II - Decision theory and Bayesian inference

We saw in the previous chapter that Bayesian inference is based on the following two principles:

- We can express our ignorance/information about the unknown parameter $\theta \in \Theta$ by a probability distribution $\pi(\theta)$ on Θ.
- We can use the Bayes rule and the likelihood $f(x|\theta)$ of an observation $x \in \mathcal{X}$ to update our prior knowledge about θ.

The first output of Bayesian inference is therefore the posterior distribution $\pi(\theta|x) \propto f(x|\theta)\pi(\theta)$.

However, we often want to derive an estimator of the unknown parameter θ; that is, a point $\widehat{\theta}$ in the parameter space that approximates θ in some sense.

The "most natural" estimators we can derive from $\pi(\theta|x)$ are:

- the posterior mean;
- the posterior median;
- the posterior mode (also called MAP for maximum a posteriori).

The goal of this chapter is to justify (or not!) these estimators from a decision theoretic perceptive.

Decision theory: General framework

Let $\mathcal{D}$ be the set of all possible decisions.

Definition 2.1 *A loss function is any function* $L : \Theta \times \mathcal{D} \to [0, +\infty)$.

Definition 2.2 *A decision rule is any mapping* $\delta : \mathcal{X} \to \mathcal{D}$.

For $\theta \in \Theta$ and $x \in \mathcal{X}$, the quantity $L(\theta, \delta(x))$ therefore gives the cost induced by the decision rule δ when we observe x.

The question of interest is then the following:

Given a loss function L, what is the "optimal" decision rule δ?

In this chapter we mainly focus on the scenario $\mathcal{D} = \Theta$. In this case, δ is an estimator of the unknown parameter θ and the Bayesian answer to the above question for different choices of loss L leads to the derivation of different Bayesian estimators of θ.

Remark: Often we address the reverse question; that is, for which (if any) loss function L is the decision rule δ optimal? This is helpful to understand in what sense δ is a good decision rule.

The frequentist approach

The frequentist approach considers the frequentist risk

$$R(\theta, \delta) = \int_{\mathcal{X}} L(\theta, \delta(x)) f(x|\theta) \mathrm{d}x.$$

Because the frequentist risk is a function of θ there exists in general no decision rule δ such that, for every decision rule $\delta' \neq \delta$, we have

$$R(\theta, \delta) \leq R(\theta, \delta'), \quad \forall \theta \in \Theta.$$

Consequently, the frequentist risk alone is not sufficient to select a particular decision rule and other criteria are needed to choose δ.

For instance, in the frequentist approach we can

1. Select the decision rule δ which minimizes the frequentist risk on a given restricted set of decision rules (e.g. the set of unbiased and linear decision rules).
2. Select a decision rule δ which is minimax; that is, such that for every decision rule $\delta' \neq \delta$ we have

$$\max_{\theta \in \Theta} R(\theta, \delta) \leq \max_{\theta \in \Theta} R(\theta, \delta').$$

3. Select a decision rule δ which is admissible.

Definition 2.3 *A decision rule δ is admissible if there exists no decision rule δ' such that*

$$R(\theta, \delta') \leq R(\theta, \delta), \quad \forall \theta \in \Theta$$

with the above inequality being strict for at least one $\theta \in \Theta$.

Admissibility and Stein's result

Admissibility seems to be a weak requirement for an estimator since e.g. the estimator δ_{θ^*} such that $\delta_{\theta^*}(x) = \theta^*$ for all $x \in \mathcal{X}$ and for a $\theta^* \in \Theta$ is in general admissible. (This is for instance the case when $L(\theta, \theta') = 0$ if and only $\theta = \theta'$ while $f(x|\theta) > 0$ for all $(x, \theta) \in \mathcal{X} \times \Theta$.)

However, Stein (1956) shows the following surprising result.

Theorem 2.1 *Let $\Theta = \mathbb{R}^d$, $f(\cdot\,|\theta)$ be the probability density function of the $\mathcal{N}_d(\theta, I_d)$ distribution and $L : \Theta \times \Theta \to [0, +\infty)$ be the quadratic loss function. Then, when $d \geq 3$, the maximum likelihood estimator (MLE) defined by $\delta_0(x) = x$, $x \in \mathbb{R}^d$, is not admissible.*

Proof: See Appendix 1.

Remark: For $d \in \{1, 2\}$, the estimator δ_0 is admissible and therefore a (surprising) corollary of Theorem 2.1 is that the aggregation of several admissible estimators of unrelated quantities is not necessarily admissible.

