UNIVERSITY OF GDAŃSK
FACULTY OF SOCIAL SCIENCES
INSTITUTE OF PHILOSOPHY

Nikodem Lewandowski

Field of Study: Philosophy
ID number: 261338

Exploring the Maximally Sensitive Priors

Master's thesis written under
the guidance of:
dr hab. Rafał Urbaniak, prof. UG

Gdańsk, 2023

Contents

Introduction

We explore the Maximum Sensitivity method (MaxSen) of selecting priors for Bayesian inference in the face of severe lack of evidence (Konek, 2013). The most popular and well-respected principle of dealing with uncertainty in such contexts is the Maximum Entropy principle (MaxEnt), which, roughly, recommends the least informative prior in a given situation. MaxSen, instead, minimizes the need of relying on the epistemic luck of true parameter values (such as chances) falling closely to their prior estimate. We start with a brief outline of MaxEnt and an explanation of Konek's anti-luck approach to priors. Next, we explore his perspective on the theoretical role of priors and illustrate how this viewpoint informs the formulation of MaxSen. Then we explore the impact of the choice of the scoring method on the recommendations made by MaxSen. It turns out that if we use Kullback-Leibler Divergence (KLD) as a distance measure (instead of Cramer-von-Mises, which Konek originally used), the recommendations are the same as that of MaxEnt. Next, we estimate how quickly the recommended prior distribution concentration increases with the planned sample size.

Chapter 1

Bayesian inference and maximum entropy

Bayesian inference is composed of three essential elements: the prior, the likelihood, and the posterior. The prior represents our belief about the parameters of interest before conditioning on data. The likelihood is the probability of observing the data given particular values of the parameters. The posterior represents the result of updating the prior with the new observations. The relation between these elements is given by Bayes' theorem, below in a simple version in which H stands for hypothesis and E for evidence:

$$P(H|E) = \frac{P(E|H)P(H)}{P(E)} \tag{Bayes' Theorem}$$

In the context of this work, we will focus on the scenario of a scientist and an experiment that she is about to conduct, set to adjudicate between some competing hypotheses $H_1, H_2, \ldots$. The prior and the posterior are probability distributions over those hypotheses. What should a prior distribution look like in such a scenario? If a scientist has some knowledge based on an earlier experiment, then perhaps the situation is not too problematic, as she should incorporate this evidence into her prior.[1] In this paper we focus on sever uncertainty: the challenge which arises when there is no evidence about the issue at hand.

Subjective Bayesians claim that if there is no reason to favor any distribution over the

[1] Eliciting probabilities even from informed experts is still a challenge (O'Hagan et al., 2006), but we do not discuss it in this paper.

hypotheses, any distribution is equally reasonable. One might suggest another, rather mysterious answer, that a scientist could base her choice of a prior on her hunch. An objectivist, instead, insists that we should exclude any strong assumptions, hunch or intuition if they do not result from our prior experience in the matter, because they might only disrupt our inferences. Therefore, the objectivist is after some kind of regulative principle that serves as a rational constraint on priors, according to which the information incorporated into the prior in severe uncertainty is in some clear sense minimal.

Typical approaches to model priors in such situations are to use a non-informative or a weakly informative prior (Gelman, 2009). The non-informative prior should be as minimally informative as possible. Roughly speaking, on this approach a rational agent under severe uncertainty should choose a distribution that maximizes information entropy, that is, minimizes the information encoded in the distribution—hence the name MaxEnt. A weakly informative prior, instead, incorporates some level of regularization, for instance, centering the marginal density for a regression parameter around 0 with standard deviation .5, assuming the data have been standardized.

The proponents of non-informative priors argue that, since we do not possess any meaningful information, the prior of choice should not convey any information whatsoever. The advocates of weakly informative priors argue that non-informative priors also contain assumptions and can often lead to the data having too much influence on the posterior, leading to overblown posterior estimates, especially when the data set is not large or contains quite a bit of noise.

There are two main classes of arguments for adopting MaxEnt, theoretical and pragmatic. The theoretical arguments are based on the premise that the central role of priors is to represent a researcher's epistemic state. If she is completely uncertain about the distribution over the hypotheses, she ought to adopt a minimally informative prior that mirrors this uncertainty. The pragmatic argument is simple and modest in nature: the use of MaxEnt recommended distributions is abundant in many empirical sciences (for instance, it is often assumed that the distribution of residuals is normal), and it works quite well.

There are some concerns. The theoretical argument MaxEnt is somewhat similar to Laplace's Principle of Insufficient Reason, and the uniformity of a distribution is not preserved under arbitrary transformations of random variables. For instance, a uniform distribution over the values of $X \in [0, 1]$ will not lead to a uniform distribution over the values of X^2, and so already some information has to be incorporated at the point when the researchers decide which variables she wants to accept uniform priors about. This seems to disagree with the idea that minimally informative priors somehow represent a pristine *tabula rasa*. Another problem is that a uniform distribution is susceptible to over-fitting. The model tends to be over-fitted when it is too flexible and too sensitive to data, in which case the noise can have to much impact on it. Such a model will adjust itself to irrelevant patterns in data, because of how easily it responds to observations. This, in fact, is closely related to what will be discussed in the paper, and we will get back to the issue later on.

While regularization is quite widely practiced, sometimes defended in terms of over-fitting-avoidance, and sometimes in terms of the researcher actually having prior reasons to think that extreme parameter values are less likely, the question of whether it can be argued for (and whether its specific form can be derived) in terms of some general principles remains open. Roughly speaking, Konek's approach is a principled approach to criticizing MaxEnt and deriving some form of regularization from general epistemological principles. The goal of our paper is to study this proposal. We will discuss the consequences of implementing MaxSen, including its dependence on the sample size. Also, we will examine different potential scoring methods to be employed within the MaxSen framework.

Some caveats before we move on. First, through the paper for computational ease, we will focus on discrete hypotheses grid-approximating a continuous parameter space. All the conceptual points we will make apply when this restriction is raised.[2] Second, the uniform distribution is recommended by MaxEnt only if nothing is known about the random variable. If, for instance, we know that we are dealing with a continuous variable with finite variance, MaxEnt (formulated in terms of differential entropy) recommends the normal distribution instead. If, instead, we know we are dealing with a discrete distribution of counts with two unordered possible events at each observation and constant expected value, the recommended distribution is binomial. In what follows, however, we

[2] Of course, there is a sense in which entropy can be applied to continuous hypothesis spaces, although it is not a fully straightforward extension of Shannon's entropy. A straightforward extension would assign infinity to any continuous distribution. While conceptually this makes sense— with infinite precision comes infinite informativeness—this approach would make a comparison of such distributions uninteresting. If we ignore the part of the continuous version of the entropy formula which tends to infinity, we are left with differential entropy. We will prefer to discretize continuous distributions with fixed levels of precision so that we can make sensible comparisons between grid-approximated continuous and other discrete distributions in a convenient manner.

will be working with finite sets of hypotheses (grid-approximating parameter values) and distributions over them, in which case the MaxEnt-recommended distribution is uniform.

Thus, for the context of this paper, we can state the recommendation given by MaxEnt a bit more clearly: in severe uncertainty about the hypotheses $H_1, \ldots, H_n$ about population frequency, the prior should be the distribution with maximal entropy, that is, it should be the uniform distribution.[3]

With this understanding of MaxEnt's recommendation for severe uncertainty, let us now focus on the alternative approach proposed by Konek. Firstly, we will explore Konek's rationale for his method, followed by an examination of the method itself.

[3] In fact, the uniform continuous distribution maximizes differential entropy as well.

Chapter 2

The anti-luck approach to priors

Konek proposes an *instrumental account of priors*. He argues that the priors have a more important role than just being used as representation of epistemic states:

> “ I suggest that the central role of priors is to help us secure accurate posterior beliefs and to minimize our need for epistemic luck in securing those beliefs. (Konek, 2013, p. 3) ”

The theoretical role of priors, on this account, is to optimize for accurate posterior beliefs while minimizing the impact of epistemic luck, if possible. Thus, Konek rejects the view that the main role of priors is to represent an agent’s epistemic state in radical uncertainty.

First, let us focus on **epistemic luck**. Having accurate posteriors can be accidental. I could have randomly favored one hypothesis without any research whatsoever and still be lucky enough to be right about it. Yet, according to the rather plausible *incomaptibility thesis* (Hetherington, 2019), epistemic luck is incompatible with knowledge. The view stems from an extensive debate surrounding Gettier-style cases in epistemology (Gettier, 1963), in which the subject, due to some coincidence, seems to have justified true beliefs, but fails to have knowledge. One of the explanations on the market is that in such cases, the subject is just epistemically lucky in being right (Engel Jr., 1992), which, arguably, undermines their claim to knowledge.

