ORFE 524: Statistical Theory and Methods
Homework 2

Zachary Hervieux-Moore

Friday 14^{th} October, 2016

Exercise 1: Let $T(X)$ be a sufficient statistic for $\mathcal{P}$. Consider the following experiment.

- Draw $X \sim P_\theta$, where $P_\theta \in \mathcal{P}$
- Compute $T(X) = t$
- Draw $X' \sim P_{X|t}$

Show that X' has the same (unconditional) distribution as X. For simplicity you can assume that all distributions are discrete.

Answer: We use the law of total probability,

$$P_{X'}(x') = \sum_{t \in T} P_{X'|t}(x'|T(X) = t) P_T(T(X) = t)$$

By assumption,

$$P_{X'|t}(x'|T(X) = t) = P_{X|t}(x'|T(X) = t)$$

Thus,

$$P_{X'}(x') = \sum_{t \in T} P_{X|t}(x'|T(X) = t) P_T(T(X) = t) = P_X(x')$$

Hence $P_{X'} = P_X$

Exercise 2: Suppose that $\Theta \subset \mathbb{R}^d$. Let $\mathcal{P} = \{P_\theta, \theta \in \Theta\}$, where P_θ has a probability density function (with respect to Lebesgue measure)

$$f^\theta(x) = h(x)l(\theta)e^{\alpha(\theta)^T T(x)}, \quad x \in \mathbb{R}^d,$$

where $T(x)$ is a r-vector. Show that $\mathcal{P}$ has an equivalent representation $\mathcal{P} = \{P_\alpha, \alpha \in \mathcal{A}\}$ for some set $\mathcal{A} \subset \mathbb{R}^r$, where the density of P_α is

$$f^\alpha(x) = h'(x)l'(\alpha)e^{\alpha^T T(x)}, \quad x \in \mathbb{R}^d,$$

In other words, you need to verify that the following holds.

- For each $\theta \in \Theta$, there exists an $\alpha \in \mathcal{A}$ such that $P_\theta = P_\alpha$
- Parameter α uniquely determines P_α, that is, for any $\alpha \neq \alpha'$, P_α and $P_{\alpha'}$ are different

Answer: To satisfy the first condition, define the set $\mathcal{A}$ as follows:

$$\mathcal{A} = \{\alpha : \alpha = \alpha(\theta), \theta \in \Theta\}$$

Thus, for all $\theta \in \Theta$, then,

$$f^\theta(x) = h(x)l(\theta)e^{\alpha(\theta)^T T(x)} = h(x)l'(\alpha)e^{\alpha^T T(x)} = f^\alpha(x)$$

Where $l'(\alpha)$ is defined to be $l(\theta)$ such that $\alpha(\theta) = \alpha$ Now, since densities integrate to 1,

$$\begin{aligned} 1 &= \int_{\mathbb{R}^d} h(x)l(\theta)e^{\alpha(\theta)^T T(x)} \\ \implies l(\theta) &= \frac{1}{\int_{\mathbb{R}^d} h(x)e^{\alpha(\theta)^T T(x)}} \end{aligned}$$

That is, $l(\theta)$ is determined uniquely by $\alpha(\theta)$. Thus if $\alpha \neq \alpha'$,

$$\begin{aligned} \alpha \neq \alpha' \iff \alpha(\theta) \neq \alpha(\theta') \implies \theta \neq \theta' \\ \iff P_\theta \neq P_{\theta'} \iff P_{\alpha(\theta)} \neq P_{\alpha(\theta')} \implies P_\alpha \neq P_{\alpha'} \end{aligned}$$

As $P_\theta = P_{\alpha(\theta)}$

Exercise 3: Let $\mathcal{P}$ be some family of distributions, and let $\mathcal{P}' \subseteq \mathcal{P}$ be a smaller family of distributions contained in $\mathcal{P}$. Suppose that T is sufficient for $\mathcal{P}$ and minimal sufficient for $\mathcal{P}'$. Show that T must also be minimal sufficient for $\mathcal{P}$.

Answer: Let T' be sufficient for $\mathcal{P}$. Then, T' is sufficient for $\mathcal{P}'$ since $\mathcal{P}' \subseteq \mathcal{P}$. However, T is minimal sufficient for $\mathcal{P}'$ so,

$$T = f(T')$$

Thus, we have for all T' sufficient for $\mathcal{P}$, $T = f(T')$. Since T is also sufficient for $\mathcal{P}$, then this is precisely the definition of minimal sufficiency. Therefore, T is also minimal sufficient for $\mathcal{P}$.

