Information geometry for bivariate distribution control

C.T.J. Dodson* and H. Wang**
*Department of Mathematics,
**Control System Centre,
University of Manchester Institute of Science and Technology,
Manchester M60 1QD, UK

ABSTRACT

The optimal control of stochastic processes through sensor estimation of probability density functions has a geometric setting via information theory and the information metric. Information theory identifies the exponential distribution as the maximum entropy distribution if only the mean is known and the gamma distribution if also the mean logarithm is known. Previously, we used the surface representing gamma models to provide an appropriate structure on which to represent the dynamics of a univariate process and algorithms to control it. In this paper we extend these procedures to gamma models with positive correlation, for which the information theoretic 3-manifold geometry has recently been formulated. For comparison we summarize also the case for bivariate Gaussian processes with arbitrary correlation.

Keywords: Information geometry, gamma distribution, Gaussian distribution, McKay bivariate gamma distribution, bivariate Gaussian distribution, B-spline, dynamics, bivariate control.

1. INTRODUCTION

Certain continuous processes involve parametric statistical models to represent dynamic stochastic features of one and two-dimensional time series or textures. In such cases the sampling of the distributed variable may reveal natural non-Gaussian behaviour to optimize. Recent work of Wang [9, 10, 11] has used B-spline bump functions to decompose an arbitrary probability density function and hence optimize its control. We used in [6] the information geometry of spaces of gamma distributions to model the control of the particle size distribution in a commercial turbid suspension. In the next section we consider the information theoretic geometry of bivariate gamma statistical models with positive correlation and offer a Riemannian space on which dynamics may be studied and optimized.

2. McKAY BIVARIATE GAMMA 3-MANIFOLD

The classical family of Mckay bivariate gamma distributions defined on $0 < x < y < \infty$ with parameters $\alpha_1, \sigma_{12}, \alpha_2 > 0$ is given by:

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

where σ_{12} is the covariance of X, Y. The correlation coefficient and marginal gamma distributions, of X and Y are given respectively, for positive x, y by :

$$\begin{aligned}
\rho(X,Y) &= \sqrt{\frac{\alpha_1}{\alpha_1+\alpha_2}} \\
f_X(x) &= \frac{(\frac{\alpha_1}{\sigma_{12}})^{\frac{\alpha_1}{2}} x^{\alpha_1-1}e^{-\sqrt{\frac{\alpha_1}{\sigma_{12}}}x}}{\Gamma(\alpha_1)}, \\
f_Y(y) &= \frac{(\frac{\alpha_1}{\sigma_{12}})^{\frac{(\alpha_1+\alpha_2)}{2}} y^{(\alpha_1+\alpha_2)-1}e^{-\sqrt{\frac{\alpha_1}{\sigma_{12}}}y}}{\Gamma(\alpha_1+\alpha_2)}.
\end{aligned}$$

These marginal functions are: the probability density of x averaged over all y values, and the probability density of y averaged over all x values, respectively. Both are gamma distributions, but note that it is not possible to choose parameters such that both marginal functions are exponential—ie with dispersion parameters unity.

Explicitly, $f_X(x)$ is a gamma density with mean $\sqrt{\alpha_1\sigma_{12}}$ and dispersion parameter α_1. Similarly, $f_Y(y)$ is a gamma density with mean $(\alpha_1+\alpha_2)\sqrt{\frac{\sigma_{12}}{\alpha_1}}$ and dispersion parameter $(\alpha_1+\alpha_2)$.

So, for the McKay distribution to be applicable to a given joint distribution for (x, y), we need to have:

- $0 < x < y < \infty$
- Covariance > 0
- (Dispersion parameter for y) $>$ (Dispersion parameter for x)

- And we expect, roughly,
  (Mean x)(Mean y) $= \frac{\alpha_1+\alpha_2}{\alpha_1}$

Using similar methods to those described in [6] for the univariate gamma distribution, Arwini and Dodson [2] provided the Mckay bivariate gamma model as a 3-manifold, equipped with Fisher information as Riemannian metric; they derived the induced geometry, i.e., the Riemann curvature tensor, the Ricci tensor, the scalar curvatures etc.

