HHS Public Access

Author manuscript

J Math Psychol. Author manuscript; available in PMC 2020 August 21.

Published in final edited form as:

J Math Psychol. 2019 August ; 91: 103–118. doi:10.1016/j.jmp.2019.04.002.

A General Approach to Prior Transformation

Simon Segert[1], **Clintin P. Davis-Stober**[2]

[1]Princeton University

[2]University of Missouri

Abstract

We present a general method for setting prior distributions in Bayesian models where parameters of interest are re-parameterized via a functional relationship. We generalize the results of Heck and Wagenmakers (2016) by considering the case where the dimension of the auxiliary parameter space does not equal that of the primary parameter space. We present numerical methods for carrying out prior specification for statistical models that do not admit closed-form solutions. Taken together, these results provide researchers a more complete set of tools for setting prior distributions that could be applied to many cognitive and decision making models. We illustrate our approach by reanalyzing data under the Selective Integration model of Tsetsos et al. (2016). We find, via a Bayes factor analysis, that the selective integration model with all four parameters generally outperforms both the three-parameter variant (omitting early cognitive noise) and the $w = 1$ variant (omitting selective gating), as well as an unconstrained competitor model. By contrast, Tsetsos et al. found the three parameter variant to be the best performing in a BIC analysis (in the absence of a competitor). Finally, we also include a pedagogical treatment of the mathematical tools necessary to formulate our results, including a simple "toy" example that illustrates our more general points.

Keywords

Bayesian Priors; Inference; Bayes Factors; Cognitive Models

1. Introduction

We present a general approach for setting prior distributions in Bayesian models when parameters of interest are re-parameterized via some (typically non-linear) functional relationship. Such cases naturally arise in cognitive modeling. For example, a model of decision making may have parameters that correspond to the probability of selecting particular choice alternatives; these parameters may, in turn, be functions of other parameters, e.g., ones relating to neural constructs such as memory and perception. As the

Correspondence should be addressed to Simon Segert. simonsegert@gmail.com (Simon Segert).

sophistication of such cognitive models increases, so too does the need for general quantitative tools to specify such models.

Our approach generalizes the work of Heck and Wagenmakers (2016) and it is instructive to revisit their motivating example. Let $\boldsymbol{\theta}$ denote a vector of real-valued, bounded parameters corresponding to a statistical model. Heck and Wagenmakers (Henceforth HW) considered psychological models that can be operationalized as order constraints of the form $\theta_i \leq \theta_j$. A prominent example is the class of multinomial processing tree models (Batchelder & Riefer, 1999). For these models, the order constraints on the $\boldsymbol{\theta}$ values encode the psychological theory and are of primary interest to the researcher. These models can be re-parameterized by replacing the $\boldsymbol{\theta}$ values via a set of auxiliary parameters, denoted $\boldsymbol{\eta}$, which are functions of the original $\boldsymbol{\theta}$ parameters (Klauer, Singmann, & Kellen, 2015; Moshagen, 2010; Singmann & Kellen, 2013). This re-parameterization yields additional sub-stantive interpretation. HW demonstrated that placing non-informative priors on the $\boldsymbol{\eta}$ parameters can lead to highly informative (and potentially nonsensical) priors on the order-constrained space defined by the original $\boldsymbol{\theta}$ parameters. The resulting Bayes factor for the order-constrained model can be highly sensitive to the priors placed on the auxiliary parameters $\boldsymbol{\eta}$. HW showed how to solve for a prior distribution on the $\boldsymbol{\eta}$ parameters that yields a uniform prior on the order-constrained parameter space defined by the $\boldsymbol{\theta}$ parameters. They termed this process *prior adjustment*. To avoid confusion with prior adjustment in the sense of Mulder (2014), we will utilize the term “prior transformation” instead.

We generalize prior transformation to the case where the constrained parameter space of allowable $\boldsymbol{\theta}$ values need not be full-dimensional within the space of all possible unconstrained $\boldsymbol{\theta}$ values and where the re-parametrization need not be one-to-one. In other words, there may be multiple or even infinitely many auxiliary parameter values which correspond to a single value of $\boldsymbol{\theta}$. In later sections we provide an illustration of exactly this case. We also present computational methods to carry out prior transformation for cases when the desired prior distribution cannot be obtained analytically and/or the model in question does not have a closed-form representation and must be simulated. This allows prior transformation to be carried out on a wide range of statistical models. We present general guidelines for carrying out these computational approximations.

We also show that the basic point made by HW applies to models that aren’t explicitly described as order-constrained models. We generalize prior transformation to the case where the constraints upon the $\boldsymbol{\theta}$ values are *implicitly* defined by a given theory and show that similar issues occur. These constraints may not be easily solvable in terms of $\boldsymbol{\theta}$ and may be highly complex and non-linear in form. As a running example, we consider the *Selective Integration (SI)* model of Tsetsos et al. (2016). SI is a computational process model of multi-attribute choice, where the probability of choosing one alternative over another is a non-linear function of the (fixed) choice attributes and four psychological parameters: (1) early cognitive noise, (2) selective attentional gating, (3) memory leak, and (4) late cognitive noise. Repeated choices are modeled within a binomial random variable framework with the SI model providing the probability of selecting one alternative over another.

To evaluate SI using Bayes factors, we would need to specify prior distributions over these four parameters[1]. If we don't have strong prior beliefs for the parameter values, we may want to proceed by letting the "data speak" and specifying a uniform or otherwise non-informative prior. While perhaps not immediately apparent, the four parameters in SI serve an auxiliary role to the binary choice probabilities, which are of primary interest, as the constraints on the binary choice probabilities define the primary parameter space of the model when considering choice frequencies. These binary choice probabilities can be directly estimated via the actual choices made by the decision maker. The four SI parameters are latent in nature and not directly measured; they are estimated through the likelihood function using binary choice data. As we show in Section 5, SI can be described as an order-constrained model on the binary choice probabilities themselves. These constraints are not easily solved for and are highly non-linear in nature.

In HW's terminology, the four SI parameters are playing the role of the auxiliary parameters η and the binary choice probabilities are playing the role of θ, which, depending upon the goals of the researcher, may require prior transformation. We demonstrate that placing uniform priors over these four SI parameters yields highly informative, and potentially nonsensical, priors over the binary choice probabilities, θ. We then demonstrate how prior transformation, in the more general contexts we consider, allow us to calculate Bayes factors for SI that are based on uniform priors at the binary choice probability level. We stress that there is nothing unique about the SI model for making our more general point. The same basic argument could be applied to well-known models such as Cumulative Prospect Theory (Tversky & Kahneman, 1992). The core idea is that for many cognitive models there are both primary and auxiliary parameter spaces and we must deal with the non-linear relationships between them (which often encodes the psychological theory) when specifying prior distributions.

For illustrative and mathematical tractability purposes, we consider the case of prior transformation for when the desired prior distribution on the primary parameters θ is uniform. However, our core results (Main Formulas 1 and 2) are general in that they could be applied to arbitrary continuous distributions over θ. Similar to the perspective offered by Lee and Vanpaemel (2017), one could specify any *informative* prior over θ and use our results to obtain the requisite priors over η. Thus, our approach provides researchers a general set of tools when carrying out prior transformation for Bayesian models.

Our main results leverage theorems from differential geometry and related fields. In Section 2, we provide a brief tutorial on this topic as it relates to our main results. Also, where appropriate, we reference sections in the Appendix, which contain additional mathematical descriptions and derivations. Readers who wish to skip the tutorial section can go directly to Section 3, which contains the primary technical results of the paper. In Section 4, we formally connect prior transformation to Bayesian statistical evidence and Bayes factors. In Section 5, we evaluate the Selective Integration model of Tsetsos et al. (2016) under prior transformation for multiple versions of the model and compare differences in the Bayes

[1]Tsetsos et al. (2016) did not specify a prior distribution for the four parameters of SI; they applied classical methods only.

factor. Finally, we end with a discussion and directions for future work. All of our code is available via an online supplement.

2. Tutorial on the interface between differential geometry and probability measures

In this section, we present a brief tutorial on the necessary concepts and results from differential geometry needed to apply our main results (Section 3). Throughout our tutorial presentation, we will use a simple toy example to illustrate the mathematics.

2.1. Toy Example

Suppose that we are astrophysicists interested in devising a statistical model of the intensity of solar flares, depending on the location on the surface of the sun. To be more concrete, let us suppose that we have somehow devised a Bayesian statistical model for the intensity of a flare on a given day, which we write as $P(\text{intensity of flare} = x|\text{location} = p)$, where $p \in \mathbb{R}^3$ is a point on the surface of the sun (relative to a coordinate system whose origin is at the center of the sun), and x is the intensity of the flare (measured in units of energy, say, joules). Clearly, for the sake of interpretability of the resulting model, it does not make sense to allow p to be an arbitrary point of $\mathbb{R}^3$. Indeed, it would be nonsensical to obtain a prediction about a "solar flare" on the surface of Mercury by blind mathematical extrapolation. Thus the surface of the sun acts as a *constrained parameter space* inside the ambient space $\mathbb{R}^3$.

For the purposes of this exposition, we assume the sun to be a perfect sphere; moreover, we will take our unit of length to be equal to the radius of the sun, so that the sun has radius equal to 1 in these units. With these conventions, we may describe our constrained parameter space mathematically as $S^2 := \{p \in \mathbb{R}^3 : \| p \| = 1\}$.

Now that we have identified the parameter space, the question arises of how to work with it. Ultimately, we are going to want to compute Bayesian evidence, which comes down to integrating a likelihood over the constrained parameter space. When the geometry of the parameter space is particularly simple, such as in this example, this step is not especially difficult, and would probably not be given much thought. As we see in our re-analysis of the SI model, however, there are situations where the geometry is not so simple and it is not so obvious how to proceed. Our objective here is therefore to examine carefully the logic of this step, so that it may be extended to these more complex situations.

In the context of this specific example, it is quite natural and sensible to switch to a representation of the sphere in terms of longitude and latitude.[2] With the exception of the two poles, every point on the sphere has a unique longitude and latitude, which, in notational contrast to the standard cartographical conventions concerning longitude and latitude, we

[20] https://figshare.com/articles/Segert_Davis-Stober_2018_Supplementary_Material_2/7289642

[2] While longitude and latitude are typically introduced only for the Earth, the logic behind them applies just as well to the sun. Indeed, the sun rotates around an axis, allowing us to define the positions of the North and South poles just as for the earth. This in turn gives a well-defined notion of latitude. The definition of longitude depends on an arbitrary choice of a semi-circular arc connecting the poles, analogous to the Prime Meridian, which we shall suppose has been made.

shall take to lie in the intervals $[0, 2\pi)$ and $(0, \pi)$, respectively. In fact, it will later turn out to be technically easier to deal with the case where the domain is an open set, so we will henceforth take the domain of the longitude to instead be $(0, 2\pi)$, in order to maintain consistency with future sections. This has the effect of excluding the prime meridian, and will not materially affect the following discussion.

We may then re-express our original likelihood $P(x|p)$ in terms of the longitude and latitude of the sun as follows: $P(x|\theta, \phi) :=^{def} P(x|p_{\theta, \phi})$, where $p_{\theta, \phi}$ denotes the unique point with the given longitude and latitude. The advantage to doing this is that the domain of the parameter values (θ, ϕ) is now simply a rectangle $(0, 2\pi) \times (0, \pi)$, with the consequence that any mathematical obstacles stemming from the geometry of the constrained parameter space have been circumvented. At this point, we might simply place a uniform prior over the parameter space $(0, 2\pi) \times (0, \pi)$ and carry out our analysis directly.

Yet, the decision to specify a point by its longitude and latitude is ultimately somewhat arbitrary. Indeed, as one learns in school, there are a multitude of different map projections, each of which may be preferred over others in certain situations. For example, let us consider the Mercator projection. To follow the exposition, it is only necessary to know that this projection specifies each point p on the sphere (again, with the exception of the poles and prime meridian) by a unique pair of numbers $(\theta, y) \in (0, 2\pi) \times (-1, 1)$ which, crucially, "live" in a different space than the longitude and latitude[3] pair $(\theta, \phi) \in (0, 2\pi) \times (0, \pi)$. Just as before, an astrophysicist might be tempted to place a uniform prior over the Mercator parameter space $(0, 2\pi) \times (-1, 1)$ and proceed to compute Bayesian evidence and other various quantities of interest.

Now the potential issue becomes clear. There is no reason why we should expect that a uniform prior on $(0, 2\pi) \times (0, \pi)$ should give us the same Bayesian evidence as a uniform prior on $(0, 2\pi) \times (-1, 1)$. This is precisely the situation described by HW. In the context of this example, HW showed that this can be addressed by solving for the Mercator projection parameters in terms of the longitude and latitude parameters. In other words, by computing the function which carries the pair (θ, ϕ) to the unique pair (θ', y') that corresponds to the same point on the sphere. HW showed that the equivalent prior can be solved for in terms of the Jacobian determinant of this function, essentially as an application of the multivariate change-of-variables formula. They termed the prior so obtained the "implied distribution."

However, the formalism of HW is not sufficient to cover all possible cases of interest. Firstly, it is not applicable to constrained spaces of positive codimension such as our example of the sphere (indeed, this is a 2-dimensional space inside of $\mathbb{R}^3$, and thus has codimension $3 - 2 = 1$). Note that in our example, we could not directly compute the implied distribution on the surface of the sphere; we could only express it indirectly via its implied distributions on the parameter spaces. Nevertheless, it is easy to conceive of a situation in which it might be more natural to work directly with a distribution on the surface of the

[3]For the benefit of the interested reader, we mention that this projection is constructed to ensure that angles on the surface of the sphere are not distorted. This can be accomplished by enclosing the earth in a cylinder around the equator, projecting outwards, and then unrolling the cylinder. However, there is a twist, in that, instead of projecting the line of latitude at angle ϕ to the height $\sin \phi$, we project this line of latitude to the height $y = \ln(\tan(\pi/4 + \phi/2))$.

sphere. For example, it would be quite sensible to want a prior which is proportional to the surface area on the sphere, and then work backwards to express it in terms of a set of parameters. If we want to carry out this procedure in complete generality, then, we will need, at the very least, a way to measure surface area of a k-dimensional constrained space in an n-dimensional ambient parameter space. By contrast, HW formulated their results only for n-dimensional subsets of n-dimensional parameter spaces.

Secondly, their approach does not handle parameter transformations that are non-invertible. To illustrate, let us consider the following situation: due to certain (hypothetical) advances in astronomical theory, we are led to postulate that the solar flare intensity in fact only depends on the latitude of the point p and that our likelihood function is a function of the latitude only. Given a prior on the space $(0, 2\pi) \times (0, \pi)$, we would like to express an equivalent prior (i.e. one that yields the same evidence value when integrated against the likelihood) in terms of the latitude only. The approach of HW is not applicable here, because the map taking a point to its latitude is not one-to-one (that is, many different points are at the same latitude). Let us denote the map taking a pair of angles $(\theta, \phi) \in (0, 2\pi) \times (0, \pi)$ to the corresponding point on the sphere by F, and let us denote the map taking a point on the sphere to its latitude by $H: S^2 \to [0, \pi]$. We can thus recast our problem as describing the implied distribution of $H \circ F: (0, 2\pi) \times (0, \pi) \to [0, \pi]$, where "$\circ$" is function composition. Alternatively, given a prior directly on S^2, we wish to compute the implied distribution under H. How can this be done? It is here that the geometric picture provides invaluable intuition and ultimately points towards a general solution.

