Information Geometric Similarity Measurement for Near-Random Stochastic Processes

Christopher T. J. Dodson and Jacob Scharcanski, *Senior Member, IEEE*

***Abstract*—We outline the information-theoretic differential geometry of gamma distributions, which contain exponential distributions as a special case, and log-gamma distributions. Our arguments support the opinion that these distributions have a natural role in representing departures from randomness, uniformity, and Gaussian behavior in stochastic processes. We show also how the information geometry provides a surprisingly tractable Riemannian manifold and product spaces thereof, on which may be represented the evolution of a stochastic process, or the comparison of different processes, by means of well-founded maximum likelihood parameter estimation. Our model incorporates possible correlations among parameters. We discuss applications and provide some illustrations from a recent study of amino acid self-clustering in protein sequences; we provide also some results from simulations for multisymbol sequences.**

***Index Terms*—Gamma models, information geometry, multisymbol sequences, random, search, stochastic process.**

I. Introduction to Gamma Models and Their Geometry

ELSEWHERE we have discussed the differential geometry of manifolds of gamma distributions and their application to various clustering problems and security testing, e.g., [5], [6], and [9]. The family of gamma distributions with event space $\Omega = \mathbb{R}^+$, parameters $\tau, \nu \in \mathbb{R}^+$ has probability density functions given by

$$f(t;\tau,\nu) = \left(\frac{\nu}{\tau}\right)^{\nu} \frac{t^{\nu-1}}{\Gamma(\nu)} e^{\frac{-t\nu}{\tau}} \quad t \in \mathbb{R}^+. \tag{1}$$

Then $\bar{t} = \tau$ is the mean and $Var(t) = \tau^2/\nu$ is the variance, so the coefficient of variation $\sqrt{Var(t)}/\tau = 1/\sqrt{\nu}$ is independent of the mean. The special case $\nu = 1$ in (1) corresponds to the situation of the random or Poisson process with mean inter-event interval τ.

For $\nu < 1$, (1) models a process that has larger variance than the random case; this corresponds to clustering since very small and very large values of t become more likely.

For *integer* $\nu = 1, 2, \ldots,$ (1) models a process that is Poisson but with intermediate events removed to leave only every ν^{th}; This would evidently have a smoothing effect for $\nu > 1$. Formally, the gamma distribution for *integer* ν is the ν-fold convolution of the exponential distribution, called also the Pearson Type III distribution.

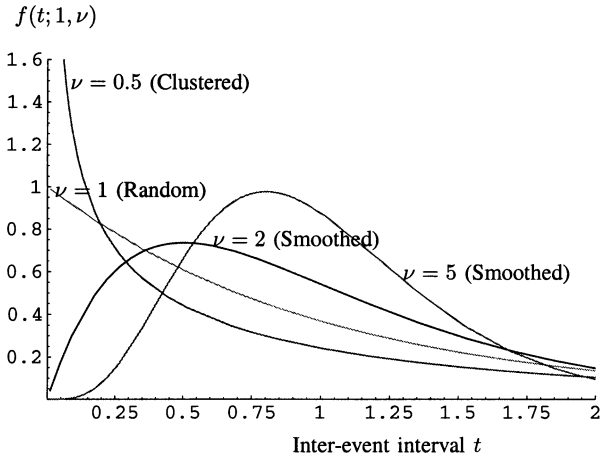


Fig. 1. Probability density functions, $f(t;\tau,\nu)$, for gamma distributions of inter-event intervals t with unit mean $\tau = 1$, and $\nu = 0.5,\ 1,\ 2,\ 5$. The case $\nu = 1$ corresponds to an exponential distribution from an underlying Poisson process; $\nu \neq 1$ represents some organization—clustering or smoothing.

Thus, gamma distributions can model a range of stochastic processes corresponding to nonindependent clustered events, for $\nu < 1$, and smoothed events, for $\nu > 1$, as well as the random case. Note that the property of having sample standard deviation independent of the mean actually characterizes gamma distributions, as shown recently by Hwang and Hu [12]. They proved, for $n \geq 3$ independent positive random variables $x_1, x_2, \ldots, x_n$ with a common continuous probability density function h, that having independence of the sample mean $\bar{x}$ and sample coefficient of variation $cv = \sigma/\bar{x}$ is equivalent to h being a gamma distribution. Fig. 1 shows some sample gamma distributions, all of unit mean, with $\nu = 0.5,\ 1,\ 2,\ 5$, thus, representing processes that are clustered, random and smoothed, respectively.

