2A

Machine Learning II
ID: 5684926 Tristan Scheidemann

2A-1.

We show that the median minimizes $E[|y(x)-t|]$:

$$
\begin{aligned}
& \iint L(y(x),t)\cdot p(x,t)dx\,dt, \\
= \quad & \underbrace{\iint L(y(x),t)\cdot p(x,t)dt\,dx}_{Fubini/Tonelli}, \\
= \quad & \int p(x)\int L(y(x),t)\cdot p(t|x)dt\,dx.
\end{aligned}
$$

With data x constant, we minimize with respect to $y(x)$. Caution: $y(x)$ is a single number $c\in\mathbb{R}$ and not a function because x is fixed, so scalar derivatives are used.

$$
\begin{aligned}
\frac{\partial}{\partial y(x)}\int_a^b L(y(x),t)\cdot p(t|x)dt \quad &= \underbrace{L(y(x),b)\cdot\frac{d}{dy(x)}b}_{=0}-\underbrace{L(y(x),a)\cdot\frac{d}{dy(x)}a}_{=0}+\int\frac{\partial}{\partial y(x)}L(y(x),t)\cdot p(t|x)dt, \\
&= \int\frac{\partial}{\partial y(x)}L(y(x),t)\cdot p(t|x)dt, \\
&= \int\frac{\partial}{\partial y(x)}|y(x)-t|\cdot p(t|x)dt, \\
&= \int \underbrace{\frac{|y(x)-t|}{y(x)-t}}_{just\ a\ number\ in\ \{-1,1\}}\cdot p(t|x)dt, \\
&= \underbrace{\int_a^{y(x)}p(t|x)dt-\int_{y(x)}^b p(t|x)dt}_{monotonicity\ of\ \frac{\partial}{\partial y(x)}L(y(x),t)}, \\
&= 0.
\end{aligned}
$$

Above immediately establishes the relationship

$$
\int_a^{y(x)}p(t|x)dt=\int_{y(x)}^b p(t|x)dt,
$$

which only holds if $y(x)$ is the conditional median of $p(t|x)$.

2A-2.

Let $\mathbf{x}$ be the data and $\mathbf{p}$ the wanted parameter. We solve the equivalent proportional problem:

$$\begin{aligned}
p(\mathbf{p}|\mathbf{x}) &\propto p(\mathbf{x}|\mathbf{p}) \cdot p(\mathbf{p}) \\
&= \prod_{i=1}^{K} p_{\mathrm{i}}^{\#\{x_n=k:x_n\in\mathbf{x}\}} \cdot \frac{\Gamma\left(\sum_{i=1}^{K}\alpha_i\right)}{\prod_{i=1}^{K}\Gamma(\alpha_i)} \prod_{i=1}^{K} p_{\mathrm{i}}^{\alpha_i-1}, \\
&\propto \prod_{i=1}^{K} p_{\mathrm{i}}^{\#\{x_n=k:x_n\in\mathbf{x}\}} \cdot \prod_{i=1}^{K} p_{\mathrm{i}}^{\alpha_i-1}, \\
&= \prod_{i=1}^{K} p_{\mathrm{i}}^{\alpha_i-1+\#\{x_n=k:x_n\in\mathbf{x}\}}, \\
&\propto Dir\big(\boldsymbol{p},(\beta_1,\dots,\beta_n)\big), \qquad with\ \beta_i = \alpha_i - 1 + \#\{x_n = k : x_n \in \mathbf{x}\}.
\end{aligned}$$

The α_i's are called pseudo counts because they simulate frequencies of classes before anything has been observed. The more "real" events $\#\{x_n = k : x_n \in \mathbf{x}\}$ happen, the less relevant pseudo counts become.

2A-3.

