Computational Statistics 11/24/21

More sampling...

Recall Bayesian modeling:

(1) We model data $X_1, \dots, X_n$ as having come from some distribution F with density $f = f(x_1, \dots, x_n; \theta_1, \dots, \theta_k)$ where $\theta_1, \dots, \theta_k$ are parameters defining the distribution. (which we want to estimate)

(2) We specific a prior distribution on $\vec{\theta}$, $f(\vec{\theta})$, which reflects our beliefs (or lack there of) on the value of $\vec{\theta}$.

(3) We compute the posterior, conditional distribution, of $\vec{\theta} \,|\, \vec{X}$ as:

$$p(\vec{\theta} \,|\, \vec{x}) = \underbrace{p(\theta_1, \dots, \theta_k \,|\, x_1 \dots x_n)}_{\text{posterior}} = \frac{f(x_1 \dots x_n \,|\, \theta_1 \dots \theta_k)\, f(\vec{\theta})}{\int f(x_1 \dots x_n \,|\, \theta_1 \dots \theta_k)\, f(\vec{\theta})\, d\vec{\theta}}$$

$$\propto \underbrace{\mathcal{L}(\vec{\theta})}_{\text{likelihood}} \, \underbrace{f(\vec{\theta})}_{\text{prior}}$$

Bayes + MCMC

Recall the Metropolis - Hastings algorithm:

With $q(y|x)$ a proposal density, construct a M. chain for sampling from f as follows:

set $X_{i+1} = \begin{cases} Y & \text{with probability } r(X_i, Y) \\ X_i & \text{with probability } 1 - r(X_i, Y) \end{cases}$

where $r(x,y) = \min\left\{ \frac{f(y)}{f(x)} \frac{q(x|y)}{q(y|x)}, 1 \right\}$

If we wish to sample from the posterior

[Recall, X is the observed data]

$$p(\theta|X) = \frac{\mathcal{L}(\theta)\ f(\theta)}{C(X)}$$

, then we wish to construct a chain $\theta_1, \theta_2, \ldots \theta_i, \ldots$ where

$$\theta_{i+1} = \begin{cases} Y & \text{with prob } r(\theta_i, Y) \\ \theta_i & \text{with prob } 1 - r(\theta_i, Y) \end{cases}$$

and $r(\theta_i, Y) = \min\left\{ \frac{p(Y|X)}{p(\theta_i|X)} \frac{q(\theta_i|Y)}{q(Y|\theta_i)}, 1 \right\}$

$$= \min\left\{ \frac{\mathcal{L}(Y) f(Y)/C(X)}{\mathcal{L}(\theta_i) f(\theta_i)/C(X)} \frac{q(\theta_i|Y)}{q(Y|\theta_i)}, 1 \right\}$$

$$= \min \left\{ \frac{\mathcal{L}(Y)\, f(Y)}{\mathcal{L}(\theta_i)\, f(\theta_i)} \frac{q(\theta_i | Y)}{q(Y|\theta_i)}, 1 \right\}$$

Note: You do <u>not</u> need to compute the normalizing constant $C(x)$ in order to sample from p using MCMC!

Furthermore: If the <u>prior</u> f was <u>uninformative</u>, i.e. <u>flat</u>, $f(x) = C$ and therefore we have

$$r(\theta_i, Y) = \min \left\{ \frac{\mathcal{L}(Y)}{\mathcal{L}(\theta_i)} \frac{q(\theta_i | Y)}{q(Y|\theta_i)}, 1 \right\}.$$

This works if θ is a scalar or vector $\vec{\theta}$.

In the case of $\vec{\theta} \in \mathbb{R}^K$, $\vec{\theta} = (\theta_1, \theta_2, \dots, \theta_K)$ (NOTATION consistent with Efron & Hastie.)

We can use <u>Gibbs Sampling</u>.

<u>Idea</u>: Reduce the generation of $\vec{\theta}$ down to a series of univariate calculations.

