INRIA

INSTITUT NATIONAL DE RECHERCHE EN INFORMATIQUE ET EN AUTOMATIQUE

# *On Bayesian Estimation in Manifolds*

Ian Jermyn

**N° 4607**

Octobre 2002

THÈME 3

Rapport de recherche

ISSN 0249-6399

# On Bayesian Estimation in Manifolds

Ian Jermyn

Thème 3 — Interaction homme-machine,
images, données, connaissances
Projets Ariana

Rapport de recherche n° 4607 — Octobre 2002 — 24 pages

**Abstract:**
It is frequently stated that the maximum *a posteriori* (MAP) and minimum mean squared error (MMSE) estimates of a continuous parameter $\gamma$ are not invariant to arbitrary "reparametrizations" of the parameter space $\Gamma$. This report clarifies the issues surrounding this problem, by pointing out the difference between coordinate invariance, which is a *sine qua non* for a mathematically well-defined problem, and diffeomorphism invariance, which is a substantial issue, and provides a solution. We first show that the presence of a metric structure on $\Gamma$ can be used to define coordinate-invariant MAP and MMSE estimates, and we argue that this is the natural and necessary way to proceed. The estimation problem and related geometrical quantities are all defined in a manifestly coordinate-invariant way. We show that the same MAP estimate results from 'density maximization' or from using an invariantly-defined delta function loss. We then discuss the choice of a metric structure on $\Gamma$. By imposing an invariance criterion natural within a Bayesian framework, we show that this choice is essentially unique. It does not necessarily correspond to a choice of coordinates. The resulting MAP estimate coincides with the minimum message length (MML) estimate, but no discretization or approximation is used in its derivation.

**Key-words:** estimation, MAP, MMSE, mean, invariance, Bayesian, manifold, metric

Unité de recherche INRIA Sophia Antipolis
2004, route des Lucioles, BP 93, 06902 Sophia Antipolis Cedex (France)
Téléphone : +33 4 92 38 77 77 — Télécopie : +33 4 92 38 77 65

# Sur l'Estimation Bayesienne dans les Variétés

**Résumé :**

Il est fréquemment dit que les estimées au sens du maximum *a posteriori* (MAP) et du minimum de l'erreur quadratique moyenne (MMSE) d'un paramètre continu $\gamma$ ne sont pas invariantes relativement aux "reparamètrisations" de l'espace des paramètres $\Gamma$. Ce rapport clarifie les questions autour de ce problème, en soulignant la différence entre l'invariance aux changements de coordonnées, qui est une condition *sine qua non* pour un problème mathématiquement bien défini, et l'invariance aux difféomorphismes, qui est une question significative, et fournit une solution. On montre d'abord que la présence d'une structure métrique sur $\Gamma$ peut être utilisée pour définir les estimées aux sens du MAP et du MMSE qui sont invariantes aux changements de coordonnées, et on explique pourquoi cela est la façon naturelle et nécessaire pour le faire. Le problème de l'estimation et les quantités géométriques qui y sont associées sont tous définis d'une façon clairement invariante aux changements de coordonnées. On montre que la même estimée au sens du MAP est obtenue en utilisant soit la 'maximisation d'une densité' soit une fonction de perte delta, définie de façon invariante. Puis, on discute le choix d'une métrique pour $\Gamma$. En imposant un critère d'invariance qui est naturel dans le cadre bayesien, on montre que ce choix est unique. Il ne correspond pas nécessairement à un choix de coordonnées. L'estimée au sens du MAP qui en résulte coincide avec l'estimée fondée sur la longueur minimum de message (MML), mais la demonstration n'utilise pas de discrétisation ou d'approximation.

**Mots-clés :** estimation, MAP, MMSE, moyenne, invariance, bayesien, variété, métrique

# Contents

# 1 Introduction

Statistical estimation is a very old field, but despite that many questions remain unanswered and debates about the best way to proceed are plentiful. From a probabilistic point of view, all the information about a quantity of interest modelled by a space $\Gamma$ is contained in a probability measure on $\Gamma$. If it is deemed necessary to single out a particular point $\gamma \in \Gamma$ for some purpose, a loss function $L : \Gamma \times \Gamma \to \mathbb{R} : (\gamma, \gamma') \mapsto L(\gamma, \gamma')$ is defined describing the cost inherent in taking the true value of the quantity to be $\gamma$ when it is in fact $\gamma'$. The mean value of the loss as a function of $\gamma$ can be computed using the probability measure, whereupon one can, for example, choose that point $\hat{\gamma} \in \Gamma$ that minimizes the mean loss as ones estimate of the true value of $\gamma$.

In the case that $\Gamma$ is a manifold[1], difficulties appear to arise in this procedure. Two popular choices for loss functions on continuous parameter spaces are a delta function and the squared difference, leading to the MAP and MMSE estimates respectively. Unfortunately, these estimates are apparently not invariant to changes in coordinates on $\Gamma$. Such a situation is untenable. The coordinates are subject to the whim of the person making the estimate; thus lack of coordinate invariance leads to the paradox that two people given the same problem can make different estimates simply by choosing to use different coordinate systems, for example, polar rather than rectangular.

The purpose of this report is to correct the above situation. We contrast coordinate invariance, which is necessary for a well-defined problem, with diffeomorphism invariance, which does possess a coordinate-invariant meaning and therefore has content. The use of the word "reparametrization" confuses the issue by conflating these two types of invariance. We describe coordinate-invariant MAP and MMSE estimates for a manifold, and in the process show how to define coordinate-invariant versions of other estimates also. The main points of the report are as follows:

1. In practice, if not in principle, the manifolds of parameters that we wish to estimate are not merely manifolds but possess a metric. Gaussian distributions rely on this fact in using an inner product, and it is necessarily true in the case of MMSE estimates, where the metric is disguised by the unjustified assumption that it is Euclidean. A metric brings together the geometric and measure-theoretic aspects of $\Gamma$ and is the essential extra structure needed for coordinate-invariant estimation.

2. Coordinate-invariant MAP estimates can be defined using the coordinate-invariant measure provided by the metric, or alternatively by using the metric to define a coordinate-invariant delta function loss.

3. Coordinate-invariant MMSE estimates can be defined using the coordinate-invariant distance provided by the metric ("Karcher mean").

4. The choice of a metric (unlike coordinate invariance, which, as stated above, is a necessary pre-condition for a well-defined problem), is a substantive issue. Every measure space possesses a natural metric. The requirement that all information about the parameters be contained

[1]Although in principle, one could have a parameter space with a topology but no manifold structure, in practice this is rarely if ever the case.

either in their correspondence with data probability measures or in the prior probability measure, leads to the requirement that the metric on $\Gamma$ be the metric induced by its embedding via the likelihood in the space of data probability measures.

5. The resulting MAP estimate turns out to coincide with the minimum message length (MML) estimate, except that no discretization of $\Gamma$ is required and no approximations are made.

The report is structured thus. In section 2, we discuss the failure of invariance for MAP estimation on manifolds and its causes. In section 3, we describe how the problem can be solved by endowing the manifold with a metric structure and argue that this is the natural solution to the problem. We use this additional structure to describe invariant MAP estimates, both from the point of view of density maximization and as arising from a delta function loss, and an invariant generalisation of MMSE estimates. In section 4, we discuss the choice of metric structure, and use a simple invariance argument to show that there is essentially a unique possibility. In section 5, we discuss the conclusions of the report and related work.

# 2 The Problem

To illustrate the problem, we examine the maximisation of a probability density function (pdf) on a manifold of dimension $m$. Let the manifold be $\Gamma$, a point in $\Gamma$ being denoted $\gamma$. We are given a probability measure $\mathbf{Q}$ on $\Gamma$, which we may view as the posterior in a MAP estimation task, although this is not important at this stage. We are also given two systems of coordinates on $\Gamma$, $\theta : \Gamma \to \mathbb{R}^m$ and $\phi : \Gamma \to \mathbb{R}^m$. (We ignore questions of topology that might require us to use more than one coordinate patch: the issue is not central to the discussion here.)

Expressed in terms of the first set of coordinates $\theta$, and the corresponding measure $d^m\theta(\gamma)$ on $\Gamma$, we find $\mathbf{Q} = Q_\theta(\theta(\gamma))d^m\theta(\gamma)$, where $Q_\theta(\theta(\gamma))$ is a function. We now separate the function $Q_\theta$ from the measure and find the argument of its maximum value $\theta_{\max} \in \mathbb{R}^m$, giving an estimate of $\gamma$, $\hat{\gamma}_\theta = \theta^{-1}(\theta_{\max})$.

We may choose to express $\mathbf{Q}$ in another coordinate system, $\phi : \Gamma \to \mathbb{R}^m$. Using the measure defined by this coordinate system, we find that $\mathbf{Q} = Q_\phi(\phi(\gamma))d^m\phi(\gamma)$. If we now follow the same procedure as before, and find the argument of the maximum value of $Q_\phi$, $\phi_{\max}$, we find another estimate, $\hat{\gamma}_\phi = \phi^{-1}(\phi_{\max})$.

