Stat 300A Theory of Statistics

Homework 1 solution

Suyash Gupta *Due on October 3, 2018*

1: Properties of exponential families

(a) Want to show that if θ_1 and θ_2 are in Θ_N, then for any $\lambda \in [0,1]$, $\lambda\theta_1 + (1-\lambda)\theta_2 \in \Theta_N$.

Observe that

$$\begin{aligned}
&\int exp\{\langle \lambda\boldsymbol{\theta}_1 + (1-\lambda)\boldsymbol{\theta}_2, \boldsymbol{T}(x)\rangle\} h(x)\mathrm{d}\boldsymbol{x} \\
&= \int exp\{\langle \lambda\boldsymbol{\theta}_1 + (1-\lambda)\boldsymbol{\theta}_2, \boldsymbol{T}(x)\rangle\} h(x)^{\lambda} h(x)^{1-\lambda}\mathrm{d}\boldsymbol{x} \\
&= \int (exp\{\langle \boldsymbol{\theta}_1, \boldsymbol{T}(x)\rangle\} h(x))^{\lambda} (exp\{\langle \boldsymbol{\theta}_2, \boldsymbol{T}(x)\rangle\} h(x))^{1-\lambda}\mathrm{d}\boldsymbol{x} \\
&\le \left(\int (exp\{\langle \boldsymbol{\theta}_1, \boldsymbol{T}(x)\rangle\} h(x))\mathrm{d}\boldsymbol{x}\right)^{\lambda} \left(\int (exp\{\langle \boldsymbol{\theta}_2, \boldsymbol{T}(x)\rangle\} h(x))\mathrm{d}\boldsymbol{x}\right)^{1-\lambda} \quad [By\ Holder's\ inequality] \\
&< \infty
\end{aligned}$$

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

(b)

$$\frac{\partial \phi}{\partial \boldsymbol{\theta}_i}(\theta) = \frac{1}{Z(\theta)} \int T_i(x) exp\{\langle \boldsymbol{\theta}, \boldsymbol{T}(x)\rangle\} h(x))\mathrm{d}\boldsymbol{x} = \mathsf{E}_{\boldsymbol{\theta}}\{T_i(\boldsymbol{X})\}$$

Now,

$$Z(\boldsymbol{\theta})\frac{\partial \phi}{\partial \boldsymbol{\theta}_i}(\theta) = \int T_i(x) exp\{\langle \boldsymbol{\theta}, \boldsymbol{T}(x)\rangle\} h(x))\mathrm{d}\boldsymbol{x} = \mathsf{E}_{\boldsymbol{\theta}}\{T_i(\boldsymbol{X})\}$$

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

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

Hence proved.

(c)

$$\begin{aligned}
&\mathsf{E}_{\boldsymbol{\theta}}\left\{\left[\frac{1}{h(\boldsymbol{x})}\frac{\partial h}{\partial x_i}(\boldsymbol{x}) + \langle \boldsymbol{\theta}, \frac{\partial \boldsymbol{T}}{\partial x_i}(\boldsymbol{x})\rangle\right] g(\boldsymbol{x})\right\} \\
&= \int \frac{\partial \mathsf{P}_{\boldsymbol{\theta}}(\boldsymbol{x})}{\partial x_i} g(x)\mathrm{d}\boldsymbol{x} \\
&= -\int \mathsf{P}_{\boldsymbol{\theta}}(\boldsymbol{x})\frac{\partial g}{\partial x_i}(\boldsymbol{x})\mathrm{d}\boldsymbol{x} \ \ [By\ Integration\ by\ parts] \\
&= -\mathsf{E}_{\boldsymbol{\theta}}\left\{\frac{\partial g}{\partial x_i}(\boldsymbol{x})\right\}
\end{aligned}$$

Hence proved.

(*d*) If $X \sim N(\mu, I)$, then trivially we have

$$\mathbb{E}\big\{(\boldsymbol{x} - \boldsymbol{\mu})\, g(\boldsymbol{x})\big\} = \mathbb{E}\big\{\nabla g(\boldsymbol{x})\big\}\,.$$

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 Σ. Let $h(\boldsymbol{x}) = g(A^{-1}\boldsymbol{x})$.

Using Stein's identity for Y and the differentiable function h., we get

$$\mathbb{E}\big\{(\boldsymbol{y} - A\boldsymbol{\mu})\, h(\boldsymbol{x})\big\} = \mathbb{E}\big\{\nabla h(\boldsymbol{x})\big\}\,.$$
$$Also,\ \ \nabla h(\boldsymbol{x}) = A^{-1}\nabla g(\boldsymbol{x})$$

Plugging in $Y = AX$, we get,

$$\mathbb{E}\big\{(A\boldsymbol{x} - A\boldsymbol{\mu})\, g(\boldsymbol{x})\big\} = A^{-1}\mathbb{E}\big\{\nabla g(\boldsymbol{x})\big\}\,.$$

Hence,

$$\mathbb{E}\big\{(\boldsymbol{x} - \boldsymbol{\mu})\, g(\boldsymbol{x})\big\} = \boldsymbol{\Sigma}\mathbb{E}\big\{\nabla g(\boldsymbol{x})\big\}\,.$$

2: Exercises on sufficient statistics

(*a*) We claim that $T(x)=\{\frac{\mathsf{p}_2}{\mathsf{p}_1}, \frac{\mathsf{p}_3}{\mathsf{p}_1}, ..., \frac{\mathsf{p}_k}{\mathsf{p}_1}\}$ is a sufficient statistic. Observe that for any j, such that $1 \leq j \leq k$, we have,

$$\mathsf{p}_j = \frac{\mathsf{p}_j}{\mathsf{p}_1}\mathsf{p}_1$$

Let $g(T(x), j) = \frac{\mathsf{p}_j}{\mathsf{p}_1}$ and $h(x) = \mathsf{p}_1$, then by Factorization theorem, we have that $T(x)$ is a sufficient statistic.