Remark: Theorem 2.1 has been extended to alternative loss functions and to non-Gaussian models.

The main message of this theorem is that there are no general guarantees that the MLE is admissible.

However, it can be shown that, in the set-up of Theorem 2.1, δ_0 is minimax for any $d \geq 1$ and is asymptotically efficient, and therefore inadmissible estimators are not necessarily bad estimators.

To sum-up: Admissible estimators are not necessarily good estimators and inadmissible estimators are not necessarily bad estimators.

The Bayesian approach

The Bayesian approach considers the posterior expected loss

$$\rho(\pi, d|x) = \int_\Theta L(\theta, d)\pi(\theta|x)\mathrm{d}\theta, \quad d \in \mathcal{D}.$$

Hence, while the frequentist approach integrates on $\mathcal{X}$, the Bayesian approach integrates on Θ. Say differently, the Bayesian approach uses the posterior distribution to integrate out the unknown quantity θ while the frequentist approach uses the likelihood to integrate out the known quantity x.

Using the posterior expected loss, we define $\delta^\pi : \mathcal{X} \to \mathcal{D}$ an estimator such that

$$\delta^\pi(x) \in \operatorname*{argmin}_{d\in\mathcal{D}} \rho(\pi, d|x), \quad \forall x \in \mathcal{X}. \tag{1}$$

Lastly, the integrated risk of δ is given by

$$r(\pi, \delta) = \int_\Theta R(\theta, \delta)\pi(\theta)\mathrm{d}\theta.$$

Definition 2.4 *A Bayes estimator associated with a prior distribution π and a loss function L is any estimator δ^π (defined in (1)) such that $r(\pi, \delta^\pi) < +\infty$. The value of $r(\pi) := r(\pi, \delta^\pi)$ is called the Bayes risk.*

Two important properties of Bayes estimators

The following result shows that if δ^π is a Bayes estimator then δ^π is a minimizer of the integrated risk; that is

$$r(\pi) \leq r(\pi, \delta), \quad \forall \delta,$$

and therefore the Bayesian risk is the minimum possible integrated risk.

Theorem 2.2 *An estimator minimising the integrated risk $r(\pi, \delta)$ can be obtained by selecting, for every $x \in \mathcal{X}$, a value $\delta(x)$ belonging to* $\operatorname{argmin}_{d \in \mathcal{D}} \rho(\pi, d|x)$.

Proof: This is a direct consequence of the fact that

$$r(\pi, \delta) = \int_\Theta R(\theta, \delta)\pi(\theta)\mathrm{d}\theta = \int_{\mathcal{X}} \rho(\pi, \delta(x)|x)m(x)\mathrm{d}x.$$

Bayes estimators are attractive beyond Bayesian statisticians because they have good frequentist properties. For instance, the next result shows that, under mild conditions, the Bayes estimator is admissible

Theorem 2.3 *If the Bayes estimator is the unique minimizer of the integrated risk then it is admissible.*

Proof: Done in class.

Other good frequentist properties of Bayes estimators are related to their asymptotic behaviours (as the number of observations goes to infinity); see Chapter 6.

The quadratic loss function

Assuming $\mathcal{D} = \Theta$, the quadratic loss function is defined by

$$L(\theta, d) = \|\theta - d\|^2, \quad (\theta, d) \in \Theta^2$$

where $\|\cdot\|$ stands for the Euclidean norm on $\mathbb{R}^d$.

Theorem 2.4 *Assume that $\mathbb{E}_\pi[\theta^T\theta|x] < +\infty$ for all $x \in \mathcal{X}$ and that Θ is a convex set. Then, the estimator δ^π associated with the quadratic loss function is unique and is the posterior expectation,*

$$\delta^\pi(x) = \mathbb{E}_\pi[\theta|x] = \frac{\int_\Theta \theta\pi(\theta)f(x|\theta)\,\mathrm{d}\theta}{\int_\Theta \pi(\theta)f(x|\theta)\,\mathrm{d}\theta}, \quad x \in \mathcal{X}.$$

Proof: Done in class.

Remark: The condition that Θ is a convex set ensures that $\mathbb{E}_\pi[\theta|x] \in \Theta$ for any $x \in \mathcal{X}$.

Proposition 2.1 *Consider the set-up of Theorem 2.4. If $r(\pi) < +\infty$ then the estimator $\delta^\pi(x) = \mathbb{E}_\pi[\theta|x]$ is admissible.*

Proof: If $r(\pi) < +\infty$ the estimator δ^π is the unique Bayes estimator associated with the quadratic loss function and the result follows from Theorem 2.3.