Two kinds of epistemic luck ought to be distinguished: *environmental* and *intervening*. Environmental luck has a passive influence on an agent’s success. The success of a pro football player in scoring a goal is explained by her skill, but of course, she was lucky in that e.g. her shoe did not fall off or that she was not hit by the water bottle thrown by a fan of the opposing team. Intervening luck, however, destroys the simplicity of the causal link between skill and success. It is the kind of luck in which for instance a ball kicked by a pro football player was hit by the water bottle in the mid-air losing its trajectory, but luckily happened to hit the referee in the head and finally landed in the gate. In such a situation, the success cannot be explained by the skill of a football player. The kind of epistemic luck that Konek suggests epistemic agents should strive to avoid is intervening luck.

Crucially, priors might be subject to intervening epistemic luck: true chances, true probability or the truth of a proposition might simply happen to fall in an appropriate sense close to the prior distribution. Priors that are more susceptible to such luck than others seem less reasonable.

What makes priors susceptible to this kind of luck? This susceptibility is linked to the distribution's *resiliency* to data. A prior that is more resilient with respect to a fairly wide range of data tends to be more luck-dependent. If a prior distribution p is resilient to a datum D, then p conditioned on D is close to p. A resilient prior is less susceptible to over-fitting and changes in its shape in the face of new data. For instance, concentrated prior distributions are far less flexible and more resilient than the uniform distribution, and so they are far more dependent on the luck that real chances happen to fall close to the posteriors reached from them. Figure 2.1 illustrates the difference between the resiliency of the uniform distribution on the left and of a more concentrated distribution on the right.

The first observation is that the uniform prior changes dramatically, and so is very susceptible to new data. On the other hand, the hunch-based prior, represented with a continuous red line on the right-hand side did not move so much the dashed green line that represents the posterior is not too far from the prior. The hunch-based prior is much more resilient to this data.

The second observation is more apparent when we compare the posterior on the second plot with the alternative posterior represented by the long-dashed blue line. This line represents the interaction of the concentrated prior with the "opposite" data (2 Heads and 8 Tails). The resulting high accuracy of the first posterior, in an example with the biased prior, comes from a lucky guess. By luck, the prior bias was close to the observed data. The resilience of a prior is correlated with its susceptibility to intervening luck, the luck that chances happened to be close to the prior distribution.

Konek argues that in adopting a prior distribution, a rational agent ought to choose the prior that best meets the primary theoretical role of priors, that is, optimizes for accuracy

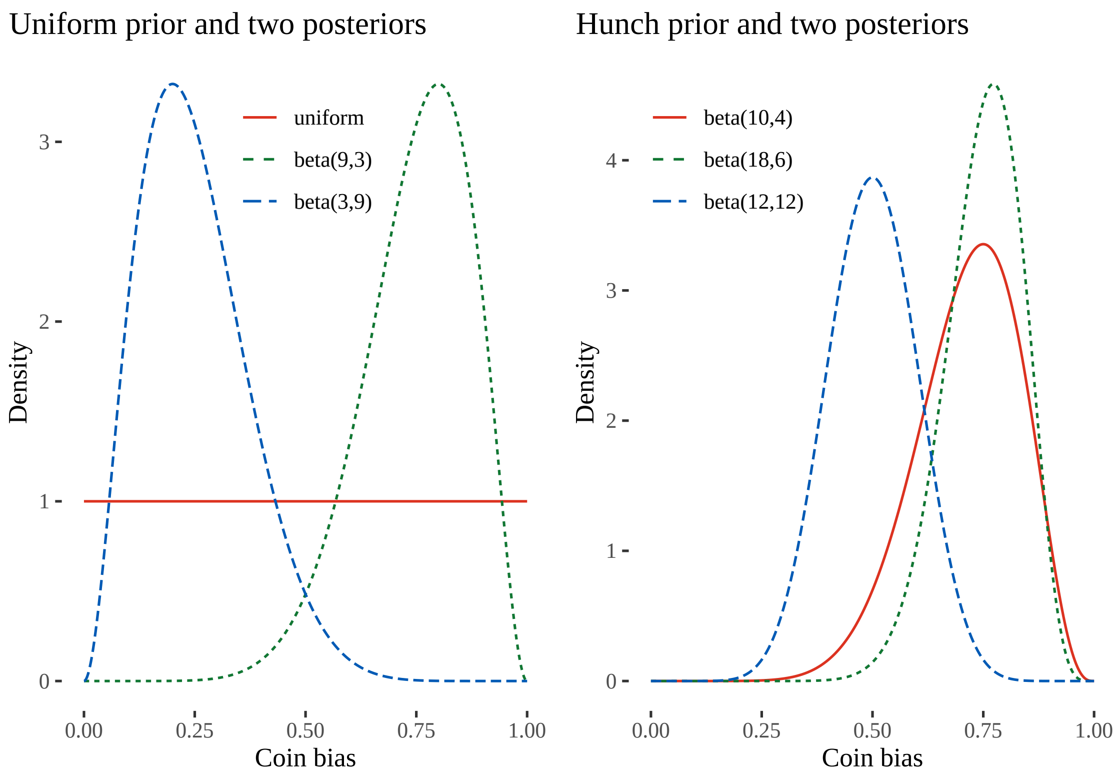

Figure 2.1: The continuous red line on the first plot is the distribution that MaxEnt recommends in the no-evidence scenario. It is the uniform beta distribution ($\alpha = \beta = 1$). The prior on the second plot represented with a continuous red line is a lucky guess, a beta distribution with $\alpha = 10$, $\beta = 4$. The short-dashed green lines represent the posteriors constructed upon observing 10 tosses resulting in 8 heads and 2 tails. The long-dashed blue lines represent alternative posteriors obtained after observing 2 heads and 8 tails.

while avoiding susceptibility to luck. So far we talked about luck, let us turn to **accuracy** now.

To make a comparison of two different distributions, Konek uses the Cramer-von Mises distance (CvM). It is, roughly speaking, calculated by taking the squared distance between two cumulative distributions.

$$\mathscr{C}(p,p_D) = \int_0^1 [P(x) - P_D(x)]^2\, dx \tag{CvM}$$

The measure is similar to the Brier score. Both methods rely on squared distance, the main differences are that the Brier score works with squared differences of probabilities and takes a sum, while the Cramer-von Mises works with squared differences in cumulative probabilities and integrates. As we want to facilitate computation without hindering insight, and to compare distributions of various shapes, including ones that assign probability 1 to a single parameter value (preferably without using Dirac-Delta), we will often follow the practice of using grid approximation with fixed precision levels even in the calculation of CvM, simply making sure that the the precision level remains the same for all the comparisons, and using summation instead of integration in such approximations. For now, we will stick to integration in order to closely follow Konek's work first.

Using (CvM) we can measure the resiliency of our two prior distributions. If we calculate these using the official definition and integration, the resulting values are $\mathscr{C}(u,u_D) = 0.093$, and $\mathscr{C}(l,l_D) = 0.001$.[1] The outcome is a distance between the prior and its posterior, which depends on two factors: the prior's shape and the data. For resilient distributions, this distance will be small, because their spiky prior shifts reluctantly, and the uniform prior will move significantly when updated by the same data. Thus, the distance between the prior and the posterior can be used to gauge the resiliency of a given prior to the data.

A posterior can be also scored by its distance to the true hypothesis (Konek uses CvM here as well)—the inaccuracy of a distribution. If we calculate this value for each particular possible chance hypothesis, thinking about the result as a function of the true chance

[1] If we use grid approximation of 1001 evenly spaced values, we obtain $\mathscr{C}_{disc}(u,u_D) = 92.832$, and $\mathscr{C}_{disc}(l,l_D) = 3.429$. While the exact values are much larger (as expected when summing a thousand squared distances), the proportional differences are similar (around 22 vs. 27), and moving to the approximation preserves ordering. Assuming that in a given calculation we stick to one of the methods, the exact values do not matter that much for the purposes of comparison.

hypothesis, we obtain what Konek calls the *objective expected inaccuracy*. Its shape is explained by the two factors:

- The internal factor—the prior's resiliency, that is, its responsiveness to new data. A prior that does not eagerly move will not adjust to new evidence, which affects its accuracy.
- The external factors, such as the proximity of the coin's real bias to the prior distribution.

