

Article

# How the Choice of Distance Measure Influences the Detection of Prior-Data Conflict

Kimberley Lek <sup>1,\*</sup>  and Rens Van De Schoot <sup>1,2</sup>

<sup>1</sup> Department of Methods and Statistics, Utrecht University, 3584 CH 14 Utrecht, The Netherlands; a.g.j.vandeschoot@uu.nl

<sup>2</sup> Optentia Research Program, Faculty of Humanities, North-West University, Vanderbijlpark 1900, South Africa

\* Correspondence: k.m.lek@uu.nl; Tel.: +31-30-253-9237

Received: 28 March 2019; Accepted: 23 April 2019; Published: 29 April 2019



**Abstract:** The present paper contrasts two related criteria for the evaluation of prior-data conflict: the Data Agreement Criterion (DAC; Bousquet, 2008) and the criterion of Nott et al. (2016). One aspect that these criteria have in common is that they depend on a distance measure, of which dozens are available, but so far, only the Kullback-Leibler has been used. We describe and compare both criteria to determine whether a different choice of distance measure might impact the results. By means of a simulation study, we investigate how the choice of a specific distance measure influences the detection of prior-data conflict. The DAC seems more susceptible to the choice of distance measure, while the criterion of Nott et al. seems to lead to reasonably comparable conclusions of prior-data conflict, regardless of the distance measure choice. We conclude with some practical suggestions for the user of the DAC and the criterion of Nott et al.

**Keywords:** prior-data conflict; distance measure; Kullback-Leibler; data agreement criterion

## 1. Introduction

Any Bayesian model consists of at least two ingredients: a prior and a sampling model for the observed data. The prior contains the a priori beliefs about the parameter(s) and can, for instance, be based on expert knowledge, if elicitation methods are used [1–6]. Consecutive inferences based on the Bayesian model are built on the assumption that the (expert) prior is appropriate for the collected data. In case of a prior data conflict, this assumption is violated. In such a conflict, the prior primarily favors regions of the parameter space that are far from the data mass. Such a conflict can be caused by the unfortunate collection of a rare data set, a problem with the prior, or both of these factors [7]. To avoid erroneous inferences, checking for prior-data conflict should be ‘part of good statistical practice’ ([8], p. 894).

In the current paper, we focus on two checks, developed for the detection of prior-data conflict. First, Bousquet (see [8]; see also [9]) suggested that the expert prior be compared with a non-informative, reference prior in his “Data Agreement Criterion” (DAC). When the expert prior has a larger distance to a reference posterior—a posterior based on the data and the non-informative, reference prior—than the non-informative prior, it can be concluded that the expert and data are in conflict. Second, Nott et al. (see [10]) suggested that the distance between the expert prior and resulting posterior be measured directly. According to this method, the expert and data are in conflict when this distance is surprising in relation to the expert’s prior predictive distribution (see also [11]).

One thing that the DAC [7] and the criterion of Nott et al. [10] have in common is that they depend on a distance measure. Currently, both the DAC and the Nott et al. criteria rely on the Kullback-Leibler divergence (heretofore abbreviated as KL). One of the advantages of the KL is that it

has an intuitive interpretation—the informative regret due to the prior choice—and some favorable analytical properties, such as its invariance to reparameterization [12]. However, the KL is not the only available option for the measurement of distance between statistical distributions. To illustrate: the “Encyclopedia of distances” [13] consists of 583 pages filled with distance measures, and their list is not even exhaustive. How the DAC and the Nott et al. criteria behave for different distance measure choices has yet to be investigated.

Therefore, the goal of the current paper is to investigate how the implementation of different distance measures in the DAC and Nott et al. criteria influence the detection of prior-data conflict. In the remainder, we first discuss both prior-data conflict criteria in detail. Thereafter, we discuss the design and results of a simulation study into the effect of the choice of distances on the conclusion of prior-data conflict, using a variety of distance measures (also see the Appendix A for an overview). We end with a general discussion and practical recommendations for users, who would like to test for a prior-data conflict.

## 2. Prior-Data Conflict Criteria

### 2.1. DAC

#### 2.1.1. Computation of Prior-Data Conflict

The Data Agreement Criterion (DAC; [7]) is based on the ratio of two distances, denoted by  $\Delta$ . The first distance is between the expert prior  $\pi_i^E(\theta)$  and the posterior  $\pi^B(\theta|\mathbf{y}_n)$ , where  $\theta$  denotes the parameter of interest, and upper script  $B$  is used for a ‘benchmark’, upper script  $E$  for an ‘expert’, and subscript  $i, 1, \dots, I$  to distinguish between experts. As reflected in the expression, ‘ $\pi^B(\theta|\mathbf{y}_n)$ ’, the posterior is based on dataset  $\mathbf{y}_n$  and prior  $\pi^B(\theta)$ . The idea is to choose  $\pi^B(\theta)$ , such that the posterior is dominated by the data  $\mathbf{y}_n$ . Following [14,15], in order to affect the posterior as little as possible, a reference prior should be chosen for  $\pi^B(\theta)$ . The second distance is between the posterior  $\pi^B(\theta|\mathbf{y}_n)$  and the non-informative prior  $\pi^B(\theta)$ . The ratio between the two distances results in the DAC:

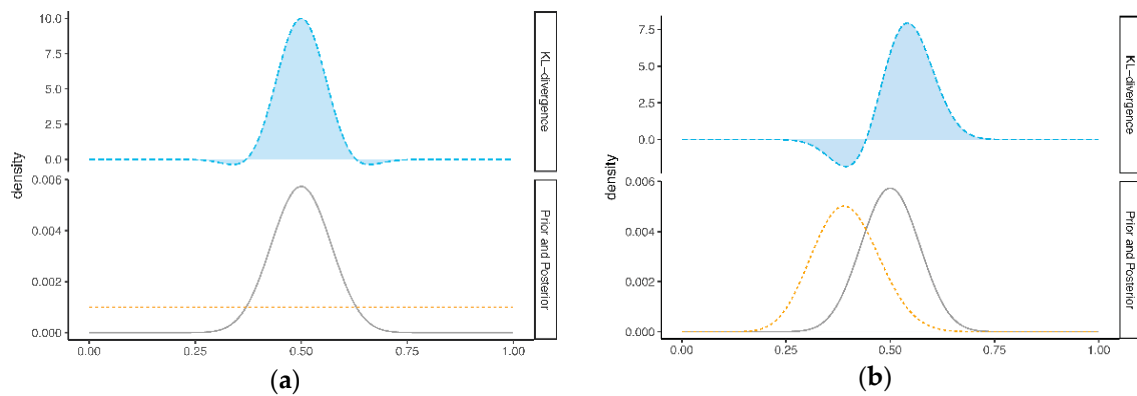
$$\text{DAC}_i = \frac{\Delta(\pi^B(\theta|\mathbf{y}_n) \parallel \pi_i^E(\theta))}{\Delta(\pi^B(\theta|\mathbf{y}_n) \parallel \pi^B(\theta))}. \quad (1)$$

In [7,9], the KL-divergence is used for  $\Delta$ , denoted here by  $\Delta_{KL}$ . The KL-divergence expresses the loss of information that occurs when we rely on the expert prior  $\pi_i^E(\theta)$ , instead of on the posterior  $\pi^B(\theta|\mathbf{y}_n)$ :

$$\Delta_{KL}(\pi^B(\theta|\mathbf{y}_n) \parallel \pi_i^E(\theta)) = \int_{\Theta} \pi^B(\theta|\mathbf{y}_n) \log \frac{\pi^B(\theta|\mathbf{y}_n)}{\pi_i^E(\theta)} d\theta. \quad (2)$$

Here,  $\Theta$  is the set of all possible values for the parameter  $\theta$ . Figure 1a illustrates the KL-divergence between the prior  $\pi^B(\theta)$  and posterior  $\pi^B(\theta|\mathbf{y}_n)$ . The lower part of this figure shows the prior  $\pi^B(\theta)$  (dashed line) and the  $\pi^B(\theta|\mathbf{y}_n)$  (solid line), and the upper part of this figure shows the corresponding KL-divergence, which is equal to the highlighted area under the curve. Figure 1b illustrates the KL-divergence between the expert prior  $\pi_i^E(\theta)$  and the posterior  $\pi^B(\theta|\mathbf{y}_n)$  (see [9]). To compute the  $\text{DAC}_i$  the KL-divergence in the upper part of Figure 1b is divided by the KL-divergence in the upper part of Figure 1a. In this example, the KL-divergence for the prior  $\pi^B(\theta)$  is 1.26, and the KL-divergence for the expert prior  $\pi_i^E(\theta)$  equals 0.89. Since this value is lower than 1.26 (the “benchmark” KL-divergence), we would, in this example, conclude that there is no prior-data conflict, according to the DAC criterion.

Note that the DAC (Equation (1)) does not necessarily have to be based on the KL-divergence (Equation (2)); any other distance or divergence measure can be substituted for  $\Delta$ .



**Figure 1.** Illustration of the Data Agreement Criterion (DAC), with (a) the Kullback-Leibler (KL)-divergence between the posterior  $\pi^B(\theta|\mathbf{y}_n)$  and non-informative prior  $\pi^B(\theta)$ , and (b) the KL-divergence between the posterior  $\pi^B(\theta|\mathbf{y}_n)$  and an expert prior  $\pi_i^E(\theta)$ . Note the difference in the y-axis scale between the upper parts and lower parts of Figure (a) and (b).

