SECONDA UNIVERSITÀ DEGLI STUDI DI NAPOLI
FACOLTÀ DI STUDI POLITICI E PER L'ALTA FORMAZIONE EUROPEA
E MEDITERRANEA «JEAN MONNET»

FIRST JOINT MEETING OF THE SOCIÉTÉ FRANCOPHONE DE CLASSIFICATION AND THE CLASSIFICATION AND DATA ANALYSIS GROUP OF THE ITALIAN STATISTICAL SOCIETY

BOOK OF SHORT PAPERS

June, 11-13, 2008 – Caserta-Italy

Edizioni Scientifiche Italiane

Scientific Committee

B. Fichet (chair)	Universite d'Aix-Marseille II
D. Piccolo (chair)	Università di Napoli Federico II
S. Balbi	Università di Napoli Federico II
F. Bartolucci	Università di Perugina
C. Biernacki	Université des Sciences et Technologies de Lille - LILLE I
H.H. Bock	University of Aachen - Germany
F.D.A.T. De Carvalho	Universidade Federal de Pernambuco
A. Chouakria-Douzal	Universite Joseph Fourier Grenoble
M. Civardi	Università degli Studi di Milano - Bicocca
L. Ferré	Université Toulouse Le Mirail
V. Esposito Vinzi	ESSEC Business school Paris
J. Gama	University of Porto - Representant of CLAD
A. Guénoche	Université d'Aix-Marseille II
G. Hébrail	ENST, Paris
P. Kuntz	LINA, Université Nantes
V. Makarenkov	Université de Montréal
S. Mignani	Alma Mater Studiorum Università di Bologna
C. Rampichini	Università degli Studi di Firenze
M. Rémon	Facultés Universitaires Notre-Dame de la Paix - Namur
N. Torelli	Università degli Studi di Trieste
R. Verde	Seconda Università di Napoli
M. Vichi	Università di Roma «La Sapienza»
C. Weihs	Universität Dortmund - Representant of GfKl
S. Zani	Università degli Studi di Parma
D. Zighed	Université de Lyon

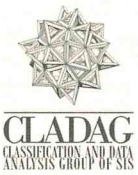
Organizing committee

Rosanna Verde (chair)	Dipartimento di Studi Europei e Mediterranei Seconda Università di Napoli
Antonio Irpino	Dipartimento di Studi Europei e Mediterranei Seconda Università di Napoli
Mario Guarracino	Istituto di Calcolo e Reti ad Alte Prestazioni Consiglio Nazionale delle Ricerche
Mauro Iacono	Dipartimento di Studi Europei e Mediterranei Seconda Università di Napoli
Elvira Romano	Dipartimento di Studi Europei e Mediterranei Seconda Università di Napoli
Antonio Balzanella	Dipartimento di Matematica e Statistica Università di Napoli Federico II
Dora Biceglia	Dipartimento di Matematica e Statistica Università di Napoli Federico II
Valentina Cozza	Dipartimento di Matematica e Statistica Università di Napoli Federico II

SECONDA UNIVERSITÀ DEGLI STUDI DI NAPOLI
FACOLTÀ DI STUDI POLITICI E PER L'ALTA FORMAZIONE EUROPEA
E MEDITERRANEA «JEAN MONNET»

FIRST JOINT MEETING OF THE SOCIÉTÉ FRANCOPHONE DE CLASSIFICATION AND THE CLASSIFICATION AND DATA ANALYSIS GROUP OF THE ITALIAN STATISTICAL SOCIETY

BOOK OF SHORT PAPERS
June, 11-13, 2008 – Caserta - Italy



Edizioni Scientifiche Italiane

First joint meeting of the Société Francophone de Classification and the Classification and Data Analysis Group of the Italian Statistical Society
Collana: Pubblicazioni della Facoltà di Studi Politici e per l'Alta Formazione Europea e Mediterranea «Jean Monnet» della Seconda Università degli Studi di Napoli
Sezione: Atti e Convegni, 5
Napoli: Edizioni Scientifiche Italiane, 2008
pp. XXIV + 456; 24 cm
ISBN 978-88-495-1656-2

80121 Napoli, via Chiatamone 7
00185 Roma, via dei Taurini 27

Internet: www.edizioniesi.it
E-mail: info@edizioniesi.it

Preface

This book contains the revised short papers presented during the first joint meeting of Société Francophone de Classification and the Classification and Data Analysis Group of the Italian Statistical Society held in Caserta at Royal Palace, June, 11-13, 2008.

