Stein's paradox and the James-Stein Estimator

Let $X_i \sim N(\mu_i, 1),\ i = 1, 2, ..., p$ be jointly independent so we have one data point for each of the p μ_i-parameters. Let $X = (X_1, ..., X_p)$ and $\mu = (\mu_1, ..., \mu_p)$. The MLE $\hat{\mu}$ for μ is $\hat{\mu}_{MLE} = X$. This estimator is inadmissable for quadratic loss.

This is a paradox which forces us to think about the meaning of admissibility, and the implications of Quadratic Loss.

Proof (of the Stein paradox): Consider the alternative estimator

$$\hat{\mu} = \left(1 - \frac{a}{\sum_i X_i^2}\right) X \quad \text{(the James-Stein estimator)}$$

We will show that if $a = p - 2$ then $R(\mu, \hat{\mu}) < R(\mu, \hat{\mu}_{MLE})$ for every $\mu \in R^n$, so that the MLE is inadmissable in this case.

First, the risk for $\hat{\mu}_{MLE}$ is

$$\begin{aligned} R(\mu, \hat{\mu}_{MLE}) &= \sum_{i=1}^{p} \mathbb{E}(|\mu_i - \hat{\mu}_{MLE,i}|^2) \\ &= \sum_{i=1}^{p} \mathbb{E}(|\mu_i - X_i|^2) \\ &= p \end{aligned}$$

recognizing $\mathrm{Var}(X_i) = 1$.

In order to calculate $R(\mu, \hat{\mu})$, it is convenient to use Stein's Lemma, for Normal RV,

$$\mathbb{E}((X_i - \mu)h(X)) = \mathbb{E}\left(\frac{\partial h(X)}{\partial X_i}\right).$$

This can be shown by integrating by parts. Noting

$$\int (x_i - \mu)e^{-(x_i-\mu_i)^2/2}dx = -e^{-(x_i-\mu_i)^2/2}$$

we have

$$\int_{-\infty}^{\infty}(x_i - \mu_i)h(x)e^{-(x_i-\mu_i)^2/2}dx_i = -h(x)e^{-(x_i-\mu_i)^2/2}\Big|_{x_i=-\infty}^{x_i=\infty} + \int_{-\infty}^{\infty}\frac{\partial h(x)}{\partial x_i}e^{-(x_i-\mu_i)^2/2}dx_i$$

The first term is zero if $h(x)$ (for eg) is bounded, giving the lemma.

Now

$$R(\mu,\hat{\mu}) = \sum_{i=1}^{p}\mathbb{E}(|\mu_i-\hat{\mu}_i|^2) \quad \text{with} \quad \hat{\mu}_i = \left(1-\frac{a}{\sum_i X_i^2}\right)X_i$$

$$\begin{aligned}\mathbb{E}(|\mu_i-\hat{\mu}_i|^2) &= \mathbb{E}(|\mu_i-X_i|^2) - 2a\mathbb{E}\left(\frac{(X_i-\mu_i)X_i}{\sum_j X_j^2}\right)\\ &\quad +a^2\mathbb{E}\left(\frac{X_i^2}{(\sum_j X_j^2)^2}\right)\end{aligned}$$

$$\mathbb{E}\left(\frac{(X_i-\mu_i)X_i}{\sum_j X_j^2}\right) = \mathbb{E}\left(\frac{\partial}{\partial X_i}\frac{X_i}{\sum_j X_j^2}\right) \quad \text{Stein's lemma}$$

$$= \mathbb{E}\left(\frac{\sum_j X_j^2 - 2X_i^2}{(\sum_j X_j^2)^2}\right) = \mathbb{E}\left(\frac{1}{\sum_j X_j^2} - 2\frac{X_i^2}{(\sum_j X_j^2)^2}\right)$$

Putting the pieces together,

$$\begin{aligned}\sum_{i=1}^{p}\mathbb{E}(|\mu_i-\hat{\mu}_i|^2) &= R(\mu,\hat{\mu}_{MLE})-(2ap-4a)\mathbb{E}\left(\frac{1}{\sum_j X_j^2}\right)+a^2\mathbb{E}\left(\frac{1}{\sum_j X_j^2}\right)\\ &= p-(2ap-4a-a^2)\mathbb{E}\left(\frac{1}{\sum_j X_j^2}\right)\end{aligned}$$

and this is less than p if $2ap-4a-a^2>0$ and in particular at $a=p-2$, which minimizes the risk over $a\in R$.

