Understanding the James Stein Estimator, as used in Brown 2008. (1)

Model 1 $\begin{cases} X_i \mid \theta_i \overset{ind}{\sim} N(\theta_i, 1) \\ \theta_i \overset{ind}{\sim} N(0, \tau^2) \end{cases}$ θ_i, τ^2 unknown

Goal Data $\{X_i\}$. Estimate θ_i via estimator $\hat{\theta}_i = \delta(X_i)$.

Naive Estimator & MLE $\boxed{\hat{\theta}_i^{(MLE)} = \delta^0(X_i) = X_i}$

Posterior Dist Derivation $p(\theta_i \mid x_i) \propto p(x_i \mid \theta_i)\, p(\theta_i)$ Bayes' Rule

$$= \mathcal{N}(x_i \mid \theta_i, 1) \cdot \mathcal{N}(\theta_i \mid 0, \tau^2)$$

$$\propto \exp\left(-\tfrac{1}{2}(x_i - \theta_i)^2\right) \cdot \exp\left(-\frac{\theta_i^2}{2\tau^2}\right)$$

$$\propto \exp\left(-\tfrac{1}{2}\left(\theta_i^2\left(1 + \tfrac{1}{\tau^2}\right) - 2\theta_i x_i\right)\right)$$

$$= \exp\left[-\tfrac{1}{2}\left(\frac{\tau^2+1}{\tau^2}\right)\left(\theta_i^2 - \left(\frac{\tau^2}{\tau^2+1}\right)2\theta_i x_i\right)\right]$$

$$\propto \exp\left(-\frac{1}{2\lambda}\left(\theta_i - \lambda x_i\right)^2\right)$$

letting $\lambda = \frac{\tau^2}{\tau^2+1}$.

Posterior $\theta_i \mid X_i \sim N(\lambda X_i, \lambda)$ where $\boxed{\lambda = \frac{\tau^2}{\tau^2+1}}$

Bayes Estimator: Posterior Mean $\boxed{\hat{\theta}_i^{(Bayes)} = \delta^*(X_i) = \mathbb{E}(\theta_i \mid X_i) = \lambda X_i}$

$\hat{\theta}_i^{(Bayes)}$ shrinks $\hat{\theta}_i^{(MLE)}$ towards 0.

Problem τ^2, and hence λ, is unknown, so we can't use $\hat{\theta}_i^{(Bayes)}$ directly. Need Empirical Bayes

$\lambda = \frac{\tau^2}{1+\tau^2}$ [redacted]

<u>Goal</u> Use Empirical Bayes estimator $\hat{\lambda}$, with $\boxed{\hat{\theta}_i^{(EB)} = \hat{\lambda} X_i}$

This is "empirical bayes" because we estimate the hyperparameter λ from the data [redacted] $\{X_i\}$.

<u>Marginal</u> $X_i \overset{ind}{\sim} N(0, 1+\tau^2)$

<u>Sum of Squares</u> $S^2 = \|X\|_2^2 = \sum_{i=1}^{p} X_i^2 \overset{d}{=} (\tau^2+1)\chi_p^2$

<u>Lemma</u> $\mathbb{E}\left(\frac{1}{\chi_p^2}\right) = \frac{1}{p-2}$ <u>Pf</u> Next page

<u>Conclusion</u>

$$[redacted]\ (\tau^2+1)\,\mathbb{E}\left(\frac{1}{S^2}\right) = [redacted]\ \mathbb{E}\left(\frac{1}{\chi_p^2}\right) = \frac{1}{p-2}$$

$$\Longrightarrow \mathbb{E}\left(\frac{p-2}{S^2}\right) = \frac{1}{\tau^2+1} = 1-\lambda$$

$\Longrightarrow$ Estimator $\hat{\lambda} = 1 - \frac{p-2}{S^2} = 1 - \frac{p-2}{\|X\|_2^2}$

$\Longrightarrow$ <u>James Stein Estimator</u> $\hat{\theta}_i^{(JS/EB)} = \hat{\lambda} X_i$

$$\boxed{\hat{\theta}_i^{(JS)} = \left(1 - \frac{p-2}{\|X\|_2^2}\right) X_i}$$

[redacted]

$V = \frac{1}{Y}$, $Y \sim \chi^2_p$. $F_V(v) = \mathbb{P}(V \le v) = P(\frac{1}{Y} \le v) = \mathbb{P}(Y \ge \frac{1}{v}) = 1 - F_Y(\frac{1}{v})$

$f_V(v) = \frac{d}{dv} F_V(v) = \frac{1}{v^2} f_Y(\frac{1}{v})$.