The scientific program of the conference included 72 contributed papers and 8 invited sessions. Moreover, 8 international renowned invited speakers kindly accepted to present their current research works concerning the core topics of Classification and Data Analysis.

The scientific meeting covered the following topics:

Association Rules
Business Intelligence
Categorical Data Analysis
Classification and Discrimination
Clustering and distance analysis
Computational Bayesian Methods
Data Mining
Data Streams and Massive data
Design of Experiments
Dimensionality Reduction
Econometrics and Statistical Finance
Econometrics and Statistical Finance
Functional Data Analysis
Genomics and Microarray Data Analysis
Graphics and Data Visualization
Imprecise Data and Fuzzy Methods
Machine Learning
Matrix Computations and Statistics
Metadata and Data Representation
Multivariate Data Analysis
Neural Networks and Genetic Algorithms
Nonparametric Statistics and Smoothing
Optimization Algorithms
Partial Least Squares
Pattern Recognition
Resampling Methods
Risk Analysis and Scoring
Robustness
Sensometrics
Spatial Statistics
Structural Equations Models
Support Vector Machine
Symbolic Data Analysis
Textual Data Analysis
and Information Retrieval
Time Series Analysis
Multiway Data Analysis

The meeting was widely supported by the Facoltà di Studi Politici e per l'Alta Formazione Europea e Mediterranea "Jean Monnet" of the Seconda Università di Napoli for scientific and organising aspects.

The initiative was patronized by Comune di Caserta and partially supported by Confindustria Caserta, Dipartimento di Studi Europei e Mediterranei - Seconda Università di Napoli, C.C.I.A.A. di Caserta and Parco Regionale del Matese.

The Organising Committee is grateful to Ente Provinciale per il Turismo, Scuola Superiore della Pubblica Amministrazione, Soprintendenza per i Beni Architettonici e per il Paesaggio per il Patrimonio Storico, Artistico ed Etnoantropologico delle Province di Caserta e Benevento for having hosted the conference sessions in the halls and the theatre of the Royal Palace of Caserta.

Special thanks are due to the members of the Scientific Committee for their contribution to the organisation of the invited sessions and the reviewing of the short papers, to the local Organising Committee for its work and specially to Antonio Irpinio and Antonio Balzanella for their hard editing activity which made possible the realisation of this book.

We hope you will enjoy the Conference as much as we enjoyed organising it.

Caserta, May, 2008

Bernard Fichet
Domenico Piccolo
Rosanna Verde

Historique de la SFC

Vers la fin des années 1970, il existait une société de classification comprenant une branche nord-américaine et une branche européenne; cette dernière était complètement anglophone. La Classification Automatique prenant une ampleur grandissante en France, Edwin Diday a pris l'initiative de créer, en 1977, en collaboration avec I.C. Lerman et S. Régnier, une société francophone indépendante, avec le soutien de l'INRIA. Ainsi fut créée en 1977 la première société, appelée à l'époque « Société Française de Classification », dédiée à la classification automatique et aux méthodes reliées d'analyse des données et de statistique. Elle avait pour but de promouvoir la communication, la collaboration, les échanges scientifiques entre tous ceux qui s'intéressent à la classification et aux méthodes reliées d'analyse des données et de statistique, tant du point de vue théorique que des applications, dans un esprit d'ouverture interdisciplinaire. Les présidents de la SFC ont été successivement S. Régnier, I.C. Lerman, M. Jambu, G. Der Megreditchian, P. Cazes, E. Diday, J.P. Barthélemy, J.P. Rasson, B. Fichet et Y. Lechevallier. Dès sa création la Société fut très active. Par exemple une session « Classification » fut traditionnellement organisée lors des « Journées de Statistiques » organisées par l'Association des Statisticiens Universitaires (ASU) ». D'autre part, suite à une proposition de l'ASU, un représentant de la SFC se joint au groupe de statisticiens invité au congrès de la Société Italienne de Statistique en 1986. De même, suite à la proposition de B. Van Cutsem, I.C. Lerman représente la SFC au sein du comité de rédaction de la revue « Statistique et Analyse des Données ». Le prix Simon Régnier, destiné à récompenser les travaux scientifiques d'un jeune chercheur, est attribué pour la première fois au cours des « Journées de Statistique » des 26-29 mai 1985. Le siège social de la SFC est transféré, en 1985, de l'INRIA au CNET.

