CMSE 820 Homework 10. Due 8 December, 2020.

The following facts from basic probability theory can be used freely in this assignment.

Proposition 1. If $\xi := (\xi_1, \ldots, \xi_n)$ is a random vector with probability density function $p_\xi(x)$ and $g : \mathbb{R}^N \to \mathbb{R}$ is a continuous function, then $g(\xi)$ is a random variable and

$$\mathbb{E}(g(\xi)) = \int_{\mathbb{R}^n} g(x) p_\xi(x) dx.$$

Corollary 1: If $g \geq 0$ everywhere, then $\mathbb{E}(g(\xi)) \geq 0$ since we always have $p_\xi \geq 0$.

Corollary 2: If we take $g(x) := c_1 x_1 + \cdots + c_n x_n$ where $c_1, \ldots, c_n$ are constants, then

$$E(c_1\xi_1 + \cdots + c_n\xi_n) = \int_{\mathbb{R}^n} (c_1 x_1 + \cdots + c_n x_n) p_\xi(x) dx$$

$$= c_1 \int_{\mathbb{R}^n} x_1 p_\xi(x) dx + \cdots + c_n \int_{\mathbb{R}^n} x_n p_\xi(x) dx = c_1 \mathbb{E}(\xi_1) + \cdots + c_n \mathbb{E}(\xi_n)$$

where $\mathbb{E}(\xi_1)$ is the expectation of the marginal distribution of ξ_1. This shows taking expectations of random variables is a linear operation.

I. Let $\xi := (\xi_1, \ldots, \xi_n)^T$ be a random vector. Prove the following statements using the definition of expectation and covariance matrix. (Caution: corollary 2 only shows taking expectations of random variables is a linear operation; it does not show taking expectations of random vectors is a linear operation. This is what you need to prove.)

(a) If $c \in \mathbb{R}^n$ is a constant vector, prove that

$$\mathbb{E}(\xi + c) = \mathbb{E}(\xi) + c,$$

$$\mathrm{Cov}(\xi + c) = \mathrm{Cov}(\xi).$$

Solution. Let ξ be a random vector and $c = (c_1, \ldots, c_n)^T$ be a constant vector both in $\mathbb{R}^n$. Let $y = \xi + c = (\xi_1 + c_1, \ldots, \xi_n + c_n)^T$. It follows that since ξ is a random vector, y is also a random vector. To take the expectation of y, we take the expectation elementwise,

$$\mathbb{E}(y) = (\mathbb{E}(y_1), \ldots, \mathbb{E}(y_n))^T.$$

Let $\mathbb{E}(y_i)$ be the expectation value of the ith element of y. Since $y_i = \xi_i + c_i$ and we know the expectation for random variables is a linear operation, it follows that $\mathbb{E}(y_i) = \mathbb{E}(\xi_i) + \mathbb{E}(c_i)$. c_i is a constant, and thus $\mathbb{E}(c_i) = c_i$. Therefore, $\mathbb{E}(y_i) = \mathbb{E}(\xi_i) + c_i$. From this,

$$\mathbb{E}(y) = (\mathbb{E}(\xi_1) + c_1, \ldots, \mathbb{E}(\xi_n) + c_n)^T.$$

Separating this into vector addition,

$$\mathbb{E}(y) = (\mathbb{E}(\xi_1), \ldots, \mathbb{E}(\xi_n))^T + (c_1, \ldots, c_n)^T,$$

thus,

$$\mathbb{E}(y) = \mathbb{E}(\xi + c) = \mathbb{E}(\xi) + c. \blacksquare$$

Consider now $\text{Cov}(\xi + c) = \text{Cov}(y)$. From the definition of covariance,

$$\text{Cov}(y) = \mathbb{E}\left[(y - \mathbb{E}(y))(y - \mathbb{E}(y))^T\right].$$