Denote by M the set of Mckay bivariate gamma distributions, that is

$$\begin{aligned} M &= \{f|f(x,y;\alpha_1,\sigma_{12},\alpha_2) \\ &= \frac{(\frac{\alpha_1}{\sigma_{12}})^{\frac{(\alpha_1+\alpha_2)}{2}} x^{\alpha_1-1}(y-x)^{\alpha_2-1}e^{-\sqrt{\frac{\alpha_1}{\sigma_{12}}}y}}{\Gamma(\alpha_1)\Gamma(\alpha_2)}, \\ & \quad y>x>0,\ \alpha_1,\sigma_{12},\alpha_2>0\} \end{aligned} \tag{2}$$

Then we have:
Global coordinates $(\alpha_1,\sigma_{12},\alpha_2)$ make M a 3-manifold with Fisher information metric $[g_{ij}]$ given by :

$$[g_{ij}] = \begin{bmatrix} \frac{-3\,\alpha_1+\alpha_2}{4\,{\alpha_1}^2}+\psi'(\alpha_1) & \frac{\alpha_1-\alpha_2}{4\,\alpha_1\,\sigma_{12}} & -\frac{1}{2\,\alpha_1} \\ \frac{\alpha_1-\alpha_2}{4\,\alpha_1\,\sigma_{12}} & \frac{\alpha_1+\alpha_2}{4\,{\sigma_{12}}^2} & \frac{1}{2\,\sigma_{12}} \\ -\frac{1}{2\,\alpha_1} & \frac{1}{2\,\sigma_{12}} & \psi'(\alpha_2) \end{bmatrix} \tag{3}$$

It follows that for small changes $d\alpha_1, d\alpha_2, d\sigma_{12}$ the element of arclength ds is given by

$$ds^2 = \sum_{ij=1}^{3} g_{ij}\,dx_i\,dx_j \tag{4}$$

with $(x_1,x_2,x_3) = (\alpha_1,\sigma_{12},\alpha_2)$. For larger separations between two bivariate gamma distributions the arclength along a curve is obtained by integration of ds.

A path through the parameter space M models the dynamics of a process change as a curve, parametrized by t in some interval $a \le t \le b$, given by

$$c : [a,b] \to M : t \mapsto (c_1(t), c_2(t), c_3(t)) \tag{5}$$

and its tangent vector $\dot{c}(t) = (\dot{c}_1(t), \dot{c}_2(t), \dot{c}_3(t))$ has norm $||\dot{c}||$ given via (3) by

$$||\dot{c}(t)||^2 = \sum_{i,j=1}^{3} g_{ij}\,\dot{c}_i(t)\,\dot{c}_j(t). \tag{6}$$

and the information length of the curve is

$$L_c(a,b) = \int_a^b ||\dot{c}(t)||\,dt \quad \text{for } a \le b. \tag{7}$$

Since arc length is often difficult to evaluate analytically, we sometimes use the 'energy' of the curve instead of length for comparison of information cost differences between nearby curves. Energy is given by integrating the square of the norm of $\dot{c}$

$$E_c(a,b) = \int_a^b ||\dot{c}(t)||^2\,dt. \tag{8}$$

Arwini and Dodson [2] gave details of the curvature tensor and its related operators, together with studies of some submanifolds of the McKay manifold, and illustrations of how the correlation influences the curvature. The McKay manifold has non-constant negative scalar curvature. This is in contrast to the manifold of bivariate Gaussian distributions, which we describe in the next section and which has constant negative curvature.

3. BIVARIATE GAUSSIAN 5-MANIFOLD

The probability density of the 2-dimensional normal distribution has the form:

$$f(x) = \frac{1}{2\pi}\,|\Sigma|^{-\frac{1}{2}}\,e^{-\frac{1}{2}(x-\mu)'\Sigma^{-1}(x-\mu)}, \tag{9}$$

where

$$x = \begin{bmatrix} x_1 \\ x_2 \end{bmatrix}, \quad \mu = \begin{bmatrix} \mu_1 \\ \mu_2 \end{bmatrix}, \quad \Sigma = \begin{bmatrix} \sigma_{11} & \sigma_{12} \\ \sigma_{12} & \sigma_{22} \end{bmatrix},$$

$$\begin{aligned} -\infty < x_1 &< x_2 < \infty, \\ -\infty < \mu_1 &< \mu_2 < \infty, \\ 0 &< \sigma_{11}, \sigma_{22} < \infty. \end{aligned}$$