La Société Francophone de Classification est membre fondateur de la Fédération Internationale des Sociétés de Classification (IFCS). En 1985, l'Assemblée Générale de la SFC ratifie la création de la Fédération Internationale des Sociétés de Classification (IFCS) dont les membres étaient les sociétés suivantes : « British Classification Society (BSC) », « Classification Society of North America (CSNA) », « Gesellschaft für Klassifikation e.V. (GfKl) », « Japanese Classification Society (JCS) », « Société Francophone de Classification (SFC) », « Italian Statistical Society (SIS) ». Pour la SFC, M. Jambu et C. Perruchet signèrent l'acte de création.

Nonparametric intensity estimation in space-time point processes and application to seismological problems

Giada Adelfio, Marcello Chiodi

Abstract Dealing with data coming from a space-time inhomogeneous process, there is often the need of semiparametric estimates of the conditional intensity function; isotropic or anisotropic multivariate kernel estimates can be used, with windows sizes **h**. The properties of the intensities estimated with this choice of **h** are not always good for specific fields of application; we could try to choose **h** in order to have good predictive properties of the estimated intensity function. Since a direct ML approach cannot be followed, we propose an estimation procedure, computationally intensive, based on the subsequent increments of likelihood obtained adding an observation at time. The first results obtained are very encouraging. Some application in statistical seismology is presented.

1 Introduction

When dealing with data coming from a space-time inhomogeneous process, like seismic data, fire data, or even desease data, there is often the need of obtaining reliable estimates of the conditional intensity function, or of the marginal intensity function. According to the field of application, intensity function can be estimated through some assessed parametric model, where parameters are estimated by Maximum Likelihood method and then intensities (conditional or marginal) are estimated using the parameter estimates. In an exploratory context , some kind of nonparametric estimation is required; we could also have this necessity if we need to assess the adequacy of an estimated parametric model; in some other model, like ETAS model (Ogata, 1988), or in a clustered intensity function (Adelfio and Chiodi, 2007), some component of the spatial intensity function is not explicited and must be estimated from data in a nonparametric way. Often, isotropic or anisotropic kernel estimates

Giada Adelfio, Marcello Chiodi
Department of Statistical and Mathematical Sciences, University of Palermo, viale delle Scienze, ed. 13, 90128, Palermo, Italy; e-mail: adelfio@dssm.unipa.it, chiodi@unipa.it

can be used, e.g. using the Silverman rule to choose the windows sizes $\mathbf{h}$ (Silverman, 1986). When the purpose of the study is the estimation of $\mathbf{h}$, we could try to choose $\mathbf{h}$ in order to have good predictive properties of the estimated intensity function. As it is known, a direct ML approach cannot be followed, unless we use a penalizing function. In the next section a predictive approach to the nonparametric estimation is presented, while in the third section some kind of application to statistical seismology is briefly sketched.

2 Intensity function and predictive likelihood

Suppose we have a general $d-$dimensional closed region, Z^d and that one of the dimension is $t \in T$, the time, or however a dimension with a *meaningful ordering* such that $Z^d = S^{d-1} \times T$. Let $\mathscr{P}$ a random collection of k points in Z^d from time t_1 until the time t_k such that $i < j \Longleftrightarrow t_i < t_j$ and each observation P_i is constitued by: $\mathbf{z}_i^{\mathrm{T}} = \left\{\mathbf{s}_i^{\mathrm{T}}, t_i\right\}$, $i = 1,2,\ldots,k$; the conditional intensity function of the process is:

$$\lambda(\mathbf{z}) = \lambda(\mathbf{s},t|H_t) = \lim_{\Delta t,\Delta\mathbf{s}\to 0} \frac{\mathrm{E}\left[\#(t,t+\Delta t;\mathbf{s},\mathbf{s}+\Delta\mathbf{s}|H_t)\right]}{\Delta t \Delta \mathbf{s}}$$

where H_t is the space-time occurrence history of the process up to time t; Δt and $\Delta\mathbf{s}$ are time and space increments; $\mathrm{E}\left[\#(t,t+\Delta t;\mathbf{s},\mathbf{s}+\Delta\mathbf{s}|H_t)\right]$ is the history-dependent expected number of events occurring in the volume $[t,t+\Delta t]\times[\mathbf{s},\mathbf{s}+\Delta\mathbf{s}]$.

