Kernal Estimation (cont.)

- should use an appropriate pointwise variance (ie, variance of gaussian curve representing a particular sample)
  - balance coverage of deadspace vs. redundant coverage of other samples' "turf"

Idea ...? - use system of equations st. each std. is dependent on each other

11/16/2017

Idea - estimate cdf instead, then derive

- order points
  - ~~prob. space value~~ order/n $\Rightarrow$ prob. space value = cdf range
  - sample value $\Rightarrow$ cdf domain
- perform fit to some function space: st.
  - $F(x) \in [0,1] \quad \forall x$
  - $F'(x) \geq 0 \quad \forall x$
  - polynomial
  - ~~harmonic series~~
  - ? something w/ fewer degrees of freedom...?
  - polynomial req's: - ~~$a_{2i} = 0 \ \forall i$?~~ ie, $e^x = 1 + 2x + \ldots$ $\int \cos x + 2 = \int 3 - \frac{x^2}{\ldots} + \frac{x^4}{\ldots}$
  - ~~Legendre polynomials, shifted on $[0,1]$~~
    - valid option
    - orthogonality $\rightarrow$ reduced complexity in curve fitting
  - $\rightarrow$ do not ensure monotonic est.
  - I-splines
    - ensure monotonic-ness
    - smoother result (based on choice of knott-count)

11/20/2017

X 11/22/2017

- Theory - let $X = \{X_1, X_2, \ldots, X_n\}$ be n i.i.d. samples of X,
  - let σ_X be an ordering s.t. $X_{\sigma_X(i)} \leq X_{\sigma_X(i+1)} \ \forall i \in [1, n-1]$,
  - let $p_{\sigma(i)} := P(X < X_{\sigma_X(i)})$
    - $\rightarrow$ the most likely value for $p_{\sigma_X(i)}$ is $\cancel{\frac{i-1}{n-1}}$ $= \frac{i-\frac{1}{2}}{n} = \frac{2i-1}{2n}$
      ~~(since $X_{\sigma(i)}$ has exactly $i-1$ samples less than it)~~
    - $\rightarrow$ specifically, $\mathcal{L}(\theta_i \mid x_i) = \theta_i^{i-1/2}(1-\theta_i)^{n-i+\frac{1}{2}}$
      - $\rightarrow \hat{\theta}_i = \cancel{\frac{i-1}{n-1}} \quad \frac{i-\frac{1}{2}}{n}$

there are $i-1$ values less than $X_{\sigma(i)}$, and $X_{\sigma(i)}$ itself is "split", with $\frac{1}{2}$ in the ">" and "<" sides, so $i-1+\frac{1}{2}$ in total

(cannot simply disregard i'th element - ie, "cherrypick" from samples)

- estimate of cdf should maximize the likelihood of the estimated probability-space value for each sample observed
  - $\rightarrow$? use weights on each sample $\propto \frac{d^2}{d\theta^2} \mathcal{L}_i(\hat{\theta}_i)$?

$i \in [1, n]$

- $\mathcal{L}_i(\theta \mid X) = \theta^{i-\frac{1}{2}}(1-\theta)^{n-i+\frac{1}{2}}$:= likelihood given that:
  - $+\frac{1}{2}$ — $i-1$ samples subceed x_i
  - 1 sample equals x_i
  - $+\frac{1}{2}$ — $n-i$ samples exceed x_i

- $$\frac{d}{d\theta}\mathcal{L}_i(\theta \mid X) = \left(\theta^{i-\frac{1}{2}}\right)\left(-(n-i+\tfrac{1}{2})(1-\theta)^{n-i-\frac{1}{2}}\right) + \left((i-\tfrac{1}{2})\theta^{i-\frac{3}{2}}\right)\left((1-\theta)^{n-i+\frac{1}{2}}\right)$$

$$= \theta^{i-\frac{3}{2}}(1-\theta)^{n-i-\frac{1}{2}}\underbrace{\left(-(n-i+\tfrac{1}{2})\theta + \underbrace{(i-\tfrac{1}{2})(1-\theta)}_{-(i-\frac{1}{2})\theta + (i-\frac{1}{2})}\right)}_{-(n-\cancel{i}+\cancel{\frac{1}{2}}+\cancel{i}-\cancel{\frac{1}{2}})\theta + i - \frac{1}{2}}$$

$$= \theta^{i-\frac{3}{2}}(1-\theta)^{n-i-\frac{1}{2}}\left(i-\tfrac{1}{2}-n\theta\right)$$