Exercise 4: Let n random variables $X = \{X_i\}_1^n \sim \mathcal{N}^n(\mu, \mu)$, where μ is a positive real number. Here $\mathcal{N}(\mu, \mu)$ is the univariate Gaussian distribution with both mean and variance equal to μ and $\mathcal{N}^n(\mu, \mu)$ is its n-th product distribution. Consider the family distributions $\mathcal{P} = \{\mathcal{N}^n(\mu, \mu), \mu > 0\}$

1) Find a minimal sufficient statistic for $\mathcal{P}$

2) In the case of $n = 1$, consider another statistic $T_0(x) = x$. Is it sufficient? Is it minimal?

Answer: First, the distribution for $X \sim \mathcal{N}^n(\mu, \mu)$ is:

$$f_\theta(x^n) = \prod_{i=1}^{n} \frac{1}{\sqrt{2\mu\pi}} e^{\frac{-(x_i-\mu)^2}{2\mu}}$$

Thus we need $f_\theta(x^n)/f_\theta(y^n)$ to be independent of μ:

$$\frac{f_\theta(x^n)}{f_\theta(y^n)} = \frac{\prod_{i=1}^{n} \frac{1}{\sqrt{2\mu\pi}} e^{\frac{-(x_i-\mu)^2}{2\mu}}}{\prod_{i=1}^{n} \frac{1}{\sqrt{2\mu\pi}} e^{\frac{-(y_i-\mu)^2}{2\mu}}} = e^{-\frac{1}{2\mu}\sum_{i=1}^{n}(x_i-\mu)^2-(y_i-\mu)^2}$$

Expanding the square,

$$= e^{-\frac{1}{2\mu}\sum_{i=1}^{n} x_i^2 - 2x_i\mu - y_i^2 - 2y_i\mu} = e^{-\sum_{i=1}^{n} \frac{x_i^2}{2\mu} - x_i - \frac{y_i^2}{2\mu} - y_i}$$

For this to be independent of μ, we need $\sum_{i=1}^{n} \frac{x_i^2}{2\mu} = \sum_{i=1}^{n} \frac{y_i^2}{2\mu}$, or:

$$\sum_{i=1}^{n} x_i^2 = \sum_{i=1}^{n} y_i^2$$

Thus, $T(X) = \sum_{i=1}^{n} x_i^2$ is a minimal sufficient statistic.

If $n = 1$, and $T_0(x) = x$. Then, $P(X = y|T = x) = 0$ if $y \neq x$ and $= 1$ if $x = y$. Thus the the law of probability simplifies to,

$$P_{X|t} = \frac{P(X = x, T = x)}{P(T = x)} = \frac{P(X = x|T = x)P(T = x)}{P(T = x)} = \frac{\frac{1}{\sqrt{2\mu\pi}} e^{\frac{-(x-\mu)^2}{2\mu}}}{\frac{1}{\sqrt{2\mu\pi}} e^{\frac{-(x-\mu)^2}{2\mu}}} = 1$$

Which is independent of μ, so $T_0(x) = x$ is sufficient. But, $T(x) = x^2$ so $T_0(x) = \sqrt{T(x)}$ but $f(x) = \sqrt{x}$ is not one-to-one, so $T_0(x)$ is not minimal.

Exercise 5: Suppose $X_1, \dots, X_n$ are i.i.d. d-dimensional Gaussian random vectors with mean μ and covariance Σ. Argue that $(\hat{\mu}, \hat{\Sigma})$ is a minimal sufficient statistic for $\mathcal{P} = \{\mathcal{N}^n(\mu, \Sigma)\}$, the family of Gaussian distribution with unknown Σ and μ.

