

Resolving the horse/donkey problem of Bayesian inference

Andrew Gelman* and Erik van Zwet†

6 Oct 2026

Abstract

We deconstruct a quip regarding a potentially disturbing aspect of Bayesian shrinkage and use it as an example to demonstrate the importance of tail behavior and predictive checking when combining conflicting information and a concern about statistical procedures that provide inappropriately precise inferences in the presence of conflict between model and data. We argue that this is a concern in statistical modeling and inference more generally.

1. The horse, the donkey, and the mule

The statistician and pharmaceutical researcher Stephen Senn characterized a Bayesian as “One who, vaguely expecting a horse and catching a glimpse of a donkey, strongly concludes he has seen a mule.”

We argue here that this is not just a funny line but represents a real problem in Bayesian inference—and, by extension, a challenge in statistical inference more generally. In this paper we express the joke in mathematical form and present a Bayesian resolution—actually, three different resolutions.

Figure 1 illustrates the scenario in continuous form, with horse-like objects having values of θ near zero, donkey-like objects near $\theta = 5$, and mule-like objects halfway between. We express “vaguely expecting a horse” as a $\theta \sim \text{normal}(0, 1)$ prior, “catching a glimpse of a donkey” as $y = 5$ in a $y \sim \text{normal}(\theta, 1)$ likelihood, and then the posterior is $\theta \sim \text{normal}(2.5, 1/\sqrt{2})$.

There are two problems with the resulting inference:

1. The prior and likelihood are conflict (as can be seen by the very small overlap of the densities in Figure 1, or mathematically by noting that $y = 5$ is more than three standard errors away from the $\text{normal}(0, \sqrt{2})$ prior predictive distribution under the model), but the Bayesian inference proceeds blithely ignoring this.
2. We might hope that the poor fit would at least be reflected in greater uncertainty, but under the normal model the posterior is actually more precise than the prior or the likelihood. This is the issue implicitly raised by Senn’s phrase “strongly concludes.”

We have drawn the graph in Figure 1 carefully, showing an overlap between prior and posterior that is small (to respect Senn’s implication that these two pieces of information are in conflict) but not completely empty (to respect that the prior is only a “vague” expectation and that the visual information is a mere “glimpse”).

If we were to increase the overlap (for example, setting the prior and data standard deviation at 5 in the graph), then the problem is too easy, because then the posterior is broad enough that it could be given the reasonable interpretation as “horse or donkey.” The strong conclusion of “mule” arises from the prior/likelihood separation. But this is a legitimate problem! In many real settings, prior and likelihood do conflict.

*Department of Statistics and Department of Political Science, Columbia University, New York.

†Leiden University Medical Center.

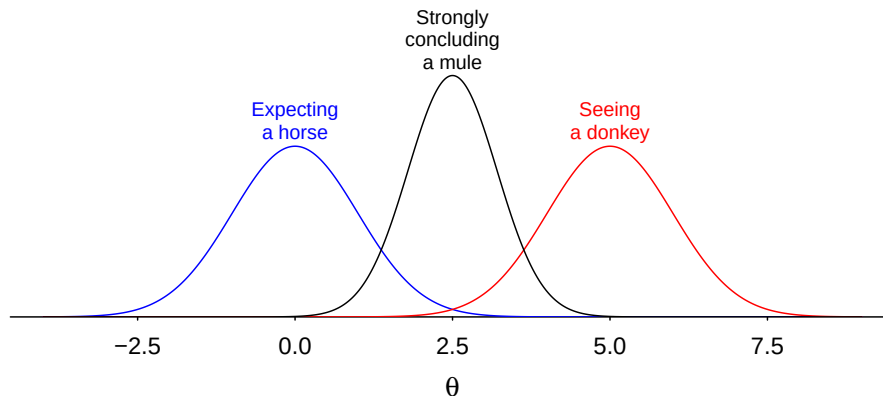


Figure 1: *Formalization of the problem. The prior (blue curve) is centered around $\theta = 0$ or “horse”; the likelihood (red curve) is centered around $\theta = 5$ or “donkey”; although the prior and data are in conflict, the Bayesian posterior confidently settles around the “mule” of $\theta = 2.5$. This is a problem with Bayesian inference, first because it does not recognize the prior-data conflict and second because it is so confident, with the posterior being narrower than the prior or the likelihood.*

The statistical literature is full of misinformed anti-Bayesian arguments (see Gelman and Robert, 2013, for a review of some of these), but there are also some legitimate problems with Bayesian inference. Underlying Senn’s quip is a serious concern, one not discussed in our earlier review on holes in Bayesian statistics (Gelman and Yao, 2021).