(*b*) The joint density of $X_1, ..., X_n$ is

$$\begin{aligned} f(x_1, ..., x_n) =& \frac{1}{\theta_2 - \theta_1}\prod_{i=1}^{k}\mathbf{1}_{\{\theta_1 \leq x_i \leq \theta_2\}} \\ =& \frac{1}{\theta_2 - \theta_1}\mathbf{1}_{\{x_{min} \geq \theta_1\}}\mathbf{1}_{\{x_{max} \leq \theta_2\}} \end{aligned}$$

By Factorization theorem, we have the result.

(*c*)

$$\begin{aligned} p_\theta =& \frac{1}{(\sigma\sqrt{2\pi})^n}exp\{-\frac{(x - A\boldsymbol{\theta})^T(x - A\boldsymbol{\theta})}{2\sigma^2}\} \\ =& \frac{1}{(\sigma\sqrt{2\pi})^n}exp\{-\frac{1}{2\sigma^2}(x^Tx + \theta^TA^TA\theta - 2x^TA\theta)\} \end{aligned}$$

Hence, again by Factorization theorem, A^Tx is a sufficient statistic.

3: Optimal linear estimation in heteroscedastic Gaussian model

Assume $\sigma_1, \dots, \sigma_d > 0$ to be known, and consider the statistical model $\mathsf{P}_\theta = \mathsf{N}(\theta\mathbf{1}, \mathbf{\Sigma})$, where $\mathbf{\Sigma} = \text{diag}(\sigma_1^2, \dots, \sigma_d^2)$, 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 $(G_i)_{i \le d} \sim_{iid} \mathsf{N}(0,1)$. (Here $\langle \boldsymbol{u}, \boldsymbol{v} \rangle = \sum_{i=1}^m u_i v_i$ denotes the usual scalar product of $\boldsymbol{u}, \boldsymbol{v} \in \mathbb{R}^m$.)

(a)

$$
\begin{aligned}
p_\theta =& \frac{1}{(\prod_{i=1}^d \sigma_i \sqrt{2\pi})^n} exp\{-\sum_{i=1}^d \left(\frac{x_i - \theta}{\sigma_i}\right)^2\} \\
=& \frac{1}{(\prod_{i=1}^d \sigma_i \sqrt{2\pi})^n} exp\{\sum_{i=1}^d \left(\frac{x_i^2}{\sigma_i^2} + \frac{\theta^2}{\sigma_i^2} - \frac{2\theta x_i}{\sigma_i^2}\right)\} \\
=& \frac{1}{(\prod_{i=1}^d \sigma_i \sqrt{2\pi})^n} exp\{\sum_{i=1}^d \frac{x_i^2}{\sigma_i^2}\} exp\{\sum_{i=1}^d \frac{\theta^2}{\sigma_i^2}\} exp\{-2\theta \sum_{i=1}^d \frac{x_i}{\sigma_i^2}\}
\end{aligned}
$$

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

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

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

$$
\begin{aligned}
\mathbb{E}(\langle \boldsymbol{a}, \boldsymbol{x} \rangle - \theta)^2 =& Var(\langle \boldsymbol{a}, \boldsymbol{x} \rangle) + (Bias(\langle \boldsymbol{a}, \boldsymbol{x} \rangle))^2 \\
=& a^T \Sigma a + \theta^2 (\langle \boldsymbol{a}, \mathbf{1} \rangle - 1)^2
\end{aligned}
$$

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

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

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 $\langle \boldsymbol{a}, \mathbf{1} \rangle = 1$.

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

Note,

$$
\begin{pmatrix} \langle \boldsymbol{a}, \boldsymbol{x} \rangle \\ \langle \boldsymbol{c}, \boldsymbol{x} \rangle \end{pmatrix} \sim N\left[\begin{pmatrix} \theta \langle \boldsymbol{a}, \mathbf{1} \rangle \\ \theta \langle \boldsymbol{c}, \mathbf{1} \rangle \end{pmatrix}, \begin{pmatrix} \boldsymbol{a}^T \Sigma \boldsymbol{a} & \boldsymbol{a}^T \Sigma \boldsymbol{c} \\ \boldsymbol{a}^T \Sigma \boldsymbol{c} & \boldsymbol{c}^T \Sigma \boldsymbol{c} \end{pmatrix}\right]
$$

Also note,

$$
\begin{aligned}
\boldsymbol{a}^T \Sigma \boldsymbol{c} &= \langle \boldsymbol{a}, \mathbf{1} \rangle = 1 \\
\boldsymbol{c}^T \Sigma \boldsymbol{c} &= \langle \boldsymbol{c}, \mathbf{1} \rangle
\end{aligned}
$$

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

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

which is free of $\boldsymbol{a}$, 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 $\langle \boldsymbol{a}, \mathbf{1} \rangle = 1$ or simply use Lagrangian Method for optimization].

(c) Let $X \sim N(\theta\mathbf{1}, \Sigma)$, take $Y = AX$ where A^{-1} is the square root of Σ, so $Y \sim N(\theta\boldsymbol{A}\mathbf{1}, I)$.

Let a_i be the ith row sum of the matrix A.

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

Again following the same argument as in part b), we have the optimal linear estimator is $\frac{\langle \boldsymbol{c}, \boldsymbol{y} \rangle}{\langle \boldsymbol{c}, A\mathbf{1} \rangle} = \frac{\langle \boldsymbol{c}, \boldsymbol{Ax} \rangle}{\langle \boldsymbol{c}, A\mathbf{1} \rangle}$.