On the geometry of learning from data: Bayes meets Hilbert

by **Miguel de Carvalho***

Abstract.—Bayes' theorem is a central result of Statistics and related fields, such as Artificial Intelligence and Machine Learning. In this note, we offer a gentle introduction to a geometric interpretation of Bayesian inference that allows one to think of priors, likelihoods, and posteriors as vectors in an Hilbert space. The given framework can be conceptualized as a geometry of learning from data, and it can be used to construct measures of agreement between these vectors. Conceptually, the geometry is tantamount to that of Pearson correlation, but where an inner product is considered over the parameter space—rather than over the sample space.

1 Introduction

This note builds on ideas from two prominent thinkers: Thomas Bayes (c. 1701–1761) and David Hilbert (1862–1943).[1] While their lives never overlapped temporally, this note shows how the work of Hilbert can be used to reinterpret Bayes' theorem and Bayesian inference from a geometric viewpoint—as well as other key statistical concepts on what we regard as a geometry of learning from data.

The Bayesian paradigm is a well-known statistical inference approach that can be used for learning from data about a parameter of statistical interest using Bayes theorem. Let $Y_1, \ldots, Y_n$ be a sequence of independent and identically distributed (iid) random variables in a measurable space $(\Omega, \mathscr{A})$ that are drawn from parametric density function

$$f_\theta(y) \equiv f(y \mid \theta),$$

with $y \in \Omega$ and $\theta \in \Theta$. The sets Ω and Θ are respectively known as *sample space* and *parameter space*.

The key goal of Bayesian inference is to learn about the distribution of the parameter θ given the data $y = (y_1, \ldots, y_n)$. It follows from Bayes theorem that,

$$p(\theta \mid y) = \frac{\pi(\theta)\ell(\theta)}{\int_\Theta \pi(u)\ell(u)\,\mathrm{d}u}. \tag{1}$$

where $\ell(\theta) = \prod_{i=1}^n f_\theta(y_i)$ is the likelihood function, and $\pi(\theta)$ is the prior density function. The density $p(\theta \mid y)$ is known as posterior density and it summarizes what we learn about θ after observing y.

The prior density can understood as a way adding prior knowledge about θ to the analysis—say, from an expert opinion, from a census, and so on—or simply as a way to "initiate the inferential machine." Quoting [9]:

> The choice of a prior distribution is necessary (as you would need to initiate the inferential machine) but there is no notion of the "optimal" prior distribution. Choosing a prior distribution is similar in principle to initializing any other sequential procedure (e.g., iterative optimization methods […] etc.). The choice of such initialization can be good or bad in the sense of the rate of convergence of the procedure to its final value, but as long as the procedure is guaranteed to converge, the choice of prior does not have a permanent impact.

[1] The key concepts and methods from this note relate with the ideas and principles in [3], which was awarded with the 2018 Lindley Prize from the International Society of Bayesian Analysis.

* School of Mathematics, University of Edinburgh; Department of Mathematics, Universidade de Aveiro
Email: Miguel.deCarvalho@ed.ac.uk

Preprint Version

CIM Bulletin (2024, to appear; also published at *Bernoulli News* with permission from the Editors)

And indeed, the posterior can be shown to converge to the true value, under rather general conditions on the prior distribution—a result known in statistical parlance as the Bernstein—von Mises theorem [11, Theorem 10.1].

The remainder of this note is organized as follows. In §2 we note that there's an hidden geometry underlying Eq. (4) that can be used to rethink Bayesian inference and to develop measures of agreement between prior, likelihood, and posterior. In §3 we illustrate how that geometry can be used for shedding light on other statistical inference concepts.

Before we get started a disclaimer is in order. To make the presentation of the key ideas more accessible, we will often use visualizations based on Cartesian representations. Yet, it is important to remember that these representations are mainly heuristic and hence should be interpreted with care.

2 The geometry of Bayesian inference

2.1 Abstract geometry

We first clarify the sense in which the term geometry will be used throughout this note. The following definition of abstract geometry can be found in [7, p. 17].

Definition 1 (Abstract geometry).— 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 $\mathscr{P} = L_2(\Theta)$ is the the space of square integrable functions, and the set of all lines is

$$\mathscr{L} = \{g + kh : g, h \in L_2(\Theta), k \in \mathbf{R}\}. \qquad (2)$$

Hence, in our setting points can be, for example, prior densities, posterior densities, or likelihoods, as long as they are in $L_2(\Theta)$. While not all priors and likelihoods are in $L_2(\Theta)$, the framework discussed herein may extend beyond $L_2(\Theta)$ with some modifications, while still allowing similar geometric interpretations as the ones provided below. See [3, §3] for details.

