Statistics review : Solutions

Semaine de pré-rentrée du master MVA

Multinomial random variables

1.

If $Z = (Z_1, ..., Z_K) \sim \mathcal{M}(\pi_1, ..., \pi_K; 1)$ we have

$$\begin{aligned} P(Z_k = 1) &= P(Z = e^{(k)}), \text{ with } e^{(k)} = (0, ..., 0, \underbrace{1}_{k}, 0, ..., 0) \\ &= \binom{1}{e^{(k)}} \prod_{j=1}^{K} \pi_j^{e_j^{(k)}} 1_{\left\{\sum_{j=1}^{K} e_j^{(k)} = 1\right\}} \\ &= \pi_k \end{aligned}$$

2.

For $(n_1, ..., n_K) \in \mathcal{N}^K$ such that $\sum_k n_k = n$, let

$$\mathcal{Z}_{n_1,..,n_K} := \left\{ (z^{(1)}, ..., z^{(n)}) \in \left(\{0;1\}^K\right)^n \mid \forall i \in \{1, \dots, n\}, \sum_{k=1}^{K} z_k^{(i)} = 1 \text{ and } \forall k \in \{1, \dots, K\}, \sum_{i=1}^{n} z_k^{(i)} = n_k \right\}$$

$$\begin{aligned} P(N_1 = n_1, ..., N_K = n_K) &= P((Z^{(1)}, ..., Z^{(n)}) \in \mathcal{Z}_{n_1,..,n_K}) \\ &= \sum_{(z^{(1)},...,z^{(n)}) \in \mathcal{Z}_{n_1,..,n_K}} P((Z^{(1)}, ..., Z^{(n)}) = (z^{(1)}, ..., z^{(n)})) \\ &= \sum_{(z^{(1)},...,z^{(n)}) \in \mathcal{Z}_{n_1,..,n_K}} P(Z^{(1)} = z^{(1)}) ... P(Z^{(n)} = z^{(n)}) \\ &= \sum_{(z^{(1)},...,z^{(n)}) \in \mathcal{Z}_{n_1,..,n_K}} \pi_1^{n_1} \pi_2^{n_2} \dots \pi_K^{n_K} \\ &= \binom{n}{n_1, ..., n_K} \prod_{k=1}^{K} \pi_k^{n_k} 1_{\{\sum_i n_i = n\}} \end{aligned}$$

This shows that $N := (N_1, ..., N_K)$ follows the distribution $\mathcal{M}(\pi_1, ..., \pi_K; n)$ since the multinomial coefficient

$$\binom{n}{n_1, ..., n_K} := \frac{n!}{n_1! \dots n_K!}$$

is exactly equal to the number of ordered partitions of $\{1, \dots, n\}$ into sets of cardinalities $n_1, \dots, n_K$.

Sufficient Statistic

We first show that the conditional independence statement implies the proposed factorization. Indeed, we have

$$p(x,t,\theta) = p(\theta|t)p(t|x)p(x)$$

but since $t = T(x)$ is assumed to be a function of x, $p(t|x) = \delta\big(t - T(x)\big)$ and

$$p(x,t,\theta) = p\big(\theta|T(x)\big)\delta\big(t - T(x)\big)p(x),$$

where we have introduced the Dirac function (more precisely the Dirac in 0), so that after marginalizing t out we obtain :

$$p(x,\theta) = p\big(\theta|T(x)\big)p(x),$$

and so

$$p(x|\theta) = \frac{p\big(\theta|T(x)\big)}{p(\theta)}p(x),$$

which is of the desired form.

We now show conversely that the factorization of $p(x|\theta)$ implies the conditional independence statement. If

$$p(x|\theta) = f\big(x, T(x)\big)\, g\big(T(x), \theta\big)$$

then

$$p(t,x,\theta) = \delta\big(T(x) - t\big)\, f\big(x, T(x)\big)\, g\big(T(x), \theta\big)\, p(\theta) = \delta\big(T(x) - t\big)\, f(x,t)\, g(t,\theta)\, p(\theta),$$

where $p(\theta)$ is the density of the prior distribution over θ with respect to a reference measure on θ.