This contains the five parameters $\mu_1, \mu_2, \sigma_{11}, \sigma_{12}, \sigma_{22}$. So our global coordinate system consists of the 5-tuples

$$\theta = (\theta^1,\theta^2,\theta^3,\theta^4,\theta^5) = (\mu_1,\mu_2,\sigma_{11},\sigma_{12},\sigma_{22}).$$

We put

$$\begin{aligned} \theta^1 &= \mu_1, \\ \theta^2 &= \mu_2, \\ \theta^3 &= \sigma_{11}, \\ \theta^4 &= \sigma_{12}, \\ \theta^5 &= \sigma_{22}. \end{aligned}$$

The information metric tensor is known [8] to be

given by:

$$[g_{ij}] = \begin{bmatrix} \frac{\theta^5}{\triangle} & -\frac{\theta^4}{\triangle} & 0 & 0 & 0 \\ -\frac{\theta^4}{\triangle} & \frac{\theta^3}{\triangle} & 0 & 0 & 0 \\ 0 & 0 & \frac{(\theta^5)^2}{2\triangle^2} & -\frac{\theta^4\theta^5}{\triangle^2} & \frac{(\theta^4)^2}{2\triangle^2} \\ 0 & 0 & -\frac{\theta^4\theta^5}{\triangle^2} & \frac{\theta^3\theta^5+(\theta^4)^2}{\triangle^2} & -\frac{\theta^3\theta^4}{\triangle^2} \\ 0 & 0 & \frac{(\theta^4)^2}{2\triangle^2} & -\frac{\theta^3\theta^4}{\triangle^2} & \frac{(\theta^3)^2}{2\triangle^2} \end{bmatrix} \quad (10)$$

where $\triangle$ is the determinant

$$\triangle = |\Sigma| = \theta^3\theta^5 - (\theta^4)^2.$$

4. SQUARE ROOT B-SPLINE APPROXIMATION TO BIVARIATE GAMMA DISTRIBUTIONS

The management and performance optimization of a stochastic process represented by gamma parametric models may then be represented through this Riemannian geometry. Other distributions may be treated similarly. In the case of simultaneous handling of two probability density functions, considered by Wang [11], we have a product of two Riemannian spaces and correlations between the two distributions would appear as a twisting of this product.

In this case the family of bivariate gamma distributions can still be considered. However, since

$$\lim_{x,y\to+\infty} f(x, y, \alpha_1, \sigma_{12}, \alpha_2) = 0 \quad (11)$$

for any arbitrarily small $\epsilon > 0$ there is a $b(\epsilon, \alpha_1, \sigma_{12}, \alpha_2) > 0$ such that the following inequality holds

$$\begin{aligned} \forall x, y &> b(\epsilon, \alpha_1, \sigma_{12}, \alpha_2) \\ |f(x, y, \alpha_1, \sigma_{12}, \alpha_2)| &\leq \epsilon. \end{aligned} \quad (12)$$

This indicates that we can use the following B-spline functions to approximate the probability density function to give

$$|\sqrt{f(x, y, \alpha_1, \sigma_{12}, \alpha_2)} - \sum_{i=1}^{n} w_i B_i(x, y)| \leq \delta \quad (13)$$

where $B_i(x, y)$ are the pre-specified bivariate basis functions defined on $\Omega = [0, b] \times [0, b]$, w_i are the weights to be trained adaptively and δ is a small number generally larger than ϵ. It has been shown that the square root approximation has the advantage of high numerical robustness in comparison with linear B-splines. Indeed, such an approximation will be used here to represent the coupled links between the three parameters and the probability density function f.