### 2.1.2. Definition of Prior-Data Conflict

As is shown by Equation (1), the distance between the prior  $\pi^B(\theta)$  and posterior  $\pi^B(\theta|\mathbf{y}_n)$  serves as a benchmark. When the distance between the expert prior  $\pi_i^E(\theta)$  and posterior  $\pi^B(\theta|\mathbf{y}_n)$  exceeds the benchmark (i.e., when  $DAC_i > 1$ ), it is concluded that there is a prior-data conflict. As stated by Bousquet (2008),  $\pi^B(\theta)$  can be seen as a fictitious, oblivious expert, perfectly in agreement with  $\mathbf{y}_n$ . When this ‘ideal’, fictitious expert  $\pi^B(\theta)$  is *more* in agreement with  $\pi^B(\theta|\mathbf{y}_n)$  than the actual expert, it is concluded that there is a prior-data conflict. According to Bousquet [7], this happens in the following two scenarios: (1) when the expert ( $\pi_i^E(\theta)$ ) favors the regions of  $\Theta$  that are far from the data mass (conflict in location), or (2) when the prior information on  $\theta$  is far more precise than the information from the data  $\mathbf{y}_n$  (conflict in information uncertainty). The latter scenario mainly arises with small data sets. Note that a prior-data conflict, as defined by Bousquet [7], does not necessarily point to a problem with the expert prior. The DAC is developed as a simultaneous check of the sampling model, and the expert data and may both have symmetric roles in the conflict.

### 2.1.3. Pros and Cons

The DAC is appealing to the applied user because of its clear, binary decision. Other than with  $p$ -values (see [10]; next section), the applied user does not have to choose a threshold him- or herself. A disadvantage, however, is the necessity to specify a suitable non-informative, benchmark prior  $\pi^B(\theta)$ . Choosing such a prior is a delicate matter, since there are many alternative candidates to be considered, which all directly influence the conclusion and thus the definition of the prior-data conflict. This is especially problematic when an improper prior is chosen. In that case, the notion of distance gets lost in the determination of  $\Delta(\pi^B(\theta|\mathbf{y}_n) \parallel \pi^B(\theta))$ . It might also feel counterintuitive that, in order to compare an expert prior  $\pi_i^E(\theta)$  with  $\mathbf{y}_n$ , the specification of another prior  $\pi^B(\theta)$  is required.

## 2.2. Criterion of Nott et al.

### 2.2.1. Computation of Prior-Data Conflict

Inspired by the work of Box see [16] and Evans and Moshonov [8,17,18], the criterion developed by Nott et al. [10] is based on the premise of Bayesian prior predictive model checking. The prior predictive distribution informs us as to which data are considered plausible a priori by the respective expert. When the collected data  $\mathbf{y}_n$  are surprising under the prior predictive (i.e., located in the tails of the prior predictive distribution), this signals that something is wrong. More formally stated, the idea

is that there is a discrepancy function  $D(\mathbf{y}_n)$  of the data  $\mathbf{y}_n$  (see [19]) and that, for the prior predictive distribution, a Bayesian  $p$ -value is computed as (see p. 3, [11]):

$$p = P(D(Y) \geq D(\mathbf{y}_n)), \tag{3}$$

where  $Y$  is a draw from the prior predictive distribution. A small  $p$ -value results when  $D(\mathbf{y}_n)$  is large (i.e., surprising), compared to  $D(Y)$ . In principle, any discrepancy function can be chosen that suits a specific model-checking goal. To check for prior-data conflict, Evans and Moshonov [8] consider  $1/m_T(T(\mathbf{y}_n))$  as their discrepancy function, where  $m_T$  is the prior predictive density of a minimal sufficient statistic  $T$ . Despite their promising results (see [17,18]), a drawback of this suggestion is that the check is not invariant in relation to the choice of the minimal sufficient statistic, when (3) is continuous (see [20]). Nott et al. [10] solve this problem elegantly by considering another discrepancy function: the distance between the prior  $\pi_i^E(\theta)$  and posterior  $\pi_i^E(\theta|\mathbf{y}_n)$ ;  $\Delta(\pi_i^E(\theta|\mathbf{y}_n) \parallel \pi_i^E(\theta))$ . The advantage of  $\Delta(\pi_i^E(\theta|\mathbf{y}_n) \parallel \pi_i^E(\theta))$  is that it depends on the data  $\mathbf{y}_n$  *only* through the posterior distribution  $\pi_i^E(\theta|\mathbf{y}_n)$ . Hence, the statistic  $\Delta(\pi_i^E(\theta|\mathbf{y}_n) \parallel \pi_i^E(\theta))$  is a function of any sufficient statistic and thus invariant in relation to the particular choice of  $T$ .

To express the distance between  $\pi_i^E(\theta)$  and  $\pi_i^E(\theta|\mathbf{y}_n)$ , Nott et al. [10] consider a class of divergences, of which the KL-divergence is a special case. This class contains the Rényi divergences of order  $\alpha$ , here denoted by  $\Delta_{R\alpha}$ .  $\Delta_{R\alpha}$  can be explained as a measure of how much beliefs change from prior to posterior, comparable to the notion of relative belief (see [21]). The earlier mentioned KL-divergence is a special case of this class, with  $\alpha \rightarrow 1$ . For the Nott et al. [10] criterion, Equation (3) can thus be rewritten as:

$$p_i = P(\Delta_{R\alpha}(\pi_i^E(\theta|Y) \parallel \pi_i^E(\theta)) \geq \Delta_{R\alpha}(\pi_i^E(\theta|\mathbf{y}_n) \parallel \pi_i^E(\theta))), \tag{4}$$

where

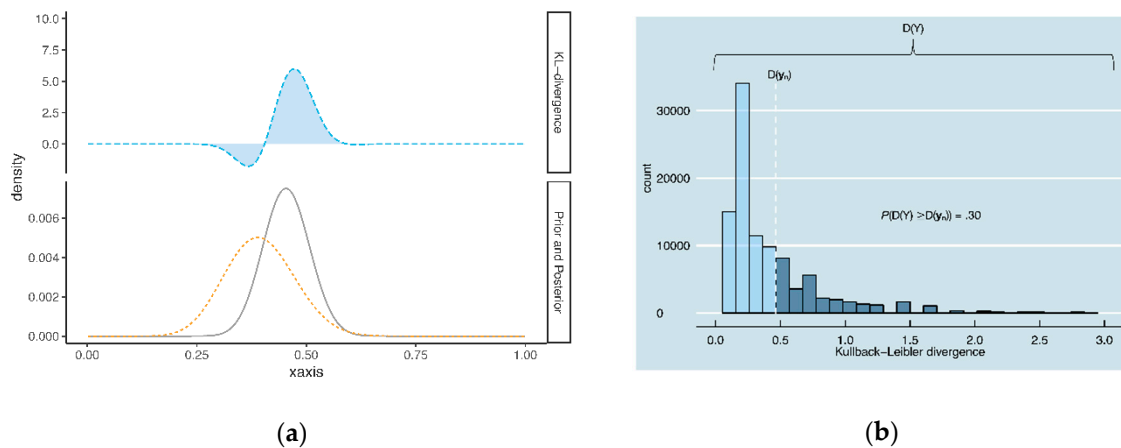
$$\Delta_{R\alpha}(\pi_i^E(\theta|\mathbf{y}_n) \parallel \pi_i^E(\theta)) = \frac{1}{\alpha - 1} \log \int_{\Theta} \pi_i^E(\theta|\mathbf{y}_n) \left\{ \frac{\pi_i^E(\theta|\mathbf{y}_n)}{\pi_i^E(\theta)} \right\}^{\alpha-1} d\theta, \tag{5a}$$

and

$$\Delta_{R\alpha}(\pi_i^E(\theta|Y) \parallel \pi_i^E(\theta)) = \frac{1}{\alpha - 1} \log \int_{\Theta} \pi_i^E(\theta|Y) \left\{ \frac{\pi_i^E(\theta|Y)}{\pi_i^E(\theta)} \right\}^{\alpha-1} d\theta. \tag{5b}$$

Figure 2 illustrates the criterion of Nott et al. for the same  $\pi_i^E(\theta)$  and  $\mathbf{y}_n$ , as used in Figure 1. In Figure 2a, the KL-divergence between the prior  $\pi_i^E(\theta)$  (lower panel; dashed line) and posterior  $\pi_i^E(\theta|\mathbf{y}_n)$  (lower panel; solid line) is illustrated. The area under the curve is equal to 0.43. In Figure 2b, this value (0.43) is compared to  $10^5$  KL-divergences, obtained by repeatedly drawing  $Y$  from the prior predictive distribution of the expert  $i$ . In this example, 30% of the KL-divergences in Figure 2b exceed 0.43 (see the dark blue bins in Figure 2b). We would therefore conclude that there is no prior-data conflict in this example.

Note that, in prior-predictive checking, the user has considerable freedom in choosing a suitable discrepancy function  $D(\mathbf{y}_n)$ . Therefore, the distance measure  $\Delta_{R\alpha}$  in the criterion of Nott et al. [10] can be replaced by any other distance or divergence measure.



