Optimal Smoothing and Mixing in MRR: Some Theoretical Results

B. A. Starnes (C-N),

Jeffrey B. Birch (VPI&SU)

James Mays (VCU)

Tim Robinson (UWYO)

5 June 2007

The Model and affiliated Assumptions

Model: $y_i = g(\mathbf{x}_i) + \varepsilon_i$, $\varepsilon_i \sim^{iid} N(0,\sigma)$, $i = 1, 2,\ldots n$

- g is continuous on C, a closed subset of $\mathbb{R}^d$
- $g(\mathbf{x}_i)$ is the *true* mean function (as if anyone could ever know that) of a variable $\mathbf{x}_i = [x_i, f_2(x_i), \ldots, f_d(x_i)]$
- The realizations of x_i are fixed uniformly (and increase uniformly) on a closed subset of $\mathbb{R}$
- Assumptions A1 – A6
  - A1-A5 are found in Mays, Birch and Starnes (2001). These generally pertain to the regularity conditions of the parametric and nonparametric estimates and are satisfied by linear parametric estimators and local polynomial nonparametric estimators
  - A6 The "leave one out" nonparametric estimate satisfies $\hat{g}^{(i)}(x_i)=\frac{\sum W_2(x_i,x_j)e_j}{\sum W_2(x_i,x_j)}$ where $e_j=y_j-y^{(p)}(x_j)$ which is satisfied for local polynomial and spline smoothing estimates. It is necessary for the DMRR results.

The Model and affiliated

Assumptions (cont.)

- R1 – R4 as found in Starnes (1999), and Robinson, Birch and Starnes (2008) are also necessary for the DMRR results. Generally speaking these ...
  - require that $V(y_i)$ at each x_i is bounded (in the instance of nonconstant variances)
  - require that the matrix D of partial derivatives of the true parametric function on $\boldsymbol{\beta}$ be of full rank for all $n \in \mathbb{N}$, and both continuous and bounded for variables x and $\boldsymbol{\beta}$ on C and $\mathbb{R}^d$ respectively

- $g(\mathbf{x}_i)$ is the *true* mean function which is based on an unknown parameter $\boldsymbol{\beta}$.

-For example our model could be written

$$y_i = \mathbf{x}_i^T\boldsymbol{\beta} + \varepsilon_i \;, \varepsilon_i \overset{iid}{\sim} N(0,\sigma),\; i = 1, 2, \ldots n$$

Here we choose the $\boldsymbol{\beta}$ estimate to minimize

$$SSE(\mathbf{b}) = \sum^n (y_i - \mathbf{x}_i^T\mathbf{b}) \text{ and obtain}$$

$$\mathbf{b} = (X^TX)^{-1}X^T\mathbf{y}, \text{ and}$$

$$\hat{g}(\mathbf{x}_i) = \mathbf{x}_i^T\mathbf{b}. \; []$$

Nonparametric Regression

- Unsure of form of g($\mathbf{x}_i$), so we estimate locally

-For example, kernel regression (Nadaraya-Watson (1964))

$$\hat{g}(x_i)=\frac{\sum K_h(x_j-x_i)Y_i}{\sum K_h(x_j-x_i)},\text{with}\,K_h(u)=\frac{K(u/h)}{h}$$

is a special case of local polynomial regression. []

Distance Measures and Asymptotics

- For i = 1, 2, ... n, and functions of $\mathbf{x}_i$
  - $\langle h_a, h_b \rangle = n^{-1}\sum(h_a(\mathbf{x}_i)h_b(\mathbf{x}_i))$
  - $\langle h_a, h_a \rangle = ||h_a||^2$
  - $\delta_n = \inf\{||\boldsymbol{g} - \boldsymbol{y}^p(\boldsymbol{\beta}^*)||: \boldsymbol{\beta}^* \in \mathbb{R}^d\}$
    - Parametric model correct ($\delta_n = 0$)
    - Parametric model incorrect ($\mathscr{L}\delta_n > 0$)
  - $\gamma_n^2 = E(||\boldsymbol{y}^{np} - \boldsymbol{g}||)^2$ = AVEMSE
  - For MRR2,

$$\gamma_n^2 = E(||\boldsymbol{y}^{np} - (\boldsymbol{g} - \boldsymbol{y}^p(\boldsymbol{\beta}^*))||)^2,\ \boldsymbol{\beta}^* \in \mathbb{R}^d$$

Distance Measures and Asymptotics (cont.)