$\boxed{\text{note:}}$ $\mathcal{L}'\left(\frac{i-\frac{1}{2}}{n} \mid x_i\right) = 0$

- $$\frac{d^2}{d\theta^2}\mathcal{L}_i(\theta \mid X) = (-n)\left(\theta^{i-\frac{3}{2}}(1-\theta)^{n-i-\frac{1}{2}}\right) + \left(i-\tfrac{1}{2}-n\theta\right)\left(\theta^{i-\frac{5}{2}}(1-\theta)^{n-i-\frac{3}{2}}\left(i-\tfrac{3}{2}+(2-n)\theta\right)\right)$$

$$= \theta^{i-\frac{5}{2}}(1-\theta)^{n-i-\frac{3}{2}}\left(-n\theta(1-\theta) + (i-\tfrac{1}{2}-n\theta)(i-\tfrac{3}{2}-n\theta+2\theta)\right)$$

$$= n\theta^2 - n\theta + i^2 - \tfrac{3}{2}i - in\theta + 2i\theta - \tfrac{1}{2}i + \tfrac{3}{4} + \tfrac{n}{2}\theta - \theta - in\theta + \tfrac{3}{2}n\theta + n^2\theta^2$$

$$= -n\theta^2 + n\theta + n^2\theta^2 - 2in\theta + 2i\theta - \theta - 2i + i^2 + \tfrac{3}{4}$$

$$= (n^2-n)\theta^2 + \underbrace{(n-2in-1+2i)\theta}_{-(n-1)(2i-1)\theta} + (i^2-2i+\tfrac{3}{4})$$

$$= \theta^{i-\frac{5}{2}}(1-\theta)^{n-i-\frac{3}{2}}\left((n^2-n)\theta^2 - (n-1)(2i-1)\theta + (i^2-2i+\tfrac{3}{4})\right)$$

- $$\frac{d^2}{d\theta^2}\mathcal{L}_i\left(\overset{=\hat{\theta}}{\frac{i-\frac{1}{2}}{n}} \,\Big|\, X\right) = \hat{\theta}^{i-\frac{5}{2}}(1-\hat{\theta})^{n-i-\frac{3}{2}}\left(n(n-1)\left(\frac{i-\frac{1}{2}}{n}\right)^2 - (n-1)(2i-1)\left(\frac{i-\frac{1}{2}}{n}\right) + i^2 - 2i + \tfrac{3}{4}\right)$$

$$= \hat{\theta}^{i-\frac{5}{2}}(1-\hat{\theta})^{n-i-\frac{3}{2}}\underbrace{\left(-\frac{n-1}{n}\underbrace{(i-\tfrac{1}{2})^2}_{i^2-i+\frac{1}{4}} + i^2 - 2i + \tfrac{3}{4}\right)}_{=\frac{1}{n}i^2 - \frac{n+1}{n}i + \underbrace{\frac{3}{4} - \frac{n-1}{4n}}_{\frac{2n+1}{4n}}}$$

$$= \underbrace{\hat{\theta}^{i-\frac{5}{2}}(1-\hat{\theta})^{n-i-\frac{3}{2}}}\left(\frac{1}{n}i^2 - \frac{n+1}{n}i + \frac{1}{2} + \frac{1}{4n}\right)$$

$$= \frac{(i-\frac{1}{2})^{i-\frac{5}{2}}(n-i+\frac{1}{2})^{n-i-\frac{3}{2}}}{n^{n-4}}\left(\frac{1}{n}i^2 - \frac{n+1}{n}i + \frac{1}{2} + \frac{1}{4n}\right)$$

$$= \frac{(i-\frac{1}{2})^{i-\frac{5}{2}}(n-i+\frac{1}{2})^{n-i-\frac{3}{2}}}{n^{n-5}}\left(i^2 - (n+1)i + \frac{2n+1}{4}\right)$$

let X be a r.v., $\{X_1, \ldots, X_n\}$ be a set of i.i.d. random variables
and let $X = \{x_1, \ldots, x_n\}$ be an instance of said set
and let $\theta_i = P(X < x_i)$

$\rightarrow$ as $n \to \infty$, $\mathcal{L}_i(\theta_i \mid X) = \theta_i^{\,i-\frac{1}{2}} \cdot (1-\theta_i)^{n-i+\frac{1}{2}}$

$$= \sqrt{\theta_i^{\,2i-1}(1-\theta_i)^{2n-2i+1}} = \sqrt{\theta_i^{\,2i-1}} \cdot \sqrt{\sum_{j=0}^{2n-2i+1} \binom{2n-2i+1}{j} (-\theta_i)^j}$$