Assuming that θ is a vector of smoothing parameters in a semiparametric context, the log-Likelihood for the point process, given the m observed values $\mathbf{z}_i$, with $m < k$, is (Daley and Vere-Jones, 2003):

$$\log L(\hat{\theta}(H_{t_m}); H_{t_m}) = \sum_{i=1}^{m} \log \lambda(\mathbf{z}_i; \hat{\theta}(H_{t_m})) - \int_{T_0}^{T_{max}} \int_{\Omega_\mathbf{S}} \lambda(\mathbf{z}; \hat{\theta}(H_{t_m}))\, d\mathbf{s}\, dt \qquad (1)$$

where $\Omega_\mathbf{S}$ is the observed space region and $(T_0 - T_{max})$ is the observed period of time and the intensities $\lambda(\cdot)$ depend on unknown parameters θ estimated by $\hat{\theta}(H_{t_m}) \equiv \hat{\theta}(\mathbf{z}_1,\mathbf{z}_2,\ldots,\mathbf{z}_i,\ldots,\mathbf{z}_m)$.

We try to find a trade-off between fitting to observed data and prediction of future data; the context of space-time point processes is different from regression problems, where we can use cross validation tecniques, or from time series context where we can compare the observed value y_{m+1} with an estimated value $\hat{y}_{m+1}$, depending only on the previous m observations. The problem does not arise if we use likelihood computed on different sets of data and with different estimates. We use variation in the likelihood to measure the capability of the observations until t_m, and estimates $\hat{\theta}(H_{t_m})$ computed only on these observations, to give information on the next observation, in the following way. Let:

$$\log L(\hat{\theta}(H_{t_m}); H_{t_{m+1}}) = \sum_{i=1}^{m+1} \log \lambda(\mathbf{z}_i; \hat{\theta}(H_{t_m})) - \int_{T_0}^{t_{m+1}} \int_{\Omega_{\mathbf{S}}} \lambda(\mathbf{z}; \hat{\theta}(H_{t_m}))\, d\mathbf{s}\, dt \quad (2)$$

be the likelihood computed on the first $m+1$ observation but using the estimates based on observation *only until* t_m; e.g. $\lambda(\mathbf{z}; \hat{\theta}(H_{t_m}))$ can be an intensity function computed by an anisotropic kernel method with a multivariate window $\hat{\theta}$ using the first m points. So we use the difference between (2) and (1) to measure the predictive information of the first m observations on the $(m+1)-th$:

$$\delta_l(\hat{\theta}(H_{t_m}); H_{t_{m+1}}) = \log L(\hat{\theta}(H_{t_m}); H_{t_{m+1}}) - \log L(\hat{\theta}(H_{t_m}); H_{t_m}) \quad (3)$$

For the sake of brevity, we report here only essential ideas with few details; a first possibility is to use $\delta_l(\hat{\theta}(H_{t_m}); H_{t_{m+1}})$ to estimate some smoothing parameter θ. In a fashion similar to cross-validation criterion we could choose $\tilde{\theta}(H_{t_m})$ which maximizes a predictive likelihood:

$$FLP_{m_1,m_2}(\hat{\theta}) = \sum_{m=m_1}^{m_2} \delta_l(\hat{\theta}(H_{t_m}); H_{t_{m+1}}): \quad FLP_{m_1,m_2}(\tilde{\theta}) \geq FLP_{m_1,m_2}(\hat{\theta}) \quad \forall \hat{\theta} \in \Theta$$

It is clear that we usually should have $m_2 = k-1$ (that is we stop at the last observation) and maybe $m_1 \approx \frac{k}{2}$ or such that $t_{m_1} - T_0 \approx \frac{T_{max}-T_0}{2}$. Another possible use of the quantities in (3) is for diagnostic purposes, but this aspect will not be introduced in the present paper. Working on these quantities, we obtained estimators (Chiodi and Adelfio, 2008) computationally expensive, which seem to give better kernel estimates of space-time intensity function with respect to classical methods, either using isotropic or anisotropic kernel function. We applied this tecnique to seismological data, although it is capable to be applied in quite different contexts.