(To be rigorous, we should not write that this is a joint density for (t,x,θ) but that it is a derivative in the sense of generalized functions of a joint probability measure over the triple (t,x,θ) ; that is, we should call for example $\mu(t,x,\theta)$ the joint measure and instead of writing $p(x,t,\theta)$ we should write $d\mu(x,t,\theta)$. However, to avoid to write things that are unnecessarily abstract we will stick to these non-rigorous notations. The reasoning is however itself rigorous.)

As a consequence we have

$$p(t,\theta) = \int_x p(t,x,\theta) = \int_x \delta\big(T(x) - t\big)\, f(x,t)\, g(t,\theta)\, dx = h(t)\, g(t,\theta)\, p(\theta).$$

(Note that here $p(t,\theta)$ is again very rigorously a density with respect to a reference measure in $\mathbb{R}^2$). For t such that $p(t,\theta) \neq 0$, we have

$$p(x|t,\theta) = \frac{p(x,t,\theta)}{p(t,\theta)} = \delta\big(T(x) - t\big)\frac{f(x,t)\, g(t,\theta)\, p(\theta)}{h(t)\, g(t,\theta)\, p(\theta)} = \delta\big(T(x) - t\big)\frac{f(x,t)}{h(t)},$$

which shows that $p(x|t,\theta) = p(x|t)$. If $p(t,\theta) = 0$, we can define $p(x|t,\theta)$ the way we want (because on a set of probability zero, its value does not matter) and in particular we may set $p(x|t,\theta) = p(x|t)$.

Method of moments vs maximum likelihood estimation

1.

a)

$$p(x_1, ..., x_n|\theta) = \frac{1}{\theta^n}\prod_{i=1}^{n} 1_{x_i \in [0;\theta]}$$

$$
\begin{aligned}
\hat{\theta}_{MLE} &= \operatorname*{argmax}_{\theta} p(x_1, ..., x_n|\theta) \\
&= \operatorname*{argmax}_{\theta} \frac{1}{\theta^n} 1_{\{(\max_i x_i) \in [0;\theta]\}} \\
&= \operatorname*{argmin}_{\theta} \theta \quad \text{s.t.} \quad \theta \geq \max_{i \in \{1,...,n\}} x_i \\
&= \max_{i \in \{1,...,n\}} x_i
\end{aligned}
$$

b)

$$
\begin{aligned}
P(\hat{\theta}_{MLE} \leq x) &= P(\forall i \in \{1, \dots, n\},\ x_i \leq x) \\
&= \left(\frac{x}{\theta}\right)^n
\end{aligned}
$$

$$
p_{\hat{\theta}_{MLE}}(z) = n\frac{z^{n-1}}{\theta^n}
$$

Thus, $\frac{\hat{\theta}_{MLE}}{\theta}$ follows a Beta distribution whose parameters are $\alpha = n$ and $\beta = 1$.

c)

We can use the given formulas :

$$
E_\theta[\hat{\theta}_{MLE}] = \theta\frac{n}{n+1}
$$

$$
\mathrm{Var}_\theta(\hat{\theta}_{MLE}) = \theta^2\frac{n}{(n+1)^2(n+2)}
$$

d)

$$
\hat{\theta}_{MO} = \frac{2}{n}\sum_{i=1}^{n} x_i
$$

$$
E_\theta[\hat{\theta}_{MO}] = \theta
$$

$$
\mathrm{Var}_\theta(\hat{\theta}_{MO}) = \frac{4}{n^2} \cdot n \cdot \frac{\theta^2}{12} = \frac{\theta^2}{3n}
$$

e)

$$
\begin{aligned}
MSE &= E[(\theta - \hat{\theta})^2] \\
&= E[(\theta - E[\hat{\theta}])^2] + E[(\hat{\theta} - E[\hat{\theta}])^2] \\
&= \begin{cases} \frac{\theta^2}{3n} & \text{for MO} \\ \frac{\theta^2}{(n+1)^2} + \frac{\theta^2 n}{(n+1)^2(n+2)} = \frac{2\theta^2}{(n+1)(n+2)} & \text{for MLE} \end{cases}
\end{aligned}
$$