These two factors are connected. Resilient priors, if they are postulated without any evidential ground, strongly rely on a special kind of intervening luck, which qualifies as an external factor. According to Konek, priors are resilient when their CvM distance to updated distribution ($\mathscr{C}(p, p_D)$) is close to zero.

However, in the selection of priors, it is not directly resiliency, but rather (expected objective) inaccuracy that plays the key role in Konek's proposal. In this approach, being as close to the truth as possible is the one and only governing principle, and the goal is to make this proximity insensitive to luck. Let us now come back to the plotted examples with coins.

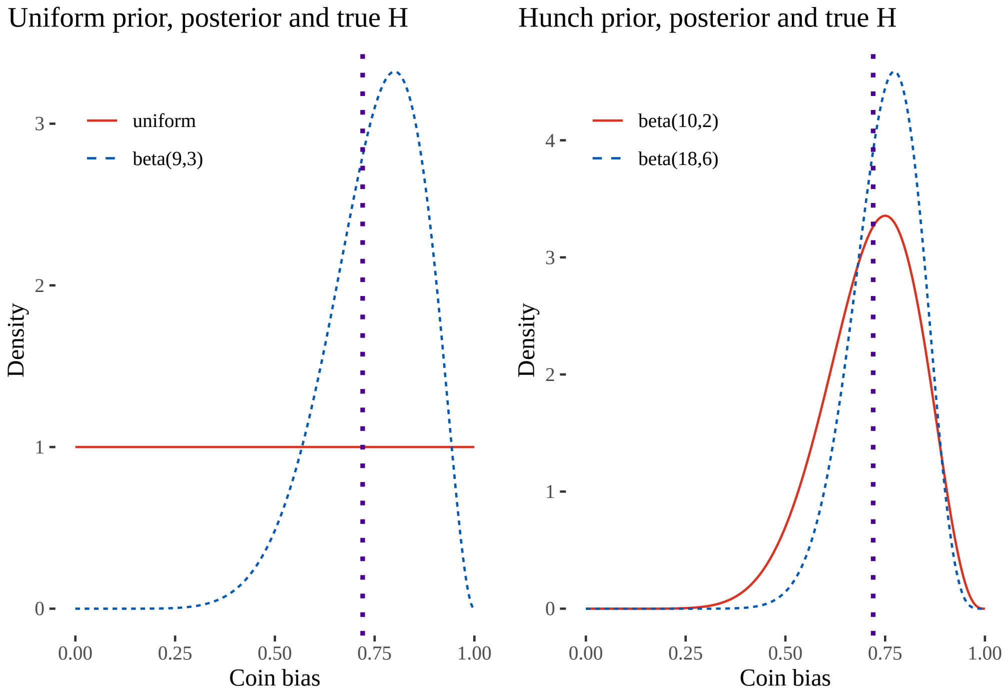

Figure 2.2: The plot on the left represents the uniform prior, its posterior and a true hyphotesis, the plot on the right starts with the biased prior. The red continuous lines represent priors, the blue short-dashed ones represent the posteriors. The straight purple dotted lines represent the true hyphotesis *h*.

The plots in Figure 2.2 illustrate our two priors, lucky and uniform, and posteriors they lead to after observing the same data (8 Tails, and 2 Heads). The purple, dotted, straight line represents the assumed true hypothesis h from the set of all possible hypotheses H, according to which *the bias of the coin is equal to 0.72*. We now use (CvM) to measure the distance between the hypotheses h and cumulative posterior distributions u_D and l_D: $\mathscr{C}(u_D, h) = 0.035$ and $\mathscr{C}(l_D, h) = 0.026$. The smaller the score, the better, and so the lucky posterior is closer to the true probability of the coin. However, this does not mean that the lucky prior is more reasonable than the uniform one. This proximity is explainable by intervening luck. In an alternative scenario, an alternative and not so lucky outcome would result in a much less accurate posterior.

For this reason, the whole spectrum of possible distances to the alternative possible true hypotheses should be taken into account. What Konek suggest is exactly this: take into account the objective expected posterior accuracy for a range of possibilities. Using the binomial distribution the formula for the (CvM) distance between a posterior distribution u_D and a true hypothesis H (distribution) conditional on the true chance being h, for D consisting of k successes in n observations is as follows:

$$\mathsf{OEI}_{\mathsf{CvM}}(u,h) = \sum_{k=0}^{n} \binom{n}{k} \times h^k \times (1-h)^{n-k} \times \mathscr{C}(u_D, H) \tag{2.1}$$

That is, to calculate $\mathsf{OEI}_{\mathsf{CvM}}(u,h)$ for each possible outcome D:

- calculate its probability using the binomial formula, assuming the true chance is h, obtaining a weight,
- update u with D, obtaining u_D and calculate the (CvM) distance of u_D from the cumulative distribution corresponding to the true chance being h,
- multiply the distances by their corresponding weights and sum them up. The shape of the objective expected posterior inaccuracy—Konek claims—helps in assessing the epistemic value of a prior, as it does not require any knowledge about what the true hypothesis actually is, and allows us to gauge the sensitivity of (in)accuracy to what the true chance is.

For instance, the OEIs for the priors we used in our running example are illustrated in Figure 2.3. The first plot represents the expected inaccuracy of a uniform prior in a scenario of 14 coin tosses (trials). The plot on the right represents the OEI of the hunch based prior

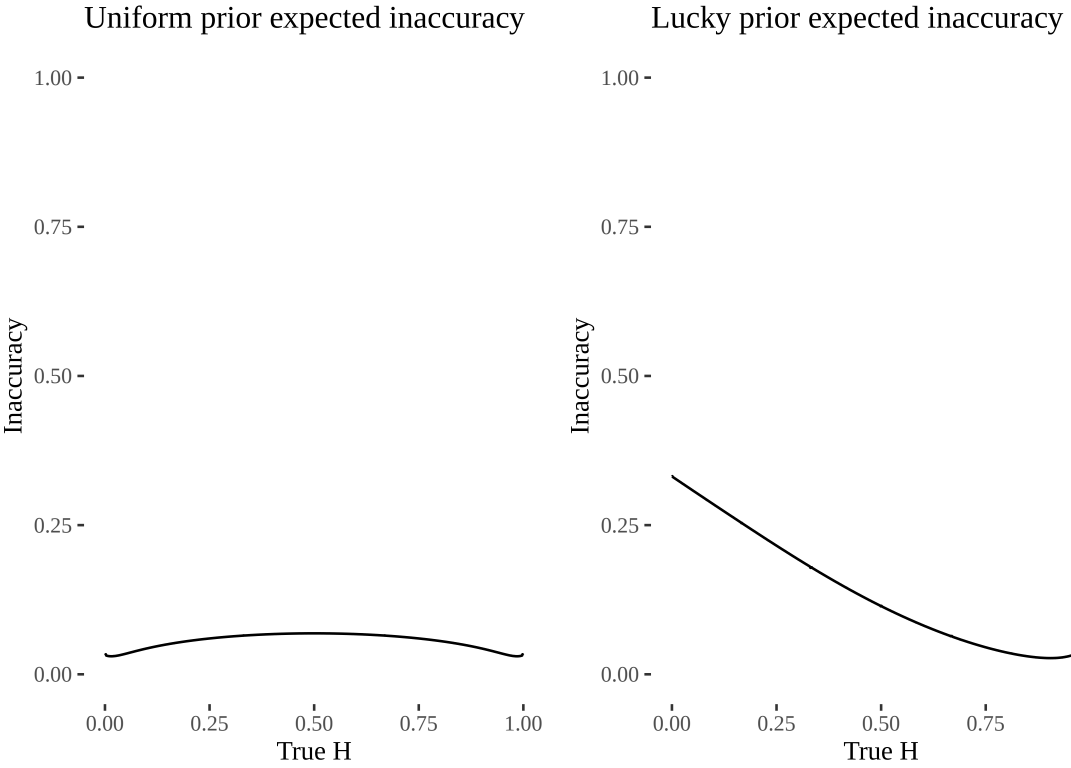

Figure 2.3: Two plots representing objective expected innacuracies of: a uniform prior on the left, and a hunch based prior on the right.

$\alpha = 10, \beta = 2$ in the same scenario. Keep in mind that these measurements represents the expected inaccuracy of posteriors built upon those priors, and not the inaccuracy of the priors themselves.