3 Examples of space-time inhomogeneous intensity functions in the seismological field

Clustering earthquakes: In many parametric contexts of space-time point processes, intensity functions are specified depending on some specific parameter without specifying all the components. For example (Adelfio et al.,2007), a seismic catalog can be the realization of a clustered inhomogeneous Poisson process, that is obtained supposing that points of the background seismicity come from a space-time Poisson process (spatially inhomogeneous) and that among these there is a number h of *mainshocks* that can generate aftershocks sequences, inhomogeneous both in space and times, and with an intensity also linked to the magnitude of the main event. The intensity function is:

$$\lambda(x,y,t;\theta) = \lambda_t \mu(x,y) + K_0 \sum_{j=1;t_j<t}^{h} g_j(x,y) \frac{\exp[\alpha(m_j - m_0)]}{(t-t_j+c_j)^{p_j}}$$

where $\theta = (\lambda_t, K_0, c_j, p_j, \alpha)$; t_j and m_j are time of the first event and magnitude of the mainshock of the cluster j, $g_j(x,y)$ is the space intensity of the cluster j and $\mu(x,y)$ is the background one; K_0 and λ_t are the weights of the clustered seismicity and of the background one, respectively; c_j and p_j are parameters of the clusters time distributions to be estimated; $g_j(x,y)$ and $\mu(x,y)$ must be estimated in a semi-parametric way.

ETAS model: One of the most important model in statistical seismology is the ETAS model (Ogata,1988), a self-exciting point process describing earthquakes catalogs as a realization of a branching or epidemic-type point process. The conditional intensity function of the ETAS model in a point x,y,t,m is defined by:

$$\lambda(x,y,t,m|\mathscr{H}_t) = J(m)(\mu(x,y) + \sum_{t_j<t} g(t-t_j)f(x-x_j, y-y_j|m_j))$$

$J(m)$ is the magnitude distribution; even if $g(\cdot)$ $f(\cdot)$ are parametrically specified, the spontaneous activity $\mu(x,y)$ must be however estimated from observed data, in a semiparametric way.

Evaluation of seismic gap: a number of statistical models with intensities $\lambda(\cdot;\theta)$ have been proposed for representing the intensity function of earthquakes. The parametric models estimation suffers by many drawbacks, often related to the definition of a reliable mathematical model from the geophysical theory and to the sensitivity of statistical estimates to the composition of the space-time region under study. Many of the disadvantages of the parametric modelling can be avoided by using also nonparametric techniques, such as those presented in this paper. So comparing $\lambda(x,y,t;\hat{\theta})$ obtained with some parametric model with a nonparametric estimate $\tilde{\lambda}(x,y,t)$ in specific region of interest in the domain of x,y,t, we can estimate the so called seismic gap.

Acknowledgements This paper and the related work have been supported by research fund of University of Palermo and by PRIN funds 2006

References

1. Adelfio G., Chiodi M. (2008) Semiparametric estimation of conditional intensity functions for space-time processes. Presented to the scientific meeting of Italian Statistical Society; Cosenza, June 2008.
2. Adelfio G., Chiodi M., De Luca L. and Luzio D. (2006) Nonparametric clustering of seismic events, Data Analysis, Classification and the Forward Search, 397 - 404.
3. Daley D.J. and Vere-Jones D. (2003) An introduction to the theory of point processes, New York: Springer-Verlag, second edition.
4. Ogata Y. (1988) Statistical models for earthquake occurrences and residual analysis for point processes, Journal of the American Statistical Association, 83, 401, 9 - 27.
5. Silverman B.W. (1986) Density Estimation for Statistics and Data Analysis, Chapman and Hall, London.