2.2 Bayes geometry

2.2.1 *The marginal likelihood is an inner product*

Suppose the goal of the inference is over a parameter θ which takes values on $\Theta \subseteq \mathbf{R}^p$. We use the geometry of the Hilbert space $\mathscr{H} = (L_2(\Theta), \langle \cdot, \cdot \rangle)$, with inner-product[2]

$$\langle g, h \rangle = \int_\Theta g(\theta) h(\theta) \, \mathrm{d}\theta, \quad g, h \in L_2(\Theta). \qquad (3)$$

Adopting the geometric terminology used in linear spaces, we denote the elements of $L_2(\Theta)$ as vectors, and assess their *magnitudes* through the use of the norm induced by the inner product in (3), i.e., $\| \cdot \| = (\langle \cdot, \cdot \rangle)^{1/2}$.

The starting point for constructing our geometry is the observation that Bayes theorem can be written using the inner-product in (2.2.1) as follows

$$p(\theta \mid y) = \frac{\pi(\theta) \ell(\theta)}{\langle \pi, \ell \rangle}, \qquad (4)$$

where $\langle \pi, \ell \rangle = \int_\Theta f(y \mid \theta) \pi(\theta) \, \mathrm{d}\theta$ is the so-called marginal likelihood. The inner product in (3) naturally leads to considering π and ℓ that are in $L_2(\Theta)$, which is compatible with a wealth of parametric models and proper priors.

As can be seen from Fig. 1, by considering p, π, and ℓ as vectors with different magnitudes and directions, Bayes' theorem essentially describes the method of reshaping the prior vector in order to derive the posterior vector. The likelihood vector amplifies or diminishes the magnitude of the prior vector, and appropriately adjusts its direction, in a way that will be clearly defined in the subsequent discussion.

The marginal likelihood $\langle \pi, \ell \rangle$ is simply the inner product between the likelihood and the prior, and thus can be interpreted as an assessment of the concordance between the prior and the likelihood. To provide a more tangible understanding, let's define the *angle measure* between the prior and the likeli-

[2] In mathematical terminology, the assertion that $\mathscr{H}$ constitutes a Hilbert space is frequently referred to as the Riesz–Fischer theorem. For a proof see [2, p. 411].

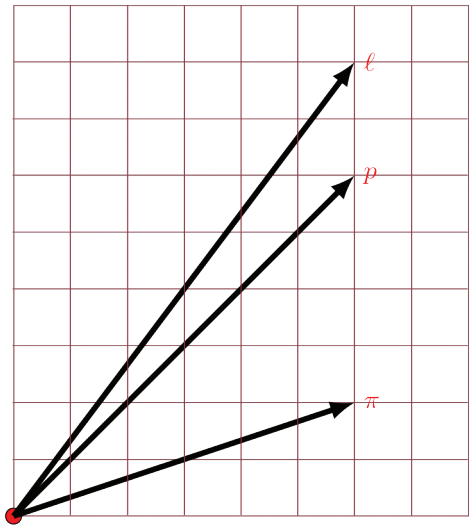

Figure 1.—Cartesian representation of vectors of interest in a Bayesian analysis.

hood as

$$\pi \angle \ell = \arccos \frac{\langle \pi, \ell \rangle}{\|\pi\| \|\ell\|}. \tag{5}$$

Since π and ℓ are nonnegative, the angle between the prior and the likelihood can only be acute or right, i.e., $\pi \angle \ell \in [0, 90^\circ]$. The closer $\pi \angle \ell$ is to 0°, the greater the agreement between the prior and the likelihood. Conversely, the closer $\pi \angle \ell$ is to 90°, the greater the disagreement between prior and likelihood. In the limiting case where $\pi \angle \ell = 90^\circ$—which implies the prior and the likelihood have all of their mass on disjoint sets—we say that the prior is orthogonal to the likelihood. Bayes theorem does not allow for a prior to be orthogonal to the likelihood as $\pi \angle \ell = 90^\circ$ implies that $\langle \pi, \ell \rangle = 0$, thus yielding a division by zero in (4).

