Lung-airway Data Interrogation via Cluster Analysis

Kung-Sik Chan[1], Eric A. Hoffman[2,3], Kun Chen[1], and Ching-Long Lin[4.5]
[1]Department of Statistics and Actuarial Science,
[2]Department of Radiology, [3]Department of Biomedical Engineering,
[4]Department of Mechanical and Industrial Engineering, [5]IIHR-Hydroscience & Engineering,
The University of Iowa, Iowa City, Iowa, U.S.A.

Abstract

Due to recent advances in the implementation of multi detector-row CT (MDCT)-based imaging and automated image analysis of the lung (Hoffman, Simon and McLennan, 2006), rather detailed *in vivo* measurements of individual human lung airways up to 12th generations have been increasingly and routinely collected by several NIH funded projects. Such microscopic airway data for each lung generally comprise thousands of variables, e.g. segment-by-segment average wall thickness, average inner area, etc. The wealth of such detailed lung-airway data raises the interesting question of how best to empirically explore the cluster structure in a general population of lung airways and/or the presence of systematic variations in the structure of lung airways across groups, e.g. normal subjects vs. subjects with certain lung disease. The presence of such variation may then be exploited for providing inference from samples to population characteristics.

The rich and complex structure inherent in human lung airways presents a steep challenge on how to untangle and amplify the relatively small phenotypical differences from the underlying strong biological variations and patterns embedded in the thousands of measurements. The challenge requires the development of (i) efficient data pre-processing methods for removal of irrelevant systematic biological variations in order to reveal and amplify certain phenotypical signals of interest and (ii) new statistical methods for analyzing high-dimensional data. We propose a new supervised learning method for addressing the aforementioned needs. Our approach builds on our recently developed sparse multivariate reduced-rank regression technique (Chen, Chan and Stenseth, 2011) and robust high dimensional classification with Bayesian variable selection method (Chen, Jiang and Tanner, 2010). With a moderate training sample, our proposed method first automatically and jointly selects a predetermined number of most important variables, based on which a parsimonious Bayesian classification rule is then estimated. The proposed method admits a very efficient implementation that provides real-time classification outcomes. We illustrate our method by utilizing data from a normative lung atlas for comparison (gathered under NIH HL-064368) with data from ongoing multi-center studies seeking to phenotype COPD and Asthma. Analyses to date, demonstrate extremely small mis-classfication error rates.

References:

Hoffman EA, Simon BA, McLennan G. (2006) State of the Art. A structural and functional assessment of the lung via multidetector-row computed tomography: phenotyping chronic obstructive pulmonary disease. Proc Am Thorac Soc. **3**: 519-32.

Chen K, Chan KS and Stenseth NC. (2011). Reduced-rank Stochastic Regression with a Sparse Singular Value Decomposition. Accepted by the Journal of Royal Statistical Society, Series B.

Chen, K, Jiang W and Tanner M.A. (2010). A note on some algorithms for the Gibbs posterior. Statistics and Probability Letters, **80**, 1234-1241.