What is df for signal, why do you want to know it, and how do you compute it?

Grace Wahba
`http://www.stat.wisc.edu/~wahba`
→ *TRLIST*

Lucent Technologies
March 3, 1998

Based on joint work with
Dong Xiang
Xiwu Lin

Abstract

We review the definition of degrees of freedom (df) for signal in the Gaussian case and its use in Mallows C_L and GCV for choosing the smoothing (penalty) parameter(s), in nonparametric regression estimates which are solutions to penalized least squares problems. In these problems the target is to minimize the predictive mean square error. Randomized trace methods for computing the df, particularly in the context of very large data sets with iterative computational solutions are discussed. Turning to non-Gaussian data, it is useful to reexamine the notion of df, and a notion of generalized df (gdf) similar to that proposed by Ye is discussed. A generalized approximate crossvalidation function (GACV, Xiang and Wahba, 1996, 1997) derived similarly as GCV is shown to be useful in choosing smoothing parameters in penalized likelihood

estimaties for Bernoulli data, when the target is the comparative Kullback-Liebler distance. The gdf, which is a part of the GACV, may be estimated by randomized trace methods. Similarly a GACV estimate has been derived for use with support vector machines, and it too, relies on the gdf for this problem, which may be computed via a randomized trace calculation.

Outline

1. Review of elementary definitions of df signal in the Gaussian case.

2. Unbiassed Risk (Mallows C_L) and GCV estimates.

3. The randomized trace estimate of df signal.

4. RanTrace in Large, Nonlinear problems

5. Generalized Approximate Crossvalidation (GACV) for non-Gaussian data

6. Application of the GACV to Support Vector Machines

7. Open problems.