Using the above proof that $\mathbb{E}(y) = \mathbb{E}(\xi) + c$ and substituting back $y = \xi + c$, we can rewrite this as

$$\text{Cov}(y) = \mathbb{E}\left[(\xi + c - (\mathbb{E}(\xi) + c))(\xi + c - (\mathbb{E}(\xi) + c))^T\right].$$

Simplifying, we see

$$\text{Cov}(y) = \mathbb{E}\left[(\xi - \mathbb{E}(\xi))(\xi - \mathbb{E}(\xi))^T\right] = \text{Cov}(\xi).$$

Hence, $\text{Cov}(\xi + c) = \text{Cov}(\xi)$. ■

(b) If $A \in \mathbb{R}^{m\times n}$ is a constant matrix, prove that

$$\mathbb{E}(A\xi) = A\mathbb{E}(\xi),$$

$$\text{Cov}(A\xi) = A\text{Cov}(\xi)A^T.$$

Solution. Let

$$A = \begin{pmatrix} a_1 \\ \vdots \\ a_m \end{pmatrix} \in \mathbb{R}^{m\times n}$$

be a matrix made of row vectors $a_1, \ldots, a_m$. Let $y = A\xi \in \mathbb{R}^m$. Since ξ is a random vector, y must also be a random vector. Let $y_i = a_i\xi$ be the ith element of y calculated as the ith row of A times the column vector ξ. We can write

$$y_i = a_{i1}\xi_1 + \cdots + a_{in}\xi_n,$$

where a_{ij} is the jth element of the ith row of A. Taking the expectation of y_i,

$$\mathbb{E}(y_i) = \mathbb{E}(a_{i1}\xi_1 + \cdots + a_{in}\xi_n).$$

Since the expectation of random variables is a linear operation, we can rewrite this as

$$\mathbb{E}(y_i) = \mathbb{E}(a_{i1}\xi_1) + \cdots + \mathbb{E}(a_{in}\xi_n).$$

a_{ij} is a constant value, and it thus follows that

$$\mathbb{E}(y_i) = a_{i1}\mathbb{E}(\xi_1) + \cdots + a_{in}\mathbb{E}(\xi_n) = a_i\mathbb{E}(\xi).$$

Therefore,

$$\mathbb{E}(y) = \begin{pmatrix} \mathbb{E}(y_1) \\ \vdots \\ \mathbb{E}(y_m) \end{pmatrix} = \begin{pmatrix} a_1\mathbb{E}(\xi) \\ \vdots \\ a_m\mathbb{E}(\xi) \end{pmatrix} = A\mathbb{E}(\xi). ■$$

Consider now $\text{Cov}(y)$. From the definition of covariance,

$$\text{Cov}(y) = \mathbb{E}\left[(y - \mathbb{E}(y))(y - \mathbb{E}(y))^T\right].$$

Using the above result that $\mathbb{E}(y) = A\mathbb{E}(\xi)$,

$$\text{Cov}(y) = \mathbb{E}\left[(y - A\mathbb{E}(\xi))(y - A\mathbb{E}(\xi))^T\right].$$

Multiplying the expression out,

$$\text{Cov}(y) = \mathbb{E}\left[yy^T - A\mathbb{E}(\xi)y^T - y(A\mathbb{E}(\xi))^T + A\mathbb{E}(\xi)(A\mathbb{E}(\xi))^T\right].$$

Substituting $y = A\xi$ back in,

$$\text{Cov}(y) = \mathbb{E}\left[A\xi\xi^T A^T - A\mathbb{E}(\xi)\xi^T A^T - A\xi(A\mathbb{E}(\xi))^T + A\mathbb{E}(\xi)(A\mathbb{E}(\xi))^T\right].$$

