Using principal component analysis and generalized linear models to predict final outcome in acute stroke

K. Ý. Jónsdóttir[1], L. Østergaard[1], and K. Mouridsen[1]
[1]Department of Neuroradiology, CFIN, Aarhus University Hospital, Århus, Denmark

Introduction Perfusion and diffusion parameters have been shown to predict the final tissue outcome in acute stroke with good accuracy. A number of multivariate, voxel-based algorithms predicting the risk of infarction of the tissue have been suggested in recent years, e.g. GLM algorithms [1], MARS [2] and instance based algorithms [3]. However, substantial noise in input images, e.g. in perfusion measurements due to deconvolution, and collinearity between predictors, leads to overfitting and bias in regression coefficients, when many MR modalities are considered simultaneously. Here, we suggest to find simple multivariate functions of the data, which accounts for most of the variation in the original data while ensuring independence between predictors. With these functions we reduce the dimension of the data while retaining an optimal separation of infarcting and non-infarcting voxels.

Theory Histograms and density estimations of individual MR markers (eg. mean transit time (MTT)) indicate that differentiation of infarcting and non-infarcting voxels is difficult because there is no clear difference in MTT distributions for the two voxel types (Fig. 1). Therefore prediction based on thresholding is almost impossible. Consequently, algorithms considering combinations of physiological parameters have been introduced. However, it has been shown that predictive algorithms can not be optimized by simply adding more physiological markers into the prediction, cf. [2].

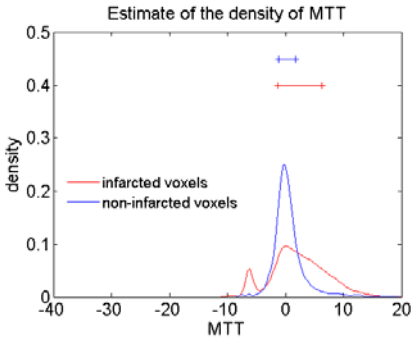

Figure 1

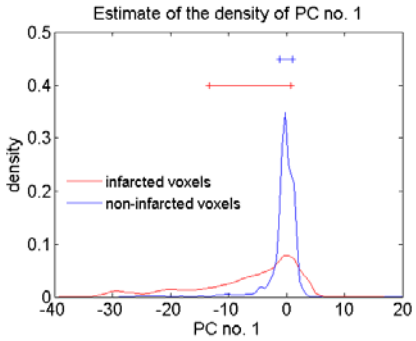

Figure 2

Principal component (PC) analysis can be used to summarize the important variations in data using only a few standardized linear combinations of the data (principal components), thereby reducing the dimension of the original data considerably. For each PC we obtain an estimate of the fraction of variation explained. PCs accounting for only minor parts of the variation are assumed to represent noise components and can be omitted in predictions. Moreover, the weights of the principal components give important information about the relevance of different physiological markers. Fig. 2 shows how voxels can be better separated using only the first principal component. This is illustrated by showing the intervals which comprise 80% of the observations in the plots in Fig. 1 and Fig. 2. We use standard GLM algorithm to combine the contributions of individual PCs (pGLM).

Materials and methods Standard perfusion, diffusion and structural images (MTT, Delay, CBF, CBV, DWI, ADC and T_2) were acquired from 11 patients with acute cerebral ischemia within 12 hours onset. All images were standardized relative to contralateral normal-appearing white matter. We used a jackknifing approach to evaluate the performance of the pGLM prediction algorithm. For each of the training sets the data of the follow up images and the contralateral white matter were used and the principal components of the data were calculated. These were used to estimate the parameters of the prediction algorithm using the first j principal components, $j = 1, \ldots 7$. Then the GLM prediction equation was used to estimate the outcome of the corresponding patient. A ROC curve was calculated and a value of AUC was obtained as a measure of predictive accuracy to evaluate the performance of the combined methods. Note that using all 7 principal components gives the same result as using the original data.

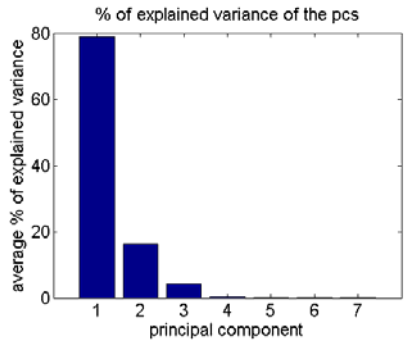

Figure 3

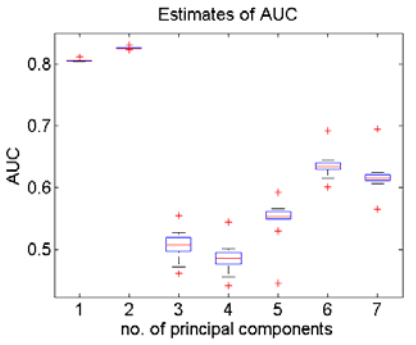

Figure 4

Results The fraction of variation explained by each of the principal components is shown in Fig. 3. The first PC accounts for 78.86% (sd=0.83%) of the variation in the original data. The second PC explains on average 16.28% (sd=0.50%) of the variation, whereas others explain less than 5%. Figure 4 shows a box plot of AUC values for the different training sets for all the seven principal components. Using only the first two principal components in the analysis gives the best performance for this data set. Note that when using more principal components the algorithm performs worse. In particular, the standard GLM is outperformed compared to the simple two PC model.

Conclusions We have shown that by reducing the dimension of the physiological data using principal component analysis we obtain substantial insight into the important variation in the data which leads to better separation of voxels based on MR imaging modalities. Our results also indicate using only a few transformations of the original data gives higher accuracy in prediction of tissue outcome. Finally, relative importance of the physiological markers may be characterized by examining the weights of the principal components.

Reference List

1. Wu, O. et al. Stroke 32:933-42, 2001.
2. Mouridsen, K. et al. ISMRM 12th international meeting and exhibition, 2004.
3. Gottrup, C. et al. Artificial Intelligence in Medicine 33:223-36, 2005.