Exercise: Show that the posterior expectation is also the Bayes estimator associated with the more general loss function

$$L(\theta, d) = (\theta - d)^T Q(\theta - d), \quad (\theta, d) \in \Theta^2$$

where Q is an arbitrary symmetric positive definite matrix.

The absolute error loss function

Assuming $\mathcal{D} = \Theta \subseteq \mathbb{R}$, the absolute error loss function is defined by

$$L(\theta, d) = |\theta - d|, \quad (\theta, d) \in \Theta^2.$$

Theorem 2.5 *Assume that $\Theta \subseteq \mathbb{R}$ is a convex set and that $\mathbb{E}_\pi[|\theta||x] < +\infty$ for all $x \in \mathcal{X}$. Then, an estimator δ^π associated with the absolute error loss function is such that, for all $x \in \mathcal{X}$, $\delta^\pi(x)$ is a median of $\pi(\theta|x)$.*

Proof: See Appendix 2.

Recall that $m \in \mathbb{R}$ is a median of $\pi(\theta|x)$ if

$$\pi\big(\{\theta : \theta \leq m\}\big) \geq \frac{1}{2}, \quad \pi\big(\{\theta : \theta \geq m\}\big) \geq \frac{1}{2}.$$

Remarks:

1. The result of Theorem 2.5 still holds if Θ is a countable set.
2. In comparison with the quadratic loss, the absolute error loss penalizes less large errors.
3. The posterior median may not be unique but always exists.
4. Even when the posterior median is not unique $\delta^\pi(x)$ is in general unique.

Exercise: Show that, for $k_1, k_2 > 0$, an estimator δ^π associated with the loss

$$L(\theta, d) = \begin{cases} k_2(\theta - d) \text{ if } \theta > d, \\ k_1(d - \theta) \text{ otherwise} \end{cases}$$

is a $k_2/(k_1 + k_2)$ quantile of the posterior distribution.

The 0–1 loss function

Assuming $\mathcal{D} = \Theta$, the 0–1 loss function is defined by

$$L(\theta, d) = 1 - \mathbb{I}_\theta(d), \quad (\theta, d) \in \Theta^2.$$

When the support of $\pi(\theta|x)$ is a countable set we have the following result:

Theorem 2.6 *Assume that the support of $\pi(\theta|x)$ is a countable set. Then, an estimator δ^π associated with the above 0–1 loss function is such that, for any $x \in \mathcal{X}$, $\delta^\pi(x)$ is a mode of $\pi(\theta|x)$.*

Proof: Done in class.

Remarks:

1. The posterior mode may not be unique.
2. When the posterior mode is not unique $\delta^\pi(x)$ is an arbitrary mode of $\pi(\theta|x)$ (and thus δ^π is not unique).
3. If $\mathrm{d}\theta$ is a continuous measure (i.e. $\pi(\theta|x)\mathrm{d}\theta$ is a continuous probability distribution) then the posterior expected loss associated with the above 0–1 loss function is one for any decision rule δ since

$$\int_\Theta (1 - \mathbb{I}_\theta(d)\pi(\theta|x)\mathrm{d}\theta = 1, \quad \forall (d, x) \in \Theta \times \mathcal{X}.$$

The MAP estimator for continuous parameter spaces

Assuming $\mathcal{D} = \Theta \subseteq \mathbb{R}^d$ we consider, for $\epsilon > 0$, the 0–1 loss function defined by

$$L_\epsilon(\theta, d) = \mathbb{I}_{\|\theta - d\| > \epsilon}, \quad (\theta, d) \in \Theta^2.$$

For $\epsilon > 0$ let δ^π_ϵ be an estimator such that

$$\delta^\pi_\epsilon(x) \in \operatorname*{argmin}_{d \in \Theta} \pi\big(\{\theta : \|\theta - d\| > \epsilon\}|x\big), \quad \forall x \in \mathcal{X}$$

and, for $x \in \mathcal{X}$, let $\delta^\pi_{\mathrm{MAP}}(x)$ be a posterior mode of $\pi(\theta|x)$; that is

$$\delta^\pi_{\mathrm{MAP}}(x) \in \operatorname*{argmax}_{\theta \in \Theta} \pi(\theta|x).$$

Then, we have the following result for the MAP estimator.

Theorem 2.7 *Let $x \in \mathcal{X}$, assume that $\pi(\theta|x)$ is continuous. Then, under some technical conditions, $\delta^\pi_{MAP}(x) = \lim_{\epsilon \to 0} \delta^\pi_\epsilon(x)$.*

Proof: See Bassett, R. and Deride J (2016). "Maximum a posteriori estimators as a limit of Bayes estimators". *Mathematical Programming*, p. 1-16.

