The James-Stein Estimator of the mean $\mu$ of $N_p(\mu, \Sigma)$.

Observe $Y_1, \ldots, Y_n$ i.i.d. $\sim N_p(\mu, \Sigma)$. Let $\overline{Y} = \frac{1}{n} \sum Y_i \sim N_p(\mu, \Sigma)$. Then:

(i) $\overline{Y}$ is the MLE of $\mu$.

(ii) $\overline{Y}$ is the minimum variance unbiased estimator (MVUE) of $\mu$:

Among all unbiased estimators $\hat{\mu}$ of $\mu$, $\overline{Y}$ minimizes the mean square error (MSE): $E_p \| \mu - \hat{\mu} \|^2$.

(iii) $\overline{Y}$ is the best (MSE) translation-equivariant estimator of $\hat{\mu}$:

$\hat{\mu}(\overline{Y} + \alpha) = \overline{Y} + \alpha \quad \forall \alpha \in \mathbb{R}^p$.

(iv) For $p = 1, 2$, $\overline{Y}$ is an admissible estimator of $\mu$: no other estimator (linear or not) has everywhere better MSE than $\overline{Y}$.

However, Stein (1956) showed that for $p \geq 3$, $\overline{Y}$ is inadmissible.

James-Stein (1962) produced a (biased, non-linear) estimator $\hat{\mu}_J$ whose MSE is everywhere less than the MSE of $\overline{Y}$, and substantially less for $\mu$ near 0. [Question: but is 0 special?] (The difference in MSE's $\to 0$ as $\|\mu\| \to \infty$.)

We shall consider only the special case where $\Sigma$ is known (for simplicity) but similar results hold if $\Sigma$ is unknown. Thus we can assume that $n = 1$.

Observe $Y \sim N_p(\mu, \Sigma)$, or known. Then $\hat{\mu} = \overline{Y}$.

The MSE of $\hat{\mu}$ is

(i) $E_p \| \hat{\mu} - \mu \|^2 = E_p \| Y - \mu \|^2 = E_p(\Sigma \Sigma^T) = \sigma^2 I_p$. 

Where $\mu$ is unknown.
The JS estimator has the form $\hat{\theta}_{JS} = a(\|Y\|) \hat{\theta}$, where $a(\|Y\|)$ is a scalar-valued function of $\|Y\|$. Note that $\hat{\theta}_{JS}$ is not translation-equivariant, but it is orthogonally equivariant (as is $\hat{\theta}$):

$$\hat{\theta}_{JS} (Y) = \hat{\theta}_{JS} (Q Y) \quad \forall Q: \text{proj orthogonale,}$$

The multiplier $a(\|Y\|)$ has the form $1 - \frac{1}{\|Y\|^2}$, hence "shrinks" $\hat{\theta}$ toward $Q$ (Sometimes, it shrinks $\hat{\theta}$ beyond $Q$, so max $\alpha(\|Y\|)$, $Q\hat{\theta}$ is preferable, but its MSE is hard to calculate analytically.) Here is a simplified version of Stein's heuristic explanation of why shrinking $\hat{\theta}$ toward $Q$ might reduce the MSE:

![Diagram](image)

By symmetry, the tangent hyperplane divides $\mathbb{R}^p$ into 2 half-spaces each having probability $\frac{1}{2}$ under $N_p(\mu, \sigma^2 I_p)$.

$$\Pr[\hat{\theta} \in \text{sphere of radius } \|Y\| \text{ in } \mathbb{R}^p] = \frac{1}{2},$$

so $\Pr[\|Y\| > \|\hat{\theta}\|] > \frac{1}{2}$. This implies that $\|Y\|$ is "too large," so to estimate $\mu$ we should "shrink" $\hat{\theta}$ toward $Q$. [but $Q$ is arbitrary!]

More precisely:
\[ \| Y \|_2^2 \sim \sigma_X^2 \left( \frac{\| M \|_2^2}{\sigma^2} \right), \] so \[ E \| Y \|_2^2 = \sigma^2 \left[ \mu + \frac{\| M \|_2^2}{\sigma^2} \right] = \mu \sigma^2 + \| Y \|_2^2 > \| M \|_2^2. \]

Thus \( \| Y \|_2^2 \) is too large an estimate of \( \| M \|_2^2 \) — in fact, \( \| Y \|_2^2 - \mu \sigma^2 \) is an unbiased estimator of \( \| M \|_2^2 \).