The problem is the following. Suppose that the two coordinate systems are related by a function $\alpha : \mathbb{R}^m \to \mathbb{R}^m$, so that $\theta(\gamma) = \alpha(\phi(\gamma))$. In this case, the measures with respect to the two coordinate systems are related by $d^m\theta(\gamma) = J[\alpha](\phi(\gamma))d^m\phi(\gamma)$, where $J[\alpha](\phi(\gamma))$ is the Jacobian of the coordinate transformation. This in turn means that the functions $Q_\theta$ and $Q_\phi$ are related by $Q_\phi(\phi(\gamma)) = Q_\theta(\theta(\gamma))J[\alpha](\alpha^{-1}(\theta(\gamma)))$.

The consequence is that the estimates obtained by maximising $Q_\theta$ and $Q_\phi$ are different, due to the presence of the Jacobian factor. Apparently our estimate of $\gamma$ depends upon the choice of coordinates, or in effect upon the whim of the person making the estimate. This may seem surprising: one thinks of the question "What is the most probable point in $\Gamma$?" and, by analogy with the discrete case, one expects an invariant answer.

The difference between the continuous and the discrete cases means however that the question being asked in the continuous case is not the previously cited one at all, but a slightly more complicated version. Given a coordinate system, $\theta$, the question being asked is "What is the infinitesimal volume element $\theta^{-1}(dz)$ in $\Gamma$ (where $dz$ is an infinitesimal coordinate volume in $\mathbb{R}^m$) that is most likely to contain the true point in $\Gamma$?"[2]. Using a different coordinate system, $\phi$ on the other hand, the question is "What is the infinitesimal volume element $\phi^{-1}(dz)$ that is most likely to contain the true point in $\Gamma$?". In general, $\theta^{-1}(dz)) \neq \phi^{-1}(dz)$. It is then clear that different answers are to be expected using different coordinate systems, because the question being asked is different in each case.

From a measure-theoretic point of view, what is happening is clear. The functions $Q_\theta$ and $Q_\phi$ are probability density functions. Any pdf is defined with respect to an underlying measure. The Radon-Nikodym derivative of the probability measure with respect to the underlying measure then gives the pdf. In the scenario just described, two different underlying measures are being used: $d^m\theta(\gamma)$ and $d^m\phi(\gamma)$. To expect them to yield the same results is unreasonable.

If one concentrates on the underlying measure, then there is no problem. In terms of $\theta$, the underlying measure is $d^m\theta(\gamma)$, while in terms of $\phi$, the same underlying measure is $J[\alpha](\phi(\gamma))d^m\phi(\gamma)$. Integration of either of these over a fixed subset of $\Gamma$ will produce the same result: they are the same measure. Using this fixed measure, the problem disappears: in terms of $\phi$, the pdf with respect to the underlying measure is $Q_\theta(\alpha(\phi(\gamma))) = Q_\theta(\theta(\gamma))$. The maxima of $Q_\theta(\alpha(\phi(\gamma)))$ with respect to $\phi$ agree completely with those of $Q_\theta(\theta(\gamma))$ with respect to $\theta$, in the sense that $\theta_{\max} = \alpha(\phi_{\max})$, which implies that $\theta^{-1}(\theta_{\max}) = \phi^{-1}(\phi_{\max})$. The points in $\Gamma$ that we find are the same. The problem is that, given an arbitrary coordinate system, we do not know which choice of coordinate is 'correct', and hence what the estimate should be. By effectively focusing on measures on $\mathbb{R}^m$, the coordinate space, rather than on underlying measures on $\Gamma$, the problem is created. How then to define a coordinate-invariant underlying measure with respect to which to take the Radon-Nikodym derivative?

# 3 Coordinate-invariant Estimates

If one wishes to discuss coordinate-invariant measures using an arbitrary set of coordinates, one must express the mathematics in a way that allows for this eventuality. Not to do so means that symbols such as $d^m\theta$ are not defined. The natural way to express integration on manifolds in a way that is manifestly free of coordinates, but that nevertheless allows the derivation of an expression in terms of an arbitrary coordinate system with the greatest of ease, is the language of forms. (Readers not familiar with this language may wish to look at appendix A, where we provide a brief introduction to forms and their uses, or at [3], which is a useful reference.) We are interested in probability measures. These can be integrated over $m$-chains, for example the whole manifold $\Gamma$, and as such are $m$-forms. In addition, they must be positive and normalised, so that they are 'probability $m$-forms'. The answer to the question at the end of the last section is then: give an $m$-form, since these are, by definition, coordinate-invariant.

[2]We use the notation $f^{-1}$ both for the inverse of a map $f : A \to B$, $f^{-1} : B \to A$, and for the pullback $f^{-1} : 2^B \to 2^A : B \supset Y \mapsto \{a \in A : f(a) \in Y\}$. Context serves to disambiguate the two usages.

In practice, the following considerations push us strongly in one direction: the introduction of a metric on the manifold $\Gamma$, and the use of a derived $m$-form as the underlying measure.

The first consideration is intuition: manifolds with a measure but no metric are strange objects. They do not correspond to our intuition of a surface or volume at all. The space of volume-preserving diffeomorphisms is much larger than the space of isometries, and allows severe distortions. An example is the mixing of two incompressible immiscible fluids. The initial "drop of oil in water" may end up smoothly distorted into dramatically different shapes. Intuitively, parameter spaces are metric: they possess rigidity. If we wish to be able to describe the geometric properties of the manifold as well as its measure-theoretic properties, a metric is necessary. In addition, it is quite hard to write down an expression for a measure on a manifold without implicitly assuming a metric. In practice, this means that metrics appear, albeit disguised, in the expressions for many probability measures. An example is the Gaussian distribution, where an inner product is used to define the exponent. An inner product on a vector space is equivalent to a constant metric, which allows identification of each tangent space with the vector space itself. In many other cases, the assumption of an Euclidean metric is made manifest by the appearance of an orthogonal inner product.

Second, whereas a metric is not strictly necessary in the case of MAP estimation, it is in the case of MMSE estimation, where its presence is again disguised by the unjustified assumption that it is Euclidean. Thus we would have to introduce this extra structure in any case to define invariant MMSE estimates. Once a metric has been introduced, the only natural underlying measure is that derived from the metric, to be defined in the sequel. Of course, one could choose to use a metric in one case and not the other, but this seems unnecessary and lacking in coherence.

What then is a metric and how does it define a measure? A metric $\mathbf{h}$ is the assignment, to each point $\gamma$ of $\Gamma$, of an inner product on the tangent space $T_\gamma\Gamma$ at $\gamma$. This is detailed in appendix A, where it is further explained how the existence of a metric allows us to map functions to $m$-forms using the Hodge star. Given a function $f$, we can thus create an $m$-form $\star_{\mathbf{h}} f$. The choice of function $f$ is dictated by consistency between the measure-theoretic and geometric aspects of the manifold. By choosing $f$ to be $\mathbb{I}$, the function identically equal to 1, the resulting $m$-form is preserved by isometries: in other words, maps that preserve length preserve volume also.

Being a form, the quantity $\mathbf{U_h} = \star_{\mathbf{h}}\mathbb{I}$ is invariantly defined. This is clear first because no coordinate system was used in its construction, but it can also be verified in detail. As described in appendix A, the expression for this form in the coordinate bases of coordinates $\theta$ is

$$\mathbf{U_h} \quad = \quad |\mathbf{h}|_\theta^{1/2} d^m\theta, \tag{1}$$

where $|\mathbf{h}|_\theta$ is the determinant of the metric components in the $\theta$ coordinate basis, and $d^m\theta$ is the coordinate basis element for the space of $m$-forms. To see the invariance of this measure explicitly, note that a change of coordinates $\alpha$ introduces a factor of $J[\alpha](\phi(\gamma))$ from $d^m\theta$, while the transformation of the determinant of the metric matrix elements from one basis to another introduces a factor of $J[\alpha](\phi(\gamma))^{-1}$. Thus, expressed in any coordinate system, the form of the measure is identical: $|\mathbf{h}|_\theta^{1/2} d^m\theta = |\mathbf{h}|_\phi^{1/2} d^m\phi$. To stress the point once again: the measure $d^m\theta(\gamma)$ has no coordinate-invariant meaning. If we try to express a measure in a general coordinate system in this way, we literally do not know what we are talking about.

## 3.1 Maximum Density Estimates

Given a probability $m$-form $\mathbf{Q}$, and another positive $m$-form $\mathbf{U}$, one defines the pdf of $\mathbf{Q}$ with respect to $\mathbf{U}$ by division:

$$Q = \frac{\mathbf{Q}}{\mathbf{U}}. \tag{2}$$

This is the equivalent of the Radon-Nikodym derivative in the language of forms. What now becomes of maximum density estimation? We simply have to use $\mathbf{U_h}$ in equation 2. If we choose a particular coordinate system $\theta$, so that $\mathbf{Q} = Q_\theta d^m\theta$, and $\mathbf{U_h} = |\mathbf{h}|_\theta^{1/2} d^m\theta$ then we have

$$Q = |\mathbf{h}|_\theta^{-1/2} Q_\theta. \tag{3}$$

The left-hand side of this equation is invariant to changes in coordinates. These will produce equal Jacobian factors in both the numerator and the denominator of equation 3, which will thus cancel out. Note also that this pdf does not result simply from a choice of coordinates. Although it may be possible to find a system of coordinates in which the determinant of the metric is constant, this is misleading in two ways. First, what is really happening is that a metric is being chosen. The naive approach really means choosing a metric whose determinant is constant in the coordinate system you already have, which is not a coordinate-invariant procedure. Second, in more than one dimension, although the determinant of the metric may be constant, it may not be possible to find a system of coordinates in which the metric itself is constant. This would imply that the manifold was flat, a statement that is coordinate-invariant and may not be true.