We stress that these two issues are not merely contrived features of our toy example. To wit, we shall encounter both of them when we consider the SI model of Tsetsos et al. (2016) in Section 5. More specifically, we will see that (in a certain restricted case), the SI model may be abstractly described as a 2-dimensional surface sitting in a three-dimensional space (under the experimental design they considered), each axis of which corresponds to a probability of a participant selecting a given alternative. Hence, the relevance of the first issue raised above. Furthermore, in their specification of the model, this surface is in fact parameterized by a function of *three* latent parameters. Hence the relevance of the second issue.

Having "pulled back the curtain" to explain the rationale guiding the construction of our toy example, let us now return to it in order to try to understand how the issues that it illustrates may be resolved in this simplified setting. We can schematically imagine the effect of the map H as squashing together all points at the same latitude. We may thus "slice up" the sun into a collection of circles (including degenerate one-point circles at the poles) of the same latitude. It is intuitively plausible that the density of the implied distribution on the "latitude space" $[0, \pi]$ should be a combination of two contributions: 1), an integral/average taken over the corresponding line of latitude, and 2) a "latitude-wise stretch factor" analogous to the Jacobian determinant that appears in the ordinary change of variables formula from multivariate calculus.

The language of differential geometry is perfectly suited to make these geometric intuitions precise, and, furthermore, fluency in this language will allow for application of the same

reasoning to far more general situations than considered in this toy example. We state the main formulas in Section 3 that make these ideas precise and general.

2.2. The definition of a manifold

All of the constrained spaces which we will consider are examples of *smooth manifolds*. Intuitively speaking, a manifold M is a space which "looks like" Euclidean space near each point. More precisely, for each point $p \in M$, we require that there exists a smooth invertible function from a subset $V \subset M$, $p \in V$, to Euclidean space of appropriate dimension, exactly like the map in the previous section $S^2 - \{\text{poles, prime meridian}\} \rightarrow (0, 2\pi) \times (0, \pi)$ which takes a point to its longitude and latitude, and is only defined for points other than the two poles and the prime meridian. Evocatively, such a map is sometimes called a "chart," although we shall not use this terminology. We may thus say that a manifold is a space such that every point is in the domain of a chart (a collection of charts whose domains cover the manifold is sometimes called an "atlas").

Let us now introduce some standard notation: for $p \in \mathbb{R}^n$ and $\varepsilon > 0$, we set $B^n_\epsilon(p) := \{q \in \mathbb{R}^n : \| p - q \| < \epsilon\}$. A subset $U \subset \mathbb{R}^n$ is said to be *open* if for any $p \in U$, we can find $\epsilon_p > 0$ with the property that $B^n_{\epsilon_p}(p) \subset U$ (the subscript is simply a reminder that the value of ϵ may depend on the point p). We can now give a formal definition.

Definition 1. *Let M be a subset of $\mathbb{R}^n$. We say that M is a manifold of dimension k, if, for any $p \in M$, we can find*

1. *an open set $U \subset \mathbb{R}^k$,*
2. *an open set $V \subset \mathbb{R}^n$ containing p, and*
3. *a smooth one-to-one map $C : U \rightarrow \mathbb{R}^n$, such that $C(U) = V \cap M$.*

Such a C is called a local parameterization[4] *near the point p.*

How does one check that a given space M is a manifold? One easy but important special case is when M is of the form $F(U)$, where $U \subset \mathbb{R}^k$ is an open set and $F : \mathbb{R}^k \rightarrow \mathbb{R}^n$ is a smooth, one-to-one function. In this case, we can just take the local parametrization to be F.

Returning to our sun example, let us consider $M = S^2$ with $n = 3$. We would like to show that M is a manifold of dimension 2. We will consider the south pole $p = (0, 0, -1)$, with the verification for other points being similar. Letting x, y and z denote the standard Cartesian coordinates, we note that on a small neighborhood of p, we can express z as a function of x and y via $z = -\sqrt{1 - x^2 - y^2}$. To match up with the definition, we may take $C : B^2_\epsilon(0) \rightarrow \mathbb{R}^3$ defined by $C(x, y) = (x, y, -\sqrt{1 - x^2 - y^2})$, where ϵ is any sufficiently small number (any $\epsilon < 1$

[4]Actually, our usage of the term "manifold" here is somewhat nonstandard. What we call manifolds would be more properly called "embedded submanifolds of Euclidean space." The more common definition of a manifold does not make reference to any ambient space. However, we will not need the additional abstraction afforded by this definition. In addition, the essential concepts are somewhat easier to grasp in this special case, and may be all extended to the more general case in a relatively straightforward way, as detailed in e.g., Boothby, (2003).

will do). In this case $C(B_\epsilon^2(0)) = \{(x, y, z) \in S^2 : x^2 + y^2 < \epsilon, z < 0\}$, which is geometrically equal to the intersection of S^2 with the open cylinder $B_\epsilon^2(0) \times (-\infty, 0)$ This cylinder plays the role of the set V in the definition of a manifold.

One can imagine that this kind of reasoning would be difficult to carry out in more complicated examples. Indeed, to find the local parametrization, we effectively had to solve for z in terms of x and y given the constraint $x^2 + y^2 + z^2 = 1$. We were lucky that the constraint had an algebraically simple form, but this need not always be the case. Fortunately, there is a general theorem that lets us bypass such arguments in cases such as this one where the manifold is defined by a constraint.

Theorem 1. *Let* $f : \mathbb{R}^n \to \mathbb{R}$ *be a smooth function and let* $c \in \mathbb{R}$. *Then* $f^{-1}(c)$ *is a smooth manifold of dimension* $n - 1$ *provided that the vector of partial derivatives* $(\partial_1 f(x), \ldots, \partial_n f(x))$ *is nonzero for every* $x \in f^{-1}(c)$.

This is a special case of the well-known Implicit Function Theorem. See Boothby (2003) for a more general statement and proof.

Let us demonstrate this theorem applied to the sphere (our "sun" example). We may write $S^2 = f^{-1}(0)$ where $f(x, y, z) = x^2 + y^2 + z^2 - 1$. The matrix of partial derivatives at the point $p = (x, y, z)$ is given by $(2x, 2y, 2z)$. If $(x, y, z) \in f^{-1}(0)$, then this vector can never be equal to (0, 0, 0) (this would contradict the assumption that $x^2 + y^2 + z^2 = 1$). This establishes, via the above theorem, that S^2 is indeed a manifold of dimension 2.

It is worth noting that many of the constrained spaces that we are likely to encounter are not quite manifolds in the sense of the formal definition we gave. For example, the unit interval [0, 1] does not satisfy the manifold condition at the endpoints. It is an example of a "manifold with boundary," which may be thought of as a manifold of dimension k with a manfiold of dimension $k-1$ "glued onto it." In this example, we have the bona-fide 1-manifold (0, 1), with the 0-dimensional manifold {0, 1} (i.e., the manifold consisting of two discrete points) glued on. Because we are ultimately interested in performing integrals over manifolds, and because the boundary pieces will always have measure zero, we will not concern ourselves too much with this issue.

2.3. Jacobians on Manifolds

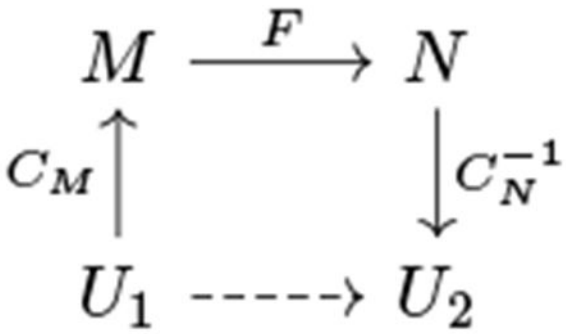
Our generalization of the results of HW, which we describe in detail in Section 3, will require a generalization of the notion of a Jacobian. Similar to the usage of the Jacobian matrix applied in the change of variables formula used by HW, we now consider how to define the Jacobian derivative of a map between manifolds. The basic idea is quite simple. First, recall that the Jacobian matrix at $p \in \mathbb{R}^n$ of a map $F : \mathbb{R}^n \to \mathbb{R}^m$ is simply the matrix of partial derivatives of the components of F, evaluated at the point p. Now consider a smooth map F: $M \to N$ where M and N are manifolds and let $p \in M$.[5] As we saw in the previous subsection, we can choose local parametrizations near $p \in M$ and near $F(p) \in N$, which

[5]When we formulate our results in section 3.1, M and N will be parameter spaces for a given model, and F will be a reparametrization function.

amounts to identifying small neighborhoods around p and $F(p)$ with open subsets of Euclidean space.

Under these identifications, we may regard F as a map between Euclidean spaces, and take its Jacobian derivative as before, thereby obtaining a matrix.

More precisely, let $C_M\colon U_1 \to M$ and $C_N\colon U_2 \to N$ be local parametrizations with $p \in C_M(U_1)$ and $F(C_M(U_1)) \subset C_N(U_2)$. Here U_1 and U_2 are subsets of Euclidean space. Then the composition[6] $C_N^{-1} \circ F \circ C_M$ is a function from U_1 to U_2, that is to say, its domain and range are both subsets of Euclidean spaces, and as such this function has a well-defined Jacobian matrix. This is illustrated by the following diagram:



Here, the dotted line denotes the composition $C_N^{-1} \circ F \circ C_M$. Note that the function C_N^{-1} is only defined on the image $C_N(U_2)$.

Finally, to compute the Jacobian of F at p, we should evaluate the jacobian of $C_N^{-1} \circ F \circ C_M$ at $C_M^{-1}(p) \in U_1$. It is essential to realize that this procedure depends upon which local parametrizations on M and N that we chose. We refer the reader to section 7.4 of the Appendix for an alternative formulation of this idea which is closer to the approach followed in many modern textbooks on differential geometry.

2.4. Measures on Manifolds and Riemannian structures

In this final subsection regarding mathematical preliminaries, we explain how to work with measures defined on manifolds. This is crucial for the body of the paper, since the priors that we will consider in this paper are most naturally regarded as measures on appropriate manifolds.

The construction of these measures is intimately tied to the way in which we decide to measure lengths and angles on the manifold. In the second half of this section, we will flesh out this connection, and see how it leads naturally to the notion of a *Riemannian structure* on a manifold which, essentially, is simply a choice of how to measure lengths and angles on the surface of the manifold.

[6]Here C_N^{-1} denotes the inverse function of C_N, not the preimage

The first step is to understand how one may "inherit" the geometry of the ambient Euclidean space to compute the volume of a (measurable) subset of a manifold, thus defining a measure supported on M which we denote by V_M. We will then explain how to compute this measure relative to a local parametrization.

Note that this notion of volume need not coincide with the Lebesgue measure on the ambient space; indeed, if $M \subset \mathbb{R}^n$ has dimension $k < n$, it can be shown that M has Lebesgue measure zero. On the other hand, if, $L \subset \mathbb{R}^2$ is a line segment, then there is a well-defined notion of length, or "1-dimensional volume" of L, even though it has measure zero, and we will show how to extend this idea to general manifolds.

Not surprisingly, we will proceed by approximating the manifold with shapes of a known volume, and then take an appropriate limit. Given a k-dimensional manifold, we choose a local parametrization C with domain $U \subset \mathbb{R}^k$. We now explain how to compute the volume of $C(U)$.

We divide U into very small cubes with edges parallel to the coordinate axes, of length Δx. Let x be the bottom left-most point of such a square. Then the image of this square will be approximately a k-dimensional parallelotope spanned[7] by the vectors $\{\Delta x(\partial_i C)(x)\}_i$.

By letting the grid get finer, it seems quite plausible that the sum of the volumes of all such paralellotopes should approach a well-defined limit. This can be made precise, although we omit the details here. Let us denote this limiting measure by V_M. To be explicit, $V_M(C(U))$ is defined by approximating $C(U)$ by a fine mesh of parallelotopes, and taking the limiting volume as the mesh sizes approaches zero.

Now, in order to complete the construction, we only need to figure out how to compute the volume of an arbitrary k-dimensional parallelotope.

Let us first introduce a standard notation:

Definition 1. *Let $v_1, \ldots, v_k \subset \mathbb{R}^n$ be an arbitrary collection of vectors. The Gram matrix of the vectors is the $k \times k$ matrix whose (i, j) entry is given by the standard dot product (v_i, v_j). We denote this matrix by $G(\{v_i\})$ or just $G(v)$ when the context is clear. The determinant of the Gram matrix is called the Gram determinant. It is always non-negative.*[8]

We have the following result.

Proposition 1. *Let $P \subset \mathbb{R}^n$ be a k-dimensional parallellotope spanned by the k vectors $v_1 \ldots, v_k$. Then the k-dimensional volume of P is given by $Vol_k(P) = \sqrt{det(G\{v_i\})}$, i.e., the square root of the Gram determinant.*

[7]Recall that a parallelotope is simply the higher-dimensional generalization of a parallelogram.

[8]To see this, note that $G(\{v_i\}) = A^T A$ where A is the matrix whose ith column is equal to v_i. If v is an eigenvector of G with eigenvalue λ, then $\lambda |v|^2 = (Gv, v) = (A^T Av, v) = (Av, Av) = |Av|^2 \geq 0$. Since G is a symmetric matrix, the determinant is equal to the product of eigenvalues, which is a product of non-negative numbers.

This is a direct generalization of the familiar fact that for a linear map $A:\mathbb{R}^n \to \mathbb{R}^n$, the image $A([0, 1]^n)$ of the unit hypercube has volume $|det(A)|$. Indeed, the image is exactly the paralellotope spanned by the columns of A. If we denote these columns by $v_1 \ldots v_n$, then the Gram matrix is simply $G(\{v_i\}) = A^T A$, and because A is a square matrix, we have $det(A^T A) = det(A^T)det(A) = det(A)^2$, by the multiplicativity property of the determinant. The proof of the proposition depends only on basic properties of volume and of the determinant, and is omitted.

Having completed the construction, let us now see how we might work with the measure V_M. Returning to our "sun" example $S^2 \subset \mathbb{R}^3$, we consider the parametrization $C: (0, 2\pi) \times (0, \pi)$ in terms of longitude and latitude. Let $U \subset (0, 2\pi) \times (0, \pi)$ be an open set; we will compute $V_M(C(U))$.

To construct the "mesh," we choose a collection $\{x_i\} = \{(\theta_i, \phi_i)\}$ of evenly-spaced grid points in U. Let us denote the spacing by Δx. In particular, the area of a square formed by 4 adjacent grid points is $(\Delta x)^2$.

Next, we cover the portion $C(U) \subset S^2$ with 2-dimensional parallelograms with origin $C(x_i)$ and spanned by $\partial_j C(x_i)\Delta x$, $j = 1, 2$. Note that, e.g. $C(x_i + \Delta x(1, 0)) \approx C(x_i) + \partial_1 C(x_i)\Delta x$, so that the parallelograms do indeed approximately touch at the edges.

Using simple trigonometry, one verifies that $C(\theta, \phi) = (\sin\phi\cos\theta, \sin\phi\sin\theta, \cos\phi)$, and by a short computation, we see $\partial_\theta C = (-\sin\phi\sin\theta, \sin\phi\cos\theta, 0)$ and $\partial_\phi C = (\cos\phi\cos\theta, \cos\phi\sin\theta, -\sin\phi)$. By the proposition, each parallelogram has area given by the square root $\sqrt{det(G(\{\partial_\theta C(x_i), \partial_\phi(C(x_i))\}))}$ of the gram determinant. A bit of tedious computation reveals this to be equal to $\sin(\phi_i)$. Our approximate volume of $C(U)$ is thus $\sum_i \sin(\phi_i)(\Delta x)^2$, and passing to the limit in the usual fashion, this becomes an integral of the function $\sin(\phi)$ over U.