Answer: Using the same method as the previous question, we want to find a $T(X^n)$ that makes $\frac{f_\theta(x^n)}{f_\theta(y^n)}$ independent of θ. We know that

$$f_\theta(x^n) = \frac{1}{\left(\sqrt{(2\pi)^d|\Sigma|}\right)^n} \exp\left(-\frac{1}{2}\sum_{i=1}^n (x_i - \mu)^T \Sigma^{-1}(x_i - \mu)\right)$$

So,

$$\begin{aligned}
\frac{f_\theta(x^n)}{f_\theta(y^n)} &= \exp\left(-\frac{1}{2}\sum_{i=1}^n (x_i - \mu)^T \Sigma^{-1}(x_i - \mu) - (y_i - \mu)^T \Sigma^{-1}(y_i - \mu)\right) \\
&= \exp\left(-\frac{1}{2}\sum_{i=1}^n x_i^T \Sigma^{-1} x_i - 2x_i^T \Sigma^{-1}\mu - y_i^T \Sigma^{-1} y_i - 2y_i^T \Sigma^{-1}\mu\right)
\end{aligned}$$

Note that $x_i^T\Sigma^{-1}x_i = \langle \mathrm{vec}(\Sigma^{-1}), \mathrm{vec}(x_i x_i^T)\rangle$ and $-2x_i^T\Sigma^{-1}\mu = \langle -2\Sigma^{-1}\mu, x_i\rangle$. Rewriting in these terms,

$$= \exp\left(-\frac{1}{2}\sum_{i=1}^n \langle \mathrm{vec}(\Sigma^{-1}), \mathrm{vec}(x_i x_i^T) - \mathrm{vec}(y_i y_i^T)\rangle + \langle -2\Sigma^{-1}\mu, x_i - y_i\rangle\right)$$

Thus, for this to be independent of Σ and μ, we need $\sum_{i=1}^n \mathrm{vec}(x_i x_i^T) = \sum_{i=1}^n \mathrm{vec}(y_i y_i^T)$ and $\sum_{i=1}^n x_i = \sum_{i=1}^n y_i$ so that both inner products will be always 0. So $T = (\sum_{i=1}^n x_i, \sum_{i=1}^n \mathrm{vec}(x_i x_i^T))$ is minimally sufficient. We'll now use a series of one-to-one transformations to get $(\hat{\mu}, \hat{\Sigma})$.

First, the $\mathrm{vec}(\cdot)$ operator is clearly a one-to-one transformation. So, $T = (\sum_{i=1}^n x_i, \sum_{i=1}^n x_i x_i^T)$ is also minimally sufficient. Now we divide by n which is one-to-one. $T = (\frac{1}{n}\sum_{i=1}^n x_i, \frac{1}{n}\sum_{i=1}^n x_i x_i^T)$. Now we use the transformation $f(T(x, y)) = T(x, y - x^2)$ which is one-to-one since $f^{-1}(T(x, y)) = T(x, y + x^2)$ is the inverse. So the following T is minimal since it is a series of one-to-one transformations of a minimal statistic,

$$T = \left(\frac{1}{n}\sum_{i=1}^n x_i, \frac{1}{n}\sum_{i=1}^n x_i x_i^T - \left(\frac{1}{n}\sum_{i=1}^n x_i\right)\left(\frac{1}{n}\sum_{i=1}^n x_i\right)\right) = \left(\hat{\mu}, \hat{\Sigma}\right)$$

Exercise 6: Let $x = \{x_i\}_{i=1}^n$ be the realization of n i.i.d. Gaussian random variables $X = \{X_i\}_1^n \sim \mathcal{N}^n(\mu, \sigma^2)$, where $\mu \in \mathbb{R}$ and $\sigma^2 > 0$, show that the maximum likelihood estimator for $\theta = (\mu, \sigma^2)$ is $(\bar{x}, S_n^2)$, where

$$\bar{x} = \frac{1}{n}\sum_{i=1}^n x_i \text{ and } S_n^2 = \frac{1}{n}\sum_{i=1}^n (x_i - \bar{x})^2$$

Answer: We write out the likelihood maximization problem,

$$L(\theta; x^n) = \prod_{i=1}^n \frac{1}{\sqrt{2\sigma^2\pi}} e^{\frac{-(x_i-\mu)^2}{2\sigma^2}} = \frac{1}{\left(\sqrt{2\sigma^2\pi}\right)^n} e^{\sum_{i=1}^n \frac{-(x_i-\mu)^2}{2\sigma^2}}$$

Now taking the log,

$$L(\theta; x^n) = \log\left(\frac{1}{\left(\sqrt{2\sigma^2\pi}\right)^n}\right) + \sum_{i=1}^n \frac{-(x_i-\mu)^2}{2\sigma^2}$$