Important remarks:

1. Because the MAP estimator is obtained as a limit of Bayes estimators (and not by minimizing a posterior expected loss for a given loss function) it is not a Bayes estimator when $\pi(\theta|x)$ is continuous.

2. Marginal MAP estimates are usually not coherent with the joint MAP estimate.

Example: The Binomial model

Recall that, for parameters $\alpha, \beta > 0$, the density of the Beta(α, β) distribution is defined

$$f_{\alpha,\beta}(x) = \frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)} x^{\alpha-1}(1-x)^{\beta-1}, \quad x \in (0,1)$$

where Γ stands for the Gamma function, i.e.

$$\Gamma(t) = \int_0^{+\infty} x^{t-1} e^{-x} \,\mathrm{d}x, \quad t > 0.$$

Proposition 2.2 *Let $(n, \alpha_0, \beta_0) \in \mathbb{N} \times \mathbb{R}_{>0} \times \mathbb{R}_{>0}$ and consider the Bayesian statistical model defined by*

$$\pi(\theta) = f_{\alpha_0,\beta_0}(\theta), \quad f(x_n|\theta) = \binom{n}{x_n} \theta^{x_n}(1-\theta)^{n-x_n},$$

where $x_n \in \{0, \dots, n\}$. Then, $\pi(\theta|x_n) = f_{\alpha_n,\beta_n}(\theta)$ with $\alpha_n = \alpha_0 + x_n$ and $\beta_n = \beta_0 + n - x_n$. Consequently, $\mathbb{E}_\pi[\theta|x_n] = \alpha_n/(\alpha_n + \beta_n)$ and, assuming $\alpha_0, \beta_0 > 1$,

$$\frac{\alpha_n - 1}{\alpha_n + \beta_n - 2} = \operatorname*{argmax}_{\theta \in (0,1)} \pi(\theta|x_n)$$

and

$$\pi\Big(\Big[0, \Big(\alpha_n - \frac{1}{3}\Big)/\Big(\alpha_n + \beta_n - \frac{2}{3}\Big)\Big]\Big| x_n\Big) \approx \frac{1}{2}.$$

Proof: Done in class (but the formula for the posterior median is admitted).

The Binomial model

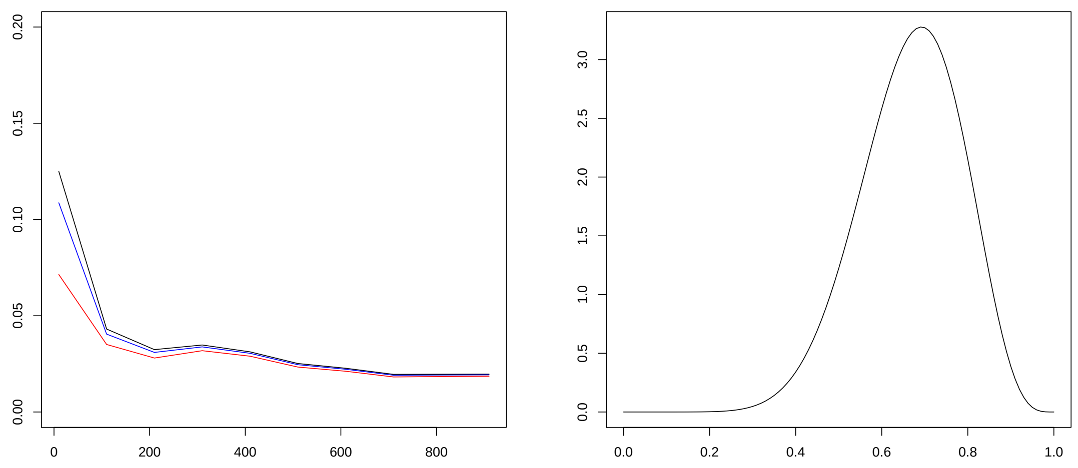

Figure 1: The left plot gives the posterior mean (black), posterior mode (red) and posterior median as a function of n while the right plot shows the prior distribution. The parameters of this latter are $(\alpha_0, \beta_0) = (1, 5)$ while $X_n \sim \text{Binomial}(n, 0.02)$.

Some lessons from the Binomial example:

1. We observe some gaps between the different Bayes estimates when n is small ($n < 200$, say).

2. However, these gaps disappear as n increases and, in fact, the different Bayes estimators converge toward the true parameter value as $n \to +\infty$.