The JSE can be motivated by an "empirical Bayesian" argument. Suppose we make the Bayesian assumption that \( M \) is also random, with prior distribution
\[
(3) \quad M \sim N_p \left( 0, \lambda^2 I_p \right). \quad \text{[we could replace 0 by any } \mu \text{]}
\]

Then we should rewrite (1) as \( Y \mid M \sim N \left( M, \sigma^2 I_p \right) \). From these two facts, (the marginal distr of \( M \) and the conditional distr of \( Y(M) \)), we can deduce that the joint distribution of \( (Y, M) \) is the following:
\[
(4) \quad \begin{bmatrix} Y \\ M \end{bmatrix} \sim N \left( \begin{bmatrix} 0 \\ 0 \end{bmatrix}, \begin{bmatrix} \sigma^2 + \lambda^2 I_p & \lambda^2 I_p \\ \lambda^2 I_p & \lambda^2 I_p \end{bmatrix} \right). \quad \text{[Exercise]}
\]

Therefore the posterior distribution of \( M \mid Y \) (i.e., the conditional distr) is
\[
(5) \quad M \mid Y \sim N \left( \begin{bmatrix} \frac{\lambda^2}{\sigma^2 + \lambda^2} Y \\ \frac{\lambda^2}{\sigma^2 + \lambda^2} \right), \begin{bmatrix} \frac{\lambda^2}{\sigma^2 + \lambda^2} & \frac{\lambda^2}{\sigma^2 + \lambda^2} \\ \frac{\lambda^2}{\sigma^2 + \lambda^2} & \frac{\lambda^2}{\sigma^2 + \lambda^2} \end{bmatrix} I_p \right) \quad \text{[Exercise]}
\]

The usual Bayes estimator (w.r.t. quadratic loss) is the posterior mean
\[
(6) \quad \hat{M} = E[M \mid Y] = \left( \frac{\lambda^2}{\sigma^2 + \lambda^2} \right) Y.
\]

This is a linear estimator of the form \( ax \) with \( a = \frac{\lambda^2}{\sigma^2 + \lambda^2} < 1 \), so
\[ E[\| M - \hat{M} \|_2^2] = E[ax]^2 = a^2 E[\| Y - M \|_2^2 + (a-1)^2 \| M \|_2^2] \to \infty \text{ as } \| M \|_2 \to \infty, \text{ hence } \hat{M} \text{ cannot dominate the MLE } \hat{M} = Y. \text{ [The MSE of } \hat{M} \text{ is less than that of } \hat{M} \]

For \( M \) near 0, i.e., for \( M \) near the mean of the assumed prior distr (3).

If the prior variance \( \lambda^2 \) is unknown, however, then we cannot use \( \hat{M} \). This raises the question: can we estimate \( \lambda^2 \), based on \( Y \)?
If so, then we can replace \( \hat{M} \) by \( \hat{M} \), where \( \hat{\lambda} \) is the estimated value. The estimator \( \hat{M} \) is called an "empirical Bayesian estimator"
because the prior distribution is estimated from the data.

To obtain $\hat{\lambda}$, note from (4) that the marginal dist. of $Y$ is

$$Y \sim N_p \left[ 0, (\sigma^2 + \lambda^2) I_p \right],$$

so

$$\|Y\|^2 \sim (\sigma^2 + \lambda^2)^2 \chi_p^2. \tag{8}$$

Now, we can rewrite $M_\lambda$ as

$$M_\lambda = \left( 1 - \frac{\sigma^2}{\sigma^2 + \lambda^2} \right) Y, \tag{9}$$

so it will suffice to estimate $\frac{1}{\sigma^2 + \lambda^2}$. But

$$\frac{1}{\|Y\|^2} \sim \left( \frac{1}{\sigma^2 + \lambda^2} \right) \frac{1}{\chi_p^2}, \tag{10}$$

so

$$E \left( \frac{1}{\|Y\|^2} \right) = \left( \frac{1}{\sigma^2 + \lambda^2} \right) E \left( \frac{1}{\chi_p^2} \right) = \left( \frac{1}{\sigma^2 + \lambda^2} \right) \left( \frac{1}{\chi_p^2} \right) \tag{11}$$

provided $p \geq 3$.

Thus, when $p \geq 3$, we may estimate $\left( \frac{1}{\sigma^2 + \lambda^2} \right)$ by $\frac{p-2}{\|Y\|^2}$ so if we replace $\frac{1}{\sigma^2 + \lambda^2}$ by this estimate in (8), we obtain the James-Stein estimator

$$\frac{1}{\|Y\|^2} \sim \left[ 1 - \frac{(p-2)\sigma^2}{\|Y\|^2} \right] Y. \tag{12}$$

[When $p=1$ or 2, $E \left( \frac{1}{\chi_p^2} \right) = \infty$ so this motivation of (12) fails. The estimator on the right of (12) is still well defined, but it will not dominate.]

