Smoothing Ordered Categorical Data

Daehak Kim

Dept. of statistical information, Catholic University of Taegu-Hyosung, Kyungsan 712-702

순서화된 범주형 자료의 평활

김 대 학

대구효성가톨릭대학교 정보통계학과

ABSTRACT - Smoothing problems for continuous data are abundant in modern statistical areas. In kernel density estimation and nonparametric regression, bandwidth selection is crucial problems than any other factors in nonparametric smoothing. In this paper, we extend the concept of smoothing techniques to the categorical data. It is interesting topics and will be a useful method for the categorical data analysis.

Key words □ nonparametric regression, smoothing, bandwidth, kernel, categorical data, multinomial distribution, sparse table

1. Smoothing and Kernel regression

The most widely used general statistical procedure is linear regression. Regression models are powerful tools for modeling a target variable \( y \) as a function of a set of predictors \( x \), allowing prediction for future values of \( y \) and construction of tests and confidence interval estimates for predictions.

Consider simple linear regression model

\[
y_i = \beta_0 + \beta_1 x_i + \varepsilon_i, \quad i = 1, 2, \ldots, n
\]

(1.1)

with the errors \( \varepsilon \) usually taken to be independent and identically distributed with zero mean and variance \( \sigma^2 \).

A more general alternative to (1.1) is the nonparametric regression model

\[
y_i = m(x_i) + \varepsilon_i
\]

(1.2)

The regression curve \( m(x) \) is the conditional expectation \( m(x) = E(Y|X = x) \). The model (1.2) removes the parametric restriction on \( m(x) \) and allows alternative structure to come through. Nadaraya(1964) and Watson(1964) independently proposed the kernel regression estimator
\[
\hat{m}(x) = \frac{\sum y_i}{h} \frac{\sum K(\frac{x-x_i}{h})}{\sum K(\frac{x-x_i}{h})}
\] (1.3)

This estimator is most natural for data using a random design.

2. Smoothing Categorical Data

Kernel estimator referred to continuous data and estimation of smooth unknown function. This is reasonable, since smoothness and continuity would seem to be naturally linked to each other.

Consider now a one dimensional categorical variable, where the sample space consists of \( K \) cells and \( n_i \) observations fall in the \( i \)th cell, \( i = 1, 2, \ldots, K \) with \( \sum n_i = n \) the sample size. The vector \( \mathbf{p} = (p_1, p_2, \ldots, p_K) \) represents the probability of an observation falling in a given cell. Simonoff (1996) gave a detailed account of smoothing methods for categorical data.

A categorical variable where the categories do not have any natural ordering is called a nominal variable. For such data smoothing is not very helpful, since it is very difficult to define how close two categories are. A categorical variable where the categories do have a natural ordering, called ordinal variable is very different matter. Such a variable can arise as a discretization of an underlying continuous variable or inherently discrete, but ordered set of categories. For such a variable, smoothing makes sense, as it is likely that the number of observations that fall in a particular cell provides information about the probability of falling in nearby cells as well. For example, if the variables represents a discretization of a continuous variable with smooth density \( f \), the probability vector \( \mathbf{p} \) also will reflect that smoothness, with \( p_i \) being close to \( p_j \) for \( i \) close to \( j \).

The laws of large numbers states that \( \bar{p}_i = n_i/n \) is consistent estimator of \( p_i \) as long as \( n \to \infty \) as \( n \to \infty \). From a practical point of view, this corresponds to the sample size \( n \) being large compared with the number of cells \( K \).

In many situations, however, such as multidimensional tables, the number of cells is close to the number of observations, resulting in many small cell count. Such a table of counts is called a sparse table. For such tables, \( \bar{p}_i \) is not a good estimator of \( p_i \), as the usual asymptotic approximations do not apply.

3. Smoothing Sparse Multinomials

Consider one dimensional table \( [n_1, n_2, \ldots, n_K] \). The standard model for this random variable is a multinomial distribution with sample size \( n \) and probability vector \( \mathbf{p} \) with log likelihood

\[
\sum n_i \log p_i, \text{ with } \sum p_i = 1 \quad (3.1)
\]

For this distribution \( E(n_i) = np_i \) and hence \( E(\bar{p}_i) = p_i \). It is helpful to think of vector \( \mathbf{p} \) as being generated from an underlying density \( f \) on \([0, 1]\) through the relation

\[
p_i = \int_{(i-1)/K}^{i/K} f(u) \, du \quad (3.2)
\]