To empathize the constant nature of x_b, we substitute $x_b = s$ and $Var[X_i] = \sigma_i^2$.
Bayes' theorem gives us:

$$f_{X_a|s}(x_a) = \frac{f(x_a, s)}{f(s)}.$$

Auxiliary calculation

By the nature of $X \sim N(\boldsymbol{\mu}, \Sigma)$, each entry x_i is normally distributed $X_i \sim N(\mu_i, \Sigma_{i,i})$, which leads to

$$f(s) = \frac{1}{\sqrt{2\pi\sigma_b^2}} \cdot e^{-\frac{1}{2\sigma_b^2}(s-\mu_b)^2}.$$

We also need the result of the following calculation later:

Because f_X exists, Σ is always invertible due to its positive definiteness:

$$\Sigma^{-1} = \frac{1}{\sigma_a^2\sigma_b^2 - Cov(X_a, X_b)^2}\begin{pmatrix} \sigma_b^2 & -Cov(X_a, X_b) \\ -Cov(X_a, X_b) & \sigma_a^2 \end{pmatrix}.$$

Thus:

$$\det|\Sigma| = \sigma_a^2\sigma_b^2 - Cov(X_1, X_2)^2 \neq 0.$$

Additionally, Σ^{-1} is positive definite as well, so $z^T\Sigma z$ is always positive and:

$$\begin{aligned}
&([x_a - \mu_a \quad s - \mu_b])\Sigma^{-1}\left(\begin{bmatrix} x_a - \mu_a \\ s - \mu_b \end{bmatrix}\right) \\
= \; &\frac{(x_a - \mu_a)^2 \cdot \sigma_b^2 - 2(x_a - \mu_a)(s - \mu_b)Cov(X_a, X_b) + (s - \mu_b)^2\sigma_a^2 \cdot}{\sigma_a^2\sigma_b^2 - Cov(X_a, X_b)^2}
\end{aligned}$$

Explicit calculation for $D = 2$:

$$\frac{f(x_a,s)}{f(s)} = \frac{\left(2\pi^{-\frac{D}{2}}\right)\cdot\det|\Sigma|^{-\frac{1}{2}}e^{-\frac{1}{2}(x-\mu)^T\Sigma^{-1}(x-\mu)}}{\frac{1}{\sqrt{2\pi\sigma_b^2}}\cdot e^{-\frac{1}{2\sigma_b^2}(s-\mu_b)^2}},$$

$$= \frac{\sqrt{2\pi\sigma_b^2}\,\det|\Sigma|^{-\frac{1}{2}}e^{-\frac{1}{2}\frac{(x_a-\mu_a)^2\cdot\sigma_b^2-2(x_a-\mu_a)(s-\mu_b)Cov(X_a,X_b)+(s-\mu_b)^2\sigma_a^2}{\sigma_a^2\sigma_b^2-Co\ (X_a,X_b)^2}}}{2\pi\cdot e^{-\frac{1}{2\sigma_b^2}(s-\mu_b)^2}},$$

$$= \frac{\sqrt{\sigma_b^2}}{\sqrt{2\pi}\sqrt{\sigma_a^2\sigma_b^2-Cov(X_1,X_2)^2}}\, e^{-\frac{1}{2}\frac{(x_a-\mu_a)^2\cdot\sigma_b^2-2(x_a-\mu_a)(s-\mu_b)Cov(X_a,X_b)+(s-\mu_b)^2\sigma_a^2}{\sigma_a^2\sigma_b^2-Cov(X_a,X_b)^2}}\, e^{\frac{1}{2\sigma_b^2}(s-\mu_b)^2},$$

$$= \frac{\sqrt{\sigma_b^2}\,e^{-\frac{1}{2}\frac{\sigma_b^2\left((x_a-\mu_a)^2\cdot\sigma_b^2-2(x_a-\mu_a)(s-\mu_b)Cov(X_a,X_b)+(s-\mu_b)^2\sigma_a^2\right)+\left(-\sigma_a^2\sigma_b^2+Cov(X_a,X_b)^2\right)(s-\mu_b)^2}{\sigma_b^2\left(\sigma_a^2\sigma_b^2-Cov(X_a,X_b)^2\right)}}}{\sqrt{2\pi}\sqrt{\sigma_a^2\sigma_b^2-Cov(X_1,X_2)^2}},$$

$$= \frac{e^{-\frac{1}{2}\frac{\sigma_b^2\left((x_a-\mu_a)^2\cdot\sigma_b^2-2(x_a-\mu_a)(s-\mu_b)Cov(X_a,X_b)+(s-\mu_b)^2\sigma_a^2\right)+\left(-\sigma_a^2\sigma_b^2+Co\ (X_a,X_b)^2\right)(s-\mu_b)^2}{\sigma_b^2\left(\sigma_a^2\sigma_b^2-Cov(X_a,X_b)^2\right)}}}{\sqrt{2\pi}\sqrt{\underbrace{\frac{\sigma_a^2\sigma_b^2-Cov(X_1,X_2)^2}{\sigma_b^2}}_{\sigma^2}}},$$

