RIT on Large Cross-Classified Datasets Models with Increasing Parameter Dimension

Research Focus: Large datasets arise naturally in many areas of science, government, and business. Typically, as the size of a dataset gets large, the complexity of questions which one addresses with it also increases. Growing parameter dimension violates the formal setting of most textbook mathematical statistics, in which parameter-dimension is fixed and sample size increases to infinity. The new setting requires a new Asymptotic Theory which explicitly recognizes the controlled growth of the parameter-space of a probability model as a function of the dataset size.

Overview: How Parameter Dimension Varies

Problems formulated in triangular-array setting, with data complexity (predictors, cross-classifying variables, or nuisance parameters) increasing with data size:

(A) *Number $p(n)$ of predictors increasing with n.*

(B) *Numbers of nuisance parameters associated with clusters increasing with n. "Two-index Asymptotics"*

(C) *Semiparametric Problems*

(A) *Number $p(n)$ of predictors increasing with n.*

In problems like econometrics, data-mining, pattern recognition, and many others, the number of potential regressions in linear or generalized-linear models is huge. Variables are selected either through automatic model selection methods (AIC, BIC and generalizations) or by specifying a large fixed set of regressors that can grow with sample size. Generic model

$$y_i = \mathbf{x}_i(n)^{tr}\ \beta(n)\ +\ \epsilon_i \quad , \qquad \epsilon_i \sim F_{(n)}$$

- Special case with $F_{(n)}(x) = \Phi(x/\sigma_n)$ is already interesting.
- Case with unknown $F_{(n)}$ is *semiparametric* regression. Natural model is $F_{(n)}(x) = G(x/\sigma_n)$ with G a fixed nuisance-function and σ_n an unknown scalar parameter.
- Logistic-regression and other GLM's could be put in similar framework.

References:

Yohai and Maronna (1979) (*M-estimation*)

Portnoy (1988) (*GLM*)

Wei (1992) (*Model selection and consistency via least-squares*)

(B) *Numbers of nuisance parameters associated with numbers of clusters or cross-classified homogeneous cells increasing with n: "Two-index Asymptotics"*

Generic model with data-size $= m(n) \cdot p(n)$:

$$y_{ij} \sim f(y; \beta, \lambda_j) \quad , \quad 1 \le i \le m(n) \ , \ 1 \le j \le p(n)$$

- Estimation of nuisance parameters λ_j decouples across different clusters j, but β is shared pop-wide.
- Techniques based on **adjusted profile likelihood** : estimate β by replacing λ_j in likelihood within cluster j by likelihood-based estimate $\tilde{\lambda}_j(\beta)$ depending on β and **then** maximizing likelihood over β

References:

Neyman, J. and E. Scott (1948)

Paper of Barndorff-Nielsen (1996) in book referenced as origin of term "two-index asymptotics".

Li, Lindsay, and Waterman (2003).

Miscellaneous papers on adjusted profile likelihoods.

Example: Cross-classified Factor analysis model

$$\mathbf{y}_{ij} = \Lambda \, \mathbf{f}_{ij} + \epsilon_{ij}, \quad f_{ij} \sim \mathcal{N}(0, I_q), \quad \epsilon_{ij} \sim \mathcal{N}(0, \sigma_j^2)$$

(C) *Semiparametric Problems*

Now the problem is that all observations are simultaneously parameterized by a finite-dimensional structural parameter β of interest and an (infinite-dimensional) nuisance-pararameter λ.

Can consider the example problem of **infinite-order regression**

$$y_i \;=\; \sum_{k=1}^{d} x_{ik}\,\beta_k \;+\; \sum_{j=1}^{\infty} x_{i,j+1}\,\gamma_j \;+\; \epsilon_i \quad , \quad \epsilon_i \sim F$$

with $\lambda = (\gamma, F)$, or 'partially linear' regression (Bhattacharya & Zhao 1997 Ann. Stat.) which can be viewed as a special case.

References:

Bhattacharya, P.K. and Zhao, P.-L. (1997)

Slud, E. and Vonta, F. (2005)

(D) *Misspecified-model Issues*

References:

Chen, Ru (2005) Thesis

White, H. (1982)