# Chapter 13
# Some Illustrations of Information Geometry in Biology and Physics

**C. T. J. Dodson**
*University of Manchester, UK*

## ABSTRACT

*Many real processes have stochastic features which seem to be representable in some intuitive sense as `close to Poisson', `nearly random', `nearly uniform' or with binary variables `nearly independent'. Each of those particular reference states, defined by an equation, is unstable in the formal sense, but it is passed through or hovered about by the observed process. Information geometry gives precise meaning for nearness and neighbourhood in a state space of processes, naturally quantifying proximity of a process to a particular state via an information theoretic metric structure on smoothly parametrized families of probability density functions. We illustrate some aspects of the methodology through case studies: inhomogeneous statistical evolutionary rate processes for epidemics, amino acid spacings along protein chains, constrained disordering of crystals, distinguishing nearby signal distributions and testing pseudorandom number generators.*

## INTRODUCTION

A question: *"We already use statistical modeling, why should we bother with information geometry? "*

Information geometry is concerned with the natural geometrization of smoothly parametrized families of discrete probability or continuous probability density functions; the naturality stems from the fact that the metric structure arises from the covariance matrix of gradients of probability. This metric yields a smooth Riemannian structure on the space of parameters, so adding the geometric concepts of curvature and arc length to the analytic tools for studying trajectories through probability distributions as statistical models evolve with time or during changes of system conditions. The development of the subject over the past 65 years has been substantially due to the work of C.R. Rao and S-I. Amari and coworkers; see for

DOI: 10.4018/978-1-61350-116-0.ch013

example (Rao, 1945; Amari, 1963; Amari, 1968; Amari, 1985; Amari et al, 1987; Amari & Nagaoka, 2000) and references therein. Information geometry and its applications remain vigorous research areas, as witness for example the series of international conferences of the same name (IGA Conference, 2010). In phenomenological modeling applications, information geometric methods complement the standard statistical tools with techniques of representation similar to those used in physical field theories where the analysis of curved geometrical spaces have contributed to the understanding of phenomena and development of predictive models.

In many statistical models of practical importance there is a small range of probability density functions that has very wide application as a result of general theorems, and the spaces of these families have just a small number of dimensions. For example, the families of Gaussian and gamma distributions and their bivariate versions are widely applied and moreover their information geometry is easily tractable, (Arwini & Dodson, 2008). In particular, the family of gamma distributions is ubiquitous in modeling natural processes that involve scatter of a positive random variable around a target state, such as for inhomogeneous populations or features of elements in a collection. The reason for this ubiquity is that a defining characteristic of the gamma distribution is for the sample standard deviation to be proportional to the sample mean. In practice, that property is commonly found to varying degrees of approximation; the case when the standard deviation *equals* the mean corresponds to the exponential distribution associated with a Poisson process, which is the fundamental reference process for statistical models. Sums of independent gamma random variables (hence also sums of independent exponential random variables) follow a gamma distribution and products of gamma random variables have distributions closely approximated by gamma distributions. We shall provide below more details about the properties of the gamma family and its associated families which include the uniform distribution, approximations to truncated Gaussians and a wide range of others.

Our case studies will show that gamma distributions model well the spacings between successive occurrences of each of the 20 different amino acids which with differing abundances lie along a protein chain (Cai et al, 2002). Figure 6illustrates the information distance in the space of gamma distributions for amino acid spacings along protein chains, measured from the exponential (Poisson) case $\kappa = 1$; intuitively we might expect that they would be scattered around the reference exponential case. In fact they all lie on the clustered side of the distribution, all have more variance than that expected by chance—the exponential case.

In typical real situations it is of interest to depict the changing state of a statistical model through the trajectory its representative distribution follows in the appropriate family of distributions, under the influence of external influences or some internal evolutionary imperative. An example is shown in Figure 1 for the integral curves starting from different initial gamma distributions of the entropy (ie the 'mean log probability density') gradient in the space of gamma distributions; there the unconstrained disordering means that the asymptote coincides with the exponential distribution, when the standard deviation equals the mean and we have maximal entropy (and hence maximal disorder). Information geometry provides the correct measures of 'information distance' along or between such trajectories, and along any other arbitrary curves, and it defines parallelism and perpendicularity as well as minimal distance curves (geodesics). Sometimes the degeneration of order is constrained by conditions in the model and then the process does not tend to the maximal entropy but only to a lower level. An example of a class of stochastic phenomena that involves the degeneration of an ensemble from more orderly to less orderly, through an external application of disruptive statistical influences, is the heating of a crystalline structure.