Shrinkage and the James-Stein Estimator for Normal Means

Guy Lebanon

In this note we show the inadmissibility of the standard estimator for normal means and introduce the James-Stein estimator. We generally follow [1] though our exposition is informal and omits some of the technical conditions. The tutorial note on Stein's unbiased risk estimator is a pre-requisite for this note.

We are concerned with estimating the mean vector of a m-dimensional Gaussian $N(\theta, I)$ given a sample $X \sim N(\theta, I)$ under squared loss $R(\theta, \delta) = \mathsf{E}_{p_\theta} \|\delta(X) - \theta\|^2$. Note that since the covariance is I we may view this problem as estimating the means of several independent one dimensional Gaussians given a single observation from each. Stein showed the unintuitive result that the standard estimator $\delta_0(X) = X$ is inadmissible when $m \geq 3$ (there exists another estimator that performs as good on all θ and better on some θ). The main component of the proof is SURE (Stein's unbiased risk estimator) which is described in an earlier tutorial note.

Proposition 1. *Assuming $X \sim N(\theta, I)$ with $dim(X) \geq 3$, the estimator $\delta_0(X) = X$ for θ is inadmissible under the squared loss function and is in fact dominated by the following estimator*

$$\delta_r(X) = \left(1 - \frac{r(\|X\|)}{\|X\|^2}\right) X \tag{1}$$

where r is a continuous non-decreasing function that satisfies $0 \leq r(t) \leq 2(m-2)$.

Proof. We recall the following result concerning $p_\theta(x) = h(x) \exp(\theta^\top x - \psi(x))$ from Stein's unbiased risk estimator (see earlier tutorial note for more details and a derivation)

$$R(\theta, \delta_1) - R(\theta, \delta_2) = \mathsf{E}_{p_\theta} \left(\|\delta_1(X)\|^2 - \|\delta_2(X)\|^2 + 2\nabla \cdot (\delta_1(X) - \delta_2(X)) + 2\frac{(\delta_1(X) - \delta_2(X))^\top \nabla h(X)}{h(X)} \right). \tag{2}$$

To apply this to our case we need to express a Gaussian with $\Sigma = I$ in exponential family notation $h(x) \exp(\theta^\top x - \psi(x))$

$$p_\theta(X) = (2\pi)^{-m/2} \exp\left(-(X-\mu)^\top (X-\mu)/2\right) = (2\pi)^{-m/2} \exp\left(\mu^\top X + X^\top X/2 - \mu^\top \mu/2\right). \tag{3}$$

Note also that $\log h = -X^\top X$, $\nabla h / h = -X$ which leads to

$$\|\delta_0\|^2 - \|\delta_r\|^2 + 2\frac{(\delta_0(X) - \delta_r(X))^\top \nabla h(X)}{h(X)} = 2r(\|X\|) - \frac{r^2(\|X\|)}{\|X\|^2} - 2\frac{r(\|X\|)}{\|X\|^2}\|X\|^2 = -\frac{r^2(\|X\|)}{\|X\|^2}$$

which together with (2) gives

$$R(\theta, \delta_0) - R(\theta, \delta_r) = \mathsf{E}_{p_\theta} \left(2\nabla \cdot \frac{r(\|X\|)X}{\|X\|^2} - \frac{r^2(\|X\|)}{\|X\|^2} \right). \tag{4}$$

We substitute the following simplifications

$$\nabla \cdot \frac{X}{\|X\|^2} = \sum_{i=1}^m \frac{\|X\|^2}{\|X\|^4} - \sum_{i=1}^m \frac{2X_i^2}{\|X\|^4} = \frac{m-2}{\|X\|^2}$$

$$\nabla r(\|X\|) = r'(\|X\|)\frac{X}{\|X\|}$$

in (5) to get

$$R(\theta, \delta_0) - R(\theta, \delta_r) = \mathsf{E}_{p_\theta}\left((2(m-2) - r(\|X\|))\frac{r(\|X\|)}{\|X\|^2} + 2\frac{r'(\|X\|)}{\|X\|}\right). \tag{5}$$

We now observe that if (as we stated above) r is non-decreasing and $0 \leq r \leq 2(m-2)$ then (5) is non-negative which implies that δ_r has a risk that is less than or equal the risk of δ_0 for all θ. Note also that if r is strictly increasing or if $0 \leq r < 2(m-2)$ then δ_r actually dominate δ_0 for some θ which proves the inadmissibility of δ_0. □

Comments

1. It may be shown that δ_0 is admissible for $k \leq 2$ i.e., δ_0 is inadmissible iff $k \geq 3$.

2. The first estimator of James and Stein was δ_r with $r \equiv m-2$ which may be shown to be the best possible constant value of r. That estimator is in fact also inadmissible and is dominated by

$$\delta(X) = \left(1 - \frac{m-2}{\|X\|^2}\right)_+ X \tag{6}$$

   (where $A_+ = A$ if $A > 0$ and 0 otherwise) which corresponds to δ_r with $r(t) = \min(t^2, m-2)$. This estimator clearly pushes the estimated value away from X towards 0 with the amount increasing with the dimensionality m and decreasing with $\|X\|$. This explains why James Stein estimators are often called Shrinkage estimators.

3. It is possible to interpret the estimation problem as estimating the means of m independent Gaussians. Viewed in this light we have a remarkable phenomenon-it is helpful to use an observation from one Gaussian when estimating the mean of another *independent* Gaussian (since the term $\|X\|$ in $[\delta(X)]_i$ contains all components not just i). A frequent example is estimating batting averages in baseball from a single game for several hitters in different teams and perhaps even different leagues. Paradoxically, it turns out to be beneficial to incorporate the performance of hitter i when estimating the batting average of hitter j. Why should an excellent hitter have their batting average estimate shrunk towards zero based on the performance of poorer hitters? This would be less surprising in a Bayesian setting where θ has a prior that ties different component together. But the setting above is completely frequentist!

4. For some choices of r the regularity conditions in the proposition do not hold. For example, the original JS estimator δ_r with $r = m-2$ is discontinuous at 0. This difficulty may be overcome by replacing $r(t)$ with $r_\epsilon(t) = \min(t/\epsilon, r(t))$ and noting that if the result holds for δ_{r_ϵ} for all $\epsilon > 0$ it should also hold for δ_r.

References

[1] L. D. Brown. *Fundamentals of Statistical Exponential Families*, volume 9 of *Lecture Notes-Monograph Series*. IMS Press, 1986.