2.2.2 *Compatibility*

The object we aim to focus next is given by a standardized inner product

$$\kappa_{\pi,\ell} = \frac{\langle \pi, \ell \rangle}{\|\pi\| \|\ell\|}. \tag{6}$$

The quantity $\kappa_{\pi,\ell} \in (0, 1]$ assesses the extent to which an expert's viewpoint aligns with the data, thereby offering an intuitive measurement of the concordance between the prior and the data.

Extending the principle in (6), for any two points in the geometry under consideration we define their compatibility as a standardized inner product.

Definition 2 (Compatibility).— The compatibility between points in the geometry under consideration is defined as

$$\kappa_{g,h} = \frac{\langle g, h \rangle}{\|g\| \|h\|}, \quad g, h \in L_2(\Theta). \tag{7}$$

Particular instances include (6) as well as:

- κ_{π_1,π_2}: which assesses the level of agreement between two experts, with respective priors π_1 and π_2.
- $\kappa_{\pi,p}$: which is a metric of the sensitivity of the posterior to the prior specification.

Example 1 (Beta-Bernoulli model).— Let

$$\begin{cases} Y_i \mid \theta \overset{\text{iid}}{\sim} \text{Bern}(\theta), & i = 1, \ldots, n, \\ \theta \sim \text{Beta}(a, b). \end{cases} \tag{8}$$

Then, $\theta \mid y \sim \text{Beta}(a^\star, b^\star)$ with $a^\star = n_1 + a$ and $b^\star = n - n_1 + b$, where $n_1 = \sum_{i=1}^n y_i$.

The compatibility between prior and likelihood for this beta–Bernoulli model is

$$\kappa_{\pi,\ell} = \frac{B(a^\star, b^\star)}{\{B(2a-1, 2b-1) B(2n_1 + 1, 2(n-n_1)+1)\}^{1/2}},$$

for $a, b > 1/2$, with $B(a, b) = \int_0^1 u^{a-1}(1-u)^{b-1}\, du$.[3] To assess how compatible the priors $\pi_1 \sim \text{Beta}(a_1, b_1)$ and $\pi_2 \sim \text{Beta}(a_2, b_2)$ are, we obtain

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

for $a_1, a_2, b_1, b_2 > 1/2$.

[3] The geometry underlying compatibility can be reframed within an Hellinger affinity context so to allow for any $a, b > 0$. See [3, §3].

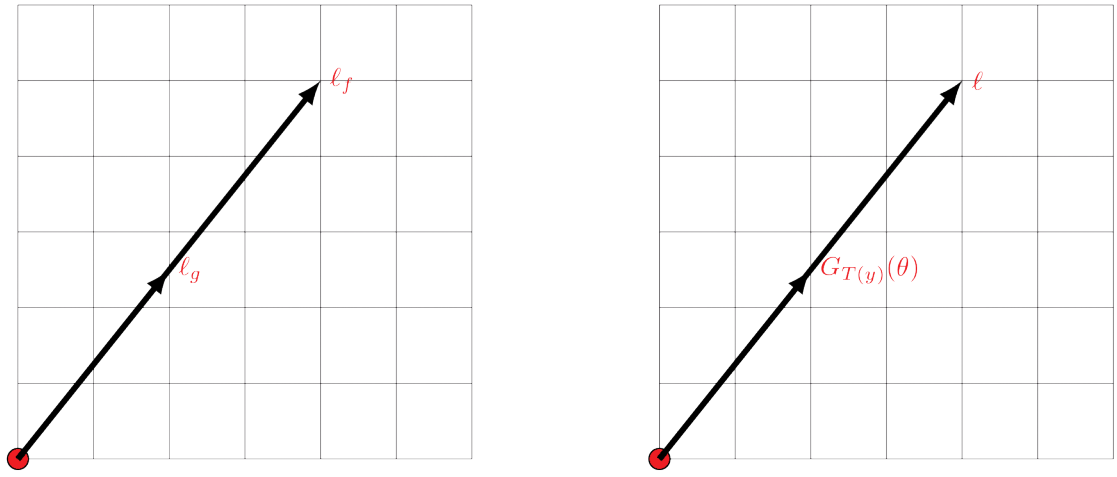

Figure 2.—Cartesian representation underlying the strong likelihood principle (left) and sufficiency (right). See §§ 3.2 and 3.3.