The log-likelihood function for a probability density function f is $l = \log f$; cf., eg [2] and [3] for more details of general results. Shannon's information theoretic entropy or "uncertainty" is given, up to a factor, by the negative of the expectation of the log-likelihood function. For the gamma densities (1) $l = \log(f(t;\tau,\nu))$ and the entropy is

$$\begin{aligned} S_f(\tau,\nu) &= -\int_0^\infty \log\left(f(t;\tau,\nu)\right) f(t;\tau,\nu)dt \\ &= \nu + (1-\nu)\frac{\Gamma'(\nu)}{\Gamma(\nu)} + \log\frac{\tau\Gamma(\nu)}{\nu}. \end{aligned} \tag{2}$$

In particular, at unit mean, the maximum entropy (or maximum uncertainty) occurs at $\nu = 1$, which is the random case, and

Manuscript received May 10, 2002; revised December 6, 2002. This work was supported by the CNPq (Conselho Nacional de Desenvolvimento Cientifico e Tecnològico, Brazil) and the British Council. This paper was recommended by Associate Editor M. Berthold.

C. T. J. Dodson is with the University of Manchester Institute of Science and Technology, Manchester M60 1QD, U.K. (e-mail: dodson@umist.ac.uk).

J. Scharcanski is with the Instituto de Informàtica, Universidade Federal do Rio Grande do Sul, Porto Alegre, RS, Brazil 91501-970 (e-mail: jacobs@inf.ufrgs.br).

Digital Object Identifier 10.1109/TSMCA.2003.809185

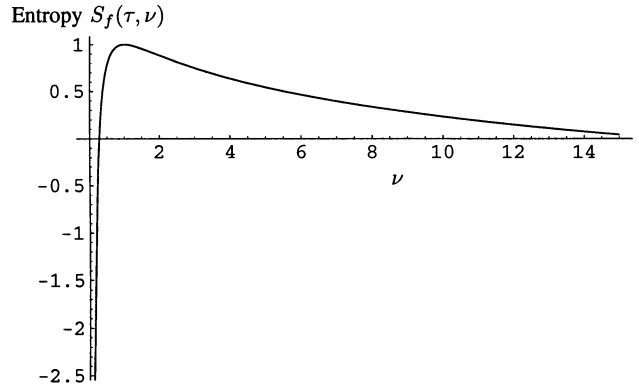


Fig. 2. Information entropy $S_f(\tau,\nu)$, for gamma distributions of inter-event intervals t with unit mean $\tau = 1$.

then $S_f(\tau,1) = 1 + \log\tau$. Fig. 2 shows a plot of $S_f(\tau,\nu)$, for the case of unit mean $\tau = 1$. So, a Poisson process of points on a line are as disorderly as possible and among all homogeneous point processes with a given density, the Poisson process has maximum entropy.

The maximum likelihood estimates $\hat{\tau},\hat{\nu}$ of τ,ν can be expressed in terms of the mean and mean logarithm of a set of independent observations $X = \{X_1, X_2, \ldots, X_n\}$. These estimates are obtained in terms of the properties of X by maximizing the log-likelihood function

$$\log lik_X(\tau,\nu) = \log\left(\prod_{i=1}^{n} p(X_i;\tau,\nu)\right)$$

with the following result that is easily applied to experimental data $\{X_1, X_2, \ldots, X_n\}$.

$$\hat{\tau} = \overline{X} = \frac{1}{n}\sum_{i=1}^{n} X_i \tag{3}$$

$$\log\hat{\nu} - \frac{\Gamma'(\hat{\nu})}{\Gamma(\hat{\nu})} = \overline{\log X} - \log\overline{X} \tag{4}$$

where $\overline{\log X} = 1/n\sum_{i=1}^{n}\log X_i$.

At each point in parameter space the covariance of partial derivatives of the log-likelihood function with respect to the parameters gives the Fisher information matrix, $[g_{ij}]$, which turns out to be positive definite. This matrix has entries the expectations

$$g_{ij} = \int_0^\infty \left(\frac{\partial l}{\partial\theta^i}\frac{\partial l}{\partial\theta^j}\right)dt = -\int_0^\infty \left(\frac{\partial^2 l}{\partial\theta^i\partial\theta^j}\right)dt \tag{5}$$

for coordinates $(\theta^i) \in \mathcal{G} = \mathbb{R}^+ \times \mathbb{R}^+$.

Since it is positive definite, $[g_{ij}]$ determines a Riemannian metric g on the parameter space $\mathcal{G}$, called the expected information metric for the parametric statistical model $\mathcal{G}$. Explicitly, the metric is given by the arc length function

$$ds^2 = \sum_{i,j} g_{ij}d\theta^i d\theta^j. \tag{6}$$

In our case, we have two parameters so we obtain a Riemannian 2-manifold and on the parameter space $\mathcal{G} = \{(\tau,\nu) \in \mathbb{R}^+ \times \mathbb{R}^+\}$ for gamma distributions, the arc length function is given by

$$ds^2 = \frac{\nu}{\tau^2}d\tau^2 + \left(\psi'(\nu) - \frac{1}{\nu}\right)d\nu^2 \quad \text{for } \tau,\nu \in \mathbb{R}^+ \tag{7}$$

where $\psi(\nu) = (\Gamma'(\nu))/(\Gamma(\nu))$ is the logarithmic derivative of the gamma function. The 1-dimensional subspace parameterized by $\nu = 1$ corresponds to all possible "random" (Poisson) processes, or equivalently, exponential distributions.

