Normative Uncertainty, Normalization, and the Normal Distribution

Philip Trammell*

November 25, 2022

I banged this out in about two days, and would definitely call it a "work in progress" except that I don't expect to keep working on it. If you want to try polishing and publishing it, or know anyone who might, let me know!

1 Introduction

Faced with a set of possible acts among which to choose, one may have normative uncertainty: uncertainty about how [morally] "choiceworthy" each act is. Under this uncertainty, one may wish to choose an act that maximizes expected choiceworthiness.

In particular, it has been argued that if your behavior under normative uncertainty does not amount to something that can be modeled as "maximizing expected choiceworthiness" across moral theories, then your behavior will have various undesirable features (MacAskill et al., 2020, ch. 2). These arguments, as far as I'm aware, are generally analogous to standard arguments—e.g. the classic von Neumann–Morgenstern (1944) arguments—for maximizing expected utility more generally (with respect to some utility function).

To maximize expected choiceworthiness, however, it is not enough to (a) assign probabilities to all possible moral theories and (b) know all there is to know about the internal structure of how each theory evaluates possible acts. One must also have a way of putting the choiceworthiness claims made by one theory on the same scale as those made by another. To use the jargon, one must be able to make "intertheoretic comparisons".

This seems hard. Suppose a_1 is the act of telling a harmless lie to create some amount of welfare and a_0 is the act of staying silent. We can straightforwardly understand that utilitarianism considers a_1 more choiceworthy than a_0. We can likewise straightforwardly understand that a view that intrinsically values honesty

*Global Priorities Institute and Department of Economics, University of Oxford.
Contact: philip.trammell@economics.ox.ac.uk.

might consider a_0 more choiceworthy than a_1. That is, we can straightforwardly understand what I have called theories' "internal structures". But how are we supposed to compare the size of the gap between a_0's choiceworthiness and a_1's choiceworthiness according to utilitarianism with the size of this gap according to the other theory?

One proposed approach is "variance normalization". This constitutes scaling the theories' sets of choiceworthiness claims, regarding the sets of acts under consideration, so that each theory's set of scaled claims has the same variance. This approach has some desirable features that some other approaches lack (Cotton-Barratt et al., 2020; MacAskill et al., 2020, ch. 4).

But variance normalization is not in general compatible with a Bayesian approach to maximizing expected choiceworthiness. One may have a prior over the distributions of choiceworthiness claims one will encounter in a given situation such that, upon Bayes-updating on the "internal structure" of each theory's choiceworthiness claims about the acts one *does* encounter, one's posterior distributions do not have the same variance. This is illustrated in §4.

As a result of this departure from Bayesianism, variance normalization has its own undesirable features. Indeed, as Savage (1954) in essence shows—though we will not expand on this here—departures from Bayesianism have undesirable features largely analogous to the undesirable features of departures from expected utility maximization highlighted by VNM. There is therefore an especially strong tension between the case for maximizing expected choiceworthiness under normative uncertainty and the case for variance normalization.

Note that the objections raised here apply to all prior-free "statistical normalization methods", to use the term introduced by Cotton-Barratt et al. (2020), not just variance normalization. We will focus on variance normalization because it is the statistical normalization method that has received the most attention to date.

Finally, note the analogy (also made in the sources cited above) to interpersonal comparisons of utility. You might believe that alternatives offer people underlying, interpersonally comparable "utility" levels (or perhaps it would be better to say "welfare" levels), but that all you ever observe is cardinal information about how people rank alternatives and the relative sizes of the gaps they place between them (e.g. on the basis of how they rank gambles). If so, everything said here about intertheoretic comparisons of choiceworthiness applies just as well to interpersonal comparisons of utility.

In any event, the structure of this note is as follows. §2 introduces a simple framework that we will use to distinguish Bayesian from non-Bayesian approaches to intertheoretic comparisons, at least under some stylized circumstances. (It is hopefully relatively clear how a more general framework could model intertheoretic comparisons under more general circumstances.) §3 shows that, within this framework, variance normalization *is* Bayesian as long one's prior is that theories' choiceworthiness claims about the possible acts at hand are drawn drawn i.i.d. from a normal distribution. §3 shows that, within this framework, variance normalization is not Bayesian in general.

2 Framework

You assign probability p to moral theory T_1 and probability $1 - p$ to moral theory T_2. There are $n > 1$ acts available to you. Your prior is that the choiceworthiness claims made by each theory for each act are independently drawn from a common, non-degenerate distribution D over the real numbers.