**Figure 2.** Illustration of the criterion of Nott et al. [9], with (a) the KL-divergence between the posterior  $\pi_i^E(\theta|\mathbf{y}_n)$  and expert prior  $\pi_i^E(\theta)$  and (b) the distribution of KL-divergences between the posterior  $\pi_i^E(\theta|Y)$  and expert prior  $\pi_i^E(\theta)$ , for  $10^5$  draws  $Y$  from the prior predictive.

2.2.2. Definition of Prior-Data Conflict

In [10], it is found that prior-data conflict exists when the observed distance between the prior  $\pi_i^E(\theta)$  and posterior  $\pi_i^E(\theta|\mathbf{y}_n)$  is surprising in relation to the corresponding prior predictive distribution. In other words, if the data  $\mathbf{y}_n$  would have been in line with the prior  $\pi_i^E(\theta)$ , we would not expect to observe such a large distance  $\Delta_{R\alpha}(\pi_i^E(\theta|\mathbf{y}_n)\pi_i^E(\theta))$ . Hence, the prior  $\pi_i^E(\theta)$  and the data  $\mathbf{y}_n$  appear to be in conflict.

Note that this definition of prior-data conflict is vastly different from the earlier definition of the DAC. First, although both rely on the KL-divergence, in the criterion of Nott et al., the prior  $\pi_i^E(\theta)$  is directly compared to its corresponding posterior  $\pi_i^E(\theta|\mathbf{y}_n)$ , instead of  $\pi^B(\theta|\mathbf{y}_n)$ , as in the DAC. The criterion of Nott et al. (2016) and the DAC therefore ask different questions: how much information do we lose when relying on  $\pi_i^E(\theta)$ , instead of  $\pi^B(\theta|\mathbf{y}_n)$  (DAC)? To what extent does our information change, when moving from  $\pi_i^E(\theta)$  to  $\pi_i^E(\theta|\mathbf{y}_n)$  (Nott et al.)? Second, the observed values  $\Delta_{KL}(\pi^B(\theta|\mathbf{y}_n)\pi_i^E(\theta))$  and  $\Delta_{R\alpha}(\pi_i^E(\theta|\mathbf{y}_n)\pi_i^E(\theta))$  are calibrated differently. In the DAC, prior-data conflict exists when more information is lost than with the benchmark prior  $\pi^B(\theta)$ . In the criterion of Nott et al., the  $p$ -value in Equation (4) calibrates the change in information by expressing how surprising this change is. Third, the uncertainty of the expert (i.e., the variance of  $\pi_i^E(\theta)$ ) is treated differently in the criterion of Nott et al. than in the DAC. In the DAC, prior-data conflict can occur when there is *no* mismatch in location (i.e., both  $\pi_i^E(\theta)$  and  $\pi^B(\theta|\mathbf{y}_n)$  place their mass primarily in the same region of  $\Theta$ ) but there *is* a mismatch in information uncertainty (i.e., the variance of  $\pi_i^E(\theta)$  is smaller than the variance of  $\pi^B(\theta|\mathbf{y}_n)$ ). In the criterion of Nott et al., the uncertainty of the expert simply reflects the variety of data that are deemed plausible by this expert, as captured in the prior predictive distribution. Therefore, when the location of the prior and posterior match, we do not find a prior-data conflict with the criterion of Nott et al., even if that prior variance is relatively small. Finally, the DAC and the criterion of Nott et al. respond differently, when little data is available. With a small sample size, the criterion of Nott et al. may fail to detect an existing prior-data conflict. A bad prior may pass the test, as the relatively large sampling variability is taken into account in the comparison with the prior predictive. In the DAC, the sampling variability is not taken into consideration;  $\pi^B(\theta|\mathbf{y}_n)$  is treated as a temporary ‘truth’. Therefore, we should be careful not to falsely interpret a prior-data conflict as a problem with the prior in a situation where there is little data. Such a conflict may very well be the result of the unavailability of representative data and a mismatch in the amount of information captured in  $\pi_i^E(\theta)$  and  $\pi^B(\theta|\mathbf{y}_n)$ .

### 2.2.3. Pros and Cons

One major advantage is that—aside from Bousquet [7]—the criterion of Nott et al. [10] does not depend on the definition of a non-informative prior  $\pi^B(\theta)$ . A possible disadvantage, however, is that the  $p$ -value in Equation (4) is often misinterpreted in practice (see [22]). The  $p$ -value in Equation (4) should be used as a measure of surprise rather than support ([20]), and not be mistaken for the probability of a (non-)conflict between  $\pi_i^E(\theta)$  and  $\pi_i^E(\theta|\mathbf{y}_n)$ . Given that the definition of prior-data conflict is quite different for the DAC and the criterion of Nott et al. (see Section 2.2.1), choosing one or the other should be based on a clear and well-informed opinion of what prior-data conflict entails.

## 3. Simulation Study

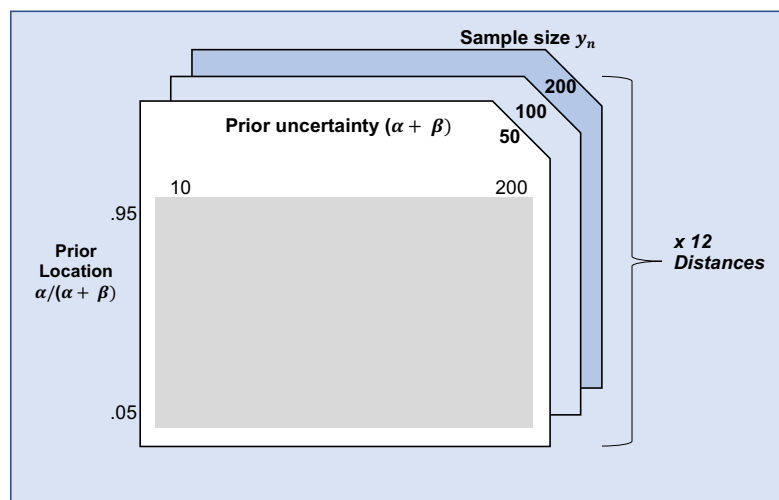
### 3.1. Goal of the Simulation

In principle, the KL-divergence in Equations (1) and (4) can be replaced by any other distance measure. The goal of our simulation is therefore to investigate the role of the distance measure in the detection of prior-data conflict. The main question is: How much does the choice of distance measure eventually impact the conclusions of prior-data conflict by the DAC and the criterion of Nott et al.? In other words: How robust are the DAC and the criterion of Nott et al. in relation to the choice of distance measure?

### 3.2. Simulation Design

#### 3.2.1. Scenario

In our simulation, we focus on one example scenario, in which  $\theta$  is a one-dimensional parameter. Specifically,  $\mathbf{y}_n \sim \text{Binomial}(n, \theta)$  and  $\pi_i^E(\theta)$  is a conjugate Beta distribution  $\text{beta}(\alpha, \beta)$  on  $[0,1]$ . In this scenario,  $\pi^B(\theta)$ —the benchmark prior in the DAC—is set to  $\text{beta}(\alpha = 1, \beta = 1)$ , and the prior predictive distribution for the criterion of Nott et al. equals a Beta Binomial, with parameters  $\alpha, \beta$  and  $n$ .  $\theta$  set to 0.5 in the population, whereas  $\alpha, \beta$  and  $n$  are varied in the simulation. Note that  $\alpha$  and  $\beta$  together express the location of the expert prior (i.e., the region with most of the prior's mass),  $\alpha/(\alpha + \beta)$ , and the (un)certainty of the expert.  $\alpha + \beta$ .  $\alpha/(\alpha + \beta)$  are set to 0.05, 0.07, 0.09,  $\dots$  0.95, respectively, and  $\alpha + \beta$  is varied from 10, 12, 14,  $\dots$ , 200. The sample size of the data,  $n$ , is varied between 50, 100 and 200. As explained in Section 3.2.3, twelve distance measures are compared. Figure 3 summarizes the simulation design.



**Figure 3.** Illustration of the simulation design. The prior location ( $\alpha/(\alpha + \beta)$ ), prior uncertainty ( $\alpha + \beta$ ), and sample size of  $\mathbf{y}_n$  were varied, as well as the distance measure used.

### 3.2.2. Steps

We started the simulation by taking 1000 samples from the respective population, with  $\theta = 0.5$  (see Section 3.2.1.) and sample sizes equal to 50. We repeated this sampling process for the other sample sizes ( $n = 100$  and  $200$ ). Subsequently, multiple expert priors  $\pi_i^E(\theta)$  were constructed, based on all possible combinations of location (i.e.,  $\alpha/(\alpha + \beta) = 0.05, \dots, 0.95$ ) and expert uncertainty (i.e.,  $\alpha + \beta = 10, 12, \dots, 200$ ). For all these expert priors, we then determined  $\pi_i^E(\theta|y_n)$  (for the criterion of Nott et al.) and  $\pi_i^B(\theta|y_n)$  (for the DAC) and calculated their respective distances to  $\pi_i^E(\theta)$  using the variety of distance measures, named in Section 3.2.3. The calculation of the DAC and  $p$ -value of the criterion of Nott et al. (2016) followed, using Equations (1) and (4), respectively. Finally, all DAC values above 1 were flagged as prior-data conflicts. For the criterion of Nott et al. (2016), we considered  $p \leq 0.05$  as indicative of a prior-data conflict. All analyses were performed using RStudio [23] and the R packages, “Philentropy” [24], “distr” and “distrEx” [25].

