On Some Geometrical Aspects of Bayesian Inference

Miguel de Carvalho†

†Joint with B. J. Barney and G. L. Page; Brigham Young University, US

THE UNIVERSITY of EDINBURGH
School of Mathematics

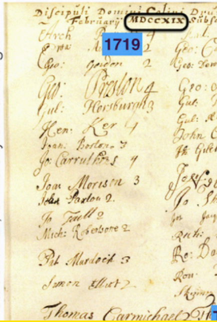
Athena SWAN Bronze Award

LONDON MATHEMATICAL SOCIETY GOOD PRACTICE SCHEME

ISBA 2018: Edinburgh, June 24–29

World Meeting of the International Society for Bayesian Analysis

THE UNIVERSITY of EDINBURGH



Student Roll, University of Edinburgh

Introduction

Motivation

- Bayesian methodologies have become main stream.
- Because of this, there is a need to develop methods accessible to 'non-experts' that assess the influence of model choices on inference.
- These will need to be:
    1. Easy to interpret.
    2. Easy to calculate.

- Bayesian methodologies have become main stream.
- Because of this, there is a need to develop methods accessible to 'non-experts' that assess the influence of model choices on inference.
- These will need to be:
  1. Easy to interpret.
  2. Easy to calculate.

Ideally: Provide a unified treatment to all pieces of Bayes theorem.

- Much work has been devoted to developing methods to assess the sensitivity of the posterior to changes in the prior and likelihood.
- The so-called prior–data conflict has been another subject which has been attracting attention (Evans and Moshonov, 2006; Walter and Augustin, 2009; Al Labadi and Evans, 2016).
- Others have investigated two competing priors to specifiy so-called weakly informative priors (Evans and Jang, 2011; Gelman et al., 2011).

- The novel contribution we intend to make is to provide a metric that is able to carry out comparisons between the:
  - prior and likelihood: to assess the prior–data agreement;
  - prior and posterior: to assess the influence that the prior has on inference;
  - prior and prior: to compare information available on competing priors.
- To be useful this metric should be:
  1. Easy to interpret.
  2. Easy to calculate.

Ideally: Provide a unified treatment to all pieces of Bayes theorem.

- To this end, we view each of the components of Bayes theorem as if they belonged to a geometry and seek to provide intuitively appealing interpretations of the norms and angles between the vectors of this geometry.
- We will show that calculating these quantities is very straightforward and can be done online.
- Interpretations are similar to those that accompany the correlation coefficient for continuous random variables.

Example (Christensen et al, 2011, pp. 26–27)

- Suppose interest lies in estimating the proportion $\theta \in [0,1]$ of US transportation industry workers that use drugs on the job. Suppose $\boldsymbol{y} = (0,1,0,0,0,0,1,0,0,0)$ and that

$$\boldsymbol{y} \mid \theta \overset{\text{iid}}{\sim} \text{Bern}(\theta), \quad \theta \sim \text{Beta}(a,b), \quad \theta \mid \boldsymbol{y} \sim \text{Beta}(a^\star, b^\star),$$

  with $a^\star = \sum Y_i + a$ and $b^\star = n - \sum Y_i + b$.
- The authors conduct the analysis picking $(a,b) = (3.44, 22.99)$.

Some key questions:

- How compatible is the likelihood with this prior choice?
- How similar are the posterior and prior distributions?
- How does the choice of Beta(a, b) compare to other possible prior distributions?

We provide a unified treatment to answer the questions above.

Storyboard
Plan of this Talk

1. Introduction (Done)
2. Bayes Geometry (Next)
3. Posterior and Prior Mean-Based Estimators of Compatibility
4. Discussion

- Suppose the inference of interest is over a parameter $\boldsymbol{\theta}$ in $\Theta \subseteq \mathbb{R}^p$.
- We work in $L_2(\Theta)$, and use the geometry of the Hilbert space

$$\mathscr{H} = (L_2(\Theta), \langle \cdot, \cdot \rangle),$$

with inner-product

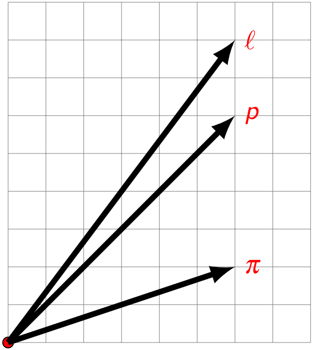
$$\langle g, h \rangle = \int_\Theta g(\boldsymbol{\theta}) h(\boldsymbol{\theta}) \, \mathrm{d}\boldsymbol{\theta}, \quad g, h \in L_2(\Theta),$$

and norm $\| \cdot \| = (\langle \cdot, \cdot \rangle)^{1/2}$.
- The fact that $\mathscr{H}$ is an Hilbert space is often known as the Riesz–Fischer theorem (Cheney, 2001, p. 411).

- Bayes theorem

$$p(\boldsymbol{\theta} \mid \mathbf{y}) = \frac{\pi(\boldsymbol{\theta}) f(\mathbf{y} \mid \boldsymbol{\theta})}{\int_{\Theta} \pi(\boldsymbol{\theta}) f(\mathbf{y} \mid \boldsymbol{\theta}) \, \mathrm{d}\boldsymbol{\theta}} = \frac{\pi(\boldsymbol{\theta})\ell(\boldsymbol{\theta})}{\langle \pi, \ell \rangle}.$$

pace1.5cm



- The likelihood vector is used to enlarge/reduce the magnitude and suitably tilt the direction of the prior vector.