You would like to maximize expected choiceworthiness across the moral theories. However, you only learn the theories' choiceworthiness claims up to affine transformation. To illustrate what this means, suppose you face three acts, labeled a_1, a_2, and a_3. Instead of knowing that T_1 assigns these acts choiceworthiness values of (say) $v_{1,1} = 8$, $v_{1,2} = 10$, and $v_{1,3} = 2$ respectively, you only learn that T_1 ranks the acts $a_2 \succ a_1 \succ a_3$ and that the gap T_1 places between a_2 and a_1 is four times the gap T_1 places between a_1 and a_3. Or, put another way, you only learn that there is some A and some $B > 0$ such that $v_{1,1} = A + 8B$, $v_{1,2} = A + 10B$, and $v_{1,3} = A + 2B$.

Your approach to maximizing expected choiceworthiness will be called Bayesian if it amounts to maximizing the expected value of the posterior implied by your prior and a Bayesian update on what you learn.

For clarity, we can imagine that this is what you know has happened:

- A list of n independent draws was taken from D. Let us denote this list $V \triangleq (v_1, ..., v_n)$. The draws, i.e. the elements of V, were not all equal.
- For each draw i, an act "a_i" was found that has objective moral choiceworthiness of v_i.
- Let $\hat{V} \triangleq (\hat{v}_1, ..., \hat{v}_n)$ be a normalization of V that preserves only cardinal information about the original draws and not information about their absolute level or scale. For example, perhaps

$$\hat{v}_i = \frac{v_i - \min(V)}{\max(V) - \min(V)}, \tag{1}$$

  in which case the normalized draws all have a maximum value of 1 and a minimum value of 0.

  With probability p, the normalized list $\hat{V}$ was put on a table labeled Table 1. Otherwise, it was put on Table 2.

- A second list of n independent draws, not all equal, was taken from D. We will call the claim-list represented by these draws the "false moral theory". It was normalized in the same way as above and placed on the other table.

When you come across the tables, what is the expected choiceworthiness, for you, of each of the n acts?

If you are a Bayesian, the expected choiceworthiness of act a_i, in light of what you have learned, is (p times the expected choiceworthiness of a_i conditional on T_1 being true) + ($1-p$ times the expected choiceworthiness of a_i conditional on T_2 being true). So you just need to work out the expected choiceworthiness of a_i conditional on each theory.

Some final notation: $v_{t,i}$ will denote the the claim about a_i made by the theory "on table t", and $\hat{v}_{t,i}$ will denote the normalized claim actually "found on table t". The lists V_t and $\hat{V}_t$ will be defined analogously. So our goal is, for each i and t, to determine $\mathbb{E}[v_{t,i}|\hat{V}_t]$.

3 Variance normalization is Bayesian given a normal prior

If D is normal, the variance of $\mathbb{E}[V_t|\hat{V}_t] \triangleq (\mathbb{E}[v_{t,1}|\hat{V}_t], ..., \mathbb{E}[v_{t,n}|\hat{V}_t])$ is independent of $\hat{V}_t$. Variance normalization is therefore Bayesian. Here is a proof sketch.

Suppose $D = N(\mu, \sigma)$.

First, to learn that V_t is some affine transformation of $\hat{V}_t$ is to learn that V_t lies on the open half-plane in $\mathbb{R}^n$ whose edge passes through the origin and which is spanned by $\mathbb{1}_n$—the n-vector of ones—and positive multiples of $\hat{V}_t$. The edge of this half-plane is the "identity line": line spanned by $\mathbb{1}_n$. All possible observations $\hat{V}_t$ therefore consist of open half-planes whose edges pass through the point at the center of the multivariate normal distribution: $\mu\mathbb{1}_n$.

Now note that the probability density of a particular list V_t of n independent draws from $N(\mu, \sigma)$ depends only on the absolute distance of V_t from this center point. $\mathbb{E}[V_t|\hat{V}_t]$ will therefore be a point on the half-plane characterized by $\hat{V}_t$ that equals $\mu\mathbb{1}_n$ plus a vector orthogonal to $\mathbb{1}_n$ that is the same distance from $\mu\mathbb{1}_n$ regardless of what half-plane we are on. This distance will be the expectation of a half-normal distribution, namely $\sigma\sqrt{2/\pi}$.

Finally, the variance of the coordinates of a point (such as $\mathbb{E}[V_t|\hat{V}_t]$) equals the squared distance of the point from the identity line. Since the nearest point on the identity line to $\mathbb{E}[V_t|\hat{V}_t]$ is $\mu\mathbb{1}_n$, the variance of $\mathbb{E}[V_t|\hat{V}_t]$ equals $2\sigma^2/\pi$ regardless of $\hat{V}_t$.

As this approach shows, variance normalization is Bayesian given any joint prior over V_t that is symmetric about a single point on the identity line. We might have the prior that V_t is drawn from the unit n-sphere, for instance.

If the draws of $v_{t,i}$ are independent draws, across i, from some distribution D, then this symmetry is guaranteed if D is normal. I am not sure whether it is guaranteed *only* if D is normal.