Computation of maximum likelihood estimators

1

a)

$$\begin{aligned}
\underset{\theta}{\operatorname{argmax}}\, P(x_1, ..., x_n|\theta) &= \underset{\theta}{\operatorname{argmax}}\ \theta^{\sum_{i=1}^n x_i} \cdot (1-\theta)^{n-\sum_{i=1}^n x_i} \\
&= \underset{\theta}{\operatorname{argmax}}\ \exp\Big(\sum_{i=1}^n x_i \log(\theta) + (n - \sum_{i=1}^n x_i)\log(1-\theta)\Big) \\
&= \underset{\theta}{\operatorname{argmax}}\ N\log(\theta) + (n-N)\log(1-\theta),
\end{aligned}$$

with $N := \sum_{i=1}^n$. Each term is continuous, strictly concave and their sum goes to $-\infty$ towards 0 and 1 so the MLE is unique.

b)

Let $l(x_1, ..., x_n|\theta) = N\log(\theta) + (n-N)\log(1-\theta)$,

$$\begin{aligned}
\frac{\partial l(x_1, ..., x_n|\theta)}{\partial \theta} &= \frac{N}{\theta} - \frac{n-N}{1-\theta} \\
&= \frac{N - n\theta}{\theta(1-\theta).}
\end{aligned}$$

Thus,

$$\hat{\theta}_{MLE} = \frac{N}{n}.$$

2.

a)

$$\begin{aligned}
P(Z_1, ..., Z_n|\,\{\pi_k\}_k) &= \pi_1^{\sum_{i=1}^n Z_{i,1}} ... \pi_K^{\sum_{i=1}^n Z_{i,K}} \\
&= \pi_1^{N_1} ... \pi_K^{N_K}
\end{aligned}$$

So $(N_1, ..., N_K)$ is a *sufficient statistic* for the sample because the likelihood depends on the data only through these quantities (see the exercise called *Sufficient statistic* for definition).

b)

$$\underset{\pi}{\operatorname{argmax}}\, \pi_1^{N_1} ... \pi_K^{N_K} = \underset{\pi}{\operatorname{argmax}}\, N_1\log(\pi_1) + ... N_K\log(\pi_K)$$

The MLE is solution of the **constrained** convex optimization problem

$$\underset{\pi \geq 0,\ \sum_{k=1}^K \pi_k = 1}{\operatorname{argmax}} \sum_{k=1}^K N_k \log(\pi_k)$$

c)

Let $L(\pi, \lambda) = \sum_{k,\ N_k > 0} N_k \log(\pi_k) - \lambda(\sum_k \pi_k - 1)$ be the associated Lagrangian.

$$\frac{\partial L}{\partial \pi_k} = \frac{N_k}{\hat{\pi}_k} - \lambda$$

So,

$$\begin{aligned}\hat{\pi}_k &= \alpha N_k, \text{ with } \alpha = \frac{1}{\lambda}\\ &= \frac{N_k}{n} \text{ since } \sum_{k=1}^{K} \hat{\pi}_k = 1.\end{aligned}$$

Note that we only introduced Lagrange multipliers for the equality constraint and not for the positivity constraints $\pi_k \geq 0$ because, the log-likelihood diverges to $-\infty$ on the edge of the domain which ensures that the constraints will be satisfied. We can indeed check that the estimators $\hat{\pi}_k$ are all non-negative.

3.

$$\begin{aligned}f(\mu + h) &= u^\top \mu + u^\top h\\ df_\mu(h) &= u^\top h\\ \nabla f(\mu) &= u\\ g(\mu + h) &= \mu^\top A \mu + \mu^\top A h + \mu^\top A^\top h + h^\top A h\\ dg_\mu(h) &= \mu^\top (A + A^\top) h\\ \nabla g(\mu) &= (A + A^\top)\mu\end{aligned}$$

a)

If Σ is fixed and positive definite,

$$\begin{aligned}\underset{\mu}{\operatorname{argmax}}\, p(x_1, ..., x_n|\mu) &= \underset{\mu}{\operatorname{argmax}} -\frac{n}{2}\log((2\pi)^d|\Sigma|) - \frac{1}{2}\sum_i (\mu^\top \Sigma^{-1}\mu - \mu^\top \Sigma^{-1} x_i - x_i^\top \Sigma^{-1}\mu + x_i^\top \Sigma^{-1} x_i)\\ &= \underset{\mu}{\operatorname{argmin}} \sum_i (\mu^\top \Sigma^{-1}\mu - \mu^\top \Sigma^{-1} x_i - x_i^\top \Sigma^{-1}\mu + x_i^\top \Sigma^{-1} x_i)\end{aligned}$$