Modeling spatial dependence in finite mixture models for disease mapping

Marco Alfó, Luciano Nieddu, and Donatella Vicari

Abstract A vast literature has recently been concerned with the analysis of variation in multivariate counts recorded across geographical areas with the aim of detecting clusters of regions with homogeneous behavior. Most of the modeling approaches have been discussed for the univariate case and only very recently spatial models have been extended to predict more than one outcome simultaneously. We extend standard finite mixture models to the analysis of multiple, spatially correlated, counts. Dependence among outcomes is modeled using a set of of correlated random effects, while the spatial structure is captured by the use of a Gibbs representation for the prior probabilities of component membership.
Keywords Multivariate Counts, Finite Mixtures, Gibbs distribution, Mode-Field Approximation.

1 Foreword

Mapping geographical variation in disease or mortality rates to generate hypotheses about possible *causes* of *variation* in relative risks and to identify potential clusters of neighboring regions characterized by homogeneous Relative Risks is a common task in epidemiological studies. Statistical models aim at providing reliable estimates of relative risks by filtering the extra-Poisson variation due to overdispersion and spatial correlation.

Finite mixture models provide a simple and effective tool to detect clusters of geographical units characterized by homogeneity in the estimated relative risks

Marco Alfó, Donatella Vicari
Dipartimento di Statistica, Probabilità e Statistiche Applicate, Sapienza - Università di Roma, P.le A. Moro, 5 - 00185 Roma, Italy e-mail: {marco.alfo;donatella.vicari}@uniroma1.it

Luciano Nieddu
Facoltá di Economia, Libera Universitá "S. Pio V", Via Delle Sette Chiese, 139 - 00145 Roma, Italy e-mail: l.nieddu@gmail.com

while being substantially different from the other clusters; see e.g. [7] or [4]. However, they rely on the assumption of independence between adjacent areas, although work has been done to account for spatial dependence (see e.g. [3], [2]).

A general approach based on finite mixture models with spatial constraints has been proposed by [5], where prior probabilities have been modeled through a hidden Markov Random Field using a Potts representation (see e.g. [6]). The model is exploited in a fully Bayesian context, assuming that the interaction parameter of the Potts representation is fixed across the whole analyzed region.

In this paper, we model prior probabilities using an inhomogeneous hidden Markov Random Field with a spatially varying interaction parameter; see [1] for a discussion on the potential negative effect of fixed interaction parameters in the context of image restoration. The choice of a spatial-dependent interaction parameter prevents from over-smoothing small-scale areas, while minimizing noise in large-scale areas; spatial information can have a strong but varying effect when the intensity process variance substantially changes across the analyzed area.

2 Mapping Multivariate Counts

We start by assuming that the analyzed region S can be partitioned into n areas: counts $o_{il}, i = 1,\ldots,n, l = 1,\ldots,L$ of observed cases for L diseases are recorded for each area. Following the usual notation for multivariate data, let $\mathbf{o}_i = (o_{i1}, o_{i2}, ..., o_{iL})^{\mathsf{T}}$ denote the vector of observed counts for the $i-th$ area in the analyzed region, $i = 1, ..., n$. Let z_{ik} denote the component indicator for the k-th component; thus, $\pi_{ik} = \Pr(z_{ik} = 1)$, and let N_i denote the neighborhood for the $i-th$ area while $\lambda_k = (\lambda_{k1}, ..., \lambda_{kL})$ denote a set of relative risk parameters for the k-th component of the finite mixture. The pseudolikelihood function can be rewritten as:

$$L(\cdot) = \prod_{i=1}^{n} \left\{ \sum_{k=1}^{K} f(\mathbf{o}_i|\lambda_k)\tau_{ik} \right\} \tag{1}$$

where $f(\mathbf{o}_i|\lambda_k)$ is the k-th class specific density (Poisson) and $\tau_{ik} = \Pr(z_{ik} = 1 \mid \mathbf{z}_j, j \in N_i) = \Pr(\lambda_i = \lambda_k \mid \lambda_j, j \in N_i), k = 1,\ldots K$ represent the join (conditional) probabilities of component membership, defined through a give Gibbs representation. We use a Strauss [8] representation to model the componer membership process.

