Mixture of Polynomials Probability Distributions for Grouped Sample Data

Barry R. Cobb

Missouri State University, Department of Management,
Springfield, Missouri, USA
BarryCobb@MissouriState.edu

Abstract. This paper describes techniques for developing a mixture of polynomials (MOP) probability distribution from a frequency distribution (also termed grouped data) summarized from a large dataset. To accomplish this task, a temporary dataset is produced from the grouped data and the parameters for the MOP function are estimated using a Bspline interpolation technique. Guidance is provided regarding the composition of the temporary dataset, and the selection of split points and order of the MOP approximation. Good results are obtained when using grouped data as compared to the underlying dataset, and this can be a major advantage when using a decision support system to obtain information for estimating probability density functions for random variables of interest.

Keywords: Bayesian information criterion, B-spline interpolation, frequency distribution, grouped data, mixture of polynomials.

1 Introduction

This paper describes the construction of a mixture of polynomials (MOP) probability density function (PDF) from a frequency distribution developed from sample data, also termed here *grouped data*. In general, the MOP function can provide a method for approximating a PDF from data in a flexible form that can be readily manipulated for mathematical calculations.

Kernel density estimation is a well-known method for assigning a PDF to empirical data [1]. However, the functional form of many kernel density estimators is not amenable to use in probabilistic graphical models, and the sample data must be retained to reproduce the density function. To construct a hybrid Bayesian network or influence diagram with continuous variables that are not exclusively Gaussian, or to build models that cannot be represented in the conditional linear Gaussian framework, a functional form that permits closed-form addition, multiplication, and integration, and a form that maintains results in the same class of functions is desirable.

Mixtures of truncated exponentials [2], mixtures of polynomials [3], and mixtures of truncated basis functions [4] are methods suggested for overcoming the

L.C. van der Gaag and A.J. Feelders (Eds.): PGM 2014, LNAI 8754, pp. 129–144, 2014.

integration problem in hybrid Bayesian network models. To estimate PDFs accurately for such models, several methods have been developed to find parameters for PDFs of continuous random variables from data [5,6,7,8].

In many applications in practice, including a supply chain management problem discussed in a related working paper [9], raw data to estimate a PDF for a continuous random variable of interest is not readily available. Managing and transferring a dataset that includes all of the empirical data can become difficult because of its size. Furthermore, the data may have to be extracted directly from a database, which may be a more difficult task than simply accessing a report from a decision support system that includes frequency distributions calculated for grouped data as part of its standard output.

This paper examines whether an adequate mixture of polynomials PDF can be estimated from a frequency distribution of grouped data without resorting to accessing the entire dataset. The approach is to create a temporary dataset and use the B-spline interpolation approach suggested by López-Cruz et al. [10]. The issues that arise when using this approach are the number of values to include in the temporary dataset, the split points to use for the MOP functions, and the order of the polynomials in the approximation. We examine each of these issues in examples where we estimate MOP density functions from known distributions using a full dataset of sample data and grouped data summarized from the full dataset. The issue of using uniform split points versus equal probability split points with the B-spline technique is examined, and this discussion is relevant regardless of whether the full dataset or summary grouped data is used to create an MOP approximation. In general, we find that acceptable MOP approximations can be created from grouped data.

The remainder of the paper is structured as follows. The next section introduces notation and definitions used throughout the paper. Section 3 reviews a B-spline technique for estimating MOP functions from data developed by López-Cruz et al. [10]. Section 4 describes the process of estimating an MOP from grouped data using the B-spline method. Section 5 applies the technique to two examples where both grouped data and the simulated underlying dataset are available to allow comparison of results. Section 6 concludes the paper.

2 Notation and Definitions

This section reviews definitions and notation used throughout the paper.

2.1 Grouped Data

We suppose that an unobserved dataset $\mathcal{D} = \{x_1, \ldots, x_{\mathcal{N}}\}$ is summarized into $\mathcal{K}$ groups. A series of $\mathcal{K}+1$ split points, $\mathbf{s} = \{\mathcal{S}_0, \mathcal{S}_1, \ldots, \mathcal{S}_{\mathcal{K}-1}, \infty\}$, defines the groups. The split points are defined without an upper bound because we often encounter cases in practice where the last group is not explicitly bounded, although if there is a finite upper bound this can be assigned as $\mathcal{S}_{\mathcal{K}}$.

A specific data point x_i is classified into group j if $\mathcal{S}_{j-1} \leq x_i < \mathcal{S}_j$. The frequency of observations in each group are denoted by $\mathbf{f} = \{\mathcal{F}_1, \ldots, \mathcal{F}_{\mathcal{K}}\}$ and