A New Feature Extraction Method for Very Non-Normal Data: Analysis of Multivariate Catch and Bycatch Data from Purse-seine Tuna Fisheries

Mihoko Minami[1] and Cleridy E. Lennert-Cody[2]

[1]The Institute of Statistical Mathematics/The Graduate University for Advanced Studies, JAPAN
[2]The Inter-American Tropical Tuna Commission, USA

We propose a new feature extraction method for very non-normal data. Our proposed method extends principle component analysis in the same manner as generalized linear model extends ordinary linear regression model. As an example, we analyze multivariate catch and bycatch data from purse-seine tuna fisheries in eastern Pacific ocean. The objective of analysis is to explore species associations and possible relationships between these associations and the environment and fishery operational factors. The catch and bycatch data contain many zero-valued observations for each variable (combinations of species and size). Thus, as an error distribution we use Tweedie distribution which has a probability mass at zero and apply Tweedie-GPCA method to the data.

Suppose we want to extract k characteristic features (components) from m dimensional data. Here we assume the mean of each variable is zero or the sample average is subtracted from the observations. PCA finds projections to minimize the mean square reconstruction error (Diamantaras and Kung, 1996). That is, under the model $\boldsymbol{Y} = \boldsymbol{M} + \boldsymbol{E}$ where $\boldsymbol{M}$ is a matrix of rank k, PCA minimizes $\sum_{i,j} E_{ij}^2$. In other words, PCA maximizes the likelihood under the model: $\boldsymbol{Y} = \boldsymbol{M} + \boldsymbol{E},\ E_{ij} \sim N(0, \sigma^2)$, i.i.d. with a constraint rank($\boldsymbol{M}) = k$.

We propose a generalized PCA (GPCA) method that extends principal component analysis in the following sense:

1. The rank of matrix $g(\boldsymbol{M})$, rather than $\boldsymbol{M}$ itself, is k where g is a monotone increasing function and $\boldsymbol{M} = \mathrm{E}[\mathrm{Y}]$.
2. Y_{ij} independently follows a distribution $f(y; M_{ij}, \sigma^2)$ in exponential family.

Features are obtained from $g(M)$ using singular value decomposition (SVD) or independent component analysis (ICA). Our proposed method is a likelihood-based method. The proportion of deviance explained can be used as a criteria for choosing the number of features k.

For catch and bycatch data, we consider Tweedie-GPCA method that uses Tweedie distribution for error and log link function. With $k = 4$, about 70% of deviance was explained by the model. The first few features for variables (species, size) appear to be associated with abundance of several species that are considered vulnerable to fisheries impacts, and associated features for sets show spatial pattern that may be related to oceanography. Some feature shows a similar spatial pattern as that of non-metric multidimensional scaling with Sorensen distance, but features of GPCA show more clear spatial patterns. These results suggest that GPCA may be useful tool for identifying areas within the region occupied by the purse-seine fishery with greater occurrence of bycatch of vulnerable species, perhaps indicating candidate areas for fishery closures to mitigate bycatch.

International Biometric Conference, Dublin, 13-18 July 2008