Simplifying,

$$\begin{aligned}
\text{Cov}(y) &= \mathbb{E}\left[A\xi\xi^T A^T - A\mathbb{E}(\xi)\xi^T A^T - A\xi(A\mathbb{E}(\xi))^T + A\mathbb{E}(\xi)(A\mathbb{E}(\xi))^T\right] \\
&= \mathbb{E}\left[A\xi\xi^T A^T - A\mathbb{E}(\xi)\xi^T A^T - A\xi\mathbb{E}(\xi)^T A^T + A\mathbb{E}(\xi)\mathbb{E}(\xi)^T A^T\right] \\
&= \mathbb{E}\left[A(\xi\xi^T - \mathbb{E}(\xi)\xi^T - \xi\mathbb{E}(\xi)^T + \mathbb{E}(\xi)\mathbb{E}(\xi)^T)A^T\right] \\
&= \mathbb{E}\left[A(\xi - \mathbb{E}(\xi))(\xi - \mathbb{E}(\xi))^T A^T\right].
\end{aligned}$$

Let $X = (\xi - \mathbb{E}(\xi))(\xi - \mathbb{E}(\xi))^T \in \mathbb{R}^{n\times n}$ such that $\text{Cov}(y) = \mathbb{E}\left[AXA^T\right]$. Let $W = AXA^T$. We can write the i,j element of W as $W_{ij} = (\sum_{k=1}^n a_{ik}x_{kj})a_{ji}$, where a_{ij} is the i,j element of A and x_{ij} is the i,j element of X. From this, $\mathbb{E}(W_{ij}) = \mathbb{E}((\sum_{k=1}^n a_{ik}x_{kj})a_{ji})$. Since the expectation is a linear operation for random variables and A is a matrix of constant values, it follows that $\mathbb{E}(W_{ij}) = (\sum_{k=1}^n a_{ik}\mathbb{E}(x_{kj}))a_{ji}$. Therefore, $\mathbb{E}(W) = A\mathbb{E}(X)A^T$. Hence,

$$\text{Cov}(y) = \text{Cov}(A\xi) = A\mathbb{E}(X)A^T = A\mathbb{E}\left[(\xi - \mathbb{E}(\xi))(\xi - \mathbb{E}(\xi))^T\right]A^T = A\text{Cov}(\xi)A^T. \blacksquare$$

(c) Prove that

$$\text{Cov}(\xi) = \mathbb{E}(\xi\xi^T) - \mathbb{E}(\xi)\mathbb{E}(\xi)^T.$$

Solution.

$$\begin{aligned}
\text{Cov}(\xi) &= \mathbb{E}\left[(\xi - \mathbb{E}(\xi))(\xi - \mathbb{E}(\xi))^T\right] \\
&= \mathbb{E}\left[\xi\xi^T - \mathbb{E}(\xi)\xi^T - \xi\mathbb{E}(\xi)^T + \mathbb{E}(\xi)\mathbb{E}(\xi)^T\right] \\
&= \mathbb{E}[\xi\xi^T] - \mathbb{E}[\mathbb{E}(\xi)\xi^T] - \mathbb{E}[\xi\mathbb{E}(\xi)^T] + \mathbb{E}[\mathbb{E}(\xi)\mathbb{E}(\xi)^T] \\
&= \mathbb{E}[\xi\xi^T] - 2\mathbb{E}(\xi)\mathbb{E}(\xi)^T + \mathbb{E}(\xi)\mathbb{E}(\xi)^T,
\end{aligned}$$

where in the last step we utilize the fact that $\mathbb{E}(\mathbb{E}(\xi)) = \mathbb{E}(\xi)$ since the expectation of a random matrix is deterministic and thus no longer a random variable. Thus,

$$\text{Cov}(\xi) = \mathbb{E}[\xi\xi^T] - \mathbb{E}(\xi)\mathbb{E}(\xi)^T. \blacksquare$$

(d) If $\xi_1, \dots, \xi_n$ are independent random variables, prove that

$$\mathbb{E}(\xi_1 \dots \xi_n) = \mathbb{E}(\xi_1) \dots \mathbb{E}(\xi_n),$$

$$\text{Cov}(\xi) \text{ is a diagonal matrix.}$$

Remark: the first equality says if $\xi_1, \dots, \xi_n$ are independent, then the expectation of the product is the same as the product of the expectations.