OEI for the uniform distribution stays fairly constant across possible true hypotheses and is symmetrical. For the lucky prior, however, we observe a high degree of inaccuracy for low values of the true probability, and inaccuracy similar to the one for the uniform prior as the true hypothesis gets closer to $h = 0.72$. This illustrates how it was a pure luck that the hunch based prior led to a slightly less inaccurate posterior. Indeed, this is possible if the true bias happens to be close to the value around with the prior is centered, but the price is that inaccuracy grows much faster than for the posteriors learned from a uniform prior, if the true chance is further from the center of the hunch-based prior distribution.

In most cases, however, it is the uniform prior and not the resilient one that will lead to a less inaccurate posterior. A general observation is that MaxEnt-recommended uniform prior is less susceptible to intervening luck when comparing to lucky spiky beta priors. However, note that the OEI for the uniform prior is not flat. The question is whether a choice of priors is possible which would make OEI even flatter. Then, from the perspective of taking the primary role of priors to be the improvement of accuracy, such a prior would minimize the impact of intervening luck, and would come recommended.

Chapter 3

Maximum sensitivity

MaxSen recommends a prior that is minimally susceptible to intervening luck. At least following the methods used by Konek, identifying the recommended prior takes some computation. The method used by Konek for beta distributions is as follows:

- For any combination from a selection of values of α and β and a given planned sample size,[1] compute the corresponding OEI.
- For each such an OEI, compute its curvature (using some sensible curvature measure).
- Use quadratic approximation of curvature as a function of α and β.
- Use the quadratic approximation to find the values of α and β that minimize this curvature.

As for each particular choice of α and β the calculation of OEI is computationally demanding, the approximation step is practically important.[2] Later on, we will argue that the choice of a quadratic function is sub-optimal once one takes into sample sizes that go above 100.

So here are the key moves that we make in our simulation-based approach, which we first use to successfully replicate Konek's claims, and are about to deploy in further investigations. First, we discretize the range of hypotheses, that is, we look at 1001 evenly spaced chance hypotheses between 0 and 1. Second, we discretize the options for shape parameters—on some occasions, the results will be good enough to guide us to decent

[1] Which prior is recommended by MaxSen depends on the sample size, we will come back to this issue.

[2] An analytic approach to the whole problem has not been developed. Perhaps trying to find an analytic method is worthwhile. But we think that a simulation-based study should precede such a search.

alpha	beta	n	CurvK
1.2	1.2	5	0.0169
1.5	1.5	10	0.0140
1.8	1.8	15	0.0103
2.0	2.0	20	0.0096

Table 3.1: MaxSen recommended beta distributions for sample size n, assuming CvM as a scoring rule and curvature measure CurvK. The grid step size is 0.1.

values, on some occasions, we may still further prefer to further approximate with a quadratic function. Third, as which prior is recommended by MaxSen depends on the expected sample size, we only consider some values of n for the sake of illustration.

Moreover, even once we restrict ourselves to this somewhat simplified set-up, there are important choices to be made. First, we need to decide which scoring rule to use. For now we will stick to Konek's rule of choice based on Cramer-von-Mises distance. Later on, we will introduce and use another rule to investigate how sensitive the recommendations are to such a choice. Second, we need to decide how to measure how uneven the OEI is. Konek's choice is rather crude, as he uses the difference between the maximum and the minimum of a given function:

$$\mathsf{curv}_k(\mathsf{prior}) = \max\left[\mathsf{OEI}(\mathsf{prior})\right] - \min\left[\mathsf{OEI}(\mathsf{prior})\right] \tag{CurvK}$$

Konek does not firmly motivate this particular method of scoring curvature, and given that it is only based on two values, a more fine-grained approach might be desirable. We will propose such a measure, which will moreover be based on information-theoretic considerations. For now, we will stick to CurvK.

Figure 3.1 illustrates a set of beta distributions and their curvature scores. This plot exemplifies how MaxSen works in practice. These distributions are candidates for a maximally sensitive prior, curvature indicates CurvK value of their OEIs (for $n = 5$ and CvM used as a scoring rule). The prior recommended by Konek's method is $\alpha = \beta = 1.2$. In fact, sensitivity is always maximized by a distribution where $\alpha = \beta$, which is also illustrated on the plot by the fact that the values are arranged in a symmetrical pattern and minimal values fall in the middle.

Quite importantly, which $\alpha = \beta$ come recommended depends on the planned sample size. In Table 3.1 we illustrate how the recommendations change with $n \in \{5, 10, 15, 20\}$.

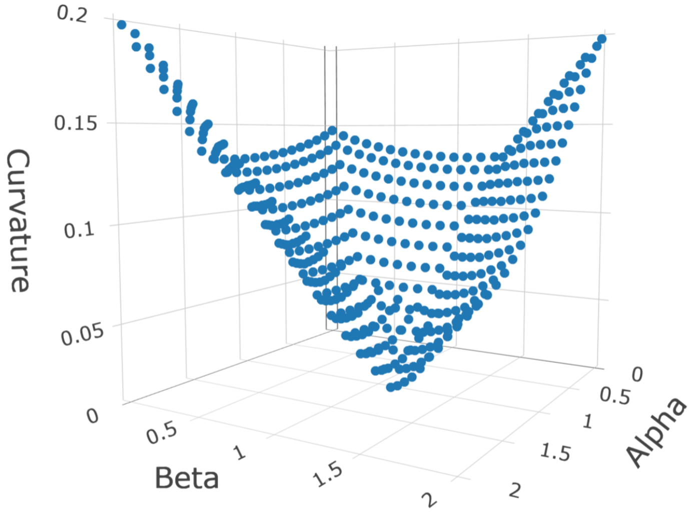

Figure 3.1: Curvature for a discretized range of possible values of α and β, restricted to (0, 2) with grid step=0.1. These values determine prior distributions, which with sample size (here, $n = 5$) and the choice of a distance measure (here, CvM) determine OEI's. These, with a curvature measure (here, CurvK) determine curvature scores, which are on the vertical axis. The OEI with the lowest curvature is the one that maximizes sensitivity in Konek's sense. In this example it is $\alpha = \beta = 1.2$.

For now, we work with small samples, but later on, we will more extensively test two hypotheses: (1) As n increases, so do the MaxSen-recommended $\alpha = \beta$. The functional form of this relation is not clear. (2) As n increases, sensitivity decreases.

Let us reflect for a moment here. Why should the prior with the lowest OEI curvature be considered maximally sensitive and rational? Konek argues, relying on his theoretical role of the priors, that we want the data to explain the posterior distribution's accuracy, and not the external factors related to a researcher's hunch. Low OEI curvature indicates that there is no such bias towards any particular chance hypothesis and indicates that the prior is more or less equally sensitive to new data whatever the actual observations will be. For this reason, arguably, it is resistant to intervening luck.

Quite interestingly, uniform priors do not maximize sensitivity when it's measured using Konek's combination of CvM-CurvK, which one might find somewhat surprising—as this suggests that the distribution that is more **equally sensitive to evidence whatever evidence will be** is not the one that is maximally sensitive to evidence in general. In the next section we investigate if this fact holds for a different scoring rule and a different curvature measure, though.

Before we move on, two more issues with MaxSen that we mention in passing. One is that due to the lack of a known analytic solution, this method requires quite a bit of computation, especially when we consider continuous beta distribution parameters and significantly larger sample sizes, and so systematic use thereof would require either analytic solutions or better approximation methods. Another is that for many applications, this is not a practically relevant recommendation, as the washing out theorem (Joyce, 2004) suggests that as soon as the data sets are interestingly large, the difference between the posteriors recommended by MaxSen and MaxEnt are not going to matter. Third, we have not even started deploying this in multi-dimensional settings where the distributions involved are not beta distributions. Even before we start talking about such issues, other interesting phenomena and problems arise, to which we now move.

Chapter 4

Choice of measures and their impact

4.1 Scoring with KLD

OEI calculations for a given prior require a distance measure. While Konek's method of choice is CvM, other mathematical tools can be used as well. One such measure, compelling due to its information-theoretic motivations is as Kullback-Leibler divergence (KLD) (Kullback and Leibler, 1951). In this section, we will investigate the consequences of using KLD instead of CvM. As it turns out, the end result is importantly different, as **MaxSen in this configuration recommends the same prior as MaxEnt.**

The Kullback-Leibler divergence (KLD), also known as relative entropy, is a commonly used tool for comparing a model distribution to a true, unknown distribution, as it measures the information lost when approximating the true distribution with the model distribution. Since we are working with grid approximation, we will continue to use summation instead of integration.[1] For two distributions P and Q over the same hypothesis space, where i ranges over all possible hypotheses, the KLD from Q to P is:

$$\mathscr{K}\mathscr{L}\mathscr{D}(P \parallel Q) = \sum_i P(i) \log \frac{P(i)}{Q(i)} \tag{KLD}$$

Arguably, given that KLD is a more theoretically grounded measure than CvM because it

[1] To apply the KLD measure, we have used the `kullback_leibler_distance()` function from the `philentropy` package.

is rooted in information theory and has a clear statistical interpretation, KLD provides a more robust and better motivated solution for comparing distributions.

