Paper Review: The Average Posterior Variance of Smoothing Spline and A Consistent Estimate of the Average Squared Error

Douglas Nychka

Group Meeting · Meimei Liu · Feb 02, 2015

Outline

Polynomial Smoothing Spline Estimation

- Consider the regression problem $Y_i = f(x_i) + \varepsilon_i$, $i = 1, \cdots, n$, where $x_i \in [0,1]$ and ε_i are i.i.d errors with mean 0 and variance σ^2. f_λ is the minimizer of

$$\frac{1}{n}\sum_{i=1}^{n}(Y_i - f(x_i))^2 + \lambda \int_0^1 (f^{(m)})^2 dx \text{ for all } f \in W_2^m[0,1]$$

where $W_2^m[0,1]$ is the Sobolev Space

$$W_2^m[0,1] = \{f : f^m \in L^2[0,1] \text{and } f^k, \\ 1 \le k \le m-1 \text{ are absolutely continuous}\}.$$

- A proper inner product makes W_2^m a RKHS with an explicit kernel. For example,

$$(f,g)=\sum_{k=0}^{m-1} f^{(v)}(0)g^{(v)}(0)+\int_0^1 f^{(m)}(x)g^{(m)}(x)dx$$

- Then the reproducing kernel $R(x,y)$ could be expressed as

$$\begin{aligned} R(x,y) &= R_0(x,y)+R_1(x,y) \\ &= \sum_{v=0}^{m-1}\frac{x^v}{v!}\frac{y^v}{v!}+\int_0^1 \frac{(x-u)_+^{m-1}}{(m-1)!}\frac{(y-u)_+^{m-1}}{(m-1)!}du \end{aligned}$$

which generate their corresponding RKHS $\mathcal{H}=\mathcal{H}_0+\mathcal{H}_1$ with

$$\mathcal{H}_0=\{f: f^{(m)}=0\}$$

$$\mathcal{H}_1=\{f: f^{(v)}(0)=0, v=0,\cdots,m-1, \int_0^1 (f^{(m)})^2dx<\infty\}$$

- Write $f \in W_2^m[0,1]$ as $f(x) = \sum_{v=0}^{m-1} d_v \frac{x^v}{v!} + \sum_{i=1}^{n} c_i R_1(x_i, x)$, Then the objective function becomes minimizing

$$(Y - Sd - Qc)^T (Y - Sd - Qc) + n\lambda c^T Qc$$

where S the $n \times m$ matrix with the (i, v)th entry $\frac{x_i^v}{v!}$ and Q the $n \times n$ matrix with the (i, j)th entry $R_1(x_i, x_j)$.

- Differentiating wrt c and d and setting the derivatives to 0, one gets that

$$\hat{Y} = f_\lambda(x) = E(f(x)|Y) = A(\lambda)Y$$

where $A(\lambda) = I - n\lambda(M^{-1} - M^{-1}S(S^T M^{-1} S)^{-1} S^T M^{-1})$ with $M = Q + n\lambda I$.

- $Var[f(x)|Y] = \sigma^2 A(\lambda)$.

Bayes Interpretation of Smoothing Spline Estimator

- So the smoothing spline estimator could be interpreted as the posterior mean when a particular Gaussian prior is placed on the unknown regression function.
- Wahba (1983) used this correspondence between a smoothing spline estimator and the posterior distribution of f to motivate 95% pointwise "confidence intervals" for $f(x_i)$ as

$$f_{\hat{\lambda}}(x_i) \pm 1.96\hat{\sigma}\sqrt{A_{ii}(\hat{\lambda})}$$

where $\hat{\lambda}$ is the minimizer of the GCV function

$$V(\lambda) = \frac{(1/n)||(I - A(\lambda))Y||^2}{((1/n)tr(I - A(\lambda)))^2}$$

- And $\hat{\sigma}^2$ is a consistent estimator of σ^2, with

$$\hat{\sigma}^2 = \frac{||(I - A(\hat{\lambda}))Y||^2}{tr(I - A(\hat{\lambda}))}$$

- The simulation results reported in Wahba (1983) indicate that the Bayesian "confidence" interval work well when evaluated by a frequentist criterion for fixed functions.
- Goal of this paper:
  Understand the statistical properties of this method.

Motivation of this Paper

- (APV) the average posterior variance is given by

$$\frac{\sigma^2}{n}\sum_{i=1}^{n} A_{ii}(\lambda) = \frac{\sigma^2 trA(\lambda)}{n}$$

- (ASE) the Average squared error is given by

$$T_n(\lambda) = \frac{1}{n}\sum_{i=1}^{n}(f_\lambda(x_i) - f(x_i))^2$$

- Wahba hypothesized that the APV is close to the expectation of ASE, i.e $E(T_n(\lambda))$.