### 3.1.1 Expression in Terms of a Delta Function Loss

Usually the maximum density estimate is regarded as derived from the use of a particular loss function, $\delta(\theta(\gamma), \theta(\gamma'))$ on $\Gamma$. Given a probability $m$-form expressed in terms of $\theta$, $Q_\theta(\theta) d^m\theta$, this leads to the familiar recipe $\hat{\gamma}_\theta = \theta^{-1}(\arg\max_\theta Q_\theta(\theta))$, in apparent contradiction to the previous discussion. From this point of view, there is no need to define a pdf at all, since we were merely integrating with respect to the probability measure. What is going on?

The answer of course involves the same concepts as above. The quantity $\delta(\theta(\gamma), \theta(\gamma'))$ is not invariantly defined, since the measure against which to integrate it has not been given. In our context, the delta function (in fact there are effectively $m$ of them) is best viewed as the identity map from $\Lambda^p\Gamma$, the space of $p$-forms on $\Gamma$, to itself. As such, it is a $p$-form at its first argument (a point in $\Gamma$) and an $(m-p)$-form at its second argument (another point in $\Gamma$). It can thus be integrated against a $p$-form to produce another $p$-form. When $p = 0$, we recover the usual delta function that evaluates a function at its first argument. In our case however, we wish to integrate the delta function against an $m$-form, and thus $p = m$. The delta function is thus an $m$-form at its first argument and a 0-form, or function, at its second argument. The result of integrating it against the posterior measure is thus an $m$-form, and to create a function that we can maximize, we need to use the Hodge star. This again introduces the factor of $|\mathbf{h}|_\theta^{-1/2}$ that we see in equation 3 and that is implicit in equation 2.

An alternative point of view is to consider the delta function as a map from $\Lambda^p\Gamma$ to $\Lambda^{(m-p)}\Gamma$, making it an $(m-p)$-form at its first argument and a $p$-form at its second argument. In order to integrate this against a $p$-form, we can use the inner product on $\Lambda^p$ described in equation 39 of appendix A. In our case, this point of view makes the delta function a 0-form (function) at its first argument and an $m$-form at its second. The result of the integration is thus a function as required for maximization, but now we find that the use of the inner product has already introduced the factor of $|\mathbf{h}|_{\theta}^{-1/2}$, thus giving the same result as in the other two methods.

There is thus no conflict between these different ways of speaking.

## 3.2 MMSE Estimates

Along with the lack of invariance of maximum density estimates, it is frequently pointed out that MMSE estimates also lack invariance under general changes of coordinates. It is equally true that the mean itself has no coordinate-invariant meaning, and for the same reasons. In both cases, one is faced with adding or subtracting certain values. If these operations are performed on the coordinate values in a particular coordinate system, they will change with a change of coordinates. Equally, one cannot add or subtract points of $\Gamma$ directly: such operations are not defined unless $\Gamma$ possesses an algebraic structure of some kind, for example, is a vector space.

In practice, what is crucial to the MMSE estimate is not the particular form of the expression, but the notion of a distance between two points in $\Gamma$. If we wish to consider MMSE estimates in general coordinate systems, we must be able to define distances in a coordinate-invariant manner. The quantity necessary to do this is a metric on $\Gamma$. A vector space is then just the special case of a constant metric.

The distance between two points $\gamma$ and $\gamma'$ in $\Gamma$ is defined as follows. We first define the notion of the length of a path, and then define the distance between two points as the length of a minimum length path between them. We will labour the explanation somewhat, in order to demonstrate the difference between coordinate invariance and invariance to diffeomorphisms, which is a coordinate-invariant and therefore content-full concept.

Let $I$ be an interval of the real line, considered as a manifold (that is, without the structure of a field). Let $p_0$ and $p_1$ be the elements of its boundary. Let $\pi : I \to \Gamma$ be an embedding of $I$ in $\Gamma$ such that $\pi(p_0) = \gamma$ and $\pi(p_1) = \gamma'$. To define the length of the path (i.e. its volume), we need a 1-form on $I$, or in other words a measure, which we will then integrate over $I$. Now however we have an invariance criterion: we must ensure that the length we calculate depends only on the image of $I$ in $\Gamma$, and not on the precise mapping of points of $I$ to points of $\Gamma$. This amounts to saying that replacing $\pi$ by $\pi\epsilon$, where $\epsilon$ is an arbitrary boundary-preserving diffeomorphism, should not change the resulting length. Note that unlike coordinate invariance on $I$, which follows as soon as we integrate over the coordinates, this condition is a substantive one. As argued in appendix A, the only way to ensure this is to construct a metric on $I$ by pulling back a metric from $\Gamma$, and then using this metric in the normal way to construct a 1-form. We thus pull back the metric $\mathbf{h}$ on $\Gamma$ to give a metric $\pi^*\mathbf{h}$ on $I$. We then use the Hodge star of this metric to map $\mathbb{I}$ to a 1-form that can be integrated on $I$. In notation:

$$l(\pi) = \int_I \star_{\pi^*\mathbf{h}} \mathbb{I}. \tag{4}$$

To illustrate the ability to derive an expression in an arbitrary coordinate system from the coordinate-invariant expression 4, we introduce a coordinate system $t : I \to \mathbb{R}$ on $I$, with a corresponding coordinate basis given by $\frac{\partial}{\partial t}(p)$, and a coordinate system $\theta$ on $\Gamma$, with a corresponding coordinate basis given by $\frac{\partial}{\partial \theta^i}(\gamma)$. In these bases, the (single) component of the pulled back metric can be found to be

$$\begin{aligned}(\pi^*\mathbf{h})_p \left(\frac{\partial}{\partial t}(p), \frac{\partial}{\partial t}(p)\right) &= \mathbf{h}_{\pi(p)} \left(\frac{d\pi^i}{dt}(p)\frac{\partial}{\partial \theta^i}(\pi(p)), \frac{d\pi^j}{dt}(p)\frac{\partial}{\partial \theta^j}(\pi(p))\right) \\ &= h_{\pi(p),ij}\frac{d\pi^i}{dt}(p)\frac{d\pi^j}{dt}(p), \end{aligned} \tag{5}$$

where $h_{ij}$ are the components of the metric $\mathbf{h}$ in the $\theta$ coordinate system. Thus the result is simply the length of the tangent vector to the path $\pi$ in the metric $\mathbf{h}$. Rewriting equation 4 in terms of this expression, we find that

$$l(\pi) = \int_a^b dt \left(h_{\pi(t),ij}\frac{d\pi^i}{dt}(t)\frac{d\pi^j}{dt}(t)\right)^{1/2}, \tag{6}$$

where we have abused notation by using the same symbol $\pi$ for the map from $I$ to $\Gamma$ and its expression in terms of coordinates. The points $a \in \mathbb{R}$ and $b \in \mathbb{R}$ are the coordinate values of $p_0$ and $p_1$ respectively.

Given the length of a path, we can now define the distance between two points as

$$d_\gamma(\gamma') = d(\gamma, \gamma') = \min_{\pi \in \Pi(\gamma,\gamma')} l(\pi), \tag{7}$$

where $\Pi(\gamma, \gamma')$ is the space of paths with endpoints $\gamma$ and $\gamma'$.

This distance is coordinate-invariant, and can be integrated as is. One can now define the coordinate-invariant form of the mean squared error, which we will call the 'mean squared distance':

$$L(\gamma) = \int_\Gamma (d_\gamma)^2 \mathbf{Q}, \tag{8}$$

where $\mathbf{Q}$ is as usual a probability $m$-form. In terms of a particular coordinate system $\theta$ on $\Gamma$, one has

$$L(\theta) = \int_{\theta(\Gamma)\subset\mathbb{R}^m} d^m\theta'\, Q_\theta(\theta') d_\theta^2(\theta, \theta'), \tag{9}$$

where $d_\theta$ is the expression for the length in terms of the given coordinates. For a general metric it is of course hard to derive an analytic expression for $d$.

Having defined the mean squared distance $L$, we now define the minimum mean squared distance (MMSD) estimate as the set of minimizers of $L(\gamma)$. In the case that the metric is Euclidean, $L$ reduces to the mean squared error, as it should. The resulting MMSD estimate is then the mean, i.e., the MMSE estimate, and is unique. In other cases, the MMSD estimate provides a generalized mean, known as the "Karcher mean", first introduced in [4] as the centre of mass on a Riemannian manifold. It is a set of points in $\Gamma$, each of which minimises the mean squared distance to every other point of $\Gamma$. Note that the set of minimizers may contain more than one point of $\Gamma$. This does not present a problem as such. It simply means that from the point of view of the mean squared distance loss function, these points are equivalent.