Dodson and Matsuzoe [9] have provided an affine immersion in Euclidean $\mathbb{R}^3$ for the Riemannian 2-manifold of gamma distributions with information metric (7). This may help in visualizing the geometric shape of the gamma manifold:

Proposition 1.1: (Dodson and Matsuzoe [9]): The coordinates $(\theta^1,\theta^2) = (\beta = \nu/\tau, \nu)$ form a natural coordinate system for the gamma manifold $\mathcal{G}$. Then $\mathcal{G}$ can be realized in Euclidean $\mathbb{R}^3$ by the graph of the affine immersion $\{h,\xi\}$ where ξ is the transversal vector field along h (cf., Amari and Nagaoka [3]):

$$h : \mathcal{G} \to \mathbb{R}^3 : \begin{pmatrix}\beta\\ \nu\end{pmatrix} \mapsto \begin{pmatrix}\beta\\ \nu\\ \log\Gamma(\nu) - \nu\log\beta\end{pmatrix},\ \xi = \begin{pmatrix}0\\0\\1\end{pmatrix}.$$

□

This immersion has been used to prove a general result which by its very qualitative nature is stable under small perturbations and hence should be useful in practice, giving confidence in the use of gamma distributions to model near random processes.

Proposition 1.2: (Arwini and Dodson [1]): Every neighborhood of an exponential distribution contains a neighborhood of gamma distributions, in the subspace topology of $\mathbb{R}^3$.

This means that in a rather precise sense, *every neighborhood of a random process on the real line has a neighborhood of processes that are represented by gamma distributions*.

It was proved elsewhere [8] that there is a Riemannian manifold $\mathcal{L}$ consisting of log-gamma distributions and isometric with $\mathcal{G}$. This log-gamma manifold $\mathcal{L}$ has several useful properties in security testing of smartcards and in modeling of galactic cluster evolution [7], [8].

Proposition 1.3: (Dodson [8]): The log-gamma probability density functions for random variable $N \in [0,1]$

$$g(N,\tau,\nu) = \frac{\frac{1}{N}^{1-\frac{\nu}{\tau}}\left(\frac{\nu}{\tau}\right)^{\nu}\left(\log\frac{1}{N}\right)^{\nu-1}}{\Gamma(\nu)} \text{ for } (\tau,\nu) \in \mathbb{R}^+ \times \mathbb{R}^+ \tag{8}$$

determine a metric space $\mathcal{L}$ of distributions with the following properties:

1) $\mathcal{L}$ contains the uniform distribution as the limit: $\lim_{\tau\to 1} g(N,\tau,1) = g(N,1,1) = 1$.
2) $\mathcal{L}$ contains approximations to truncated Gaussian distributions.
3) $\mathcal{L} \equiv \mathcal{G}$ is an isometry of the Riemannian manifold of gamma distributions with information-theoretic metric.□

In Fig. 3 are shown examples of the log-gamma distributions corresponding gamma distributions in Fig. 1. In fact, the log-gamma family (8) arises from the gamma family (1) for the nonnegative random variable $t = \log 1/N$, or equivalently, $N = e^{-t}$. So, the gamma and log-gamma families of distributions have a common differential geometry through the information metric and the exponential distributions in $\mathcal{G}$ map onto

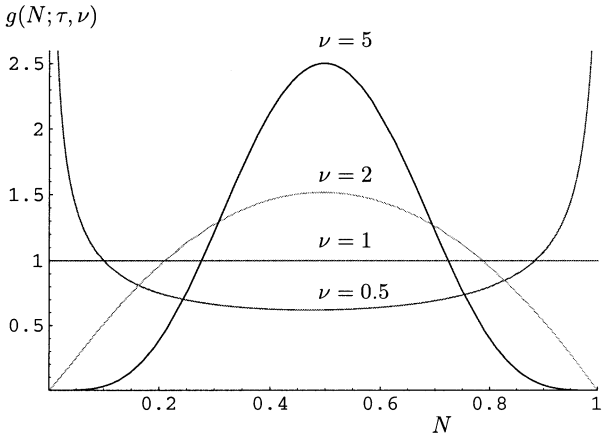


Fig. 3. Log-gamma probability density functions $g(N;\tau,\nu)$, with central mean $\bar{N} = 0.5$, and $\nu = 0.5,\ 1,\ 2,\ 5$. The case $\nu = 1$ is the uniform distribution, $\nu < 1$ corresponds to clustering in the underlying spatial process; conversely, $\nu > 1$ corresponds to smoothing.

the uniform distribution in $\mathcal{L}$, giving further topological properties through the isometry [1].

A. Correlation

Clearly, in certain bivariate stochastic processes we may expect that there will arise departures from randomness that incorporate correlation between the variables. So it is natural to consider bivariate gamma distributions.