### 3.2.3. Distance Measures

The KL-divergence, used in both the DAC and the criterion of Nott et al., is an instance of an  $f$ -divergence, also known as an Ali-Silvey divergence ([26]). The more general Rényi divergences, considered in Nott et al., are related to this class. One of the distance measures we investigate in our simulation (the total variation distance) belongs to the class of  $f$ -divergences (see [27]). We also consider distance measures that are *not* part of this class: Hellinger, Kolmogorov, Euclidean, Manhattan, Sorensen, Intersection, Harmonic mean, Bhattacharyya, Divergence, Jeffreys and Jensen-Shannon (leading to twelve distance measures in total). See [28] for more information on these distance measures.

## 4. Results

As stated in Section 3.1, the main goal of the simulation was to investigate the robustness of the DAC and the criterion of Nott et al. in relation to the choice of the twelve distance measures under study. Below, we split the results of the simulation into two parts. In the first part, we look at how the conclusion of prior-data conflict varies with different choices of distance measures, when we vary the expert priors for a specific sample  $y_n$  (see Figure 3). Ideally, whether or not the varying expert priors are found to be in conflict with the specific sample should not depend on the choice of distance measure. In this first part, we act as if the fixed, single sample is ‘the truth’. Of course, the sample is drawn from a population. Therefore, in the second part, we look at varying samples  $y_n$  and specific expert priors. This way, we can investigate how sampling variability influences the behavior of the DAC and the criterion of Nott et al., when different distance measures are used.

### 4.1. Robustness in Relation to the Choice of Distance Measure: Specific Sample and Varying Expert Priors

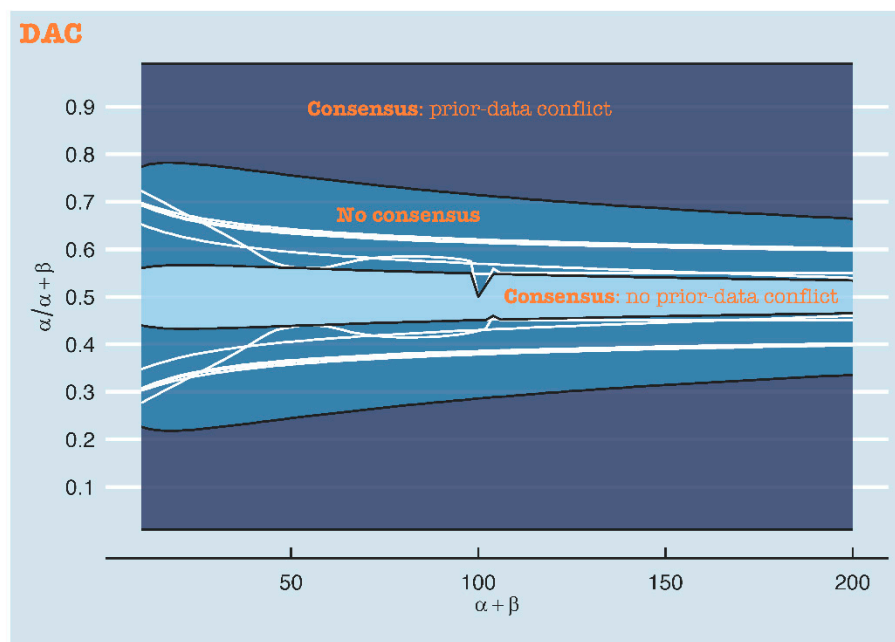
In this section, we investigate varying expert priors to determine whether they are in conflict with one specific sample,  $y_n \sim \text{Binomial}(n = 100, \theta = 0.5)$ . To limit the number of plots here, a part of the plots is shown in the Appendix A. Specifically, the plots in the Appendix A visualize the DAC and  $p$ -values for varying expert priors, for each of the twelve distance measures separately. The x-axes show the expert prior uncertainty ( $\alpha + \beta$ ) and the y-axis prior location ( $\alpha/(\alpha + \beta)$ ), such that, for every combination of uncertainty and location, the DAC and  $p$ -value can be read from the plots. The contour lines show which expert priors lead to comparable DAC and  $p$ -values, such as  $\text{DAC} = 0.2, 0.3, \dots, 1.1$  and  $p = 0.05, 0.1, \dots, 0.9$ . The most interesting contour lines are  $\text{DAC} = 1$  and  $p = 0.05$ , since these lines distinguish the expert priors for which a prior-data conflict and no prior-data conflict is concluded. As seen in the DAC plots, the Jeffreys divergence leads to the most lenient decision of a prior-data conflict. Indeed, when Jeffreys divergence is used, relatively many expert priors lead to a DAC value below 1 (i.e., many combinations of expert uncertainty and location fall in the area of DAC values  $< 1$ ). Kolmogorov, on the other hand, leads to the most stringent decision of prior-data conflict. All divergence-based measures show problematic behavior. Using ‘divergence’, DAC values below 1

are solely found for a small group of expert priors. In the criterion of Nott et al., divergence shows divergent behavior, especially for expert priors with a low  $\alpha + \beta$ . To a lesser extent, this problematic behavior at low values of  $\alpha + \beta$  is also found when using the Euclidean distance in the criterion of Nott et al. Apart from the divergence and Euclidean distance, the criterion of Nott et al. generally shows behavior more comparable to the  $p$ -value for different distance measures than the DAC.

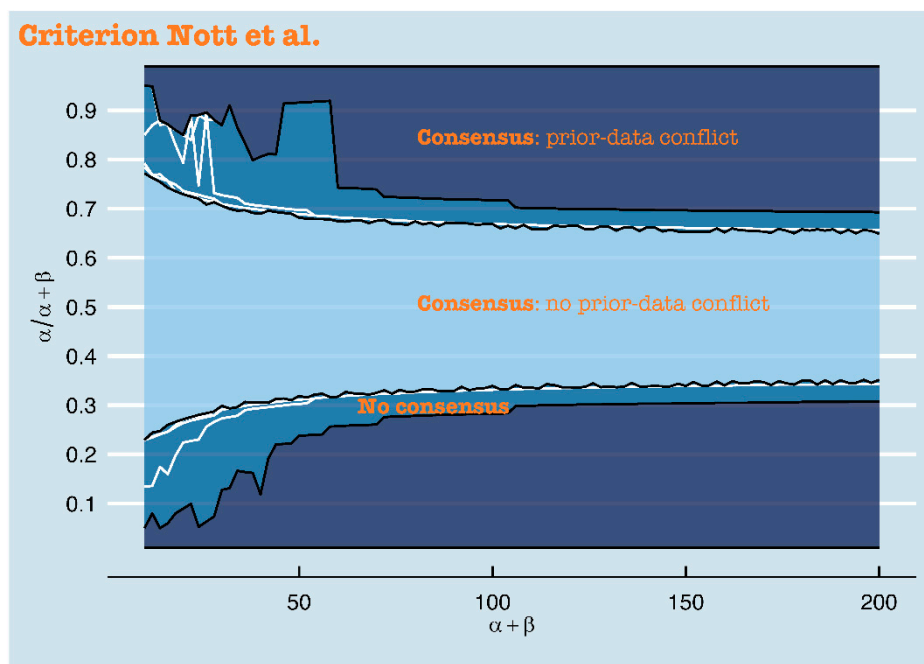
Figure 4a combines the twelve DAC plots from the Appendix A into one plot, in which only the contour lines 'DAC = 1' are displayed. By doing so, these DAC = 1 boundaries can more easily be compared over the twelve distance measures. Figure 4b does the same for the twelve  $p$ -value plots from the Appendix A, for the contour lines  $p = 0.05$ . The contour lines (DAC = 1 and  $p = 0.05$ ) of the distance measure which leads to the most stringent conclusion of prior-data conflict, mark the area (i.e., the combinations of expert prior uncertainty and location) for which all distance measures agree that there is *no* prior-data conflict. In Figure 4a,b, this area is colored light blue. As stated before, in Figure 4a, this area is determined by Kolmogorov. The contour lines (DAC = 1 and  $p = 0.05$ ) belonging to the distance measure which leads to the most lenient conclusion of prior-data conflict, mark the area for which all distance measures agree that there *is* a prior-data conflict. As stated before, in Figure 4a, this area is determined by Jeffreys divergence. In Figure 4a,b, this area is colored dark blue. The area in between the contour lines of the most lenient and most stringent distance measures illustrates the area of no consensus, i.e., the expert priors for which some distance measures would lead to a conclusion of prior-data conflict and others not. The larger this area is, the more influence the choice of distance measure has on the conclusion of prior-data conflict. The white lines in this area illustrate the exact DAC = 1 and  $p = 0.05$  contour lines of the distance measures. Comparing Figure 4a,b clearly shows that a different choice of distance measure more profoundly influences the conclusion of the DAC than the criterion of Nott et al. When used in the criterion of Nott et al., the distance measures primarily disagree with each other for the expert priors that have  $\alpha/(\alpha + \beta)$  far from 0.5 (0.1, 0.2, 0.8 and 0.9) and a low  $\alpha + \beta$  (smaller than 50). The outer line separating the area with no consensus from the area with consensus on prior-data conflict is formed by the Divergence distance. Of the twelve distance measures under study here, this distance measure thus leads to the most lenient decision of prior-data conflict, at least in this specific scenario.