# 4 Choice of Metric

We have argued that introducing a metric on the manifold $\Gamma$ is the natural, indeed essential in the case of MMSE estimation, step for achieving coordinate-invariant and therefore meaningful estimates. Given such a metric, we have shown how to use it to define coordinate-invariant MAP and MMSE estimates, and in general how to construct manifestly coordinate-invariant expressions for estimates. We now turn to the question that we have been conspicuously avoiding thus far. How is one to choose a metric on $\Gamma$? Is this entirely problem-dependent, or is there a 'natural' choice in every case?

So far, we have been dealing with a manifold $\Gamma$ and a probability measure $\mathbf{Q}$ on this manifold. This simplified the discussion and notation. We turn now to the case that is usually of interest: when $\mathbf{Q}$ is a posterior probability measure derived from a likelihood and a prior.

## 4.1 Likelihoods

We introduce a second space, $X$, the data space. We shall view $X$ as a manifold, but it can equally well be discrete. On $X$, one can define the space of measures, $\mathcal{M}(X)$, which is a vector space if we allow signed measures. The space of probability measures, $\mathcal{S}(X)$, is a subset of the cone of positive measures. We will talk mainly about $\mathcal{M}(X)$, with normalisation understood. We are free to choose coordinates on $\mathcal{M}(X)$ as on any manifold. One choice is to describe measures as $n$-forms, in which case the space $\mathcal{S}(X)$ becomes the space of probability $n$-forms. Other choices are possible. We discuss another useful choice in section 4.3.1 below.

We are given a likelihood, $\tilde{\sigma}$. Perhaps the most general way to view a likelihood is as follows. Given a (measurable) map between two spaces, $f : X \to Y$, we can construct the pushforward map $f_* : \mathcal{M}(X) \to \mathcal{M}(Y)$. The action of the pushforward $f_*\mathbf{M}$ of a measure $\mathbf{M} \in \mathcal{M}(X)$ on a function $a$ on $Y$ is given by: $(f_*\mathbf{M})(a) = \mathbf{M}(f^*a) = \mathbf{M}(af)$. In the case that $X = Y \times Z$, and $f$ is the projection onto one of the factors, this is just marginalisation.

A likelihood is then a linear, positive, normalisation-preserving map $\tilde{\sigma} : \mathcal{M}(Y) \to \mathcal{M}(X)$ such that $f_*\tilde{\sigma} = \mathrm{id}_{\mathcal{M}(Y)}$[3].

In the case of Bayes' theorem, we have that $X = \Gamma \times X$, $Y = \Gamma$ and $f = \pi_\Gamma$, the canonical projection. The likelihood takes $\mathcal{M}(\Gamma)$ to $\mathcal{M}(\Gamma \times X)$. We thus have the following diagram:

$$\begin{tikzcd} \Gamma \arrow[r, "\delta"] & \mathcal{M}(\Gamma) \arrow[r, shift left, "\tilde{\sigma}"] & \mathcal{M}(\Gamma \times X) \arrow[l, shift left, "\pi_{\Gamma *}"] \arrow[r, shift right, "\pi_{X *}"'] & \mathcal{M}(X) \arrow[l, shift right, "\tilde{\sigma}_{\mathbf{q}}"'] & X \arrow[l, "\delta"'] \end{tikzcd} \tag{10}$$

where the $\pi$ are the canonical projections from $\Gamma \times X$ to its components and $\delta$ is the map that takes each point of its source to the delta function measure at that point in the target measure space.

[3]The traditional definition of conditional probability can be recovered from this definition by considering $Y$ to be the two-point space $\{0, 1\}$. Then any point of $X$ is mapped to one or the other point of $Y$, and a likelihood is the assignment to each point of $Y$ of a measure on $X$. For each point $b \in Y$, this measure must be concentrated on $f^{-1}(b)$. Given an initial measure on $X$, and its marginalisation to $Y$, there is then a unique likelihood that will take the marginalised measure back to the original measure on $X$. This is the conditional probability.

Bayes' theorem then functions as follows. Given a prior measure $\mathbf{q} \in \mathcal{M}(\Gamma)$, it is mapped to $\tilde{\sigma}(\mathbf{q}) \in \mathcal{M}(\Gamma \times X)$. There is then a unique linear, positive, normalisation-preserving map $\tilde{\sigma}_{\mathbf{q}}$ : $\mathcal{M}(X) \to \mathcal{M}(\Gamma \times X)$ that satisfies:

$$\begin{aligned} \pi_{X*}\tilde{\sigma}_{\mathbf{q}} &= \mathrm{id}_{\mathcal{M}(X)} \\ \tilde{\sigma}_{\mathbf{q}}\pi_{X*}\tilde{\sigma}(\mathbf{q}) &= \tilde{\sigma}(\mathbf{q}). \end{aligned} \tag{11}$$

Now, given a measurement, that is, a point $x \in X$, our knowledge about the measurement is described by the delta measure at $x$, $\delta_x$. We can thus form $\tilde{\sigma}_{\mathbf{q}}(\delta_x)$, a measure in $\mathcal{M}(\Gamma \times X)$. Now we can apply $\pi_{\Gamma*}$ to create a measure in $\mathcal{M}(\Gamma)$. This is the posterior measure. The abstract nature of this construction, using only the spaces of measures and linear maps, means that the construction of the posterior is independent of any underlying measures, and coordinate-invariant.

Note that by combining $\pi_{X*}$, $\tilde{\sigma}$ and $\delta$, we can form the map $\sigma = \pi_{X*}\tilde{\sigma}\delta$ from $\Gamma$ to $\mathcal{M}(X)$. This corresponds to the normal notion of likelihood as a parameterized probability measure on $X$.

## 4.2 An Invariance Criterion

Having set up this structure, we now present an argument for a particular choice of metric on $\Gamma$. The argument rests on one simple idea: that all information about the parameters not contained in the data be contained in the prior measure, or in other words, that all information that distinguishes one point of $\Gamma$ from another should come either from their correspondences with data probability measures (condition 1) or from the prior measure on $\Gamma$ (condition 2). It is the probability measures on $X$ alone that determine the relationship between the points in $\Gamma$ and the observations represented by points in $X$, and the way that these measures are parameterized serves to determine the meaning of the points in $\Gamma$ and not the other way around. Any other information in addition to the data we have at hand should be described by the prior. Any metric that we choose on $\Gamma$ should respect this principle, and not introduce any extra information about points in $\Gamma$.

The fact that it is not the identity of individual points in $\Gamma$ that is important, but merely their correspondence with data probability measures, means that it is only the image of $\Gamma$ in $\mathcal{M}(X)$ that counts. The conclusion from this line of argument is that any map from $\Gamma$ to $\mathcal{M}(X)$ that has the same image as $\sigma$ should produce the same inference results, and in particular the same MAP and MMSE estimates. We thus require that the results of inference be invariant to replacements of $\sigma$ by $\sigma\epsilon$, where $\epsilon : \Gamma \to \Gamma$ is a diffeomorphism. This diffeomorphism invariance, although superficially similar to a change of coordinates, is defined independently of any change of coordinates, and as such is a substantive restriction.

There are only two ways to achieve this aim. One is to pick a particular representative of the equivalence class of maps $\{\sigma\epsilon\}$ and to define a metric on the corresponding copy of $\Gamma$. This metric can then be pulled back to other members of the equivalence class using the maps $\epsilon$. Although this will satisfy condition 1, the selection of a particular member of the equivalence class to be endowed with a particular metric implies that we already know something about the points in $\Gamma$ independently of their correspondence with probability measures on $X$. Otherwise, how could we know to which points of $\Gamma$ to assign which values of the metric? This is exactly the type of information that should be included in the prior, and thus the procedure described in this paragraph violates condition 2.

The second approach is to pull back a metric from $\mathcal{M}(X)$ to each equivalent copy of $\Gamma$ using $\sigma\epsilon$. Such a metric automatically satisfies the consistency conditions introduced by the maps $\epsilon$ between members of the equivalence class: $\sigma\epsilon^*\mathbf{g} = \epsilon^*\sigma^*\mathbf{g}$, where $\mathbf{g}$ is a metric on $\mathcal{M}(X)$, and thus our results will depend solely on the image of $\Gamma$ in $\mathcal{M}(X)$. In addition, we were not required to pick a particular member of the class *a priori*, since each member of the equivalence class gets its own consistent metric induced by its own likelihood map. Thus both condition 1 and condition 2 are satisfied.

We are thus in a position to define a metric and underlying $m$-form on $\Gamma$ that satisfies the invariance criterion stated at the beginning of this section. We lack only one thing: a metric on $\mathcal{M}(X)$ to pull back.