Using similar heuristic arguments, one can convince oneself that for any manifold M with local parametrization C, the value of $V_M(C(U))$ is given by $\int_U \sqrt{det(G(\{\partial_j(C(x))\}_j))}dx$ i.e., the integral of the square root of the Gram determinant of the partial derivatives of C.[9]

We now turn to the notion of a Riemannian structure on a manifold, which was already implicit in our construction of the measure V_M. To begin, let us note that the construction of V_M could be generalized substantially at the cost of introducing a small amount of conceptual overhead.

Indeed, returning to the proposition, we see that, in defining the measure V_M, we only needed to measure the volume of parallelotopes, and via the proposition, we saw that these volumes could be expressed only in terms of inner products of vectors. Let us imagine repeating the construction, except replacing (v_i, v_j) everywhere with a different inner product on $\mathbb{R}^n$.

[9] In contrast to the Δx notation in the preceding paragraph, "dx" here denotes an integral against the (possibly, multi-dimensional) Lebesgue measure on U.

Any such inner product is necessarily of the form (v_i, Av_j) where A is a symmetric, positive definite matrix. In fact, we can even allow A to depend on the origin of the vectors v_i, v_j. Thus, given two vectors v and w with the same origin p, we may define a "generalized inner product" $(v, A(p)w)$, and repeat the construction of the measure V, replacing the old inner product everywhere with this new inner product. Such a matrix-valued function A is called a Riemannian metric on $\mathbb{R}^n$, and denoted g for historical reasons.[10] Let us denote the measure obtained by measuring the volumes of parallelotopes in this way by V_M^g. This measure is called the *Induced Riemannian volume form on M*. By retracing the above arguments, it may be easily seen that $V_M^g(C(U)) = \int_U \sqrt{det(G^g(\{\partial_j(C(x))\}_j))}dx$., where the (i, j) entry of the matrix $G^g(\{\partial_j(C(x)\}_j))$ is equal to $g_{C(x)}(\partial_i(C(x), \partial_j(C(x))$.

The practical-minded reader may wonder what is to be gained by this abstraction. The justification is that non-Euclidean metrics are in fact encoutered "in the wild," even if it is not immediately apparent. Perhaps the prototypical example of this is the Jeffreys prior, which is the measure induced by the so-called Fisher-Rao metric on the ambient space (Amari, 1985). Alternatively, this formalism may be used when the axes of the space are naturally regarded as having different units, in which case the matrix A can be taken to be diagonal, with the entries being conversion factors to a common scale.

Finally, we mention that it is possible to define a Riemannian metric directly on a manifold, without reference to such a metric on the ambient space. To do so, we simply need, for each point $p \in M$, an inner product $g_p(v, w)$, where v and w are any two vectors whose origins are at p and which are tangent to M. (More precisely, v and w should lie in the *tangent space* T_pM, which is formally defined in Section 7.4 of the Appendix). In addition, g_p is required to depend smoothly on p, in a certain precise sense. A manifold with such structure is called a *Riemannian manifold*.

3. General Prior transformation

3.1. Setup and general formulas

Let us consider a Bayesian statistical model that is expressed in terms of a parameter vector $\theta \in \mathbb{R}^n$ which is itself defined in terms of another parameter vector $\eta \in \mathbb{R}^m$ by a relation of the form $\theta = F(\eta)$ for some smooth function F.[11]

As an example, and a concrete way to connect to HW, we could consider θ as a vector of choice probabilities applied to a set of n binary decision tasks, while considering η as a vector of theoretical latent parameters that govern the choice probabilities (e.g., "loss aversion"). Abstractly speaking, F may also be regarded as a (possibly many-to-one) reparametrization function, and conversely, any reparametrization function also may be regarded as a model in this sense.

[10] More precisely, the correspondence with our notation is given by $g_p(v, w) := (v, A(p)w)$.

[11] Note that the letter F was used previously in section 2.3 to denote a smooth map between manifolds. This notational reprisal is deliberate, as it will shortly become apparent how the present usage of F is equivalent to the previous usage.

We suppose that $\boldsymbol{\eta}$ is constrained to lie in a compact subset $D \subset \mathbb{R}^m$. The most convenient case, and the one which is relevant to our examples, is when D is a rectangle $\prod_{i=1}^{m} [a_i, b_i]$. We refer to the set D as the *auxiliary parameter space*, and the image $Im(F) \subset \mathbb{R}^n$ as the *primary parameter space* of the model.[12]

Observe that any prior distribution $\boldsymbol{\pi}$ on the auxiliary parameter space induces a distribution $F_*\pi$ on the primary parameter space defined by the formula $(F_*\pi)(U) = \boldsymbol{\pi}(F^{-1}U)$, for each measurable $U \subset Im(F)$.[13]

This distribution is called the *pushforward of* $\boldsymbol{\pi}$ *by* F, or the *implied distribution on* $\boldsymbol{\theta}$, to use HW's terminology. An equivalent definition which we find more convenient is that the pushforward measure $F_*\pi$ is the unique measure for which the equation

$$\int f(\boldsymbol{\theta})(F_*\pi)(d\boldsymbol{\theta}) = \int f(F(\boldsymbol{\eta}))\pi(d\boldsymbol{\eta}),$$

holds for any bounded measurable $f: Im(F) \to \mathbb{R}$.[14] See Appendix section 7.3 for a brief discussion and additional examples of pushforward measures.

To perform Bayesian analysis, it is necessary to specify a prior distribution $\boldsymbol{\pi}$. We consider two possible ways to do this:

1. Specify a prior on the auxiliary parameter space, and push it forward by F to obtain the corresponding distribution on the primary parameter space.
2. Specify a prior on the primary parameter space, and try to "reverse engineer" an appropriate prior on the auxiliary parameter space which pushes forward to this distribution.

The discrepancy between these two approaches was pointed out in the special case of multinomial processing tree models by HW. Using our notation, they considered the case of $\boldsymbol{\eta} \in [0,1]^n, F(\boldsymbol{\eta})_i = \boldsymbol{\Pi}_{j \geq i}\eta_j, 1 \leq i \leq n$. They showed how to construct a prior on $\boldsymbol{\eta}$ whose pushforward is equal to the uniform istribution on $Im(F) = \{(\theta_1, \ldots, \theta_n) \in [0,1]^n : \theta_1 \leq \ldots \leq \theta_n\}$, thus carrying out Option (2), for the uniform distribution with support equal to $Im(F)$. Observe that in this (HW's) setup, $n = m$.

As we now show, their argument can be adapted to a wide class of smooth functions F. Our approach differs in several respects. First, whereas they started with a primary parameter space of a known geometry (implied by a multinomial processing tree model), our starting

[12]Technical remark: Because D is not an open set, some care is required in saying what it means for the function F to be "smooth." To be precise, we will require that F is continuous on D, and is smooth when restricted to the open cube $\pi_{i=1}^{m}(a_i, b_i)$.

[13]Here, and in the rest of the paper, we use the notation $F^{-1}U$ to denote the preimage, defined by $\{x \in Domain(F) : F(x) \in U\}$. In case U consists of a single point p, we abuse notation slightly and denote the preimage by $F^{-1}(p)$. This should not be confused with the inverse function, which, in general, will not exist.

[14]A word on notation: We will make use of both equivalent notations $\int h\mu$ and $\int h(x)\mu(dx)$ for the integral of a function h against a measure μ, the latter in cases in which one could possibly lose track of which variable corresponds to which measure, and the former when there is little possibility for such confusion.

point is the parametrization function F, and the resulting primary parameter space $Im(F)$ may be very difficult or impossible to express as a system of equalities and inequalities. Second, we do not need to assume that the image of F is full-dimensional in the codomain, nor do we need to assume F is one-to-one.

We must define what is meant by a uniform prior on the primary parameter space. Let us fix a Riemannian metric on $\mathbb{R}^n$. We use the fact that the primary parameter space is an immersed submanifold in $\mathbb{R}^n$, and therefore inherits a Riemannian metric through restriction of this metric as described in Section 2.4 (see also Chavel, 2006). In general, for any such submanifold M, we will denote the corresponding measure[15] on M by V_M. For the primary parameter space, we will use the notation $V_{prim} = V_{Im(F)}$.

Our aim is to construct a distribution μ on D for which $F_*\mu = V_{prim}$. In other words, μ will be a prior distribution on the auxiliary parameter space which induces the uniform distribution on the primary parameter space. Fix a Riemann metric on D, and let $V_{aux} = V_D$ denote the corresponding measure.

We now recall the well-known *co-area formula* of Riemannian Geometry (see Chavel, 2006, for a precise statement and proof), which generalizes both the basic change of variables formula (used by HW for prior transformation when $n = m$) and Fubini's Theorem.

Coarea Formula 1. *Let M and N be Riemannian manifolds and $F: M \to N$ a surjective submersion.*[16] *Then for any measurable function $f: M \to \mathbb{R}$, one has*

$$\int_N \left(\int_{F^{-1}(p)} f V_{F^{-1}(p)} \right) V_N(dp) = \int_M J_F(q) f(q) V_M(dq)$$

where J_F is the generalized Jacobian determinant[17] *defined as $J_F = \sqrt{det(Jac(F) Jac(F)^T)}$, det is the usual determinant of a linear operator, and $V_{F^{-1}(p)}$ is the measure induced on the preimage $F^{-1}(p)$ by the Riemannian metric on M. The formula also holds when $dim(F^{-1}(p)) = 0$, provided one interprets the integrals over these sets as discrete sums.*

For readers familiar with the usual calculus change of variables formula, the coarea formula is quite similar with the generalized Jacobian determinant, $J_F(q)$, replacing the usual Jacobian determinant. See Section 7.1 of the appendix for more details on the generalized Jacobian determinant.

This formula allows us to compute the pushforward (i.e., implied distribution) of any continuous distribution on the auxiliary parameter space, even if the mapping is neither one-

[15] This need not be a probability measure, although it could be made so by dividing by $V_{prim}(Im(F))$ with no change to the following analysis. In what follows, we do not need to assume the measures are normalized.

[16] This means that at every point $p \in M$, the differential $Jac(F)_p: T_pM \to T_{F(p)}N$ is a surjective linear map.

[17] The transpose here is taken with respect to the inner products induced on the tangent spaces by the metric, that is $(Jac(F)v, w)_N = (v, Jac(F)^T w)_M$ for all v and w. Note that for a fixed choice of bases, the matrix of $Jac(F)^T$ is not necessarily equal to the transpose of the matrix of $Jac(F)$. This is discussed further in section 7.1 of the Appendix.

to-one or onto. To do this, fix a measurable function $h: D \to \mathbb{R}$ and consider the measure hV_{aux}. We shall compute the density $\frac{d(F_*(hV_{aux}))}{dV_{prim}}$.

Consider an arbitrary measurable function $f: Im(F) \to \mathbb{R}$ and apply the coarea formula to the function $f(F(\eta))h(\eta)/J_F(\eta)$ to see

$$\int_{Im(F)}\left(\int_{F^{-1}\theta}\frac{f(F(\eta))h(\eta)}{J_F(\eta)}V_{F^{-1}\theta}(d\eta)\right)V_{prim}(d\theta) = \int_D f(F(\eta))h(\eta)V_{aux}(d\eta).$$

Note that for each $\eta \in F^{-1}\theta$, we have $f(F(\eta)) = f(\theta)$ (by definition of the inverse image). So we can pull this term out of the integral to see that

$$\int_{Im(F)} f(\theta)\left(\int_{F^{-1}\theta}\frac{h(\eta)}{J_F(\eta)}V_{F^{-1}\theta}(d\eta)\right)V_{prim}(d\theta) = \int_D f(F(\eta))h(\eta)V_{aux}(d\eta).$$

Since f was arbitrary, we have established the following result.

Main Formula 1. *The pushforward of the measure hV_{aux} is given by*

$$F_*(hV_{aux})(d\theta) = \left(\int_{F^{-1}(\theta)}\frac{h(\eta)}{J_F(\eta)}V_{F^{-1}(\theta)}(d\eta)\right)V_{prim}(d\theta).$$

Applying the above formula to the function $h(\eta) = \frac{J_F(\eta)g(F(\eta))}{Vol(F^{-1}(F(\eta)))}$, and noting that the Riemannian volume of the inverse image is given by $Vol(F^{-1}\theta) = \int_{F^{-1}\theta}V_{F^{-1}\theta}(d\eta)$, we conclude the following result.

Main Formula 2. *For any measurable function $f: M \to \mathbb{R}$, the measure*

$$\mu(d\eta) = \frac{J_F(\eta)g(F(\eta))}{Vol(F^{-1}(F(\eta)))}V_{aux}(d\eta),$$

satisfies $F_*\mu = gV_{prim}$.

If the denominator looks strange, we remind the reader that in general, $F^{-1}(F(\eta)) \neq \eta$, as F need not be one-to-one.

Finally, we observe that if F is not one-to-one, then there are infinitely many measures μ such that $F_*\mu = fV_{prim}$. Indeed, we see from the Main Formula 1 that the Radon-Nikodym derivative $\frac{d(F_*(hV_{aux}))}{dV_{prim}}(\theta) = \int_{F^{-1}\theta}\frac{h(\eta)}{J_F(\eta)}V_{F^{-1}\theta}(d\eta)$ depends only on the integral of h/J_F over the sets $F^{-1}(\theta)$, so changing the function g in such a way that each of these integrals is unchanged will have no effect on the pushforward. In this way, Main Formulas 1 and 2 generalize the main results of HW. Indeed, even if the desired prior on the primary parameter space were an arbitrary distribution, i.e., not the uniform distribution, Main

Formulas 1 and 2 describe the relationship necessary for specifying the correct prior on the auxiliary parameter space to obtain it.

3.2. Special Cases

The purpose of this section is to show how Main Formula 2 directly generalizes several familiar formulas, in particular, the multivariate change of variables formula as well as the formula for the area of a parametric surface.

In particular we will make the following simplifying assumptions for the remainder of this section: (1) F is injective, and (2), the ambient metric on the codomain $\mathbb{R}^n$ of F is the standard Euclidean metric. Observe that Assumption (1) requires that the dimension of the auxiliary parameter space not exceed the dimension of the codomain. This assumption also implies that F is a parametrization of the manifold $Im(F)$.

We will make use of the formula for J_F derived in Section 7.1 of the Appendix.

Before we begin the calculations, let us observe that each pre-image $F^{-1}(\boldsymbol{\theta})$ consists of a single point. Thus, strictly speaking, these sets are not Riemannian manifolds, and the measure $V_{F^{-1}\theta}$ is accordingly not defined. However, as we explained in the statement of the Coarea Formula, if M is a single point, the formula still holds provided that we define V_M so that $V_M(M) = 1$. In particular, under this convention, $Vol(F^{-1}\theta) = V_{F^{-1}\theta}(F^{-1}\theta) = 1$ for each θ. After we have finished our calculations, it will be apparent that this is the only sensible choice.