Maxwell's Theorem gets us part of the way to an "only if". It tells us that the joint distribution of n independent random variables is symmetric with respect to *all* rotations about some point $\vec{\mu} \in \mathbb{R}^n$ iff the random variables are all normally distributed with means $\{\mu_i\}$ and a common variance. Here, since we are assuming that the random variables are identically distributed, $\vec{\mu} = \mu \mathbb{1}_n$ for some scalar μ.

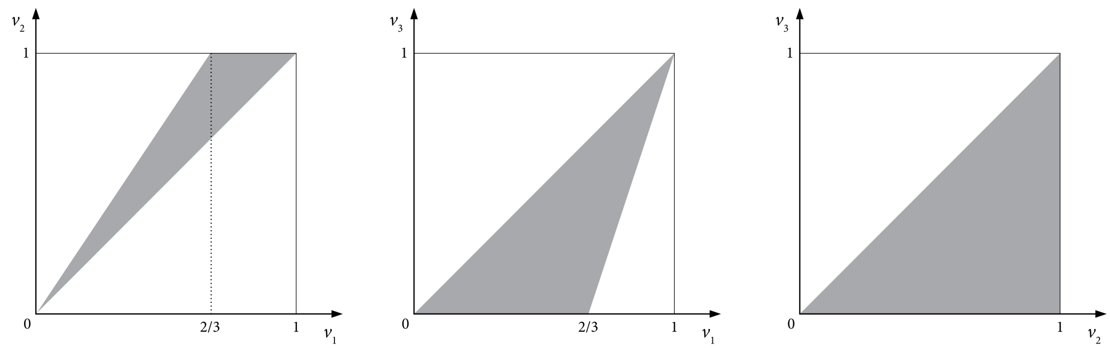
Maxwell's Theorem doesn't get us all the way to an "only if". This is because to justify variance normalization, though, what we need is weaker than full symmetry about $\mu \mathbb{1}_n$, in two ways. First, we only need to consider rotations around the identity line, not *all* rotations around $\mu \mathbb{1}_n$. Second, we only need *the variance of the coordinates of the mean* of the half-plane to be rotation-independent, rather than the whole distribution on the half-plane to be rotation-independent.

Nevertheless, my guess is that normality is necessary for variance normalization to be Bayesian in the above framework. If I show that it is—or, more generally, if I characterize the prior under which the relevant symmetry holds—I'll add that in here. (Either way, I expect this is something already well-known in statistics.) In any event, variance normalization is certainly not Bayesian in general, as we will now show.

4 Variance normalization is not Bayesian in general

Suppose $D = U[0,1]$ and $n = 3$. Suppose both theories rank the three acts such that $v_{t,2} > v_{t,1} > v_{t,3}$, but $v_{1,1} - v_{1,3} = 2(v_{1,2} - v_{1,1})$ whereas $v_{2,1} - v_{2,3} = v_{2,2} - v_{2,1}$.

Let us consider the case geometrically. Our prior over V_1 is a uniform distribution over the unit cube, $[0,1]^3$. Learning $\hat{V}_1$—i.e. learning that $v_{2,1} - v_{2,3} = 2(v_{2,2} - v_{2,1}) > 0$—informs us that we lie in the triangle with vertices $(0,0,0)$, $(\frac{2}{3},1,0)$, $(1,1,1)$. The projections of this triangle onto all three planes that pass through the axes look like this:



From here it is straightforward to calculate that the posterior expectations of $v_{1,1}$, $v_{1,2}$, and $v_{1,3}$ are 5/9, 2/3, and 1/3, respectively.

Our prior over V_2 is also a uniform distribution over the unit cube, $[0,1]^3$. Learning $\hat{V}_2$—that $v_{2,1} - v_{2,3} = v_{2,2} - v_{2,1} > 0$—informs us that we lie on the triangle with vertices $(0,0,0)$, $(\frac{1}{2},1,0)$, $(1,1,1)$. The side projections of this triangle on all three dimensions look like those above, but with 1/2 in place of 2/3 in the first and second projections. The posterior expectations of $v_{2,1}$, $v_{2,2}$, and $v_{2,3}$ are 1/2, 2/3, and 1/3, respectively.

The variance of $(5/9, 2/3, 1/3)$ is around 0.0288. The variance of $(1/2, 2/3, 1/3)$ is different: around 0.0278.

5 References

Cotton-Barratt, Owen, William MacAskill, and Toby Ord, "Statistical Normalization Methods in Interpersonal and Intertheoretic Comparisons," *Journal of Philosophy*, 2020, *117(2)*, 61–95.

MacAskill, William, Toby Ord, and Krister Bykvist, *Moral Uncertainty*, Oxford: Oxford University Press, 2020.

Savage, Leonard Jimmie, *The Foundations of Statistics*, New York: Wiley, 1954.

von Neumann, John and Oskar Morgenstern, *Theory of Games and Economic Behavior*, Princeton: Princeton University Press, 1944.