Motivation of this Paper

- Let λ^0 be the minimizer of $E(T_n(\lambda))$ for $\lambda \in [0, \infty)$, then Wahba's conjecture is: if $f \in W_2^m[0,1]$, then

$$\frac{\sigma^2 trA(\lambda^0)/n}{E(T_n(\lambda^0))} = \kappa(1 + o(1)) \text{as } n \to \infty$$

  for some $\kappa \in [1, (\frac{2m}{2m-1})(\frac{4m}{4m+1})]$.

- Wahba's conjecture is important since it links a frequency quantity with a functional of the posterior distribution.

- But λ was not fixed in the simulations, it was determined by GCV.

- In this paper, give a proof for a version of Wahba's conjecture that accounts for the adaptive choice of the smoothing parameter λ.

Theorem 1.1

Theorem

Suppose that the observation points $\{x_i\}_{1\leq i\leq n}$ *are a random sample from a distribution with density function g s.t. g is strictly positive on [0,1]. If* $E|\varepsilon_i|^8 < \infty$, $\hat{\lambda}$ *is the minimizer of* $V(\lambda)$ *restricted to* $[\lambda_n, \infty)$ *with* $\lambda_n \sim n^{-4m/5}$, $f \in W_2^{2m}$ *with* $m \geq 2$, *and f satisfies the natural boundary conditions*

$$f^{(k)}(0) = f^{(k)}(1) = 0,\ m \leq k \leq 2m-1$$

then

$$\frac{\hat{\sigma}^2 trA(\hat{\lambda})/n}{E(T_n(\lambda^0))} \xrightarrow{P} K \, as \, n \to \infty$$

where $K = (\frac{2m}{2m-1})(\frac{4m}{4m+1})$.

From Speckman (1983) and Cox (1984), under the condition of Theorem 1.1, if $S_n^2 = 1/n \sum_{i=1}^n \varepsilon_i^2$, then

$$\frac{V(\hat{\lambda}) - S_n^2}{E(T_n(\lambda^0))} \xrightarrow{P} 1 \text{ as } n \to \infty$$

Suppose $\hat{S}_n^2$ is an estimator of S_n^2 s.t. $S_n^2 - \hat{S}_n^2 = o_p(E(T_n(\lambda^0)))$. Then a natural estimate of $E(T_n(\lambda^0))$ is

$$\hat{T}_n = V(\hat{\lambda}) - \hat{S}_n^2$$

Theorem 1.2

Theorem

Under the same hypotheses as Thm 1.1, if

$$\hat{S}_n^2 = \frac{||(I - A(\hat{\lambda}))Y||^2}{tr(I - CA(\hat{\lambda}))} with\, C = 2 - \frac{1}{K}$$

then

$$\hat{S}_n^2 - S_n^2 = o_p(E(T_n(\lambda^0)))\; as\, n \to \infty$$

Remark of Thm 1.2

- Theorem 1.1 follows easily from Theorem 1.2. With some algebra we have

$$\hat{T}_n = V(\hat{\lambda}) - \hat{S}_n^2 = (2-C)\left[\frac{\hat{\sigma}^2 trA(\hat{\lambda})}{n}\right]\left[\frac{1+\beta_n^2/(2-C)}{(1-C\beta_n)(1-\beta_n)^2}\right]$$

  where $\beta_n = trA(\hat{\lambda})/n$.
  - Thus $\hat{T}_n$ is proportional to the estimated APV.
  - Also, from Thm 1.2 and Speckman (1983), Cox (1984), $\hat{T}_n/E(T_n(\lambda^0)) \xrightarrow{P} 1$.
  - The second bracketed term converges to 1 in probability as $n \to \infty$ could be proved in Lemma.

Remark of Thm 1.2

- Recall $\hat{\sigma}^2 = \frac{||(I-A(\hat{\lambda}))Y||^2}{tr(I-A(\hat{\lambda}))}$ is also an estimation of σ, the only difference between $\hat{\sigma}^2$ and $\hat{S}_n^2$ is the constant C in the denominator.
- Although $\hat{\sigma}^2 - S_n^2 = o_p(1)$, under the hypothesis of Theorem 1.1, $(\hat{\sigma}^2 - S_n^2)/E(T_n(\lambda^0)) \xrightarrow{P} \mathscr{E}$ where $\mathscr{E} \neq 1$.
So $V(\hat{\lambda}) - \hat{\sigma}^2$ will not be a consistent estimator of $E(T_n(\lambda^0))$.