3. One reason for the phenomenon described in 2. is that the posterior distribution is approximatively Gaussian when n is large (see Chapter 6).

Appendix 1: Proof of Theorem 2.1

To prove Theorem 2.1 we will need the following result known as "Stein's lemma".

Lemma 2.1 *Let $Z \sim \mathcal{N}_1(0,1)$ and $h : \mathbb{R} \to \mathbb{R}$ be a differentiable function such that $\mathbb{E}[h'(Z)] < +\infty$. Then,*

$$\mathbb{E}[Zh(Z)] = \mathbb{E}[h'(Z)].$$

Proof: We have

$$\begin{aligned}
\mathbb{E}[Zh(Z)] &= \int_{-\infty}^{+\infty} zh(z)\frac{1}{\sqrt{2\pi}}e^{-\frac{z^2}{2}}\mathrm{d}z \\
&= \int_{-\infty}^{+\infty} zh(z)\frac{1}{\sqrt{2\pi}}e^{-\frac{z^2}{2}}\mathrm{d}z - h(0)\mathbb{E}[Z] \\
&= \int_{-\infty}^{+\infty} z(h(z)-h(0))\frac{1}{\sqrt{2\pi}}e^{-\frac{z^2}{2}}\mathrm{d}z \\
&= \int_{-\infty}^{+\infty} z\bigg(\int_0^z h'(u)\mathrm{d}u\bigg)\frac{1}{\sqrt{2\pi}}e^{-\frac{z^2}{2}}\mathrm{d}z
\end{aligned} \tag{2}$$

while

$$\begin{aligned}
\mathbb{E}[h'(Z)] &= \int_{-\infty}^{0} h'(z)\frac{1}{\sqrt{2\pi}}e^{-\frac{z^2}{2}}\mathrm{d}z + \int_{0}^{+\infty} h'(z)\frac{1}{\sqrt{2\pi}}e^{-\frac{z^2}{2}}\mathrm{d}z \\
&= \frac{1}{\sqrt{2\pi}}\int_{-\infty}^{0} h'(z)\bigg(\int_{-\infty}^{z} -ue^{-\frac{u^2}{2}}\mathrm{d}u\bigg)\mathrm{d}z \\
&+ \frac{1}{\sqrt{2\pi}}\int_{0}^{+\infty} h'(z)\bigg(\int_{z}^{+\infty} ue^{-\frac{u^2}{2}}\mathrm{d}u\bigg)\mathrm{d}z.
\end{aligned} \tag{3}$$

We now study the two integrals that appear on the right-hand side of the second equality sign.

Appendix 1: Proof of Theorem 2.1 (continued)

Using Fubini's theorem,

$$\begin{aligned}
&\frac{1}{\sqrt{2\pi}}\int_{-\infty}^{0} h'(z)\Big(\int_{-\infty}^{z} -ue^{-\frac{u^2}{2}}\mathrm{d}u\Big)\mathrm{d}z \\
&\qquad = \int_{-\infty}^{0}\left(\int_{-\infty}^{0}\frac{1}{\sqrt{2\pi}}h'(z)\mathbf{1}_{(-\infty,z]}(u)\mathrm{d}z\right)(-u)e^{-\frac{u^2}{2}}\mathrm{d}z\mathrm{d}u \\
&\qquad = \int_{-\infty}^{0}\left(\int_{u}^{0}\frac{1}{\sqrt{2\pi}}h'(z)\mathrm{d}z\right)(-u)e^{-\frac{u^2}{2}}\mathrm{d}u \\
&\qquad = \int_{-\infty}^{0}\left(\int_{0}^{u}\frac{1}{\sqrt{2\pi}}h'(z)\mathrm{d}z\right)ue^{-\frac{u^2}{2}}\mathrm{d}u
\end{aligned}$$

and, similarly, one can easily check that

$$\frac{1}{\sqrt{2\pi}}\int_{0}^{+\infty} h'(z)\Big(\int_{z}^{+\infty} ue^{-\frac{u^2}{2}}\mathrm{d}u\Big)\mathrm{d}z = \int_{0}^{+\infty}\left(\int_{0}^{u}\frac{1}{\sqrt{2\pi}}h'(z)\mathrm{d}z\right)ue^{-\frac{u^2}{2}}\mathrm{d}u.$$

Together with (3), this shows that

$$\mathbb{E}[h'(Z)] = \int_{-\infty}^{+\infty} u\left[\int_{0}^{u}\frac{1}{\sqrt{2\pi}}h'(z)\mathrm{d}z\right]e^{-\frac{u^2}{2}}\mathrm{d}u = \mathbb{E}[Zf(Z)]$$

where the second equality is due to (2). This concludes the proof of Lemma 2.1.