- $$\frac{d}{d\theta} \mathcal{L}_i(\theta \mid X) = \frac{1}{2\mathcal{L}_i(\theta|X)} \cdot \left(\frac{2i-1}{\theta} \mathcal{L}_i(\theta|X)^2 + \theta^{2i-1} \sum_{j=1}^{2n-2i+1} j \binom{2n-2i+1}{j} (-\theta)^{j-1} \right)$$

($= 0$ for $\theta = \hat{\theta}_i$)

- $$\frac{d^2}{d\theta^2} \mathcal{L}_i(\theta|X) = \frac{2i-1}{2\theta} \cdot \mathcal{L}_i(\theta|X) + \frac{1}{2\mathcal{L}_i(\theta|X)} \cdot \theta^{2i-1} \cdot \sum_{j=1}^{2n-2i+1} \left(j \cdot \binom{2(n-i)+1}{j} (-\theta)^{j-1} \right)$$

$$= \left(\frac{2i-1}{2\theta} \cdot \mathcal{L}_i'(\theta|X) - \frac{2i-1}{2\theta^2} \mathcal{L}_i(\theta|X) \right) + \left(-\frac{\mathcal{L}_i'(\theta|X)}{2\mathcal{L}_i(\theta|X)^2} \cdot \theta^{2i-1} \cdot \sum_{j=1}^{2n-2i+1} \left(j \cdot \binom{2n-2i+1}{j} (-\theta)^{j-1} \right) \right)$$
$$+ \left(\frac{2i-1}{2\mathcal{L}_i(\theta|X)} \cdot \theta^{2i-2} \cdot \sum_{j=1}^{\cdots} (\ldots) \right) + \left(\frac{\theta^{2i-1}}{2\mathcal{L}_i(\theta|X)} \sum_{j=2}^{2n-2i+1} \left(j(j-1) \binom{2n-2i+1}{j} (-\theta)^{j-2} \right) \right)$$

$$= \left(\frac{2i-1}{2\theta} \right) \left(\mathcal{L}_i'(\theta|X) - \frac{1}{\theta} \mathcal{L}_i(\theta|X) \right) + \left(\frac{\theta^{2i-2}}{2\mathcal{L}_i(\theta|X)} \right) \left(\frac{-\theta \mathcal{L}_i'(\theta|X)}{\mathcal{L}_i(\theta|X)} + 2i-1 \right) \sum_{j=1}^{2n-2i+1} \left(j \binom{2n-2i+1}{j} (-\theta)^{j-1} \right)$$
$$+ \frac{\theta^{2i-1}}{2\mathcal{L}_i(\theta|X)} \sum_{j=2} j(j-1) \binom{2n-2i+1}{j} (-\theta)^{j-2}$$

$$= \frac{2i-1}{2\theta} \left(\mathcal{L}_i'(\theta|X) - \frac{1}{\theta} \mathcal{L}_i(\theta|X) \right) + \frac{\theta^{2i-1}}{2\mathcal{L}_i(\theta|X)} \left[\left(\frac{2i-1}{\theta} - \frac{\mathcal{L}_i'(\theta|X)}{\mathcal{L}_i(\theta|X)} \right) \sum_{j=1}^{2n-2i+1} \binom{2n-2i+1}{j} j (-\theta)^{j-1} + \sum_{j=2}^{2n-2i+1} \binom{2n-2i+1}{j} j(j-1)(-\theta)^{j-2} \right]$$

- $$\frac{d^2}{d\theta^2} \mathcal{L}_i(\hat{\theta}_i|X) = -\frac{2i-1}{2\hat{\theta}^2} \mathcal{L}_i(\hat{\theta}|X) + \frac{\hat{\theta}^{2i-1}}{2\mathcal{L}_i(\hat{\theta}|X)} \left(\frac{2i-1}{\hat{\theta}} \sum_{j=1}^{2n-2i+1} \binom{2n-2i+1}{j} j (-\hat{\theta})^{j-1} + \sum_{j=2}^{2n-2i+1} \binom{2n-2i+1}{j} (j-1) \ldots \right.$$
$$= \frac{\hat{\theta}^{2i-1}}{2\mathcal{L}_i(\hat{\theta}|X)} \left(\frac{2i-1}{\hat{\theta}} \cdot (2n-2i+1) + \sum_{j=2}^{2n-2i+1} \binom{2n-2i+1}{j} j (-\theta)^{j-2} (j - \cancel{1} - 2i + 1) \right) - \frac{2i-1}{2\hat{\theta}^2} \mathcal{L}_i(\hat{\theta}|X)$$