For only a relatively small area, the distance measures agree that there is no prior-data conflict, when used in the DAC (see Figure 4a). Other than in the criterion of Nott et al., differences in conclusions of (no) prior-data conflict are not restricted to certain combinations of prior location ( $\alpha/(\alpha + \beta)$ ) and uncertainty ( $\alpha + \beta$ ).





(a)



(b)

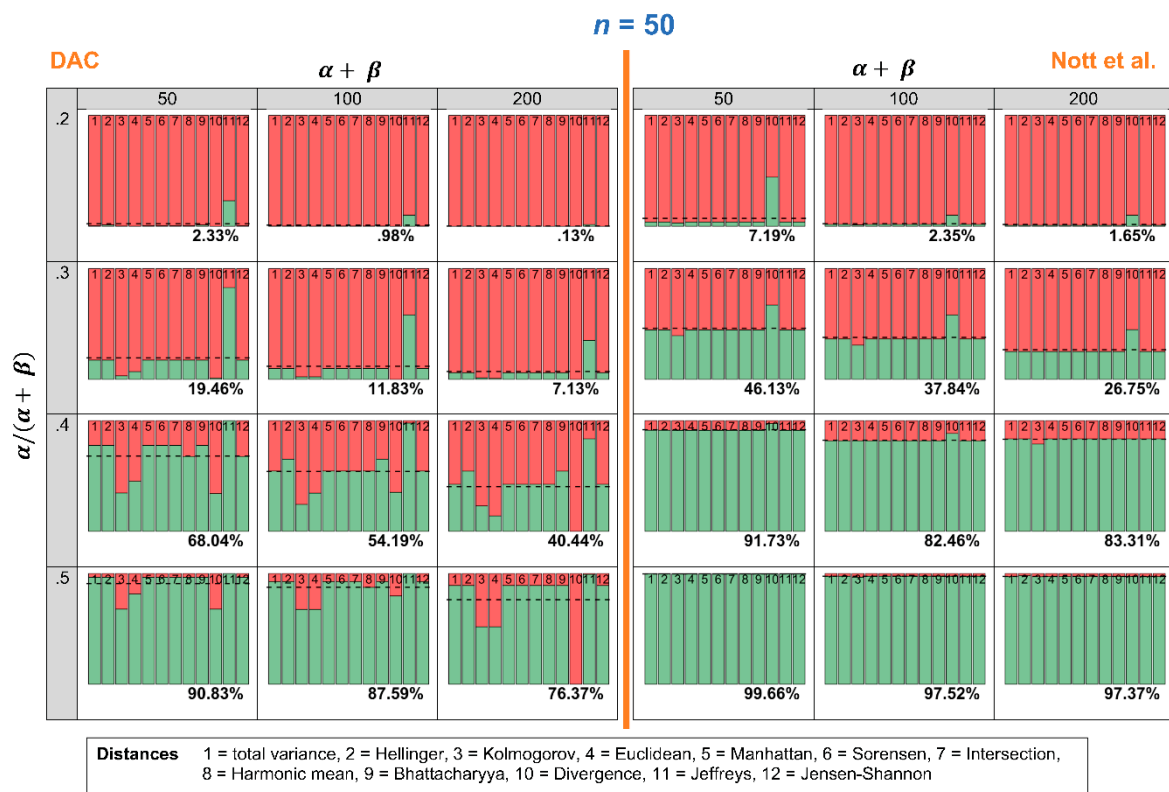
**Figure 4.** The white lines in this figure show the  $DAC = 1$  (a) and  $p < 0.05$  (criterion of Nott et al.) (b) boundaries for the twelve distance measures, when the sample is fixed, and  $\alpha + \beta$  (x-axis) and  $\alpha/(\alpha + \beta)$  (y-axis) are varied. The darkest blue area illustrates the expert priors (i.e., the combination of  $\alpha + \beta$  and  $\alpha/(\alpha + \beta)$ ) for which all distance measures would lead to a conclusion of prior-data conflict. The lightest blue area shows the expert priors that consistently lead to a conclusion of no prior-data conflict. The area in between these two areas is inconclusive and depends on the choice of distance measure.

4.2. Robustness in Relation to the Choice of Distance Measure: Specific Expert Prior and Varying Samples

In this section, we investigate how many times the DAC and  $p$ -values lead to a conclusion of prior-data conflict, when repeated samples are drawn, and the expert prior is fixed. Figure 5a–c show

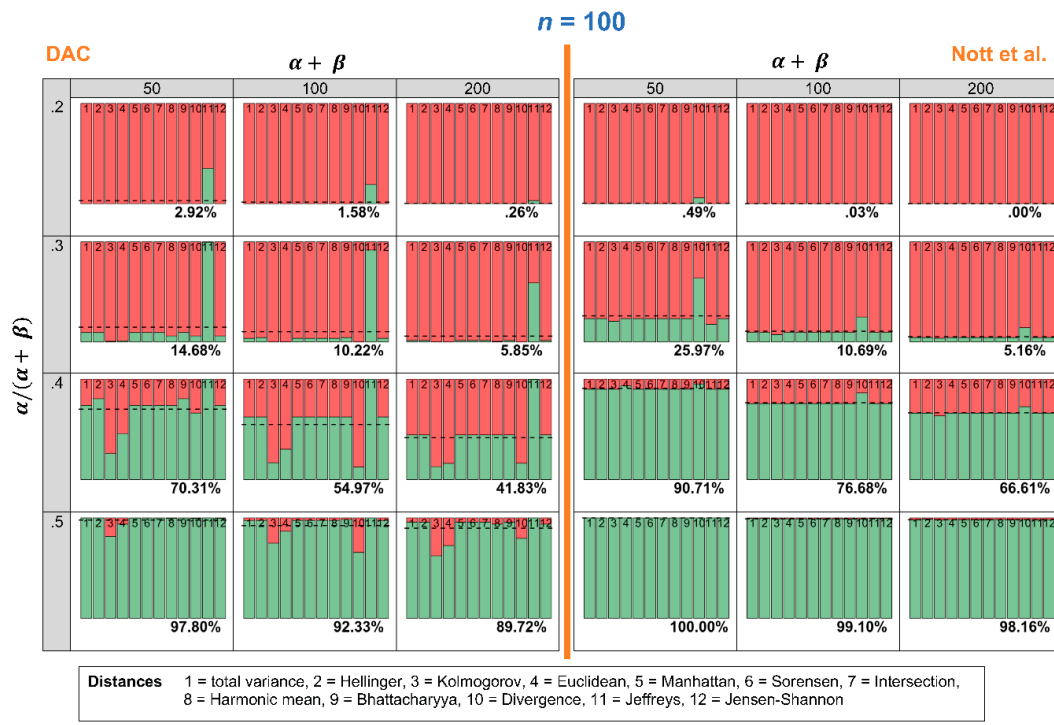
an overview of the results. The stacked histograms show the number of times the DAC and the criterion of Nott et al. do (red color) or do not (green color) reach a conclusion of prior-data conflict. Every stacked histogram is based on a Beta prior, with a certain combination of  $\alpha/(\alpha + \beta)$  (row) and  $\alpha + \beta$  (column) and 1000 repeated samples  $y_n \sim \text{Binomial}(n, \theta)$  with  $\theta = 0.5$  and  $n = 50$  (Figure 5a),  $n = 100$  (Figure 5b) or  $n = 200$  (Figure 5c). Every bin of the stacked histogram corresponds to one of the 12 distance measures. The first bin, for instance, corresponds with the total variance distance, and the second one corresponds with the Hellinger distance. When a bin within a stacked histogram is completely red, this means that *all* 1000 repeated samples led to a conclusion of prior-data conflict, when this specific distance measure was used. The dashed horizontal line and the corresponding percentage printed at the bottom of each histogram show the average number of times the distance measures lead to a conclusion of no prior-data conflict. To ease comparison, the left side of Figure 5a–c shows the results for the DAC, and the right side of Figure 5a–c shows the results for the criterion of Nott et al.

The most important conclusion that can be drawn from Figure 5a–c is that the criterion of Nott et al. is less sensitive to the choice of the twelve distance measures than the DAC. Comparing the right side (criterion Nott et al.) with the left side (DAC) of Figure 5a–c, the coloring of the bins is more comparable over the twelve distance measures when the Nott et al. criterion rather than the DAC is used. Only one of the twelve distance measures shows problematic behavior: “Divergence” (bin nr. 10). When “Divergence” is used as a distance measure in the criterion of Nott et al., the conclusion of prior-data conflict becomes more lenient. When  $\alpha/(\alpha + \beta) = 0.2$ ,  $\alpha + \beta = 50$  and  $n = 50$ , for instance, almost half of the 1000 repeated samples would result in a conclusion of “no prior-data conflict”. Using any of the other distance measures would, however, lead to a conclusion of prior-data conflict in almost all samples.