## 4.3 Metrics on $\mathcal{M}(X)$

The first thing we must do is to define what we mean by the tangent space to $\mathcal{M}(X)$. Since we are using $n$-forms as coordinates on $\mathcal{M}(X)$, and since the space of signed measures is linear, it is easy to see that a tangent vector to $\mathcal{M}(X)$ can be identified with an $n$-form. If we restrict attention to $\mathcal{S}(X)$, this $n$-form must integrate to zero to preserve normalisation. Then, at a point $\mathbf{T} \in \mathcal{M}(X)$, an inner product between two tangent vectors $\mathbf{v}_1$ and $\mathbf{v}_2$ is given by

$$\mathbf{g}(\mathbf{v}_1, \mathbf{v}_2) = \int_X \mathbf{T}\frac{\mathbf{v}_1}{\mathbf{T}}\frac{\mathbf{v}_2}{\mathbf{T}}, \tag{12}$$

where we have identified the abstract tangent vectors $\mathbf{v}$ with their expression as $n$-forms. Note that the divisions are well-defined because $\mathbf{T}$ is positive. The justifications for this choice as the only reasonable metric on $\mathcal{M}(X)$ are many, and we do not re-iterate them here. Interested readers can consult, for example, [1].

### 4.3.1 Distances in $\mathcal{S}(X)$

As a brief aside, we discuss distances in $\mathcal{M}(X)$ and $\mathcal{S}(X)$ themselves. Given a metric on a manifold such as equation 12, it is in principle merely a question of applying known machinery to calculate the distance between any two points. In practice, this is often an analytically intractable procedure. It is thus remarkable that the metric given above on the space $\mathcal{M}(X)$ is in fact Euclidean in a suitable coordinate system, and that the calculation of the geodesic distance between two measures, or between two probability measures, is therefore trivial.

To see this, we change coordinates on the space $\mathcal{M}(X)$, from $n$-forms to 'half-densities'. These have the property that the product of two of them forms an $n$-form, which can then be integrated over $X$. Half-densities are thus the 'square roots' of $n$-forms. We can use the new coordinate system to define a new coordinate basis for the tangent space to $\mathcal{M}(X)$. In terms of this new coordinate system, the metric in equation 12 becomes (up to an irrelevant constant factor)

$$\mathbf{g}(\mathbf{v}_1, \mathbf{v}_2) = \int_X \mathbf{v}_1^{1/2}\mathbf{v}_2^{1/2}, \tag{13}$$

where the $\mathbf{v}^{1/2}$ are the components of the tangent vectors in the half-density coordinate system. Thus the metric is Euclidean in this coordinate system. This is even clearer if an underlying measure $\mu$ on $X$ is given. Then the half-densities can be expressed in terms of functions $\mathbf{v}^{1/2}(x)$, and the above equation becomes simply the $L^2$ inner product with respect to $\mu$. Note however that the definition does not depend on such a choice.

The simplicity of equation 13 means that calculating the distance between two points in $\mathcal{M}(X)$ or $\mathcal{S}(X)$ is very easy. We do not need to follow the prescription of section 3.2. Distance in $\mathcal{M}(X)$ is simply the Hellinger distance. Distance in $\mathcal{S}(X)$ is calculated as follows.

In terms of the half-density coordinates, $\mathcal{S}(X)$ is simply the submanifold of $\mathcal{M}(X)$ such that $\langle \mathbf{T}_1 | \mathbf{T}_1 \rangle = \int_X \mathbf{T}_1^{1/2} \mathbf{T}_1^{1/2} = 1$, or in other words the unit sphere. The distance between the points $\mathbf{P}_1$ and $\mathbf{P}_2$ in $\mathcal{S}(X)$ is then simply

$$d(\mathbf{P}_1, \mathbf{P}_2) = \cos^{-1} \langle \mathbf{P}_1 | \mathbf{P}_2 \rangle. \tag{14}$$

This is of course closely related to the Bhattacharyya distance, defined by $-\ln \langle \mathbf{P}_1 | \mathbf{P}_2 \rangle$. Note that the Hellinger distance is the length of the straight line in $\mathcal{M}(X)$ joining the two measures concerned. Thus for probability measures, it is equal to the length of the chord of the unit sphere joining the two measures. In contrast, the distance defined in equation 14 is the distance *in* the unit sphere: the length of the shortest arc joining the two probability measures.

## 4.4 Pullback to $\Gamma$

Using the embedding $\sigma$ of $\Gamma$ in $\mathcal{M}(X)$, we can pull the metric on $\mathcal{M}(X)$ back to $\Gamma$. The definition of the pullback of the metric acting on two tangent vectors $u$ and $v$ in $T_\gamma \Gamma$ is as before

$$\begin{aligned} \mathbf{h}_\sigma(u, v) &= (\sigma^* \mathbf{g})_\gamma(u, v) \\ &= \mathbf{g}_{\sigma(\gamma)}(\sigma_*(u), \sigma_*(v)), \end{aligned} \tag{15}$$

where $\sigma_* : T_\gamma \Gamma \to T_{\sigma(\gamma)} \mathcal{M}(X)$ is the tangent (derivative) map. This expression is coordinate-invariant. If we wish to know the matrix elements of $\mathbf{h}_\sigma = \sigma^* \mathbf{g}$ in the basis determined by a system of coordinates, $\frac{\partial}{\partial \theta^i}$, on $\Gamma$, we must evaluate $\mathbf{h}_\sigma$ on these basis elements. The result is

$$\mathbf{h}_{\sigma,\gamma} \left( \frac{\partial}{\partial \theta^i}(\gamma), \frac{\partial}{\partial \theta^j}(\gamma) \right) = \int_X \sigma_\theta \frac{1}{\sigma_\theta} \frac{\partial \sigma_\theta}{\partial \theta^i} \frac{1}{\sigma_\theta} \frac{\partial \sigma_\theta}{\partial \theta^j}, \tag{16}$$

where we denote by $\sigma_\theta$ the value of the likelihood at the point $\gamma$ with coordinates $\theta$. We thus find that the components of the induced metric form the Fisher information matrix. The coordinate-invariant measure on $\Gamma$ is then given by $\mathbf{U}_\sigma = |\mathbf{h}_\sigma|_\theta^{1/2} d^m \theta$.

Some discussion is in order here, but we postpone it until section 5.

## 4.5 MAP Estimates

MAP estimation is now simply a question of using equation 2 with $\mathbf{Q}$ equal to the posterior measure and $\mathbf{U}$ equal to $\mathbf{U}_\sigma$.

Note that the introduction of a prior probability apparently prevents the estimate from being invariant under replacement of $\sigma$ by $\sigma\epsilon$. The solution to this problem is the following. The prior probability is assigned to one member of the equivalence class $\{\sigma\epsilon\}$ based on knowledge of the parameters that is independent of current data. It can then be pushed forward to other copies of $\Gamma$ using $\epsilon^{-1}$. Note that this violates condition 2 as it should, but that it does not violate condition 1.

## 4.6 MMSD Estimates

In section 3.2, we defined a coordinate-invariant version of the mean squared error estimate, which we called the MMSD estimate. Having defined a metric on $\Gamma$ above, we can now use it to calculate distances in $\Gamma$, and hence to define the MMSD estimate. In general, this is a difficult task that is not tractable analytically, although approximations may be available. In simple examples however, one can compute the distance function $d(\gamma, \gamma')$ analytically. We give an example in the subsection below.

Note that the result obtained by this procedure is not necessarily the same as computing the distance in equation 14 between the measures $\sigma(\gamma)$ and $\sigma(\gamma')$, since the paths considered in the computation of $d$ are constrained to lie in the image of $\Gamma$. In fact, the approach based on loss functions that are distances on $\mathcal{S}(X)$ or $\mathcal{M}(X)$ rather than on $\Gamma$ is somewhat peculiar, in that it ignores the structure of the model altogether.