3 Further perspectives and insights

The roadmap for this section is as follows. §3.1 notes that a variational representation of the posterior density naturally fits our geometry. §§3.2 and 3.3 are related with collinearity; it follows from §2, whenever the symbol "$\propto$" is used in a Bayesian setting it simply implies that two likelihoods, priors or posteriors are collinear. Finally, §3.4 notes the similarities between the geometry of compabitility and that of Pearson correlation.

3.1 Donsker–Varadhan representation

The celebrated Donsker–Varadhan representation shows that the posterior density is the solution to a variational problem with search domain $\mathscr{P}(\Theta)$; here and below, $\mathscr{P}(\Theta)$ is the space of probability density functions that can be defined over Θ and $l(\theta) = \log \ell(\theta)$ is the log likelihood. Specifically, the Donsker–Varadhan representation is given by

$$p(\theta \mid y) = \arg \min_{q \in \mathscr{P}(\Theta)} [-\mathrm{E}_q\{l(\theta)\} + \mathrm{KL}(q, \pi)], \quad (9)$$

where E_q and KL are respectively the prior expectation and Kullback–Leibler divergence, that is,

$$E_q\{l(\theta)\} = \int_\Theta l(\theta)\, q(\theta)\, \mathrm{d}\theta,$$
$$\mathrm{KL}(q, \pi) = \int_\Theta q(\theta) \log\{q(\theta)/\pi(\theta)\}\, \mathrm{d}\theta.$$

A geometric interpretation of (3.1) follows from elementary properties of inner products,

$$\begin{aligned} p(\theta \mid y) &= \arg \min_{q \in \mathscr{P}(\Theta)} -\langle q, l\rangle + \langle q, \log(q/\pi)\rangle \\ &= \arg \max_{q \in \mathscr{P}(\Theta)} \langle q, l\rangle - \langle q, \log(q/\pi)\rangle \\ &= \arg \max_{q \in \mathscr{P}(\Theta)} \langle q, \mathrm{DV}_q\rangle, \end{aligned} \quad (10)$$

where DV_q is what we refer to as the *Donsker–Varadhan likelihood ratio*,

$$\mathrm{DV}_q(\theta) \equiv \log[\ell(\theta)/\{q(\theta)\pi(\theta)\}]. \quad (11)$$

Loosely, (10) implies that the posterior density is the density in $\mathscr{P}(\Theta)$ which is most lined up with the Donsker–Varadhan likelihood ratio in (11).

3.2 Collinearity, I: likelihood principle

Let ℓ_f and ℓ_g be the likelihoods based on observing $y \sim f$ and $y^* \sim g$, respectively. The strong likelihood principle states that if

$$\ell_f(\theta) = f(\theta \mid y) \propto g(\theta \mid y^*) = \ell_g(\theta),$$

then the same inference should be drawn from both samples. According to our geometry, this means that likelihoods with the same direction yield the same inference. For instance, the Bernoulli likelihood of the model from Example (1) is

$$\ell_f(\theta) = \prod_{i=1}^n \theta^{y_i}(1-\theta)^{n-y_i} = \theta^{\sum_{i=1}^n y_i}(1-\theta)^{n-\sum_{i=1}^n y_i},$$