- Define the angle measure between the prior and the likelihood as

$$\pi \angle \ell = \arccos \frac{\langle \pi, \ell \rangle}{\|\pi\| \|\ell\|}.$$

- Define the angle measure between the prior and the likelihood as

$$\pi \angle \ell = \arccos \frac{\langle \pi, \ell \rangle}{\|\pi\| \|\ell\|}.$$

- Since π and ℓ are nonnegative, $\pi \angle \ell \in [0, 90^\circ]$.
- Bayes theorem is incompatible with a prior being orthogonal to the likelihood as

$$\pi \angle \ell = 90^\circ \Rightarrow \langle \pi, \ell \rangle = 0,$$

thus leading to a division by zero.
- Our first target object of interest is given by a standardized inner product

$$\kappa_{\pi,\ell} = \frac{\langle \pi, \ell \rangle}{\|\pi\| \|\ell\|},$$

which quantifies how much an expert's opinion agrees with the data, thus providing a natural measure of prior–data agreement.

Definition (Millman and Parker, 1991, p. 17)

An abstract geometry $\mathscr{A}$ consists of a pair $\{\mathscr{P}, \mathscr{L}\}$, where the elements of set $\mathscr{P}$ are designed as points, and the elements of the collection $\mathscr{L}$ are designed as lines, such that:

1. For every two points $A, B \in \mathscr{P}$, there is a line $l \in \mathscr{L}$.
2. Every line has at least two points.

- Our abstract geometry of interest is $\mathscr{A} = \{\mathscr{P}, \mathscr{L}\}$, where $\mathscr{P} = L_2(\Theta)$ and

$$\mathscr{L} = \{g + kh, : g, h \in L_2(\Theta)\}.$$

- In our setting points are, for example, prior densities, posterior densities, or likelihoods, as long as they are in $L_2(\Theta)$.

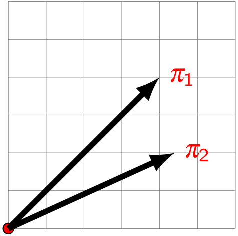
- Lines are elements of $\mathscr{L}$, so that for example if g and h are densities, line segments in our geometry consist of all possible mixture distributions which can be obtained from g and h, i.e.:

$$\{\lambda g + (1-\lambda)h : \lambda \in [0,1]\}.$$

- Vectors in $\mathscr{A} = \{\mathscr{P}, \mathscr{L}\}$ are defined through the difference of elements in $\mathscr{P} = L_2(\Theta)$.
- If $g, h \in L_2(\Theta)$ are vectors then we say that g and h are collinear if there exists $k \in \mathbb{R}$, such that $g(\boldsymbol{\theta}) = kh(\boldsymbol{\theta})$.
- Put differently, we say g and h are collinear if $g(\boldsymbol{\theta}) \propto h(\boldsymbol{\theta})$, for all $\boldsymbol{\theta} \in \Theta$.

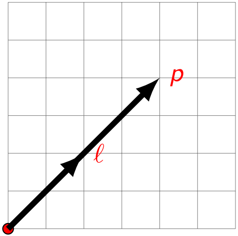
- Two different densities π_1 and π_2 cannot be collinear:

  If $\pi_1 = k\pi_2$, then $k = 1$, otherwise $\int \pi_2(\boldsymbol{\theta})\,\mathrm{d}\boldsymbol{\theta} \neq 1$.



- A density can be collinear to a likelihood:

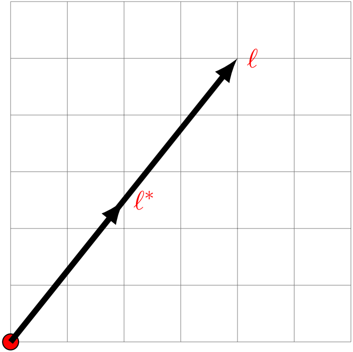
  If the prior is uniform $p(\boldsymbol{\theta} \mid \boldsymbol{y}) \propto \ell(\boldsymbol{\theta})$.



- Our geometry is compatible with having two likelihoods being collinear.
- This can be used to rethink the strong likelihood principle that states that if

$$\ell(\boldsymbol{\theta}) = f(\boldsymbol{\theta} \mid \mathbf{y}) \propto f(\boldsymbol{\theta} \mid \mathbf{y}^*) = \ell^*(\boldsymbol{\theta}),$$

then the *same* inference should be drawn from both samples.



pace0.5cm According to our geometry the strong likelihood principle reads:

"Likelihoods with the same direction should yield the same inference."

Definition (Compatibility)

The compatibility between points in the geometry under consideration is the mapping $\kappa : L_2(\Theta) \times L_2(\Theta) \to [0,1]$ defined as

$$\kappa_{g,h} = \frac{\langle g, h \rangle}{\|g\| \|h\|}, \quad g, h \in L_2(\Theta).$$

pace-.1cmPearson correlation coefficient *vs.* compatibility

$$\begin{cases} \langle X, Y \rangle = \int_\Omega XY \, \mathrm{d}P, \\ X, Y \in L_2(\Omega, \mathbb{B}_\Omega, P), \end{cases}$$

Definition (Compatibility)

The compatibility between points in the geometry under consideration is the mapping $\kappa : L_2(\Theta) \times L_2(\Theta) \to [0, 1]$ defined as

$$\kappa_{g,h} = \frac{\langle g, h \rangle}{\|g\| \|h\|}, \quad g, h \in L_2(\Theta).$$

pace-.1cmPearson correlation coefficient *vs.* compatibility

$$\begin{cases} \langle X, Y \rangle = \int_\Omega XY \, \mathrm{d}P, \\ X, Y \in L_2(\Omega, \mathbb{B}_\Omega, P), \end{cases}$$

instead of

Definition (Compatibility)