Pf (lemma)

$X = Z^2$, $Z \sim N(0,1)$, has χ^2_1 distribution, which for $x > 0$ has CDF

$$F_X(x) = \mathbb{P}(X \le x) = \mathbb{P}(Z^2 \le x) = P(|Z| \le \sqrt{x}) = P(-\sqrt{x} \le Z \le \sqrt{x})$$

$$= \int_{-\sqrt{x}}^{\sqrt{x}} \frac{1}{\sqrt{2\pi}} e^{-\frac{t^2}{2}} dt = 2\int_0^{\sqrt{x}} \frac{1}{\sqrt{2\pi}} e^{-\frac{t^2}{2}} dt.$$

Then $X \sim \chi^2_1$ has density since $\Gamma(1/2) = \sqrt{\pi}$.

$$f_X(x) = \frac{d}{dx} F_X(x) = \frac{2}{\sqrt{2\pi}} e^{-\frac{x}{2}} \cdot \frac{1}{2\sqrt{x}} = \frac{1}{2^{1/2}\Gamma(\frac{1}{2})} x^{-1/2} e^{-x/2}, \quad x > 0$$

Hence $\chi^2_1 \sim \text{Gamma}(\frac{1}{2}, \frac{1}{2})$.

MGF of $\text{Gamma}(\alpha, \beta)$ is $M_G(t) = \mathbb{E}e^{Gt} = (1 - \frac{t}{\beta})^{-\alpha}$

Hence $M_{\chi^2_1}(t) = (1-2t)^{-1/2}$

Hence $M_{\chi^2_p}(t) = \prod_{k=1}^{p} M_{(\chi^2_1)_k}(t) = \prod_{k=1}^{p} (1-2t)^{-1/2} = (1-2t)^{-p/2} \Longrightarrow \chi^2_p \sim \text{Gamma}(\frac{p}{2}, \frac{1}{2})$.

Hence χ^2_p has density $f_{\chi^2_p}(y) = \frac{(1/2)^{p/2}}{\Gamma(p/2)} y^{\frac{p}{2}-1} e^{-\frac{y}{2}}$.

So, a Random Variable $X \sim \frac{1}{\chi^2_p}$ has density $f(x) = \frac{(1/2)^{p/2}}{\Gamma(p/2)} x^{-\frac{p}{2}-1} e^{-\frac{1}{2x}}$, $x > 0$.

It is a proper distribution, so $\int_0^\infty x^{-\frac{p}{2}-1} e^{-\frac{1}{2x}} dx = \Gamma(p/2)\, 2^{p/2}$

Hence $$\mathbb{E}X = \int_0^\infty x f(x) dx = \int_0^\infty x \frac{(1/2)^{p/2}}{\Gamma(p/2)} x^{-\frac{p}{2}-1} e^{-\frac{1}{2x}} dx$$

$$= \frac{(1/2)^{p/2}}{\Gamma(p/2)} \int_0^\infty x^{-(\frac{p-2}{2})-1} e^{-\frac{1}{2x}} dx = \frac{(1/2)^{p/2}}{\Gamma(p/2)} \cdot \Gamma\left(\frac{p-2}{2}\right) 2^{\frac{p-2}{2}}$$

$$= \frac{1}{2} \frac{\Gamma(\frac{p}{2}-1)}{\Gamma(\frac{p}{2})} = \frac{1}{2} \frac{\Gamma(\frac{p}{2}-1)}{(\frac{p}{2}-1)\Gamma(\frac{p}{2}-1)} \quad \text{using } \Gamma(s) = (s-1)\Gamma(s-1).$$