We have shown that the obvious (MLE) estimator is inadmissable, and we do better to use an estimator in which data for different μ_i influence the value at other μ_i. This is surprising, given that the data are independent.

We have not shown that the James Stein estimator is admissible.

Empirical Bayes

Bayes estimators have good risk properties (for example, the posterior mean is usually admissible for quadratic loss).

In Empirical Bayes, we use Bayesian reasoning to find estimators which can then be used in classical frequentist inference.

However, Bayes estimators may be hard to compute (for example, the posterior mean is an integral, or sum), particularly for hierarchical models. EB uses a particular strategy to simplify hierarchical models.

Recall the setup for Bayesian inference for hierarchical models.

$$\begin{aligned} X &\sim f(x;\theta) \\ \theta &\sim \pi(\theta;\psi) \\ \psi &\sim g(\psi) \end{aligned}$$

Our prior for θ has a parameter ψ which also has a prior. The posterior is

$$\pi(\theta,\psi|x) \propto L(\theta;x)\pi(\theta;\psi)g(\psi)$$

If we want minimim risk for quadratic loss (for eg) we should use

$$\hat{\theta} = \int \theta\pi(\theta,\psi|x)d\theta d\psi$$

The EB trick is to avoid doing ψ-integrals by replacing ψ with an estimate $\hat{\psi}$, derived from the data, and consider the model

$$\begin{aligned} X &\sim f(x;\theta) \\ \theta &\sim \pi(\theta;\hat{\psi}) \end{aligned}$$

This EB approximation to the full posterior 'chops off' a layer of the hierarchy. The reduced model has posterior

$$\hat{\pi}(\theta|x) \propto L(\theta; x)\pi(\theta; \hat{\psi}),$$

and a Bayes estimator $\hat{\theta}_{EB}$ is calculated using $\hat{\pi}(\theta|x)$. For example, for quadratic loss,

$$\hat{\theta} = \int \theta \hat{\pi}(\theta|x) d\theta.$$

We still need an estimator for ψ. There are several choices. We can use the MLE $\hat{\psi} = \arg\max_{\psi} p(x|\psi)$ for ψ in the marginal likelihood

$$p(x|\psi) = \int L(\theta; x)\pi(\theta; \psi) d\theta.$$

Method of moments estimators are used also.

Example The James-Stein estimator is an EB-estimator.

Data $x_i \sim N(\theta_i, 1)$, $i = 1, \ldots, p$ (so one observation x_i for each parameter θ_i). The MLE for θ_i is simply $\hat{\theta}_{MLE,i} = x_i$. Construct an EB estimator for quadratic loss.

Suppose the prior is $\theta_i \sim N(0, \tau^2)$ (we have some freedom here, as we are mainly interested in the risk-related properties of the final estimator). If we knew τ we have (completing the square)

$$\theta_i | (x_i, \tau) \sim N\left(\frac{x_i \tau^2}{1+\tau^2}, \frac{\tau^2}{1+\tau^2}\right).$$

To get an estimate for τ we compute the marginal distribution for X_i given τ, which is $X_i \sim N(0, \tau^2 + 1)$. The MLE for τ is

then $\hat{\tau}^2 = \frac{1}{p}\sum_i X_i^2 - 1$, and this gives

$$\begin{aligned}\hat{\theta}_{EB,i} &= \frac{X_i\hat{\tau}^2}{1+\hat{\tau}^2} \\ &= \left(1 - \frac{p}{\sum_i X_i^2}\right) X_i\end{aligned}$$

which is the James-Stein estimator

$$\hat{\theta}_{JS,i} = \left(1 - \frac{a}{\sum_i X_i^2}\right) X_i$$

with $a = p$. This isn't the minimum risk JS estimator for quadratic loss, (that is $a = p - 2$) but it already beats the MLE for all θ. (to get the best JS estimator (with $a = p - 2$) use a method of moments estimator for τ. See Young and Smith Section 3.5)

Example

Data $x_i \sim$ Poisson(θ_i), $i = 1, \ldots, n$ (so one observation x_i for each parameter θ_i). The MLE for θ_i is simply $\hat{\theta}_{MLE,i} = x_i$. Construct an EB estimator for quadratic loss.