Appendix 1: Proof of Theorem 2.1 (continued)

We now prove Theorem 2.1. To this end remark first that

$$R(\theta,\delta_0)=\int_{\mathbb{R}^d}\sum_{i=1}^{d}(\theta_i-x_i)^2 f(x|\theta)\mathrm{d}x=\sum_{i=1}^{d}\mathbb{E}_\theta[(X_i-\theta_i)^2]=d,\quad \forall\theta\in\Theta.$$

Next, let $\delta^{JS}:\mathbb{R}^d\to\mathbb{R}^d$ be defined by

$$\delta^{JS}(x)=x-\frac{d-2}{\|x\|^2}x,\quad x\in\mathbb{R}^d$$

and we now compute $R(\theta,\delta^{JS})$ for all $\theta\in\mathbb{R}^d$ and $d\geq 2$.

Let $d\geq 2$ and $\theta\in\mathbb{R}^d$. Then,

$$\begin{aligned}
R(\theta,\delta^{JS}) &= \int_{\mathbb{R}^d}\left\|\theta-x+\frac{d-2}{\|x\|^2}x\right\|^2 f(x|\theta)\mathrm{d}x\\
&= \int_{\mathbb{R}^d}\|\theta-x\|^2 f(x|\theta)\mathrm{d}x+(d-2)^2\int_{\mathbb{R}^d}\frac{\|x\|^2}{\|x\|^4}f(x|\theta)\mathrm{d}x\\
&\quad+2(d-2)\int_{\mathbb{R}^d}(\theta-x)^T\frac{x}{\|x\|^2}f(x|\theta)\mathrm{d}x\\
&= R(\theta,\delta_0)+(d-2)^2\int_{\mathbb{R}^d}\frac{1}{\|x\|^2}f(x|\theta)\mathrm{d}x\\
&\quad+2(d-2)\sum_{i=1}^{d}\frac{(\theta_i-x_i)x_i}{\|x\|^2}f(x|\theta)\mathrm{d}x\\
&= R(\theta,\delta_0)+(d-2)^2\mathbb{E}_\theta\left[\frac{1}{\|X\|^2}\right]\\
&\quad+2(d-2)\sum_{i=1}^{d}\mathbb{E}_\theta\left[\frac{(\theta_i-X_i)X_i}{\|X\|^2}\right].
\end{aligned}\tag{4}$$

Appendix 1: Proof of Theorem 2.1 (end)

To proceed further let $i\in\{1,\dots,d\}$, $x_{-i}\in\mathbb{R}^{d-1}$ and $h_{i,x_{-i}}:\mathbb{R}\to\mathbb{R}$ be defined by

$$h_{i,x_{-i}}(z)=\frac{z+\theta_i}{\|(z+\theta_i,x_{-i})\|^2},\quad z\in\mathbb{R}.$$

We have

$$h'_{i,x_{-i}}(z)=\frac{\|(z+\theta_i,x_{-i})\|^2-2(z+\theta_i)^2}{\|(z+\theta_i,x_{-i})\|^4},\quad z\in\mathbb{R}$$

and thus $\mathbb{E}[h'_{i,x_{-i}}(Z)]<+\infty$ when $Z\sim\mathcal{N}_1(0,1)$. Therefore, using Lemma 2.1 we have (with '$X_{-i}=X$ without component i')

$$\begin{aligned}\mathbb{E}_\theta\left[\frac{(\theta_i-X_i)X_i}{\|X\|^2}\Big|X_{-i}=x_{-i}\right]&=-\mathbb{E}_\theta\big[(X_i-\theta_i)h_{i,x_{-i}}(X_i-\theta_i)\big]\\&=-\mathbb{E}_\theta\left[\frac{\|(X_i,x_{-i})\|^2-2X_i^2}{\|(X_i,x_{-i})\|^4}\right]\\&=\mathbb{E}_\theta\left[\frac{2\,X_i^2}{\|(X_i,x_{-i})\|^4}-\frac{1}{\|(X_i,x_{-i})\|^2}\right].\end{aligned}$$

