the model assumptions presented in Sect. 2 a decoupling of noise and signal in