Let's compute the gradient of the log-likelihood,

$$\nabla l(\mu) = \sum_i 2\Sigma^{-1}\mu - 2\Sigma^{-1} x_i$$

This gives us,

$$\hat{\mu}_{MLE} = \frac{1}{n}\sum_i x_i$$

b)

If μ is fixed and $\Lambda = \Sigma^{-1}$, since

$$\begin{aligned}\sum_i (x_i-\mu)^\top \Sigma^{-1}(x_i-\mu) &= \operatorname{tr}\left(\sum_i (x_i-\mu)^\top \Sigma^{-1}(x_i-\mu)\right)\\ &= \operatorname{tr}\left(\sum_i \Sigma^{-1}(x_i-\mu)(x_i-\mu)^\top\right)\end{aligned}$$

We have :

$$p(x_1,...,x_n|\Sigma) = \frac{1}{((2\pi)^d|\Sigma|)^{\frac{1}{2}}}\exp(-\frac{1}{2}\operatorname{tr}(\hat{\Sigma}\Lambda))$$

d)

$$\langle A, B+H\rangle_F - \langle A,B\rangle_F = \langle A,H\rangle_F$$

e)

$$f : A \to \log(|A|)$$

Let $H = (h_{i,j})_{i,j} \in R^{n^2}$

$$\begin{aligned}|I+H| &= \sum_{\sigma\in S_n}\operatorname{sgn}(\sigma)\prod_{i=1}^n(1_{\sigma(i)=i}+h_{i,\sigma(i)})\\ &= \prod_{i=1}^n(1+h_{i,i}) + \sum_{\sigma\in S_n\setminus I_n}\operatorname{sgn}(\sigma)\prod_{i=1}^n(1_{\sigma(i)=i}+h_{i,\sigma(i)})\\ &= 1+\sum_{i=1}^n h_{i,i} + O(|H|^2)\end{aligned}$$

$$d|.|_I(H) = tr(H)$$

A symmetric positive definite, H symmetric such that $A+H$ positive definite :

$$\begin{aligned}\log\det(A+H) &= \log(\det A.\det(I+A^{-1}H))\\ &= \log\det A + \log\det(I+A^{-1}H)\\ &= \log\det A + \operatorname{tr}(A^{-1}H)\end{aligned}$$

$$d|\ .|_A(H) = \operatorname{tr}(A^{-1}H)$$
$$\nabla f(A) = (A^{-1})^\top$$

f)

$$\log p(x_1,...,x_n|\Lambda) = -\frac{d}{2}\log((2\pi) + \frac{1}{2}\log|\Lambda| - \frac{1}{2}\operatorname{tr}(\hat{\Sigma}\Lambda)$$
$$\nabla f(\Lambda) = \frac{1}{2}\Lambda^{-1} - \frac{1}{2}\hat{\Sigma}$$
$$\Lambda_{MLE} = \hat{\Sigma}^{-1}$$
$$\Sigma_{MLE} = \hat{\Sigma}$$

Indeed, if $\hat{\theta} \in \operatorname{argmax}_\theta f(\theta)$ and $\hat{\theta} = \phi(\hat{\alpha})$ then $\hat{\alpha} \in \operatorname{argmax}_\alpha f(\phi(\alpha))$.

g)

$$p(x_1, ..., x_n|\mu, \Sigma) = \frac{1}{((2\pi)^d|\Sigma|)^{\frac{1}{2}}} \exp\Big(- \frac{1}{2}\sum_i (x_i - \mu)^\top \Sigma^{-1}(x_i - \mu)\Big)$$

If $\hat{\Sigma}$ is not invertible, let's write $\hat{\Sigma} = U \operatorname{diag}(\lambda_1, ..., \lambda_K, 0, ..., 0)\, U^\top$ with U an orthogonal matrix and set $\Lambda_\alpha = U \operatorname{diag}(\lambda_1, ..., \lambda_K, \alpha, ..., \alpha)^{-1} U^\top$.

$$\log p(x_1, ..., x_n|\Lambda_\alpha) = -\frac{d}{2}\log((2\pi) + \frac{1}{2}\log|\Lambda_\alpha| - \frac{1}{2}\operatorname{tr}(\hat{\Sigma}\Lambda_\alpha)$$

The second term goes to ∞ with α while the two others are constant, so the log-likelihood is unbounded. In practice, the maximum likelihood estimator is extended by continuity to these case ; the obtained estimator can also be though of as the maximum likelihood estimators for Gaussian densities on the subspace spanned by $\{x_1, \dots, x_n\}$.