The compatibility between points in the geometry under consideration is the mapping $\kappa : L_2(\Theta) \times L_2(\Theta) \to [0,1]$ defined as

$$\kappa_{g,h} = \frac{\langle g, h \rangle}{\|g\| \|h\|}, \quad g, h \in L_2(\Theta).$$

pace-.1cmPearson correlation coefficient *vs.* compatibility

$$\begin{cases} \langle X, Y \rangle = \int_\Omega XY \, \mathrm{d}P, \\ X, Y \in L_2(\Omega, \mathbb{B}_\Omega, P), \end{cases} \qquad \underline{\text{instead of}} \qquad \begin{cases} \langle g, h \rangle = \int_\Theta g(\boldsymbol{\theta}) h(\boldsymbol{\theta}) \, \mathrm{d}\boldsymbol{\theta}, \\ g, h \in L_2(\Theta). \end{cases}$$

pace-.2cm Note that:

- $\kappa_{\pi,\ell}$: prior-data agreement. pace0.05cm
- $\kappa_{\pi,p}$: sensitivity of the posterior to the prior specification. pace0.05cm
- κ_{π_1,π_2}: compatibility of different priors [coherency of opinions of experts].

Bayes Geometry

Norms and their Interpretation

- $\kappa_{\pi,\ell}$ is comprised of function norms: How do we interpret norms?
- In some cases the norm of a density is linked to the variance.

Example

Let $U \sim \text{Unif}(a,b)$ and let $\pi(x) = (b-a)^{-1} I_{(a,b)}(x)$. Then,

$$\|\pi\| = 1/(12\sigma_U^2)^{1/4},$$

where the variance of U is $\sigma_U^2 = 1/12(b-a)^2$.

Example

Let $X \sim \text{N}(\mu, \sigma_X^2)$ with known variance σ_X^2. It can be shown that

$$\|\phi\| = \{\int_{\mathbb{R}} \phi^2(x;\mu,\sigma_X^2)\,\mathrm{d}\mu\}^{1/2} = 1/(4\pi\sigma_X^2)^{1/4}.$$

Proposition

Let $\Theta \subset \mathbb{R}^p$ *with* $|\Theta| < \infty$ *where* $|\cdot|$ *denotes the Lebesgue measure. Consider* $\pi : \Theta \to [0, \infty)$ *a probability density with* $\pi \in L_2(\Theta)$ *and let* $\pi_0 \sim \mathit{Unif}(\Theta)$ *denote a uniform density on* Θ*, then*

$$\|\pi\|^2 = \|\pi - \pi_0\|^2 + \|\pi_0\|^2.$$

Proposition

Let $\Theta \subset \mathbb{R}^p$ with $|\Theta| < \infty$ where $|\cdot|$ denotes the Lebesgue measure. Consider $\pi : \Theta \to [0, \infty)$ a probability density with $\pi \in L_2(\Theta)$ and let $\pi_0 \sim Unif(\Theta)$ denote a uniform density on Θ, then

$$\|\pi\|^2 = \|\pi - \pi_0\|^2 + \|\pi_0\|^2.$$

- This interpretation cannot be applied to Θ's that do not have finite Lebesgue measure as there is no corresponding proper Uniform distribution.

Proposition

Let $\Theta \subset \mathbb{R}^p$ with $|\Theta| < \infty$ where $|\cdot|$ denotes the Lebesgue measure. Consider $\pi : \Theta \to [0, \infty)$ a probability density with $\pi \in L_2(\Theta)$ and let $\pi_0 \sim Unif(\Theta)$ denote a uniform density on Θ, then

$$\|\pi\|^2 = \|\pi - \pi_0\|^2 + \|\pi_0\|^2.$$

- This interpretation cannot be applied to Θ's that do not have finite Lebesgue measure as there is no corresponding proper Uniform distribution.
- Yet, the notion that the norm of a density is a measure of its peakedness may be applied whether or not Θ has finite Lebesgue measure.

- To see this,

- To see this, evaluate $\pi(\theta)$ on a grid $\theta_1 < \cdots < \theta_D$

- To see this, evaluate $\pi(\theta)$ on a grid $\theta_1 < \cdots < \theta_D$ and consider the vector

$$p = (\pi_1, \ldots, \pi_D),$$

with $\pi_d = \pi(\theta_d)$ for $d = 1, \ldots, D$.

- To see this, evaluate $\pi(\theta)$ on a grid $\theta_1 < \cdots < \theta_D$ and consider the vector

$$p = (\pi_1, \ldots, \pi_D),$$

with $\pi_d = \pi(\theta_d)$ for $d = 1, \ldots, D$.

- The larger the norm of the vector p, the higher the indication that certain components would be far from the origin

- To see this, evaluate $\pi(\theta)$ on a grid $\theta_1 < \cdots < \theta_D$ and consider the vector

$$p = (\pi_1, \ldots, \pi_D),$$

with $\pi_d = \pi(\theta_d)$ for $d = 1, \ldots, D$.

- The larger the norm of the vector p, the higher the indication that certain components would be far from the origin—that is, $\pi(\theta)$ would be peaking for certain θ in the grid.

- To see this, evaluate $\pi(\theta)$ on a grid $\theta_1 < \cdots < \theta_D$ and consider the vector

$$p = (\pi_1, \ldots, \pi_D),$$

with $\pi_d = \pi(\theta_d)$ for $d = 1, \ldots, D$.

- The larger the norm of the vector p, the higher the indication that certain components would be far from the origin—that is, $\pi(\theta)$ would be peaking for certain θ in the grid.

- Now, think of a density as a vector with infinitely many components (its value at each point of the support) and replace summation by integration to get the L_2 norm.

