Challenges in Applied Statistical Model Building in Time and Space With Very Large Data Sets

Grace Wahba

All TR's since late 93 up in

http://www.stat.wisc.edu/~wahba

Canberra, Australia, July 2, 1996

Collaborators:

Dong Xiang

Zhen Luo

Jianjian Gong

Message

Penalized likelihood methods provide an important paradigm for statistical model building, nonparametric regression and data analysis in many fields. However, new numerical methods must be developed in order to be able to efficiently apply these methods to large ($n = 10^3$) to extremely large ($n = 10^6$) data sets, as obtain in e. g. in demographic medical studies and environmental data analysis.

The numerical methods should be tailored to the statistical problem at hand so that both the goals, of numerical efficiency, and statistical optimality are simultaneously met. Sometimes an efficient numerical method can be developed which at the same time implements some statistically desireable goal.

Recent developments include good methods for solving very large penalized log likelihood problems approximately by using a judicious choice of basis functions, and numerical methods for implementing GCV, unbiased risk and related methods for methods for choosing multiple smoothing parameters without using matrix decompositions. To be useful with extremely large data sets, the methods must work well in the context of iterative solutions to mildly nonlinear optimization problems.

In this talk we will describe one of these recently developed methods in detail, and briefly refer to several others.

Bibliography:

(1) Optimal stopping of the conjugate gradient algorithm-recent results.

Wahba, G., Johnson, D., Gao, F. & Gong, J. (1995), 'Adaptive tuning of nummerical weather prediction models: randomized GCV in three and four dimensional data assimilation', *Mon. Wea. Rev.* **123**, 3358–3369.

(1) Optimal stopping of the conjugate gradient algorithm-early results.

Wahba, G. (1987), Three topics in ill posed problems, *in* H. Engl & C. Groetsch, eds, 'Proceedings of the Alpine-U.S. Seminar on Inverse and Ill Posed Problems', Academic Press, pp. 37–51.

Hutchinson, M. (1989), 'A stochastic estimator for the trace of the influence matrix for Laplacian smoothing splines', *Commun. Statist.-Simula.* **18**, 1059–1076.

Girard, D. (1989), 'A fast 'Monte-Carlo cross-validation' procedure for large least squares problems with noisy data', *Numer. Math.* **56**, 1–23.

(2) Adaptive selection of basis functions, HAS (Hybrid Adaptive Splines) and clustering-recent results.

Luo, Z. & Wahba, G. (1995), Hybrid adaptive splines, Technical Report 947, Dept. of Statistics, University of Wisconsin, Madison WI, to appear, J.A.S.A.

Xiang, D. (1996), Model Fitting and Testing for Non-Gaussian Data with a Large Data Set, PhD thesis, Technical Report 957, University of Wisconsin-Madison, Madison WI.

(2) Adaptive selection of basis functions-early references.

Hutchinson, M. (1984), A summary of some surface fitting and contouring programs for noisy data, Technical Report ACT 84/6, CSIRO Division of Mathematics and Statistics, Canberra, Australia. Many papers since.

Wahba, G. (1980), Spline bases, regularization, and generalized cross validation for solving approximation problems with large quantities of noisy data, *in* W. Cheney, ed., 'Approximation Theory III', Academic Press, pp. 905–912.

(3) GACV and the randomized trace algorithm for Bernoulli data.

Xiang, D. & Wahba, G. A generalized approximate cross validation for smoothing splines with non-Gaussian data, Technical Report 930, Dept. of Statistics, University of Wisconsin, Madison, WI. 1994. to appear, *Statistica Sinica*

Xiang, D. (1996), Model Fitting and Testing for Non-Gaussian Data with a Large Data Set, PhD thesis, Technical Report 957, University of Wisconsin-Madison, Madison WI.

(4) Analyzing large, nearly regular data sets with smoothing spline ANOVA and the backfitting algorithm.

Luo, Z. (1996), Backfitting in Smoothing Spline ANOVA and an application to historical global temperature data, PhD thesis, Dept. of Statistics, University of Wisconsin, Madison, WI.

Wahba, G. & Luo, Z. (1996), Smoothing spline ANOVA fits for very large, nearly regular data sets, with application to historical global climate data, *in* C. Micchelli, ed., 'Festschrift in Honor of Ted Rivlin', Balzer Press, pp. xx –to appear.