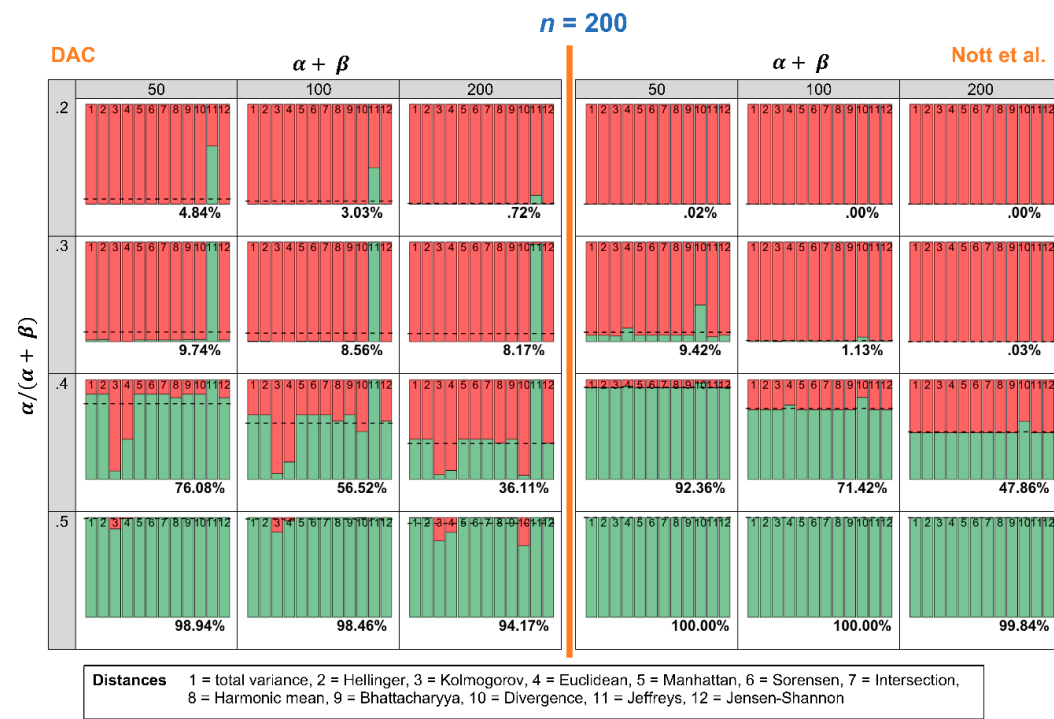


(a)

Figure 5. Cont.



(b)



(c)

**Figure 5.** This Figure shows how many times, out of 1000 repeated samples  $y_n \sim \text{Binomial}(n, \theta)$  of  $\theta = 0.5$  and  $n = 50$  (a),  $n = 100$  (b) or  $n = 200$  (c), the twelve distance measures led to a conclusion of prior-data conflict. Every stacked histogram is based on an expert prior, with a certain combination of  $\alpha/(\alpha + \beta) = 0.2, 0.3, 0.4$  or  $0.5$  (rows) and  $\alpha + \beta = 50, 100$  or  $200$  (columns). The red bins correspond to a conclusion of prior-data conflict, and the green bins correspond to a conclusion of no prior-data conflict.

## 5. Conclusions

The goal of the current paper was to investigate how the implementation of different distance measures in the DAC and Nott et al. criteria influence the detection of prior-data conflict. Overall, the criterion of Nott et al. seems less sensitive to the choice of distance measure than the DAC.

In the criterion of Nott et al., the user seems to have more freedom in the choice of distance measure. With the exception of one distance measure, all distance measures led to a comparable conclusion of prior-data conflict (see again Figure 5). This is advantageous, as even the simplest (and fastest computable) distance measures (for instance, the Manhattan distance) can be chosen, without changing the interpretation of prior-data conflict. In our specific scenario, we did see that some distance measures led to a more lenient treatment of experts, who were very insecure and incorrect.

Concerning the DAC, substituting the KL-divergence (see Equation (2)) for another distance measure is only encouraged, when the KL results in issues. One reason to *not* use the KL-divergence in the DAC, for instance, is that it can lead to computational problems (see [27]). In case of such computational problems, the best strategy is to search for a distance measure without these computational issues and with comparable behavior, so as to not jeopardize the interpretation of the DAC. Plotting the size of the DAC (or the criterion of Nott et al.), compared to the characteristics of the (possible) expert priors, as we did in Figure 4 and in the Appendix A, can help to spot potential problems with the distance measure. Such a plot can also help to check whether the distance measure leads to a relatively strict or relatively lenient conclusion of prior-data conflict.

The results of our study are limited to a specific example scenario and a limited set of twelve distance measures. We chose this limited scope, as it was impossible to investigate the hundreds of possible distance measures and anticipate all possible scenarios. Our study can, therefore, best be read as an example of how the choice of distance measure *may* influence the detection of prior-data conflict with the DAC and the criterion of Nott et al. Whether or not the detection of prior-data conflict depends on the choice of distance measure in another scenario should be checked beforehand by the investigator.

**Author Contributions:** Conceptualization, K.L. and R.V.D.S.; Formal analysis, K.L.; Funding acquisition, K.L. and R.V.D.S.; Methodology, K.L.; Supervision, R.V.D.S.; Writing – original draft, K.L.

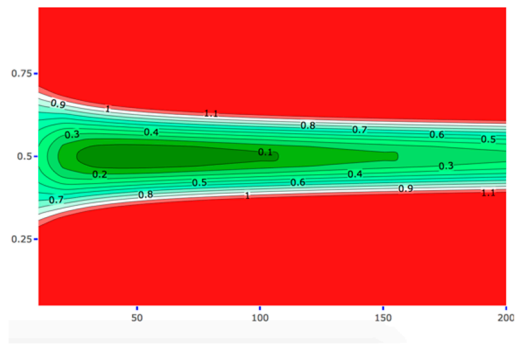
**Funding:** This research was funded by the Netherlands Organization for Scientific Research (NWO). Specifically, Kimberley Lek was funded by a NWO-talent grant, grant number [406-15-062] and Rens Van De Schoot by a NWO-Vidi grant, grant number [452-14-006].

**Conflicts of Interest:** The authors declare no conflict of interest.

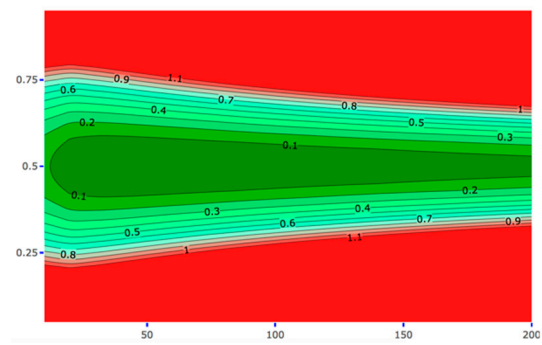
### Appendix A DAC- and $p$ -Values Separately for Each of the Twelve Distance Measures, for Fixed Sample and Varying Expert Priors

This appendix provides background information for Figure 4. In Figure 4, we showed the DAC = 1 and  $p = 0.05$  boundaries of all twelve distance measures in one plot. We did this to compare the distance measures' behavior for varying expert priors, when the sample  $y_n$  is fixed. In this Appendix, we show the DAC- (upper part) and  $p$ -values (lower part) for each of the twelve distance measures separately, for the expert priors with varying  $\alpha + \beta$  (x-axis) and  $\alpha/(\alpha + \beta)$  (y-axis).