Bayes Geometry

Example (On-the-job drug usage toy example, cont. 1)

From the example in the Introduction we have $\theta \mid \boldsymbol{y} \sim \text{Beta}(a^\star, b^\star)$ with $a^\star = a + \sum Y_i = a + 2$ and $b^\star = b + n - \sum Y_i = b + 8$. The norm of the prior, posterior, and likelihood are respectively given by

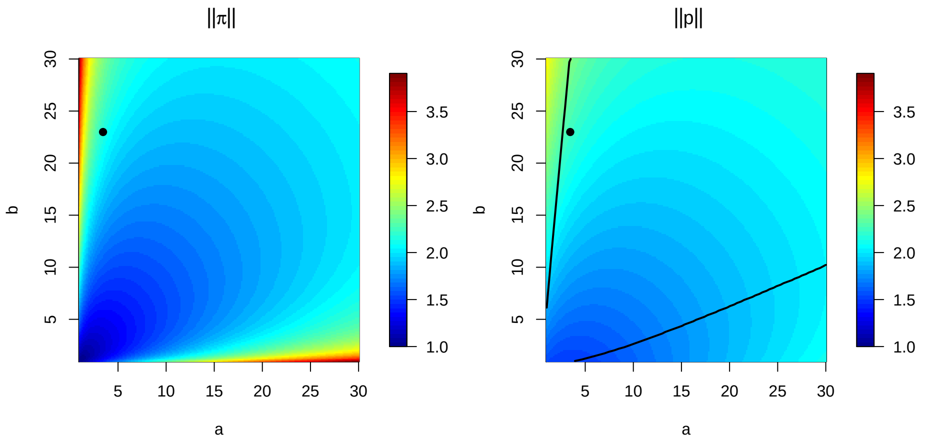
$$\|\pi(a, b)\| = \frac{\{B(2a-1, 2b-1)\}^{1/2}}{B(a, b)}, \quad a, b > 1/2,$$

and

$$\|p(a, b)\| = \|\pi(a^\star, b^\star)\|.$$

Bayes Geometry

Prior and Posterior Norms: On-the-Job Drug Usage Toy Example



pace-.5cm

Figure: Prior and posterior norms for on-the-job drug usage toy example. The black dot

Bayes Geometry

Angles Between Other Vectors

Considering κ, it follows that

$$\kappa_{\pi,\ell}(a,b) = B(a^\star, b^\star)\{B(2a-1,2b-1)B(2\sum Y_i + 1, 2(n - \sum Y_i) + 1)\}^{-1/2}.$$

As mentioned, we are not restricted to use κ only to compare π and ℓ.

Example (On-the-job drug usage toy example, cont. 2)

Extending a previous example, we calculate

$$\begin{aligned}\kappa_{\pi,p} = {} & B(\sum Y_i + 2a - 1, n - \sum Y_i + 2b - 1) \\ & \times \{B(2a-1,2b-1) \\ & \times B(2\sum Y_i + 2a - 1, 2n - 2\sum Y_i + 2b - 1)\}^{-1/2},\end{aligned}$$

and for $\pi_1 \sim \mathrm{Beta}(a_1, b_1)$ and $\pi_2 \sim \mathrm{Beta}(a_2, b_2)$,

$$\kappa_{\pi_1,\pi_2} = \frac{B(a_1 + a_2 - 1, b_1 + b_2 - 1)}{\{B(2a_1 - 1, 2b_1 - 1)B(2a_2 - 1, 2b_2 - 1)\}^{1/2}}.$$

Bayes Geometry

Compatibility: On-the-Job Drug Usage Toy Example

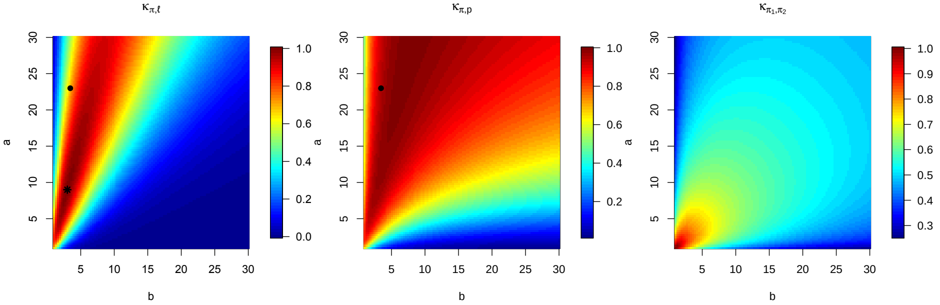


Figure: Compatibility (κ) for on-the-job drug usage toy example. In (i) and (ii) the black dot corresponds to $(a,b) = (3.44, 22.99)$ (values employed by Christensen et al. 2011, pp. 26–27).

Definition (Max-compatible prior)

Let $\boldsymbol{y} \sim f(\cdot \mid \boldsymbol{\theta})$, and let $\mathscr{P} = \{\pi(\boldsymbol{\theta} \mid \boldsymbol{\alpha}) : \boldsymbol{\alpha} \in \mathscr{A}\}$ be a family of priors for $\boldsymbol{\theta}$.