Let $\vec{\theta}_{(k)} = (\theta_1, \dots, \theta_{k-1}, \theta_{k+1}, \dots, \theta_K) \in \mathbb{R}^{K-1}$ and

$g_{(k)}$ be conditional density of $\theta_k | \vec{\theta}_{(k)}, \vec{x}$ ← the data

$$\theta_k | \vec{\theta}_{(k)}, \vec{x} \sim g_{(k)}(\theta_k | \vec{\theta}_{(k)}, \vec{x}).$$

Given $\vec{\theta}^{(0)}, \ldots, \vec{\theta}^{(i)}$, to generate the components of

$\vec{\theta}^{(i+1)}$ merely draw from

$$\theta_k^{(i+1)} \sim g_{(k)}\left(\theta_k \mid \vec{\theta}_{(k)}^{(i)}, \vec{x}\right) \quad \text{for } k=1,\ldots,K.$$

I.e. use the previous step's values of $\theta_j^{(i)}$ to condition

and draw $\theta_k^{(i+1)}$.

(See Efron & Hastie § 13.4 for worked example.)

Just a note: Imagine we have $X_1,\ldots,X_n \sim \text{IID } N(\mu,\tau)$

and we wish to sample from the posterior for μ, τ.

We aim to generate a sequence of $\mu^{(1)}, \tau^{(1)}, \mu^{(2)}, \tau^{(2)}, \ldots$

$$p(\mu,\tau \mid \vec{x}) \propto \mathcal{L}(\mu,\tau)\, f(\mu,\tau)$$

$$f(x,y) = f(x \mid y)\, f(y)$$

And to sample, we need

$$p(\mu \mid \tau, \vec{x}) = \frac{p(\mu,\tau \mid \vec{x})}{p(\tau)}$$

$$p(\tau \mid \mu, \vec{x}) = \frac{p(\mu,\tau \mid \vec{x})}{p(\mu)}$$

need to compute them, or be smart about choosing f as conjugate prior, etc...

M-H is more general algorithm...

Monte Carlo Variance Reduction

For any MC method used to estimate

$$I = \int f(x)\,dx \quad , \quad \text{the variance scales as } Var(\hat{I}) = \frac{C}{n}.$$

All one can do is make C smaller.

Analytic Transformations

Consider computing $P(X > 2)$ with $X \sim$ Cauchy.

$$\mathbb{P}(X>2) = I = \int_2^\infty \frac{1}{\pi(1+x^2)}\,dx$$

$$= \int_{-\infty}^\infty \mathbb{1}_{(x>2)} \frac{1}{\pi(1+x^2)}\,dx$$

$$\Rightarrow \hat{I} = \frac{1}{n}\sum_1^n \mathbb{1}_{(x>2)}(x_i)$$

where x_i is a draw from Cauchy distribution.

Alternatively: $$I = \int_0^{1/2} \frac{1/y^2}{\pi(1+1/y^2)}\,dy$$ (by change of variables)

Let $$\tilde{I} = \frac{1}{n}\sum_1^n \frac{1/y_i^2}{\pi(1+1/y_i^2)}$$ where $y_i \sim U(0,1/2)$

$$\Rightarrow \text{one can show } \frac{var(\tilde{I})}{var(\hat{I})} \approx .001$$

Analytic methods for reducing the variance are the best!

Importance Sampling

Any integral $I = \int f(x)\,dx$ can be rewritten as:

$$I = \int \frac{f(x)}{p(x)}\, p(x)\, dx$$

probability density $\int p(x)\,dx = 1$, $p \geq 0$.

Then $\hat{I} = \frac{1}{N} \sum_{j=1}^{N} \frac{f(x_i)}{p(x_i)}$ when $x_i \sim$ IID from p.

Intuitively, to compute $\int f$ we should sample where $|f|$ is large $\rightarrow$ how should we choose p? $\hat{I}$ is clearly unbiased, we should minimize $Var(\hat{I})$.

$$Var(\hat{I}) = \frac{1}{N} Var\left(\frac{f(x)}{p(x)}\right) \quad \text{where } X \sim p.$$

$$Var\left(\frac{f(x)}{p(x)}\right) = \mathbb{E}\left(\frac{f^2(x)}{p^2(x)}\right) - \mathbb{E}\left(\frac{f(x)}{p(x)}\right)^2$$

$$\mathbb{E}\left(\frac{f(x)}{p(x)}\right)^2 = \left(\int \frac{f(x)}{p(x)} p(x)\,dx\right)^2 = \underbrace{\left(\int f(x)\,dx\right)^2}_{\text{does not depend on choice of } p!}.$$