- A stochastic sequence $\{X_n\}$ is said to be $O_P(1)$ if for every $0 < \eta < 1$, there exist constants $K(\eta)$, $n(\eta) \in \mathbb{R}$, such that for $n \geq n(\eta)$,

$$P\{|X_n| \leq K(\eta)\} \geq 1 - \eta$$

- A stochastic sequence $\{X_n\}$ is said to be $O_P(b_n)$ if

$$\{X_n / b_n\} = O_P(1)$$

- This notation is the stochastic parallel to "*O notation*" in the Analysis framework.

- A simple example is given by the CLT; a special case of the Weak LLN. Here, if repeated samples of increasing size n are taken of an r.v. X ~?(μ, σ) (with σ finite) then we would write

$$|\bar{X}_n - \mu| = O_P(n^{-.5}). []$$

Model Robust Regression

- Improve regression function estimate of g by combining parametric and non-parametric estimates with a mixing parameter
- Two forms of Model Robust Regression: MRR1 and MRR2 (developed at Virginia Tech)
- Risk consists of possible drawbacks involving weaknesses of either estimate
- The MRR2 Procedure was developed in 1995 by James Mays and Jeffrey B. Birch
- The parametric estimate (designated as y^{p}) is always involved
- Nonparametrically smoothed residuals (nonparametric estimate designated as y^{np})

The MRR2 Estimate

- Form: $\hat{g}(x_i) = y_i^p + \lambda y_i^{np}$
- Mixing parameter: $0 \leq \lambda \leq 1$
  - If $\lambda = 0$, parametric model is correct
  - If $\lambda = 1$, nonparametric estimate is added since parametric model is inadequate.
  - The size of λ represents the degree to which the parametric estimate is improved
  - The optimal mixing parameter obtained via simple calculus is given by

    $\lambda^* = \langle \mathbf{y}^{np}, \boldsymbol{g} - \mathbf{y}^p \rangle / ||\mathbf{y}^{np}||^2$

  - The optimal data driven mixing parameter is given by

    $\hat{\lambda}^* = \langle \mathbf{y}^{np}, \mathbf{y} - \mathbf{y}^p \rangle / ||\mathbf{y}^{np}||^2$

Theorem 1

Given the model and affiliated assumptions A1 – A5 above ...

$$|| \hat{\lambda}\, {}^{*}\mathbf{y}^{np} + \mathbf{y}^{p} - \boldsymbol{g} || = O_P(\gamma_n) \text{ if } \mathscr{L}\ \delta_n > 0$$

and

$$|| \hat{\lambda}\, {}^{*}\mathbf{y}^{np} + \mathbf{y}^{p} - \boldsymbol{g} || = O_P(n^{-.5}) \text{ if } \delta_n = 0.$$

Thoughts

- When the parametric model is correct, the MRR2 estimate converges asymptotically at the same rate as a parametric estimate. Otherwise the estimate converges at the rate of the nonparametric estimate
- We achieve the same result for the data driven and theoretical asymptotically optimal λ
- Asymptotically, MRR2 achieves the “Golden result of Model Robust Regression”

Dual Model Robust Regression

- The DMRR Procedure was developed in 1997 by Tim Robinson and Jeffrey B. Birch at Virginia Tech
- In the case of nonconstant variances, this method provides a regression function estimate of *g* in both mean and variance cases by combining parametric and non-parametric estimates with a mixing parameter in each case
- Utilizes the form of Model Robust Regression known as MRR1 (developed at Virginia Tech in the late 1980's) for the variance estimate and MRR2 for the means estimate
- Risk consists of possible drawbacks involving weaknesses of either estimate, which in this instance are interrelated
- The process of estimation involves a 10 step IRLS algorithm that estimates the mean and variance alternatively
- We will designate the true mean function by g_μ and the true variance function by g_σ