Bayesian estimation

1.

a)

$$\begin{aligned} p(\pi|\alpha, n) &= p(n|\alpha, \pi) \cdot \frac{p(\pi|\alpha)}{p(n|\alpha)} \\ &\propto \prod_{k=1}^K \pi_k^{n_k} \frac{\Gamma(\alpha_1 + ... + \alpha_K)}{\prod_{k=1}^K \Gamma(\alpha_k)} \prod \pi_k^{\alpha_k - 1} \\ &\propto \prod_{k=1}^K \pi_k^{n_k + \alpha_k - 1} \end{aligned}$$

b)

We denote by $\triangle$ the canonical simplex $\triangle := \big\{u \in \mathbb{R}_+^K \mid \sum_{k=1}^K u_k = 1\big\}$. we then have

$$E[\pi_j|Z] = \int_\triangle \pi_j\, p(\pi|Z)\, d\pi$$

Let us consider a fixed value for $j \in \{1, \dots, K\}$ and define $\beta_k = \alpha_k + n_k$ for all k.

$$\begin{aligned} E[\pi_j|Z] &= \frac{\Gamma(\beta_1 + \ldots + \beta_K)}{\prod_{k=1}^K \Gamma(\beta_k)} \int_\triangle \pi_j \prod_{k=1}^K \pi_k^{\beta_k - 1} d\pi \\ &= \frac{\Gamma(\beta_1 + \ldots + \beta_K)}{\prod_{k=1}^K \Gamma(\beta_k)} \cdot \frac{\Gamma(\beta_j + 1)\prod_{k\neq j}\Gamma(\beta_k)}{\Gamma(\beta_1 + \ldots + \beta_K + 1)} \\ &= \frac{\Gamma(\beta_j + 1)}{\Gamma(\beta_j)} \cdot \frac{\Gamma(\beta_1 + \ldots + \beta_K)}{\Gamma(\beta_1 + \ldots + \beta_K + 1)} \\ &= \frac{\beta_j}{\beta_1 + \ldots + \beta_K} \\ &= \frac{\alpha_j + n_j}{\alpha_{\text{tot}} + n}, \end{aligned}$$

with $\alpha_{\text{tot}} = \alpha_1 + \ldots + \alpha_K$.

c)

$$E[\pi_1|n_1,n_2] = \frac{n_1+1}{n_1+n_2+2}$$

Without smoothing, if $x_1 = 0$ or $x_2 = 0$, which is common if π_1 is close to 0 or 1, the maximum likelihood estimator estimates that $p_i = 0$ even though $p_i > 0$. This is a major problem because then the probability of some non-zero event is assessed to be equal to 0, which makes all probabilistic reasonings fail.

2.

$$\mathcal{P} = \{p(x|\theta), \theta \in \Theta\}$$
$$\Pi = \{p_\alpha(\theta), \alpha \in \mathcal{A}\}$$

Π is a conjugate family of distributions for $\mathcal{P}$ if for all $p_\alpha \in \Pi$, there exists $p_{\alpha'} \in \Pi$ such that we can write $p_\alpha(\theta|x) = p_{\alpha'}(\theta)$.

$$p_{Bernoulli}(x|\theta) = \theta^x(1-\theta)^{1-x}$$

The family of beta distributions $p_{\alpha,\beta}(\theta) \propto \theta^{\alpha-1}(1-\theta)^{\beta-1}$ is a conjugate family of distributions for the family of Bernoulli distributions.

$$p_{Poisson}(x|\lambda) = \frac{\lambda^{\sum_{i=1}^n x_i} e^{-n\lambda}}{\prod_{i=1}^n (x_i!)}$$

The family of gamma distributions $p_{\alpha,\beta}(\lambda) \propto \lambda^{\alpha-1}e^{-\beta\lambda}$ is a conjugate family of distributions for the family of Poisson distributions.

$$p_{exp}(x|\mu) \propto \exp\Big(-\sum_i (\mu - x_i)^t \Sigma^{-1}(\mu - x_i) - (\mu - \mu_0)^t \Sigma_0^{-1}(\mu - \mu_0)\Big)$$

The family of gaussian distributions with the given covariance and unknown mean is a conjugate family of distributions for the family of gaussian random variables with fixed known covariance and unknown mean (cf. ex. 3).