Kibble's bivariate gamma distribution has been used in a variety of applications [13], but from our viewpoint it suffers from the disadvantage that its two marginal gamma distributions have a common dispersion parameter ν. Moreover, the calculation of the Fisher metric and its information geometry is intractable.

McKay's bivariate gamma distribution [14] is given by the density function

$$f(x,y;\alpha_1,\sigma_{12},\alpha_2) = \frac{\left(\frac{\alpha_1}{\sigma_{12}}\right)^{\frac{\alpha_1+\alpha_2}{2}} x^{\alpha_1-1}(y-x)^{\alpha_2-1}e^{-\sqrt{\frac{\alpha_1}{\sigma_{12}}}y}}{\Gamma(\alpha_1)\Gamma(\alpha_2)} \tag{9}$$

defined on $0 < x < y < \infty$ with parameters $\alpha_1, \sigma_{12}, \alpha_2 > 0$. Where σ_{12} is the covariance of x and y. This has the limitations that it constrains the random variables to the octant $0 < x < y < \infty$ and to have nonnegative covariance. The information geometry of this density function yields a Riemannian 3-manifold which has been studied by Arwini and Dodson [1] and will be reported elsewhere since the details of the geometry are rather cumbersome.

In the sequel, to circumvent these difficulties in developing easily applied information geometry of bivariate gamma manifolds, we introduce the notion of warped products of statistical manifolds. A simple direct product geometry like $\mathcal{G} \times \mathcal{G}$ represents the case when we have two independent stochastic processes subordinate to gamma distributions. Warped products allow us to create a new geometry from any pair of manifolds by blending them through a warping function which can represent interaction; no interaction remains as the special case for independent processes. Such methods are used in the pseudo- Riemannian geometry of general relativistic spacetime, cf., Beem *et al.* [4].

II. Curves and Distances in $\mathcal{G}$

In the manifold $\mathcal{G}$ of gamma models for the distribution of intervals between events, we use the Riemannian metric to measure information distances between pairs of points. In a neighborhood of a given point we can obtain a locally bilinear approximation to this distance. From (7) for small variations $\Delta\tau, \Delta\nu$, near $(\tau_0,\nu_0) \in \mathcal{G}$; it is approximated by

$$\Delta s_{\mathcal{G}} \approx \sqrt{\frac{\nu_0}{\tau_0^2}\Delta\tau^2 + \left(\psi'(\nu_0) - \frac{1}{\nu_0}\right)\Delta\nu^2}. \tag{10}$$

As ν_0 increases from 1, the factor $(\psi'(\nu_0) - (1)/(\nu_0))$ decreases monotonically from $(\pi^2)/(6) - 1$. So, in the information metric, the difference $\Delta\tau$ has increasing prominence over $\Delta\nu$ as the standard deviation reduces with increasing ν_0—corresponding to increased temporal smoothing of event scheduling.

In particular, near the exponential distribution, where $(\tau_0,\nu_0) = (1,1)$, (10) is approximated by

$$\Delta s_{\mathcal{G}} \approx \sqrt{\Delta\tau^2 + \left(\frac{\pi^2}{6} - 1\right)\Delta\nu^2}. \tag{11}$$

For a practical implementation we need to obtain rapid estimates of distances in larger regions than can be represented by quadratics in incremental coordinates. This can be achieved using the result of Dodson and Matsuzoe [9] that established geodesic foliations of the gamma manifold. Now, a geodesic curve is locally minimal and so a network of two nonparallel sets of geodesics provides a mesh of upper bounds on distances by using the triangle inequality about any point. Such a geodesic mesh is shown in Fig. 4 using the geodesic curves $\tau = \nu$ and $\nu =$ constant, which foliate $\mathcal{G}$, as described in [9].

Explicitly, the arc length along the geodesic curves $\tau = \nu$ from $(\tau_0 = \nu_0, \nu_0)$ to $(\tau = \nu, \nu)$ is

$$\left|\frac{d^2\log\Gamma}{d\nu^2}(\nu) - \frac{d^2\log\Gamma}{d\nu^2}(\nu_0)\right|$$

and the distance along curves of constant $\nu = \nu_0$ from (τ_0,ν_0) to (τ,ν_0) is

$$\left|\nu_0 \log\frac{\tau_0}{\tau}\right|.$$

The functions involved in these latter two distances can be obtained numerically so at any given parameter values they can be substituted directly. In Fig. 4, we use the base point $(\tau_0,\nu_0) = (20,1) \in \mathcal{G}$ and combine the above two arc lengths of the geodesics to obtain an upper bound on distances from (τ_0,ν_0) as

$$\begin{aligned} &Distance\,[(\tau_0,\nu_0),(\tau,\nu)] \\ &\quad \le \left|\frac{d^2\log\Gamma}{d\nu^2}(\nu) - \frac{d^2\log\Gamma}{d\nu^2}(\nu_0)\right| + \left|\nu_0\log\frac{\tau_0}{\tau}\right|. \end{aligned} \tag{12}$$

III. Product Geometries

In a practical application of the above differential geometry we can measure departures from randomness in the gamma manifold $\mathcal{G}$. Equivalently, in the log-gamma manifold $\mathcal{L}$ we

can measure differences between approximations to truncated Normal distributions or departures from a uniform distribution.