Ultimately, we think Senn’s horse/donkey/mule problem can be resolved Bayesianly—to paraphrase the twentieth-century politician Al Smith, we believe that all the ills of Bayesian inference can be cured by more Bayesian inference—but, as with democracy, the best way to do so depends on goals and context. That is the subject of the present paper.

2. Potential solutions

2.1. Long-tailed prior

Instead of using the normal distribution, the Bayesian can express his “vague expectation” with a prior with flatter-than-normal tails. Figure 2 shows the result with a Cauchy, or t distribution with 1 degree of freedom. The resulting inference is now a weak conclusion of “donkey,” which represents an improvement on the original normal model, first because the prior and data no longer strongly conflict, second because the result is not an awkward “mule” compromise, and third because the posterior distribution is not so narrow.

2.2. Long-tailed likelihood

Alternatively, the Bayesian could express the potentially-ambiguous “glimpse” as a more ambiguous likelihood function. Figure 3 shows the result with a normal prior and Cauchy likelihood. Now the posterior is close to “horse,” which makes sense in that if the glimpse is uncertain, any contradiction will be resolved in favor of the prior—but, again, without yielding a narrow posterior or Senn’s “strong conclusion.”

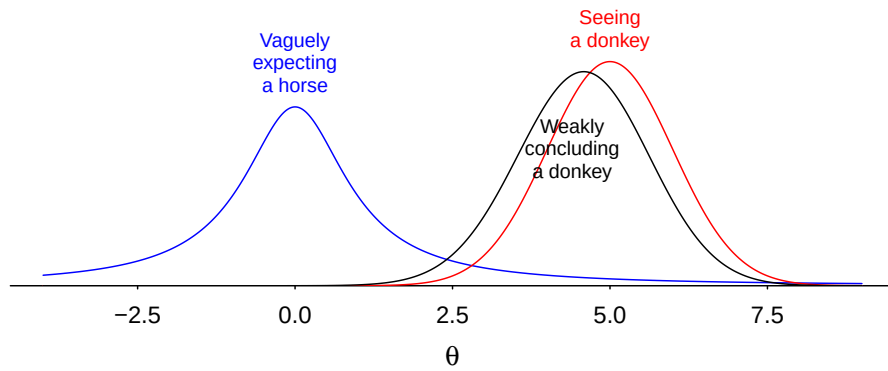


Figure 2: *Cauchy prior and normal likelihood combine to yield a posterior inference that is much closer to the data, or “donkey,” avoiding conflict and the uncomfortable “mule” conclusion.*

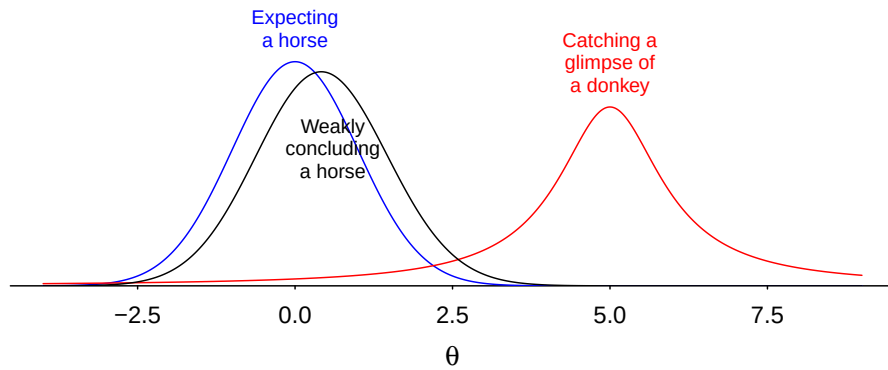


Figure 3: *Normal prior and Cauchy likelihood combine to yield a posterior inference that is much closer to the prior, or “horse,” avoiding conflict and the uncomfortable “mule” conclusion.*

