Computational Statistics and Data Analysis 71 (2014) 14–29

Contents lists available at ScienceDirect

Computational Statistics and Data Analysis

journal homepage: www.elsevier.com/locate/csda

Clustering longitudinal profiles using P-splines and mixed effects models applied to time-course gene expression data

N. Coffey [a,*], J. Hinde [b], E. Holian [b]

[a] School of Mathematical Sciences/Systems Biology Ireland, University College Dublin, Ireland

[b] School of Mathematics, Statistics and Applied Mathematics, National University of Ireland, Galway, Ireland

ARTICLE INFO

Article history:
Received 29 June 2012
Received in revised form 6 April 2013
Accepted 6 April 2013
Available online 12 April 2013

Keywords:
Longitudinal profiles
Time-course gene expression
Clustering
Mixed effects model
Finite mixture model

ABSTRACT

Longitudinal data is becoming increasingly common and various methods have been developed to analyze this type of data. Profiles from time-course gene expression studies, where cluster analysis plays an important role to identify groups of co-expressed genes over time, are investigated. A number of procedures have been used to cluster time-course gene expression data, however there are many limitations to the techniques previously described. An alternative approach is proposed, which aims to alleviate some of these limitations. The method exploits the connection between the linear mixed effects model and P-spline smoothing to simultaneously smooth the gene expression data to remove any measurement error/noise and cluster the expression profiles using finite mixtures of mixed effects models. This approach has a number of advantages, including decreased computation time and ease of implementation in standard software packages.

1. Introduction

Multivariate clustering methods such as *k*-means clustering (Hartigan and Wong, 1978), hierarchical clustering (Eisen et al., 1998; Spellman et al., 1998), clustering on self-organizing maps (Kohonen, 1997; Tamayo et al., 1999), finite mixture models (Fraley and Raftery, 2002; McLachlan et al., 2002, 2003, 2006), fuzzy *c*-means clustering (Futschik and Carlisle, 2005) and tight clustering (Tseng and Wong, 2005) have been useful to reduce the dimensionality of gene expression data and identify groups of co-expressed genes and consequently co-regulated genes (since genes with similar expression profiles tend to be controlled by the same regulatory mechanisms). See Thalamuthu et al. (2006) and Hand and Heard (2005) for a full discussion and comparison of numerous clustering techniques. Time-course gene expression studies involve measuring the expression levels of thousands of genes repeatedly through time and result in extremely high-dimensional data. The methods mentioned above treat the sequence of measurements for each gene as a vector of distinct points and thus an arbitrary permutation of the elements of the sequence will not affect the clustering results. However, the time-ordering of the data and the corresponding clusters obtained is an important consideration in time-course gene expression studies. In addition, time-course gene expression data exhibit problems such as missing values, unequal sampling times and/or large measurement errors. Many of the techniques mentioned above have difficulties handling missing values, require identical sampling times for all genes or fail to account for the correlation between measurements made on the same gene over time. This has led to the development of techniques such as CAGED (Ramoni et al., 2002), Hidden Markov Models (Schliep et al., 2005), Bayesian mixture models (Wakefield et al., 2003), mixtures of linear mixed effects models (Celeux et al., 2005; Ng et al., 2006; Qin and Self, 2006; Nueda et al., 2007), clustering based on shape similarity (Hestilow and Huang, 2009) and

* Corresponding author.
E-mail address: norma.coffey@ucd.ie (N. Coffey).

0167-9473/$ – see front matter © 2013 Elsevier B.V. All rights reserved.
http://dx.doi.org/10.1016/j.csda.2013.04.001

clustering of time-course data using SOMs (Chen, 2009). However, these techniques do not facilitate the removal of noise from the measured data thus ignoring any smoothness that may be evident in the gene expression profiles. As a result, curve-based clustering methods have been employed to cluster time-course gene expression data. Such methods assume that gene expression over time is a continuous process and thus can be represented by a continuous smooth curve or function. Treating the expression profiles as continuous functions ensures that missing values and irregularly sampled data can be handled appropriately, measurement error can be removed using smoothing techniques, and correlation between measurements made over time on the same gene can also be accounted for.