Taking the partial derivatives,

$$\frac{\partial L(\theta; x^n)}{\partial \mu} = \sum_{i=1}^n \frac{x_i - \mu}{\sigma^2}$$
$$\frac{\partial L(\theta; x^n)}{\partial \sigma^2} = -\frac{n}{\sigma^2} + \sum_{i=1}^n \frac{(x_i-\mu)^2}{2\sigma^4}$$

Setting these to 0 and solving for the parameters,

$$\sum_{i=1}^n \frac{x_i - \mu}{\sigma^2} = 0 \implies \mu = \frac{1}{n}\sum_{i=1}^n x_i = \bar{x}$$
$$-\frac{n}{2\sigma^2} + \sum_{i=1}^n \frac{(x_i-\mu)^2}{2\sigma^4} = 0 \implies \sum_{i=1}^n \frac{(x_i-\bar{x})^2}{\sigma^4} = \frac{n}{\sigma^2}$$
$$\implies \sigma^2 = \frac{1}{n}\sum_{i=1}^n (x_i - \bar{x})^2 = S_n^2$$

Which is a maximum since $L_{\mu\mu}L_{\sigma\sigma} > L_{\mu\sigma}^2$ and $L_{\mu\mu} < 0$ and hence satisfies Sylvester's criterion.

Exercise 7: Let $x = \{x_i\}_1^n$ be i.i.d. realizations of a random variable $\xi \sim \text{Uniform}([0,\theta])$, where $\theta > 0$. We have shown that the maximum likelihood estimator for θ is $\hat{\theta} = \max_{x_i} x_i$. Show that

1) $\hat{\theta}$ has a density with respect to Lebesgue measure

2) $\hat{\theta}$ is biased

Note: In question 1), you only need to show that the cumulative distribution function of $\hat{\theta}$ is absolutely continuous (easier than showing Lebesgue domination).

Answer: The density $\hat{\theta}$ is given as

$$
\begin{aligned}
P_{\hat{\theta}} &= P(\max(X_1, \ldots, X_n) \in [x, x+\epsilon]) \\
&= \sum_{i=1}^{n} P(X_i \in [x, x+\epsilon]) P(\text{all others} < x) \\
&= nP(X_1 \in [x, x+\epsilon]) P(X_2 < x) \cdots P(X_n < x) \text{ since i.i.d.} \\
&= nf(x)F(x)^{n-1} = \frac{n}{\theta}\left(\frac{x}{\theta}\right)^{n-1} = \frac{nx^{n-1}}{\theta^n}, \quad x \in [0,\theta]
\end{aligned}
$$

Thus, the CDF is

$$\int_0^x \frac{nt^{n-1}}{\theta^n} dt = \frac{x^n}{\theta^n}$$

Which is a polynomial and hence continous and therefore absolutely continuous. Now to show biasness,

$$E[\hat{\theta}] = \int_0^\theta \frac{nx^n}{\theta^n} dx = \frac{n}{n+1} \cdot \frac{\theta^{n+1}}{\theta^n} = \frac{n}{n+1}\theta \neq \theta$$

Thus, $\hat{\theta}$ is biased.

Exercise 8: This exercise is about the expectation-maximication (EM) algorithm. Derive an EM algorithm for a mixture of K Gaussians with diagonal covariance matrices. In other words, suppose that we observe n i.i.d. observations $\{x_i\}_{i=1}^n$ of d-dimensional random vectors

$$X \sim \sum_{\ell=1}^{K} \frac{1}{K} N(\mu_\ell, \Sigma_\ell), \text{ where } \Sigma_\ell = \mathrm{diag}(\sigma_{\ell,1}^2, \ldots, \sigma_{\ell,d}^2)$$

Derive the EM algorithm that estimates the parameters $\{\mu_\ell, \Sigma_\ell\}_{\ell=1}^K$

Answer: We write the expectation step,

$$Q(\theta, \theta') = E_{Z|X}^{\theta'}[\log f_\theta(x, z)] = E_{Z|X}^{\theta'}[\log \frac{1}{K} N(\mu_\ell, \Sigma_\ell)]$$

$$= \sum_{\ell=1}^{K} \sum_{i=1}^{n} \left[-\frac{1}{2}(x_i - \mu_\ell)^T \Sigma_\ell^{-1} (x_i - \mu_\ell) - \log \left(\frac{1}{K\sqrt{(2\pi)^d |\Sigma_\ell|}} \right) \right] \cdot P_{Z|X}^{\theta'}$$