3.

a)

As the product of two gaussian distributions, the a posteriori distribution is still a gaussian $N(\hat{\mu}_{PM}, \omega)$. On the one hand,

$$\exp\Big(-\frac{(\mu - \hat{\mu}_{PM})^2}{2\omega^2}\Big) = \exp\Big(-\frac{\mu^2}{2\omega^2} + \frac{2\mu\hat{\mu}_{PM}}{2\omega^2} - \frac{\hat{\mu}_{PM}^2}{2\omega^2}\Big)$$

On the other hand,

$$\exp\Big(\sum -\frac{(x_i-\mu)^2}{2\sigma^2} - \frac{(\mu_0 - \mu)^2}{2\tau^2}\Big)$$
$$= \exp\Big(-\mu^2.(\frac{n}{2\sigma^2} + \frac{1}{2\tau^2}) + 2\mu(\sum \frac{x_i}{2\sigma^2} + \frac{\mu_0}{2\tau^2}) - (\sum \frac{x_i^2}{2\sigma^2} + \frac{\mu_0^2}{2\tau^2})\Big)$$

By identification

$$\frac{1}{\omega^2} = \frac{n}{\sigma^2} + \frac{1}{\tau^2},$$

and we get the posterior mean :

$$\hat{\mu}_{PM} = \frac{\sum \frac{x_i}{\sigma^2} + \frac{\mu_0}{\tau^2}}{\frac{n}{\sigma^2} + \frac{1}{\tau^2}}.$$

b)

$$\hat{\mu}_{PM} = \frac{\sum x_i}{n} \cdot \lambda_n + \mu_0 \cdot (1 - \lambda_n) \quad \text{, with} \quad \lambda_n = \frac{\frac{n}{\sigma^2}}{\frac{n}{\sigma^2} + \frac{1}{\tau^2}}$$

d)

In this case and if $\mu_0 = 0$:

$$\underset{\mu}{\text{argmin}} - \ \log p(x_1, ..., x_n|\mu) + \frac{\lambda}{2}\mu^2 = \underset{\mu}{\text{argmin}} - \ \log p(x_1, ..., x_n|\mu) \cdot p(\mu) \text{ for } \lambda = \frac{1}{\tau^2}$$

Thus the MAP estimator can be viewed as a minimizer of the log-likelihood with some ridge regularization.

e)

$$\mu_{MAP} = \mu_{PM}$$

This property doesn't hold for a Bernoulli distribution with a Beta prior.

f)

$$E[\log\ p(X'|\nu, \sigma)] = E[-\frac{(X' - \nu)^2}{2\sigma^2}]$$

g)

$$\begin{aligned} R(\nu) &= E[(\nu - X')^2] \\ &= E[(\nu - E[X'] + E[X'] - X')^2] \\ &= E[(\nu - E[X'])^2] + E[(E[X'] - X')^2] + 2E[(\nu - E[X'])(E[X'] - X')] \\ &= (\mu - \nu^2) + \text{Var}(X') \end{aligned}$$

h)

$$\begin{aligned} E_{D_n}[\mathcal{E}(\hat{\mu})] &= E_{D_n}[R(\hat{\mu}) - R(\mu)] \\ &= E_{D_n}[(\hat{\mu} - \mu)^2] \\ &= E_{D_n}[(\hat{\mu} - E_{D_n}[\hat{\mu}])^2] + (E_{D_n}[\hat{\mu}] - \mu)^2 \end{aligned}$$

i)

$$\begin{aligned} E_{D_n}[\mathcal{E}(\mu_{MLE})] &= E_{D_n}[(\mu_{MLE} - E_{D_n}[\mu_{MLE}])^2] + 0 \\ &= \frac{\sigma^2}{n} \end{aligned}$$

$$
\begin{aligned}
E_{D_n}[\mathcal{E}(\mu_{MAP})] &= E_{D_n}[(\mu_{MAP} - E_{D_n}[\mu_{MAP}])^2] + (E_{D_n}[\mu_{MAP}] - \mu)^2 \\
&= \frac{\frac{n}{\sigma^2}}{(\frac{n}{\sigma^2} + \frac{1}{\tau^2})^2} + \frac{\frac{(\mu-\mu_0)^2}{\tau^4}}{(\frac{n}{\sigma^2} + \frac{1}{\tau^2})^2}
\end{aligned}
$$