$$= \frac{1}{2} \cdot \frac{2}{p-2} = \frac{1}{p-2} \quad \square$$

<u>Shocking Thm</u> For $p \geq 3$, the James Stein estimator $\hat{\theta}^{(JS)}$ everywhere dominates the MLE $\hat{\theta}^{(MLE)}$ in terms of expected total squared error. That is, for all choices of θ,

$$\mathbb{E}_\theta \| \hat{\theta}^{(JS)} - \theta \|_2^2 < \mathbb{E}_\theta \| \hat{\theta}^{(MLE)} - \theta \|_2^2$$

<u>Note</u> This is a frequentist, not a Bayesian, result. $\hat{\theta}^{(JS)}$ is superior no matter what one's prior beliefs about θ may be.

<u>Pf</u> Notice $(\hat{\theta}_i - \theta_i)^2 = (X_i - \hat{\theta}_i)^2 - (X_i - \theta_i)^2 + 2(\hat{\theta}_i - \theta_i)(X_i - \theta_i)$.

Sum over $i = 1, \ldots, p$ and take Expectations:

$$\mathbb{E}_\theta \| \theta - \hat{\theta} \|_2^2 = \mathbb{E}_\theta \| X - \hat{\theta} \|_2^2 - p + 2 \sum_{i=1}^{p} \mathrm{cov}_\theta(\hat{\theta}_i, X_i)$$

cov_θ indicates covariance under $X \sim N_p(\vec{\theta}, I_p)$.

<u>Lemma</u> $\mathrm{cov}_\theta(\hat{\theta}_i, X_i) = \mathbb{E}_\theta\left(\frac{\partial \hat{\theta}_i}{\partial X_i}\right)$ (See <u>Stein's Lemma</u> on next page) (Just integration by parts)

Thus

$$\mathrm{cov}_\theta(\hat{\theta}_i, X_i) = \mathbb{E}_\theta \frac{\partial}{\partial X_i}\left[\left(1 - \frac{p-2}{\sum_{k=1}^{p} X_k^2}\right) X_i\right]$$

$$= \mathbb{E}_\theta\left[\left(1 - \frac{p-2}{\|X\|_2^2}\right) + X_i \left(\frac{(p-2)\, 2 X_i}{\|X\|_2^4}\right)\right]$$

$$= 1 - (p-2)\, \mathbb{E}\left(\frac{1}{\|X\|_2^2}\right) + 2(p-2)\, \mathbb{E}_\theta\left(\frac{X_i^2}{\|X\|_2^4}\right)$$

$\longrightarrow$

Thus $2\sum_{i=1}^{p} \mathrm{cov}_\theta(\hat{\theta}_i, X_i) = 2\left[p - p(p-2)\,\mathbb{E}\left(\frac{1}{\|X\|_2^2}\right) + 2(p-2)\,\mathbb{E}\left(\frac{\|X\|_2^2}{\|X\|_2^4}\right)\right]$

$$= 2p - 2(p-2)^2\,\mathbb{E}\left(\frac{1}{\|X\|_2^2}\right)$$

$$= 2p - 2\,\mathbb{E}_\theta\left[\frac{(p-2)^2}{\|X\|_2^2}\right]$$

and $\mathbb{E}_\theta\|X - \hat{\theta}^{(JS)}\|_2^2 = \sum_{i=1}^{p} \mathbb{E}_\theta\left(X_i - \hat{\theta}_i^{(JS)}\right)^2$

$$= \sum_{i=1}^{p} \mathbb{E}_\theta\,(p-2)^2\frac{X_i^2}{\|X\|_2^4} = \mathbb{E}_\theta\left(\frac{(p-2)^2}{\|X\|_2^2}\right)$$

Therefore $\mathbb{E}_\theta\|\theta - \hat{\theta}^{(JS)}\|_2^2 = \mathbb{E}_\theta\left(\frac{(p-2)^2}{\|X\|_2^2}\right) - p + 2p - 2\,\mathbb{E}_\theta\left(\frac{(p-2)^2}{\|X\|_2^2}\right)$

$$= p - \mathbb{E}_\theta\left(\frac{p-2}{\|X\|_2^2}\right)$$

$$< p = \sum_{i=1}^{p} 1 = \sum_{i=1}^{p}\mathbb{E}_\theta(X_i - \theta_i)^2 = \mathbb{E}_\theta\|\theta - \hat{\theta}^{(MLE)}\|_2^2 \quad \square$$

assuming $p \geq 3$.