3 Parameter Estimation

The component labels are unobservable and therefore have to be conside as missing data and this naturally leads to the use of the EM algorithm. T hypothetical space of the complete data is given by $(\mathbf{o}, \mathbf{z}) = (\mathbf{o}_i, \mathbf{z}_i, i \in S)$. Start

from a multinomial distribution for the $\mathbf{z}_i$s, the log-pseudolikelihood for the complete data can be written as:

$$\ell_c(\cdot) \propto \sum_{i \in S} \sum_{k=1}^{K} z_{ik} \{\log(\tau_{ik}) + \log f_{ik}\}, \tag{2}$$

where $f_{ik} = f(\mathbf{o}_i \mid \lambda_k) = \prod_l f(\mathbf{o}_{il} \mid \lambda_{lk})$. In the E-step of the EM algorithm, we define the log pseudo-likelihood for *observed* data by taking the expectation of the log-pseudolikelihood for *complete* data over the unobservable component label vector $\mathbf{z}_i$ given the observed data $\mathbf{o}$ and the current PML estimates of model parameters, say $\lambda^{(t)}$. The conditional expectation of the complete log-pseudolikelihood given the observed data $\mathbf{o}$ is expressed by the function:

$$Q^{(t)}(\cdot) \propto \sum_{i \in S} \sum_{k=1}^{K} w_{ik}^{(t)} \{\log f_{ik} + \log(\tau_{ik})\} \tag{3}$$

Maximizing $Q^{(t)}(\cdot)$, we obtain the following PML estimates for the parameters of the univariate Poisson distributions in the k-th class:

$$\hat{\lambda}_{kl}^{(t)} = \frac{\sum_{i \in S} w_{ik}^{(t)} o_{il}}{\sum_{i \in S} w_{ik}^{(t)} E_{il}} \quad k = 1, \ldots, K, \; l = 1, \ldots, L \tag{4}$$

which are well known results from ML in finite mixtures. The interaction parameters of the Gibbs distribution that account for spatial dependence, say β_k, give info on the strength of the process describing how the cluster membership of area k is influenced by the memberships of the neighboring areas. They come from the solutions of the following M-step equations:

$$\frac{\partial Q^{(t)}}{\partial \beta_k} = \frac{\partial}{\partial \beta_k} \sum_{i \in S} \sum_{k=1}^{K} w_{ik}^{(t)} \log(\tau_{ik}) = 0, \quad k = 1, \ldots, K-1 \tag{5}$$

which represent weighted sums of pseudo-likelihood equations for the Strauss automodel with weights given by $w_{ik}^{(t)}$. At the t-th step of the EM algorithm the current parameter estimates are $\tilde{\mathbf{z}}_i^{(t-1)}$, $\lambda^{(t-1)}$ and $\beta^{(t-1)}$, where $\tilde{\mathbf{z}}_i^{(t-1)}$ represents the estimated field at the $(t-1)$-th step of the algorithm and is considered fixed. Let us define (by sequential updating):

$$\tilde{\mathbf{z}}_i^{(t)} = \text{argmax}_{\mathbf{z}_i} \Pr(\mathbf{z}_i \mid \mathbf{o}_i, \beta^{(t-1)}) \simeq \text{argmax}_{\mathbf{z}_i} \Pr(\mathbf{z}_i \mid \tilde{\mathbf{z}}_j^{(t-1)}, j \in N_i, \mathbf{o}_i, \beta^{(t-1)}). \tag{6}$$

Given the current field approximation, say $\tilde{\mathbf{z}}^{(t)}$ (which is assumed fixed during the M-step), the modified EM algorithm is as follows:

E-step $\tilde{w}_{ik}^{(t)} \propto f(\mathbf{o}_i \mid \mathbf{z}_i, \lambda_k^{(t-1)}) \Pr(\mathbf{z}_i \mid \tilde{\mathbf{z}}_j^{(t)}, j \in N_i, \beta_k^{(t-1)})$

M-step $\lambda_k^{(t)} = \text{argmax}_{\lambda_h} \sum_h \tilde{w}_{ih}^{(t)} \log f_{ih}, \quad \beta_k^{(t)} = \text{argmax}_h \sum_h \tilde{w}_{ih}^{(t)} \log[\Pr(\mathbf{z}_i \mid \beta_h)]$

where the term $\Pr(\mathbf{z}_i \mid \beta_h)$ is approximated by $\Pr(\mathbf{z}_i \mid \tilde{\mathbf{z}}_j^{(t)}, j \in N_i, \beta_h)$. In the E-step the 2nd right-hand side term must be computed using the new configuration $\mathbf{z}^{(t)}$ and the *old* parameter vector $\beta^{(t-1)}$. The Q function is therefore Mode-Field approximated by:

$$\begin{aligned} Q_{MF}^{(t)}(\cdot) &\propto \sum_{i,k} \tilde{w}_{ik}^{(t)} \log\left[f(\mathbf{o}_i \mid \lambda_k) \Pr(\mathbf{z}_i \mid \tilde{\mathbf{z}}_j^{(t)}, j \in N_i, \beta) \right] = \\ &= \sum_{i,k} \tilde{w}_{ik}^{(t)} \left\{ \log f_{ik} + \log(\tau_{ik}^{(t)}) \right\} \end{aligned} \tag{7}$$