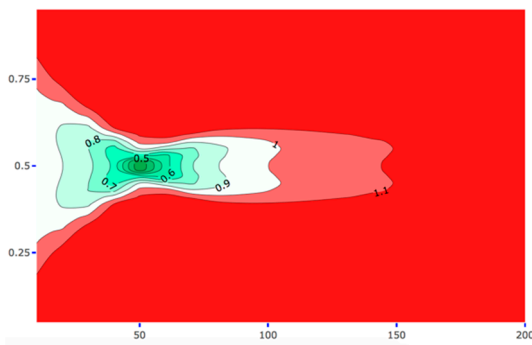
DAC



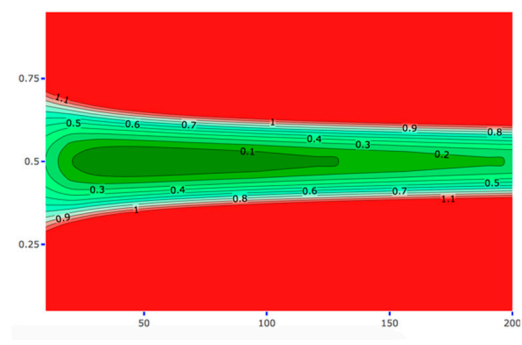
Jensen-Shannon



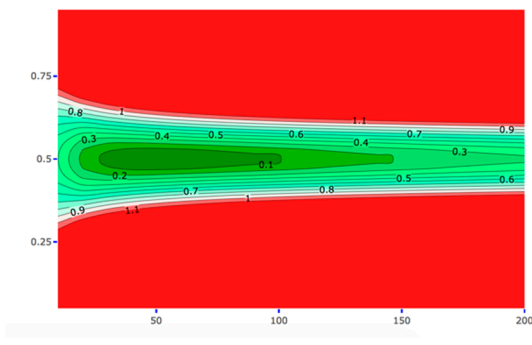
Jeffreys



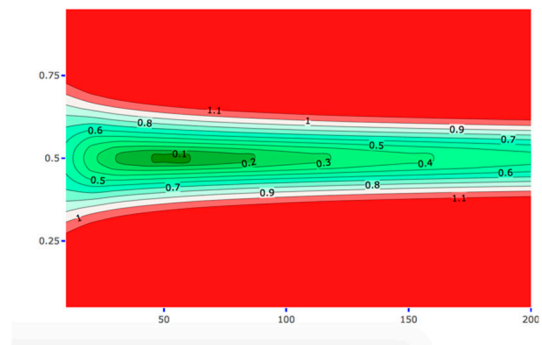
Divergence



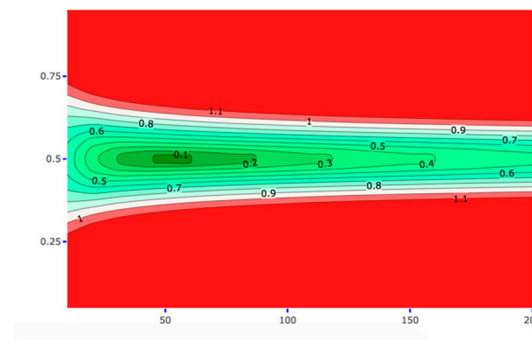
Bhattacharyya



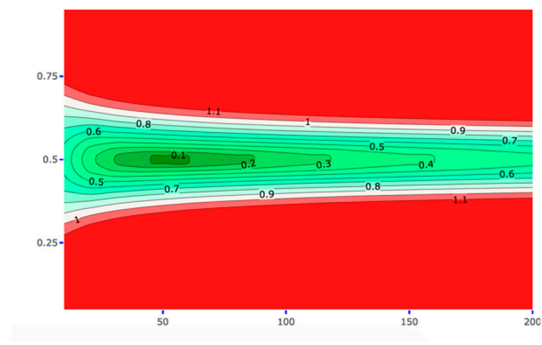
Harmonic mean



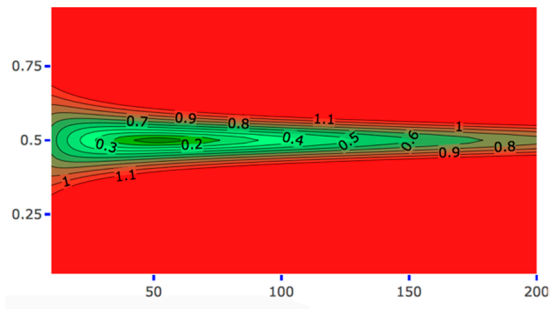
Intersection



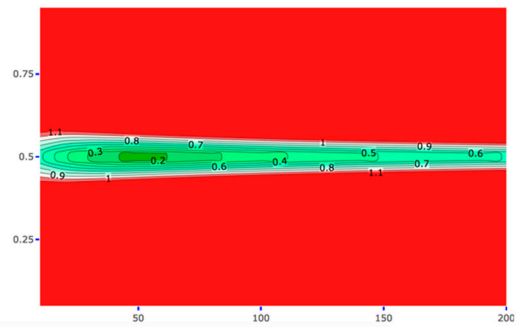
Sorensen



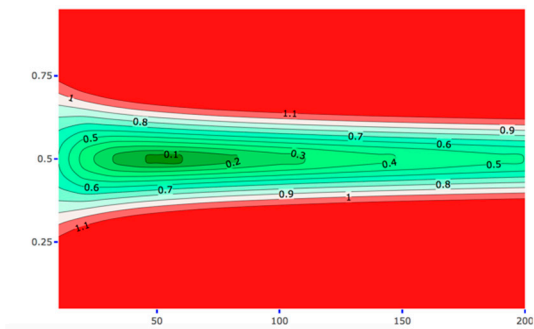
Manhattan



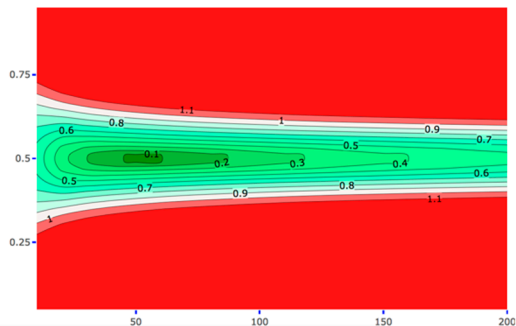
**Euclidean**



**Kolmogorov**

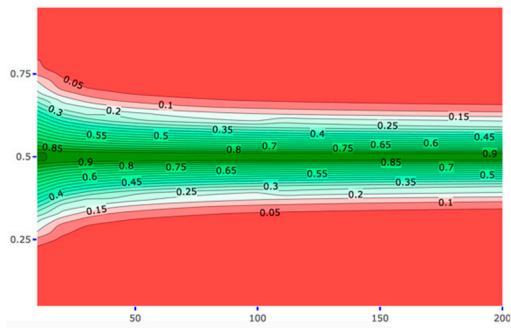


**Hellinger**

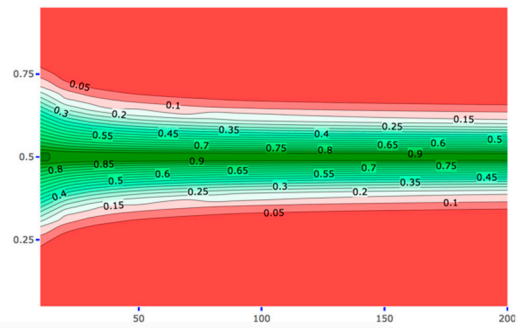


**Total variance**

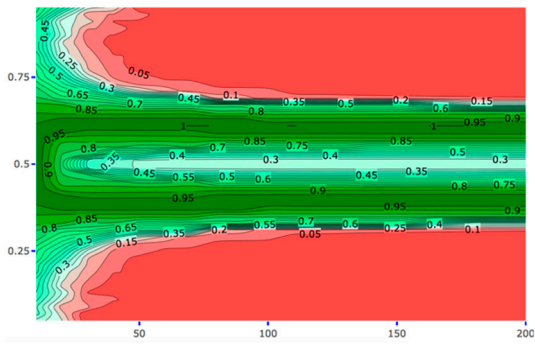
*Criterion Nott et al.*



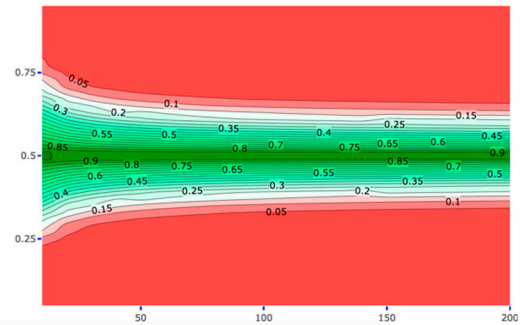
**Jensen-Shannon**



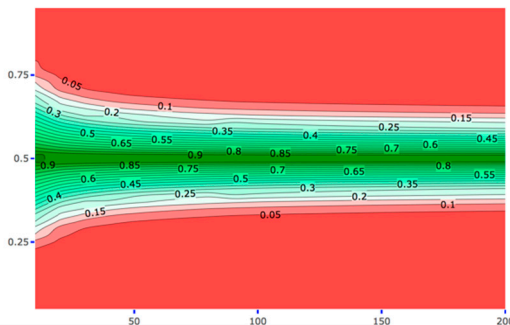
**Jeffreys**



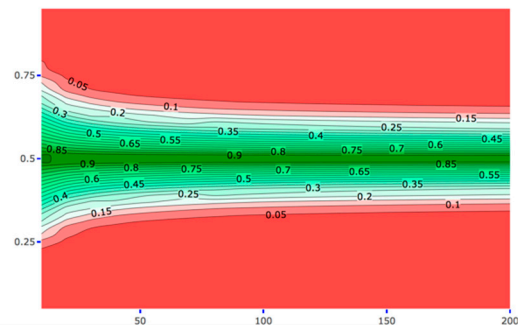
**Divergence**



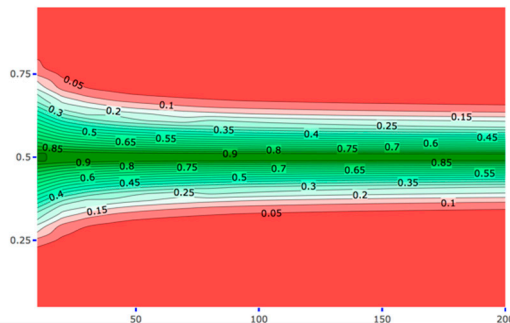
**Bhattacharyya**



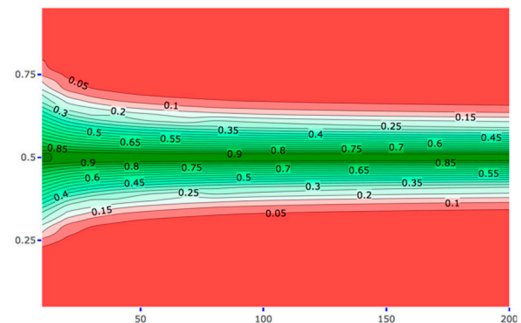
**Harmonic mean**



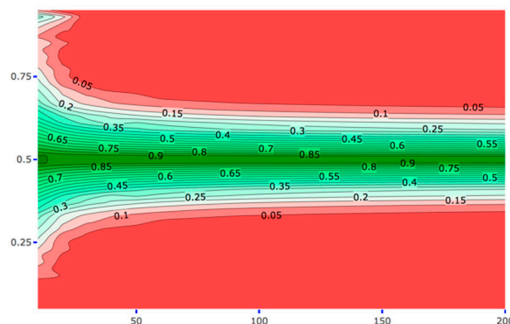
**Intersection**



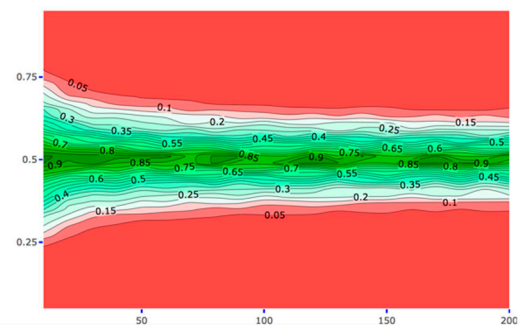
**Sorensen**



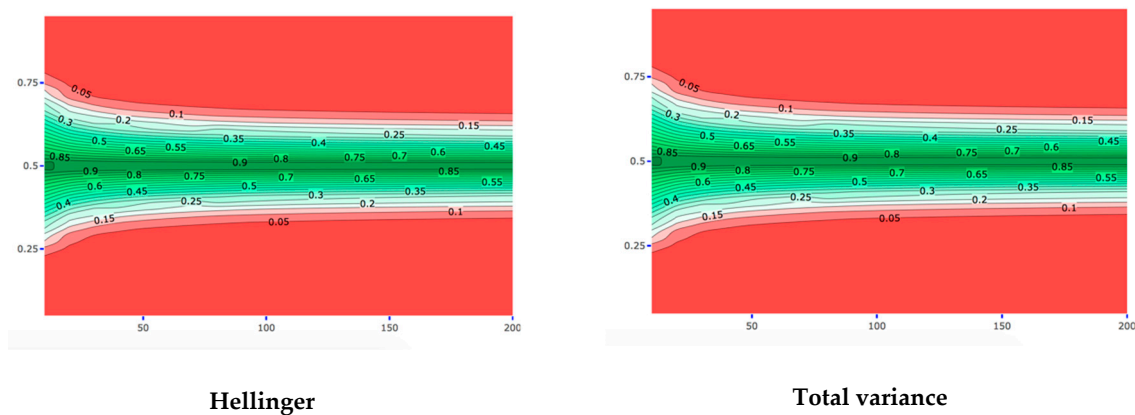
**Manhattan**



**Euclidean**



**Kolmogorov**



## References

- O'Hagan, A.; Buck, C.E.; Daneshkhah, A.; Eiser, J.R.; Garthwaite, P.H.; Jenkinson, D.J.; Oakley, J.E.; Rakow, T. *Uncertain Judgements: Eliciting Experts' Probabilities*, 3rd ed.; John Wiley & Sons: London, UK, 2006.
- Kuhnert, P.M.; Martin, T.G.; Griffiths, S.P. A guide to eliciting and using expert knowledge in Bayesian ecological models. *Ecol. Lett.* **2010**, *13*, 900–914. [[CrossRef](#)] [[PubMed](#)]
- O'Leary, R.; Low Choy, S.; Murray, J.V.; Kynn, M.; Denham, R.; Martin, T.G.; Mengersen, K. Comparison of three expert elicitation methods for logistic regression on predicting the presence of the threatened brush-tailed rock-wallaby *Petrogale penicillata*. *Environmetrics* **2009**, *20*, 379–398. [[CrossRef](#)]
- Martin, T.G.; Burgman, M.A.; Fidler, F.; Kuhnert, P.M.; Low-Choy, S.; McBride, M.; Mengersen, K. Eliciting expert knowledge in conservation science. *Conserv. Biol.* **2011**, *26*, 29–38. [[CrossRef](#)] [[PubMed](#)]
- Zondervan-Zwijnenburg, M.; Van De Schoot-Hubeek, W.; Lek, K.; Hoijtink, H.; Van De Schoot, R. Application and evaluation of an expert judgment elicitation procedure for correlations. *Front. Psychol.* **2017**, *8*, 1–15. [[CrossRef](#)] [[PubMed](#)]
- Lek, K.M.; Van De Schoot, R. Development and evaluation of a digital expert elicitation method aimed at fostering elementary school teachers' diagnostic competence. *Front. Educ.* **2018**, *3*, 1–14. [[CrossRef](#)]
- Bousquet, N. Diagnostics of prior-data agreement in applied Bayesian analysis. *J. Appl. Stat.* **2008**, *35*, 1011–1029. [[CrossRef](#)]
- Evans, M.; Moshonov, H. Checking for prior-data conflict. *Bayesian Anal.* **2006**, *1*, 893–914. [[CrossRef](#)]
- Veen, D.; Stoel, D.; Schalken, N.; Mulder, K.; Van De Schoot, R. Using the data agreement criterion to rank experts' beliefs. *Entropy* **2018**, *20*, 592. [[CrossRef](#)]