–

$$= \frac{e^{-\frac{1}{2}\frac{\sigma_b^2\left((x_a-\mu_a)^2\cdot\sigma_b^2-2(x_a-\mu_a)(s-\mu_b)Cov(X_a,X_b)+(s-\mu_b)^2\sigma_a^2\right)+\left(-\sigma_a^2\sigma_b^2+Cov(X_a,X_b)^2\right)(s-\mu_b)^2}{\sigma_b^2\left(\sigma_a^2\sigma_b^2-Cov(X_a,X_b)^2\right)}}}{\sqrt{2\pi\sigma^2}},$$

$$= \frac{e^{-\frac{1}{2}\frac{\sigma_b^2\left((x_a-\mu_a)^2\cdot\sigma_b^2-2(x_a-\mu_a)(s-\mu_b)Cov(X_a,X_b)+(s-\mu_b)^2\sigma_a^2\right)-\sigma_a^2\sigma_b^2(s-\mu_b)^2+Cov(X_a,X_b)^2(s-\mu_b)^2}{\sigma_b^2\left(\sigma_a^2\sigma_b^2-Cov(X_a,X_b)^2\right)}}}{\sqrt{2\pi\sigma^2}},$$

$$= \frac{e^{-\frac{1}{2}\frac{\sigma_b^2\left((x_a-\mu_a)^2\cdot\sigma_b^2-2(x_a-\mu_a)(s-\mu_b)Cov(X_a,X_b)+\cancel{(s-\mu_b)^2\sigma_a^2-\sigma_a^2(s-\mu_b)^2}+\frac{Cov(X_a,X_b)^2}{\sigma_b^2}(s-\mu_b)^2\right)}{\sigma_b^2\left(\sigma_a^2\sigma_b^2-Cov(X_a,X_b)^2\right)}}}{\sqrt{2\pi\sigma^2}},$$

$$= \frac{e^{-\frac{1}{2}\frac{\left((x_a-\mu_a)^2-\frac{2(x_a-\mu_a)(s-\mu_b)Cov(X_a,X_b)}{\sigma_b^2}+\frac{Cov(X_a,X_b)^2}{\sigma_b^4}(s-\mu_b)^2\right)}{\sigma^2}}}{\sqrt{2\pi\sigma^2}},$$

$$= \frac{e^{-\frac{1}{2}\frac{\left((x_a-\mu_a)-\frac{Cov(X_a,X_b)}{\sigma_b^2}(s-\mu_b)\right)^2}{\sigma^2}}}{\sqrt{2\pi\sigma^2}},$$

$$= \frac{e^{-\frac{1}{2}\frac{(x_a-\mu)^2}{\sigma^2}}}{\sqrt{2\pi\sigma^2}}.$$

The conditional distribution $f_{X_a|s}$ is univariately normally distributed with

$$\begin{aligned}\mu &= \mu_a - \frac{Cov(X_a,X_b)}{\sigma_b^2}(s-\mu_b)\\ &= \mu_a + \frac{\sigma_a}{\sigma_b}\kappa(s-\mu_b), \qquad with\ \kappa = \frac{Cov(X_a,X_b)'}{\sigma_a\sigma_b}\end{aligned}$$

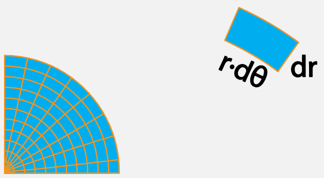
$$\sigma^2 = \frac{\sigma_a^2\sigma_b^2 - Cov(X_1, X_2)^2}{\sigma_b^2}.$$

2A-4.

Before we start we need to know the following prerequisites:

Auxiliary calculation

The area of a single infinitesimal d-dimensional piece of $f(r, \boldsymbol{\theta})$ is $r^{d-1} d\theta_1 \cdot \ldots \cdot d\theta_{d-1} \cdot dr$. This is trivially an d-dimensional extension of the two-dimensional case shown below:



.

