Issues in Statistical Model Building Where a 'True' Model is Not Well Defined

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

June 26, 1996

Collaborators:
Yuedong Wang
Chong Gu
Ron Klein, MD
Barbara Klein, MD
Zhen Luo
Jianjian Gong
Donald R. Johnson
Joseph Tribbia

We consider model building via multivariate spline ANOVA and related penalized likelihood methods with multiple smoothing and tuning parameters and possibly messy model selection issues from the point of view of three different contexts:

1. Risk factor estimation in demographic medical studies.

2. Descriptive methods for environmental data, scattered in time and space.

3. Ingestion of data into dynamical systems equations, as in atmospheric and oceanic modeling.

We describe how a common paradigm involving reproducing kernel hilbert spaces, cross-validation, unbiased risk and maximum likelihood methods can be fruitfully applied in all of these contexts. We note that similar issues of parsimony, accuracy, predictability and interpretability of the results arise in all three contexts. However in the above applications there are important differences depending on the sample sizes available, the nature of the data quality, and last, but definitely not least, differing evaluation and interpretability and requirements which are tied to the context and the goals of the potential user. In all of the above examples a 'true' model is not really completely well defined, but the concept of a model is useful nevertheless.

Bibliography: See also

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

(1) Risk factor estimation.

Wahba, G., Wang, Y., Gu, C., Klein, R. & Klein, B. (1995*b*), 'Smoothing spline ANOVA for exponential families, with application to the Wisconsin Epidemiological Study of Diabetic Retinopathy', *Ann. Statist.* **23**, 1865-1895.

Wahba, G., Gu, C., Wang, Y. & Chappell, R. (1995*a*), Soft classification, a. k. a. risk estimation, via penalized log likelihood and smoothing spline analysis of variance, *in* D. Wolpert, ed., 'The Mathematics of Generalization, Santa Fe Institute Studies in the Sciences of Complexity, Proc. Vol. XX', Addison-Wesley, Reading, MA, 329-360.

(2) Environmental data in time and space.

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, to appear.

Luo, Z. (1996), Backfitting in Smoothing Spline ANOVA, and an Application to Historical Global Temperature Data, PhD thesis, to appear.

(3) Data + dynamical systems equations.

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.

Gong, J. Wahba, G., Johnson, D.,& Tribbia, J. (1996), 'Adaptive tuning of nummerical weather prediction models: simultaneous estimation of weighting, smoothing and physical parameters, in preparation.

See also

Wahba, G. (1995), Generalization and regularization in nonlinear learning systems, *in* M. Arbib, ed., 'Handbook of Brain Theory and Neural Networks', MIT Press, 426-430.

Wahba, G. (1990), Spline Models for Observational Data, SIAM. 169 + xii pp.

Following are the subset of the slides which were typed. Handwritten slides not included.

Model Selection Issues
Part (1) Risk Factor Estimation

- Given the terms to be included in the model choose the smoothing parameters. GRKPACK-Gu's Unbiased Risk Estimate. GACV possible. (Xiang and Wahba).

- Which terms should be included? Note: in the WESDR study many potential predictor variables were recorded.

- Needed: Good and efficient prescreening procedures. (Marginal fits, NN's?, parametric contingency tables?)

Model selection methods:

- Direct comparison of UBR or GACV not ‘fair’ for nested models, need to ‘charge’ for adding another subspace - how?, or is this important ?

- Leaving out x%. Use KL-distance. When are two different models ‘significantly different’ based on the KL-distance from the data left out.

- Gu’s cosine diagnostics.

- Component-wise Bayesian confidence intervals.

- Bootstrap confidence intervals.

- Component of no ‘practical’ significance.

- Test of null model hypothesis. Xiang and Wahba(1995). Neyman tests, many authors.

- Test for additive model. (Z. Chen, 1993, Barry, 1993, Eubank et al 1995)

- Subjective judgments, interpretability. In the case of medical demographic data, predictor variables are related - which came first, the chicken or the egg?

Model Selection Issues
Part (2) Environmental Data in Time and Space

No question about the number of terms in the model, each has a meaning to the climatologist and all are present from a scientific point of view.

Questions

What does the 'global average' mean? We have used Lebesgue integral with uniform measure to define the average in a function space. Is this reasonable? Given the sparsity of data in some regions, should some other definition be used to give more meaning to the 'change in the global average' with time.

Should the temperature be corrected for altitude, or, should this be taken into account in the averaging. (A high altitude station which is reporting irregularly may bias the estimate of 'warming'.)

'Global warming' estimates may be sensitive to the choice of smoothing parameters. (Luo used a subjective near-interpolation choice for the main effects and GCV for the other terms.)

Choice of penalty functionals (RK's) - does this make a difference? Linear part of the time trend is not penalized, should the quadratic part also not be penalized? What about R_2?

Note: (i) These questions are only partly statistical !! (ii) There is only one 'realization' of this process in our lifetimes!!

Model Selection Issues
Part (3) Data + dynamical systems equations.

Choice of J_M and J_S, which in practice (numerical weather prediction, atmospheric data reanalysis) will contain various kinds of smoothing, weighting and physical parameters, to be selected in some manner. There is a lot of prior data and physical information that could be brought to bear in designing J_S, ('balance between wind and heights', 'energy distribution in the atmosphere'). How do you penalize deviation from the dynamical systems model when there is not a lot of understanding as to how the model differs from nature? (If the discrepancy was understood, it would be fixed!)

For numerical weather prediction, the evaluation criteria is usually not an issue, the goal is to improve the prediction. However, trade-offs between short and long term adaptivity can be an issue. Other issues include sensitivity to various kinds of tuning parameters, estimability of physical parameters in M_t, F_t.

Message

Penalized likelihood in general and smoothing spline ANOVA in particular provide a paradigm which can be used in a variety of areas of scientific research. These methods, as well as essentially all nonparametric regression methods have have smoothing or bandwidth parameters which in some general sense control the tradeoff between goodness of fit to the data, and complexity of the model (degrees of freedom for signal). These issues were not really discussed in this talk, but unbiased risk, generalized cross validation, leaving out one or several, bootstrapping, and maximum likelihood are widely accepted methods that have found successful use in many applications. However, we have attempted to show that broader and more difficult model selection solutions may need to incorporate valid subject matter information to get the best answers.