In fact all of these types of comparison between such distributions—or empirical sampling of them—arise in the cost function for approximating a given stochastic process. In general, however, we have n distributed parameter sets to optimize. First we consider the case where our n search parameters all come from the joint families of gamma and log-gamma distributions and are independent of one another. Then we have a product manifold $\mathcal{P}$ of dimension $2n$ with n pairs of coordinates $\{(\tau_i, \nu_i) | i = 1, 2, \ldots, n\}$. So, $\mathcal{P}$ consists of a product of γ copies of $\mathcal{G}$ and λ copies of $\mathcal{L}$, where $\gamma, \lambda \geq 0$ and $\gamma + \lambda = n$. Hence,

$$\mathcal{P} = \mathcal{G}^{\gamma} \times \mathcal{L}^{\lambda} \tag{13}$$

with the n-fold direct product metric of (I.7)

$$\begin{aligned} ds^2_{\mathcal{P}} = &\sum_{i=1}^{\gamma} \left(\frac{\nu_i}{\tau_i^2} d\tau_i^2 + \left(\psi'(\nu_i) - \frac{1}{\nu_i} \right) d\nu_i^2 \right) \\ &+ \sum_{i=\gamma+1}^{n} \left(\frac{\nu_i}{\tau_i^2} d\tau_i^2 + \left(\psi'(\nu_i) - \frac{1}{\nu_i} \right) d\nu_i^2 \right) \\ &\text{for } \tau_i, \nu_i \in \mathbb{R}^+ \text{ but } \nu_i \geq 1 \text{ for } i > \gamma. \end{aligned} \tag{14}$$

We note that each component space in such products contributes two dimensions.

A. Warped Products and Correlation

More intricate products arise in applications of geometry to physics, as discussed for example in Beem *et al.* [4]. A *warped product* of two Riemannian manifolds (X, g) with coordinates (x_i) and (Y, h) with coordinates (y_i) is a manifold $(X \times Y, g \times_f h)$ with coordinates $(z_i) = ((x_i), (y_i))$ under the metric $g \times_f h$ has the form

$$g \times_f h_{ij}\, dz^i dz^j = g_{ij}\, dx^i dx^j + f(x_i)\, h_{ij}\, dy^i dy^j$$

for some positive warping function, f defined on X. It is possible that correlation may be represented to some extent by a suitable choice of warping function in a warped product of statistical manifolds; this is under investigation.

Meanwhile, it seems that some empiricism may be needed to introduce correlation between variables in the manifold $\mathcal{P}$. One way would be to modify the direct product metric by introducing symmetrically off-diagonal terms g_{ij}, $i \neq j$, while preserving positive definiteness. These off-diagonal terms could be bounded by $\pm\epsilon$, say, and ranked in absolute size by the relative strengths of the corresponding correlations.

B. Example of Two-Fold Products

Let us take for illustration the submatrix of the metric g_{ij} for $i, j = 1, 2, \ldots, 4$; so it applies to one of the spaces $\mathcal{G}^2$, $\mathcal{G} \times \mathcal{L}$, $\mathcal{L} \times \mathcal{G}$, $\mathcal{L}^2$, as part of the n-fold product $\mathcal{P}$. Then this part of the metric tensor matrix g_{ij} will have the form

$$[M_{12}] = \begin{pmatrix} \frac{\nu_1}{\tau_1^2} & 0 & \rho_{13} & \rho_{14} \\ 0 & \psi'(\nu_1) - \frac{1}{\nu_1} & \rho_{23} & \rho_{24} \\ \rho_{13} & \rho_{23} & \frac{\nu_2}{\tau_2^2} & 0 \\ \rho_{14} & \rho_{24} & 0 & \psi'(\nu_2) - \frac{1}{\nu_2} \end{pmatrix} \tag{15}$$

and the information distance arc length element from this component of the metric tensor will be given by

$$ds^2 = X[M_{12}]X^T \tag{16}$$

where $X = (d\tau_1 d\nu_1 d\tau_2 d\nu_2)$.

Here, the off-diagonal terms $\epsilon\rho_{ij}$ are symmetric and consist of the product of the correlation coefficient $\rho_{ij} = \rho_{ji}$ between the two parameter spaces and the scale control ϵ. The scaling value ϵ must be chosen such that $\det[M_{12}] > 0$, to ensure that positive definiteness is preserved. The maximum likelihood estimates should be used for the parameter values (τ_i, ν_i) obtained from measured data histograms.