Additionally, to convert a function $f(r, \boldsymbol{\theta})$ from hyperspherical coordinates into cartesian coordinates $f(\mathbf{x})$, we use the following trigonometric conversion:

$$\mathbf{x} = \begin{pmatrix} x_1 \\ x_2 \\ \ldots \\ x_{d-1} \\ x_d \end{pmatrix} = \begin{pmatrix} r\cos\theta_1 \\ r\sin\theta_1\cos\theta_2 \\ \ldots \\ r\sin\theta_1\sin\theta_2\sin\theta_3\ldots\sin\theta_{d-3}\sin\theta_{d-2}\cos\theta_{d-1} \\ r\sin\theta_1\sin\theta_2\sin\theta_3\ldots\sin\theta_{d-3}\sin\theta_{d-2}\sin\theta_{d-1} \end{pmatrix},$$

i.e.

$$f(\mathbf{x}) = f(r\cos\theta_1, r\sin\theta_1\cos\theta_2, \ldots, r\sin\theta_1\sin\theta_2\sin\theta_3\ldots\sin\theta_{d-3}\sin\theta_{d-2}\sin\theta_{d-1}).$$

Lastly, let

$$K(\boldsymbol{\theta}) = \cos\theta_1{}^2 + (\sin\theta_1\cos\theta_2)^2 + \cdots \\ + (\sin\theta_1\sin\theta_2\sin\theta_3\ldots\sin\theta_{d-3}\sin\theta_{d-2}\sin\theta_{d-1})^2.$$

Note: $\boldsymbol{\theta}$ describe points on the unit d-sphere, so it is no surprise that $\|K(\boldsymbol{\theta})\|^2 = 1$ for all $\boldsymbol{\theta}$, because the radius of the unit sphere is 1.

The centered sphere is described by $B_0(r) \coloneqq \{x_1^2 + \cdots + x_d^2 \leq r^2 : x_i \in \mathbb{R}\}$.

Armed with this knowledge, $P\big(B_0(r)\big)$ becomes:

$$\int_0^r \int_0^{2\pi} \ldots \int_0^{\pi} \underbrace{(2\pi)^{-\frac{D}{2}} e^{-\frac{1}{2}\|\mathbf{x}\|^2}}_{pdf\ normal\ dist.} \underbrace{r^{d-1} d\theta_1 \ldots d\theta_{d-1} dr}_{infinitismal\ area} = \int_0^r \int_0^{2\pi} \ldots \int_0^{\pi} (2\pi)^{-\frac{D}{2}} e^{-\frac{1}{2}r^2 \underbrace{K(\boldsymbol{\theta})}_{=1}} r^{d-1} d\theta_1 \ldots d\theta_{d-1} dr,$$

$$= \int_0^r r^{d-1} (2\pi)^{-\frac{D}{2}} e^{-\frac{1}{2}r^2} \underbrace{\int_0^{2\pi} \ldots \int_0^{\pi} d\theta_1 \ldots d\theta_{d-1}}_{Surface\ Area\ unit\ n-sphere\ S_D} dr,$$

$$= \int_0^r S_D r^{d-1} (2\pi)^{-\frac{D}{2}} e^{-\frac{1}{2}r^2} dr.$$

Ergo $p(r)dr = S_D r^{d-1}(2\pi)^{-\frac{D}{2}} e^{-\frac{1}{2}r^2}$.

Now we are looking for the maximum density $\max_r p(r)$:

$$\frac{d}{dr}\left[\log r^{d-1} + \log e^{-\frac{1}{2}r^2}\right] = \frac{(d-1)r^{d-2}}{r^{d-1}} - r = 0.$$

$\Leftrightarrow (d-1) = r^2$.

Because radii are non-negative, we have a maximum at $\sqrt{d-1}$.

Now if we set $\|\mathbf{x}\| = \sqrt{d-1}$, we get

$$\frac{p(\mathbf{x})}{p(0)} = \frac{(2\pi)^{-\frac{D}{2}} e^{-\frac{1}{2}\|\mathbf{x}\|^2}}{(2\pi)^{-\frac{D}{2}} e^{-\frac{1}{2}\|\mathbf{0}\|^2}} = \frac{e^{-\frac{1}{2}(d-1)}}{e^{-\frac{1}{2}}} = e^{-\frac{d}{2}}.$$

2A-5.