wheras that of the Binomial model for $n_1 = \sum_{i=1}^n y_i$ is

$$\ell_g(\theta) = \binom{n}{n_1}\theta^{n_1}(1-\theta)^{n-n_1},$$

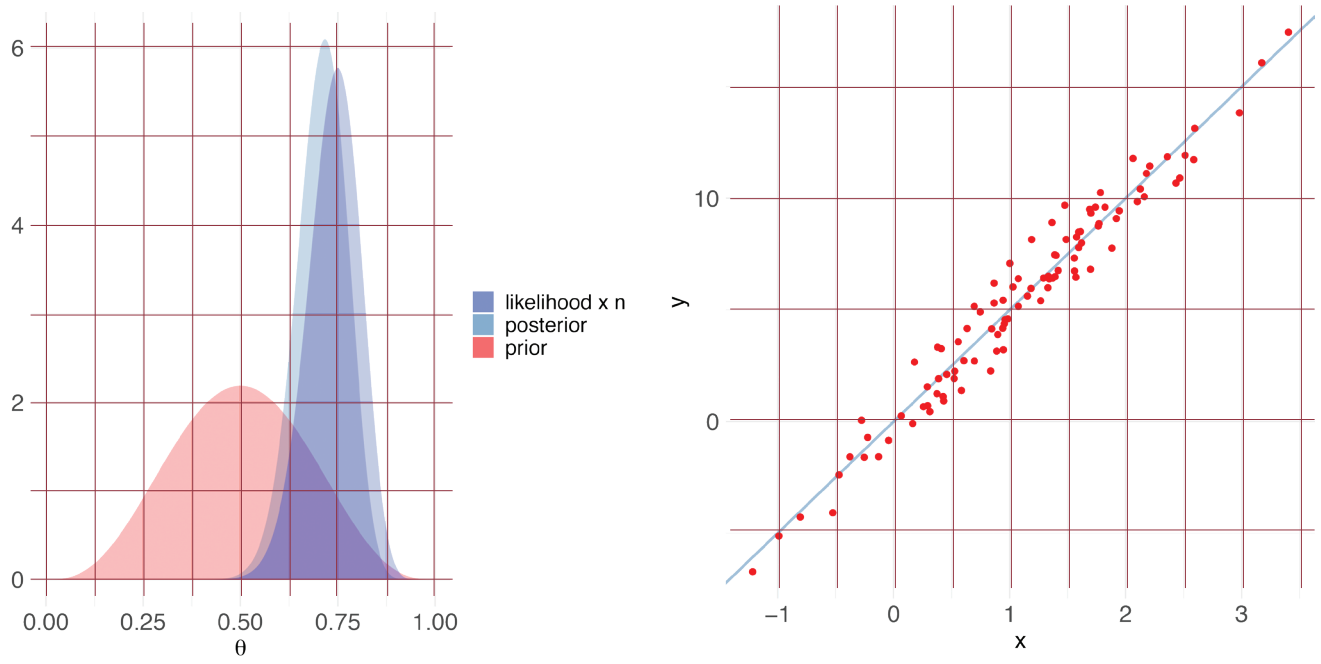

Figure 3.—Left: Prior, posterioar, and likelihood for beta–binomial specification from Example 1 with $(a,b) = (4, 4)$, $n = 40$, and $n_1 = 30$ so that, for example, $\kappa_{\pi,l} = 0.41$. Right: Simulated data from bivariate normal distribution with $\rho_{X,Y} = 0.98$.

with $\binom{a}{b}$ denoting the binomial coefficient. Trivially,

$$\ell_f(\theta) \propto \ell_g(\theta),$$

and hence ℓ_f and ℓ_g are collinear.

3.3 Collinearity, II: sufficiency

Roughly speaking, a sufficient statistic is one that contains all the information that is required to learn about θ.[4] The geometry from §2.2 can also be used to rethink a celebrated characterization of sufficient statistics in a geometric fashion.

Theorem 3 (Neyman factorization).— Suppose that $Y = (Y_1, \dots, Y_n)$ has a joint density function or a frequency function $f_\theta(y)$. Then $T(Y)$ is sufficient for θ iff there exists a function of that statistic, $G_{T(y)}(\theta)$, that is collinear to $\ell(\theta)$, that is,

$$\ell(\theta) \propto G_{T(y)}(\theta).$$

See, for instance, [6, §4] for a nongeometrical formulation of this classical result. Let's illustrate this on a well-known example.

Example 2.— Let $Y_1, \dots, Y_n \overset{\text{iid}}{\sim} \text{Uniform}(0, \theta)$. It can be easily shown that

$$\ell(\theta) = \prod_{i=1}^{n} \frac{1}{\theta} \mathbf{1}_{[0,\theta]}(y_i) \propto \frac{1}{\theta^n} \mathbf{1}_{[0,\theta]}\{T(y)\} \equiv G_{T(y)}(\theta),$$

where $T(y) = \max\{y_1, \dots, y_n\}$ and $\mathbf{1}_A$ is the indicator function.

3.4 Compatibility vs Pearson correlation

Compatibility in Definition 2 follows the same construction principles as the Pearson correlation coefficient, which is based on the inner product

$$\langle X, Y \rangle = \int_\Omega XY \, \mathrm{d}P, \quad X, Y \in L_2(\Omega, \mathbb{B}_\Omega, P), \quad (12)$$

instead of the inner product in (3). Recall that Pearson correlation is defined as

$$\rho_{X,Y} = \frac{\text{cov}(X, Y)}{\text{sd}(X)\,\text{sd}(Y)},$$

and it can be understood as a cosine of $X \angle Y$ in a similar fashion as (5)—but with "cov" and "sd" denoting the covariance (inner product) and standard deviation (norm), respectively. And indeed, just like the cosine function, $\rho_{X,Y} \in [-1, 1]$.