To calculate J_F, we endow D with the standard Euclidean metric. Consider now the bases $e_i = \partial_{\eta_i}$ and $f_i = Jac(F)e_i = \partial_{\eta_i}F$ of the tangent spaces $T_{\boldsymbol{\eta}}D$ and $T_{F(\boldsymbol{\eta})}Im(F)$. Note that both the matrix representation $jac(e, f)$ of $Jac(F)$ and the Gram matrix $G(e)$ are equal to the identity matrix.[18] Therefore, the formula in Section 7.1 of the Appendix for J_F simplifies to

$$J_F = \sqrt{det(G(f))}.$$

Recall that $G(f)_{ij} = (f_i, f_j)_{Im(F)}$. However, we assumed that the metric on $Im(F)$ was the restriction of the Euclidean metric on the ambient space, so we have $G(f)_{ij} = (f_i, f_j)$ where f_i and f_j are now considered as vectors in $\mathbb{R}^n$, and $(\cdot, \cdot)$ denotes the standard dot product. Therefore, we obtain the following result.

Main Formula 3. *Under the assumptions* (1) *and* (2) *at the beginning of Subsection 3.2, we have*

$$J_F(\boldsymbol{\eta}) = \sqrt{det(G(\{\partial_{\eta_i}F\}_i))}$$

where $G(\{\partial_{\eta_i}F\}_i)$ denotes the Gram matrix of the vectors $\{\partial_{\eta_i}F\}_i$, as defined in section 2.4.

[18] The general definitions of $Jac(F)$ and G are given in Appendix Section 7.4 and Section 2.4, respectively.

Now, let us further specialize to the case when $m = n = Im(F)$ (i.e, F is locally invertible). Consider the classical Jacobian matrix $jac_{ij} = \partial_{\eta_j} F_i$. We observe that

$$\begin{aligned}(\partial_{\eta_i} F, \partial_{\eta_j} F) &= \sum_k \partial_{\eta_i} F_k \partial_{\eta_j} F_k \\ &= \sum_k (jac)^t_{ik} (jac)_{kj} \\ &= ((jac)^t jac)_{ij}.\end{aligned}$$

giving us the matrix equality $G(f) = (jac)^t jac$. Since $m = n$, jac is a square matrix. Recalling basic properties of determinants, we obtain

$$\sqrt{det(G(f))} = \sqrt{det(jac)^2} = |det(\partial_{\eta_j} F_i)|.$$

By Main Formulas 2 and 3, we have shown

$$F_*(|\det(\partial_i F_j)(\eta)| d\eta) = d\theta.$$

Recalling the definition of the pushforward measure, this is equivalent to saying that for any measurable $g: Im(F) \to \mathbb{R}$, we have

$$\int_{Im(F)} g(\theta) d\theta = \int_D g(F(\eta)) \mid det(\partial_i F_j)(\eta) \mid d\eta,$$

which is nothing other than the standard change of variables formula. This special case coincides with the situation considered by HW, albeit using our notation.

The next case of interest is when $m = 2$, $n = 3$.

In this case, $G(f)$ is a 2×2 matrix, so we can simply compute the determinant:

$$\begin{aligned}det(G(f)) &= G(f)_{11} G(f)_{22} - G(f)_{12} G(f)_{21} \\ &= |\partial_{\eta_1} F|^2 |\partial_{\eta_2} F|^2 - (\partial_{\eta_1} F, \partial_{\eta_2} F)^2 \\ &= |\partial_{\eta_1} F \times \partial_{\eta_2} F|^2,\end{aligned}$$

where $\times$ is the usual cross product in $\mathbb{R}^3$.

Thus we obtain

$$F_*(|(\partial_{\eta_1} F \times \partial_{\eta_2} F(\eta))| d\eta) = V_{prim},$$

where V_{prim} is the measure on $Im(F)$ induced by the Euclidean metric on $\mathbb{R}^3$. This is nothing other than the formula for the area of a parametric surface.

Finally, although not immediately relevant to our purposes, it is also interesting to note that Fubini's Theorem can be deduced directly from the Coarea formula simply by letting $D = \mathbb{R}^2$ and $F(x, y) = x$.

3.3. Discrete Distributions

It is worth noting that the same procedure in Section 3.1 can be carried out for discrete distributions as well. That is, assume D is a finite set and let $F: D \to \mathbb{R}$ be a function. Then the preimages $F^{-1}y$ are disjoint sets and $\cup_{y \in Im(F)} F^{-1}y = D$. One immediately deduces the following.

Coarea Formula 2. *For any function $f: D \to \mathbb{R}$, and measure π on D, one has*

$$\sum_{x \in D} f(x)\pi(x) = \sum_{y \in Im(F)} \left(\sum_{x \in F^{-1}y} f(x)\pi(x) \right).$$

Arguing as before, we apply the formula to the function $g(F)$ for some function $g: Im(F) \to \mathbb{R}$ to see

$$\sum_{x \in D} g(F(x))\pi(x) = \sum_{y \in Im(F)} g(y) \left(\sum_{x \in F^{-1}y} \pi(x) \right).$$

Taken together, we obtain the following result.

Main Formula 4. *In the above setting, the pushforward $F_*\pi$ is given by* $(F_*\pi)(y) = \sum_{x \in F^{-1}y} \pi(x)$ for $y \in Im(F)$. *Moreover, given any measure ν on $Im(F)$, we have $F_*\pi = \upsilon$ where* $\pi(x) = \frac{\upsilon(F(x))}{cardinality(F^{-1}F(x))}$.

3.4. Discrete Approximations

The advantage in simplicity of the discrete formula over the continuous analogue is obvious. In practice, one may employ a scheme such as the following to approximate a continuous computation by a discrete one. Further, such approximations could be applied when the model in question doesn't have an analytic expression and must be simulated.

Suppose we are given a continuous measure hV_{aux} on the auxiliary space and we wish to compute the pushforward.

Let us divide the range $Im(F)$ into a finite set $\mathbb{F}$ of disjoint (full-dimensional) subsets which cover $Im(F)$. We consider the measure $\pi: \mathbb{F} \to \mathbb{R}, \pi(X) = (F_*(hV_{aux}))(X)$. By definition of the pushforward, we see that

$$\pi(X) = \int_{F^{-1}X} h(\eta) V_{aux}(d\eta).$$

We have made some progress because now instead of having to integrate over the zero-measure sets $F^{-1}\boldsymbol{\theta}$, each with its own measure $V_{F^{-1}\boldsymbol{\theta}}$, we need only integrate over finitely many positive measure sets $F^{-1}X$, all with the same known measure V_{aux}. These integrals can in turn be approximated by a straightforward Monte Carlo procedure. Specifically, we maintain an array $[n_X]_{X \in \mathbb{F}}$ and initialize $n_X = 0$ for each X. We then take successive samples $\boldsymbol{\eta}_i$ from the distribution hV_{aux}. For each sample, we determine which set X contains $F(\boldsymbol{\eta}_i)$ and then increase n_X by 1. After any number of samples, the estimate $\tilde{\pi}(X)$ for $\boldsymbol{\pi}(X)$ is simply

$$\tilde{\pi}(X) = \frac{n_X}{\sum_{X \in \mathbb{F}} n_X}.$$

As a basic consistency check, one may ask about the quality of the approximation as the discretization gets finer, in the sense that $\sup_{X \in \mathbb{F}} Diam(X) \to 0$, where $Diam(X)$ is the maximal distance between any two points of X. Choose a sample point $x_X \in X$ for each $X \in \mathbb{F}$ and consider the measure $\upsilon_X = \sum_{X \in \mathbb{F}} \delta_{x_X} \pi(X)$. One would hope that this measure converges to the true measure $F_*(hV_{aux})$ in some sense. In particular, one might hope for convergence in distribution; that is that $\int f d\upsilon_X \to \int f d(F_*(hV_{aux}))$ for all bounded measurable f. Unfortunately this does not hold in general. Indeed, if $Im(F) = \mathbb{R}$ and $F_*(hV_{aux})$ is the Lebesgue measure, then this would imply that the Riemann integral of every measurable function is equal to its Lebesgue integral, which is false. However, we do get convergence if we instead assume that f is continuous (or even that the set of discontinuities is countable), which is probably sufficient for most applications.[19]

In anticipation of Section 4, we note that the result of this procedure can also be used to construct an approximation to the uniform measure on $Im(F)$. To do this, we set $\mathbb{F}_{>0} = \{X \in \mathbb{F} : \tilde{\pi}(X) > 0\}$. Then consider the measure $\upsilon : \mathbb{F} \to \mathbb{R}$ where $\upsilon(X)$ is given by

$$\upsilon(X) = \begin{cases} \dfrac{V_{Im(F)}(X)}{\sum_{Y \in \mathbb{F}_{>0}} V_{Im(F)}(Y)} & X \in \mathbb{F}_{>0}, \\ 0 & \text{otherwise}. \end{cases}$$

This converges to $V_{Im(F)}$ in the same sense as before.

Of course, we still have to calculate $V_{Im(F)}(X)$ for each such X. If $Im(F)$ is not full-dimensional in the codomain, this might be challenging. But if $Im(F)$ is full-dimensional (i.e., F is full-rank), then we can simply choose $\mathbb{F}$ to be a collection of small boxes, in which case the computation of $V_{Im(F)}(X)$ is trivial (assuming the ambient metric is Euclidean).

We have included an interactive demonstration of this technique in a Jupyter Notebook in the online supplement.[20]

[19] Indeed, it is not hard to show that, for any probability measure μ, and continuous,compactly-supported function f the integral $\int f d\mu$ is the limit of any sequence of Riemann sums for which the maximum diameter of any set in the partition tends to zero.

Finally, we remark that this technique becomes intractable for ambient spaces of sufficiently large dimension. Indeed, a d-dimensional grid containing only 10 bins per dimension would have a total of 10^d boxes. Nonetheless, this procedure can be put to good use when the dimensions of the spaces in question are not too large, as we will see in our re-analysis of the SI model.

4. Statistical Inference

4.1. Computation of Evidence

Let us return to the setup of Subsection 3.1, in which we have a functional relation of the form $\boldsymbol{\theta} = F(\boldsymbol{\eta})$. Equivalently, the conditional distribution $p(\boldsymbol{\theta}|\boldsymbol{\eta})$ is a delta function $\delta_{F(\boldsymbol{\eta})}(\theta)$ centered at $F(\boldsymbol{\eta})$. We will now assume in addition that we have a likelihood function $L_{\boldsymbol{y}}(\theta) = P(\boldsymbol{y}|\theta)$, where $\boldsymbol{y}$ is the observed data. The motivating example is when $\boldsymbol{\theta}$ is a vector of success probabilities for a collection of binary decision problems, and L is the standard binomial likelihood. For a fixed prior $\boldsymbol{\pi}$ on $\boldsymbol{\eta}$, the Bayesian evidence of $\boldsymbol{y}$ is by definition $P_{\boldsymbol{\pi}}(\boldsymbol{y})$. To express this in terms of L, we first marginalize over $\boldsymbol{\eta}$, and then apply a further marginalization to $P(\boldsymbol{y}|\boldsymbol{\eta})$, in order to express it in terms of $\boldsymbol{\theta}$. Thus:

$$\begin{aligned} P_{\pi}(\boldsymbol{y}) &= \int_D P(\boldsymbol{y}|\boldsymbol{\eta})\pi(d\boldsymbol{\eta}) \\ &= \int_D \left(\int P(\boldsymbol{y}|\theta) P(\theta|\boldsymbol{\eta}) d\theta \right) \pi(d\boldsymbol{\eta}) \\ &= \int_D \left(\int P(\boldsymbol{y}|\theta) \delta_{F(\boldsymbol{\eta})}(\theta) d\theta \right) \pi(d\boldsymbol{\eta}) \\ &= \int_D P(\boldsymbol{y}|F(\boldsymbol{\eta}))\pi(d\boldsymbol{\eta}) \\ &= \int_D L_{\boldsymbol{y}}(F(\boldsymbol{\eta}))\pi(d\boldsymbol{\eta}) . \end{aligned}$$

By the change of variables formula, we have

$$P_{\pi}(\boldsymbol{y}) = \int_{Im(F)} L_{\boldsymbol{y}}(\theta)(F_* \pi)(d\theta) .$$

Thus we see the importance of the pushforward measure $F_*\pi$ for the purposes of evidence computation. The above formula shows that the evidence can only "see" the measure $F_*\pi$, and not the measure $\boldsymbol{\pi}$ itself. Accordingly, it is imperative to understand the pushforward of a given measure $\boldsymbol{\pi}$, which is given by our Main Formula 1. Conversely, it is sometimes the case that we want to set a prior ν directly on $Im(F)$, but that the geometry of the space $Im(F)$ is sufficiently complex to make a direct integral over the space difficult to compute, while the geometry of D is trivial (as in our examples to follow). In this case, one would want to instead carry out the integral in terms of $\boldsymbol{\eta}$. To do so, one must use a prior $\boldsymbol{\pi}$ with the correct pushforward $F_*\pi = \nu$. Doing so ensures that one gets the same answer as if the integral were computed directly over the space $Im(F)$; hence, our Main Formula 2.

4.2. Relation to Predictive Distributions

A reader with a background in Bayesian statistics might wonder about the relation between the pushforward measure $F_*\pi$ and the Prior Predictive Distribution.

Returning to the setting of Subsections 3.1 and 3.5, consider a prior distribution $\boldsymbol{\pi}$ on the parameters $\boldsymbol{\eta} \in \mathbb{R}^m$, a smooth function $F(\boldsymbol{\eta}) \in \mathbb{R}^m$ and a likelihood function $L_{\boldsymbol{y}}(\theta), \theta \in \mathbb{R}^n$. The Prior Predictive Distribution $PP(\boldsymbol{y})$ is by definition the likelihood of $\boldsymbol{y}$ averaged over the prior. Thus, we have[21]

$$PP(\boldsymbol{y}) = \int_{\mathbb{R}^m} L_{\boldsymbol{y}}(F(\boldsymbol{\eta}))\pi(d\boldsymbol{\eta}).$$

The pushforward $F_*\pi$ can also be expressed in a similar way. Consider the function $M_{\boldsymbol{\theta}}(\boldsymbol{\eta}) = \delta_{F(\boldsymbol{\eta})}(\boldsymbol{\theta})$. It is straightforward to verify that

$$(F_*\pi)(d\boldsymbol{\theta}) = \int_{\mathbb{R}_m} M_{\boldsymbol{\theta}}(\boldsymbol{\eta})\pi(d\boldsymbol{\eta})$$

We see that PP and $F_*\pi$ are both averages of appropriate functions with respect to the prior $\boldsymbol{\pi}$. They differ in the stage at which the average is taken. While the measure PP averages over both of the steps $\boldsymbol{\eta} \rightarrow \boldsymbol{\theta}$ and $\boldsymbol{\theta} \rightarrow \boldsymbol{y}$, the measure $F_*\pi$ averages only over the step $\boldsymbol{\eta} \rightarrow \boldsymbol{\theta}$ and cannot "see" the data-generating step $\boldsymbol{\theta} \rightarrow \boldsymbol{y}$. Thus, the pushforward is appropriate for studying relations between the quantities $\boldsymbol{\eta}$ and $\boldsymbol{\theta}$ which are purely internal to the theory, while the PP is appropriate if we are also interested in the process by which the data are generated, which is codified by the function $L_{\boldsymbol{y}}$.

Finally, there is a precise mathematical relation between PP and $F_*\pi$. In words, PP is obtained by averaging the likelihood L with respect to $F_*\pi$. That is,

$$PP(\boldsymbol{y}) = \int_{\mathbb{R}^n} L_{\boldsymbol{y}}(\theta)(F_*\pi)(d\theta).$$

This can be seen by applying the change of variables formula to the defining equation for PP.

