Singular Learning Theory Problems: Day 1

Shaowei Lin

21 May 2013

1. **Toric Models**.
   Let $A = (a_{ij})$ be a nonnegative integer $d \times k$ matrix where all the column sums are equal:

$$\sum_{i=1}^{d} a_{i1} = \sum_{i=1}^{d} a_{i2} = \cdots = \sum_{i=1}^{d} a_{ik}.$$

   Given a column $a_j = (a_{1j}, a_{2j}, \ldots, a_{dj})$ and $\omega = (\omega_1, \omega_2, \ldots, \omega_d)$, let ω^{a_j} denote the monomial $\prod_{i=1}^{d} \omega_i^{a_{ij}}$. We use these monomials to define a statistical model whose distribution is given by

$$p(j|\omega) = \frac{1}{Z} c_j \omega^{a_j} \quad \text{for all } j = 1, \ldots, m$$

   where the parameters ω varies over the positive orthant $\mathbb{R}^d_{>0}$, the c_j are positive real constants and Z is the normalization constant $Z = \sum_{j=1}^{k} c_j \omega^{a_j}$. Such models are called *toric models* or *log linear models*. An example is the "biased coin toss" model described in the lecture today.

   Now, suppose we have discrete random variables X_1, X_2 and Y_1, Y_2, Y_3 where the X_i are identically distributed with s states each, the Y_j are identically distributed with t states each, and all the X_i and Y_j are mutually independent.

   (a) Write down parametric equations defining the independence model for $(X_1, X_2, Y_1, Y_2, Y_3)$.

   (b) Given an empirical distribution $\hat{q} = (\hat{q}_{i_1 i_2 j_1 j_2 j_3}) \in \mathbb{R}^k_{\geq 0}$ where $\sum \hat{q}_{i_1 i_2 j_1 j_2 j_3} = 1$, what is the maximum likelihood estimate $\hat{\omega}$ for the independence model as a function of $\hat{q}$?

   (c) (Hard) Prove that for general toric models, the maximum likelihood distribution $\hat{p}$ satisfies

$$A\hat{p} = A\hat{q}.$$

2. **132 Schizophrenic Patients**.
Evans-Gilula-Guttman(1989) studied schizophrenic patients for connections between recovery time (in years Y) and frequency of visits by relatives.

	$2 \leq Y < 10$	$10 \leq Y < 20$	$20 \leq Y$	*Totals*
Regularly	43	16	3	*62*
Rarely	6	11	10	*27*
Never	9	18	16	*43*
Totals	*58*	*45*	*29*	**132**

The *independence model* for this data is a toric model which says that the recovery time and visit frequencies are independent random variables.

Evans, Gilula and Guttman wanted to find out if the data can be explained by a mixture of two independence models, i.e. there is a hidden variable with two states (e.g. male and female).

(a) Write down the parametric equations defining this mixture model.

(b) Write down a nontrivial implicit equation defining this mixture model.

(c) (Hard) Are there any other implicit equations? Can you prove it?

(d) (Hard) Using a computer, find the maximum likelihood estimate numerically.

(e) (Hard) Find the MLE algebraically, i.e. in terms of roots of univariate polynomial.