Taking the gradient with respect to μ to maximize it,

$$\nabla_{\mu_\ell} Q(\theta, \theta') = \nabla_{\mu_\ell} \sum_{\ell=1}^{K} \left[-\frac{1}{2}(x - \mu_\ell)^T \Sigma_\ell^{-1} (x - \mu_\ell) \right] \cdot P_{\theta'}(z = \ell | x_i)$$

$$= -\frac{1}{2} \sum_{\ell=1}^{K} \sum_{i=1}^{n} \nabla_{\mu_\ell} \left[x^T \Sigma_\ell^{-1} x - 2\mu_\ell^T \Sigma_\ell^{-1} x + \mu_\ell^T \Sigma_\ell^{-1} \mu_\ell \right] \cdot P_{\theta'}(z = \ell | x_i)$$

$$= \sum_{i=1}^{n} [\Sigma_\ell^{-1} x - \Sigma_\ell^{-1} \mu_\ell] \cdot P_{\theta'}(z = \ell | x_i)$$

Setting to 0 and solving for μ_ℓ yields

$$\mu_\ell = \frac{\sum_{i=1}^n P_{\theta'}(z = \ell | x_i) x_i}{\sum_{i=1}^n P_{\theta'}(z = \ell | x_i)}$$

Similarly for Σ_ℓ,

$$\Sigma_\ell = \frac{\sum_{i=1}^n P_{\theta'}(z = \ell | x_i)(x_i - \mu_\ell)^T (x_i - \mu_\ell)}{\sum_{i=1}^n P_{\theta'}(z = \ell | x_i)}$$

Exercise 9: This exercise relates maximum likelihood estimation with information theory. Let $\mathcal{P} = \{P_\theta : \theta \in \Theta\}$ be a family of probability measures on $\mathcal{X}$ with density f_θ w.r.t. to some measure σ. Assume for simplicity that all f_θ have the same support. Let $X = \{X_i\}_{i=1}^n$ be n i.i.d. random variables with $X_i \sim P_{\theta_0} \in \mathcal{P}$ for each i, where θ_0 is unknown. Let $x = \{x_i\}_{i=1}^n$ be the realization of X.

1) Let $L(\theta; x)$ be the likelihood function; express $\mathbb{E}[n^{-1} \log L(\theta; X)]$ in terms of information measures (entropy and/or K-L divergence)

2) For any fixed θ, give a simple unbiased estimator of $\mathbb{E}[n^{-1} \log L(\theta; X)]$. Suppose this estimate is close to $\mathbb{E}[n^{-1} \log L(\theta; X)]$ (for instance for sufficiently large n), explain in simple terms (nothing rigorous here), how MLE might be interpreted as minimizing some notion of distance between distributions.

3) Derive a simple form for the K-L divergence between two multivariate Gaussians $\mathcal{N}(\mu_1, \Sigma)$ and $\mathcal{N}(\mu_2, \Sigma)$. Here $\mu_1, \mu_2 \in \mathbb{R}^d$ and $\Sigma \in \mathbb{R}^{d\times d}$ is positive definite.

4) Suppose now that $\mathcal{P} = \{\mathcal{N}(\mu, \Sigma), \mu \in \mathbb{R}^d\}, \Sigma$ fixed. Conclude that maximizing $\mathbb{E}[n^{-1} \log L(\theta; X)]$ is the same as minimizing some distance measure in parameter space.

Answer:

1) First note that, $L(\theta; x) = \prod_{i=1}^n P_\theta$. Thus, we have,

$$\mathbb{E}[n^{-1} \log L(\theta; X)] = \mathbb{E}[n^{-1} \sum_{i=1}^n \log P_\theta] = \mathbb{E}[\log P_\theta]$$
$$= \mathbb{E}[\log P_\theta - \log P_{\theta_0} + \log P_{\theta_0}] = \mathbb{E}[\log \frac{P_\theta}{P_{\theta_0}} + \log P_{\theta_0}]$$
$$= -D(P_{\theta_0}||P_\theta) - H(P_{\theta_0})$$

2) A simple unbiased estimator is $\frac{1}{n}\sum_{i=1}^n \log L(\theta; x_i)$. Unbiasedness is immediate from taking the expectation. Using this, we can see that MLE is equivalent to attempting to minimize $D(P_{\theta_0}||P_\theta)+H(P_{\theta_0})$ which is again equivalent to minimizing $D(P_{\theta_0}||P_\theta)$ since this is the only term dependent on θ. Divergence is a metric that measures the "distance" between distributions. So, MLE can be thought of minimizing the distance between distributions.

