# 1: Properties of exponential families

(a) Want to show that if $\theta_1$ and $\theta_2$ are in $\Theta_N$, then for any $\lambda \in [0,1]$, $\lambda \theta_1 + (1 - \lambda) \theta_2 \in \Theta_N$.

Observe that

$$
\int \exp\{\langle \lambda \theta_1 + (1 - \lambda) \theta_2, T(x) \rangle \} h(x) dx \\
= \int \exp\{\langle \lambda \theta_1 + (1 - \lambda) \theta_2, T(x) \rangle \} h(x)^\lambda h(x)^{1-\lambda} dx \\
= \int \exp\{\langle \theta_1, T(x) \rangle \} h(x)^\lambda \exp\{\langle \theta_2, T(x) \rangle \} h(x))^{1-\lambda} dx \\
\leq \left( \int \exp\{\langle \theta_1, T(x) \rangle \} h(x) dx \right)^\lambda \left( \int \exp\{\langle \theta_2, T(x) \rangle \} h(x) dx \right)^{1-\lambda} \quad \text{[By Holder’s inequality]} \\
< \infty
$$

That $\phi$ is a convex function follows from exactly the same line of arguments as above.

(b)

$$
\frac{\partial \phi}{\partial \theta_i}(\theta) = \frac{1}{Z(\theta)} \int T_i(x) \exp\{\langle \theta, T(x) \rangle \} h(x) dx = E_\theta\{T_i(X)\}
$$

Now,

$$
Z(\theta) \frac{\partial \phi}{\partial \theta_i}(\theta) = \int T_i(x) \exp\{\langle \theta, T(x) \rangle \} h(x) dx = E_\theta\{T_i(X)\}
$$

Differentiating the above equation w.r.t. $\theta_j$ and then dividing both sides by $Z(\theta)$, we get

$$
\frac{\partial \phi}{\partial \theta_i}(\theta) \frac{\partial \phi}{\partial \theta_j}(\theta) + \frac{\partial^2 \phi}{\partial \theta_i \partial \theta_j}(\theta) = \frac{1}{Z(\theta)} \int T_i(x) T_j(x) \exp\{\langle \theta, T(x) \rangle \} h(x) dx
$$

Hence proved.

(c)

$$
E_\theta \left\{ \frac{1}{h(x)} \frac{\partial h}{\partial x_i}(x) + \langle \theta, \frac{\partial T}{\partial x_i}(x) \rangle \right\} g(x)
$$

$$
= \int \frac{\partial P_\theta(x)}{\partial x_i} g(x) dx
$$

$$
= - \int P_\theta(x) \frac{\partial g}{\partial x_i}(x) dx \quad \text{[By Integration by parts]} \\
= -E_\theta \left\{ \frac{\partial g}{\partial x_i}(x) \right\}
$$

Hence proved.
(d) If \( X \sim N(\mu, I) \), then trivially we have
\[
\mathbb{E}\{(x - \mu)g(x)\} = \mathbb{E}\{\nabla g(x)\}.
\]
where we applied Stein’s identity component wise.

Now, let \( X \sim N(\mu, \Sigma) \), and whiten \( X \).

Take \( Y = AX \) where \( A^{-1} \) is the square root of \( \Sigma \). Let \( h(x) = g(A^{-1}x) \).

Using Stein’s identity for \( Y \) and the differentiable function \( h \), we get
\[
\mathbb{E}\{(y - A\mu) h(x)\} = \mathbb{E}\{\nabla h(x)\}.
\]
Also,
\[
\nabla h(x) = A^{-1} \nabla g(x)
\]
Plugging in \( Y = AX \), we get,
\[
\mathbb{E}\{(Ax - A\mu)g(x)\} = A^{-1} \mathbb{E}\{\nabla g(x)\}.
\]

Hence,
\[
\mathbb{E}\{(x - \mu)g(x)\} = \Sigma \mathbb{E}\{\nabla g(x)\}.
\]

# 2: Exercises on sufficient statistics

(a) We claim that \( T(x) = \{\frac{p_j}{p_1}, \frac{p_2}{p_1}, \ldots, \frac{p_k}{p_1}\} \) is a sufficient statistic. Observe that for any \( j \), such that \( 1 \leq j \leq k \), we have,
\[
p_j = \frac{p_j}{p_1}p_1
\]
Let \( g(T(x), j) = \frac{p_j}{p_1} \) and \( h(x) = p_1 \), then by Factorization theorem, we have that \( T(x) \) is a sufficient statistic.

(b) The joint density of \( X_1, \ldots, X_n \) is
\[
f(x_1, \ldots, x_n) = \frac{1}{\theta_2 - \theta_1} \prod_{i=1}^{k} \mathbf{1}_{\{\theta_1 \leq x_i \leq \theta_2\}}
\]
\[
= \frac{1}{\theta_2 - \theta_1} \mathbf{1}_{\{x_{\min} \geq \theta_1\}} \mathbf{1}_{\{x_{\max} \leq \theta_2\}}
\]
By Factorization theorem, we have the result.