It is essential to note that from this point on, we will assume that the beta distribution that maximizes sensitivity is always symmetrical, i.e., $\alpha = \beta$. We will not include skewed beta distributions since we did not observe any unsymmetrical beta distribution that would be suggested prior using any of the methods we have used. However, please be aware that this is an assumption, and we do not have a proof of this at the moment.

Let us run some examples to compare what happens when we switch CvM for KLD in the deployment of MaxSen. Figure 4.1 illustrates some OEI curvature measurements (with CurvK) for a range of sample sizes. Since computations become cumbersome for larger values of n, we selected a few values not exceeding 100. For each of those values, we will identify the $\alpha = \beta$ parameters that minimize the curvature.

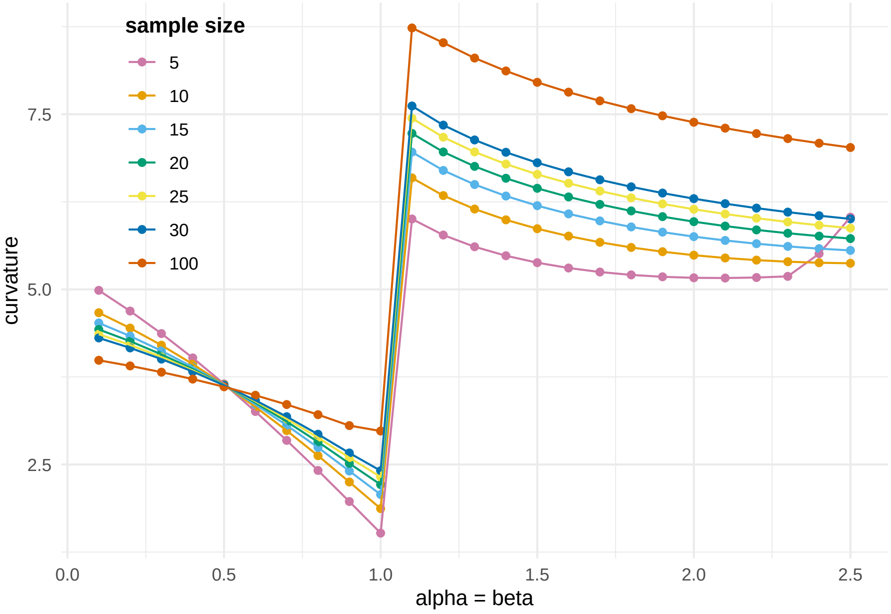

Figure 4.1: The plot shows seven different MaxSen searches for given sample sizes (n) with a grid approximation of 0.1. For all sample sizes, the recommended prior is $\alpha = \beta = 1$ since it consistently exhibits the lowest OEI curvature. The points represent real measurements, while the lines are included only for enhanced visibility.

Figure 4.1 illustrates an intriguing phenomenon: MaxSen with KLD, instead of CvM, recommends the uniform priors ($\alpha = \beta = 1$) as they maximize sensitivity, at least for $n \in \{5, 10, 15, 20, 25, 30, 100\}$. While we do not have a general proof that MaxSen with KLD recommends a uniform prior for every sample size, it holds true for every sample size we tested.[2] This suggests that MaxSen's recommendations are not stably (in the sense of being preserved under a sensible range of scoring rules) different from that of MaxEnt.

4.2 Scoring-rule/curvature configurations

In the previous section, we observed that the selection of a scoring rule significantly impacts whether MaxSen recommends regularization instead of following MaxEnt's recommendations. Now, let's explore another parameter: the curvature measure. Konek utilizes CurvK, which captures the difference between the maximum and minimum values of OEI. However, this measure is not very fine-grained as it relies on only two data points of OEI. To address this limitation, we examine the implications of replacing CurvK with an information-theoretic curvature measure, denoted as CurvE. The objective of this section is to compare the results obtained by combining the choice of curvature measure with the selection of a scoring rule in all configurations: KLD/CvM with CurvK/CurvE.

CurvE measures how much less entropy the given distribution has compared to a uniform distribution:

$$\mathsf{curv}_e = 1 - \frac{H[\mathrm{OEI}(\mathsf{prior})]}{H(\mathsf{uniform})} \tag{CurvE}$$

where $H[\mathrm{OEI}(\mathsf{prior})]$ is the entropy of normalized OEI corresponding to the prior in question, and $H(\mathsf{uniform})$ is the entropy of a uniform distribution as calculated on the same discretized grid. CurvE ranges from 0 to 1, with higher values indicating a greater deviation from the uniform distribution and, therefore, greater curvature. CurvE, being sensitive to the entire shape of the distribution appears to be a more justifiable measure of curvature than CurvK.

Figure 4.2 illustrates three priors recommended by different methods for a sample size of $n = 5$. Two variations of MaxSen, one utilizing CurvK and the other incorporating CurvE, employ CvM as a distance measure. The third prior is recommended by MaxEnt. CurvK recommends a prior of $\mathsf{beta}(1.19, 1.19)$, CurvE recommends $\mathsf{beta}(1.32, 1.32)$, and MaxEnt recommends a uniform distribution. On the left side of the figure, we can see the distributions of their OEIs while on the right side, the beta distributions of those priors

[2] We tested a sequence of values ranging from 0 to 100 with an increment of 5, as well as additional values of 150, 200, and 500.

are displayed.

The overall shape of the OEIs for these priors is relatively similar. One observation is that the marginal values of the uniform prior are less inaccurate compared to MaxSen CurvK. On the other hand, the CurvE prior exhibits a significant discrepancy in accuracy between its marginal values and the middle values. This comparison highlights the functioning of CurvK, Konek's curvature measuring method. One common characteristic of CurvE and the uniform prior is the large difference between their maximum and minimum values, which CurvK aims to minimize. Therefore, the CurvK prior has the smallest possible difference between these two values.

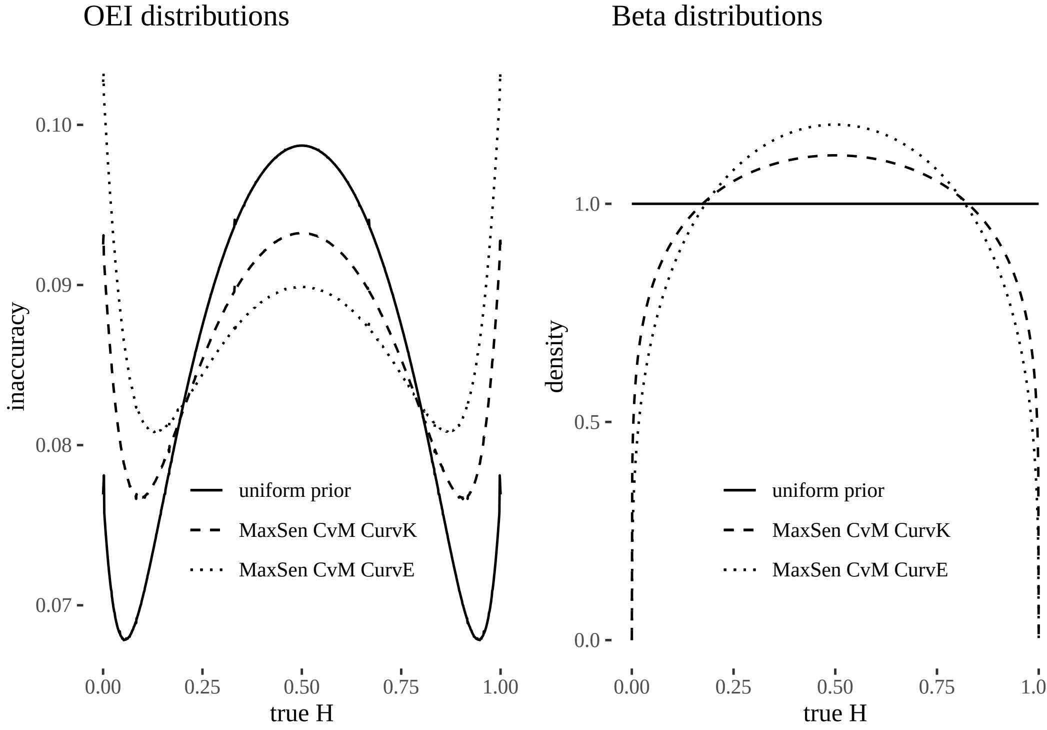

Figure 4.2: Shape comparison of three OEIs is shown on the left. The dashed line represents MaxSen CurvK ($\alpha = \beta = 1.19$), the dotted line represents MaxSen CurvE ($\alpha = \beta = 1.32$), and the continuous line represents the uniform prior OEI recommended by MaxEnt. The OEI was obtained with an approximation to 0.01, and the number of trials, n, is equal to 5. The plot on the right represents the beta distributions of those priors. The dashed line corresponds to Konek's MaxSen, the continuous line represents the uniform distribution, and the dotted line represents the CurvE MaxSen prior.