Solution. If $\xi_1, \dots, \xi_n$ are independent random variables, it follows that their cumulative distribution function is

$$F_\xi(x_1, \dots, x_n) = \prod_{i=1}^{n} F_{\xi_i}(x_i)$$

for constant $x_1, \dots, x_n$. Therefore,

$$p_\xi(x) = \prod_{i=1}^{n} p_{\xi_i}(x_i) = \prod_{i=1}^{n} \frac{\partial F_{\xi_i}(x_i)}{\partial x_i}.$$

Therefore,

$$\mathbb{E}(\xi_1 \dots \xi_n) = \int_{\mathbb{R}^n} \xi_1 \dots \xi_n \prod_{i=1}^{n} \frac{\partial F_{\xi_i}(x_i)}{\partial x_i} dx = \prod_{i=1}^{n} \int_{-\infty}^{\infty} \xi_i \frac{\partial F_{\xi_i}(x_i)}{\partial x_i} dx_i = \prod_{i=1}^{n} \mathbb{E}(\xi_i). \blacksquare$$

Consider $\text{Cov}(\xi_i, \xi_j)$ as the i, j element of the covariance matrix $\text{Cov}(\xi)$.

$$\begin{aligned}
\text{Cov}(\xi_i, \xi_j) &= \mathbb{E}\left[(\xi_i - \mathbb{E}(\xi_i))(\xi_j - \mathbb{E}(\xi_j))\right] \\
&= \mathbb{E}(\xi_i \xi_j) - 2\mathbb{E}(\xi_i)\mathbb{E}(\xi_j) + \mathbb{E}(\xi_i)\mathbb{E}(\xi_j) \\
&= \mathbb{E}(\xi_i \xi_j) - \mathbb{E}(\xi_i)\mathbb{E}(\xi_j).
\end{aligned}$$

If $i \neq j$, then for independent ξ_i, ξ_j,

$$\text{Cov}(\xi_i, \xi_j) = \mathbb{E}(\xi_i)\mathbb{E}(\xi_j) - \mathbb{E}(\xi_i)\mathbb{E}(\xi_j) = 0.$$

If $i = j$, then,

$$\text{Cov}(\xi_i, \xi_i) = \mathbb{E}(\xi_i^2) - \mathbb{E}(\xi_i)^2 = \text{Var}(\xi_i).$$

Thus, $\text{Cov}(\xi)$ is a diagonal matrix where the diagonal elements are the variance of $\xi_1, \dots, \xi_n$. $\blacksquare$

II. Given a random vector $\xi := (\xi_1, \dots, \xi_n)^T$, we say a deterministic vector $d \in \mathbb{R}^n$ is a realization of ξ if d is a possible outcome of ξ. We say ξ is an n-dimensional Gaussian random vector denoted by $\xi \sim N(\mu, C)$ if its probability density function is

$$p_\xi(x; \mu, C) = \frac{1}{\sqrt{(2\pi)^n \det(C)}} e^{-\frac{1}{2}(x-\mu)^T C^{-1}(x-\mu)}, x \in \mathbb{R}^n.$$

Here, $\mu \in \mathbb{R}^n$ is a vector, $C \in \mathbb{R}^{n \times n}$ is a symmetric positive definite matrix, $\det(C)$ denotes the determinant of C, and C^{-1} is the matrix inverse of C. It is easy to verify that $\mathbb{E}(\xi) = \mu$ and $\text{Cov}(\xi) = C$.

Suppose $\xi \sim N(\mu, C)$ and we know C but not μ. We can compute an estimator for μ, known as the maximum likelihood estimator (MLE) μ^{mle}, from a single realization d of ξ. Here, μ^{mle} is defined by

$$\mu^{mle} := \arg\max_{\mu} p_\xi(d; \mu, C).$$

If $p_\xi > 0$ everywhere, this definition is equivalent to

$$\mu^{mle} := \arg\min_{\mu} -\log p_\xi(d; \mu, C).$$

Compute μ^{mle} in terms of d.