**THEOREM**: $E_{\hat{\lambda}} \|\hat{M}_\lambda - M\|^2 \leq E_{\hat{\lambda}} \|\hat{M}_\lambda - M\|^2 \equiv p\sigma^2$ for every $M \in \mathbb{R}^p$ ($p \geq 3$).

**Proof**: We will prove a more general result. Define

$$\hat{M}_c = \left[ 1 - \frac{c\sigma^2}{\|Y\|^2} \right] Y, \tag{13}$$

[c a constant].

We shall show that for every $0 < c < 2(p-2)$,

$$E_{\hat{\lambda}} \|\hat{M}_c - M\|^2 \leq p\sigma^2 \text{ for every } M \in \mathbb{R}^p, \tag{14}$$

and that the left side of (14) is minimized when $c = p-2$, i.e., when $\hat{M}_c = \hat{M}_JS$. 


Since \( \| \mathbf{z} \|^2 = \frac{\mathbf{z} \cdot \mathbf{z}}{\| \mathbf{z} \|^2} \), we have

\[
E \| \mathbf{z} \|^2 - \| \mathbf{z} \|^2 = E \| \mathbf{z} \|^2 - \frac{\sigma_{\mathbf{z}}^2 \| \mathbf{z} \|^2}{\| \mathbf{z} \|^2} = \frac{E \| \mathbf{z} \|^2 - 2\sigma_{\mathbf{z}}^2}{\| \mathbf{z} \|^2} + \sigma_{\mathbf{z}}^2 E \left[ \frac{1}{\| \mathbf{z} \|^2} \right].
\]

(15)

Now \( \mathbf{z} \sim N_p (\mathbf{0}, \sigma^2 I_p) \), so

\[
E \left[ \frac{\| \mathbf{z} \|^2}{\| \mathbf{z} \|^2} \right] = \frac{1}{\sigma^2} \sum_{i=1}^p E \left[ \frac{\| z_i \|^2}{\| \mathbf{z} \|^2} \right] = \frac{1}{\sigma^2} \sum_{i=1}^p \int_{\mathbb{R}^p} (z_i - \mu_i)^2 \frac{1}{\| \mathbf{z} \|^2} \text{d} \mathbf{z},
\]

where

\[
\tilde{g}(\mathbf{y} - \mathbf{m}) = \frac{1}{\sqrt{2\pi} \sigma} e^{-\frac{1}{2} (\mathbf{y} - \mathbf{m})^T \mathbf{y}}.
\]

(17)

Therefore, from (15),

\[
\Delta(\mathbf{M}) = \frac{1}{\| M \|^2} E \| \mathbf{z} \|^2 - \| \mathbf{z} \|^2 = \sigma^2 \left[ 2(p-1) - c \right] E \left[ \frac{1}{\| \mathbf{z} \|^2} \right].
\]

(20)

So \( \Delta(\mathbf{M}) > 0 \) when \( 0 < c < 2(p-1) \), and \( \Delta(\mathbf{M}) \) is maximized for all \( \mu \) when \( c = (p-1) \). This completes the proof of the theorem.
Remarks. A. When $C = p - 2$ so $\hat{\mu}_c = \hat{\mu}_{JS}$, the reduction in total MSE provided by $\hat{\mu}_{JS}$ is

$$\Delta(\mu) = (p - 2)^2 \sigma^2 E \left[ \frac{1}{M Y / \mu \mid} \right] = (p - 2)^2 \sigma^2 E \left[ \frac{1}{N \mu (S)} \right],$$

where $S = \|\hat{M}\|^2 / \sigma^2$. Since the noncentral $\chi^2_p (S)$-distribution can be shown to have monotone likelihood ratio in $S$, the expectation $E [1 / \chi^2_p (S)]$ is decreasing in $S$, so $\Delta(\mu)$ is decreasing in $\|\mu\|$. Thus the maximum reduction in total MSE provided by $\hat{\mu}_{JS}$ occurs when $\mu = 0$ and is given by

$$\Delta(0) = (p - 2)^2 \sigma^2 E \left[ \frac{1}{\chi^2_p \mu} \right] = (p - 2) \sigma^2,$$

which is quite large for large $p$. On the other hand, as $\|\mu\| \to \infty$, $\chi^2_p (S) \to \infty$ and $E [1 / \chi^2_p (S)] \to 0$ [this requires verification], so

$$\Delta(\mu) \to 0 \quad \text{as} \quad \|\mu\| \to \infty.$$

Thus, the total MSE of $\hat{\mu}_{JS}$, i.e., $E \|\hat{\mu}_{JS} - \mu\|^2$, depends on $\|\mu\|$ as follows:

\[ \begin{array}{c|c}
\|\mu\| & E \|\hat{\mu}_{JS} - \mu\|^2 \\
\hline
2 \sigma^2 & \text{Fig. 1} \\
0 & \text{Fig. 2}
\end{array} \]

Hence, the JS estimator $\hat{\mu}_{JS}$ "dominates" the LSE $\hat{\mu}$: its total MSE is always smaller.