Suppose the prior for θ_i's is iid Exponential(λ).

$$\begin{aligned} p(x_i \mid \lambda) &= \int_0^\infty \frac{e^{-\theta_i}\theta_i^{x_i}}{x_i!}\lambda e^{-\lambda\theta_i} d\theta_i \\ &= \left(\frac{1}{1+\lambda}\right)^{x_i} \frac{\lambda}{1+\lambda} \end{aligned}$$

Given λ the $x_i's$ are iid geometric $(\lambda/(1+\lambda))$. The MLE of λ based on $x_1, \ldots, x_n$ is $\widehat{\lambda} = 1/\bar{x}$, where $\bar{x} = \frac{1}{n}\sum_1^n x_i$.

Now, under the EB simplification, set $\lambda = \hat{\lambda}$, so that

$$\hat{\pi}(\theta|x) \propto L(\theta; x)\pi(\theta; \hat{\psi}) \;=\; \prod_i e^{-\theta_i}\theta_i^{x_i}\hat{\lambda}e^{-\hat{\lambda}\theta_i}$$

and we recognise $\theta_i|x \sim \Gamma(x_i+1, \hat{\lambda}+1)$ in this EB approximation. This leads to an estimator

$$\begin{aligned}\hat{\theta}_{EB,i} &= \int \theta_i \hat{\pi}(\theta_i|x)d\theta_i \\ &= \frac{x_i+1}{\hat{\lambda}+1} = \bar{x}\frac{x_i+1}{\bar{x}+1}\end{aligned}$$

We can rewrite this

$$\hat{\theta}_{EB,i} = x_i + \frac{\bar{x}}{\bar{x}+1}(\bar{x} - x_i)$$

showing that this EB estimator shrinks the estimates towards the common mean.

Computationally intensive methods

In Bayesian inference, the things we need to know are often given in terms of intractable integrals or sums.

If $x \sim f(x;\theta)$ and we have a prior $\pi(\theta)$ on $\theta \in \Theta$ and we want to measure the evidence for $\theta \in S$, $S \subset \Theta$ then we might calculate

$$P(\theta \in S|x) = \int_S \pi(\theta|x)d\theta$$

where $\pi(\theta|x) \propto f(x;\theta)\pi(\theta)$ is the posterior density. The posterior mean is commonly the most useful point estimate. There we have to compute

$$\mathbb{E}_{\theta|x}\theta = \int_\Theta \theta\pi(\theta|x)d\theta.$$

In the examples so far we have restricted our choice of prior to keep these calculations tractable.

Monte Carlo estimation

$P(\theta \in S|x) = \mathbb{E}_{\theta|x}\mathbb{I}(\theta \in S)$ and $\mathbb{E}_{\theta|x}\theta$ are both averages. If we have N iid samples $\theta^{(i)} \sim \pi(\theta|x),\ i = 1, 2, ..., N$ distributed according to the posterior density and we want to estimate a finite mean $\mu_f = \mathbb{E}_{\theta|x} f(\theta)$ for some given function f, and we take

$$\bar{f}_N = \frac{1}{N} \sum_i f(\theta^{(i)})$$

then

$$\bar{f}_N \xrightarrow{P} \mu_f,$$

that is, $\bar{f}_N$ is consistent for estimation of μ_f (in fact there is usually a CLT). But how do we generate the samples $\theta^{(i)} \sim \pi(\theta|x),\ i = 1, 2, ..., N$?

Markov Chain Monte Carlo estimation

The idea in MCMC is to simulate an ergodic Markov Chain $\theta^{(i)}$ $i = 1, 2, ..., N$ that has $\pi(\theta|x)$ as its stationary distribution. The sequence of simulated θ-values $\theta^{(i)}$ $i = 1, 2, ..., N$ is correlated. However, the ergodic theorem for Markov chains from Part A tells us that

$$\frac{1}{N}\sum_i f(\theta^{(i)}) \xrightarrow{P} \mu_f$$

so we can use the Markov Chain sequence to form estimates. The key here is that we can often write down an ergodic Markov chain that is easy to simulate, and has stationary distribution $\pi(\theta|x)$.

Here $\pi(\theta|x)$ is called the target, and we will drop the "$|x$" while we are setting up the theory.

Markov Chains Suppose Θ is finite. Let

$$P^{(n)}(\theta, \theta') = P(\theta^{(t+n)} = \theta' | \theta^{(t)} = \theta)$$

give the n-step transition probability in a homogeneous Markov chain with transition probability $P(\theta, \theta')$.