### 4.6.1 MMSD estimate of variance

As an example of the minimum squared distance estimate, consider the data space $X = \mathbb{R}^n$, corresponding to $n$ independent experiments, and a Gaussian family of product measures on $n$, for the sake of argument with zero mean. The parameter space $\Gamma$ is isomorphic to $\mathbb{R}^+$: we use coordinates $\lambda \in \mathbb{R}$ on this space, where $\lambda$ is the standard deviation. The likelihood $\sigma$ is then given by

$$\sigma_\lambda = \sigma(\tilde{\delta}_\lambda) = d^n x \, (2\pi\lambda^2)^{-n/2} \exp -\frac{(x,x)}{2\lambda^2}, \tag{17}$$

where $(\cdot, \cdot)$ denotes the Euclidean inner product on $\mathbb{R}^n$. Derivation of the Fisher information then shows that the inner product between tangent vectors $u$ and $v$ in $T_\gamma\Gamma$, where the point $\gamma$ has coordinate $\lambda$, is (up to a constant factor)

$$\mathbf{h}_\sigma(u, v) = \left(\frac{n}{\lambda}\right)^2 u^\lambda v^\lambda, \tag{18}$$

where the superscript $\lambda$ denotes the component with respect to the coordinate basis $\frac{\partial}{\partial\lambda}$. Thus the infinitesimal distance $ds$ between the points with coordinates $\lambda$ and $\lambda + d\lambda$ is given by

$$ds^2 = \left(\frac{n}{\lambda}\right)^2 d\lambda^2. \tag{19}$$

This is easily integrated to give the distance between two points with coordinates $\lambda_0$ and $\lambda_1$ (assume $\lambda_1 > \lambda_0$):

$$d(\lambda_0, \lambda_1) = n \ln(\frac{\lambda_1}{\lambda_0}). \tag{20}$$

The MMSD estimate of $\lambda$ is therefore given by considering the following mean loss under the posterior distribution $\mathbf{Q}$ for $\lambda$:

$$L(\lambda) = \int_0^\infty d\lambda' Q(\lambda') n^2 (\ln\lambda - \ln\lambda')^2. \tag{21}$$

Differentiation with respect to $\lambda$ then shows that the minimum squared distance estimate of $\lambda$, $\hat{\lambda}$, is given by

$$\hat{\lambda} = \exp\langle \ln\lambda\rangle_{\mathbf{Q}}, \tag{22}$$

where $\langle\cdot\rangle_{\mathbf{Q}}$ indicates expectation using the measure $\mathbf{Q}$. Note that $\langle\ln\lambda\rangle_{\mathbf{Q}} \neq \ln\langle\lambda\rangle_{\mathbf{Q}}$ in general and that therefore the estimate is not simply the mean of $\lambda$ as would have been obtained by assuming a Euclidean metric.

The mean of $\ln\lambda$ can be calculated in the case that the prior on $\lambda$ is taken to be Jeffreys' prior. It is given in terms of coordinates by

$$\langle\ln\lambda\rangle_{\mathbf{Q}} = \frac{1}{2}\left[\ln\left(\frac{1}{2}(x,x)\right) - \psi\left(\frac{1}{2}n\right)\right], \tag{23}$$

where $\psi$ is the function

$$\psi(z) = \frac{d}{dz}\ln\Gamma(z) \tag{24}$$

and $\Gamma$ is the Gamma function $\Gamma(z) = \int_0^\infty dt\, t^{z-1}e^{-t}$. Thus

$$\hat{\lambda} = \sqrt{\frac{(x,x)}{2}}e^{-\frac{1}{2}\psi(n/2)}. \tag{25}$$

To first order, $\psi(z) = \ln(z)$, so that the estimate becomes:

$$\begin{aligned}\hat{\lambda}_{\text{cl}} &= \sqrt{\frac{(x,x)}{2}}e^{-\frac{1}{2}\ln(n/2)} \\ &= \sqrt{\frac{(x,x)}{n}},\end{aligned} \tag{26}$$

the classical result. To the next order, $\psi(z) = \ln(z) - \frac{1}{2z}$. This introduces the first corrections to the classical result:

$$\hat{\lambda} = e^{\frac{1}{2n}}\hat{\lambda}_{\text{cl}}. \tag{27}$$

As the number of experiments approaches infinity, the correcting factor approaches unity and we recover the classical result. For small amounts of data however, there are corrections:

$$\hat{\lambda} \simeq \hat{\lambda}_{\text{cl}}(1 + \frac{1}{2n}). \tag{28}$$

### 4.6.2 General case in one dimension

The form of the above estimate is quite general in the one-dimensional case. Consider that we have derived the metric on $\Gamma$, $\mathbf{h}$. The distance between two points $\gamma_0$ and $\gamma_1$ is then given according to the general discussion in section 3. In a general coordinate system, $\theta$, this can be written

$$\begin{aligned} d(\gamma, \gamma') &= \int_{t_0}^{t_1} dt \left( h(\pi(t)) \left( \frac{d\pi}{dt}(t) \right)^2 \right)^{1/2} \\ &= \int_{\theta_0}^{\theta_1} d\theta\, h^{1/2}(\theta), \end{aligned} \tag{29}$$

where $\pi(t_{0,1}) = \gamma_{0,1}$, $\theta_{0,1} = \theta(\gamma_{0,1})$ and $h$ is the (single) component of the metric $\mathbf{h}$ in the $\theta$ coordinate system. Note that there is no need for a minimization in one dimension. All paths with the same endpoints belong to the same equivalence class under the action of (boundary- and orientation-preserving) diffeomorphisms of $I$. Now let $F(\theta)$ be the inverse derivative of $h^{1/2}$. The (signed) distance between the two points is now $d(\theta_1, \theta_0) = F(\theta_1) - F(\theta_0)$. Including this in equation 8, differentiating $L$ and equating to zero then gives the result that

$$F(\hat{\theta}) = \langle F \rangle_{\mathbf{Q}}, \tag{30}$$

and thus that

$$\hat{\theta} = F^{-1} \langle F \rangle_{\mathbf{Q}}. \tag{31}$$

In more than one dimension of course the problem is a great deal more complicated, since there is an infinity of equivalence classes, and the minimization means solving a partial differential equation for the geodesics.

# 5 Discussion and Related Work

There is a significant amount of work on the geometry of probability measure spaces from the point of view of classical statistics. [10] first introduced the pulled back metric defined in section 4.4. The work of [1] brought these ideas to prominence, while [7] and [5] provide more recent treatments. For the most part, this work has focused on asymptotics and other issues of importance to classical statistics, while the Bayesian approach using prior and posterior probabilities and loss functions has largely been ignored. For example, many of the estimation problems considered are connected to the notion of a "true" distribution and the fact that it might not lie in the image of $\Gamma$. A great deal of extra structure is introduced in this approach (non-metric connections for example), which seems from a Bayesian point of view to be unnecessary and rather inelegant, while the simpler structures considered here are ignored. For example, [7] say that the Riemannian distance is not of statistical significance, and that the mean in a manifold cannot be calculated; all that is possible is an analysis of the way in which the value of the mean, calculated in coordinates, changes with the coordinates. As we have seen however, the Riemannian metric precisely allows the definition of a natural, coordinate-invariant generalisation of the mean. [9] develops some basic statistical tools for Riemannian manifolds, and applies these ideas in various ways to problems in computer vision. The approach is not Bayesian however, and in particular the choice of a metric and the relation with estimation problems, including the use of the metric measure as an underlying measure for MAP estimation, are not considered.

MML inference was developed in [11], [12] and [13]. A discussion of its relationship with the standard Bayesian approach and of its invariance properties can be found in the above papers and in [8]. The literature on MML inference frequently cites the invariance of MML estimates as one reason to prefer them to MAP estimates. The above analysis shows that this is not a special property of MML estimates, or a deep problem with MAP estimates. Indeed, the issue is not one of MAP estimation *per se*. Lack of invariance is a consequence of not describing the quantities of interest in $\Gamma$ in a coordinate-invariant, and hence meaningful, way. To do this, one must recognise that a metric is lurking in the definition of both MAP and MMSE estimates, and indeed in any useful discussion of $\Gamma$, and that making it explicit is a necessary condition for meaningful definitions in arbitrary coordinate systems. Once done, the definition of coordinate-invariant estimates is an immediate consequence of the geometry. Although equation 2 with the pulled-back metric as underlying measure is formally the same as that for MML estimates, unlike MML methods, no discretization of $\Gamma$ is needed, and no approximations are made. In fact, the above derivation throws light on the procedure used in deriving MML estimates, which from this point of view appears to be a roundabout way of defining an underlying measure by first discretizing the manifold and then considering the volume of each cell.

The fact that we are discussing the geometry of $\Gamma$ and not a particular form of estimate means that the analysis presented here is more general than MML however. By recognising the necessity of an explicit metric on $\Gamma$ for inference, the way is open for the definition of coordinate-invariant loss functions of many different types. Here we have given the example of a coordinate-invariant MMSE estimate, the MMSD estimate, but whenever defining a loss function on a parameter space, the issues described here must be taken into account.

## 5.1 Discussion of choice of metric

In section 4, we came to the conclusion that the only choice of metric that satisfied the two conditions mentioned at the beginning of that section, was the metric induced by pullback from $\mathcal{M}(X)$. To recap: the metric and its associated underlying measure should not introduce information about $\Gamma$. Such information should be contained in one of two sources: the correspondence between points of $\Gamma$ and points of $\mathcal{M}(X)$, and the prior measure. The first leads to the idea that the metric on diffeomorphically related copies of $\Gamma$ should be related by pullback, while the second eliminates the possibility of choosing a metric on one fixed copy of $\Gamma$ and then pulling it back to the other copies, since this implies that we must be able to assign a value of the metric to particular points in $\Gamma$ *a priori*, which in turn implies that we must know something about the "identity" of these points beyond the information contained in the prior. Hence the result given.

Note that this argument is somewhat different to that normally used for Jeffreys' prior, or rather is a clarification and a refinement of that argument, which essentially boils down to proving that this prior is invariant under "reparametrizations". First, the emphasis is on the metric as providing $\Gamma$ with geometry, and not on the measure, which is a derived quantity. Second, coordinate invariance is not an issue: the abstract way in which the geometry is described does not rely on a particular choice of coordinate system. Equation 2, for example, is coordinate-invariant for any choice of metric. Instead the emphasis is on diffeomorphism invariance: our results should not depend on which copy of $\Gamma$ we use, since this merely "shuffles" the points of $\Gamma$ without changing their correspondence with points of $\mathcal{M}(X)$.