Definition (Max-compatible prior)

Let $\boldsymbol{y} \sim f(\cdot \mid \boldsymbol{\theta})$, and let $\mathscr{P} = \{\pi(\boldsymbol{\theta} \mid \boldsymbol{\alpha}) : \boldsymbol{\alpha} \in \mathscr{A}\}$ be a family of priors for $\boldsymbol{\theta}$. If there exists $\boldsymbol{\alpha}^*_{\boldsymbol{y}} \in \mathscr{A}$, such that $\kappa_{\pi,\ell}(\boldsymbol{\alpha}^*_{\boldsymbol{y}}) = 1$, the prior $\pi(\boldsymbol{\theta} \mid \boldsymbol{\alpha}^*_{\boldsymbol{y}}) \in \mathscr{P}$ is said to be max-compatible

Definition (Max-compatible prior)

Let $\boldsymbol{y} \sim f(\cdot \mid \boldsymbol{\theta})$, and let $\mathscr{P} = \{\pi(\boldsymbol{\theta} \mid \boldsymbol{\alpha}) : \boldsymbol{\alpha} \in \mathscr{A}\}$ be a family of priors for $\boldsymbol{\theta}$. If there exists $\boldsymbol{\alpha}_{\boldsymbol{y}}^* \in \mathscr{A}$, such that $\kappa_{\pi,\ell}(\boldsymbol{\alpha}_{\boldsymbol{y}}^*) = 1$, the prior $\pi(\boldsymbol{\theta} \mid \boldsymbol{\alpha}_{\boldsymbol{y}}^*) \in \mathscr{P}$ is said to be max-compatible, and $\boldsymbol{\alpha}_{\boldsymbol{y}}^*$ is said to be a max-compatible hyperparameter.

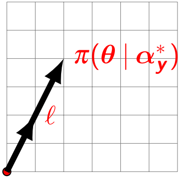
Definition (Max-compatible prior)

Let $\boldsymbol{y} \sim f(\cdot \mid \boldsymbol{\theta})$, and let $\mathscr{P} = \{\pi(\boldsymbol{\theta} \mid \boldsymbol{\alpha}) : \boldsymbol{\alpha} \in \mathscr{A}\}$ be a family of priors for $\boldsymbol{\theta}$. If there exists $\boldsymbol{\alpha}^*_{\boldsymbol{y}} \in \mathscr{A}$, such that $\kappa_{\pi,\ell}(\boldsymbol{\alpha}^*_{\boldsymbol{y}}) = 1$, the prior $\pi(\boldsymbol{\theta} \mid \boldsymbol{\alpha}^*_{\boldsymbol{y}}) \in \mathscr{P}$ is said to be max-compatible, and $\boldsymbol{\alpha}^*_{\boldsymbol{y}}$ is said to be a max-compatible hyperparameter.

- The max-compatible hyperparameter, $\boldsymbol{\alpha}^*_{\boldsymbol{y}}$, is by definition a random vector, and thus a max-compatible prior density is a random function.
- Geometrically: A prior is max-compatible iff it is collinear to the likelihood in the sense that

$$\kappa_{\pi,\ell}(\boldsymbol{\alpha}^*_{\boldsymbol{y}}) = 1 \quad \text{iff} \quad \pi(\boldsymbol{\theta} \mid \boldsymbol{\alpha}^*_{\boldsymbol{y}}) \propto \ell(\boldsymbol{\theta})$$

.



Example (Beta–Binomial)

Let $\sum_{i=1}^{n} Y_i \sim \text{Bin}(n, \theta)$, and suppose $\theta \sim \text{Beta}(a, b)$. It can be shown that the max-compatible prior is $\pi(\theta \mid a^*, b^*) = \beta(\theta \mid a^*, b^*)$, where $a^* = 1 + \sum_{i=1}^{n} Y_i$, and $b^* = 1 + n - \sum_{i=1}^{n} Y_i$, so that

$$\widehat{\theta}_n = \arg \max_{\theta \in (0,1)} f(\mathbf{y} \mid \theta) = \bar{Y} = \frac{a^* - 1}{a^* + b^* - 2} =: m(a^*, b^*).$$

Theorem

Let $\mathbf{y} \sim f(\cdot \mid \boldsymbol{\theta})$, and let $\mathscr{P} = \{\pi(\boldsymbol{\theta} \mid \boldsymbol{\alpha}) : \boldsymbol{\alpha} \in \mathscr{A}\}$ be a family of priors for $\boldsymbol{\theta}$. Suppose there exists a max-compatible prior $\pi(\boldsymbol{\theta} \mid \boldsymbol{\alpha}_{\mathbf{y}}^*) \in \mathscr{P}$, which we assume to be unimodal. Then,

$$\widehat{\boldsymbol{\theta}}_n = \arg \max_{\boldsymbol{\theta} \in \Theta} f(\mathbf{y} \mid \boldsymbol{\theta}) = m_\pi(\boldsymbol{\alpha}_{\mathbf{y}}^*) := \arg \max_{\boldsymbol{\theta} \in \Theta} \pi(\boldsymbol{\theta} \mid \boldsymbol{\alpha}_{\mathbf{y}}^*).$$

Example (Exp–Gamma)

In this case the max-compatible prior is given by $f_\Gamma(\theta \mid a^*, b^*)$ where $(a^*, b^*) = (1 + n, \sum_{i=1}^n Y_i)$. The connection with the ML estimator is the following

$$\widehat{\theta} = \arg\max_{\theta \in \Theta} f(\boldsymbol{y} \mid \theta) = \frac{n}{\sum_{i=1}^n Y_i} = \frac{a^* - 1}{b^*} =: m_2(a^*, b^*).$$

Example (Exp–Gamma)

In this case the max-compatible prior is given by $f_\Gamma(\theta \mid a^*, b^*)$ where $(a^*, b^*) = (1 + n, \sum_{i=1}^n Y_i)$. The connection with the ML estimator is the following

$$\widehat{\theta} = \arg\max_{\theta \in \Theta} f(\boldsymbol{y} \mid \theta) = \frac{n}{\sum_{i=1}^n Y_i} = \frac{a^* - 1}{b^*} =: m_2(a^*, b^*).$$

Example (Poisson–Gamma)

In this case the max-compatible prior is $f_\Gamma(\theta \mid a^*, b^*)$, where $(a^*, b^*) = (1 + \sum_{i=1}^n Y_i, n)$. The max-compatible hyperparameter in this case is different from the one in the previous example, but still

$$\widehat{\theta} = \arg\max_{\theta \in \Theta} f(\boldsymbol{y} \mid \theta) = \bar{Y} = \frac{a^* - 1}{b^*} =: m_2(a^*, b^*).$$