On the beta distribution plot in Figure 4.2, the priors appear markedly different. The shape of the uniform prior is a straight line parallel to the x-axis. In contrast, the MaxSen priors exhibit a bell-like shape, with marginal values at 0 and the majority of values slightly

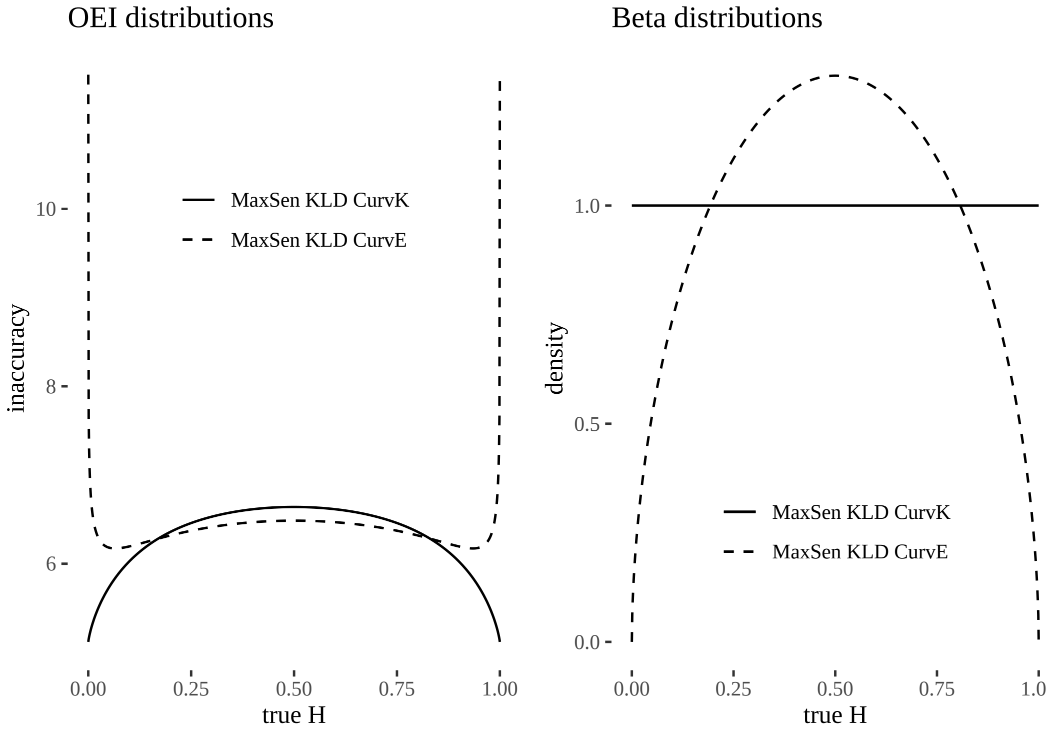

Figure 4.3: Shape comparison of two OEIs is shown on the left. The dashed line represents MaxSen KLD CurvK ($\alpha = \beta = 1$), and the continuous line represents MaxSen KLD CurvE ($\alpha = \beta = 1.55$) expected inaccuracy distribution. The OEI was obtained by rounding to two decimal places, and the sample size is 5. The plot on the right represents the beta distributions of those priors.

above the uniform distribution. The CurvK and CurvE distributions share a similar shape. The bell-like shape of the MaxSen prior may be the most optimal for adjusting its position based on new data, which could explain why the MaxSen method does not recommend the uniform prior.

Now, let's consider a similar test, but this time we will not use the CvM as a measure of distance. Instead, we will utilize the Kullback-Leibler divergence (KLD). In the visualizations shown in Figure 4.3, once again, the left side displays the distribution of OEIs, while the right side showcases the beta distributions of the respective priors. Here, you can observe two priors recommended for $n = 5$ using CurvK and CurvE, both of which employ KLD. It is important to note that there are only two priors this time, as we discovered in the previous section that MaxSen KLD with CurvK recommends uniform priors (at least for the tested values of n).

The CurvE OEI exhibits an unusual shape where its marginal values are significantly distant from the rest. It is likely that all priors generated by MaxSen KLD have these outlier-like marginal values, which is why CurvK, being sensitive to disparities between the maximum and minimum values, recommends the uniform prior. The beta distributions plot is less surprising. The CurvE prior is similar to the priors we observed before, with the only difference being that it appears to be more peaked, and more values are above 1 (which is indicated by the uniform prior) compared to the previous examples

To see how the different approaches to MaxSen change the recommendations, refer to Figure 4.4, which shows the various n sizes outcomes for all the configurations mentioned in this section. One striking observation is that the recommended α and β values seem to increase with the number of observations, n. This is a problem that will be explored in the next section.

4.3 Impact of sample size on MaxSen priors

In the previous sections, we observed that one of the characteristics of MaxSen recommended priors is that the $\alpha = \beta$ parameters increase with the planned sample size (this is not the case when using MaxSen with KLD and CurvK, as this combination always recommends the uniform prior). This similarity to classical hypothesis testing methods, which are sensitive to stopping intentions, may raise concerns.

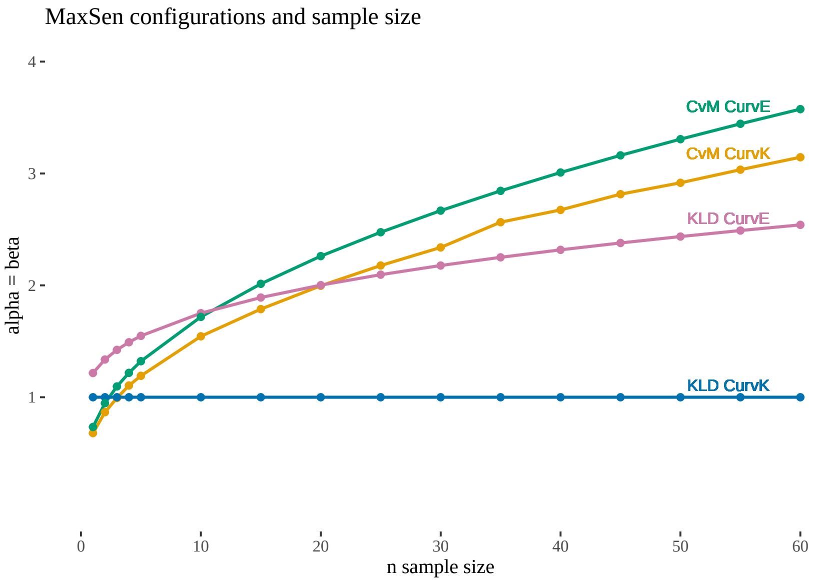

Figure 4.4: The four lines represent 4 MaxSen methods and the recommended $\alpha = \beta$ depending on the n sample size. Except for KLD CurvK, all the methods recommend priors whose parameters increase with the sample size. The smoothed lines goo through the point values obtained for $n = \{1,2,3,4,5,10,15,20,25,30,35,40,45,50,55,60\}$ a (non-analytic) gradient descent.

In many cases, the sample size is not known prior to the experiment and can be difficult to predict. Some experimental designs (Kruschke, 2015) might make it even in principle impossible to predict the sample size prior to the experiment (a simple example is flipping a coin until observing 30 heads—you can't predict how many tosses you will need to make). In such situations, where the stopping intention influences the data, researchers following Konek's recommendation would be forced to predict the sample size in order to construct a prior distribution. A significant concern arises when the method leads to a concentration of the prior distribution for large sample sizes. Many real-life applications of Bayesian methods involve much larger sample sizes than those we have discussed, making it crucial to investigate what MaxSen (and its variations) recommend for large values of n.