<u>Stein's lemma</u> Let $X \sim N(\theta, \sigma^2)$, g differentiable function, $\mathbb{E}|g'(X)| < \infty$.

Then $\boxed{\mathbb{E}\left[g(X)(X-\theta)\right] = \sigma^2 \, \mathbb{E}\, g'(X)}$

<u>Pf</u>

$$\mathbb{E}\left[g(X)(X-\theta)\right] = \frac{1}{\sqrt{2\pi}\sigma}\int_{-\infty}^{\infty} g(x)(x-\theta)\, e^{-\frac{(x-\theta)^2}{2\sigma^2}}\,dx.$$

$$\left[\begin{array}{ll} u = g(x) & dv = (x-\theta)\, e^{-\frac{(x-\theta)^2}{2\sigma^2}} \\ du = g'(x)\,dx & v = -\sigma^2 e^{-\frac{(x-\theta)^2}{2\sigma^2}} \end{array}\right]$$

$$= \underbrace{\frac{1}{\sqrt{2\pi}\sigma}\left[-\sigma^2 g(x)\, e^{-\frac{(x-\theta)^2}{2\sigma^2}}\right]_{-\infty}^{\infty}}_{=0 \text{ since } \mathbb{E}|g'(X)| < \infty} + \sigma^2 \int_{-\infty}^{\infty} g'(x)\frac{e^{-\frac{(x-\theta)^2}{2\sigma^2}}}{\sqrt{2\pi}\sigma}\,dx$$

$$= \sigma^2\, \mathbb{E}\, g'(X)$$

$\square$

Model 2 $\begin{cases} X_i \mid \theta_i \overset{ind}{\sim} N(\theta_i, \sigma_i^2) \\ \theta_i \overset{ind}{\sim} N(0, \tau^2) \end{cases}$ σ_i^2 known, $i = 1, \ldots, p$

Data Transformation $\begin{cases} \tilde{X}_i = \dfrac{X_i}{\sigma_i} \\ \tilde{\theta}_i = \theta_i / \sigma_i \end{cases}$ so that $\begin{cases} \tilde{X}_i \mid \tilde{\theta}_i \overset{ind}{\sim} N(\tilde{\theta}_i, 1) \\ \tilde{\theta}_i \overset{ind}{\sim} N(0, \tau^2/\sigma_i^2) \end{cases}$

We know the James Stein estimator in this situation!

$$\widehat{\tilde{\theta}}_i^{(JS)} = \left(1 - \frac{p-2}{\|\tilde{X}\|_2^2}\right) \tilde{X}_i = \left(1 - \frac{p-2}{\sum_{k=1}^{p} X_k^2/\sigma_k^2}\right) \frac{X_i}{\sigma_i}$$

Define $\hat{\theta}_i^{(JS)} = \sigma_i \widehat{\tilde{\theta}}_i^{(JS)}$

Then $\boxed{\hat{\theta}_i^{(JS)} = \left(1 - \dfrac{p-2}{\sum_{k=1}^{p} X_k^2/\sigma_k^2}\right) X_i}$

this is the Batting Avg model used in Brown (2008)

$$\boxed{\text{Model 3} \quad \begin{cases} X_i \mid \theta_i \overset{ind}{\sim} N(\theta_i, \sigma_i^2) \\ \theta_i \overset{ind}{\sim} N(\mu, \tau^2) \end{cases} \quad \begin{matrix} \sigma_i^2 \text{ known} \\ i=1,\dots,p \end{matrix}}$$

