2nd International Symposium on
Information Geometry and its Applications
December 12-16, 2005, Tokyo

GEOMETRY OF MULTISCALE BOOTSTRAP RESAMPLING

HIDETOSHI SHIMODAIRA

The bootstrap probability (BP), the frequency of bootstrap replicates supporting a hypothesis, is often used as a confidence level in complicated data analysis. The hierarchical clustering analysis is a typical example; the BP value is calculated for each cluster by counting how many times that the cluster appeared in thousands of bootstrapped trees. However, BP value is biased, and often leads to false positives. In fact, BP is only first order accurate asymptotically in term of unbiased tests. In this talk, I will explain a newly developed multiscale bootstrap method for calculating approximately unbiased (AU) probability value, which is third order accurate and asymptotically equivalent to the double bootstrap or the p*-formula. The key idea of the algorithm is to alter the sample size of the replicated dataset from that of the observed dataset. The BP values are calculated for several sample sizes, and then BP is plotted against sample size; a very accurate AU p-value is obtained from the slope of the curve, instead of the BP value itself. Geometry in the parameter space, such as the curvature of the boundary of the hypothesis region and the distance from the data-point to the boundary, plays an important role. This new algorithm has already been used widely in Bioinformatics applications.

References

[1] H. Shimodaira and M. Hasegawa. CONSEL: for assessing the confidence of phylogenetic tree selection, *Bioinformatics*, **17**, 1246-1247, 2001.

[2] H. Shimodaira. An approximately unbiased test of phylogenetic tree selection, *Systematic Biology*, **51**, 492-508, 2002.

[3] H. Shimodaira. Approximately unbiased tests of regions using multistep-multiscale bootstrap resampling, *Annals of Statistics*, **32**, 2616-2641, 2004.