The use of the underlying measure of the pulled-back metric does not force us to use Jeffreys' prior as an uninformative prior; in fact, we may assume that all of our priors are proper, and still use the metric and its measure in estimation problems. Thus the large amount of previous work (for example, [2]—see [6] for an extensive discussion and bibliography) on the choice of priors is not directly relevant to our discussion here.

Note also that some of the problems associated with Jeffreys' prior do not appear when we are talking about an underlying measure. Normalization is not necessary since the underlying measure is not a probability distribution. Second, the procedure advocated here suggests that we should first eliminate nuisance parameters using whatever prior information we possess, to obtain a likelihood on the parameter of interest, and only then derive the metric by pullback. Thus the various paradoxes associated with the non-commutativity of the derivation of Jeffreys' prior and marginalization do not arise.

As has been noted in the literature on MML estimates, the appearance of the density term in the denominator of equation 2 means that if we do use Jeffreys' prior as the prior measure, then MAP estimation reduces to maximum likelihood. Since it is well-known that maximum likelihood estimates are coordinate-invariant, we see concrete confirmation of the invariance of equation 2 in this special case. The conflict between classical and Bayesian approaches to parameter estimation is thus to some extent resolved.

Our argument for the metric and underlying measure on $\Gamma$ does not depend on group-theoretic considerations. Nevertheless, the metric is compatible with these considerations, as is Jeffreys prior, because of the following simple argument. Let $X$ be a manifold with metric $\mathbf{h}$, and $Y$ be embedded in $X$ by $f$. Suppose we have two group actions $\beta_X : G \times f(Y) \to f(Y)$ and $\beta_Y : G \times Y \to Y$. Note that the group action on $X$ need only be defined for the image of $Y$: it may for example be induced by the group action on $Y$ itself. If we have:

$$\begin{tikzcd} Y \arrow[r, "f"] & f(Y) \\ Y \arrow[u, "\beta_Y(g)"] \arrow[r, "f"'] & f(Y) \arrow[u, "\beta_X(g)"'] \end{tikzcd} \tag{32}$$

then, if $G$ acts by isometries on $X$, endowing $Y$ with the metric $f^*\mathbf{h}$ ensures that $f$ is an isometry also. Therefore, $G$ must act by isometries on $Y$. If $Y$ is $G$ itself, this ensures that the underlying measure induced by the metric $f^*\mathbf{h}$ is a Haar measure.

Finally, an information-theoretic intuition is interesting. In computing the MAP estimate, it is equivalent to maximize the logarithm of equation 3. Naturally the logarithm consists of the difference of two terms: the logarithm of the posterior density and the logarithm of the underlying density. The role of the underlying density is the following. The information that we possess should presumably be that amount of information that we possess beyond 'ignorance'. If our expression for 'ignorance' does not possess the value 'zero' (i.e. the identity) in the algebra in which we add and subtract information, then the information that we possess beyond 'ignorance' should be the difference between the algebraic element representing our knowledge, and the algebraic element representing 'ignorance'. In view of the "uninformative" nature of the underlying measure that we are using, the MAP estimate can thus consistently be thought of as finding that point in $\Gamma$ with maximum information.

# A Forms

We provide a short introduction to the language of forms. A good reference is [3]. Briefly, differential forms are antisymmetric, multilinear functionals on products of vector spaces. For manifolds they are defined pointwise on the tangent space at each point and then required to satisfy smoothness properties. They also allow a beautiful theory of integration on manifolds, and in this capacity they are thought of as *co-chains*, linear functionals on the vector space of chains in a manifold. Their advantages are great concision and uniformity of notation; independence of basis or coordinates; manifest invariance to diffeomorphisms and other transformations; and generality. In bringing together integration and geometry in one notation, they are ideal for our discussion.

We are given a manifold $\Gamma$. From here, we can define the tangent space at each point, $T_\gamma\Gamma$ using a number of approaches. The result is intuitively clear however, so we will not go into detail. We can bring all the tangent spaces together in the 'tangent bundle', $T\Gamma$. This is another manifold, each point of which can be thought of as a pair: a point $\gamma$ in $\Gamma$ and a vector in $T_\gamma\Gamma$. There is a canonical projection from $T\Gamma$ to $\Gamma$ supplied by forgetting the tangent vector. At each point $\gamma$, the tangent space $T_\gamma\Gamma$ has a dual space, $T^*_\gamma\Gamma$, the space of linear maps from $T_\gamma\Gamma$ to $\mathbb{R}$. These can be combined to form the co-tangent bundle, $T^*\Gamma$. A 'vector field' is a 'section' of the tangent bundle: a map from $\Gamma$ to $T\Gamma$ whose left inverse is the canonical projection.

We can also form product bundles, in which the 'extra space' at each point $\gamma$ is the product of copies of the tangent space: thus each point in $T^p\Gamma$ can be thought of as a point $\gamma$ and an element of $\otimes^p T_\gamma\Gamma$. Now at each point, we can define higher dual spaces: $T^{*p}_\gamma\Gamma = \otimes^p T^*_\gamma\Gamma$ is the space of multilinear functions on $\times^p T_\gamma\Gamma$. In particular, we can restrict attention to the antisymmetric linear functions: those that change sign under the interchange of any two arguments. These are antisymmetric direct products of the co-tangent space, denoted $\wedge^p T^*_\gamma\Gamma$. Their combination into a bundle is denoted $\wedge^p T^*\Gamma$. A section of $\wedge^p T^*\Gamma$ defines, for each point $\gamma$, an element of $\wedge^p T^*_\gamma\Gamma$. Sections of $\wedge^p T^*\Gamma$ are known as 'forms', and $p$ is the 'degree' of the form. We denote the space of $p$-forms $\Lambda^p\Gamma$. Forms of degree $p$ and $q$ can be multiplied to give forms of degree $p + q$. Because the product of co-tangent spaces is antisymmetric, all forms of degree higher than $m$, the dimensionality of the manifold, are zero. $0$-forms are functions on $\Gamma$.

In order to express vectors and forms more easily, it is convenient to introduce bases for the various spaces. This is easily done using a coordinate system $\theta : \Gamma \to \mathbb{R}^m$. A basis for $T_\gamma\Gamma$ is then the set of $\frac{\partial}{\partial\theta^j}(\gamma)$. A dual basis for $T^*_\gamma\Gamma$ is then the set of $d\theta^i(\gamma)$, which acts on the basis of $T_\gamma\Gamma$ as

$$d\theta^i(\gamma)(\frac{\partial}{\partial\theta^j}(\gamma)) = \delta_{ij}. \tag{33}$$

Taking the collection of these bases all over $\Gamma$, we have bases for the spaces of vector fields and 1-forms. Now we can form bases for the various power bundles. For example, a basis for the space of 2-forms is given by the set $d\theta^i(\gamma) \wedge d\theta^j(\gamma)$, where $\wedge$ denotes the antisymmetric product. We will denote the basis element $d\theta^1(\gamma) \wedge \ldots \wedge d\theta^m(\gamma)$ of the space of $m$-forms (there is only one - if the indices are not different, antisymmetry of the product means the result is zero) by $d^m\theta(\gamma)$. The sign of this basis element (or in other words, the order of the factors of $d\theta^i$ that it contains) defines an 'orientation' on the manifold, in the sense that a basis for the tangent spaces, when acted upon by the form, will give either a positive or negative result depending on its orientation in the traditional sense of right- and left-handed coordinate systems. Given an orientation in this sense, a basis for the tangent spaces is either 'oriented' or not. Not all manifolds admit a global orientation. We consider only orientable manifolds.

Given another manifold $Y$, and a map $\sigma : Y \to \Gamma$, we define the 'tangent map' or 'derivative map' at a point $y \in Y$, $\sigma_* : T_yY \to T_{\sigma(y)}\Gamma$ as follows. A point $(y, u) \in TY$ is taken to $(\sigma(y), \sigma_* u) \in T\Gamma$, where, in terms of coordinates $\theta^i$ on $\Gamma$ and $\phi^\alpha$ on $Y$, in which $u = u^\alpha \frac{\partial}{\partial\phi^\alpha}$, we have

$$\sigma_* u \quad = \quad (\sigma_* u)^i \frac{\partial}{\partial\theta^i} = u^\alpha \frac{\partial\sigma^i}{\partial\phi^\alpha} \frac{\partial}{\partial\theta^i}, \tag{34}$$

where $\sigma^i = \theta^i(\sigma)$. We also introduce the convention that repeated indices, one up, one down, are summed over.

Using this map, we can define the 'pullback' $\sigma^*\mathbf{A}$ of a form $\mathbf{A} \in \Lambda^p\Gamma$ (or in fact of any member of a power of a co-tangent space, whether antisymmetric or not) as

$$\sigma^*\mathbf{A}_y(u, v, \ldots) \quad = \quad \mathbf{A}_{\sigma(y)}(\sigma_* u, \sigma_* v, \ldots). \tag{35}$$

Thus the action of a pulled back form on tangent vectors is defined by the action of the original form on the tangent vectors pushed forward by the tangent map.