5. Example using the Selective Integration Model

In this section, we present an illustration of prior transformation using the Selective Integration (SI) model of Tsetsos et al. (2016). Our motivation for this example is twofold. As a qualitative generalization of HW, the SI model does not place *explicit* order-constraints upon choice probabilities. We use the term "explicit" in the sense that it is not

[21] Note that although this bears a striking resemblance to the formula for the evidence of the result $\boldsymbol{y}$, it is not the same thing. While the Evidence is calculated for a specific empirical result $\boldsymbol{y}$ (and thus is a real number), the Prior Predictive Distribution is instead a probability distribution over the space of all possible results. Evaluating this distribution (or its density, in the case of a continuous distribution) at the point $\boldsymbol{y}$ gives the evidence of $\boldsymbol{y}$.

straightforward to describe the set of viable binary choice probabilities defined by the SI model as a set of non-linear inequality constraints. In other words, the allowable set of binary choice probabilities forms a smooth manifold in the appropriate probability space and it is not straightforward to geometrically describe this manifold. While this, and other models of choice, are not explicitly described as (potentially non-linear) constraints on parameters, the general prior transformation insights by HW hold to these cases as well. Secondly, we consider three versions of SI (the same versions considered by Tsetsos et al., 2016). All three versions of SI have the property that the corresponding mapping between auxiliary and primary parameter spaces is not one-to-one. One version of SI we consider, with no selective gating, admits a closed-form solution for the adjusted prior in terms of the auxiliary parameters. The other two versions do not, and therefore provide an illustration of our computational methods via discrete approximation.

We re-analyze all choice data from Experiment 4 in Tsetsos et al. (2016) for three versions of SI under both ‘adjusted’ and ‘unadjusted’ priors. We show that prior transformation has a major impact on the resulting Bayes factors for these data.

5.1. Selective Integration Definition

Tsetsos et al. (2016) presented a computational model of choice termed the *selective integration* model. This model is defined recursively as a sequential sampling process. Said simply, decision makers repeatedly sample information about sets of presented choice alternatives. This information is then integrated via accumulator functions. Cognitive as well as perceptual noise are explicitly modeled within this framework. A key aspect to the model is that choice alternative information from an inferior alternative is systematically downweighted relative to the superior alternative, i.e., selectively integrated. For example, if two plane tickets have the same price for the same destination but one ticket is first class then the economy class information for the second ticket will be downweighted further as a “loss.”

The following definition specifies the selective integration model for a binary forced choice between two options, A and B. Let $\mathcal{S}$ denote a finite set of choice alternatives. Let t denote the current discrete time step. Let $s_A(t)$, $s_B(t)$ denote stimulus intensities (e.g., bar heights in pixels) at t, and let $\rho_A(t), \rho_B(t)$ be standard normal random variables independent of each other and independent across different values of t. Let $\boldsymbol{X}_A(t)$, $\boldsymbol{X}_B(t)$ denote the corrupted subjective representations of the stimuli, due to *(early) noise* with noise parameter σ, defined by

$$\boldsymbol{X}_A(t) = \sigma \cdot \rho_A(t) + s_A(t), \qquad \boldsymbol{X}_B(t) = \sigma \cdot \rho_B(t) + s_B(t)$$

Let $w \in [0, 1]$ be a *selective gating parameter* and define a *gain function* ψ as the step-function

$$\psi(x) = \begin{cases} 1, & if\, x > 0, \\ w & if\, x < 0. \end{cases}$$

Define

$$\boldsymbol{I}_A(t) = \psi(\boldsymbol{X}_A(t) - \boldsymbol{X}_B(t)) \cdot \boldsymbol{X}_A(t), \qquad \boldsymbol{I}_B(t) = \psi(\boldsymbol{X}_B(t) - \boldsymbol{X}_A(t)) \cdot \boldsymbol{X}_B(t),$$

to serve as momentary inputs to *accumulators* $\boldsymbol{Y}_A(t)$ and $\boldsymbol{Y}B(t)$ that integrate the attribute values of the two sequences A and B according to the difference equations

$$\boldsymbol{Y}_A(t) = (1 - \lambda) \cdot \boldsymbol{Y}_A(t-1) + \boldsymbol{I}_A(t) + \xi \cdot \zeta_A(t),$$
$$\boldsymbol{Y}_B(t) = (1 - \lambda) \cdot \boldsymbol{Y}_B(t-1) + \boldsymbol{I}_B(t) + \xi \cdot \zeta_B(t),$$

with λ an *integration leak,* ξ *late noise*[22], and let $\zeta_A(t)$, $\zeta_B(t)$ be two standard normal random variables independent of each other and across different values of t. The accumulators are initialized as

$$\boldsymbol{Y}_A(0) = \boldsymbol{Y}_B(0) = 0$$

At the end of the presentation of stimuli, the model chooses sequence A if $\boldsymbol{Y}_A(t) > \boldsymbol{Y}_B(t)$, chooses sequence B if $\boldsymbol{Y}_B(t) > \boldsymbol{Y}_A(t)$, and chooses either with equal probability if $\boldsymbol{Y}_A(t) = \boldsymbol{Y}_B(t)$.

5.2. Base Model Specifications

We will analyze three different versions of the SI model described above, each with two different prior distributions. By a "version," we simply mean a specification of a subset of the four-dimensional auxiliary parameter space to which the model is constrained, leaving the prior distribution unspecified.

The first such version is simply the full model from the previous section, with an additional constraint on the maximum value of the parameters. This is done to make it possible to define a prior which is uniform over the parameters of the model. The bounds $0 \leq \lambda, w \leq 1$ are required by the model, but σ and ξ can a priori take any positive value. We impose the bounds $0 \leq \sigma, \xi \leq 20$. The second version is obtained from the first simply by requiring $\sigma = 0$ (no early noise), leaving the other bounds unchanged. The final version is obtained from the first by requiring $w = 1$ (no selective gating), again leaving the other bounds unchanged. For the first two of the above models, we estimated the value of the Bayesian evidence for the prior which is uniform over the corresponding primary parameter space using the procedure outlined in Subsection 3.4. We used a $50 \times 50 \times 50$ grid on the image space $[0, 1]^3$ of choice probabilities, and 10 million samples from the auxiliary parameter space. To estimate the vector of probabilities corresponding to each such sample, we used a Monte Carlo simulation with 10000 iterations.

In the case of $w = 1$, as we will see in the next section, an exact formula for the probabilities is available, as well as for the prior distribution which implies a uniform distribution on the primary parameter space. Therefore, the evidence values can be expressed as integrals over

[22]We assume $0 \leq \xi \in \mathbb{R}$.

the parameters of known functions (i.e., the product of the binomial likelihood with the prior density), which we estimated using a simple Monte Carlo algorithm, again with 10 million iterations.

The primary parameter spaces of the first two models can be approximated by simulation; that is, we randomly select a point in the auxiliary parameter space, and then plot the corresponding 3-tuple of probabilities. A consequence is that the spatial density of sample points in the primary parameter space approximates the pushforward of the uniform measure. The results are shown in Figure 1.

5.3. Prior Specification for Restricted Selective Integration

In case $w = 1$, Davis-Stober et al. (2017) showed that the choice probabilities P_{AB}, P_{BC}, P_{AC} can be expressed explicitly in terms of the remaining three parameters. We denote by F the corresponding function. Thus:

$$(P_{AB}, P_{BC}, P_{AC}) = F(\lambda, \sigma, \xi),$$

see Davis-Stober et al. (2017) for the explicit formula.

We are therefore in the situation of Subsection 3.1, in which the value of a parameter vector is given as a function of another parameter vector. In fact, the variables ξ and σ appear only in the combination $\sqrt{\xi^2 + \sigma^2}$, so we define the variable $\alpha = \sqrt{\sigma^2 + \xi^2}$.

As in the previous section, we used the bounds $0 \leq \lambda \leq 1$, $0 \leq \sigma, \xi \leq 20$.[23]

To conclude this section, we illustrate visually the difference between the two prior distributions (i.e., uniform over auxiliary parameter space vs. uniform over primary parameter space).

To sample from the primary parameter space when the auxiliary parameters are uniformly distributed, we can simply sample from the auxiliary space, then apply the function F to obtain a point on the primary parameter space. The distribution of the points so obtained approximates the pushforward of the uniform prior by F. The results are shown in Figure 2.

By contrast, we now construct a prior ν_1 on the parameters λ and α which implies a uniform distribution on the primary parameter space; thus $F_* \nu_1 = V_{prim}$. This can be done[24] with Main Formulas 2 and 3. The density is shown in Figure 3. Intuitively, the value of this density describes the amount that F "stretches" a small neighborhood around that point in

[23]The number 20 for the bounds on the ξ and σ parameters is somewhat arbitrary, and, unlike the bound on λ, does not reflect a theoretical constraint. In fact, the form of the adjusted prior can provide guidance on such parameter bounds. For example, the sharp peak in the density of Figure 3 suggests that the choice of cutoff will not dramatically affect the shape of the transformed prior, provided that the cutoff is sufficiently large that it encompasses most of the mass of the peak. Bear in mind that this figure was constructed to have a uniform implied distribution on the primary parameter space. Other transformed priors may have very different forms and the bounds on ξ and σ may matter for such cases.

[24]In fact, the hypotheses under which these formulas were derived are not strictly met, because the Jacobian determinant is not positive everywhere. However, the prior distribution constructed in this way will still have the required property, because it will, by construction, assign zero probability to the region on which the determinant vanishes.

the auxiliary space. A large value of the density means that a small neighborhood around the point covers a disproportionately large area in the primary parameter space.

For comparison, we turn to $\frac{d(F_*\nu_2)}{dV_{prim}}$, i.e., the density of the pushforward measure, where ν_2 is uniform over the three remaining model parameters. This is done in two steps. Consider the function $H(\lambda, \alpha)$, which is obtained from F by replacing $\sqrt{\sigma^2+\xi^2}$ with α, and $G(\lambda,\sigma,\xi) = (\lambda, \sqrt{\sigma^2+\xi^2})$. By construction $F = H \circ G$. We make use of the formula

$$(H \circ G)_* \nu_2 = H_*(G_* \nu_2)$$

which is a general property of the pushforward, and not specific to this situation.

Let us ignore the bounds $\sigma, \xi \leq 20$ for now (i.e., we take ν_2 to be an improper prior). We will explain how to incorporate them shortly.

A straightforward calculation shows the Jacobian matrix is given by

$$Jac_G(\lambda,\sigma,\xi) = \begin{pmatrix} 1 & 0 & 0 \\ 0 & \frac{\sigma}{\sqrt{\sigma^2+\xi^2}} & \frac{\xi}{\sqrt{\sigma^2+\xi^2}} \end{pmatrix},$$

and J_G^2 is simply the determinant of the product of this matrix with its transpose; thus $J_G = 1$.

By Main Formula 1, we see that $(G_*\nu_2)(d\lambda, d\alpha) = Vol(G^{-1}(\lambda,\alpha))d\lambda d\alpha$. Recalling that σ and ξ were constrained to be positive, we see that $G^{-1}(\lambda, \alpha)$ is geometrically a quarter-circle with radius α, so has volume $\pi\alpha/2$. Thus $(G_*\nu_2)(d\lambda, d\alpha) = (\pi\alpha/2)d\lambda d\alpha$.

We next turn to the function H. By the calculations in Subsection 3.2, J_H is given by the magnitude of the cross product of the columns of the corresponding Jacobian matrix. Note that the value of J_H is precisely what is plotted in Figure 3, modulo a constant multiplicative factor.

As can be seen in Figure 3, there are regions of the auxiliary parameter space in which the Jacobian determinant of H vanishes, making direct application of Main Formula 1 problematic. To proceed, we partition the space into two disjoint regions $C = \{J_H > 0\}$ and $D = \{J_H = 0\}$. We may write[25] $G_*\nu_2 = (G_*\nu_2)_{|C} + (G_*\nu_2)_{|D}$, and $H_*(G_*\nu_2) = H_*((G_*\nu_2)_{|C}) + H_*((G_*\nu_2)_{|D})$ Then since $J_H > 0$ on C, we can apply the Main Formulas to compute the pushforward of $(G_*\nu_2)_{|C}$. The measure $H_*((G_*\nu_2)_{|D})$ is absolutely singular, which means it is supported on a null set of V_{prim}.[26]

[25] Recall that the restriction of a measure μ to a measurable subset A is defined by $\mu_{|A}(X) = \mu(A \cap X)$

[26] More abstractly, partitioning the auxiliary parameter space thusly gives the Lebesgue decomposition of the pushforward.

To summarize, a vanishing Jacobian determinant over a non-null set implies the pushforward measure is singular. Dividing up the auxiliary parameter space according to whether or not the Jacobian determinant vanishes has the effect of decomposing the pushforward into the sum of a continuous measure and an absolutely singular measure. We will now use Main Formulas 1 and 2 to compute the density of the continuous part.

Because $H_{/C}$ is one-to-one, each preimage $H^{-1}(p), p \in [0,1]^3$ consists of a single point (or is empty). Accordingly, it follows from Main Formula 1 and the expression for $G_*\nu_2$ that

$$\frac{d(H_*(G_*\nu_2)|C)}{dV_{prim}}(H(\lambda,\alpha)) = \frac{\pi\alpha}{2J_{H(\lambda,\alpha)}},$$

provided $(\lambda,\alpha) \in C$. Therefore, the density of the continuous part of $F_*\pi$, evaluated at $H(\lambda, \alpha)$, is given by $\frac{\pi\alpha}{2J_{H(\lambda,\alpha)}}$.

To incorporate the bound $\sigma,\xi \leq 20$, the only difference is that $\pi a/2$ must be replaced with $Vol(\{\pi^2 + \xi^2 = a, 0 \leq \sigma,\xi \leq 20\})$, which is geometrically the intersection of a quarter-circle with a square. Accordingly, the value of this volume remains $\pi a/2$ provided that $\alpha \leq 0$, but is smaller than $\pi a/2$ for larger[27] values of a, and zero if $\alpha \geq \sqrt{2} * 20$. In addition, one must multiply by a normalizing factor to ensure that ν_2 integrates to 1. Since exact knowledge of the density of $F_*\nu_2$ is not required for our analysis, we content ourselves with the computation for the improper version of ν_2, and leave the details of the extension to the case where ν_2 has bounded support to the interested reader.

As for the singular part, numerical simulations suggest that it is supported along the "ridge" at the bottom of the primary parameter space, as visible in Figure 2.

5.4. Data Illustration

In the previous sections, we described the 3 base models, and the two priors used for each model. For each combination of the 21 subjects and 6 stimuli orderings, we computed the Bayesian evidence for each of the 6 models. Thus, there were $3 \times 2 \times 21 \times 6 = 756$ values in total. In addition, for each subject and permutation combination, we computed the evidence for the "encompassing model" in which (P_{AB}, P_{BC}, P_{AC}) is chosen uniformly at random from the unit cube. Since the likelihood function is a product of binomials, it is easy to check that the evidence for this model is a product of beta functions and binomial coefficients. Using this as our comparison model, we obtained Bayes factors for each of the 6 models, 21 subjects, and 6 stimuli orderings. See also Heck, Hilbig, and Moshagen (2017) for a similar approach. We used the data from "cyclic trials" in Tsetsos et. al.'s Experiment 4.[28] In this experiment, partipants were presented with sequences of choice stimuli. These stimuli were comprised of bar graphs indicating attributes of job candidates. After a fixed number of stimuli presentations, participants were asked to select the candidate who had, on

[27]The exact value for the case $20 \leq \alpha \leq \sqrt{2} * 20$ may be worked out with simple calculus.