Solution.

$$\begin{aligned}
-\log p_\xi(d; \mu, C) &= -\log\left(\frac{1}{\sqrt{(2\pi)^n \det(C)}} e^{-\frac{1}{2}(d-\mu)^T C^{-1}(d-\mu)}\right) \\
&= -\log\left(\frac{1}{\sqrt{(2\pi)^n \det(C)}}\right) + \frac{1}{2}(d-\mu)^T C^{-1}(d-\mu).
\end{aligned}$$

In order to minimize this expression, it suffices to solve

$$\frac{\partial}{\partial \mu}(-\log p_\xi(d; \mu, C)) = 0.$$

The first term is a constant that does not depend on μ. Therefore, we need only consider

$$\frac{\partial}{\partial \mu}\left(\frac{1}{2}(d-\mu)^T C^{-1}(d-\mu)\right) = 0.$$

$$\begin{aligned}
\frac{\partial}{\partial \mu}\left(\frac{1}{2}(d-\mu)^T C^{-1}(d-\mu)\right) &= 0 \\
\frac{\partial}{\partial \mu}\left(\frac{1}{2}(d^T-\mu^T)(C^{-1}d - C^{-1}\mu)\right) &= 0 \\
\frac{\partial}{\partial \mu}\left(\frac{1}{2}(d^T C^{-1} d + \mu^T C^{-1}\mu - \mu^T C^{-1} d - d^T C^{-1}\mu)\right) &= 0 \\
\frac{1}{2}\left(0 + 2C^{-1}\mu - C^{-1}d - C^{-1}d\right) &= 0 \\
C^{-1}\mu - C^{-1}d &= 0 \\
\mu &= d.
\end{aligned}$$

Taking a second derivative, we see

$$\frac{\partial^2}{\partial \mu^2}(-\log p_\xi(d; \mu, C)) = C^{-1}.$$

Since C is symmetric positive definite, then C^{-1} is also positive definite. This implies that $\mu = d$ is a minimum.

Therefore, the μ that minimizes $-\log p_\xi(d; \mu, C)$ is when $\mu = d$, and thus $\mu^{mle} = d$. ■

III. Consider the linear model with noise

$$\eta := X^T\beta^* + \epsilon$$

where $X \in \mathbb{R}^{p\times n}$ is a deterministic sample matrix, $\beta^* \in \mathbb{R}^p$ is an unknown deterministic parameter, $\epsilon = (\epsilon_1, \dots, \epsilon_n)^T$ is a random vector with $\epsilon \sim N(0, \sigma^2 I_n)$, where I_n is the $n \times n$ identity matrix and $\sigma > 0$ is a constant (see Problem #2 for the definition of $N(0, \sigma^2 I_n)$).

(a) Prove that

$$\eta \sim N(X^T\beta^*, \sigma^2 I_n).$$

Solution.

$$\begin{aligned}\mathbb{E}(\eta) &= \mathbb{E}(X^T\beta^* + \epsilon)\\ &= \mathbb{E}(X^T\beta^*) + \mathbb{E}(\epsilon)\\ &= X^T\beta^* + 0\\ &= X^T\beta^*.\end{aligned}$$

$$\begin{aligned}\mathrm{Cov}(\eta) &= \mathbb{E}\left[(X^T\beta^* + \epsilon - \mathbb{E}(X^T\beta^* + \epsilon))(X^T\beta^* + \epsilon - \mathbb{E}(X^T\beta^* + \epsilon))^T\right]\\ &= \mathbb{E}\left[\epsilon\epsilon^T\right]\\ &= \mathbb{E}\left[(\epsilon - 0)(\epsilon - 0)^T\right]\\ &= \mathbb{E}\left[(\epsilon - \mathbb{E}(\epsilon))(\epsilon - \mathbb{E}(\epsilon))^T\right]\\ &= \mathrm{Cov}(\epsilon) = \sigma^2 I_n.\end{aligned}$$

Therefore, η is a random variable with mean $X^T\beta^*$ and covariance $\sigma^2 I_n$. Since ϵ is a normally distributed random vector and X and β are deterministic, it follows that η is also a normally distributed random vector. Thus, $\eta \sim N(X^T\beta^*, \sigma^2 I_n)$.■