- In many situations closed form estimators of κ and $\|\cdot\|$ are not available.
- This leads to considering algorithmic techniques to obtain estimates.
- As most Bayes methods resort to using MCMC methods it would be appealing to express $\kappa_{\cdot,\cdot}$ and $\|\cdot\|$ as functions of posterior expectations and employ MCMC iterates to estimate them.
- For example, $\kappa_{\pi,p}$ can be expressed as

$$\kappa_{\pi,p} = E_p \, \pi(\boldsymbol{\theta}) \left[E_p \left\{ \frac{\pi(\boldsymbol{\theta})}{\ell(\boldsymbol{\theta})} \right\} E_p\{\ell(\boldsymbol{\theta})\pi(\boldsymbol{\theta})\} \right]^{-1/2},$$

where $E_p(\cdot) = \int_\Theta \cdot \, p(\boldsymbol{\theta} \mid \boldsymbol{y}) d\boldsymbol{\theta}$.

- A natural Monte Carlo estimator would then be

$$\hat{\kappa}_{\pi,p} = \frac{1}{B}\sum_{b=1}^{B}\pi(\boldsymbol{\theta}^{b})\left[\frac{1}{B}\sum_{b=1}^{B}\frac{\pi(\boldsymbol{\theta}^{b})}{\ell(\boldsymbol{\theta}^{b})}\frac{1}{B}\sum_{b=1}^{B}\ell(\boldsymbol{\theta}^{b})\pi(\boldsymbol{\theta}^{b})\right]^{-1/2},$$

  where $\boldsymbol{\theta}^{b}$ denotes the bth MCMC iterate of $p(\boldsymbol{\theta} \mid \boldsymbol{y})$.
- Consistency of such an estimator follows trivially by the ergodic theorem and the continuous mapping theorem, but there is an important issue regarding its stability.

- Unfortunately, the previous estimator includes an expectation that contains $\ell(\boldsymbol{\theta})$ in the denominator and therefore (29) inherits the undesirable properties of the so-called harmonic mean estimator (Newton and Raftery, 1994).
- It has been shown that even for simple models this estimator may have infinite variance (Raftery et al. 2007), and has been harshly criticized for, among other things, converging extremely slowly.
- As argued by Wolpert and Schmidler (2012, p. 655):

> *"the reduction of Monte Carlo sampling error by a factor of two requires increasing the Monte Carlo sample size by a factor of $2^{1/\varepsilon}$, or in excess of $2.5 \cdot 10^{30}$ when $\varepsilon = 0.01$, rendering [the harmonic mean estimator] entirely untenable."*

Posterior and Prior Mean-Based Estimators of Compatibility

Solution

- An alternate strategy is to avoid writing $\kappa_{\pi,p}$ as a function of harmonic mean estimators and instead express it as a function of posterior and prior expectations. For example, consider

$$\kappa_{\pi,p} = E_p \, \pi(\boldsymbol{\theta}) \left[\frac{E_\pi\{\pi(\boldsymbol{\theta})\}}{E_\pi\{\ell(\boldsymbol{\theta})\}} \, E_p\{\ell(\boldsymbol{\theta})\pi(\boldsymbol{\theta})\} \right]^{-1/2},$$

  where $E_\pi(\,\cdot\,) = \int_\Theta \cdot \, \pi(\boldsymbol{\theta}) \, \mathrm{d}\boldsymbol{\theta}$.

- Now the Monte Carlo estimator is

$$\tilde{\kappa}_{\pi,p} = \frac{1}{B} \sum_{b=1}^{B} \pi(\boldsymbol{\theta}^b) \left\{ \frac{B^{-1}\sum_{b=1}^{B} \pi(\boldsymbol{\theta}_b)}{B^{-1}\sum_{b=1}^{B} \ell(\boldsymbol{\theta}_b)} \frac{1}{B} \sum_{b=1}^{B} \ell(\boldsymbol{\theta}^b)\pi(\boldsymbol{\theta}^b) \right\}^{-1/2},$$

  where $\boldsymbol{\theta}_b$ denotes the bth draw of $\boldsymbol{\theta}$ from $\pi(\boldsymbol{\theta})$, which can also be sampled within the MCMC algorithm.

Posterior and Prior Mean-Based Estimators of Compatibility

Illustration

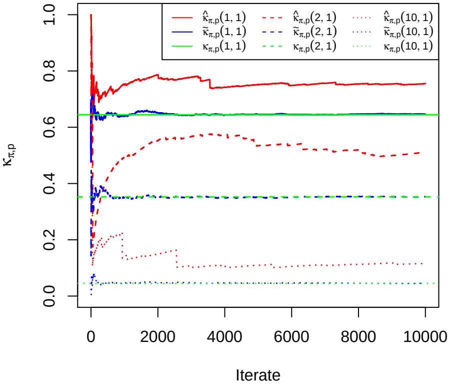


Figure: Running point estimates of prior-posterior compatibility, $\kappa_{\pi,p}$, for the on-the-job drug usage toy example. Green lines correspond to the true $\kappa_{\pi,p}$ values, blue represents $\tilde{\kappa}_{\pi,p}$ and red denotes $\hat{\kappa}_{\pi,p}$.

