Abstract

In several application fields today − genomics and proteomics are examples − we need models for selecting a small subset of useful features from high-dimensional data, where the useful features are both *rare* and *weak*, this being crucial for e.g. supervised classification of sparse high-dimensional data. A preceding step is to detect the presence of useful features, *signal detection*. This problem is related to testing a very large number of hypotheses, where the proportion of false null hypotheses is assumed to be very small. However, reliable signal detection will only be possible in certain areas of the two-dimensional sparsity-strength parameter space, the *phase space*.

In this report, we focus on two families of distributions, $\mathcal{N}$ and χ^2. In the former case, features are supposed to be independent and normally distributed. In the latter, in search for a more sophisticated model, we suppose that features depend in blocks, whose empirical separation strength asymptotically follows the non-central χ^2_ν-distribution.

Our search for informative features explores Tukey's *higher criticism (HC)*, which is a *second-level significance testing procedure*, for comparing the fraction of observed significances to the expected fraction under the global null.

Throughout the phase space we investgate the estimated error rate, $\widehat{\mathrm{Err}}$ = (#Falsely rejected H_0+ #Falsely rejected H_1)/#Simulations, where H_0: absence of informative signals, and H_1: presence of informative signals, in both the $\mathcal{N}$-case and the χ^2_ν-cases, for $\nu = 2, 10, 30$.

In particular, we find, using a feature vector of the approximately same size as in genomic applications, that the analytically derived *detection boundary* is too optimistic in the sense that close to it, signal detection is still failing, and we need to move far from the boundary into the *success region* to ensure reliable detection. We demonstrate that $\widehat{\mathrm{Err}}$ grows fast and irregularly as we approach the detection boundary from the success region.

In the χ^2_ν-case, $\nu > 2$, no analytical detection boundary has been derived, but we show that the empirical success region there is smaller than in the $\mathcal{N}$-case, especially as ν increases.