INFORMATION THEORY AND FLEXIBLE HIGH-DIMENSIONAL NON-LINEAR FUNCTION ESTIMATION

Andrew R. Barron

YALE UNIVERSITY
DEPARTMENT OF STATISTICS

Presentation, November 12, 2011

at the Info-Metrics Institute, American University, Washington, DC

Outline

- Data, Model
- Combining Non-linearly Parameterized Terms
- Penalized Likelihood Criteria, Minimum Description Length
- Statistical Risk Determination
- Computation
- Adaptive Annealing
- Summary

Data

- Data: (X_i, Y_i), $i = 1, 2, \ldots, n$
- Inputs: explanatory variables X_i in a unit cube in R^d
- Random design: independent $X_i \sim P$
- Output: response variable Y_i in R
- Relationship: $Y_i = f(X_i) + \epsilon_i$
- Noise: ϵ_i independent $N(0, \sigma^2)$
- Function: f unknown

Non-linear Dictionaries

- Build functions $f_m(x) = \sum_{j=1}^{m} c_j \phi_d(\theta_j, x)$ in the span of a dictionary $\Phi = \{\phi_d(\theta, \cdot) : \theta \in \Theta\}$
- Product Bases

$$\phi_d(\theta, x) = \phi_1(\theta_1, x_1)\,\phi_1(\theta_2, x_2) \cdots \phi_1(\theta_d, x_d)$$

- Ridge Bases (as in projection pursuit regression)

$$\phi_d(\theta, x) = \phi_1(\theta_0 + \theta_1 x_1 + \theta_2 x_2 + \ldots + \theta_d \theta_1)$$

- Examples of activation functions $\phi(z) = \phi_1(z)$ for
  - Perceptron networks: $\phi(z) = 1_{\{z>0\}}$
  - Sigmoidal networks: $e^z/(1+e^z)$
  - Sinusoidal models: $cos(z)$
  - Hinging hyperplanes: $(z)_+$
  - Quadratic splines: 1, z, $(z)_+^2$
  - Cubic splines: 1, z, z^2, $(z)_+^3$
  - Polynomials: $(z)^q$

Notation

- Response vector: $Y = (Y_i)_{i=1}^n$ in R^n
- Dictionary vectors: $\Phi_{(n)} = \left\{ (\phi_d(\theta, X_i))_{i=1}^n \ : \ \theta \in \Theta \right\}$
- Sample squared norm: $\|f\|_{(n)}^2 = \frac{1}{n}\sum_{i=1}^n f^2(X_i)$
- Population squared norm: $\|f\|^2 = \int f^2(x) P(dx)$
- Normalized dictionary condition: $\|\phi\| \leq 1$ for $\phi \in \Phi$

Functions represented in span of non-linear dictionary

- Variation of f w.r.t. Φ

$$V_\Phi(f) = \|f\|_\Phi = \inf\{V \;:\; f/V \in conv(\pm\Phi)\}$$

- E.g. $f(x) = \sum_j c_j \phi(\theta_j, x)$ has $\|f\|_\Phi = \sum_j |c_j|$, the ℓ_1 norm of the coefficients in representation of f in the span of Φ
- E.g. $f(x) = \int e^{i\,\theta^T x}\,\tilde{f}(\theta)\,d\theta$ (Fourier representation) Then $\|f\|_\Phi$ is given by an L_1 spectral norm:

$$V_{cos}(f) = \int_{R^d} |\tilde{f}(\theta)|\,d\theta$$

$$V_{step}(f) = \int |\tilde{f}(\theta)|\;\|\theta\|_1\,d\theta$$

$$V_{q-spline}(f) = \int |\tilde{f}(\theta)|\;\|\theta\|_1^{q+1}\,d\theta$$

Penalized Likelihood, Minimum Description Length

- Description Length divided by sample size:

  $\frac{1}{n}\left[\ \log 1/\text{likelihood} + \text{ Model Complexity Penalty }\right]$
- Control of number of terms:

$$\frac{\|Y - f_m\|^2_{(n)}}{2\sigma^2} \quad + \frac{m}{n} \log N_{\Phi,1/n}$$

  where the penalty is typically of order $\frac{md}{n} \log n$
- Control of the ℓ_l norm of coefficients:

$$\frac{\|Y - f\|^2_{(n)}}{2\sigma^2} \quad + \quad \lambda_n \|f\|_\Phi$$

$$\lambda_n = \sqrt{\frac{2 \log N_{\Phi,1/n}}{n}}$$

- Optimize the above criteria to yield estimators $\hat{f}$ and $\hat{f}_{\hat{m}}$

