ICM 1998, Berlin, Aug. 18–27

Abstracts of Plenary and Invited Lectures

Section:

12. Probability and Statistics

1991 MS Classification: 62G07,62G20,62C20
Johnstone, Iain, Stanford University, USA
Oracle Inequalities and Nonparametric Function Estimation

Statistical theory aims in part to articulate when and why certain applied methods of data analysis succeed. With emergence of large, often instrumentally acquired datasets, recent decades have seen a focus on "nonparametric" models in which the number of model parameters grows with the size of available data.

A common caricature is to posit observed data $y \in \mathcal{R}^n$ with structure $y = K\theta + \epsilon z$, Here θ is an unknown function which one desires to "estimate" or "recover", K is a linear operator and $z \in \mathcal{R}^n$ is a vector of standard Gaussian noise of known scale ϵ. An estimator $\hat{\theta}(y)$ is evaluated in terms of, for example, its mean squared error, or "risk" $R(\hat{\theta}, \theta) = E_\theta|\hat{\theta}(y) - \theta|^2$, where E_θ denotes expectation with respect to the probability distribution of y given θ and $|\cdot|$ denotes Euclidean distance.

A class of models $\{\mathcal{M}_J, J \in \mathcal{J}\}$ is given. Typically $\mathcal{M}_J$ are subsets, often linear, of $\mathcal{R}^n$ to which θ might belong. For each model J, there is a "natural" estimator $\hat{\theta}_J$ with risk $R(\hat{\theta}_J, \theta)$. For a fixed, unknown θ, the "ideal risk" would be $\inf_J R(\hat{\theta}_J, \theta)$. This risk would be attainable with the aid of an oracle who divulges the optimal J. Lacking access to oracles, we seek estimators $\hat{\theta}(y)$ with risk properties that mimick the ideal risk as closely as possible.

Here is one example of such an inequality. Suppose that K is the identity on $\mathcal{R}^n$ and that the data model is expressed in terms of coefficients in an orthonormal basis $\{\psi_i\}$:

$$y_i = \theta_i + \epsilon z_i \qquad\qquad i = 1, \ldots n.$$

Here z_i are independent Gaussian noises of mean zero and variance one. Let an estimator $\hat{\theta}_\lambda$ be built componentwise via *soft thresholding* of the observed data: $\hat{\theta}_{\lambda,i}(y)$ equals 0 if $|y_i| \le \lambda$ and equals $y_i - \lambda\,\mathrm{sgn} y_i$ otherwise. Set $\lambda = \epsilon\sqrt{2\log n}$. Then, whatever be $\theta \in \mathcal{R}^n$ and $\epsilon > 0$,

$$E_\theta \sum_1^n (\hat{\theta}_{\lambda,i}(y) - \theta_i)^2 \le (2\log n + 1)[\epsilon^2 + \sum_1^n \min(\theta_i^2, \epsilon^2)].$$

From the inequality, one learns, for example (i) the ideal risk $\sum_1^n \min(\theta_i^2, \epsilon^2)$ connects compressibility of the signal θ in basis $\{\psi_i\}$ with the quality of estimation/recovery. It also establishes a link with approximation theory, (ii) near minimax optimality properties simultaneously over a class of parameter spaces Θ, and (iii) the factor $2\log n$ cannot be improved: a heuristic limiting the detection of sparse signals in Gaussian noise shows why.
Time permitting, we will discuss two other such oracle inequalities: (i) a linear shrinkage inequality involving the classical James-Stein estimator, and
(ii) an oracle inequality for "empirical atomic decomposition" in which the orthobasis $\{\psi_i, i = 1, \dots n\}$ is replaced by a possibly highly redundant *dictionary*, such as might be approriate when the class of models for the signal θ is not well represented in terms of unconditional bases.