ROBUST TWO-STAGE MODELLING WITH FINITE MIXTURE CONDITIONAL DISTRIBUTIONS

Inna Chervoneva[1], Mark Vishnyakov[1], Boris Iglewicz[2]
[1]Division of Biostatistics, Department of Pharmacology and Experimental Therapeutics,
Thomas Jefferson University, Philadelphia, Pennsylvania 19107, U.S.A.
[2]Department of Statistics, Temple University, Philadelphia, Pennsylvania 19122, U.S.A.

The motivation for the proposed methodology was an animal study that investigates the role of leucine-rich repeat proteins in the process of collagen fibril development. The mechanisms involved in tendon extracellular matrix assembly are investigated, in part, by studying decomposition of the collagen fibril diameter distributions into subpopulations with different characteristics and functional roles. Statistical modeling of fibril diameters as finite mixtures of normal subpopulations provides insight into the mechanisms regulating collagen fibrillogenesis. Fibril diameter data are collected from multiple animals and microscopic fields per animal, with rather large samples of fibril diameters per microscopic field. The unusual feature of these data is substantial variability in shape and modality of fibril distributions among microscopic fields even from the same animal.

Here, we propose a hierarchical model with multiple levels of random effects and conditional distributions modeled as finite mixtures of normal components. This model is developed to support the most meaningful analysis of fibril diameter distributions, but it is also a general statistical tool, which allows accommodating multilevel clustered data with variety of non-Gaussian conditional distributions, since most continuous univariate distributions may be well approximated with a relatively small number of normal components. We use a two-stage estimation approach, which is appropriate for data with sufficiently large number of continuous measures per cluster in the lowest level of clustering. In the first stage, we use density power divergence estimators (Basu et al, 1998; Fujisawa and Eguchi, 2005) to fit robust 2 and 3-component normal mixture models to tendon fibril diameters in individual microscopic fields. Then resulting microscopic field-specific parameter estimates for means, standard deviations and mixing proportions are modeled in a second stage linear mixed effects (LME) model (Chervoneva et al, 2006). This analysis approach is feasible and expected to perform well for these data because large numbers (200-400) of observations per each microscopic field are available.

Robust rather than maximum likelihood modeling of conditional distributions is important because these conditional distributions are often contaminated with outliers either due to genetic alterations (e.g. abnormally large fused fibrils) or cross sections through the tapered ends of fibrils. Modeling robust estimates for means, standard deviations and mixing proportions in the second stage LME model increases precision and provides superior means for detecting genotype differences in individual mixture components. Analyses of the tendon fibril diameters from wild type and decorin deficient mice demonstrate the advantage of robust stage 1 estimation. In particular, we are able to identify a biologically important difference in the component of immature fibril intermediates with the smallest diameters.