The simplest case is perhaps that of a relationship between the two mean values, τ_1, τ_2. For this suppose that all of the ρ_{ij} are zero except $\rho_{13} = \rho$, say. Then, we have to control the size of ρ in order to have $\det[M_{12}] > 0$, namely

$$\begin{vmatrix} \frac{\nu_1}{\tau_1^2} & 0 & \rho & 0 \\ 0 & \psi'(\nu_1) - \frac{1}{\nu_1} & 0 & 0 \\ \rho & 0 & \frac{\nu_2}{\tau_2^2} & 0 \\ 0 & 0 & 0 & \psi'(\nu_2) - \frac{1}{\nu_2} \end{vmatrix} > 0. \tag{17}$$

But we know that the product of diagonal terms is positive because this is the determinant for the trivial product space, i.e. with $\rho = 0$. Hence, the constraint reduces to

$$\rho^2 < \frac{\nu_1}{\tau_1^2} \frac{\nu_2}{\tau_2^2} = \frac{1}{Var_1\, Var_2} \tag{18}$$

$$-\frac{\sqrt{\nu_1 \nu_2}}{\tau_1 \tau_2} < \rho < +\frac{\sqrt{\nu_1 \nu_2}}{\tau_1 \tau_2} \tag{19}$$

and so the magnitude of ρ is bounded by the reciprocal of the geometric mean of the variances of the two marginal gamma distributions in the product. This bound could be estimated once the domain of interest was established.

C. Representing Multimodal Distributions

A large class of distributions arise in practical situations as bimodal or multimodal histograms. A typical situation is that of several disjoint symmetric peaks. We can easily handle the case when the peaks all resemble gamma or log-gamma shaped distributions; we just multiply the metric contribution of each peak by the total probability fraction represented by that peak.

Suppose that an observed data set has a histogram $\mathcal{H}$ with k peaks giving respective fractional contributions $p_1, p_2, \ldots, p_k$ to the total probability. If each peak is well represented by a gamma or log-gamma distribution, then there will be a $2k$-dimensional subspace corresponding to such histograms and its metric will be

$$ds^2_{\mathcal{H}} = \sum_{i=1}^{i=k} p_i \left(\frac{\nu_i}{\tau_i^2} d\tau_i^2 + \left(\psi'(\nu_i) - \frac{1}{\nu_i} \right) d\nu_i^2 \right) \tag{20}$$

with $0 \leq p_i \leq 1$ and $\sum_{i=1}^{i=k} p_i = 1$.

IV. Applications

A number of applications arise rather naturally from the observation that the gamma and log-gamma distributions have a natural role in representing departures from randomness, uniformity and Gaussian behavior in stochastic processes. We

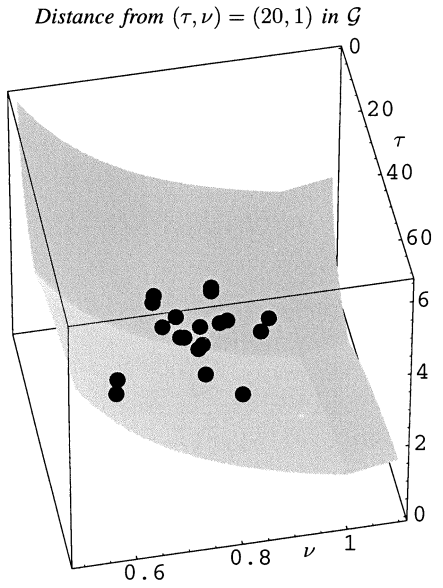


Fig. 4. Distances in the space of gamma models, using a geodesic mesh. The surface height represents upper bounds on distances from $(\tau, \nu) = (20, 1)$, the random case with mean $\tau = 20$. Depicted also are 20 data points for a set of amino acid sequences with clustering to differing degrees.

have begun studies of several such situations, some are outlined below.

A. Characterizing Self-clustering of Amino Acids

The data plotted on the distance surface in Fig. 4 comes from measurements of occurrencies of individual amino acids along a protein chain within the Saccharomyces cerevisiae genome, see Cai *et al.* [5]. If amino acids are distributed randomly within a sequence then they follow a Poisson process and a histogram of the number of observations of each gap size will follow a negative exponential distribution. Our techniques show that this is not the case and that all 20 amino acids tend to cluster, all having $\nu < 1$. In other words, the frequencies of short gap lengths tends to be higher and the variance of the gap lengths is greater than expected by chance. In this application we have a one-dimensional (1-D) space $\mathcal{G}$ where the intervals are between successive occurrencies of a given amino acid, for all 20 possible amino acids. The maximum-likelihood parameters were obtained for gamma fits to the interval distribution for each amino acid.