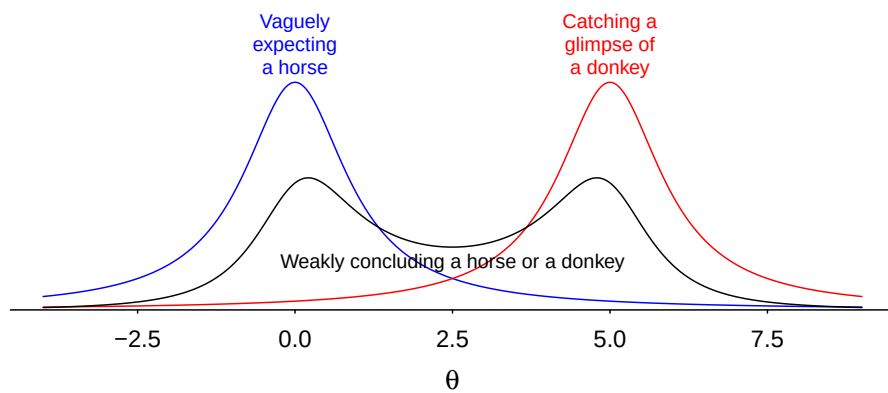


Figure 4: *Cauchy prior and Cauchy likelihood combine to yield a bimodal “horse or donkey” posterior inference.*

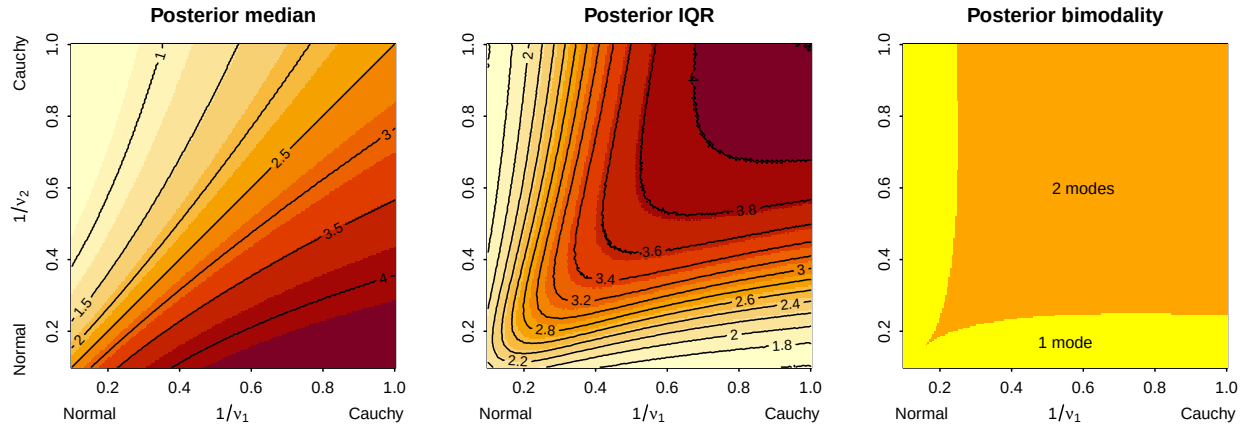


Figure 5: *Posterior median, interquartile range, and bimodality from combining $\theta \sim t_{\nu_1}(0, 1)$ prior and data $y = 5$ with $y|\theta \sim t_{\nu_2}(\theta, 1)$ likelihood, as a function of $1/\nu_1$ and $1/\nu_2$. The lower-left, lower-right, upper-left, and upper-right corners correspond to Figures 1, 2, 3, and 4, respectively.*

2.3. Long-tailed prior and likelihood

What happens if both prior and likelihood are Cauchy distributed? By symmetry, the posterior will then be centered at the “mule” value of 2.5, but, as shown in Figure 4, the distribution is bimodal, returning the implication that the unknown object is horse-like or donkey-like, but we’re not sure which.

2.4. Intermediate models

We can continuously map between the above scenarios by applying a $t_{\nu_1}(0, 1)$ prior distribution and a $t_{\nu_2}(\theta, 1)$ prior, so that Figures 1–4 correspond to $(\nu_1, \nu_2) = (\infty, \infty)$, $(1, \infty)$, $(\infty, 1)$, and $(1, 1)$, respectively. We work out the inferences on a grid of (ν_1, ν_2) and display the results in Figure 5 showing the posterior median, interquartile range, and unimodality or bimodality as a function of μ_1 and μ_2 . We summarize using the median and interquartile range rather than the mean and standard deviation because of the long tails of the t distribution when degrees of freedom are low. Beyond the details here, we think these graphs represent a useful way of understanding inferences from a family of potential models.