Some of the earliest examples of curve-based clustering in time-course gene expression analyses are described in Bar-Joseph et al. (2003), Luan and Li (2003) and James and Sugar (2003) who used linear combinations of basis functions (e.g. cubic splines, B-splines, etc.) to model the mean expression profile in each cluster and cluster the estimated basis function coefficients. Other authors such as Kim et al. (2008) used a linear combination of Fourier basis functions to represent the expression profiles for clustering and Kim and Kim (2008) clustered based on the derivative coefficients of a Fourier series. Song et al. (2007) determined the functional principal components (FPCs) using basis function expansions and clustered based on the FPC scores, while Leng and Müller (2006) represented the expression profiles using a linear combination of FPCs and performed functional logistic regression of the scores to classify the expression profiles into groups. However, to estimate the cluster mean curves the methods used in these papers required choosing K, the number of basis functions, and the knots, the join points for these functions. Choosing an optimal value for the number of basis functions (or equivalently knots) is a complex problem and it is difficult to control the degree of smoothing applied to the data. Using too many basis functions results in over-fitting and lack of smoothness in the estimated expression profiles, while using too few under-fits and over-smooths the data. One solution is to use smoothing spline regression, where a knot is placed at each unique time point and the resulting over-fitting is controlled by adding a penalty term to the optimization criterion. In the statistics literature, a common approach is to penalize the curvature (i.e. the integrated squared second derivative) of the expression curve estimate (see Wahba, 1990; Green and Silverman, 1994; Ramsay and Silverman, 2005, for full details). The trade-off between fit to the data and smoothness is controlled by a smoothing parameter λ and an optimal value for λ can be chosen using methods such as cross-validation (CV), or generalized cross-validation (GCV). In time-course gene expression studies, Ma et al. (2006) used smoothing spline regression to cluster gene expression profiles and called their method SSClust. Ma et al. (2008) extended this to a Bayesian setting and Ma and Zhong (2008) included additional covariates in the clustering algorithm. Déjean et al. (2007) used smoothing spline regression to estimate the derivatives of the gene expression profiles before clustering based on the principal component scores of the discretized derivative functions. Tarpey (2007) and Kayano et al. (2010) clustered time-course data by transforming non-orthogonal basis functions to orthogonal basis functions (using singular value decomposition (SVD) and the Cholesky decomposition respectively) in conjunction with k-means and SOMs respectively, to ensure that clustering the basis function coefficients was equivalent to clustering the raw data. A major drawback of clustering using smoothing spline regression is the high computational overhead associated with these methods. Since smoothing splines usually use the same number of basis functions as unique time points, and the optimization criterion requires numerically evaluating the integral associated with the penalty term, computations can become cumbersome as the sample size and/or the number of unique observation times increases. In addition, smoothing spline clustering requires choosing an optimal value for λ for each cluster, which also increases the computational burden required to fit the model.

In contrast, this paper uses penalized spline (P-spline) smoothing, as discussed by Eilers and Marx (1996) and Ruppert et al. (2003), to model the gene expression profiles in each cluster. P-spline smoothing is a low-rank smoothing method that requires using a relatively large number, K, of basis functions, but still less than the number of unique time points encountered. As stated previously, when the number of basis functions is large, unconstrained estimates of the mean expression profile leads to over-fitting of the data and a fit that includes excessive amounts of noise. P-spline smoothing retains all of the basis functions, but constrains their influence using a discrete penalty on the estimated coefficients. Again the trade-off between fit to the data and smoothness is controlled by the smoothing parameter λ. Representing the smoothing problem using P-splines reduces the dimensionality of the problem thus reducing the computational burden. According to Ruppert (2002), P-splines are also relatively insensitive to the number of basis functions selected once enough basis functions are used. A strategy for choosing the number and location of the knots is given in Section 2.1. P-splines are also easy to compute since the penalty is discrete, rather than a continuous integral as with smoothing splines. Furthermore, representing the penalized smoothing problem as a linear mixed effects model has numerous additional advantages. In a clustering context, writing the smoothing problem as a mixed effects model provides a framework for simultaneously determining a smooth estimate of the mean expression profile in each cluster, determining estimates of the gene-specific expression profiles within a cluster through the use of additional random effects (e.g. a random intercept for each gene), including additional covariates, and clustering expression profiles using mixtures of mixed effects models. An optimal value for the smoothing parameter λ can be chosen automatically via restricted maximum likelihood (REML) further reducing the computational overhead, and the model fitting can be implemented using standard statistical software packages. This paper uses the linear mixed effects model representation of P-spline smoothing to cluster time-course gene expression profiles. While the methodology is presented in the context of time-course gene expression data, it can be applied to any longitudinal dataset where cluster analysis is required.

The remainder of the paper is outlined as follows. Section 2 describes how to represent the raw time-course gene expression data for a particular gene as a smooth curve and discusses how to implement the P-spline smoothing problem as