Computational Statistics and Data Analysis 87 (2015) 73–83

Contents lists available at ScienceDirect

Computational Statistics and Data Analysis

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

A Bayesian hierarchical model for estimating and partitioning Bernstein polynomial density functions[☆]

Charlotte C. Gard[a,*], Elizabeth R. Brown[b]

[a] Department of Economics, Applied Statistics and International Business, New Mexico State University, Las Cruces, NM, USA

[b] Vaccine and Infectious Disease and Public Health Sciences Divisions, Fred Hutchinson Cancer Research Center, Department of Biostatistics, University of Washington, Seattle, WA, USA

ARTICLE INFO

Article history:
Received 26 March 2014
Received in revised form 14 January 2015
Accepted 26 January 2015
Available online 3 February 2015

Keywords:
Bayesian hierarchical models
Bernstein polynomials
Birth–death MCMC
Density estimation
Mixture models
Partitioning

ABSTRACT

A Bayesian hierarchical model for simultaneously estimating and partitioning probability density functions is presented. Individual density functions are flexibly modeled using Bernstein densities, which are mixtures of beta densities whose parameters depend only on the number of mixture components. A prior distribution is placed on the number of mixture components, and the mixture weights are expressed as increments of a distribution function G. A Dirichlet process prior is placed on G and the parameters of the Dirichlet process, the baseline distribution and the precision parameter, are treated as random. A mixture of a product of beta densities is used to partition subjects into groups, with subjects in the same group sharing information via a common baseline distribution. Inference is carried out using Markov chain Monte Carlo. A computing algorithm based on the constructive definition of the Dirichlet process is offered, for both a fixed number of groups and an unknown number of groups. When the number of groups is unknown, a birth–death algorithm is used to make inference regarding the number of groups. The model is demonstrated using radiologist-specific distributions of percent mammographic density.

1. Introduction

Increasingly, it is common to observe a distribution of measurements for each of a number of subjects in a population at a single point in time. In imaging studies, for example, histograms of pixel gray-level, or brightness, values are often generated from digital or digitized images. Distributions of relative fluorescence or light scatter intensity are products of flow cytometry. It is often of interest to flexibly estimate the probability density function for each subject based on his or her observed distribution and to identify groups of subjects who have similar density functions. Such groupings may provide insight into data generating mechanisms or be of interest for further modeling. Straightforward analytical methods to jointly estimate probability density functions for multiple subjects in a population nonparametrically and to simultaneously partition density functions into groups are not currently available.

Our interest in this statistical problem is motivated by an analysis of mammographic breast density. Breast density measures the amount of radiographically dense tissue in a woman's breast and is a well-known risk factor for breast cancer

[☆] Supplementary material related to this article is available online at http://dx.doi.org/10.1016/j.csda.2015.01.016.

[*] Correspondence to: MSC 3CQ, New Mexico State University, P.O. Box 30001, Las Cruces, NM 88003-8001, USA. Tel.: +1 575 646 5500; fax: +1 575 646 1915.

E-mail address: cgard@nmsu.edu (C.C. Gard).

http://dx.doi.org/10.1016/j.csda.2015.01.016
0167-9473/© 2015 Elsevier B.V. All rights reserved.