(b) Suppose $\sigma > 0$ is known yet β^* is unknown. Given a realization y of η, the maximum likelihood estimate β^{mle} is defined by

$$\beta^{mle} := \arg\min_{\beta} -\log p_\eta(y; X^T\beta, \sigma^2 I_n).$$

Prove that β^{mle} is exactly the least square estimator β^{ls} which we discussed before for the linear model with noise $y = X^T\beta^* + \epsilon$. Remark: this result gives another characterization of the least square estimator from the statistical point of view.

Solution. Let $C = \sigma^2 I_n$. We write the probability density function of η as

$$p_\eta(y; X^T\beta, C) = \frac{1}{\sqrt{(2\pi)^n \det(C)}} e^{-\frac{1}{2}(y - X^T\beta)^T C^{-1}(y - X^T\beta)}.$$

To solve for β^{mle} , it suffices to solve

$$\frac{\partial}{\partial\beta}\left(-\log p_\eta(y; X^T\beta, \sigma^2 I_n)\right) = 0.$$

$$\frac{\partial}{\partial \beta}\left(-\log p_\eta(y; X^T\beta, \sigma^2 I_n)\right) = 0$$

$$\frac{\partial}{\partial \beta}\left(-\log\left(\frac{1}{\sqrt{(2\pi)^n \det(C)}}\right) + \frac{1}{2}(y - X^T\beta)^T C^{-1}(y - X^T\beta)\right) = 0.$$

The first term carries no dependency on β, therefore,

$$\frac{\partial}{\partial \beta}\left(\frac{1}{2}(y - X^T\beta)^T C^{-1}(y - X^T\beta)\right) = 0.$$

Expanding and taking the derivative

$$\begin{aligned}
\frac{\partial}{\partial \beta}\left(\frac{1}{2}(y^T C^{-1} y - \beta^T X C^{-1} y - y^T C^{-1} X^T \beta + \beta^T X C^{-1} X^T \beta)\right) &= 0 \\
-XC^{-1}y + XC^{-1}X^T\beta &= 0 \\
-X(\sigma^2 I_n)^{-1} y + X(\sigma^2 I_n)^{-1} X^T \beta &= 0 \\
-Xy + XX^T\beta &= 0 \\
(XX^T)^{-1}Xy &= \beta.
\end{aligned}$$

The β that minimizes $-\log p_\eta(y; X^T\beta, \sigma^2 I_n)$ is thus $\beta^* = (XX^T)^{-1}Xy$, which is exactly the least square solution we discussed from before. ■

IV. Consider the linear model with noise

$$\eta = X^T\beta + \epsilon.$$

Here $X \in \mathbb{R}^{p\times n}$ is a deterministic sample matrix, $\beta \sim N(0, C_\beta)$ is a p-dimensional Gaussian random vector with symmetric positive definite $C_\beta \in \mathbb{R}^{p\times p}, \epsilon \sim N(0, C_\epsilon)$ is an n-dimensional Gaussian random vector with symmetric positive definite $C_\epsilon \in \mathbb{R}^{n\times n}$, β and ϵ are independent.

In the lectures, we derived the maximum a posteriori estimator (MAP) β^{map} for β under the assumption that $C_\beta = \sigma_\beta^2 I_p$ and $C_\epsilon = \sigma_\epsilon^2 I_n$ for some $\sigma_\beta > 0$, $\sigma_\epsilon > 0$. In this problem, we extend the result to the general positive definite C_β and C_ϵ. Recall tat β^{map} is defined as

$$\beta^{map} := \arg\max_t p_{\beta|\eta}(t|s).$$

Prove that

$$\beta^{map} = (XC_\epsilon^{-1}X^T + C_\beta^{-1})^{-1}XC_\epsilon^{-1}\eta.$$

Solution. Suppose

$$p_\beta(t; 0, C_\beta) = \frac{1}{\sqrt{(2\pi)^p \det(C_\beta)}} e^{-\frac{1}{2}t^T C_\beta^{-1} t}$$

and

$$p_\epsilon(r; 0, C_\epsilon) = \frac{1}{\sqrt{(2\pi)^n \det(C_\epsilon)}} e^{-\frac{1}{2}r^T C_\epsilon^{-1} r}.$$