Recall the definitions of irreducibility

$$\forall \theta, \theta' \in \Theta, \exists n \text{ such that } P^{(n)}(\theta, \theta') > 0,$$

aperiodicity

$$\forall \theta \in \Theta, \gcd\{n : P^{(n)}(\theta, \theta) > 0\} = 1$$

from Part A. The probability distribution $\pi(\theta)$ is a stationary distribution of a Markov chain if

$$\sum_{\theta \in \Theta} \pi(\theta) P(\theta, \theta') = \pi(\theta').$$

Detailed balance

$$\pi(\theta)P(\theta,\theta') = \pi(\theta')P(\theta',\theta)$$

is an easy sufficient condition for a distribution to be stationary.

Theorem If $\theta^{(t)},\ t = 0, 1, 2, ...$ is an irreducible, aperiodic Markov chain on a finite space Θ, satisfying detailed balance with respect to a target distribution $\pi(\theta)$ on Θ, and $f : \Theta \to R$ is a function with finite mean and variance in π, then

$$\frac{1}{N}\sum_i f(\theta^{(i)}) \xrightarrow{P} \mathbb{E}_\pi f(\theta)$$

[This is essentially a watered down version of the Ergodic Theorem of Part A, with positive recurrence replaced by detailed balance]

The Metropolis algorithm

This is a Markov Chain Monte Carlo (MCMC) algorithm which simulates a transition probability that automatically respected detailed balance with respect to a given target distribution, so the target distribution is the stationary distribution. If the chain is also irreducible and aperiodic, then it will (be ergodic and so) generate the samples we need for our estimates.

Suppose at step t the Markov chain state is $\theta^{(t)} = \theta$. We need an algorithm that generates a random value for $\theta^{(t+1)}$. The strategy is to propose a candidate state θ' according to a fixed symmetric proposal distribution $q(\theta'|\theta)$. We accept the candidate with acceptance probability $\alpha(\theta'|\theta)$ and otherwise the chain stays in the same state.

Metropolis algorithm for target $\pi(\theta)$.

Suppose $\theta^{(t)} = \theta$. Simulate a value for $\theta^{(t+1)}$ as follows.

1. Simulate a candidate state $\theta' \sim q(\theta'|\theta)$ (a simple random change to the state satisfying $q(\theta'|\theta) = q(\theta|\theta')$).

2. With probability $\alpha(\theta'|\theta)$, set $\theta^{(t+1)} = \theta'$ (accept) and otherwise set $\theta^{(t+1)} = \theta$ (reject). Here

$$\alpha(\theta'|\theta) = \min\left\{1, \frac{\pi(\theta')}{\pi(\theta)}\right\}$$

This acceptance probability is chosen so that the overall transition probability from θ to θ'

$$P(\theta, \theta') = q(\theta'|\theta)\alpha(\theta'|\theta)$$

satisfies detailed balance with respect to target distribution $\pi(\theta)$.

Let us check DB. Assume WLOG $\pi(\theta) \geq \pi(\theta')$.

$$\begin{aligned} P(\theta, \theta')\pi(\theta) &= q(\theta'|\theta) \min\left\{1, \frac{\pi(\theta')}{\pi(\theta)}\right\} \pi(\theta) \\ &= q(\theta'|\theta)\pi(\theta') \\ &= q(\theta|\theta') \min\left\{1, \frac{\pi(\theta)}{\pi(\theta')}\right\} \pi(\theta') \\ &= P(\theta', \theta)\pi(\theta'), \end{aligned}$$

since $\min\left\{1, \frac{\pi(\theta')}{\pi(\theta)}\right\} = \pi(\theta')$ and $\min\left\{1, \frac{\pi(\theta)}{\pi(\theta')}\right\} = 1$.

We have shown that the transition probability for the Metropolis update has $\pi(\theta)$ as a stationary distribution. If the chain is also irreducible and aperiodic, it will generate the samples we need. These last two conditions have to be checked separately for each application (but are often obvious).

Example 1

Recall the polling example. In a population N are red, $m - N$ are blue m. We sample k and find $X = x$ red. The prior on N is uniform 0 to m. The posterior for $N = n$ is

$$\pi(n \mid x) \propto n^x (m-n)^{k-x}, \quad n \in \{0, 1, 2, ..., m\}$$

In this example we can work out posterior probabilities exactly. Let us use MCMC to estimate posterior probabilities, and check it matches the exact result (within Monte Carlo error).

We need to write an MCMC algorithm simulating a Markov chain $N^{(t)}, t = 0, 1, 2, ...$ with the property that $N^{(t)} \sim \pi(n \mid x)$.