(c)
\[
p_\theta = \frac{1}{(\sigma \sqrt{2\pi})^n} e^{\exp\{-(x - A\theta)^T(x - A\theta)\}}
\]
\[
= \frac{1}{(\sigma \sqrt{2\pi})^n} e^{\exp\{-\frac{1}{2\sigma^2}(x^T x + \theta^T A^T A \theta - 2x^T A \theta)\}}
\]
Hence, again by Factorization theorem, \( A^T x \) is a sufficient statistic.
# 3: Optimal linear estimation in heteroscedastic Gaussian model

Assume $\sigma_1, \ldots, \sigma_d > 0$ to be known, and consider the statistical model $P_\theta = N(\theta \mathbf{1}, \Sigma)$, where $\Sigma = \text{diag}(\sigma_1^2, \ldots, \sigma_d^2)$, and $\theta \in \Theta = \mathbb{R}$ (with $\mathbf{1}$ denoting the all-ones vector). In other words, $X_i = \theta + \sigma_i G_i$ where $(G_i)_{i \leq d} \sim \text{iid } N(0, 1)$. (Here $\langle \mathbf{u}, \mathbf{v} \rangle = \sum_{i=1}^m u_i v_i$ denotes the usual scalar product of $\mathbf{u}, \mathbf{v} \in \mathbb{R}^m$.)

(a) 

\[
p_\theta = \frac{1}{(\prod_{i=1}^d \sigma_i \sqrt{2\pi})^n} \exp\left\{ -\sum_{i=1}^d \left( \frac{x_i - \theta}{\sigma_i} \right)^2 \right\} \\
= \frac{1}{(\prod_{i=1}^d \sigma_i \sqrt{2\pi})^n} \exp\left\{ \sum_{i=1}^d \left( \frac{x_i^2}{\sigma_i^2} + \frac{\theta^2}{\sigma_i^2} - \frac{2\theta x_i}{\sigma_i^2} \right) \right\} \\
= \frac{1}{(\prod_{i=1}^d \sigma_i \sqrt{2\pi})^n} \exp\left\{ \sum_{i=1}^d \frac{x_i^2}{\sigma_i^2} \right\} \exp\left\{ -2\theta \sum_{i=1}^d \frac{x_i}{\sigma_i^2} \right\}
\]

So, by Factorization theorem, $l(x) = \langle c, x \rangle$ is a sufficient statistic where $c = \left( \frac{1}{\sigma_1^2}, \ldots, \frac{1}{\sigma_d^2} \right)$

(b) We use Rao-Blackwell theorem to do this problem.

First observe that $\langle \mathbf{a}, \mathbf{1} \rangle = 1$, as

\[
\mathbb{E}(\langle \mathbf{a}, x \rangle - \theta)^2 = \text{Var}(\langle \mathbf{a}, x \rangle) + (\text{Bias}(\langle \mathbf{a}, x \rangle))^2 \\
= a^T \Sigma a + \theta^2 (\langle \mathbf{a}, \mathbf{1} \rangle - 1)^2
\]

So, if $\langle \mathbf{a}, \mathbf{1} \rangle \neq 1$, then we can not avoid having the corresponding minimum value as $\infty$.

Now, for any $\mathbf{a}$ such that $\langle \mathbf{a}, \mathbf{1} \rangle = 1$, consider the Rao-Blackwell estimator $\mathbb{E}(\langle \mathbf{a}, x \rangle | \langle c, x \rangle)$ corresponding to the estimator $\langle \mathbf{a}, x \rangle$, where we have used the sufficiency of $\langle c, x \rangle$.

By Rao-Blackwell theorem, the corresponding Rao-Blackwell estimator has lower risk.

So, it suffices to consider only the Rao-Blackwell estimators , i.e., we only to optimize among all Rao-Blackwell estimators such that $\langle \mathbf{a}, \mathbf{1} \rangle = 1$.

Let us find the explicit form of $\mathbb{E}(\langle \mathbf{a}, x \rangle | \langle c, x \rangle)$.

Note,

\[
\left( \begin{array}{c} \langle \mathbf{a}, x \rangle \\ \langle c, x \rangle \end{array} \right) \sim N \left[ \left( \begin{array}{c} \theta (\langle \mathbf{a}, \mathbf{1} \rangle) \\ \theta (\langle c, \mathbf{1} \rangle) \end{array} \right), \left( \begin{array}{cc} a^T \Sigma a & a^T \Sigma c \\ a^T \Sigma c & c^T \Sigma c \end{array} \right) \right]
\]

Also note,

\[
a^T \Sigma c = \langle \mathbf{a}, \mathbf{1} \rangle = 1 \\
c^T \Sigma c = \langle c, \mathbf{1} \rangle
\]

Using the expression for conditional mean for bivariate Normal, we have,

\[
\mathbb{E}(\langle \mathbf{a}, x \rangle | \langle c, x \rangle) = \frac{\langle c, x \rangle}{\langle c, \mathbf{1} \rangle}
\]

which is free of $\mathbf{a}$, so it is the required optimal linear estimator.

[Remark: You can also go about the problem by first invoking Rao-Blackwell Theorem and then observing that $\langle \mathbf{a}, \mathbf{1} \rangle = 1$ or simply use Lagrangian Method for optimization].
(c) Let $X \sim N(\theta 1, \Sigma)$, take $Y = AX$ where $A^{-1}$ is the square root of $\Sigma$, so $Y \sim N(\theta A 1, I)$.

Let $a_i$ be the $i$th row sum of the matrix $A$.

It can be easily verified that the sufficient statistic is again of the form $l(y) = \langle c, y \rangle$, where $c = (a_1, \ldots, a_d)$.

Again following the same argument as in part b), we have the optimal linear estimator is

$$\frac{\langle c, y \rangle}{\langle c, A 1 \rangle}.$$