General Theorem 2

Theorem 2 includes the results of Thm 1.1 and Thm 1.2. First introduce conditions F1-F3.

- Let G_n denote the empirical distribution for the design points, $\{x_i\}$. Consider two cases:
  - Case A (Designed knots) There is a distribution function G s.t. $\sup_{v\in[0,1]}|G_n(v) - G(v)| = O(\frac{1}{n})$.
  - Case B (Random knots) $\{x_i\}$ is a random from a distribution with c.d.f G.

In either case assume that $g = (d/dv)G$ is strictly positive on [0,1], and $g \in C^\infty[0,1]$.

Conditions for Thm 2

- (F1) $E|\varepsilon|^{2+v} < \infty$ with
  Case A (Designed knots) $v > 4m - 1$
  Case B (Random knots) $v > 2(8m - 3)/5$
- (F2) $\lambda \in [\lambda_n, \infty)$
  Case A (Designed knots) $\lambda_n \approx n^{-4m/5} log(n)$
  Case B (Random knots) $\lambda_n \approx n^{-2m/5} log(n)^m$
- (F3) There is a $\gamma > 0$, s.t.

$$\frac{1}{n}\sum_{i=1}^{n}(Ef_\lambda(x_i) - f(x_i))^2 = \gamma\lambda^2(1 + o(1))$$

uniformly for $lambda \in [\lambda_n, \infty)$

Theorem 2

Theorem

Under (F1)-(F3), if $\hat{S}_n^2 = \frac{||(I-A(\hat{\lambda}))Y||^2}{tr(I-CA(\hat{\lambda}))}$, *and* $\hat{T}_n = V(\hat{\lambda}) - \hat{S}_n^2$, *then*

$$S_n^2 - \hat{S}_n^2 = o_p(E(T_n(\lambda^0)))$$

$$\frac{\hat{T}_n}{E(T_n(\lambda^0))} \xrightarrow{P} 1$$

$$\frac{\hat{\sigma}^2 trA(\hat{\lambda})/n}{E(T_n(\lambda^0))} \xrightarrow{P} K$$

$$where\, K = (\frac{2m}{2m-1})(\frac{4m}{4m+1})\, as\, n \to \infty$$

Sketch of Proof

Let $m_k(\lambda) = \frac{1}{n} tr[A(\lambda)^k]$, and $\mu_k(\lambda) = \alpha I_k \frac{\lambda^{-1/2m}}{n}$ where $I_k = \int_0^\infty \frac{dv}{(1+v^{2m})^k}, \alpha = \frac{\pi}{\int_0^1 (g(v))^{1/2m} dv}$ for $k = 1, 2$.
Let $\mu_k = \mu_k(\lambda^0), m_k = m_k(\lambda^0), A = A(\lambda^0)$, then

$$\hat{S}_n^2 - S_n^2 = [\hat{S}_n^2 - \frac{(1/n)||(I-A)Y||^2}{1-Cm_1}] + [\frac{(1/n)||(I-A)Y||^2 - (1-Cm_1)S_n^2}{1-Cm_1}$$

Since $Y = f + \varepsilon$, and the second term can be expanded by adding and subtracting $-C\mu_1\sigma^2 + 2\mu_1\sigma^2 - \mu_2\sigma^2$ in the numerator to give

$$\frac{R_1 + R_2 + R_3 + \tau}{1 - Cm_1}$$

Need to show each term in the numerator is $o_p(E(T_n(\lambda^0)))$.

- $R_1 = C(m_1 S_n^2 - \mu_1 \sigma^2)$ is $o_p(E(T_n(\lambda^0)))$ by Lemma 3.1.
- $R_2 = -2(\frac{1}{n}\varepsilon' A \varepsilon - \mu_1 \sigma^2) + (\frac{1}{n}\varepsilon' A^2 \varepsilon - \mu_2 \sigma^2)$ is $o_p(E(T_n(\lambda^0)))$ by Lemma 3.4.
- $R_3 = \frac{1}{n} f'(I - A)^2 \varepsilon$ is $o_p(E(T_n(\lambda^0)))$ by Lemma 3.5.
- $\tau = \frac{1}{n}||(I - A)f||^2 - \sigma^2(2 - C)\mu_1 + \sigma^2 \mu_2$ is $o_p(E(T_n(\lambda^0)))$ by Lemma 3.6.