Data Transformation $\begin{cases} \tilde{X}_i = X_i - \mu \\ \tilde{\theta}_i = \theta_i - \mu \end{cases}$ so that $\begin{cases} \tilde{X}_i \mid \tilde{\theta}_i \overset{ind}{\sim} N(\tilde{\theta}_i, \sigma_i^2) \\ \tilde{\theta}_i \overset{ind}{\sim} N(0, \tau^2) \end{cases}$

We know the James Stein estimator in this situation!

$$\hat{\tilde{\theta}}_i^{(JS)} = \left(1 - \frac{p-2}{\sum_{k=1}^{p} \tilde{X}_k^2/\sigma_k^2}\right)\tilde{X}_i = \left(1 - \frac{p-2}{\sum_{k=1}^{p} (X_k-\mu)^2/\sigma_k^2}\right)(X_i - \mu)$$

Define $\hat{\theta}_i^{(JS)} = \mu + \hat{\tilde{\theta}}_i^{(JS)}$

Then $$\boxed{\hat{\theta}_i^{(JS)} = \mu + \left(1 - \frac{p-2}{\sum_{k=1}^{p} (X_k-\mu)^2/\sigma_k^2}\right)(X_i - \mu)}$$

Note $\hat{\theta}_i^{(JS)}$ shrinks the MLE X_i towards the mean μ.

Empirical Bayes Because μ is unknown, we estimate μ from the data as $\hat{\mu}$, and use $\hat{\mu}$ in place of μ in the above $\hat{\theta}_i^{(JS)}$

<u>Model</u> $\begin{cases} X_i \mid \theta_i \overset{ind}{\sim} N(\theta_i, \sigma_i^2) \\ \theta_i \overset{ind}{\sim} N(\mu, \tau^2) \end{cases}$ σ_i^2 Known

<u>Marginal</u> $X_i \overset{ind}{\sim} N(\mu, \tau^2 + \sigma_i^2)$

<u>Goal</u> Obtain estimate $\hat{\mu}$ of μ to use in $\hat{\theta}_i^{(JS)}$ [empirical bayes!]

<u>MLE $\hat{\mu}_{MLE}$</u>

i^{th} likelihood $P(X_i) = \frac{1}{\sqrt{2\pi(\tau^2+\sigma_i^2)}} e^{-\frac{(X_i-\mu)^2}{2(\tau^2+\sigma_i^2)}}$

Full log-likelihood $\ell(\vec{X} \mid \mu, \tau^2, \vec{\sigma}^2) = \sum_{i=1}^{p} \log P(X_i) = -\frac{1}{2} \sum_{i=1}^{p} \frac{(X_i-\mu)^2}{(\tau^2+\sigma_i^2)} + C(\tau^2, \vec{\sigma}^2)$

$$\frac{\partial \ell}{\partial \mu} = \sum_{i=1}^{p} \frac{X_i - \mu}{\tau^2+\sigma_i^2}$$

$$\frac{\partial \ell}{\partial \mu}(\hat{\mu}) = 0 \implies \sum_{i=1}^{p} \frac{X_i}{\tau^2+\sigma_i^2} = \hat{\mu} \sum_{i=1}^{p} \frac{1}{\tau^2+\sigma_i^2}$$

$$\implies \boxed{\hat{\mu}_{MLE} = \frac{\sum X_i/(\tau^2+\sigma_i^2)}{\sum 1/\tau^2+\sigma_i^2}}$$

However, τ^2 is unknown, so we must use an estimate $\hat{\tau}^2$ for τ^2.

Brown simply uses $\tau^2 = 0$, to make life simple, yielding the estimate

$$\boxed{\hat{\mu}_1 = \frac{\sum X_i/\sigma_i^2}{\sum 1/\sigma_i^2}}$$

This corresponds to the model $\begin{cases} X_i \mid \theta_i \sim N(\theta_i, \sigma_i^2) \\ \theta_i \equiv \mu \end{cases}$ which is to assume each batter has a common batting average mean.