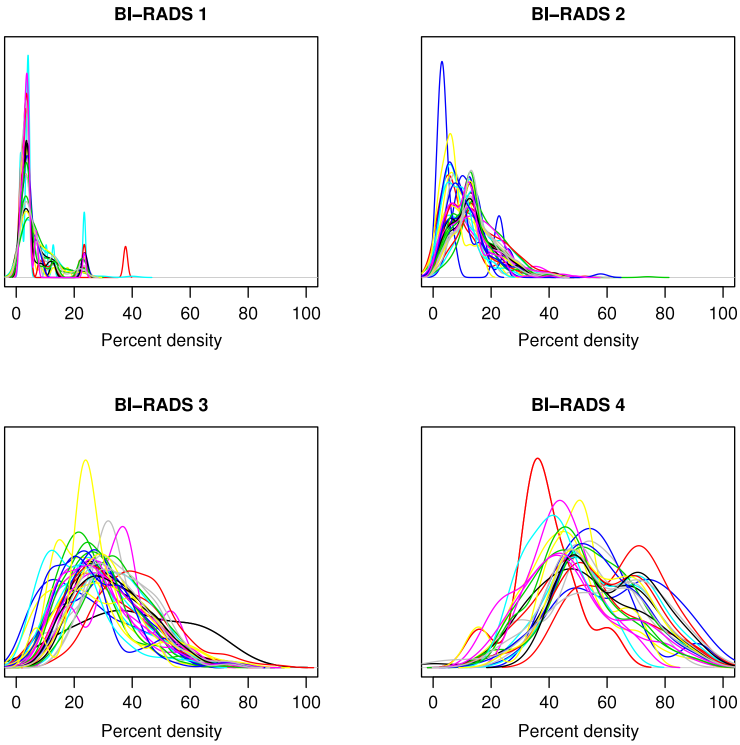

Fig. 1. Radiologist-specific kernel density estimates for percent density data for mammograms interpreted as BI-RADS density 1, 2, 3, and 4.

(McCormack and dos Santos Silva, 2006). In clinical settings, radiologists classify breast density according to the four-category Breast Imaging Reporting and Data System (BI-RADS) lexicon (American College of Radiology, 2003), with density 1 representing breasts that are "almost entirely fat, <25% dense", density 2 breasts with "scattered fibroglandular densities, 25%–50% dense", density 3 breasts that are "heterogeneously dense, 51%–75% dense", and density 4 breasts that are "extremely dense, >75% dense". Percent mammographic density, which is often ascertained in research settings using computer-assisted methods, measures the percentage of the total area of a woman's breast that is occupied by dense tissue and, as such, takes values between 0 and 100. Our data consist of 14,414 BI-RADS density interpretations by 49 radiologists and corresponding percent density measurements. Fig. 1 plots radiologist-specific kernel density estimates for the percent density data for mammograms interpreted as BI-RADS density 1, 2, 3, and 4, for radiologists with ten or more percent density measurements corresponding to a particular BI-RADS density category. It is of interest to model radiologist-specific distributions of percent density within categories of BI-RADS density and to identify radiologists who are similar in their assignment of BI-RADS density relative to underlying percent density (i.e., radiologists whose distributions of percent density are similar).

In our example, we could envision using finite mixture models to model subject-specific (radiologist-specific, in our case) density functions, like those in Fig. 1, then using a heuristic method such as k-means (MacQueen, 1967) to partition density functions into groups based on the estimated model parameters. However, we would prefer an approach that simultaneously estimates and partitions density functions and, in doing so, shares information across subjects. Zhou and Wakefield (2006) presented a Bayesian hierarchical model for simultaneously estimating and partitioning time course gene expression data. Gene trajectories were modeled using a first-order random walk model and partitioned into groups based on the model parameters. The authors treated the number of groups as random, using a birth–death algorithm (Stephens, 2000) for computation. We could take a similar approach, using mixture models to model subject-specific density functions then partitioning subjects into a random number of groups based on the parameters of the mixture models. However, we would prefer to partition subjects based on their entire density functions rather than on features of their density functions. Rodriguez et al. (2008) proposed a nested Dirichlet process mixture model for estimating and partitioning probability distributions. The model allows for both the clustering of subjects with similar distributions and the clustering of observations within subjects, the latter of which is not of interest in our setting. As noted by Rodriguez et al. (2008), the Dirichlet process is conservative in adding new clusters, favoring allocation to existing clusters as the number of observations increases. We would prefer that cluster sizes be balanced, a priori.

We present a Bayesian hierarchical model for simultaneously estimating and partitioning probability density functions. Our approach utilizes finite mixture models, which provide both a means for flexible density estimation and a means for exploring data for group structure. At the individual level, we use Bernstein densities (Petrone, 1999a,b; Petrone and