[28]The full results and Matlab code are available at figshare.com/articles/Segert_Davis-Stober_2017_Supplementary_Material_/5686648

average, superior attributes. The choice stimuli were designed to induce violations of the transitivity axiom. We refer readers to Tsetsos et. al (2016) for the complete experimental descriptions.

Figure 4 shows the distribution of Bayes factors thus obtained, broken down by model and prior. As per Jeffrey's (1961) classification, we considered a Bayes Factor above 3.16 to be substantial evidence in favor of the model in question over the encompassing model, a value between 1 and 3.16 to be very weak evidence in favor, a value between .316 and 1 to be very weak evidence against, and a value less than .316 to be substantial evidence against.

As is demonstrated by the figure, adjusting the priors can make a difference in the overall performance of the model. As can be seen from the figure, the strength of evidence in the $\sigma = 0$ variant was enhanced by switching from a prior that is uniform over the auxiliary parameter space to one that is uniform over the primary parameter space space. By contrast, this same transformation had the opposite effect in the $w = 1$ model. Finally, the distribution of Bayes factors showed almost no change in the full model. This is consistent with Figure 1, which shows that samples drawn from the pushforward appear to be nearly uniformly distributed over the primary parameter space.

Next, we assess model performance at the individual-level. We calculated overall Bayes factors for each subject, for each model, by taking the product of Bayes factors across experimental conditions - similar to the "group Bayes factor" of Prince et al. (2012). This allows us to select a best-fitting model for each subject via the usual Bayesian model comparison by Bayes factor. For models defined by a uniform distribution over the auxiliary parameters, all 21 participants were best described by the full SI model. For models defined by a uniform distribution over the primary parameters, 15 subjects were best described by the full SI model, while the remaining 6 were best described by the three parameter SI model. No subject was best described by the encompassing (unconstrained) model nor any model with $w = 1$. These results support Tsetsos et al.'s (2016) general conclusions, as gating ($w < 1$) was a necessary component for high-performing models. Our results provide some nuance to these conclusions as not all subjects appear to be using early noise in the same way under our analysis. We note that this is the first analysis of these data to directly compare the variants of the SI model to an unconstrained competitor model. This is also the first analysis of the SI model using the dis-aggregated data from Experiment 4. As argued by Davis-Stober et al. (2016), aggregating the data across experimental conditions, for each individual, can lead to data artifacts - see also the reply by Tsetsos et al. (2016b).

5.5. Parameter Estimation

Finally, we briefly consider the question of obtaining point estimates λ^* and a^* of the parameters of the restricted model. There are (at least) two nonequivalent ways to do this. We may maximize either the posterior density $\pi|D$, giving the standard *maximum a posteriori* estimate, or we may maximize the density $\frac{d(F_*(\pi|D))}{dV_{prim}}$ of the pushforward of this measure, expressed in terms of λ and a.

The corresponding densities are plotted in Figure 5 for a single subject in a single permutation setting.[29] The corresponding parameter estimates are then given by maximizing these two functions.

6. Discussion

The question of prior selection is one that has obvious implications for practitioners of Bayesian statistics. Many prescriptions for such a question, e.g., maximizing entropy subject to certain constraints or choosing a uniform distribution, depend upon the specific parameterization of a model[30], which can be problematic when there is not a canonical parametrization. In this paper, we have demonstrated a general procedure for describing how a given choice of prior transforms under a wide class of re-parameterizations, and demonstrated explicitly how the results obtained thereby differ from simply applying a uniform prior to both parametrizations. Our results provide a comprehensive set of tools for researchers to construct priors, over either the primary or auxiliary parameter space, which are equivalent to a specified prior over the other parameter space, thereby generalizing the core results of HW.

In our illustration using the Selective Integration model of Tsetsos et al., we solved for the prior distribution on the auxiliary parameter space that yielded a uniform prior over the primary parameter space. In this case, it happens that both the primary parameter space and the space of empirical choice frequencies (data outcomes) are both subsets of the same space. This connects, at least conceptually, to the work of Chandramouli and Shiffrin (2016), who developed a framework for Bayesian inference where priors are placed at the level of data/outcomes. Our results are not limited to the usage of non-informative priors and could be used to construct informative priors for inference, along the lines described by Lee and Vanpaemel (2017).

Considerations of prior transformation can arise in many contexts. For example, Danaher et al. (2012) developed a Bayesian model in which the parameters of interest were constrained to lie within a polyhedral space. To enable efficient posterior sampling, they re-parameterized these values using a Minkowski-Weyl decomposition. Said simply, using this decomposition, any value within this polyhedral space can be represented as a simple linear sum of alternative parameters. Viewed within our results, the parameters corresponding to the Minkowski-Weyl decomposition serve as the auxiliary parameters. Our results make explicit the relationship between these parameter spaces and allow solving for priors over the Minkowski-Weyl decomposition that yield desired priors over the original (primary) parameters.

---

[29] More definitely, the data are from subject 4,in the identity permutation setting.

[30] More explicitly, suppose our model is parametrized by a single parameter $x \in [0, 1]$. If we were to set a prior on x by maximizing the entropy, we would have $x \sim Unif(0, 1)$.
However, if another researcher has the same theory, but instead chooses to parametrize it by the parameter $y = x^2 \in [0, 1]$, and defines a prior on y in the same way, then his prior will differ from ours even after accounting for the re-parametrization, in the sense that the pushforward of our prior by the function $f(x) = x^2$ will not be equal to his prior. In fact, while his prior will be uniform over y, it is easy to check that $\frac{d(f_* Unif(0, 1))}{dy} = \frac{1}{2\sqrt{y}}$.

The primary contribution of our paper regards the setting of priors over re-parameterized spaces so that the researcher obtains their desired model. Within the contexts we considered, the parameter space of interest will be constrained, either linearly or, more generally, non-linearly. We did not go into details on how to efficiently calculate Bayes factors over such constrained spaces. For that, we refer the reader to Hoijtink (2011). Equally important is determining the inequality constraints themselves, i.e., developing the order-constrained model in which to test, - see Mulder (2016) and Dittrich, Leenders, and Mulder (2017) for further discussion on "informative hypotheses" and related concepts.

Finally, we comment on what some readers may consider to be the relatively high level of mathematical sophistication with which we formulated our results. In particular, some readers may wonder whether the same goal could be achieved with less machinery. We reiterate from Section 2.1 that our use of this machinery was necessitated by the following two features of the Tsetsos model: (1) The parameter space has positive codimension relative to the ambient space, and (2) The parameter transformation is not one-to-one, and in fact is infinitely-many-to-one. Furthermore, as we demonstrated in Section 2.1, these features, far from being idiosyncrasies specific to the Tsetsos model, can in fact arise quite naturally in much simpler settings. As such, we believe that this example provides a convincing demonstration of both the necessity and the power of the differential-geometric tools which we have introduced.

Further directions might include the development of efficient sampling techniques for measures derived from a Riemannian volume form, especially on manifolds which are not parametrized simply by a single coordinate patch.

Acknowledgments

This work was supported by NSF grants SES 14-59866 (PI: C. Davis-Stober) and NIH grant (K25AA024182, PI: C. Davis-Stober). Any opinions, findings, conclusions, or recommendations expressed in this publication are those of the authors and do not necessarily reflect the views of their funding agencies or universities.

7.: Mathematical appendix

7.1. Expression for the Jacobian

The goal of this appendix is to derive a "local" expression for J_F in terms of arbitrary bases of the tangent spaces.

Consider bases e_i and f_i of $T_{\eta}D$ and $T_{F(\eta)}Im(F)$ respectively. Let $jac(e, f)$ denote the matrix of $Jac(F)$ with respect to these bases. In other words, $Jac(F)e_i = \sum_j jac(e,f)_{ji} f_j$.

We denote by N the matrix of $Jac(F)^T$ with respect to the same bases. Then the matrix of $Jac(F)Jac(F)^T$ with respect to the basis f_i is given by $j(e, f)N$. Since the determinant does not depend on the choice of basis, we have $J_F = det(MN)$. It remains only to express N in terms of $jac(e, f)$. A word of caution: N is not necessarily equal to the matrix transpose of $jac(e, f)$.

In preparation for what follows, we define the *Gram matrices* $G(e)_{ij} = (e_i, e_j)_{aux}$ and $G(f)_{ij} = (f_i, f_j)_{prim}$ where the subscripts refer to the Riemannian metrics on the auxiliary and primary parameter spaces, respectively.

Recall that the transpose is defined by the equation

$$(v, Jac(F)(\eta)^T w)_{aux} = (Jac(F)(\eta)v, w)_{prim},$$

for any $v \in T_\eta D$ and $w \in T_{F(\eta)} Im(F)$. In particular,

$$(e_j, Jac(F)(\eta)^T f_i)_{aux} = (Jac(F))(\eta)e_j, f_i)_{prim}.$$

Recall that N denotes the matrix of $Jac(F)^T$ with respect to the bases e_i and f_j; thus $Jac(F)^T f_i = \sum_k N_{ki} e_k$ Taking the dot product of both sides with e_j, we see

$$\begin{aligned}(e_j, Jac(F)^T f_i)_{aux} &= (Jac(F)(\eta)e_j, f_i)_{prim} \\ &= \sum_k N_{ki} G(e)_{kj} = (N^t G(e))_{ij}\end{aligned}$$

On the other hand, we have

$$\begin{aligned}(Jac(F)(\eta)e_j, f_i)_{prim} &= \sum_k jac(e|f)_{kj}(f_k|f_i) \\ &= \sum_i jac(e|f)_{kj} G(f)_{ki} \\ &= (jac(e,f)^t G(f))_{ji} \\ &= (jac(e,f)^t G(f))^t_{ij} \\ &= (G(f)jac(e,f))_{ij}.\end{aligned}$$

Here the lower-case t denotes the matrix transpose, and should not be confused with the operator transpose, denoted by upper case T.

We therefore obtain the matrix equation $N^t G(e) = G(f)jac(e, f)$ for N, giving $N = (G(f)jac(e, f)G(e)^{-1})^t = G(e)^{-1}jac(e, f)^t G(f)$ (we have used the fact that the inverse of a symmetric matrix is again symmetric). Therefore, we obtain the formula:

$$J_F = det(Jac(F)Jac(F)^T) = \sqrt{det(jac(e,f)G(e)^{-1}jac(e,f)^t G(f))}.$$

7.2. Worked Example

In order to illustrate the techniques to those unfamiliar, we provide a worked example which is slightly more involved than the special cases discussed in the text. In particular, this example will deal with the case where the pre-images are manifolds, rather than single points.

Consider $D = \{(x, y) \in \mathbb{R}^2 : x^2 + y^2 \leq 1\}$ and let $F : D \to [0, 1]^2$, $F(x, y) = (x^2 + y^2, (x^2 + y^2)^2)$. Note that F is geometrically equal to the segment $\{(x, y) : y = x^2, x \in [0, 1]\}$.

We will suppose that the ambient space $[0, 1]^2$ is endowed with the Fisher-Rao metric g_{FR}, which may be written as follows

$$\begin{aligned}(g_{FR})_{(p,q)}((\partial_x)_{(p,q)}, (\partial_x)_{(p,q)}) &= \frac{1}{p(1-p)} \\ (g_{FR})_{(p,q)}((\partial_y)_{(p,q)}, (\partial_y)_{(p,q)}) &= \frac{1}{q(1-q)} \\ (g_{FR})_{(p,q)}((\partial_x)_{(p,q)}, (\partial_y)_{(p,q)}) &= 0.\end{aligned}$$

(More precisely, it is the Fisher-Rao metric, considering (p, q) to be the vector of success probabilities of two independent Bernoulli distributions with one trial). In what follows, we will drop the subscript for notational clarity.

We compute $F_x = 2x(1, 2(x^2 + y^2))$, $F_y = 2y(1, 2(x^2 + y^2))$. The tangent space $T_{F(x,y)}Im(F)$ is given by the span of these; thus we may take $(1, 2(x^2 + y^2))$ as a basis for this space.

To describe the corresponding measure V_{prim} on $Im(F)$, consider the "interval" $I_{a,b} = \{(x, y) : x = y^2, x \in [a, b]\} \subset Im(F)$. We will compute $V_{prim}(I_{a,b})$.

To do so, we consider the parametrization $\gamma : [a, b] \to I_{a,b}, \gamma(t) = (t, t^2)$. The metric at the point $\gamma(t)$ is given by $g_{11} = \frac{1}{t(1-t)}, g_{22} = \frac{1}{t^2(1-t^2)}, g_{12} = 0$. Since $\gamma'(t) = \partial_x + 2t\,\partial_y$, we see

$$|\gamma'(t)|_{FR} = \sqrt{\frac{1}{t(1-t)} + \frac{4t^2}{t^2(1-t^2)}},$$

and $Vprim(I_{a,b}) = \int_a^b |\gamma'(t)|_{FR}\, dt$.

We take V_{aux} to be simply the uniform measure on D.

Consider now the jacobian $Jac(F) : T_{(x,y)}D \to T_{F(x,y)}Im(F)$. We compute the matrix with respect to the bases $e_i = \partial_x, \partial_y$ and $f = (1, 2(x^2 + y^2))$ of the two spaces. Our computations above show that $F_i = Jac(e_i) = 2x_i f$, so the matrix is simply $jac_{x,y} = (2x, 2y)$.

Since we put the Euclidean metric on D, the Gram matrix $G(e)$ is the identity. Note that the Fisher-Rao metric at the point $F(x, y)$ is given by $g_{11} = \frac{1}{(x^2+y^2)(1-(x^2+y^2))}, g_{22} = \frac{1}{(x^2+y^2)^2(1-(x^2+y^2)^2)}, g_{12} = 0$, so

$$\begin{aligned}G(f) &= g_{11} + 4(x^2 + y^2)^2 g_{22} \\ &= \frac{1}{(x^2+y^2)(1-(x^2+y^2))} + \frac{4}{1-(x^2+y^2)^2} \\ &= \frac{5(x^2+y^2)+1}{(x^2+y^2)-(x^2+y^2)^3}\end{aligned}$$

By our result, we have

$$J_F(x,y) = \sqrt{det(jacG(f)jac^tG(e)^{-1})}$$
$$= \sqrt{4(x^2+y^2)G(f)}$$
$$= 2\sqrt{\frac{5(x^2+y^2)+1}{1-(x^2+y^2)^2}}.$$

Finally, we observe that $F^{-1}(a, b)$ is a circle with radius $\sqrt{a}$ and accordingly has length $2\pi\sqrt{a}$. Appealing to our Main Formula 2, we conclude that the measure

$$\mu = \frac{1}{2\pi\sqrt{(x^2+y^2)}}2\sqrt{\frac{5(x^2+y^2)+1}{1-(x^2+y^2)^2}}dxdy,$$

satisfies $F_*\pi = V_{prim}$, that is to say, that $\mu(F^{-1}U) = V_{prim}(U)$ for every measurable U. The direct verification of this result is a straightforward and instructive exercise, which we leave to the reader.