Proposition

The following equalities hold:

$$\|p\|^2 = \frac{E_p\{\ell(\boldsymbol{\theta})\pi(\boldsymbol{\theta})\}}{E_\pi\,\ell(\boldsymbol{\theta})}, \quad \|\pi\|^2 = E_\pi\,\pi(\boldsymbol{\theta}), \quad \|\ell\|^2 = E_\pi\,\ell(\boldsymbol{\theta})\,E_p\left\{\frac{\ell(\boldsymbol{\theta})}{\pi(\boldsymbol{\theta})}\right\},$$

$$\kappa_{\pi_1,\pi_2} = E_{\pi_1}\,\pi_2(\boldsymbol{\theta})\Big[E_{\pi_1}\,\pi_1(\boldsymbol{\theta})\,E_{\pi_2}\,\pi_2(\boldsymbol{\theta})\Big]^{-1/2}, \quad \kappa_{\pi,\ell} = E_\pi\,\ell(\boldsymbol{\theta})\left[E_\pi\,\pi(\boldsymbol{\theta})\,E_\pi\,\ell(\boldsymbol{\theta})\,E_p\left\{\frac{\ell(\boldsymbol{\theta})}{\pi(\boldsymbol{\theta})}\right\}\right]^{-1/2},$$

$$\kappa_{\pi,p} = E_p\,\pi(\boldsymbol{\theta})\left[\frac{E_\pi\,\pi(\boldsymbol{\theta})}{E_\pi\,\ell(\boldsymbol{\theta})}\,E_p\,\{\ell(\boldsymbol{\theta})\pi(\boldsymbol{\theta})\}\right]^{-1/2}, \quad \kappa_{\ell,p} = E_p\,\ell(\boldsymbol{\theta})\left[E_p\left\{\frac{\ell(\boldsymbol{\theta})}{\pi(\boldsymbol{\theta})}\right\}E_p\,\{\ell(\boldsymbol{\theta})\pi(\boldsymbol{\theta})\}\right]^{-1/2},$$

$$\kappa_{\ell_1,\ell_2} = E_\pi\,\ell_2(\boldsymbol{\theta})\,E_{p_2}\left\{\frac{\ell_1(\boldsymbol{\theta})}{\pi(\boldsymbol{\theta})}\right\}\left[E_\pi\{\ell_1(\boldsymbol{\theta})\}E_{p_1}\left\{\frac{\ell_1(\boldsymbol{\theta})}{\pi(\boldsymbol{\theta})}\right\}E_\pi\,\ell_2(\boldsymbol{\theta})\,E_{p_2}\left\{\frac{\ell_2(\boldsymbol{\theta})}{\pi(\boldsymbol{\theta})}\right\}\right]^{-1/2}.$$

Draft

Conditionally accepted, *Bayesian Analysis*

Bayesian Analysis (2018) **00**, Number 0, pp. 1

On the Geometry of Bayesian Inference

Miguel de Carvalho*, Garritt L. Page†, and Bradley J. Barney†

Abstract. We provide a geometric interpretation to Bayesian inference that allows us to introduce a natural measure of the level of agreement between priors, likelihoods, and posteriors. The starting point for the construction of our geometry is the observation that the marginal likelihood can be regarded as an inner product between the prior and the likelihood. A key concept in our geometry is that of compatibility, a measure which is based on the same construction principles as Pearson correlation, but which can be used to assess how much the prior agrees with the likelihood, to gauge the sensitivity of the posterior to the prior, and to quantify the coherency of the opinions of two experts. Estimators for all the quantities involved in our geometric setup are discussed, which can be directly computed from the posterior simulation output. Some examples are used to illustrate our methods, including data related to on-the-job drug usage, midge wing length, and prostate cancer.

Keywords: Bayesian inference, Geometry, Hellinger affinity, Hilbert space, Marginal likelihood.

Miguel.deCarvalho@ed.ac.uk

- We discussed a natural geometric framework to Bayesian inference which motivated a simple, intuitively appealing measure of the agreement between priors, likelihoods, and posteriors: compatibility (κ).
- In this geometric framework, we also discuss a related measure of the "informativeness" of a distribution, $\|\cdot\|$.
- We developed MCMC-based estimators of these metrics that are easily computable and, by avoiding the estimation of harmonic means, are reasonably stable.
- Our concept of compatibility can be used to evaluate how much the prior agrees with the likelihood, to measure the sensitivity of the posterior to the prior, and to quantify the level of agreement of elicited priors.

- To streamline the talk, I have focused on priors which are on $L_2(\Theta)$.
- Yet there are examples of priors that are not in $L_2(\Theta)$. A simple example is that of the Jeffreys prior for the Beta-Binomial, Beta$(1/2, 1/2)$, whose norm will be infinity.
- Our geometric construction is still able to consider densities not in $L_2(\Theta)$, because as documented in the paper, the following approach is still nested in our setup

$$\kappa_{\sqrt{g},\sqrt{h}} = \int_{\Theta} g(\boldsymbol{\theta})^{1/2} h(\boldsymbol{\theta})^{1/2} \, \mathrm{d}\boldsymbol{\theta}.$$

- To streamline the talk, I have focused on priors which are on $L_2(\Theta)$.
- Yet there are examples of priors that are not in $L_2(\Theta)$. A simple example is that of the Jeffreys prior for the Beta-Binomial, Beta$(1/2, 1/2)$, whose norm will be infinity.
- Our geometric construction is still able to consider densities not in $L_2(\Theta)$, because as documented in the paper, the following approach is still nested in our setup

$$\kappa_{\sqrt{g},\sqrt{h}} = \int_\Theta g(\boldsymbol{\theta})^{1/2} h(\boldsymbol{\theta})^{1/2} \, \mathrm{d}\boldsymbol{\theta}.$$

- This approach results in a κ that continues being a metric that measures agreement between two elements of a geometry, but loses direct connection with Bayes theorem.

The End

Thanks!

MLNet

Agarawal, A., and Daumé, III, H. (2010), "A Geometric View of Conjugate Priors," *Machine Learning*, 81, 99–113.