3) THe K-L divergence between these two distributions is:

$$\int_{\mathbb{R}} \frac{1}{\sqrt{(2\pi)^k|\Sigma|}} e^{-\frac{1}{2}(x-\mu_1)^T\Sigma^{-1}(x-\mu_1)} \log \frac{\frac{1}{\sqrt{(2\pi)^k|\Sigma|}} e^{-\frac{1}{2}(x-\mu_1)^T\Sigma^{-1}(x-\mu_1)}}{\frac{1}{\sqrt{(2\pi)^k|\Sigma|}} e^{-\frac{1}{2}(x-\mu_2)^T\Sigma^{-1}(x-\mu_2)}} dx$$

$$= \int_{\mathbb{R}} \frac{1}{\sqrt{(2\pi)^k|\Sigma|}} e^{-\frac{1}{2}(x-\mu_1)^T\Sigma^{-1}(x-\mu_1)} - \frac{1}{2}\Big[(x-\mu_1)^T\Sigma^{-1}(x-\mu_1)$$

$$-(x-\mu_2)^T\Sigma^{-1}(x-\mu_2)\Big] dx$$

$$= -\frac{1}{2}\mathbb{E}[\mathrm{tr}\left((x-\mu_1)^T\Sigma^{-1}(x-\mu_1)\right) - \left((x-\mu_2)^T\Sigma^{-1}(x-\mu_2)\right)]$$

$$= -\frac{1}{2}\left(\mathrm{tr}\left(\mathbb{E}[(x-\mu_1)^T(x-\mu_1)]\Sigma^{-1}\right) - \mathbb{E}\left[(x-\mu_2)^T\Sigma^{-1}(x-\mu_2)\right]\right)$$

$$= -\frac{1}{2}\left(\mathrm{tr}(\Sigma\Sigma^{-1}) - \mathbb{E}\left[(x+\mu_1-\mu_1-\mu_2)^T\Sigma^{-1}(x+\mu_1-\mu_1-\mu_2)\right]\right)$$

$$= -\frac{1}{2}\Big(\mathrm{tr}(I_d) - \mathbb{E}\left[(\mu_1-\mu_2)^T\Sigma^{-1}(\mu_1-\mu_2)\right]$$

$$-\mathbb{E}\left[\mathrm{tr}((x-\mu_2)^T\Sigma^{-1}(x-\mu_2))\right]\Big)$$

$$= -\frac{1}{2}\left(d - (\mu_1-\mu_2)^T\Sigma^{-1}(\mu_1-\mu_2) - \mathbb{E}\left[\mathrm{tr}((x-\mu_2)^T(x-\mu_2)\Sigma^{-1})\right]\right)$$

$$= -\frac{1}{2}\left(d - (\mu_1-\mu_2)^T\Sigma^{-1}(\mu_1-\mu_2) - \mathrm{tr}(\Sigma\Sigma^{-1})\right)$$

$$= \frac{1}{2}(\mu_1-\mu_2)^T\Sigma^{-1}(\mu_1-\mu_2)$$

4) Using parts 2) and 3), we can see that maximizing $\mathbb{E}[n^{-1}\log L(\theta; X)]$ is equivalent to minimizing the K-L divergence of the two. In this case, it comes down to minimizing $\frac{1}{2}(\mu_\theta - \mu_{\theta_0})^T\Sigma^{-1}(\mu_\theta - \mu_{\theta_0})$ and so you are trying to minimize the distance between parameters μ_θ and μ_{θ_0}.

Exercise 10: Suppose we have data $(x_i, y_i)_{i=1}^n$, where $x_i \in \mathbb{R}^d$. The Ridge estimator of the linear model $Y = X^T\beta + \epsilon, \mathbb{E}[\epsilon] = 0, \epsilon \perp\!\!\!\perp X$, is defined as the minimizer of the following problem:

$$\min_{\beta \in \mathbb{R}^d} \sum_{i=1}^n (y_i - x_i^T\beta)^2 + \lambda\|\beta\|^2, \quad \lambda \geq 0,$$

where $\|\cdot\|$ is the Euclidean norm in $\mathbb{R}^d$.