Statistical Risk Bounds

Bounds on the population accuracy of function estimates when the true function is f^*

$$E\|\hat{f}_m - f^*\|^2 \;\leq\; \|f_m - f^*\|^2 + \frac{cmd}{n}\log n$$

$$E\|\hat{f}_{\hat{m}} - f^*\|^2 \;\leq\; \min_m\{\|f_m - f^*\|^2 + \frac{cmd}{n}\log n\}$$

$$E\|\hat{f} - f^*\|^2 \;\leq\; \min_f\{\|f - f^*\|^2 + \lambda_n\|f\|_\Phi\}$$

$$E\|\hat{f} - f^*\|^2 \;\leq\; \|f^*\|_\Phi\sqrt{\frac{2d\log n}{n}}$$

Relaxed Greedy Algorithm and LASSO

- Initialize $\hat{f}_0 = 0$. For step k, have the previous fit $\hat{f}_{k-1}$.
- Optimize the new term: Maximize the inner product with the residuals $res_i = Y_i - \hat{f}_{k-1}(X_i)$ to obtain the new ϕ and its parameter vector $\hat{\theta}_k$

$$\mathrm{argmax}_\theta \frac{1}{n} \sum_{i=1}^{n} res_i \, \phi(\theta, X_i)$$

- Update the fit:

$$\hat{f}_k = \alpha \hat{f}_{k-1} + \beta \phi$$

- Obtain the coefficients α, β by either least squares or by ℓ_1 penalized least squares:
  - $\min \| Y - \alpha \hat{f}_{k-1} - \beta \phi_k \|^2$ for the relaxed greedy algorithm
  - $\min \| Y - \alpha \hat{f}_{k-1} - \beta \phi_k \|^2 + \lambda_n \big(|\alpha| V_{k-1} + |\beta| \big)$ for ℓ_1 pen pursuit (LASSO)

Alternative: Forward Stepwise Regression or Orthogonal Matching Pursuit

- Initialize $\hat{f}_0 = 0$. For step k, have the previous fit $\hat{f}_{k-1}$.
- Optimize the new term: Maximize the inner product with the residuals $res_i = Y_i - \hat{f}_{k-1}(X_i)$ to obtain the new ϕ and its parameter vector $\hat{\theta}_k$

$$\text{argmax}_\theta \frac{1}{n} \sum_{i=1}^{n} res_i \, \phi(\theta, X_i)$$

- Alternative fit update: $\hat{f}_k$ in span$\{\phi_1, \ldots, \phi_{k-1}, \phi_k\}$
- with coefficients achieving

$$\min_{c_1, c_2, \ldots, c_k} \|Y - \sum_{j=1}^{k} c_j \phi_j\|^2$$

Stopping Criteria

Choice of final k:

- $k = m$ fixed, e.g. m equal a const multiple of $\sqrt{n/(d \log n)}$
- $k = \hat{m}$ chosen by MDL with the relaxed greedy algorithm or forward stepwise regression
- May choose a larger final number of steps m between $\hat{m}$ and n, for implementation of LASSO with control on closeness to the solution

Computation Bound

- Bound the accuracy of greedy computation at step m

  For relaxed greedy and forward stepwise regression

$$\|Y - \hat{f}_m\|^2_{(n)} \;\leq\; \inf_f \{\|Y - f\|^2_{(n)} + \frac{4\|f\|^2_\Phi}{m}$$

Computation Bound (ℓ_1 penalized case)

- Bound on accuracy of ℓ_1 penalized optimization at step m

  For ℓ_1 penalized greedy pursuit (implementation of LASSO)

$$[\|Y-\hat{f}_m\|^2_{(n)}+\lambda\|\hat{f}_m\|_\Phi] \leq \inf_f \left\{ [\|Y - f\|^2_{(n)} + \lambda\|f\|_\Phi] + \frac{4\|f\|^2_\Phi}{m} \right\}$$

Computation bound with rough choices of ϕ)

- Accuracy of computation with rough choice of ϕ each step
- Choose ϕ to achieve $\frac{1}{n}\sum_{i=1}^{n} res_i \phi(X_i)$ at least $(1/C)J_{max}$
- For relaxed greedy and forward stepwise regression

$$\|Y - \hat{f}_m\|_{(n)}^2 \ \le\ \inf_f \{\|Y - f\|_{(n)}^2 + \frac{4C^2\|f\|_\Phi^2}{m}$$

- $C > 1$ is the approximate optimization factor (e.g. $C = 2$)

Non-linear Optimization Step

- maximize

$$J(\theta) = \frac{1}{n}\sum_{i=1}^{n} r_i\,\phi(\theta, X_i)$$

- There may be exponentially many peaks for θ in R^{d+1}
- Exact optimization is NP hard for certain dictionaries Φ (e.g. the neural net case with the step activation function)
- Seek Choices of flexible non-linear dictionaries $\Phi = \{\phi(\theta, x)\}$ for which optimization to within a constant factor is possible by stochastic search strategies