Aitchison, J. (1971), "A Geometrical Version of Bayes' Theorem," *The American Statistician*, 25, 45–46.

Al Labadi, L., and Evans, M. (2016), "Optimal Robustness Results for Relative Belief Inferences and the Relationship to Prior-Data Conflict," *Bayesian Analysis*, in press.

Bartle, R., and Sherbert, D. (2010), *Introduction to Real Analysis* (4th ed.), New York: Wiley.

Berger, J. (1991), "Robust Bayesian Analysis: Sensitivity to the Prior," *Journal of Statistical Planning and Inference*, 25, 303–328.

Berger, J., and Berliner, L. M. (1986), "Robust Bayes and Empirical Bayes Analysis with ε-Contaminated Priors," *The Annals of Statistics*, 14, 461–486.

Berger, J. O., and Wolpert, R. L. (1988), *The Likelihood Principle*, In *IMS Lecture Notes*, Ed. Gupta, S. S., Institute of Mathematical Statistics, vol. 6.

Christensen, R., Johnson, W. O., Branscum, A. J., and Hanson, T. E. (2011), *Bayesian Ideas and Data Analysis*, Boca Raton: CRC Press.

Cheney, W. (2001), *Analysis for Applied Mathematics*, New York: Springer.

Evans, M., and Jang, G. H. (2011), "Weak Informativity and the Information in one Prior Relative to Another," *Statistical Science*, 26, 423–439.

Evans, M., and Moshonov, H. (2006), "Checking for Prior-Data Conflict," *Bayesian Analysis*, 1, 893–914.

Gelman, A., Jakulin, A., Pittau, M. G., and Su, Y. S. (2011), "A Weakly Informative Default Prior Distribution for Logistic and other Regression Models," *Annals of Applied Statistics*, 2, 1360–1383.

Giné, E., and Nickl, R. (2008), "A Simple Adaptive Estimator of the Integrated Square of a Density," *Bernoulli*, 14, 47–61.

Grogan, W., and Wirth, W. (1981), "A New American Genus of Predaceous Midges Related to

Palpomyia and Bezzia (Diptera: Ceratopogonidae)," *Proceedings of the Biological Society of Washington*, 94, pp. 1279–1305.

Hastie, T., Tibshirani, R., and Friedman, J. (2008), *Elements of Statistical Learning*, New York: Springer.

Hoff, P. (2009), *A First Course in Bayesian Statistical Methods*, New York: Springer.

Hunter, J., and Nachtergaele, B. (2005), *Applied Analysis*, London: World Scientific Publishing.

Kyung, M., Gill, J., Ghosh, M., and Casella, G. (2010), "Penalized Regression, Standard Errors, and Bayesian Lassos," *Bayesian Analysis*, 5, 369–412.

Knight, K. (2000), *Mathematical Statistics*, Boca Raton: Chapman & Hall/CRC Press.

Lavine, M. (1991), "Sensitivity in Bayesian Statistics: The Prior and the Likelihood," *Journal of the American Statistical Association* **86** 396–399.

Lenk, P. (2009), "Simulation Pseudo-Bias Correction to the Harmonic Mean Estimator of Integrated Likelihoods," *Journal of Computational and Graphical Statistics*, 18, 941–960.

Lopes, H. F., and Tobias, J. L. (2011), "Confronting Prior Convictions: On Issues of Prior Sensitivity and Likelihood Robustness in Bayesian Analysis," *Annual Review of Economics*, 3, 107–131.

Millman, R. S., and Parker, G. D. (1991), *Geometry: A Metric Approach with Models*, New York: Springer.

Newton, M. A., and Raftery, A. E. (1994), "Approximate Bayesian Inference with the Weighted Likelihood Bootstrap (With Discussion)," *Journal of the Royal Statistical Society,* Series B, 56, 3–26.

Pajor, A., and Osiewalski, J. (2013), "A Note on Lenk's Correction of the Harmonic Mean Estimator," *Central European Journal of Economic Modelling and Econometrics*, 5, 271–275.

Park, T., and Casella, G. (2008), "The Bayesian Lasso," *Journal of the American Statistical Association*, 103, 681–686.

Raftery, A. E., Newton, M. A., Satagopan, J. M., and Krivitsky, P. N. (2007), "Estimating the

Integrated Likelihood via Posterior Simulation using the Harmonic Mean Identity," In *Bayesian Statistics*, Eds. Bernardo, J. M., Bayarri, M. J., Berger, J. O., Dawid, A. P., Heckerman, D., Smith, A. F. M., and West, M., Oxford University Press, vol. 8.

Ramsey, J. O., and Silverman, B. W. (1997), *Functional Data Analysis*, New York: Springer-Verlag.

Scheel, I., Green, P. J., and Rougier, J. C. (2011), "A Graphical Diagnostic for Identifying Influential Model Choices in Bayesian Hierarchical Models," *Scandinavian Journal of Statistics*, 38, 529–550.

Shortle, J. F., and Mendel, M. B. (1996), "The Geometry of Bayesian Inference," In *Bayesian Statistics*, eds. Bernardo, J. M., Berger, J. O., Dawid, A. P., and Smith, A. F. M., Oxford University Press, vol. 5, pp. 739–746.

Walter, G., and Augustin, T. (2009), "Imprecision and Prior-Data Conflict in Generalized Bayesian Inference," *Journal of Statistical Theory and Practice*, 3, 255–271.

Wolpert, R., and Schmidler, S. (2012), "α-Stable Limit Laws for Harmonic Mean Estimators of Marginal Likelihoods," *Statistica Sinica*, 22, 655–679.

Zhu, H., Ibrahim, J. G., and Tang, N. (2011), "Bayesian Influence Analysis: A Geometric Approach," *Biometrika*, 98, 307–323.