j)

$$\mathcal{R}_\pi(MLE) = \frac{\sigma^2}{n}$$

$$\mathcal{R}_\pi(PM) = \frac{\frac{\sigma^2}{n}}{(1 + \frac{\sigma^2}{n\tau^2})}$$

l)

$$E_{\mu\sim\pi,D_n}[(\hat{\mu} - \mu)^2] = E_{D_n}[E_{\mu\sim\pi}[(\hat{\mu} - \mu)^2|D_n]]$$

The inner quantity is minimized for every possible D_n by using the posterior mean.

Bregman divergence

1.

$$D_F(p, q) = \langle p, p\rangle - \langle q, q\rangle - 2\langle q, p\rangle + 2\langle q, q\rangle = \langle p - q, p - q\rangle$$

2.

$$(\nabla H(q))_i = -\log q_i - 1$$

$$
\begin{aligned}
D_H(p, q) &= \sum p_i \log p_i - \sum q_i \log q_i - \sum(\log q_i + 1)(p_i - q_i) \\
&= \sum p_i(\log p_i - \log q_i) \\
&= KL(p, q)
\end{aligned}
$$

3.

We assume the loss is differentiable.

$$F(\mu) = E_X[l(\mu, X)] - E_X[l(\mu^*, X)]$$

1 Ridge regression and PCA

1.

a)

$$X = USV^\top,\ X^- = VS^-U^\top,\ S = \mathrm{diag}(s_i)$$

$$\begin{aligned} XX^-X &= USV^\top VS^-U^\top USV \\ &= U\,\mathrm{diag}(1_{s_i \neq 0})SV^\top)V^\top \\ &= USV^\top = X \end{aligned}$$

b)

$$\begin{aligned} (X^\top X)^- &= (VS^2V^\top)^- \\ &= V(S^-)^2V^\top \end{aligned}$$

$$X^-(X^-)^\top = V(S^-)^2V^\top$$

$$\begin{aligned} (X^\top X)^-X^\top &= X^-(X^-)^\top X^\top \\ &= VS^-S^-SU^\top \\ &= VS^-U^\top \end{aligned}$$

c)

$$X^\top XX^- = VSU^\top USV^\top VS^-U^\top = VSU^\top$$

d)

First suppose $X = S$. Let w be a solution to the normal equation.

$$s_{i,i}^2 w_i = s_i \Rightarrow w_i = s_i^- y \text{ if } s_i \neq 0$$

$$(S^\top y)_i = 0 \text{ if } s_i = 0$$

So it is true if $X = S$.
For any X, let $\tilde{w} = V^\top w$, $\tilde{y} = U^\top y$.
w is a solution to the normal equation iff $\tilde{w}$ is a solution of $S^2\tilde{w} = S\tilde{y}$

2.

$X = USV^\top$, U and V are square matrices. The columns of U and the columns of V are called the left-singular vectors and right-singular vectors of X.

a)

Since U and V are orthogonal matrices, we have for all $w \in R^n$: $\|Xw\| = \|USV^\top w\| = \|SV^\top w\|$

$$\underset{w \in R^n, \|w\|=1}{\operatorname{argmax}} \|Xw\| = V^\top \underset{w \in R^n, \|w\|=1}{\operatorname{argmax}} \|Sw\|$$

b)

$$\hat{\Sigma} = \frac{1}{n} V S^2 V^\top$$

c)

$$\begin{aligned}
\sum_i (c_{j,i} - \bar{c}_j)^2 &= \sum_i ((Xv_j)_i - \sum_k (Xv_j)_k)^2 \\
&= \sum (Xv_j)^2 \\
&= s_j^2 \cdot 1
\end{aligned}$$

d)

$$XX^\top X v_j = USV^\top V^\top SUUSV^\top v_j = US^3 V v_j = s_{j,j}^2 X v_j$$

3.