2.5. Continuous mixtures

Above we show how switching from normal to t distributions addresses the lack of overlap between prior and likelihood. There are other ways of achieving this goal. Most simply we could just increase the width of the normal distributions, but this would be cheating if the prior scale were decided after the data were seen, or if the variation in the data model depended on the prior. We can, however, salvage this idea by considering the scale parameters of the model to be unknown, replacing the fixed $\text{normal}(0, 1)$ and $\text{normal}(\theta, 1)$ distributions by $\text{normal}(0, \sigma_1)$ and $\text{normal}(\theta, \sigma_2)$, and supplying prior distributions for σ_1 and σ_2 . If these priors are inverse-gamma distributions with scale 1 and degrees of freedom ν_1 and ν_2 , this reduces to the t model discussed in Section 2.4.

This mixture interpretation also gives insight as to how the model with low degrees of freedom behaves so gracefully when the prior and likelihood are far apart: a big difference between the data and the prior mean represents evidence that the prior variance or the data variance is unexpectedly

large, and this propagates back to yield a weaker (less precise) posterior inference. It is also possible to construct a long-tailed continuous distribution using a finite mixture, for example, $p(\theta) = 0.95 \text{normal}(\theta|0, 1) + 0.05 \text{normal}(\theta|0, 10)$. In the present paper we focus on the t family for simplicity, but in applied settings it can make sense to use a mixture prior fit to historical data (van Zwet and Gelman, 2022).

2.6. Discrete mixtures

So far we have considered “horse” to be a null hypothesis and “donkey” to be a somewhat arbitrary positive value. Another approach is to model the discrete classification problem directly, with three kinds of animal and some quantitative trait θ that is (stochastically) ordered along the axis from horse to mule to donkey. The data model is $y \sim \text{normal}(0, 1)$, the prior expectation of “horse” can be expressed as $\theta|\text{horse} \sim \text{normal}(0, 1)$, and the observation of “donkey” corresponds to $y = 5$. The resulting inference is $p(\theta | \text{horse}, y = 5) = \text{normal}(\theta | 2.5, 1/\sqrt{2})$. The problem is that this posterior distribution is more compatible with a mule than a horse.

There are three resolutions of the prior-data conflict:

1. The prior, $p(\theta|\text{horse})$, was too narrow, so we replace it with a wide-tailed t distribution, implying that even a strongly donkey-like parameter value is compatible with an underlying horse.
2. The measurement model was too narrow, so we replace it with wide-tailed t distribution, implying that the observed $y = 5$ is compatible with a more typically horse-like θ near 0.
3. The model was incomplete. We failed to specify the data models $p(\theta|\text{mule})$ and $p(\theta|\text{donkey})$, and the prior probabilities $\text{Pr}(\text{horse})$, $\text{Pr}(\text{mule})$, and $\text{Pr}(\text{donkey})$.

If the goal is to estimate θ and the labels for the three animals are merely symbolic, we prefer the continuous mixture model which directly addresses the problem. If the goal is to (probabilistically) classify the animal, then it makes sense to directly model the probability of each animal.

2.7. Posterior predictive check

Another way into the horse/donkey problem is to include model checking in our workflow. Under the normal-normal model, the posterior predictive distribution for a new measurement is $y^{\text{rep}}|y \sim \text{normal}(2.5, \sqrt{1.5})$, and the observed measurement of 5 is more than two standard deviations in the tail of this distribution, which is a cause for concern. In this case, the model-checking Bayesian who vaguely expected a horse and then caught a glimpse of a donkey could suspect some problem with his expectations, or with his visual identification, or both. The inference of a mule remains, but it is subject to revision with future modeling. In contrast, with the Cauchy prior, likelihood, or both, the data are consistent with the posterior predictive distribution.

It is also possible to do a prior predictive check, but that is less relevant here, as our concern is with the reasonableness of the posterior. For example, consider the case of a $t_4(0, 1)$ prior with a $\text{normal}(y, 1)$ likelihood, and data $y = 5$. The overlap between prior and posterior is small and the data is in the tail of the prior predictive distribution with a two-sided p -value of 0.01: the prior predictive check shows a problem. But the data fit fine within the posterior predictive distribution, which in this case has a mean of 4.0 and standard deviation of 1.5. The model fits the data fine in this sense.

2.8. How to choose a single model?

As the different solutions in Figures 1–5 demonstrate, the horse/donkey problem is underspecified, with the answer depending strongly on the degrees of freedom in the prior and likelihood. The inference also depends on the center and scale of the prior and data, but that is expected. The shape of a distribution is much harder to identify, suggesting a fundamental indeterminacy in the conclusion.

This can be taken as a drawback of Bayesian inference, but non-Bayesian approaches have their own problems. Taking the likelihood by itself yields the unambiguous conclusion of “donkey,” but this may not be appropriate if the observation is a mere “glimpse.” Ignoring the prior is indeed an easy way to avoid prior-data conflict, but as statisticians in the real world we want to face our incoherences, not sweep them under the rug.