By Bayes' Theorem,

$$p_{\beta|\eta}(t|s) = \frac{p_{\eta|\beta}(s|t)p_\beta(t)}{p_\eta(s)},$$

and thus
$$\log p_{\beta|\eta}(t|s) = \log p_{\eta|\beta}(s|t) + \log p_\beta(t) - \log p_\eta(s).$$
In order to maximize $p_{\beta|\eta}(t|s)$, it suffices to minimize $-\log p_{\beta|\eta}(t|s)$ by solving
$$\frac{\partial}{\partial t}(-\log p_{\beta|\eta}(t|s)) = 0.$$
We showed previously that $\eta \sim N(X^T\beta, \sigma_\epsilon^2 I_n)$. It follows for the more general case that $\eta \sim N(X^T\beta, C_\epsilon)$. If $\beta = t$, then
$$p_{\eta|\beta}(s|t) = \frac{1}{\sqrt{(2\pi)^n}\det(C_\epsilon)} e^{-\frac{1}{2}(s-X^Tt)^T C_\epsilon^{-1}(s-X^Tt)}.$$
$p_\eta(s)$ has no dependence on t, and thus it will go to 0 in the derivative. Therefore,
$$\begin{aligned}
\frac{\partial}{\partial t}(-\log p_{\beta|\eta}(t|s)) &= 0\\
\frac{\partial}{\partial t}\left(-\log\left(\frac{1}{\sqrt{(2\pi)^n}\det(C_\epsilon)}\right) + \frac{1}{2}(s-X^Tt)^TC_\epsilon^{-1}(s-X^Tt) - \log\left(\frac{1}{\sqrt{(2\pi)^p}\det(C_\beta)}\right) + \frac{1}{2}t^TC_\beta^{-1}t\right) &= 0\\
\frac{\partial}{\partial t}\left(\frac{1}{2}(s-X^Tt)^TC_\epsilon^{-1}(s-X^Tt) + \frac{1}{2}t^TC_\beta^{-1}t\right) &= 0\\
\frac{\partial}{\partial t}\left(\frac{1}{2}(s-X^Tt)^T(C_\epsilon^{-1}s - C_\epsilon^{-1}X^Tt) + \frac{1}{2}t^TC_\beta^{-1}t\right) &= 0\\
\frac{\partial}{\partial t}\left(\frac{1}{2}(s^T-t^TX)(C_\epsilon^{-1}s - C_\epsilon^{-1}X^Tt) + \frac{1}{2}t^TC_\beta^{-1}t\right) &= 0\\
\frac{\partial}{\partial t}\left(\frac{1}{2}(s^TC_\epsilon^{-1}s - t^TXC_\epsilon^{-1}s - s^TC_\epsilon^{-1}X^Tt + t^TXC_\epsilon^{-1}X^Tt) + \frac{1}{2}t^TC_\beta^{-1}t\right) &= 0\\
-XC_\epsilon^{-1}s + XC_\epsilon^{-1}X^Tt + C_\beta^{-1}t &= 0\\
(XC_\epsilon^{-1}X^T + C_\beta^{-1})t - XC_\epsilon^{-1}s &= 0.
\end{aligned}$$

Therefore, the t that minimizes $-\log p_{\beta|\eta}(t|s)$ is $t^* = (XC_\epsilon^{-1}X^T + C_\beta^{-1})^{-1}XC_\epsilon^{-1}s$. The general maximum likelihood a posteriori estimate of β is therefore $\beta^{map} = (XC_\epsilon^{-1}X^T + C_\beta^{-1})^{-1}XC_\epsilon^{-1}\eta$. ■