Towards a general theory of statistical inference

Chris Goddard
January 18, 2015

Abstract

It is a common problem in statistics to determine the appropriate heuristic to select from a set of hypotheses (or equivalently, models), prior to optimising that model to fit the data. In this short note I sketch a technique based on the construction of an information in order to compute the optimal model within a given model space and given data.

We assume from the outset the following: we are dealing with the class of linear models. That is, if our data can be massaged into some form of continuous and/or discrete variables x_i, $i = 1, ..., N$ for some finite N, then we define a model P to be an N by N matrix P_{ij} such that the response Y_j to an input X_i will be given by

$$Y_j = P(d)_{ji} X_i$$

where d_{ki}, $k = 1, ..., M$ are the set of M data points that we already know over the set of variables x_i which are used to optimise P by linear regression.

So that seems relatively straightforward. But there is a problem here, or rather two problems:

- What is the optimal choice for N?, and
- We are implicitly assuming that the space of data M is Euclidean. What if it is not? For instance, what if it is endowed with some general Riemannian metric σ? Then what is the optimal choice for σ?

The first question is a standard problem, which is the problem of *overfit*. If we use too many variables to descibe our data set D, then the predictive power of our model P will be hamstrung; in particular, the modelled response Y to a new input X will tend to have more error than if we are not overzealous with fitting our model to the existing data.

In a way, both of these problems can be dealt with, if we allow σ to be a degenerate bilinear form, or a Riemann-Cartan metric over some general infinite dimensional idealised variable space, M. Then presumably if we can compute σ,

there will be some set of spanning geodesic eigenbasis for the vector space at each point, or a set of normal coordinates for M. From this we should then theoretically be able to back-deduce a choice of natural coordinates to describe the data.

So then, let us instead merely consider a metric σ on some idealised space M. σ_{ij} represents the failure of our space of coordinates M to be iid, ie, σ_{ij} is to represent the correlation between directions i and j at a point in M.

Consider now the space of degenerate infinite matrices, ie $GL(\infty)$. Call this our model space, A.

Define a statistical distribution function f that takes a point in data space and assigns a distribution of models to that point, such that the integral is 1 (ie, it is a pdf), and also such that it is continuous and differentiable. In other words, let $f(m, a)$ where $m \in M$ and $a \in A$ be this function st

$$\int_A f(m, a)da = 1$$

for each and every m in M.

Then, define a *section* of this space to be a choice of correlation metric σ with $f(m, a) = \delta(\sigma(m) - a)$, where σ is a Riemann-Cartan metric and δ is the Dirac delta function.

Now define an information

$$J(f) = \int_A \int_M f(m, a)(ln\partial f(m, a))^2 dmda$$

Theorem 0.1. *This information, which I will call by abuse of terminology the* Akaike information*, satisfies the Cramer-Rao equality, that is,* $J(f) \geq 0$.

Proof. (sketch). ln, the inverse of the exponential mapping from $TM \to M$, can be viewed as a 1-category derivative ∂_1, much as ∂ can be viewed as a 0-category derivative ∂_0. Hence, the Fisher information:

$$I(f) = \int_A \int_M f(m, a)(\partial_0\partial_1 f(m, a))^2 dmda$$

is almost exactly analogous to

$$J(f) = \int_A \int_M f(m, a)(\partial_1\partial_0 f(m, a))^2 dmda$$

related by a Z_2 symmetry. This is not a coincidence, and happens to be because these information functionals form separate information theories under the umbrella of first order cybernetics (which can also be viewed equivalently as a 3-tensor construction for its simplest level of abstraction), but that is getting perhaps a bit further afield than absolutely necessary in this instance.

The basic point here is that the proof of the Cramer-Rao inequality for the Fisher information easily extends to a proof for the Akaike information, as the flipping of zeroth and first cat derivatives does not alter or change the flow of the arguments therein.

□

Moreover, we have another observation, that in many respects that this functional which I refer to as the Akaike information is a primitive for the invariants used in the Akaike information criterion (hence the name). Central to the AIC is the measure $ln(L)$ where L is a maximised value of the likelihood function, λ - hence, intuitively, we expect that since L is 'vaguely' obtained from $\partial\lambda = 0$, that $ln\partial\lambda$ is a more 'primitive' measure, and, in order to make things dimensionally make sense - as well as in terms of choice of which of the two information measures for first order cybernetics - that $\lambda(ln\partial\lambda)^2$ is a natural integrand / density for a 'deeper' AIC-like understanding of model selection.

To make this slightly more concrete, one would need to indicate how the Kullback-Leibler information which was used as a heuristic for the choice of the AIC in the first place corresponds to a '0-cat' view of model selection, and show how this naturally extends to the '1-cat' view of model selection embodied by the choice of functional above. Alternatively, one could think of things in the following way: while the Fisher information allows one to measure how quickly things are changing within a model in a local manner, and thereby obtain some understanding of the dynamics of a particular system, through reversal of the order of derivatives one considers more how things are changing as one moves through the space of models locally around a particular selected model in the case of the Akaike information.

Since the Akaike information satisfies the Cramer-Rao inequality, this leads us to a compelling conclusion:

By optimising the Akaike information, we are optimising for the optimal choice of model, given our assumptions about the nature of the space of models we are dealing with in the first place. I think that this is quite important, and if you, the reader, take little else away from this piece, this would not be a bad snippet to bear in mind, as it is the key observation that I

wished to communicate in this short article. Note, of course, that this observation is contingent on the choice of the space of models in question. I have been assuming $GL(\infty)$ endowed with a Riemann-Cartan metric of first order correlations (mild departure from assuming that the variables are iid). Then there naturally arises the further question as to how to choose from a set of sets of models, the optimal set of models in which to apply the above criterion. Naturally, there should be some tool or natural way to think about this, maybe in a cybernetic fashion. However, there comes a point where one receives diminishing returns (dependent, of course, on the quality and scope of one's data), and such considerations are beyond the scope of this document.

There is one final result that might be useful for computation:

It can be demonstrated that the Akaike information functional for a section $f(m, a) = \delta(\sigma(m) - a)$ reduces to

$$J(\sigma) = \int_M R_{\hat{\sigma}}(m) dm$$

where $\hat{\sigma}$ is the Riemann-Cartan metric dual to σ over M. I will not provide a proof of this statement here, but the details translate roughly from a similar result that holds for the Fisher information for a section in a similar situation.

This has the simple consequence that to optimise σ, we should solve the equation $R_{\hat{\sigma}}(m) = 0$ for σ, using the data d to calibrate the process. This should allow us to determine which variables are important, find some appropriate normal basis, and then perform a standard linear regression to fit σ to d. Then, to determine our response Y to an input X, we would calculate $Y = \sigma X$.

In many respects, this general approach could be viewed as a generalisation of other approaches, such as, in the case of linear models, an extension of the technique of ridge regression / Tikhonov regression, or the use of penalty functions to approximate or forecast data.