But then, what to do? Ultimately the answer must be to set up a model based on one’s understanding of the problem—the population of animals under study when setting up the prior and the measurement process when setting up the likelihood.

One might suggest not choosing any single model but rather considering the distribution of models above and averaging over them. This could work in a hierarchical setting where multiple observations and multiple animals could give us the opportunity to estimate aspects of the prior and data models, but it is not so useful given only one observation on a single animal.

If you look carefully at the labels on the priors and likelihoods in Figure 1–4, you will see that we used the terms “Expecting” and “seeing” for the normal models and “Vaguely expecting” and “Catching a glimpse” to emphasize the greater uncertainties in the Cauchy distribution. Arguably, Senn’s phrasing suggests something like the model in Figure 4 in which both the prior and likelihood have long tails.

3. Discussion

3.1. Discrete and continuous models

It might seem perverse to use a continuous parameterization to model the essentially discrete problem of distinguishing between different animal species, but we think this is in the spirit of Senn’s argument in which Bayesian inference yields an intermediate value rather than discrete probabilities of horse or donkey. Beyond this particular example, we generally use continuous models to when studying opinion changes, treatment effects, drug dosing parameters, environmental risks, and other applications, and the balancing of prior and likelihood is very much a live question in these problems.

As discussed in Section 2.6, there is a mismatch between the Bayesian model that’s implicitly assumed in the quip (normal data and normal likelihood) and the task of identifying which animal has been seen. We choose to resolve this by considering the horse/donkey problem not as a discrete classification problem but as a continuous inference problem in which the prior and likelihood contradict. Alternatively, we could think of this a discrete classification problem where the underlying parameter is discrete (horse or donkey) and the data are continuous (some measure of horse or donkey resemblance). But in that case you’d want a prior on the discrete parameter. And then the problem becomes trivial with no paradox: you have prior probability p_1 of horse, p_2 of mule, and p_3 of donkey, and then the data give you three likelihoods, you multiply them together, and you renormalize, and you’re done. For the paradox to be interesting, the underlying parameter needs to be continuous, which is why we are giving it this treatment.

3.2. Implications for statistical practice

We recommend that a Bayesian who is vaguely expecting a horse and catching a glimpse of a donkey should draw a tentative inference along the horse/mule/donkey spectrum that depends on assumptions about the population under study and the measurement process. The Bayesian should make these assumptions transparent and should be aware of the importance of tail behavior in settings of prior/likelihood conflict (a point also discussed in Section 5.10 of Gelman et al., 2026). Stephen Senn’s unexpectedly deep quip is helpful in dramatizing this sensitivity.

Thinking about Bayesian analysis more generally, three lessons come to mind. First, posterior computation will not in itself flag discrepancies between model and data; for that, predictive checking can be useful. Second, posterior inferences can be sensitive to the tail properties of the prior and likelihood, and so we should think hard about what knowledge and assumptions are going into them. Third, continuous mixture modeling can automatically resolve prior/likelihood conflict; this is related to well-known ideas about Bayesian robustness (Box and Tiao, 1962), but here we are focusing in particular on the way in which, under the t or other scale-mixture models, posterior inference becomes less certain when the data are far from the prior, so that Senn’s “strong conclusion” is avoided. This does not imply that the normal-model should never be used, just that it does not have this property of yielding more equivocal inferences under model misfit, and this puts more of a burden on checking the fit of the different components of this model. Finally, seeing something unexpected is an opportunity to learn about the world (such as the presence of mules) and to expand your model.

References

- George E. P. Box and George C. Tiao (1962). A further look at robustness via Bayes’s theorem. *Biometrika* **49**, 419–432.
- Andrew Gelman, Aki Vehtari, Richard McElreath, et al. (2026). *Bayesian Workflow*. CRC Press.
- Andrew Gelman and Christian Robert (2013). “Not only defended but also applied”: The perceived absurdity of Bayesian inference (with discussion). *American Statistician* **67**, 1–17.
- Andrew Gelman and Yuling Yao (2021). Holes in Bayesian statistics. *Journal of Physics G: Nuclear and Particle Physics* **48**, 014002.
- Stephen Senn (1993). The devil’s drug development dictionaries. <https://www.senns.uk/wdict.html>
- Erik van Zwet and Andrew Gelman (2022). A proposal for informative default priors scaled by the standard error of estimates. *American Statistician* **76**, 1–9.