Component specific parameters for the Poisson densities are estimated according to (4) and the M-step is broken down into two substeps. The first concerns the parameters of the Strauss model once the posterior probability field has been approximated, while the second corresponds to estimation of parameters for (component-specific) Poisson densities. A Newton-type ML algorithm is used to obtain the solutions of equation (5). Solving these equations for given weights and updating the weights for given parameter estimates defines an EM algorithm. The E and M steps are alternated repeatedly until convergence, which is obtained with a sequence of approximated likelihood values which is bounded from above.

The proposed approach will be illustrated using two real datasets.

References

[1] Aykroyd, R. G., and Zimeras, S. (1999), Inhomogeneous Prior Models for Image Reconstruction. *Journal of the American Statistical Association*, 94: 934–946.

[2] Alfò, M. and Vitiello, C. (2003). Finite mixture models fot the analysis of geographical variation in disease rates. *Statistical Methods and Applications*, 12, 93–108.

[3] Biggeri, A., Dreassi, E., Lagazio, C., and Böhning, D. (2002). A transitional non-parametric maximum pseudo-likelihood estimator for disease mapping. *Computational Statistics and Data Analysis*, 41:617–629.

[4] Böhning, D., Dietz, E., and Schlattmann, P. (2000). Space-time mixture modelling of public health data. *Statistics in Medicine*, 19:2333–2344.

[5] Green, P.J., Richardson, S. (2002). Hidden Markov models and disease mapping. *Journal of the American Statistical Association*, 97:1–16.

[6] Kindermann, R. and Snell, J. L. (1999). *Markov Random Fields and Their Applications*, volume 1. American Mathematical Society, Providence, Rhode Island.

[7] Schlattmann, P. and Bohning, D. (1993). Mixture models and disease mapping. *Statistics in Medicine*, 12:943–950.

[8] Strauss, D. J. (1977). Clustering on coloured lattices. *Journal of Applied Probability*, 14:135–143.

An Enhancement of the Plaid Model Algorithm

Angelo M. Mineo and Luigi Augugliaro

Abstract Microarrays have become a standard tool for studying gene functions. For example, we can investigate if a subset of genes shows a coherent expression pattern under different conditions. The plaid model, a model-based biclustering method, can be used to incorporate the addiction structure used for the microarray experiment. In this paper we describe an enhancement for the plaid model algorithm based on the theory of the false discovery rate.

1 Introduction

There has been considerable recent interest in the analysis of microarray experiments. A typical microarray experiment investigates thousands of genes, recording their expression level over tens of samples. A *bicluster* identifies a group of genes and an associated group of samples on which the genes are characterized by a similar expression level. This fact may indicate a common biological function. Several clustering methods have been developed in recent years in order to identify a bicluster, such as gene-shaving [2], EMMIX-GENE [5], EMMIX-WIRE [7], spectral biclustering [3] and the plaid model [4], among the others. The plaid model is a model-based clustering method that can be used to study structured microarray experiments, for this reason it is usually preferred to the other methods. Aim of this paper is to present an enhancement of the plaid model algorithm proposed in [9], in order to reduce the uncertainty related to the parameters used in the pruning step to remove ill-fitted genes and samples. To increase the interpretation and the accuracy

Angelo M. Mineo
Dipartimento di Scienze Statistiche e Matematiche University of Palermo, Viale delle Scienze, 90128 Palermo, e-mail: elio.mineo@dssm.unipa.it

Luigi Augugliaro
Dipartimento di Scienze Statistiche e Matematiche University of Palermo, Viale delle Scienze, 90128 Palermo, e-mail: Augugliaro@dssm.unipa.it