The methodology here allows representation of the departures from randomness of the processes that allocate gaps between occurrences of each amino acid. Fig. 4 shows information distances in the space of gamma models, using a geodesic mesh; the surface height represents upper bounds on distances from $(\tau, \nu) = (20, 1)$, the random case with mean $\tau = 20$. Depicted also are the 20 data points for the set of amino acid sequences; these show clustering to differing degrees.

The data for Fig. 4 consisted of sequences with of the order of 10^5 occurrencies and from this the maximum-likelihood parameters were obtained. Here we have then a reduction of some three million experimentally determined amino acid positions to just 20 points and the qualitative result that all amino acids within the Saccharomyces cerevisiae genome exhibit self-clustering. We might expect that such stable stochastic information in these long sequences encodes important features that may be relevant in genetic analysis.

TABLE I
TYPICAL SIMULATION RESULTS FOR 4-SYMBOL SEQUENCES OF LENGTH 10,000, WITH THE SYMBOLS HAVING ABUNDANCE DISTRIBUTIONS: I UNIFORM AND II EXPONENTIAL. THESE DATAPOINTS ARE PLOTTED IN FIG. 5

	I Uniform			*II Exponential*		
	Probability	ν	τ	*Probability*	ν	τ
A	0.25	1.27	4.03	0.45	1.75	2.22
B	0.25	1.39	3.98	0.28	1.38	3.59
C	0.25	1.29	4.00	0.17	1.16	5.89
D	0.25	1.32	3.99	0.10	1.05	9.80

B. Stochastic Similarity for Multisymbol Sequences

An application of the gamma manifold $\mathcal{G}$ would be to provide a structural model for stochastic features of intervals between consecutive occurrence of symbols through multi-symbol sequences. If the intervals between occurrences of a given symbol exhibit the property that their coefficient of variation is independent of the mean, then their distribution may be modeled by a gamma distribution. Clustering $(\nu < 1)$ would occur when the symbol has greater frequency in certain sections; smoothing $(\nu > 1)$ would occur for symbols that are more regularly spaced than at random.

In order to illustrate how the metric might benefit the study of stochastic sequences of symbols, we have developed a simulator which generates a wide range of such sequences, of arbitrary length and with arbitrarily many symbols. The probability of occurrence of symbols is either uniformly distributed over symbol types or not; if it is not uniform, then we can represent the ranked probability values by a triangular-type distribution—exponential serves well enough.

We extract some information from such simulated sequences by computing the maximum likelihood estimate of gamma distribution parameters (τ_k, ν_k) for each symbol k.

Sample results from sequences of length 10 000 using four symbols are shown in Table I for uniform and exponential abundance distributions, symbols being chosen independently with replacement. Fig. 5 shows the results, illustrating the distances in the space of gamma models, using a geodesic mesh. The surface height represents upper bounds on distances from $(\tau, \nu) = (4.7, 1)$, the random case with mean $\tau = 4.7$. Depicted also are data points from Table I for two sample simulations of sequences of length 10 000 with four symbols. The small points near the center are from a uniform distribution of symbol abundances; the four larger points are from an exponential distribution of abundances.

We see from Fig. 5 that both processes yield sequences of symbols all exhibiting more smoothing than random, namely all have $\nu > 1$. In the case of the nonuniform abundances, we observe, as expected that the mean interval τ between occurrences of a symbol decreases with increasing abundance, essen-

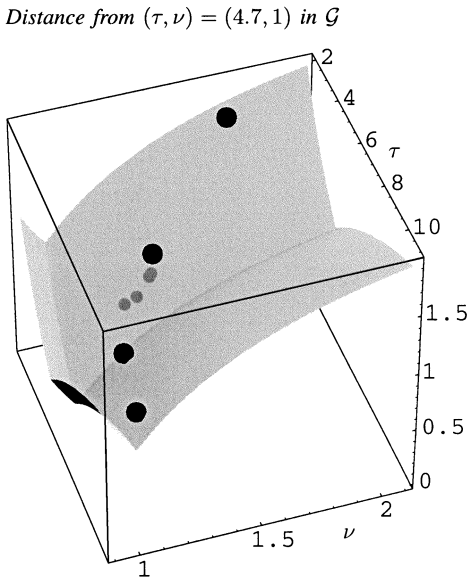


Fig. 5. Distances in the space of gamma models, using a geodesic mesh. The surface height represents upper bounds on distances from $(\tau, \nu) = (4.7, 1)$, the random case with mean $\tau = 4.7$. Depicted also are data points from Table I for two sample simulations of sequences of length 10 000 with four symbols. The small points near the center are from a uniform distribution of symbol abundances; the four larger points are from an exponential distribution of abundances.

tially one is a reciprocal of the other. The other parameter, ν, increases also with abundance, in a systematic way.

The extraction of such features from long multisymbol sequences might be of value in monitoring and managing information flow through large networks. In such cases, dynamic management might benefit from knowledge of qualitative and partially quantitative properties of datastream flow simply by exploiting stable stochastic features.