Figure 4.5 illustrates some MaxSen priors (using CvM and CurvK) and provides information on the relationship between sample size and beta distribution parameters. Those calculations are represented with points. It is important to note that these calculations were done with the gradient descent algorithm, introducing some degree of uncertainty to the values. At least in some proximity to the values for which approximations have been obtained, the relationship between the shape parameters and the sample sizes is close to exponential (models with a few formulaic shapes, such as logarithmic or polynomials were also tested, but displayed worse performance). A Bayesian model that predicts the $\alpha = \beta$ based on the sample size n is visualized as a line, where the orange line represents the mean prediction and the gray area represent the 89% highest posterior density interval (HPDI) of the predictions.

The approximating learned function that best describes the growth of the recommended α and β parameters with the sample size (the exponent $1/2.5$ is the result of an optimization) n is:

$$f(x) = 0.61 \cdot n^{\frac{1}{2.5}}$$

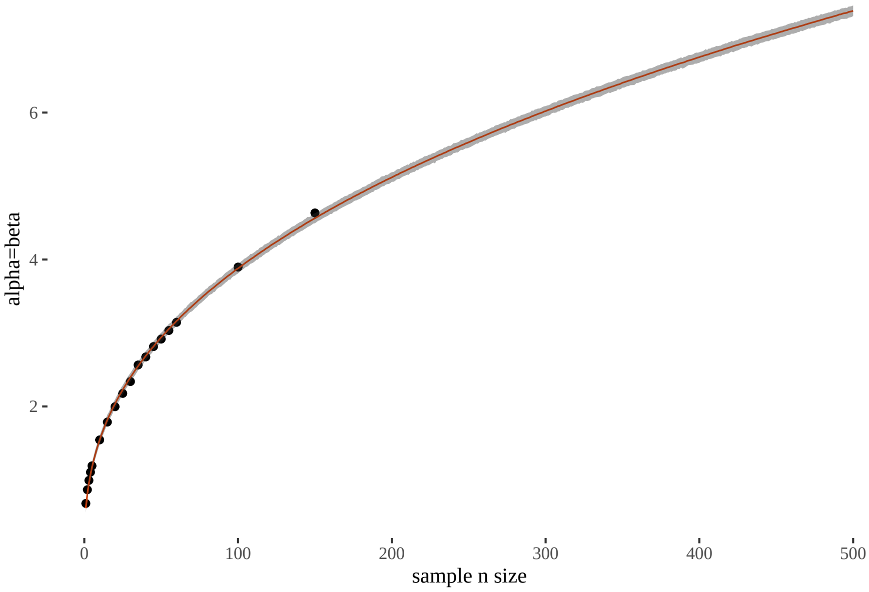

Figure 4.5: MaxSen (CvM, CurvK) results for different sample sizes are depicted in the plot. Each point represents an optimized value of $\alpha = \beta$, ranging up to 150 with irregular breaks. The line represents a learned exponential model predicting the observed trend. The gray area represents the 89% highest posterior density interval (HPDI), while the continuous line represents the mean prediction.

The increase in the shape parameters poses a problem, as larger sample sizes lead to priors that are more concentrated around the center. While MaxSen recommends such priors, their plausibility seems counter intuitive in a scenario of complete uncertainty. It is unclear why a rational agent would consider a hypothesis in the center to be much more probable as the planned sample size increases. If one wishes to avoid this consequence while adhering to MaxSen, the only viable option is the KLD-CurvK combination. However, this combination does not justify regularization, as it recommends the uniform prior. Perhaps another combination of a scoring rule and curvature measure could yield better results. The question remains as to what extent we should trust principles that yield divergent recommendations based on the choice of relatively straightforward measurement methods.

Chapter 5

Conclusions

We have discussed the Maximum Sensitivity principle for choosing priors under radical uncertainty. The method focuses on the primary theoretical role of priors: mitigating reliance on epistemic luck. Konek argues that a rational agent should adopt priors that are maximally sensitive to achieve minimally luck-dependent priors. This sensitivity is measured by first calculating the objective expected posterior inaccuracy OEI of a prior and selecting the prior with the flattest OEI curvature—where distances are measured using the Cramér-von Mises (CvM) and the range of possible values (CurvK) is the curvature measure.

Having explored Konek's view, we tested four variations of the MaxSen method. We introduced the Kullback-Leibler Divergence (KLD) as the distance measure and CurvE as a curvature measure with better theoretical support. The four configurations are: CvM CurvK, CvM CurvE, KLD CurvK, and KLD CurvE. Two key observations result:

1. With the exception of KLD CurvK, all the methods recommend priors in which the sample size n grows with $\alpha = \beta$. The concentration of the prior, influenced directly by the $\alpha = \beta$ parameter, increases with n. We explored this issue further in a separate section.
2. KLD CurvK consistently recommends uniform priors, suggesting that in a certain configuration, the goal of maximizing entropy and maximizing sensitivity are equivalent. Therefore, in some configurations, the goal of maximizing entropy and maximizing sensitivity coincide.

Next section, we examined the problem of the recommended $\alpha = \beta$ prior parameter growing with the sample size n. Our simulations indicate that for the classical MaxSen method (CvM CurvK), the relationship between $\alpha = \beta$ and the sample size follows an exponential growth pattern. The consequences of this finding are:

1. The $\alpha = \beta$ prior parameters dependence on sample size n results in the constraint that n must be known or assumed prior to the experiment. This can be problematic because different experiments employ various stopping principles.
2. A larger choice of $\alpha = \beta$ leads to a more concentrated prior, which seems unreasonable in cases of complete uncertainty, where having a prior with a very high concentration is not desirable. Such cases occur when the sample size n is large, which is not uncommon in real-life data analysis.

In general, it is not obvious which configuration of MaxSen makes the most sense, and decisions in this dimension lead to different recommendations. The fact that MaxSen in one fairly sensible configuration (KLD CurvK) recommends uniform priors suggests that MaxEnt can be interpreted as a sensitivity-maximizing rule to start with. Moreover, from the options we considered, it is the only one that is resilient to the problem of growing $\alpha = \beta$.

Further research is needed to extend the MaxSen method to more complex modeling choices.

Appendix A

Information and maximum entropy

To describe how MaxEnt works, let us first talk about Shannon information and information entropy. Imagine that you are about to toss a coin with the bias of 0.8 towards Tails once. Clearly, observing Tails surprises you less than the opposite outcome. In the sense in which we will use the word "information", seeing Tails provides less information than seeing Heads. In general, how much information we gain by observing an outcome depends on how much surprise it produces. As Shannon information reflects the amount of surprise when the outcome is observed, it can be understood as a measure of surprise. Shannon postulated three basic desiderata that uncertainty (that is, in some sense, the reverse of information) ought to satisfy:

- **Continuity**: the uncertainty changes continuously with the probabilities of the outcomes.
- **Increase**: As the number of possible outcomes increases, so does uncertainty.
- **Additivity**: the uncertainty about a combination of outcomes of different observations should be the sum of the separate uncertainties.

Shannon proved that his measure of uncertainty (which we will introduce very soon) is the only definition which fulfills all of these constraints (Shannon et al., 1949), up to a transformation.

The surprise of an outcome value x (amount of surprise when the outcome is observed) is $1/p(x)$. Further, to ensure additivity, Shannon used the logarithmic function.[1]

$$h(x) = \log(1/p(x))$$

By using a simple rule that $\log(\frac{a}{b}) = \log(a) - \log(b)$, and the fact that $\log(1) = 0$ we end up with:

$$h(x) = -\log(p(x)) \tag{Information}$$

For example, let's compare the amount of information (surprise) of the following two possible outcomes of a coin:

$$\begin{aligned} h(x_1) &= -\log(0.8) = 0.22 \\ h(x_2) &= -\log(0.2) = 1.61 \end{aligned}$$

Observing an outcome (e.g. Tails) when its probability is only 0.2 is much more surprising and therefore more informative than the other, much more likely outcome.