1) Show that there exists some $\lambda \geq 0$, such that the above minimization problem has a unique minimizer.

2) Derive the minimizer when the unique minimizer exists.

Answer:

1) First write the Ridge regressor in matrix notation,

$$\min_{\beta \in \mathbb{R}^d} (y - X^T\beta)^T(y - X^T\beta) + \lambda\|\beta\|^2$$

We take the gradient of the above expression with respect to β,

$$\nabla_\beta (y - X^T\beta)^T(y - X^T\beta) + \lambda\|\beta\|^2$$
$$= -2Xy + 2XX^T\beta + 2\lambda\beta$$

Convex in β, so set to 0 and minimize,

$$Xy = (XX^T + \lambda I)\beta$$

Note that we can invert $(XX^T + \lambda I)$ since XX^T is positive semi-definite and λI is positive defintie. So their sum is positive definite and hence invertible,

$$\beta = (XX^T + \lambda I)^{-1}Xy$$

Note, $\lambda = 0$ can only happen if X is invertible. The minimizer is unique since it is a solution to a linear equation.

2) the minimizer β is shown above.

Exercise 11:

1) For Ridge regression, derive a MAP interpretation. That is come up with a proper Bayesian setting where the MAP estimator corresponds to the Ridge estimator. You can consider a fixed design setting.

2) Reduce the general polynomial model

$$Y = \text{poly}_k(X) + \epsilon, \quad \mathbb{E}[\epsilon] = 0, \quad \epsilon \perp\!\!\!\perp X,$$

to the linear model and derive a solution. Note that $\text{poly}_k(X), x \in \mathbb{R}^d$ is any polynomial of some degree $k (k \geq 1)$, i.e.

$$\text{poly}_k(X) = \sum_{\ell \in \mathbb{N}^d : \sum \ell_i \leq k} w_\ell x^\ell$$

Answer:

1) Let us assume we have $Y = X^T\beta + \epsilon, \quad \epsilon \sim \mathcal{N}(0, \sigma^2), \quad \epsilon \perp\!\!\!\perp X$. This means that,

$$f_{y_i|\beta_i,x_i} \sim \mathcal{N}(x_i^T\beta, \sigma^2)$$

By the fortune of hindsight, we pick the priori distribution to be

$$f_{\beta|x} \sim \mathcal{N}(0, I\frac{\sigma^2}{\lambda})$$

Thus, we write the MAP problem,

$$\arg\max_\beta f_{\beta|y,x} = \frac{f_{y|\beta,x}f_{\beta|x}}{f_{y|x}} \propto f_{y|\beta,x}f_{\beta|x} \propto e^{\sum_{i=1}^n -\frac{(y_i - x_i^T\beta)^2}{2\sigma^2}} \cdot e^{-\frac{\lambda}{\sigma^2}\beta^T\beta}$$

Now, taking the log,

$$\arg\max_\beta \sum_{i=1}^n -\frac{(y_i - x_i^T\beta)^2}{2\sigma^2} - \frac{\lambda}{\sigma^2}\beta^T\beta$$

Which flipping the signs and taking the min instead,

$$\arg\min_\beta \sum_{i=1}^n (y_i - x_i^T\beta)^2 + \lambda\|\beta\|^2$$

Which is the Ridge regression problem as desired.

2) We can choose to write the polynomial as

$$
\begin{aligned}
Y &= AX + \epsilon \\
A &= \begin{bmatrix} w^{\ell_(0,0,\ldots,0)} & w^{\ell_(0,1,\ldots,0)} & \ldots & w^{\ell_(k,0,\ldots,0)} & \ldots \end{bmatrix} \\
X^T &= \begin{bmatrix} x^{\ell_(0,0,\ldots,0)} & x^{\ell_(0,1,\ldots,0)} & \ldots & x^{\ell_(k,0,\ldots,0)} & \ldots \end{bmatrix}
\end{aligned}
$$

You can think of these vectors as all the possible combinations of the terms in the polynomials and the A matrix as the coefficients.

Where the ℓ's satisfy $\sum \ell_i \leq k$. Now we can use the same techniques shown in question 10 and Ridge regression to solve for the solution.