On the other hand, the MSE of the Bayes estimator $\hat{\mu}_{\lambda}$ in (6) (shown in Figure 2) exceeds that of $\hat{\mu}$ for sufficiently large $\|\mu\|$: 

$$E \|\hat{\mu}_{\lambda} - \mu\|^2 = E \|a \hat{Y} - \mu\|^2 = E \|a(Y - \mu) + (a - 1) \mu\|^2 = a^2 E \|Y - \mu\|^2 + (a - 1)^2 \|\mu\|^2 \to \infty,$$

where $a = \frac{\lambda \mu}{\sigma^2 + \lambda < 1}$, so $E \|\hat{\mu}_{\lambda} - \mu\|^2 \to \infty$ as $\|\mu\| \to \infty$. 

\[ \begin{array}{c|c}
\|\mu\| & E \|a \hat{Y} - \mu\|^2 \\
\hline
0 & \text{Fig. 2}
\end{array} \]
B. The JS estimator $\hat{\theta}_{JS}$ does not dominate the MLE $\hat{\theta}$ in the strong sense

$$E_{\theta}^{\left(\hat{\theta}_{JS} - \theta\right)^2} - E_{\theta}^{\left(\hat{\theta}_{MLE} - \theta\right)^2} \equiv \Sigma_{MLE} - \Sigma_{JS}$$

is not positive semidefinite. The theorem on p. 14 states that $\Sigma_{MLE} > \Sigma_{JS}$, but it can be shown that some diagonal elements of $\Sigma_{JS}$ (i.e., $E_{\theta}^{\left(\hat{\theta}_{JS} - \theta\right)^2}$) can exceed the corresponding diagonal elements of $\Sigma_{MLE}$ (i.e., $E_{\theta}^{\left(\hat{\theta} - \theta\right)^2}$).

C. The JS estimate has the form $a(\|X\|)^{-1}$, where $a(\|X\|) = \left[1 - \frac{\|X\|^2}{\|X\|^2} \right]$. Since $a(\|X\|) < 1$, $\hat{\theta}_{JS}$ "shanks" $\theta$ toward $0$ (mirroring the behavior of the Bayes estimator $\hat{\theta}_B$, which shanks $\theta$ toward $\theta_0$ the prior mean). However, it can occur that $a(\|X\|) < 0$, (whenever $\|X\|^2 < (p-2)\sigma^2$), in which case $\hat{\theta}_{JS}$ "shanks" $\theta$ too far. It can be shown that the modified JS estimator $\hat{\theta}_{JS}^* = a^*(\|X\|)\theta$ dominates $\hat{\theta}_{JS}$ (hence $\hat{\theta}_{JS}^*$ dominates $\hat{\theta}$), where $a^*(\|X\|) = \max a(\|X\|), 0$.

D. If $\sigma^2$ is unknown but we have an estimator $\hat{\sigma}^2$ of $\sigma^2$ such that $\hat{\sigma}^2$ is independent of $\theta$ and $\hat{\sigma}^2 \sim \sigma^2/\kappa$ for some $\kappa$, then estimators of $\mu$ of the form $[1 - \frac{\hat{\sigma}^2}{\|X\|^2}] \theta$ will dominate $\hat{\mu}$ for suitable choices of $\kappa$.

E. As already mentioned, if we replace the prior (3) by $N\left(M_0, \sigma^2 I_p\right)$, we are led to the "JS" estimator that shanks $\theta$ toward $M_0$, namely

$$\hat{\theta}_{JS}(M_0) \equiv \left[1 - \frac{(p-2)\sigma^2}{\|X\|^2}\right] \left(\frac{\|X\|^2}{\Sigma M_0} + M_0\right).$$

The choice of $M_0$ should not be made arbitrarily, but should be based on prior information available. If no such info. is available, the JS estimator should not be used (this is my personal opinion).

F. It may be preferable to shrink toward a linear subspace, e.g., toward $(\bar{y}, \ldots, \bar{y})$. Use p-3.