Compatibility is however defined for priors, posteriors, and likelihoods in $L_2(\Theta)$ equipped with the inner product (3), whereas Pearson correlation works with random variables in $L_2(\Omega, \mathbb{B}_\Omega, P)$ equipped with the inner product (12).

Fig. 3 sheds light on the different uses of compatibility and Pearson correlation. For example, $\kappa_{\pi,\ell}$ mea-

[4] Recall that a statistic $T = T(Y)$ is sufficient for θ if, $P(Y \in A \mid T = t)$ does not depend on θ, for all t in the range of T and for all sets A.

sures the agreement between likelihood and prior density, whereas $\rho_{X,Y}$ assesses the degree of linear association between random variables X and Y. The value $\kappa_{\pi,\ell} = 0.41$ is in line with the moderate overlap between prior and likelihood visible in Fig. 3. The value of $\rho_{X,Y} = 0.98$ is in line with the strong positive association between the random variables X and Y that can be seen in Fig. 3.

4 Closing remarks

This note offers a gentle introduction to geometrical aspects underlying the Bayesian paradigm that can be used for defining metrics of agreement between priors, likelihoods and posteriors as well as to rethink other concepts and results related with learning from data.

Geometrical interpretations are commonplace in Statistics and related fields—including for example that of Pearson correlation [15], least squares and LASSO (Least Absolute Shrinkage and Selection Operator) [10], and information geometry [1]; also, the geometry of multivariate analysis is well-known [13]. Many well-known geometrical insights concentrate on the *geometry of data* itself, whereas the focus of this note has been on the *geometry of learning from data*. Despite the long tradition of geometrical interpretations of statistical concepts, the view of the Bayesian paradigm along the lines of this note is relatively novel and it has been pioneered by [3] and [5].

Beyond geometry, topology and algebra hava also recently introduced a variety of insights and novel paradigms to the practice of learning from data—leading to the fields of topological data analysis [12] and algebraic statistics [4, 14].

Finally, we note that the geometrical view of the Donsker–Varadhan representation in (10) consists of a variational maximum inner product problem, and that nonvariational versions of such problems are of interest in the Machine Learning literature [8].

Acknowledgements

I am thankful to the Editor for the invitation. This note has benefited from the input and insight of a variety of discussants, including B. Barney, S. Beentjes, V. Inácio, G. Page, and V. Palacios.

References

[1] S. I. Amari, *Information Geometry and its Applications*, Springer, New York, 2016.

[2] W. Cheney, *Analysis for Applied Mathematics* Springer, New York, 2001

[3] M. de Carvalho, G. Page and B. Barney, On the geometry of Bayesian inference, *Bayesian Anal.* **14** (2019) 1013–1036.

[4] M. Drton, B. Sturmfel and S. Sullivant, *Lectures on Algebraic Statistics*, Springer, New York, 2009

[5] S. Kurtek and K. Bharath, Bayesian sensitivity analysis with the Fisher–Rao metric, *Biometrika* **102** (2015) 601–616.

[6] K. Knight, *Mathematical Statistics*, Chapman & Hall/CRC, Boca Raton, 2001.

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

[8] S. Mussmann and S. Ermon, Learning and inference via maximum inner product search, *Proceedings of Machine Learning Research* **48** (2016) 2587–2596.

[9] B. J. Reich and S. K. Ghosh, *Bayesian Statistical Methods*, Chapman & Hall/CRC, Boca Raton, 2019.

[10] J. Taylor, The geometry of least squares in the 21st century, *Bernoulli* **19** (2013) 1449–1464.

[11] A. W. Van der Vaart, *Asymptotic Statistics*, Cambridge University Press, Cambridge UK, 1998.

[12] L. Wasserman, Topological data analysis, *Annual Review of Statistics and Its Application* **5** (2018) 501–532.

[13] T. D. Wickens, *The Geometry of Multivariate Statistics*, Psychology Press, New York & London, 2014.

[14] S. Watanabe, *Algebraic Geometry and Statistical Learning Theory*, Cambridge University Press, Cambridge UK, 2009.

[15] D. Williams, *Probability with Martingales*, Cambridge University Press, Cambridge UK, 1991.