Then, because this equality holds for any $x_{-i}\in\mathbb{R}^{d-1}$,

$$\mathbb{E}_\theta\left[\frac{(\theta_i-X_i)X_i}{\|X\|^2}\right]=2\,\mathbb{E}_\theta\left[\frac{X_i^2}{\|X\|^4}\right]-\mathbb{E}_\theta\left[\frac{1}{\|X\|^2}\right],\quad\forall i\in\{1,\dots,d\}$$

and thus

$$\sum_{i=1}^d\mathbb{E}_\theta\left[\frac{(\theta_i-X_i)X_i}{\|X\|^2}\right]=-(d-2)\mathbb{E}_\theta\left[\frac{1}{\|X\|^2}\right].$$

Then, using (4), it follows that for any $d\geq 2$ we have

$$R(\theta,\delta^{JS})=R(\theta,\delta_0)-(d-2)^2\mathbb{E}_\theta\left[\frac{1}{\|X\|^2}\right],\quad\forall\theta\in\mathbb{R}^d$$

and the proof is complete.

Remark: The above computations are not valid when $d=1$ since in this case $R(\theta,\delta^{JS})=+\infty$ for any $\theta\in\Theta$.

Appendix 2: Proof of Theorem 2.5

To prove Theorem 2.5 let $d \in \Theta$, $x \in \mathcal{X}$ and assume that $\Theta = \mathbb{R}$ (the extension to an arbitrary convex set being trivial). Below we use the shorthand $\pi(\mathrm{d}\theta|x) = \pi(\theta|x)\mathrm{d}\theta$ and $\int_{\mathbb{R}} f(y)\mathrm{d}y$ denotes the (improper) Riemman integral of f on $\mathbb{R}$.

Then, because $\mathbb{E}_\pi[|\theta|\, x] < +\infty$, we have

$$\begin{aligned}\rho(d) := \rho(\pi, d|x) = &\int_\Theta \mathbf{1}_{(-\infty,d]}(\theta)(d-\theta)\pi(\mathrm{d}\theta|x)\\ &+ \int_\Theta \mathbf{1}_{(d,+\infty)}(\theta)(\theta-d)\pi(\mathrm{d}\theta|x)\end{aligned}$$

where (using Fubini's theorem for the third equality)

$$\begin{aligned}\int_{-\infty}^{d} \pi(\{\theta:\ \theta \leq y\}|x)\mathrm{d}y &= \int_{-\infty}^{d}\Big(\int_\Theta \mathbf{1}_{(-\infty,y]}(\theta)\pi(\mathrm{d}\theta|x)\Big)\mathrm{d}y\\ &= \int_{-\infty}^{d}\Big(\int_\Theta \mathbf{1}_{(-\infty,y]}(\theta)\mathbf{1}_{(-\infty,d]}(\theta)\pi(\mathrm{d}\theta|x)\Big)\mathrm{d}y\\ &= \int_\Theta \mathbf{1}_{(-\infty,d]}(\theta)\Big(\int_{-\infty}^{d}\mathbf{1}_{(-\infty,y]}(\theta)\mathrm{d}y\Big)\pi(\mathrm{d}\theta|x)\\ &= \int_\Theta \mathbf{1}_{(-\infty,d]}(\theta)\Big(\int_{\theta}^{d}\mathrm{d}y\Big)\pi(\mathrm{d}\theta|x)\\ &= \int_\Theta \mathbf{1}_{(-\infty,d]}(\theta)(d-\theta)\pi(\mathrm{d}\theta|x).\end{aligned}$$

Remark: Funini's theorem can be used because $\mathbb{E}_\pi[|\theta|\, x] < +\infty$.

Appendix 2: Proof of Theorem 2.5 (continued)

Similarly (using again Fubini's theorem for the third equality),

$$\begin{aligned}
\int_d^{+\infty} \pi(\{\theta:\ \theta > y\}|x)\mathrm{d}y &= \int_d^{+\infty} \left(\int_\Theta \mathbf{1}_{(y,+\infty)}(\theta)\pi(\mathrm{d}\theta|x)\right)\mathrm{d}y \\
&= \int_d^{+\infty} \left(\int_\Theta \mathbf{1}_{(y,+\infty)}(\theta)\mathbf{1}_{(d,+\infty)}(\theta)\pi(\mathrm{d}\theta|x)\right)\mathrm{d}y \\
&= \int_\Theta \mathbf{1}_{(d,+\infty)}(\theta)\left(\int_d^{+\infty} \mathbf{1}_{(y,+\infty)}(\theta)\mathrm{d}y\right)\pi(\mathrm{d}\theta|x) \\
&= \int_\Theta \mathbf{1}_{(d,+\infty)}(\theta)\left(\int_d^{\theta} \mathrm{d}y\right)\pi(\mathrm{d}\theta|x) \\
&= \int_\Theta \mathbf{1}_{(d,+\infty)}(\theta)(\theta-d)\pi(\mathrm{d}\theta|x)
\end{aligned}$$

so that

$$\rho(d) = \int_{-\infty}^{d} \pi(\{\theta:\ \theta \leq y\}|x)\mathrm{d}y + \int_d^{+\infty} \pi(\{\theta:\ \theta > y\}|x)\mathrm{d}y.$$