7.3. Definition of the Pushforward measure

While this is a standard notion in measure theory, we realize this its formal definition might look mysterious on a first reading. Considering how crucial this concept is to our results, we shall illustrate through examples how this concept in fact corresponds to something very natural and simple, and we hope that after reading this section the reader will be able to see the connection between the formal definition and this simple operation.

Firstly, consider a mound of dirt sitting on the unit square $[0, 1]^2$. The volume of dirt above a region $U \subset [0, 1]^2$ is $\int_U h(x, y)dxdy$, where $h(x, y)$ denotes the height in meters of the mound at the given point. Assume for convenience that the total volume of dirt is equal to 1 cubic meter. Taking an equivalent perspective, this mound defines a probability distribution μ_{mound} on $[0, 1]^2$, where $\mu_{mound}(U) = \int_U h$.

Imagine that there are very sturdy walls located along the x and y axes. Now further imagine driving a bulldozer into the mound in such a way that the metal plate in the front of the vehicle ends up parallel and very close to the wall along the x-axis, thereby squashing the mound into a very thin sheet lying along this wall. Allowing ourselves some pedagogic license, we will suppose that we have driven the bulldozer "infinitely close" to the wall, so that the 3-dimensional mound becomes a 2-dimensional sheet. Let $g(x)$ denote the height of the sheet so obtained at the point $x \in [0, 1]$. As before, we consider this sheet as defining a probability meausre μ_{sheet} defined on $[0, 1]$, where $\mu_{sheet}(V) = \int_V g(x)dx$ is the total *area* of dirt lying above the region $V \subset [0, 1]$.[31] Let $F(x, y) \in [0, 1]$ denote the position along the wall that a particle at the location $F(x, y) \in [0, 1]^2$ will end up at after we have pushed it with the bulldozer (evidently, $F(x,y) = x$). Then the measure μ_{sheet} corresponding to the

[31]The reader may verify that g does indeed integrate to 1, by applying Fubini's theorem.

squashed dirt mound is equal to the pushforward $F_*\mu_{mound}$ of the measure corresponding to the original mound.

For an alternative example in the discrete setting, let *Cities* denote the set of all American cities (which we take to also include all small towns, villages, etc.). Define a measure μ_{city} on this set by letting $\mu_{city}(C)$ denote the population of the city C. Similarly, let *States* denote the set of American states, and define the measure μ_{states} by setting $\mu_{states}(S)$ to be the population of the state S. Then $\mu_{state} = F_*\mu_{city}$ where F: *Cities* → *States* is the function whose value on a city is equal to the state in which it is located. We leave the reader to contemplate how this interpretation may be extended to other similar binning or coarsening procedures.

7.4. Alternative formulation of Jacobians on manifolds

In this subsection, we present an alternative presentation of defining Jacobians on manifolds. The main issue which this formulation addresses is the dependence of the previous definition on the choice of local coordinate systems. From a conceptual standpoint, removing this dependence is more satisfying, because it makes clear that the Jacobian is an *intrinsic* quantity depending only the geometry of the situation (i.e., the manifolds M and N, in addition to the map F), and not on how we choose to describe it (i.e., the local parametrizations C_M and C_N).

The conceptual leap needed to formulate this idea is to consider the Jacobian to be a *linear operator* rather than a matrix. To a manifold $M \subset \mathbb{R}^k$ and a point $p \in M$, we will define the *tangent space* T_pM to be the linear subspace of $\mathbb{R}^k$ which best approximates M near p. The Jacobian operator $Jac(F)(p): T_pM \to T_{F(p)}N$ of F at p is then the the best linear approximation of the function F near the point p.

Definition 1. *If $M \subset \mathbb{R}^n$ is a manifold, then the tangent space T_pM is given by $Span\{\partial_i C(0)\}$ where C is any local parametrization near p, and $\partial_i C(0)$ is the ordinary partial derivative, considered, as a vector in $\mathbb{R}^n$. It can be shown that the subspace so obtained, does not depend on the choice of C.*

Given a manifold M, one may wonder how to compute the tangent space. If $M = F(U)$ where $U \subset \mathbb{R}^k$ is an open set and $F: \mathbb{R}^n$ is a smooth one-to-one map, then this is easy: we have $T_pM = Span\{\partial_i F(F^{-1}(p))\}_{i=1}^k$, that is, the span of the ordinary partial derivatives of the function F, evaluated at the point $F^{-1}(p) \in U$. For a space like the sphere, which is instead defined by a constraint, it is less obvious how to proceed. We state an extension to the Implicit function theorem which addresses this:

Proposition 1. *Let $M = f^{-1}(0)$ where $f: \mathbb{R}^n$ is a smooth function with the property that the vector $(\partial_1 f(x), \ldots, \partial_n f(x))$ is nonzero for every $x \in f^{-1}(0)$.*

Then the tangent space at any point $p \in M$ is given by $T_pM = ker(\{\partial_i f(p)\}_i)$, i.e. the kernel of the Jacobian matrix of f.

As in the previous section, this result generalizes to the case when M is defined by multiple simultaneous constraints, we refer to Boothby (2003) for the details.

Applying this result to the sphere, we see that if $p = (x, y, z) \in S^2$, then $T_pS^2 = Ker(2x, 2y, 2z)$. In other words, T_pS^2 is equal to the orthogonal complement of the vector connecting p to the origin, as in accord with geometric intution.

Having seen how to define the spaces on which the Jacobian operates, let us now give its formal definition:

Definition 2. *Let $F: M \to N$ be a map between manifolds, and let $p \in M$. The Jacobian of F at p, $Jac(F)(p): T_pM \to T_{F(p)}N$ is the linear operator defined as follows:*

1. *Choose local parametrizations $C_M: U_1 \to M$ and $C_N: U_2 \to N$ such that $p \in C_M(U_1)$ and $F(p) \in C_N(U_2)$*
2. *Compute the ordinary Jacobian matrix of the function $C_N^{-1} \circ F \circ C_M: U_1 \to U_2$ evaluated at the point $C_M^{-1}(p) \in U_1$.*
3. *Recall from the definition of the tangent space that the sets $b_M = \{\partial_i C_M\}$ and $b_N = \{\partial_i C_N\}$ are bases of T_pM and $Tp(_p)N$ (here the partial derivatives are evaluated at the points C_M^{-1} (p and $C_N^{-1}(F(p))$, respectively).*
4. *Define Jac(F)(p) to be the linear operator whose matrix with respect to the bases b_M and b_N from step 3 is equal to the matrix from step 2.*

It can be checked that the linear operator so obtained does not depend on the parametrizations chosen in part 1.

Finally, let us explicitly compute a Jacobian matrix. Let us consider the map $F: S^2 \to \mathbb{R}$ defined by $F(x, y, z) = x + y + z^2$ for $(x,y,z) \in S^2$. We will compute the Jacobian of this map at the south pole SP.

As in the definition, let us denote $M = S^2$ and $N = \mathbb{R}$.

To perform the computation, we simply follow the recipe given in the definition.

1. As described in the previous section, we may take the local parametrization of S^2 near the pole to be the function $C_M: B^2_{1/2}(0) \to S^2, C_M(x, y) = (x, y, -\sqrt{1 - x^2 - y^2}$. For the codomain, we may take the trivial parametrization $C_N: \mathbb{R} \to \mathbb{R}, C_N(x) = x$.

2. The function $\tilde{F} = C_N^{-1} \circ F \circ C_M$ is given by $\tilde{F}(x, y) = F(C_M(x, y)) = x + y + (1 - x^2 - y^2)$. By standard calculus, the Jacobian matrix of this function at the point (x, y) is given by $(1 - 2x, 1 - 2y)$. We need to evaluate this matrix at the point $(x, y) = C_M^{-1}(SP) = (0, 0)$. We denote the matrix so obtained by *jac*; thus *jac* = (1, 1).

3. We have $\partial_1 F(0, 0) = (1, 0, 0)$ and $\partial_2 F(0, 0) = (0, 1, 0)$. Thus $b_M = \{(1, 0, 0), (0, 1, 0)\}$ is a basis for the tangent space $T_{SP}S^2$. Similarly, $b_N = \{1\}$ is a basis for $T_{F(SP)}\mathbb{R}$ Let us denote the elements of these bases by b_M^1, b_M^2 and b_N^1.

4. The Jacobian derivative $Jac(F)(p): T_{SP}S^2 \to T_{F(SP)}\mathbb{R}$ is uniquely defined by the requirements that $Jac(F)(SP)(b_M^1) = jac_{1,1}b_N^1$ and $Jac(F)(SP)(b_M^2) = jac_{1,2}b_N^1$, where $jac_{i,j}$ denote the entries of the matrix $jac = (1, 1)$ from above. In this case, $Jac(F)(SP)$ maps each element of the basis $\{(1, 0, 0), (0, 1, 0)\}$ of $T_{SP}S^2$ to $1 \in T_{F(SP)}\mathbb{R}$

We invite the reader to consider different parametrizations in Step 1, and verify that this same linear operator is obtained.

8. References

[1]. Amari S (1985). Differential-geometrical methods in statistics. Springer-Verlag, Berlin.

[2]. Batchelder WH, & Riefer DM (1999). Theoretical and empirical review of multinomial process tree modeling. Psychonomic Bulletin & Review, 6, 57–86. [PubMed: 12199315]

[3]. Boothby WM (2003). An introduction to differentiable manifolds and Riemannian geometry. Academic press.

[4]. Chandramouli SH, & Shiffrin RM (2016). Extending Bayesian induction. Journal of Mathematical Psychology, 72, 38–42.

[5]. Chavel I (2006). Riemannian geometry: a modern introduction. Cambridge university press.

[6]. Danaher MR, Roy A, Chen Z, Mumford SL, & Schisterman EF (2012). Minkowski–Weyl priors for models with parameter constraints: an analysis of the biocycle study. Journal of the American Statistical Association, 107, 1395–1409. [PubMed: 27099406]

[7]. Davis-Stober CP, Brown N, Park S, & Regenwetter M (2017). Recasting a biologically motivated computational model within a Fechnerian and random utility framework. Journal of Mathematical Psychology, 77, 156–164. [PubMed: 28827888]

[8]. Davis-Stober CP, Park S, Brown N, & Regenwetter M (2016). Reported violations of rationality may be aggregation artifacts. Proceedings of the National Academy of Sciences, 113, E4761–E4763.

[9]. Dittrich D, Leenders RTA, & Mulder J (2017). Network autocorrelation modeling: A Bayes factor approach for testing (multiple) precise and interval hypotheses. Sociological Methods & Research, . doi:10.1177/0049124117729712.

[10]. Heck DW, Hilbig BE, & Moshagen M (2017). From information processing to decisions: Formalizing and comparing psychologically plausible choice models. Cognitive psychology, 96, 26–40. [PubMed: 28601709]

[11]. Heck DW, & Wagenmakers E-J (2016). Adjusted priors for Bayes factors involving reparameterized order constraints. Journal of Mathematical Psychology, 73, 110–116.

[12]. Hoijtink H (2011). Informative hypotheses: Theory and practice for behavioral and social scientists. Chapman and Hall/CRC.

[13]. Jeffreys H (1961). The theory of probability. Oxford University Press.

[14]. Klauer KC, Singmann H, & Kellen D (2015). Parametric order constraints in multinomial processing tree models: An extension of Knapp and Batchelder (2004). Journal of Mathematical Psychology, 64, 1–7.

[15]. Lee MD, & Vanpaemel W (2017). Determining informative priors for cognitive models. Psychonomic Bulletin & Review, (pp. 1–14).

[16]. Moshagen M (2010). multitree: A computer program for the analysis of multinomial processing tree models. Behavior Research Methods, 42, 42–54. [PubMed: 20160285]

[17]. Mulder J (2014). Prior adjusted default Bayes factors for testing (in)equality constrained hypotheses. Computational Statistics & Data Analysis, 71, 448–463.

[18]. Mulder J (2016). Bayes factors for testing order-constrained hypotheses on correlations. Journal of Mathematical Psychology, 72, 104–115.

[19]. Prince M, Brown S, & Heathcote A (2012). The design and analysis of state-trace experiments. Psychological Methods, 17, 78–99. [PubMed: 22040373]

[20]. Singmann H, & Kellen D (2013). Mptinr: Analysis of multinomial processing tree models in R. Behavior Research Methods, 45, 560–575. [PubMed: 23344733]

[21]. Tsetsos K, Moran R, Moreland J, Chater N, Usher M, & Summerfield C (2016). Economic irrationality is optimal during noisy decision making. Proceedings of the National Academy of Sciences of the United States of America, 113, 3102–3107. [PubMed: 26929353]

[22]. Tsetsos K, Moran R, Moreland JC, Chater N, Usher M, & Summerfield C (2016). Reply to Davis-Stober et al.: Violations of rationality in a psychophysical task are not aggregation artifacts Proceedings of the National Academy of Sciences, (pp. E4764–E4766).

[23]. Tversky A, & Kahneman D (1992). Advances in prospect theory: Cumulative representation of uncertainty. Journal of Risk and Uncertainty, 5, 297–323.

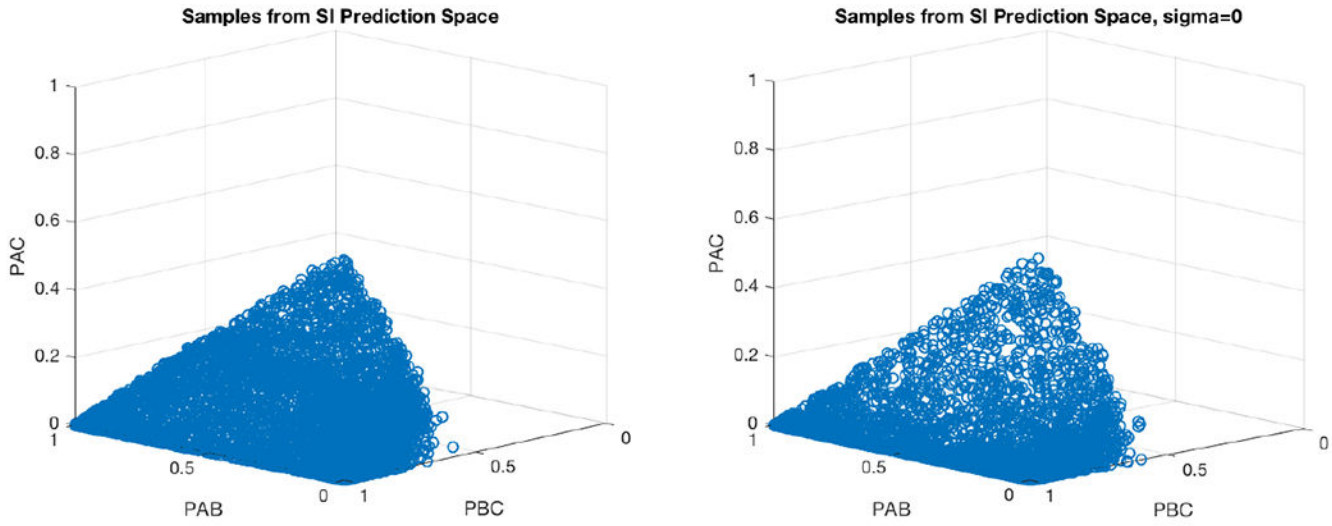


Figure 1:
Samples from the SI primary parameter space (identity stimulus order)