As another example, let's calculate the amount of information that results from an observation of ten tosses of our biased coin. In scenario A, we observe 8 Tails and 2 heads. In scenario B, we observe 1 Tail and 9 heads.

$$h(A) = \left[\left(\sum_{i=1}^{8} \log(1/0.8) \right) + \left(\sum_{i=1}^{2} \log(1/0.2) \right) \right] \approx 5 \tag{Scenario A}$$

$$h(B) = \left[\log(1/0.8) + \left(\sum_{i=1}^{9} \log(1/0.2) \right) \right] \approx 14.71 \tag{Scenario B}$$

The frequency in the sequence of tosses in scenario A was much closer to the bias of the coin, so the result was much less surprising and informative. Scenario B is unlikely to occur. In Figure A.1 the plot on the left illustrates the probability of all possible outcomes of ten tosses of our biased coin, and the plot on the right displays Shannon information $-\log(p(x))$ of these outcomes when observed. Clearly, probabilities negatively correlate with the amount of information that a result carries.

[1] Originally, he used base two logarithm and information was measured in bits. In this paper we will use natural logarithm (with base e) instead, because it is widely used in Bayesian inference. The outcome of such a measure of information results in a unit of information called nats.

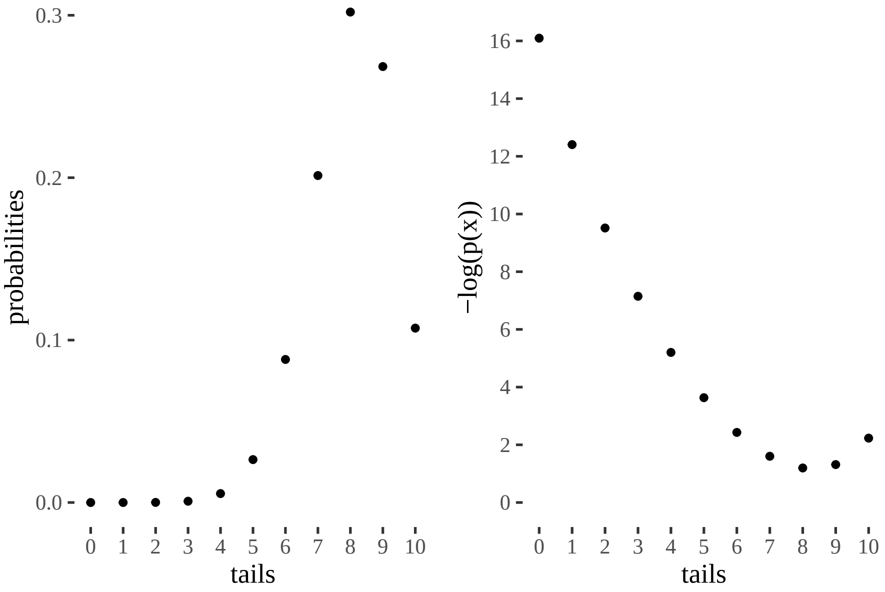

Figure A.1: Probabilities obtained using the binomial distribution (left) and Shannon information (right) of various possible outcomes of ten tosses of a coin with bias .8.

Entropy is a measure of uncertainty associated with a range of hypotheses and a distribution over them, which is the expected Shannon information (Shannon, 1948):

$$H(p) = -\Sigma_i p(x_i) \times \log(p(x_i)) \qquad \text{(Entropy)}$$

This is simply a probability-weighted average Shannon information.

Here is another intuition that might assist one in understanding the notion of entropy. The entropy of a probability distribution corresponds to the number of ways in which it can be obtained. The most uniform distribution is achievable in the largest number of ways. To illustrate, consider at the two plots in Figure A.2. They show two possible outcomes of 10 coin tosses identified in terms of counts, together with the numbers of possible sequences (ways) that would result in these counts.

The distribution that can be achieved in the largest number of ways is the uniform distribution: 5 Tails and 5 Heads. Shannon's entropy follows the count of possible sequences: the larger the number of ways that a distribution can be obtained, the larger the entropy is going to be. The entropies of the two distributions corresponding to the frequencies illustrated above ($p_1(Tails = .8)$, and $p_2(Tails) = .5$) are:

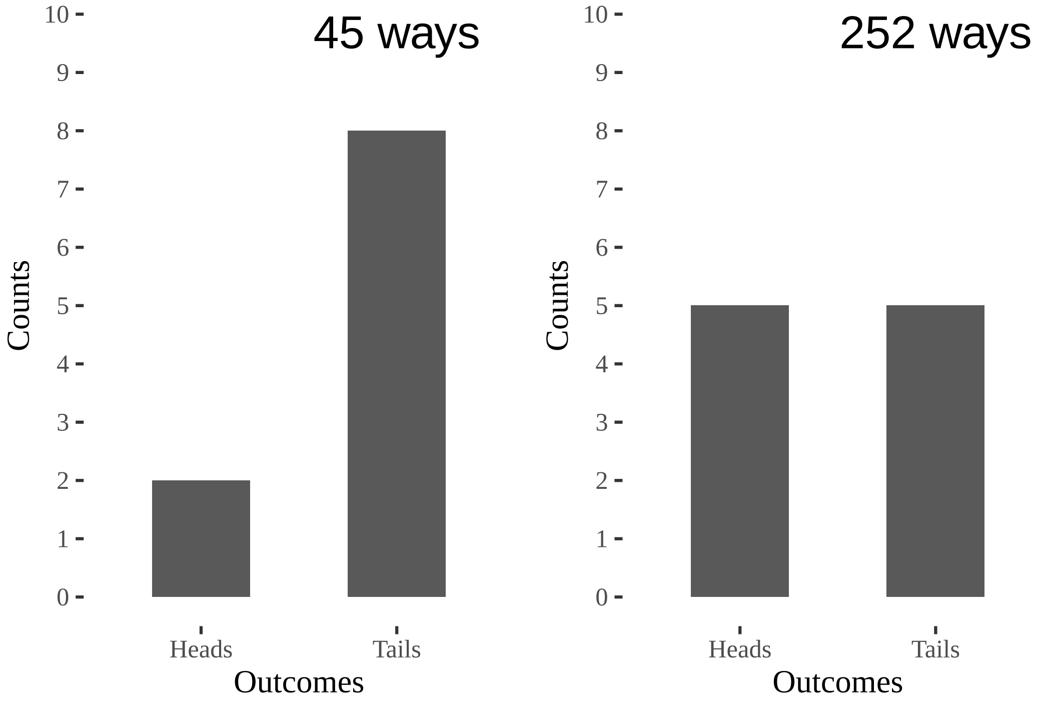

Figure A.2: The number of ways that a distribution can be obtained is calculated using the binomial coefficient.

$$H(p_1) = 0.5$$
$$H(p_2) = 0.693$$

As expected, the distribution with larger entropy is the uniform one (in fact, it maximizes entropy).

List of Figures

Bibliography

Engel Jr., M. (1992). Is epistemic luck compatible with knowledge? *The Southern Journal of Philosophy*, 30(2):59–75.

Gelman, A. (2009). Bayes, jeffreys, prior distributions and the philosophy of statistics. *Statistical Science*, 24(2):176–178.

Gettier, E. L. (1963). Is justified true belief knowledge? *Analysis*, 23(6):121–123.

Hetherington, S. (2019). *The Luck/Knowledge Incompatibility Thesis*, chapter chapter26. Routledge.

Joyce, J. M. (2004). Bayesianism. In *The Oxford Handbook of Rationality*. Oxford University Press.

Konek, J. P. (2013). *New Foundations for Imprecise Bayesianism.* PhD thesis.

Kruschke, J. (2015). *Doing Bayesian Data Analysis (Second Edition).* Academic Press, Boston.

Kullback, S. and Leibler, R. A. (1951). On information and sufficiency. *The annals of mathematical statistics*, 22(1):79–86.

O'Hagan, A., Buck, C. E., Daneshkhah, A., Eiser, J. R., Garthwaite, P. H., Jenkinson, D. J., Oakley, J. E., and Rakow, T. (2006). Uncertain judgements: eliciting experts' probabilities.

Shannon, C. (1948). A mathematical theory of communication. *The Bell System Technical Journal*, 27:379–423.

Shannon, C., E, S., Weaver, W., of Illinois (Urbana-Champaign). Press, U., Blahut, R., and Hajek, B. (1949). *The Mathematical Theory of Communication.* Number t. 1 in Illini books. University of Illinois Press.

DECLARATION

I, the undersigned, declare that the submitted thesis has been prepared by myself, and does not infringe copyrights, interests, material rights of any person, and the use of materials produced by generative tools of artificial intelligence took place to the extent agreed with the thesis supervisor.

29.06.2023

Nikodem Lewandowski