Then, using Leibniz integral rule,

$$\begin{aligned}
\rho'(d) &= \pi(\{\theta:\ \theta \leq d\}|x) - \pi(\{\theta:\ \theta > d\}|x) \\
&= 2\pi(\{\theta:\ \theta \leq d\}|x) - 1.
\end{aligned}$$

Let $d^* \in \mathbb{R}$ be such that $\pi(\{\theta:\ \theta \leq d^*\}|x) \geq 1/2$ and remark that, since the mapping $y \mapsto |y|$ is convex on $\mathbb{R}$, the mapping $d \mapsto \rho(d)$ is convex on $\mathbb{R}$ (see Problem Sheet 1, Problem 5). Hence,

$$\rho(d) \geq \rho(d^*) + \rho'(d^*)(d-d^*) \geq \rho(d^*), \quad \forall d > d^*. \tag{5}$$

Appendix 2: Proof of Theorem 2.5 (continued)

Next, because (5) holds for any d^* such that $\rho'(d^*) \geq 0$, this inequality holds in particular for

$$d^* = \min\{d \in \mathbb{R} : \pi(\{\theta :\ \theta \leq d\}|x) \geq 1/2\}. \tag{6}$$

(Note that d^* is well defined since the mapping $d \mapsto \pi(\{\theta :\ \theta \leq d\}|x)$ is right continuous.)

Let $d < d^*$. Then, there exists an $\epsilon_d > 0$ such that, for all $\epsilon \in (0, \epsilon_d)$, we have $d \leq d^* - \epsilon < d^*$ and thus

$$\rho(d) \geq \rho(d^* - \epsilon) + \rho'(d^* - \epsilon)(d - (d^* - \epsilon)) \geq \rho(d^* - \epsilon), \quad \forall \epsilon \in (0, \epsilon_d).$$

Then, because the mapping $d \mapsto \rho(d)$ is continuous on $\mathbb{R}$ (because it is convex on this set), together with (5) this shows that d^* defined in (6) is such that

$$\rho(d) \geq \rho(d^*), \quad \forall d \neq d^*.$$

Hence, $d^* \in \operatorname{argmin}_{d\in\Theta} \rho(d)$.

To conclude the proof it remains to show that d^* is a median of $\pi(\theta|x)\mathrm{d}\theta$; that is,

$$\pi(\{\theta :\ \theta \leq d^*\}|x) \geq 1/2, \quad \pi(\{\theta :\ \theta \geq d^*\}|x) \geq 1/2 \tag{7}$$

where the first inequality holds by the definition of d^*.

Appendix 2: Proof of Theorem 2.5 (end)

We show the second inequality in (7) by contradiction and assume that $\pi(\{\theta : \theta \geq d^*\}|x) < 1/2$.

In this case

$$\begin{aligned} \pi(\{\theta : \theta \leq d^*\}|x) &= \pi(\{\theta : \theta < d^*\}|x) + \pi(\{\theta : \theta = d^*\}|x) \\ &> \frac{1}{2} + \pi(\{\theta : \theta = d^*\}|x) \end{aligned} \tag{8}$$

and we consider below the two possible cases.

1. $\pi(\{\theta : \theta = d^*\}|x) = 0$. In this case, (8) imples that

$$\pi(\{\theta : \theta \leq d^*\}|x) > 1/2$$

   and thus there exists a $d' < d^*$ such that $\pi(\{\theta : \theta \leq d'\}|x) \geq 1/2$, contradicting the definition of d^*. (Such a d' exists because the mapping $d \mapsto \pi(\{\theta : \theta \leq d\}|x)$ is continuous at d^* when $\pi(\{\theta : \theta = d^*\}|x) = 0$.)

2. $\pi(\{\theta : \theta = d^*\}|x) > 0$. In this case,

$$\pi(\{\theta : \theta \leq d^*\}|x) > \pi(\{\theta : \theta < d^*\}|x) > \frac{1}{2}$$

   and again (since the first equality is strict) there exists a $d' < d^*$ such that $\pi(\{\theta : \theta \leq d'\}|x) \geq 1/2$, contradicting the definition of d^*.

Therefore, (8) holds and the proof is complete.