Consider the first term — by Jensen's Inequality:

$$\mathbb{E}\left(\frac{f^2(x)}{p^2(x)}\right) \geq \mathbb{E}\left(\frac{|f(x)|}{p(x)}\right)^2$$

$$= \left(\int |f(x)|\, dx\right)^2$$

$$\Rightarrow \quad \mathbb{E}\left(\frac{f^2(x)}{p^2(x)}\right) \geq \left(\int |f(x)|\, dx\right)^2$$

This bound can be achieved by choosing:

$$p(x) = \frac{|f(x)|}{\int |f(x)|\, dx}$$

then

$$\mathbb{E}\left(\frac{f^2(x)}{p^2(x)}\right) = \mathbb{E}\left(\frac{f^2(x)}{|f(x)|^2}\left(\int |f(x)|\, dx\right)^2\right)$$

$$= \left(\int |f(x)|\, dx\right)^2 \mathbb{E}(1) = \left(\int |f(x)|\, dx\right)^2$$

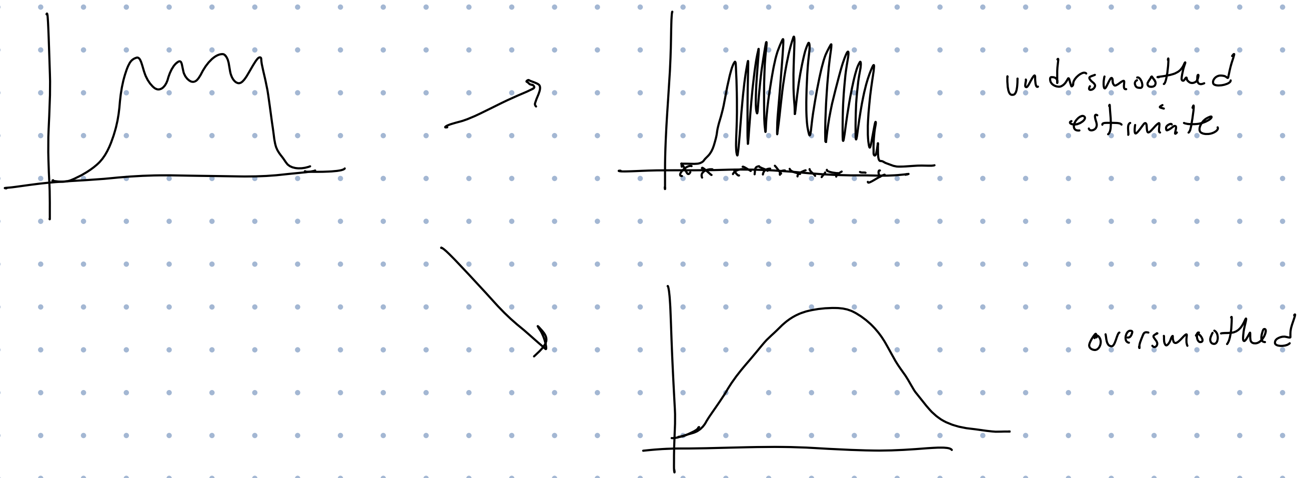
So choose p close to $|f|$, i.e. so that $\frac{|f|}{p} \approx 1$.

Density Estimation

Setup: Observe data $X_1, \dots, X_n \sim F$, and the density is $f = F'$.

Goal: Estimate f using as few assumptions as possible.

$\Rightarrow$ Still a smoothing problem:



One possible measure of the error is the L^2 error:

$$\text{Loss} = L = \int \left(\hat{f}(x) - f(x) \right)^2 dx$$

$$= \| \hat{f} - f \|^2$$

$$= \underbrace{\| \hat{f} \|^2 - 2 (\hat{f}, f)}_{J} + \| f \|^2 = J + \| f \|^2 = J + C.$$

$$(\hat{f}, f) = \text{inner product of } \hat{f} \text{ with } f$$

$$= \int \hat{f}(x) \, f(x) \, dx$$

$$= \int \hat{f}(x) \, dF(x)$$

$$= \mathbb{E} (\hat{f}(x))$$

Goal is to estimate J.

As before, denote by $\hat{f}_{(-i)}$ the estimator obtained by leaving out X_i:

Def: CV estimate of the risk:

$$\hat{J} = \|\hat{f}\|^2 - \frac{2}{n} \sum \hat{f}_{(-i)}(X_i)$$

Histograms

Assume we are estimating f on $[0,1]$, set $h = \frac{1}{m}$, then we have bins $B_1 = [0,h)$, $B_2 = [h, 2h)$, .., $B_j = [(j-1)h, jh]$.