a)

$$\begin{aligned}
\sum_i \|y^{(i)} - (c_{1:k}^{(i)})^\top w\|^2 &= \sum_i \|y^{(i)} - (x_i^\top v_j)_{j=1:k} w\|^2 \\
&= \|y - (XV_{1:k})w\|^2 \\
&= \|y - U_k S_k w\|^2
\end{aligned}$$

with $U_k = (u_1, ..., u_k)$, $S_k = \operatorname{diag}(s_1, ..., s_k)$. So,

$$\tilde{w} = ((U_k S_k)^\top U_k S_k)^{-1} (U_k S_k)^\top y = S_k^{-1} U_k^\top y.$$

b)

$$\tilde{w}^\top (\langle x - \bar{x}_0, v_1 \rangle, ..., \langle x - \bar{x}_0, v_k \rangle) = \sum_j \hat{w}_j \langle x - \bar{x}_0, v_j \rangle = \langle x - \bar{x}_0, \sum_j \frac{1}{s_j} \langle u_j, y \rangle v_j \rangle$$

c)

$$
\begin{aligned}
w_R &= (X^\top X + \lambda I_p)^{-1} X^\top y \\
&= (V(S^2 + \lambda I_p)V^\top)^{-1} V S U^\top y \\
&= V \operatorname{diag}(\frac{s_i}{s_i^2 + \lambda}) U^\top y \\
&= \sum_{j=1}^{m} \frac{s_j}{s_j^2 + \lambda} \langle u_j, y \rangle v_j
\end{aligned}
$$

d)

The coefficients for $j > k$ will vanish.

e)

$$
\begin{aligned}
\langle X^\top y, v_j \rangle &= \langle V S U^\top y, v_j \rangle \\
&= (S U^\top Y)_J \\
&= s_j \langle u_j, y \rangle
\end{aligned}
$$

f)

Andrei Tikhonov and Karl Pearson

Area under the curve and Mann-Whitney U statistic

a)

Let C_0 be the set of elements that belong to class 0, C_1 be the set of elements that belong to class 1.

$$
\begin{aligned}
rTP(b) &= \frac{P(s(x) > b, x \in C_1)}{P(x \in C_1)} \\
&= \frac{P(s(x) > b).P(x \in C_1)}{P(x \in C_1)}, \text{ since } s(x) \text{ doesn't depend on } x. \\
&= 1 - F(b)
\end{aligned}
$$

$$
\begin{aligned}
rFP(b) &= \frac{P(s(x) > b, x \in C_0)}{P(x \in C_0)} \\
&= \frac{P(s(x) > b).P(x \in C_0)}{P(x \in C_0)} \\
&= 1 - F(b)
\end{aligned}
$$

Hence,

$$
AUC = \frac{1}{2}
$$

b)

WLOG, suppose $s(x_1) < ... < s(x_n)$, $s(y_1) < ... < s(y_m)$. On the one hand,

$$
\begin{aligned}
U &= \sum_{i=1}^{n} |\{s(z) | z \in D_N \cup D_P \text{ and } s(z) \leq s(x_i)\}| - \frac{n(n+1)}{2} \\
&= \sum_{i=1}^{n} |\{s(z) | z \in D_N \text{ and } s(z) \leq s(x_i)\}| + \sum_{i=1}^{n} \underbrace{|\{s(z) | z \in D_P \text{ and } s(z) \leq s(x_i)\}|}_{=i} - \frac{n(n+1)}{2} \\
&= \sum_{i=1}^{n} |\{s(z) | z \in D_N \text{ and } s(z) \leq s(x_i)\}|
\end{aligned}
$$

On the other hand,

$$rTP(s(x_i)) = 1 - \frac{i}{n}$$

and

$$
\begin{aligned}
rFP(s(x_i)) &= \frac{|\{s(z) | z \in\in D_N \text{ and } s(z) > s(x_i)\}|}{m} \\
&= 1 - \frac{|\{s(z) | z \in D_N \text{ and } s(z) \leq s(x_i)\}|}{m}
\end{aligned}
$$

so

$$
\begin{aligned}
\frac{U}{m.n} &= \frac{1}{n} \sum_{i=1}^{n} \big(1 - rFP(s(x_i))\big) \\
&= \sum_{i=1}^{n} \big(1 - rFP(s(x_i))\big).(rTP(s(x_i)) - rTP(s(x_{i-1}))) \text{ with } x_0 = -\infty \\
&= AUC
\end{aligned}
$$