- Nott, D.J.; Xueou, W.; Evans, M.; Englert, B.-G. Checking for prior-data conflict using prior to posterior divergences. *arXiv* **2016**, arXiv:1611.00113.
- Hoijtink, H.; van de Schoot, R. Testing small variance priors using prior-posterior predictive p values. *Psychol. Methods* **2018**, *23*, 561–569. [[CrossRef](#)] [[PubMed](#)]
- Kullback, S.; Leibler, R.A. On information and sufficiency. *Ann. Math. Stat.* **1951**, *1*, 79–86. [[CrossRef](#)]
- Deza, M.M.; Deza, E. *Encyclopedia of Distances*; Springer: Berlin, Germany, 2009.
- Bernardo, J.M. Reference posterior distributions for Bayesian inference. *J. Royal Stat. Soc. B Methodol.* **1979**, *41*, 113–147. [[CrossRef](#)]
- Berger, J.O.; Bernardo, J.M.; Sun, D. The formal definition of reference priors. *Ann. Stat.* **2009**, *37*, 905–938. [[CrossRef](#)]
- Box, G. Sampling and Bayes' inference in scientific modelling and robustness. *J. Royal Stat. Soc. A* **1980**, *143*, 383–430. [[CrossRef](#)]
- Evans, M.; Jang, G.-H. A limit result for the prior predictive applied to checking for prior-data conflict. *Stat. Probab. Lett.* **2011**, *81*, 1034–1038. [[CrossRef](#)]
- Evans, M.; Moshonov, H. Checking for prior-data conflict with hierarchically specified priors. In *Bayesian Statistics and its Applications*; Upadhyay, A., Singh, U., Dey, D., Eds.; Anamaya Publishers: New Delhi, India, 2007; pp. 145–159.
- Cox, D.R.; Hinkley, D.V. *Theoretical Statistics*; Chapman and Hall: London, UK, 1974.



20. Jang, G.-H. Invariant procedures for model checking, checking for prior-data conflict and Bayesian inference. Ph.D. Thesis, University of Toronto, Toronto, ON, Canada, 2010.
21. Evans, M. *Measuring Statistical Evidence Using Relative Belief*; Taylor & Francis: Didcot, UK, 2015.
22. Bernardo, J.M.; Smith, A.F.M. *Bayesian Theory*; Wiley Series in Probability and Statistics; John Wiley & Sons Ltd.: New York, NY, USA, 2000; Paperback.
23. RStudio Team. RStudio: Integrated Development for R. RStudio Inc.: Boston, MA, USA, 2016. Available online: <http://www.rstudio.com> (accessed on 9 April 2019).
24. Drost, H.-G. Package 'Philentropy', Similarity and Distance Quantification between Probability Functions. Available online: <https://github.com/HajkD/philentropy> (accessed on 9 April 2019).
25. Ruckdeschel, P.; Kohl, M.; Stabla, T.; Camphausen, F. S4 Classes for Distributions. *R News* **2006**, *6*, 2–6. Available online: <https://CRAN.R-project.org/doc/Rnews/> (accessed on 9 April 2019).
26. Ali, S.M.; Silvey, S.D. A general class of coefficients of divergence of one distribution from another. *J. Royal Stat. Soc. Ser. B* **1966**, *28*, 131–142. [[CrossRef](#)]
27. Liese, F.; Vajda, I. On divergences and informations in statistics and information theory. *IEEE Trans. Inf. Theory* **2006**, *52*, 4394–4412. [[CrossRef](#)]
28. Cha, S.-H. Comprehensive survey on distance/similarity measures between probability density functions. *Int. J. Math. Model. Methods Appl. Sci.* **2007**, *1*, 300–307.



© 2019 by the authors. Licensee MDPI, Basel, Switzerland. This article is an open access article distributed under the terms and conditions of the Creative Commons Attribution (CC BY) license (<http://creativecommons.org/licenses/by/4.0/>).