Denote by $Y_j = \#X_i$'s in bin j.

$\hat{p}_j = Y_j/n$. ← estimate probability of ending up in bin j

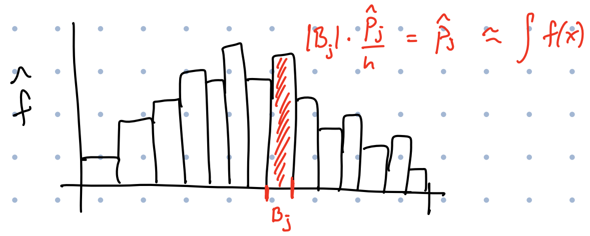
$p_j = \int_{B_j} f(x)\,dx$ ← true probability of landing in bin j.

$= \mathbb{P}(X \in B_j)$.

Histogram estimator: $\hat{f}(x) = \sum_{j=1}^{m} \frac{\hat{p}_j}{h} \mathbb{1}(x \in B_j)$.

Maybe a surprising factor?

Why not just $\hat{p}_j$?



$$\mathbb{E}(\hat{f}(x)) = \frac{\mathbb{E}(\hat{p}_j)}{h} = \frac{p_j}{h} = \frac{1}{h}\int_{B_j} f(x)\,dx \approx \frac{1}{h} f(x) \cdot h = f(x).$$

Thm $\mathbb{E}(\hat{f}(x)) = \frac{p_j}{h}$ for $x \in B_j$

$$\mathrm{Var}(\hat{f}(x)) = \frac{p_j(1-p_j)}{n h^2}$$

and the risk can be computed as:

Thm Assume that f' is "absolutely continuous" and $\int (f')^2 < \infty$, then

$$R(\hat{f}, f) = \frac{h^2}{12} \int (f'(x))^2 dx + \frac{1}{nh} + O(h^2) + O\left(\frac{1}{n}\right)$$

and for fixed n, the minimum occurs at

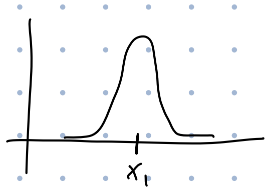
$$h_* = \frac{1}{n^{1/3}} \left(\frac{6}{\int f'^2 dx} \right)^{1/3} \sim \frac{1}{n^{1/3}} \quad \text{and}$$

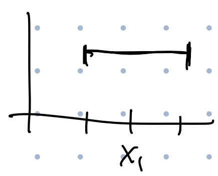
then $R(\hat{f}, f) \sim C \frac{1}{n^{2/3}}$.

Remember: Since f is not known, minimize $\hat{J}$ instead since it can actually be computed.

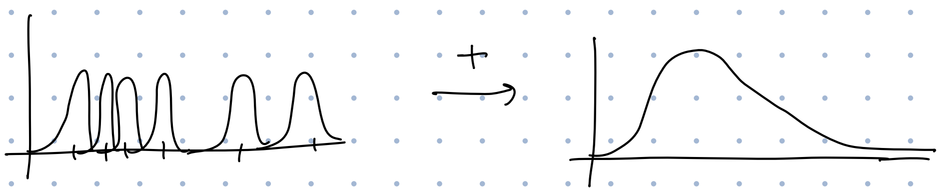
Kernel Density Estimators

If you had only one data point, x_1, what would you do?





Idea: Place a local kernel at each data point, and sum:



How wide should the kernel be?

Def: Kernel density estimator:

$$\hat{f}(x) = \frac{1}{n} \sum_{i=1}^{n} \frac{1}{h} K\left(\frac{x - x_i}{h}\right)$$

$\int K = 1$

$\int x K = 0$

$\sigma_k^2 = \int x^2 K < \infty$

Thm If f is continuous at x, and $h \to 0$, $nh \to \infty$, then $\hat{f}(x) \xrightarrow{\mathbb{P}} f(x)$:

$$\lim_{n \to \infty} \mathbb{P}\left(| \hat{f}(x) - f(x) | > \varepsilon \right) = 0$$

Thm Let $R(x) = \mathbb{E}\left((f(x) - \hat{f}(x))^2 \right)$ be the risk at x. Then

$$R(x) = \frac{1}{4} \sigma_k^4 h^4 f''(x)^2 + \frac{f(x)}{nh} \int K^2(x)\, dx + O\left(\frac{1}{n}\right) + O(h^6).$$