Metropolis algorithm targeting $\pi(n|x)$. See R code.
Suppose $N^{(t)} = n$. Simulate a value for $N^{(t+1)}$ as follows.

1. Toss a coin and set $n' = n + 1$ on a Head and otherwise $n' = n - 1$ (so $q(n+1|n) = q(n|n+1) = 1/2$).

2. With probability

$$\alpha(n'|n) = \min\left\{1, \frac{\pi(n'|x)}{\pi(n|x)}\right\}$$

set $N^{(t+1)} = n'$ and otherwise set $N^{(t+1)} = n$ (reject).

Note that $\alpha = 0$ if $n' = -1, m+1$ (since $\pi(n'|x) = 0$ in those cases) and otherwise

$$\alpha(n'|n) = \min\left\{1, \frac{(n')^x (m-n')^{k-x}}{n^x (m-n)^{k-x}}\right\}.$$

Example 2. In Q4 of PS3, we wrote down the posterior distribution for a hierarchical model, but were unable to do anything with it, as the integrals were intractable.

The observation model is $X_i \sim B(n_i, p_i), i = 1, 2, ..., k$, the prior $\pi(\theta|a, b)$ is $\mathsf{Beta}(a, b)$ and the prior on a, b is the improper prior $\pi(a, b) \propto 1/ab$ on $a, b > 0$. The posterior density is

$$\pi(p, a, b|x) \propto \frac{1}{ab} \prod_{i=1}^{k} p_i^{x_i}(1 - p_i)^{n_i - x_i} \frac{p_i^{a-1}(1 - p_i)^{b-1}}{B(a, b)},$$

with $B(a, b) = \dfrac{\Gamma(a)\Gamma(b)}{\Gamma(a + b)}$.

Note: we used an improper prior - we need to check this is a proper posterior distribution.

Use Monte Carlo to sample

$$(p^{(t)}, a^{(t)}, b^{(t)}) \sim \pi(p, a, b|x), \quad t = 1, 2, ..., T$$

and use the samples to estimate any posterior expectations of interest.

For example, if we want to estimate a using the posterior mean

$$\hat{a} = \frac{1}{T} \sum_t a^{(t)}.$$

or the posterior probability that $p_1 > c$,

$$\overline{\pi(p_1 > c|x)} = \frac{1}{T} \sum_t \mathbb{I}(p_1^{(t)} > c).$$

Metropolis algorithm targeting $\pi(p, a, b|x)$. See R code.

Suppose $\theta^{(t)} = \theta = (p, a, b)$. Simulate $\theta^{(t+1)}$ as follows.

1. Pick a component of θ, one of $(p_1, , ..., p_k)$, a or b and propose an updated value.

    1.i If we choose p_i, propose $p_i' \sim U(0, 1)$.

    1.ii If we choose a propose $a' \sim U(a - w, a + w)$

    1.iii If we choose b propose $b' \sim U(b - w, b + w)$,

with $w > 0$ a fixed number. This generates a new state θ' with one component changed. Clearly, $q(\theta'|\theta) = q(\theta|\theta')$.

2. With probability $\alpha(\theta'|\theta)$ set $\theta^{(t+1)} = \theta'$ and otherwise set $\theta^{(t+1)} = \theta$ (reject).

Exercise Verify the following formulae.

If we update a p_i (so $\theta' = ((p_1, ..., p_{i-1}, p_i', p_{i+1}, ..., p_k), a, b)$)

$$\alpha(\theta'|\theta) = \min\left\{1, \frac{(p_i')^{a+x_i-1}(1-p_i')^{b+n_i-x_i-1}}{(p_i)^{a+x_i-1}(1-p_i)^{b+n_i-x_i-1}}\right\}$$

whilst if we update a (so that $\theta' = (p, a', b)$)

$$\alpha(\theta'|\theta) = \min\left\{1, \frac{(p_1p_2...p_k)^{a'}B(a,b)^k}{(p_1p_2...p_k)^{a}B(a',b)^k}\frac{a}{a'}\right\}$$

and finally b (where $\theta' = (p, a, b')$)

$$\alpha(\theta'|\theta) = \min\left\{1, \frac{[(1-p_1)(1-p_2)...(1-p_k)]^{b'}B(a,b)^k}{[(1-p_1)(1-p_2)...(1-p_k)]^{b}B(a,b')^k}\frac{b}{b'}\right\}$$

Now run the R code `mcmc.R` accompanying this lecture.