Non-linear Optimization Step

- maximize

$$J(\theta) = \frac{1}{n} \sum_{i=1}^{n} r_i \, \phi(\theta, X_i)$$

- Seek Choices of flexible non-linear dictionaries $\Phi = \{\phi(\theta, x)\}$ for which optimization to within a constant factor is possible by stochastic search strategies
- Try running a Markov Chain, initialized with $\theta \sim p_0(\theta)$ or $p_\epsilon(\theta)$, targeting having long-run distribution

$$p_\gamma(\theta) = \frac{1}{c_\gamma} \exp\{\gamma \, J(\theta)\}$$

- Gain γ of order $d \log d$ would be sufficient for outcomes with $J(\theta) \geq (1/2) J_{max}$ with high probability

Markov Chain Optimization

- FAILURE in high-dimensions of methods that rely on transitions designed for invariance
  - Metropolis-Hastings
  - Simulated Annealing
  - Diffusion with gradient drift

$$d\theta(t) = Drift_t(\theta(t))dt + dB(t)$$

$$\theta(t+\delta) = \theta(t) + Drift_t(\theta(t)\delta + Z(t)\sqrt{\delta}$$

  - Gradient drift

$$Drift_t(\theta) = \frac{1}{2}\nabla \log p_\gamma(\theta) = \frac{\gamma}{2}\nabla J(\theta)$$

- Time until distribution is near $p_\gamma(\theta)$ is exponential in $\gamma \times depth_J$

Optimization by Adaptive Annealing

- SUCCESS for certain Φ of Adaptive Annealing
- Modify the Markov Chain so that the distribution tracks $p_{\gamma_t}(\theta)$ with increasing γ_t
  - A stochastic diffusion with modified drift accomplishes the desired evolution

$$d\theta(t) = Drift_t(\theta(t))dt + dB(t)$$

$$\theta(t+\delta) = \theta(t) + Drift_t(\theta(t)\delta + Z(t)\sqrt{\delta}$$

  - The modified drift is a local gradient plus a simple global change function

$$Drift_t(\theta) = \frac{\gamma}{2}\nabla J(\theta) + change_t(\theta),$$

where we may set $change_t(\theta) = a_t\,\theta$

- Starting from $\gamma_0 = \epsilon$, it tracks

$$p_{\gamma_t}(\theta) = (1/c_{\gamma_t})\exp\{\gamma_t J(\theta)\}$$

Optimization by Adaptive Annealing

- Adaptive Annealing tracks $p_{\gamma_t}(\theta)$ with increasing γ_t
- Stochastic diffusion using drift of the form

$$Drift_t(\theta) = \frac{\gamma}{2}\nabla J(\theta) + change_t(\theta)$$

  solves the Kolmogorov, Fokker-Planck PDE governing the relationship between the drift and the desired evolution of the marginal density of the state θ,

$$\frac{\partial}{\partial t} p_{\gamma_t}(\theta) = -\nabla^T \left(Drift_t(\theta) p_{\gamma_t}(\theta) \right) + \frac{1}{2}\nabla^T \nabla p_{\gamma_t}(\theta)$$

  when $\phi(\theta, x)$ is a ridge polynomial or ridge spline

$$\phi(\theta, x) = \phi(\theta_0 + \theta_1 x_1 + \ldots \theta_d x_d)$$

  with $\phi(z)$ equal to $(z)^q$ or $(z)_+^q$ with $q \geq 2$.
- May set $change_t(\theta) = a_t \theta$ with $a_t = (\log \gamma_t)'/q$.

Optimization by Adaptive Annealing

Clarification:

- For normalizeability, e.g. with $q = 2$, may use

$$p_\gamma(\theta) = \frac{1}{c_\gamma} exp\left\{\gamma\left[\frac{1}{n}\sum r_i\,(\theta^T X_i)_+^2 - \lambda\|\theta\|^2\right]\right\}$$

- Both $(\theta^T X_i)_+^2$ and $\|\theta\|^2$ have the property that they are recovered by taking the inner product of θ with their gradient. Likewise

$$\theta^T \nabla J(\theta) = q\, J(\theta)$$

for

$$J(\theta) = \frac{1}{n}\sum_{i=1}^{n} r_i\,(\theta^T X_i)_+^q \; - \; \lambda\|\theta\|^q$$

- These identities determine the suitability of a multiple of θ as an ingredient in the drift to solve the PDE.

Refinement:

- More general change functions of the form

$$change_t(\theta) = a_t\theta + G_t(\theta)/p_{\gamma_t}(\theta)$$

where

$$\nabla G_t(\theta) = c_t\, p_{\gamma_t}(\theta)$$

also solve the PDE.

- For instance the following is an acceptable choice

$$G_t(\theta) = \int_0^{\theta_0} p_{\gamma_t}(\tilde{\theta}_0, \theta_1, \ldots, \theta_d) d\theta_0$$

expressible as a sum of one-dimensional Gaussian integrals.

- These more general solutions provide more freedom in setting γ_t with favorable growth properties.

Summary

- Ridge splines with adaptive annealing
- Plus an information-theoretic criterion based on ℓ_1 penalized greedy pursuit
- Provides flexible high-dimensional function estimation that is fast and accurate