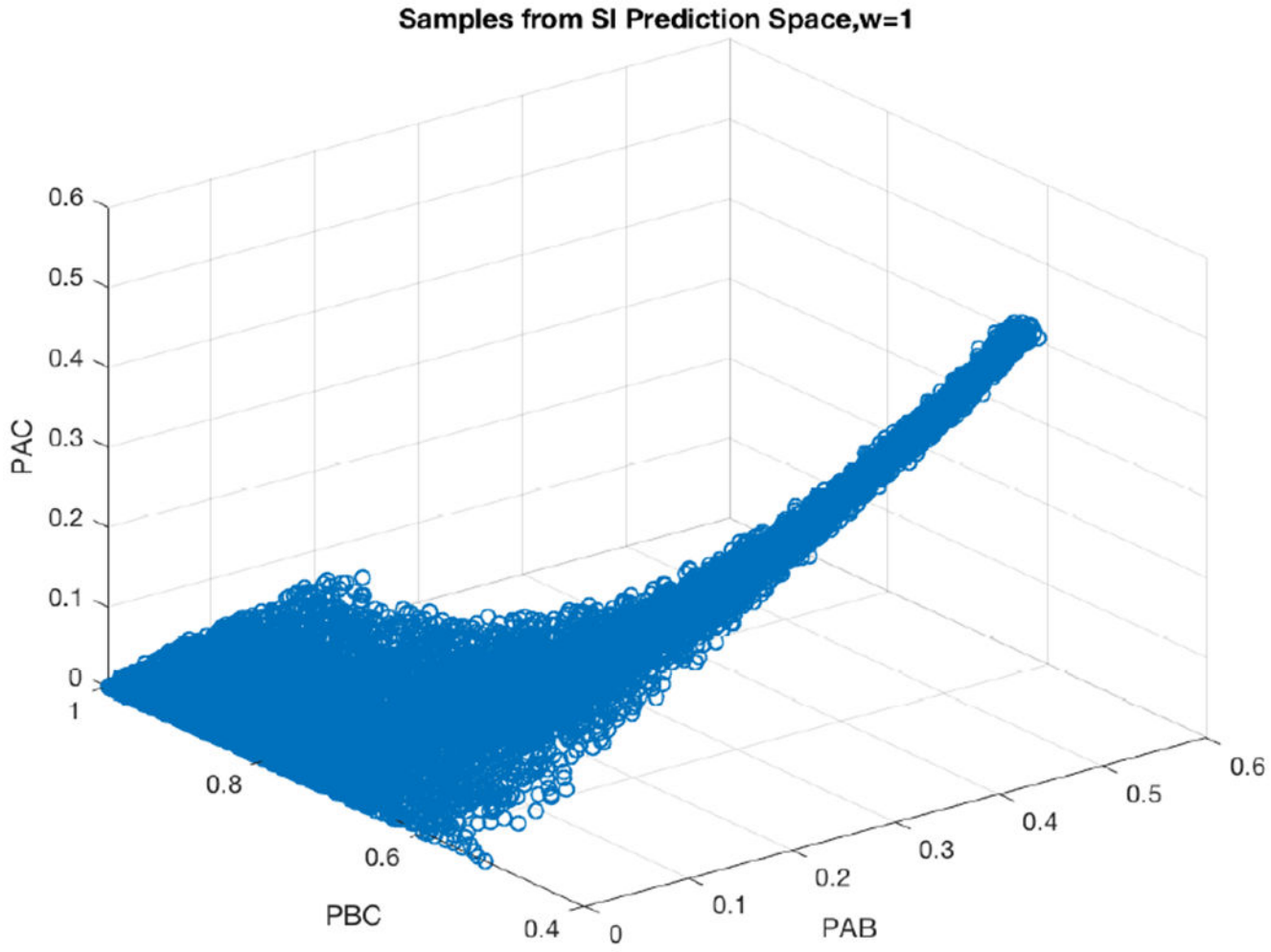


Figure 2:
Samples taken from the pushforward of a uniform measure (for the identity stimuli ordering). The pushforward is singular along the “ridge” where the surface meets the PBC axis.

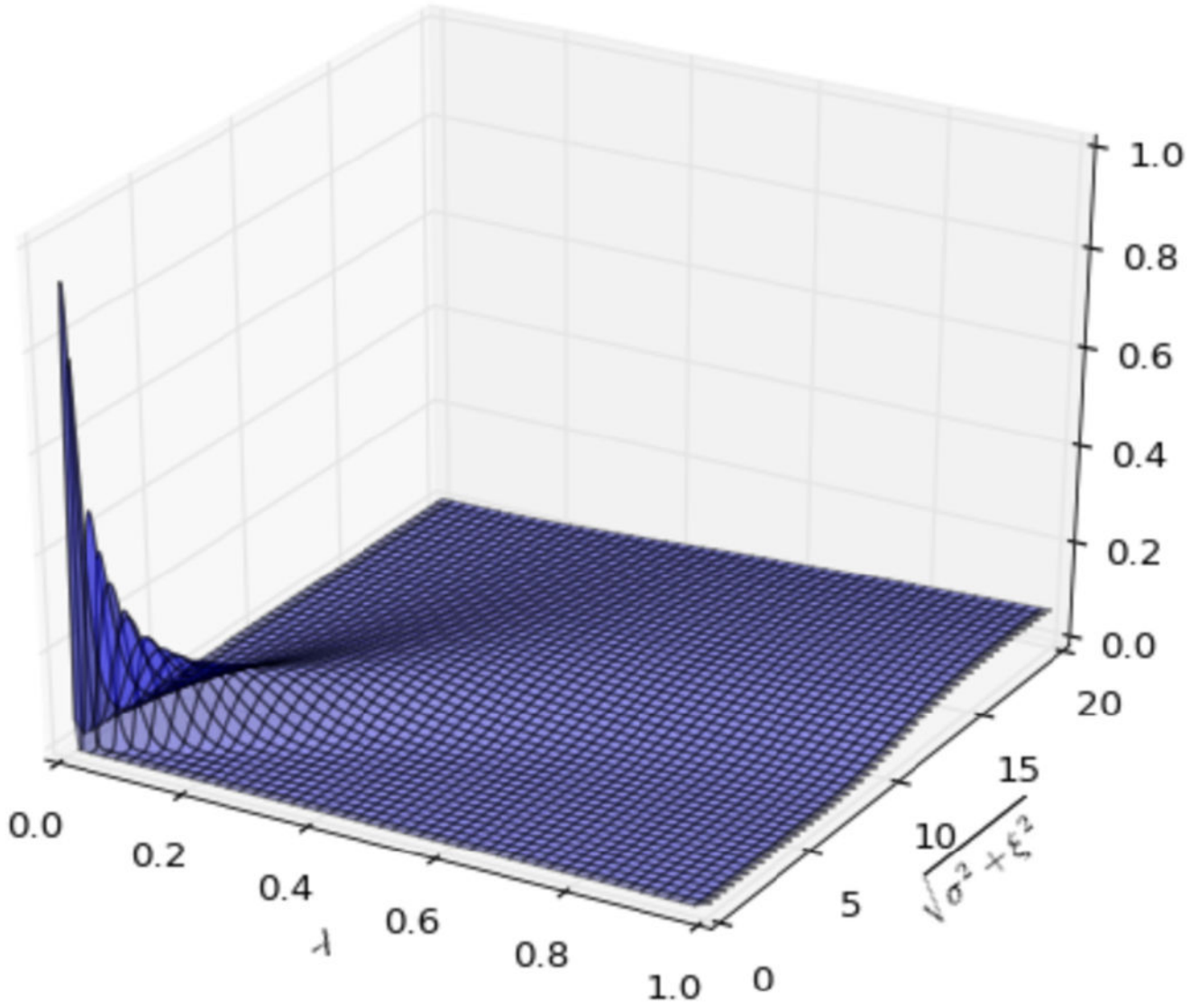


Figure 3:
The density of the measure ν, which implies a uniform distribution on the primary parameter space, for the SI model with $w = 1$

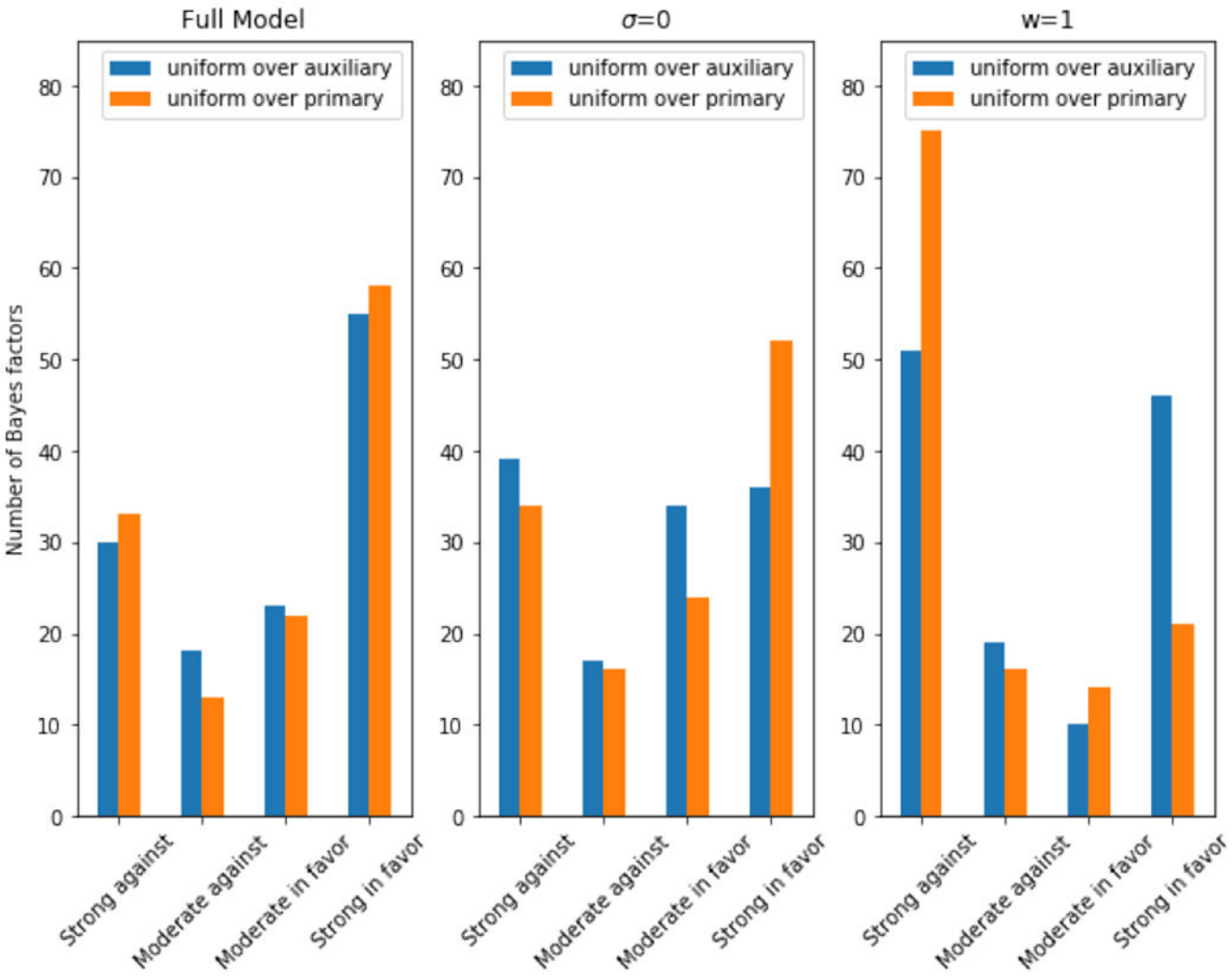


Figure 4:
Distribution of Strength of Evidence against Encompassing Model, by Model and Prior

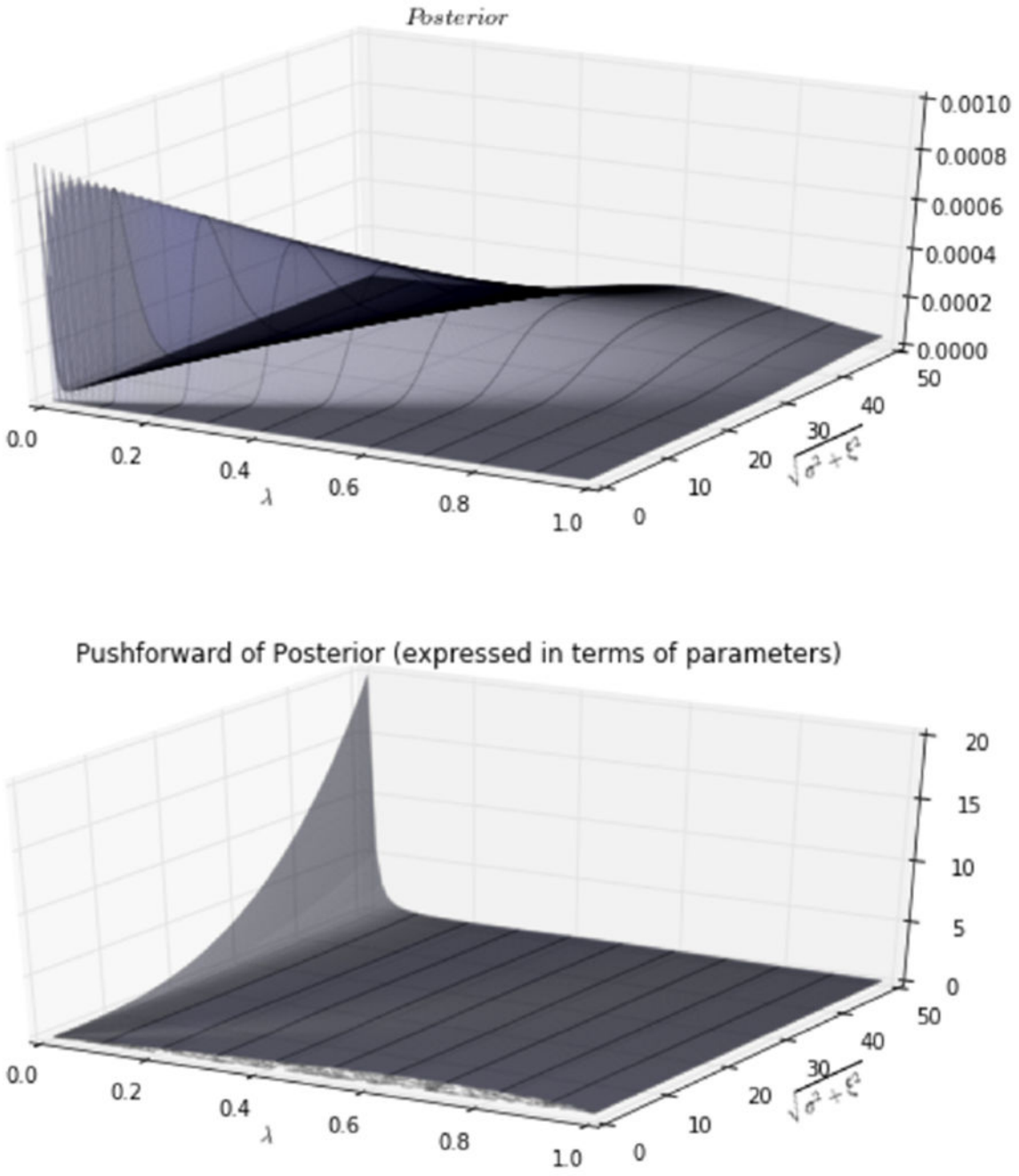


Figure 5:
Density of Posterior vs. Density of Pushforward of Posterior