V. Concluding Remarks

We have offered arguments to support the opinion that the gamma and log-gamma distributions have a natural role in representing departures from randomness, uniformity and Gaussian behavior in stochastic processes. We show also how the information geometry provides a surprisingly tractable Riemannian manifold and product spaces thereof, on which may be represented the evolution of a stochastic process or the comparison of different processes, by means of well-founded maximum likelihood parameter estimation.

References

[1] K. Arawini and C. T. J. Dodson. (2002) Information geometric neighborhoods of randomness and geometry of the McKay bivariate gamma 3-manifold. [Online]. Available: http://www.ma.umist.ac.uk/kd/PREPRINTS/gamran.pdf.

[2] S.-I. Amari, *Differential Geometrical Methods in Statistics*. Berlin, Germany: Springer-Verlag, 1985.

[3] S. Amari and H. Nagaoka, *Methods of Information Geometry, American Mathematical Soc.,*. New York: Oxford Univ. Press, 2000.

[4] J. K. Beem, P. E. Ehrlich, and K. L. Easley, *Global Lorentzian Geometry*, 2nd ed. New York: Marcel Dekker, 1996.

[5] Y. Cai, C. T. J. Dodson, O. Wolkenhauer, and A. J. Doig, "Information-theoretic analysis of protein sequences show that amino acids self cluster," *J. Theor. Biol.*, vol. 218, no. 4, pp. 409–418, 2002.

[6] C. T. J. Dodson, "Information geodesics for communication clustering," *J. Statist. Comput. Simulat.*, vol. 65, pp. 133–146, 2000.

[7] C. T. J. Dodson, "Spatial statistics and information geometry for parametric statistical models of galaxy clustering," *Int. J. Theor. Phys.*, vol. 38, no. 10, pp. 2585–2597, 1999.

[8] C. T. J. Dodson, "Geometry for stochastically inhomogeneous spacetimes," *Nonlinear Anal.*, vol. 47, pp. 2951–2958, 2001.

[9] C. T. J. Dodson and H.Matsuzoe Matsuzoe, "An affine embedding of the gamma manifold," *Appl. Sci.*, vol. 5, no. 1, pp. 1–6, 2003.

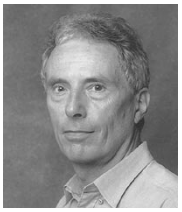
[10] C. T. J. Dodson and T. Poston, *Tensor Geometry*, 2nd ed. New York: Springer-Verlag, 1991.

[11] A. Gray, *Modern Differential Geometry of Curves and Surfaces*, 2nd ed. Boca Raton, FL: CRC, 1998.

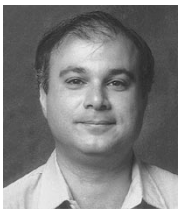
[12] T.-Y. Hwang and C.-Y. Hu, "On a characterization of the gamma distribution: the independence of the sample mean and the sample coefficient of variation," *Ann. Inst. Statist. Math.*, vol. 51, no. 4, pp. 749–753, 1999.

[13] W. F. Kibble, "A two variate gamma-type distribution," *Sankhyā*, vol. 5, pp. 137–150, 1941.

[14] K. V. Mardia, *Families of Bivariate Distributions.* London, U.K., 1970.



Christopher T. J. Dodson has been a Professor in the Department of Mathematics, University of Manchester Institute of Science and Technology, Manchester, U.K., since 1996. From 1989 to 1996, he was NSERC Abitibi-Price Senior Research Chair, University of Toronto, Tornoto, ON, Canada. From 1969 to 1989, he was with the Department of Mathematics, Lancaster University, U.K. His research interests include differential geometry, stochastic geometry and applications to spacetime structure, stochastic processes, and information systems. He has recently written several books, including *A User's Guide to Algebraic Topology* (with P.E. Parker), (Norwell, MA: Kluwer, 1997), *Tensor Geometry, Graduate Texts in Mathematics 120* (Berlin, Germany: Springer-Verlag, 1991, 1997), and *An Engineered Stochastic Structure* (Atlanta, GA: Tappi Press, 1994).



Jacob Scharcanski (SM'03) received the B.Eng. degree in electrical engineering and the M.Sc. degree in computer science in 1981 and 1984, respectively, from the Federal University of Rio Grande do Sul, Brazil, and the Ph.D. degree in systems design engineering, University of Waterloo, Waterloo, ON, Canada in 1993.

His main areas of interest are image processing and analysis, pattern recognition, and visual information retrieval. He has lectured at the University of Toronto, Toronto, ON, University of Guelph, Guelph, ON, Canada, Universtiy of East Anglia and Universtiy of Manchester, Manchester, U.K., as well as at several Brazilian Universities. He authored and coauthored more than 60 publications in Journals and Conferences. He also held research and development positions in the Brazilian and North-American Industry. Currently, he is an Associate Professor with the Institute of Informatics, Federal University of Rio Grande do Sul, Porto Alegre, RS, Brazil.