The DMRR Estimate

- Form: $\hat{g}_{\mu}(x_i) = y_i^{p(EWLS)} + \lambda_{\mu} y_i^{np(LLR)}$
- Form: $\hat{g}_{\sigma}(x_i) = (1-\lambda_{\sigma}) y_i^{p(GLM)} + \lambda_{\sigma} y_i^{np(LLR)}$
- Mixing parameter: $0 \leq \lambda \leq 1$
  - The optimal data driven mixing parameter for the mean estimate is given by

$$\hat{\lambda}_{\mu}^{*} = \langle \mathbf{y}^{np(LLR)}, \mathbf{y} - \mathbf{y}^{p(EWLS)} \rangle / ||\mathbf{y}^{np(LLR)}||^2$$

  - The optimal data driven mixing parameter for the variance estimate is given by

$$\hat{\lambda}_{\sigma}^{*} = \frac{\langle \mathbf{y}^{p(GLM)} - \mathbf{y}^{np(LLR)}, \mathbf{y}^{p(GLM)} - \mathbf{e}^2 \rangle}{||\mathbf{y}^{p(GLM)} - \mathbf{y}^{np(LLR)}||^2}$$

where $\mathbf{e}^2$ is the vector of squared residuals in the variance estimate case

Theorem 2

Given the model and affiliated assumptions A1 – A6 above ...

$$|| \hat{\lambda}_{\mu}^{*}\mathbf{y}^{np(LLR)} + \mathbf{y}^{p(EWLS)} - \boldsymbol{g}_{\mu} || = O_P(\gamma_n)$$

if $\mathcal{L}\ \delta_n > 0$

and

$$|| \hat{\lambda}_{\mu}^{*}\mathbf{y}^{np(LLR)} + \mathbf{y}^{p(EWLS)} - \boldsymbol{g}_{\mu} || = O_P(n^{-.5})$$

if $\delta_n = 0$.

Theorem 3

Given the model and affiliated assumptions A1 – A6
and requirements R1 - R4 above ...

$$|| \hat{\lambda}_{\sigma}^{*}\mathbf{y}^{np(LLR)} + (1 - \hat{\lambda})\mathbf{y}^{p(GLM)} - \boldsymbol{g}_{\sigma}|| = O_P(\gamma_n)$$

if $\mathscr{L}\ \delta_n > 0$

and

$$|| \hat{\lambda}_{\sigma}^{*}\mathbf{y}^{np(LLR)} + (1 - \hat{\lambda})\mathbf{y}^{p(GLM)} - \boldsymbol{g}_{\sigma}|| = O_P(n^{-.5})$$

if $\delta_n = 0$.

Thoughts

- When the parametric model is correct, the MRR2 estimate converges asymptotically at the same rate as a parametric estimate. Otherwise the estimate converges at the rate of the nonparametric estimate
- Once again we achieve the same result for the data driven and theoretical asymptotically optimal λ
- Asymptotically, DMRR achieves the "Golden result of Model Robust Regression" as in the MRR1 and MRR2 cases
- It is important to note that the asymptotic results do not pertain to the number of iterations, but rather the number of observations

References

Bishop, Y. M. M., Feinberg, S. E., and Holland, P. W. (1975). *Discrete Multivariate Analysis,Theory and Practice*. Cambridge, MA: MIT Press.

Burman, P., and Chaudhuri, P. (1992), "A Hybrid Approach to Parametric and Nonparametric Regression", *Technical Report No. 243*, Division of Statistics, University of California-Davis, Davis, CA, USA.

Mays, J., Birch, J., and Starnes, B. (2001), "Model Robust Regression: Combining Parametric, Nonparametric and Semiparametric Methods", *Journal of Nonparametric Statistics*, **13**, 245-277.

Starnes, B., and Birch, J. (2000), "Asymptotic Results for Model Robust Regression", *Journal of Statistical Computation and Simulation*, **66**, 19-33.

Robinson, T., Birch, J., and Starnes, B. (2008), "A Semiparametric Approach to Dual Modeling", *Journal of Statistical Planning and Inference*, article to appear.