As well as antisymmetric products of co-tangent spaces, we can form symmetric products. If at each point $\gamma$, we form the space of symmetric, bilinear functions on $T_\gamma\Gamma \times T_\gamma\Gamma$, which we will denote $T^*_\gamma\Gamma \vee T^*_\gamma\Gamma$, we can again form a product bundle $T^*\Gamma \vee T^*\Gamma$. A 'metric' $\mathbf{h}$ on $\Gamma$ is a positive (semi-)definite section of this bundle: to each point $\gamma$, it assigns a positive (semi-)definite element of $T^*_\gamma\Gamma \vee T^*_\gamma\Gamma$, or in other words, an inner product on $T_\gamma\Gamma$.

In a particular coordinate basis $\frac{\partial}{\partial\theta^i}(\gamma)$, the metric has components, given by

$$h_{\gamma,ij} = \mathbf{h}_\gamma(\frac{\partial}{\partial\theta^i}(\gamma), \frac{\partial}{\partial\theta^j}(\gamma)). \tag{36}$$

The matrix elements of the metric at each point $\gamma$ possess a determinant, which we will write $|\mathbf{h}|_\theta(\theta(\gamma))$.

Using the metric $\mathbf{h}$, we can define a canonical isomorphism, the 'Hodge star' $\star_{\mathbf{h}}$, between $\Lambda^p\Gamma$ and $\Lambda^{m-p}\Gamma$. We show here its action for $p = 0$ and $p = m$ only, since that is all we will need. We choose coordinates $\theta^i$ (nothing will depend on this choice). Let $f$ be a 0-form, and $\mathbf{A} = Ad^m\theta$ be an $m$-form ($A$ is a function—the component of $\mathbf{A}$ in the basis $d^m\theta$). Then we have

$$\star_{\mathbf{h}} f = |\mathbf{h}|^{1/2} f d^m\theta \tag{37}$$

$$\star_{\mathbf{h}}\mathbf{A} = |\mathbf{h}|^{-1/2} A, \tag{38}$$

where we have suppressed arguments and reference to the coordinate system in the definition of the determinant for clarity.

The Hodge star can be used to define an inner product on each $\Lambda^p\Gamma$. Since $\star_{\mathbf{h}}\mathbf{A}$ is an $(m-p)$-form if $\mathbf{A}$ is a $p$-form, the quantity $\mathbf{A} \star_{\mathbf{h}} \mathbf{B}$ for two $p$-forms is a $m$-form, and can be integrated on $\Gamma$:

$$\langle \mathbf{A}, \mathbf{B} \rangle = \int_\Gamma \mathbf{A} \star_{\mathbf{h}} \mathbf{B}. \tag{39}$$

We can define 'positive' $m$-forms as those whose action on oriented bases produces a positive result. It is equivalent to say that their dual under the action of the Hodge star is a positive function. A 'probability $m$-form' is a positive $m$-form whose integral over $\Gamma$ is equal to 1. We can divide $m$-forms by positive $m$-forms. For a $m$-form $\mathbf{A}$ and a positive $m$-form $\mathbf{B}$, the value of $\frac{\mathbf{A}}{\mathbf{B}}$ is that unique function $f$ such that $\mathbf{A} = f\mathbf{B}$. This division is the analogue of the Radon-Nikodym derivative for forms.

On an $m$-dimensional manifold, $m$-forms can be integrated in the way that the notation suggests. For a $m$-form $\mathbf{A} = Ad^m\theta$, we have that

$$\int_{\Omega\subset\Gamma} \mathbf{A} = \int_{\theta(\Omega)} A(\theta) d^m\theta, \tag{40}$$

where we have used the same symbol $A$ for the function and its expression in terms of coordinates.

To integrate a $p$-form $\mathbf{A}$ over a $p$-dimensional submanifold embedded in $\Gamma$, $Y \overset{\sigma}{\rightarrowtail} \Gamma$, one first pulls the form back to the embedded manifold and then integrates:

$$\int_{\sigma(Y)} \mathbf{A} = \int_Y \sigma^*\mathbf{A}. \tag{41}$$

These definitions highlight the second way of interpreting forms: as 'co-chains'. A $p$-chain in $\Gamma$ is (roughly speaking) a linear combination of $p$-dimensional rectangles embedded in the manifold. The space of linear functions on the space of $p$-chains (the co-chains) can be identified with $\Lambda^p\Gamma$.

We will have cause to integrate a function $f$ over a $p$-dimensional submanifold $Y \overset{\sigma}{\rightarrowtail} \Gamma$ of $\Gamma$. This is slightly different from the case of integrating a $p$-form. One first pulls the function back to $Y$ and then uses a metric on $Y$ to convert the function into a $p$-form that can be integrated over $Y$:

$$\int_{\sigma(Y)} f = \int_Y \star_{\mathbf{h}}\sigma^* f, \tag{42}$$

where by definition $(\sigma^* f)(y) = f(\sigma(y))$, and $\mathbf{h}$ is a metric on $Y$.

However, since we are interested in the submanifold in $\Gamma$ and not $Y$ itself, we are really considering an equivalence class of embeddings $\{f\epsilon\}$, where $\epsilon : Y \to Y$ is a diffeomorphism, with the same image. The result of our integration should be independent of the representative in this equivalence class, and this means that the metric on $Y$ must vary with the representative. If no representative is distinguished, the only way to achieve this invariance is to pull back a metric $\mathbf{g}$ on $\Gamma$ to $Y$, and use this metric to define the Hodge star:

$$\int_{\sigma(Y)} f \quad = \quad \int_Y \star_{\sigma^*\mathbf{g}} \sigma^* f. \tag{43}$$

# References


[1] S. Amari, *Differential-geometrical methods in statistics*, Lecture Notes in Statistics, vol. 28, Springer-Verlag, Berlin, 1985.

[2] J. M. Bernardo, *Reference posterior distributions for bayesian inference*, J. Roy. Statist Soc. Ser. B **41** (1979), 113–147.

[3] Y. Choquet-Bruhat, C. DeWitt-Morette, and M. Dillard-Bleick, *Analysis, manifolds and physics*, Elsevier Science, Amsterdam, The Netherlands, 1996.

[4] H. Karcher, *Riemannian center of mass and mollifier smoothing*, Comm. Pure Appl. Math. **30** (1977), 509–541.

[5] R. E. Kass and P. W. Vos, *Geometrical foundations of asymptotic inference*, John Wiley & Sons, 1997.

[6] R. E. Kass and L. Wasserman, *Formal rules for selecting prior distributions: A review and annotated bibliography*, J. Amer. Statist. Assoc. **91** (1996), 343–1370.

[7] M. K. Murray and J. W. Rice, *Differential geometry and statistics*, Monographs on Statistics and Applied Probability, vol. 48, Chapman and Hall, London, 1993.

[8] J. J. Oliver and R. A. Baxter, *MML and bayesianism: Similarities and differences*, Tech. Report 206, Department of Computer Science, Monash University, Clayton, Vic. 3168 Australia, 1995.

[9] X. Pennec, *Probabilities and statistics on riemannian manifolds: Basic tools for geometric measurements*, Proc. Nonlinear Signal and Image Processing (NSIP'99) (Antalya, Turkey) (A.E. Cetin, L. Akarun, A. Ertuzun, M.N. Gurcan, and Y. Yardimci, eds.), vol. 1, IEEE-EURASIP, June 1999, pp. 194–198.

[10] C. R. Rao, *Information and accuracy attainable in the estimation of statistical parameters*, Bull. Calcutta Math. Soc. **37** (1945), 81–91.

[11] C. S. Wallace and D. M. Boulton, *An information measure for classification*, Comp. J. **11** (1968), 185–195.

[12] ———, *An invariant bayes method for point estimation*, Classification Soc. Bull. **3** (1975), 11–34.

[13] C. S. Wallace and P. R. Freeman, *Estimation and inference by compact coding*, J. R. Statist. Soc. B **49** (1987), no. 3, 240–265.

Unité de recherche INRIA Sophia Antipolis
2004, route des Lucioles - BP 93 - 06902 Sophia Antipolis Cedex (France)

Unité de recherche INRIA Lorraine : LORIA, Technopôle de Nancy-Brabois - Campus scientifique
615, rue du Jardin Botanique - BP 101 - 54602 Villers-lès-Nancy Cedex (France)
Unité de recherche INRIA Rennes : IRISA, Campus universitaire de Beaulieu - 35042 Rennes Cedex (France)
Unité de recherche INRIA Rhône-Alpes : 655, avenue de l'Europe - 38330 Montbonnot-St-Martin (France)
Unité de recherche INRIA Rocquencourt : Domaine de Voluceau - Rocquencourt - BP 105 - 78153 Le Chesnay Cedex (France)

Éditeur
INRIA - Domaine de Voluceau - Rocquencourt, BP 105 - 78153 Le Chesnay Cedex (France)
http://www.inria.fr
ISSN 0249-6399