Albert-Ludwigs-Universität Freiburg
Physikalisches Institut

Johannes Berg

Information Theory: From Statistical Physics to Quantitative Biology

8. exercise class – 28. January 2009

1. Maximum likelihood and Bayes' theorem

Following the model discussed in the lecture, generate a vector of i.i.d random entries $x_1, x_2, \ldots, x_N$. $x_i = 1$ is chosen with probability λ, and $x_i = 0$ with probability $1 - \lambda$. For each $x_i = 1$, pick E_i randomly from a distribution $Q(E)$, and for $x_i = 0$ from $P(E)$. $Q(E)$ and $P(E)$ might be Gaussian ensembles of variance one and mean zero and one, respectively.

The task is to infer the *hidden information* λ and $\{x_i\}$. (Of course, during this computer experiment, *you* know this information.)

Background: The setup mimics transcription factor binding sites on DNA. In some cases, the binding energy of a stretch of DNA to a transcription factor is known as a function of the sequence. Then we can hope to identify functional binding sites, since binding sites have a higher binding energy than stretches of DNA which do not bind to transcription factors (non-binding sites). Of course binding energies both of binding sites and of non-binding sites vary. Suppose the energies E of binding sites have a given distribution $Q(E)$, those of non-binding sites have a distribution $P_0(E)$.

a) Write down the likelihood given $\{E_i\}$ as a function of λ. For a given set $\{E_i\}$, plot the likelihood against λ and compare the position $\lambda^\star$ of its maximum with the value of λ you used to generate the data for both small and large values of N. (You may find it easier to plot the logarithm of the likelihood.)

b) Use Bayes' theorem to evaluate the probability $Pr(Q|E)$ that a given value of E was generated from the ensemble $Q(E)$. Compare the result to the fraction of i with $x_i = 1$ as a function of E. (A practical way is to bin the values of E according to a discrete grid).