1. Maximum Likelihood Method: consider $n$ random samples from a multivariate normal distribution, $X_i \in \mathbb{R}^p \sim \mathcal{N}(\mu, \Sigma)$ with $i = 1, \ldots, n$.

   (a) Show the log-likelihood function
   
   $$ l_n(\mu, \Sigma) = -\frac{n}{2} \text{trace}(\Sigma^{-1}S_n) - \frac{n}{2} \log \det(\Sigma) + C, $$

   where $S_n = \frac{1}{n} \sum_{i=1}^{n} (X_i - \mu)(X_i - \mu)^T$, and some constant $C$ does not depend on $\mu$ and $\Sigma$;

   (b) Show that $f(X) = \text{trace}(AX^{-1})$ with $A, X \succeq 0$ has a first-order approximation,
   
   $$ f(X + \Delta) \approx f(X) - \text{trace}(X^{-1}A'X^{-1}\Delta) $$

   hence formally $df(X)/dX = -X^{-1}AX^{-1}$ (note $(I + X)^{-1} \approx I - X$);

   (c) Show that $g(X) = \log \det(X)$ with $A, X \succeq 0$ has a first-order approximation,
   
   $$ g(X + \Delta) \approx g(X) + \text{trace}(X^{-1}\Delta) $$

   hence $dg(X)/dX = X^{-1}$ (note: consider eigenvalues of $X^{-1/2}\Delta X^{-1/2}$);

   (d) Use these formal derivatives with respect to positive semi-definite matrix variables to show that the maximum likelihood estimator of $\Sigma$ is
   
   $$ \hat{\Sigma}_n^{MLE} = S_n. $$

A reference for (b) and (c) can be found in Convex Optimization, by Boyd and Vandenberghe, examples in Appendix A.4.1 and A.4.3:


2. Shrinkage: Suppose $y \sim \mathcal{N}(\mu, I_p)$. 

---

A reference for (b) and (c) can be found in Convex Optimization, by Boyd and Vandenberghe, examples in Appendix A.4.1 and A.4.3:

(a) Consider the Ridge regression
\[
\min_{\mu} \frac{1}{2} \| y - \mu \|^2 + \frac{\lambda}{2} \| \mu \|^2.
\]
Show that the solution is given by
\[
\hat{\mu}^{\text{ridge}} = \frac{1}{1 + \lambda y_i}.
\]
Compute the risk (mean square error) of this estimator. The risk of MLE is given when \( C = I \).

(b) Consider the LASSO problem,
\[
\min_{\mu} \frac{1}{2} \| y - \mu \|^2 + \lambda \| \mu \|_1.
\]
Show that the solution is given by Soft-Thresholding
\[
\hat{\mu}^{\text{soft}}_i = \mu^{\text{soft}}(y_i; \lambda) := \text{sign}(y_i)(|y_i| - \lambda)_+.
\]
For the choice \( \lambda = \sqrt{2 \log p} \), show that the risk is bounded by
\[
\mathbb{E} \| \hat{\mu}^{\text{soft}}(y) - \mu \|^2 \leq 1 + (2 \log p + 1) \sum_{i=1}^p \min(\mu_i^2, 1).
\]
Under what conditions on \( \mu \), such a risk is smaller than that of MLE? Note: see Gaussian Estimation by Iain Johnstone, Lemma 2.9 and the reasoning before it.

(c) Consider the \( l_0 \) regularization
\[
\min_{\mu} \| y - \mu \|^2 + \lambda^2 \| \mu \|_0,
\]
where \( \| \mu \|_0 := \sum_{i=1}^p I(\mu_i \neq 0) \). Show that the solution is given by Hard-Thresholding
\[
\hat{\mu}^{\text{hard}}_i = \mu^{\text{hard}}(y_i; \lambda) := y_i I(|y_i| > \lambda).
\]
Rewriting \( \hat{\mu}^{\text{hard}}(y) = (1 - g(y))y \), is \( g(y) \) weakly differentiable? Why?

(d) Consider the James-Stein Estimator
\[
\hat{\mu}^{\text{JS}}(y) = \left( 1 - \frac{\alpha}{\| y \|^2} \right) y.
\]
Show that the risk is
\[
\mathbb{E} \| \hat{\mu}^{\text{JS}}(y) - \mu \|^2 = \mathbb{E} U_{\alpha}(y)
\]
where \( U_{\alpha}(y) = p - (2\alpha(p - 1) - \alpha^2)/\| y \|^2 \). Find the optimal \( \alpha^* = \arg \min_{\alpha} U_{\alpha}(y) \). Show that for \( p > 2 \), the risk of James-Stein Estimator is smaller than that of MLE for all \( \mu \in \mathbb{R}^p \).
(e) In general, an odd monotone unbounded function $\Theta : \mathbb{R} \rightarrow \mathbb{R}$ defined by $\Theta_{\lambda}(t)$ with parameter $\lambda \geq 0$ is called shrinkage rule, if it satisfies

- [shrinkage] $0 \leq \Theta_{\lambda}(|t|) \leq |t|$;
- [odd] $\Theta_{\lambda}(-t) = -\Theta_{\lambda}(t)$;
- [monotone] $\Theta_{\lambda}(t) \leq \Theta_{\lambda}(t')$ for $t \leq t'$;
- [unbounded] $\lim_{t \to \infty} \Theta_{\lambda}(t) = \infty$.

Which rules above are shrinkage rules?

3. Necessary Condition for Admissibility of Linear Estimators. Consider linear estimator for $y \sim \mathcal{N}(\mu, \sigma^2 I_p)$

$$\hat{\mu}_C(y) = Cy.$$ 

Show that $\hat{\mu}_C$ is admissible only if

(a) $C$ is symmetric;
(b) $0 \leq \rho_i(C) \leq 1$ (where $\rho_i(C)$ are eigenvalues of $C$);
(c) $\rho_i(C) = 1$ for at most two $i$.

These conditions are satisfied for MLE estimator when $p = 1$ and $p = 2$. 


4. *James Stein Estimator for $p = 1, 2$ and upper bound:

If we use SURE to calculate the risk of James Stein Estimator,

$$R(\hat{\mu}^{JS}, \mu) = \mathbb{E}U(Y) = p - \mathbb{E}_\mu \left(\frac{(p - 2)^2}{\|Y\|^2}\right) < p = R(\hat{\mu}^{MLE}, \mu)$$

it seems that for $p = 1$ James Stein Estimator should still have lower risk than MLE for any $\mu$. Can you find what will happen for $p = 1$ and $p = 2$ cases?

Moreover, can you derive the upper bound for the risk of James-Stein Estimator?

$$R(\hat{\mu}^{JS}, \mu) \leq p - \frac{(p - 2)^2}{p - 2 + \|\mu\|^2} = 2 + \frac{(p - 2)\|\mu\|^2}{p - 2 + \|\mu\|^2}.$$