Since all the basis functions are pre-specified, different values of $\{\alpha_1, \sigma_{12}, \alpha_2\}$ will generate different sets of weights. As such, the approximation (4.16) should be further represented as

$$\sqrt{f(x, y, \alpha_1, \sigma_{12}, \alpha_2)} = \sum_{i=1}^{n} w_i B_i(x, y) + e \quad (14)$$

where $|e| \leq \delta$. In Wang ([12]), the following transformed representation has been used

$$\sqrt{f(x, y, \alpha_1, \sigma_{12}, \alpha_2)} = C(x, y)V_k + h(V_k)B_n(x, y) \quad (15)$$

to guarantee that

$$\int_0^b \int_0^b f(x, y, \alpha_1, \sigma_{12}, \alpha_2)dxdy = 1$$

where $V_k = (w_1, w_2..., w_{n-1})^T$ constitutes a vector of independent weights, and $h(.)$ is a known nonlinear function of V_k.

With this format, the relationship between V and $\{\alpha_1, \sigma_{12}, \alpha_2\}$ can be formulated using (15).

5. CONTROL OF THE DISTRIBUTION

Once such square root models are formulated, the next question arising from many practical situations is to see how the parameters of a distribution can be selected so that the actual distribution is made as close as possible to a given one. This is a control problem and has many applications in particulate processing.

It is assumed that the initial bivariate gamma distribution is characterised by $\{\alpha_1(0), \sigma_{12}(0), \alpha_2(0)\}$ and the desired distribution is presented by $\{\alpha_1(f), \sigma_{12}(f), \alpha_2(f)\}$, then focus will be made on the evaluation of how the weight vector V_k will behave in tuning the initial distribution to the final distribution. Since the weight vector V_k is directly related to the three parameters in the bivariate gamma distribution, the tuning rule for the weight vector V_k will be used. Let $g(x, y)$ represent the probability density of the bivariate gamma distribution with parameter set

$$\{\alpha_1(f), \sigma_{12}(f), \alpha_2(f)\}.$$

This means that an effective trajectory for V_k should be chosen to minimise the following performance function

$$J = \frac{1}{2} \int_0^b \int_0^b g(x, y) log \frac{K(x, y)}{g(x, y)} dxdy \quad (16)$$

with $K(x,y) = C(x,y)V_k + h(V_k)B_n(x,y)$.

This leads to the following application of the gradient rule

$$V_k = V_{k-1} - \eta \frac{\partial J}{\partial V}|_{V=V_{k-1}} \tag{17}$$

where $k = 0, 1, 2, \ldots$ represents the sample number and $\eta > 0$ is a pre-specified learning rate.

Using the relationship between the weight vector and the three parameters, the adaptive tuning of the actual parameters in the pdf of (15) can be readily formulated.

Now, for two probability density functions p and p' on an event space Ω, the function

$$KL(p,p') = \int_\Omega \log \frac{p(x)}{p'(x)} p(x)dx \tag{18}$$

is called the *Kullback-Leibler divergence* or *relative entropy*. In our situation, we could instead of (16), consider (18) as the performance function; explicitly this would be:

$$W = \frac{1}{2}\int_0^b \int_0^b g(x,y) log \frac{g(x,y)}{K(x,y)} dxdy \tag{19}$$

with $K(x,y) = (C(x,y)V_k + h(V_k)B_n(x,y)$.

References

[1] S-I. Amari. **Differential Geometrical Methods in Statistics** Springer Lecture Notes in Statistics 28, Springer-Verlag, Berlin 1985.

[2] Khadiga Arwini and C.T.J. Dodson. Information geometric neighbourhoods of randomness and geometry of the McKay bivariate gamma 3-manifold. Preprint (2002).

[3] C.T.J. Dodson. Gamma manifolds and stochastic geometry. In: **Proceedings of the Workshop on Recent Topics in Differential Geometry**, Santiago de Compostela 16-19 July 1997. *Public. Depto. Geometría y Topología* 89 (1998) 85-92.

[4] C.T.J. Dodson and W.W. Sampson. Modeling a class of stochastic porous media. *Appl. Math. Lett.* 10, 2 (1997) 87-89.

[5] R.A. Fisher. Theory of statistical estimation. *Proc. Camb. Phil. Soc.* 122 (1925) 700-725.

[6] C.T.J. Dodson and H. Wang. Iterative approximation of statistical distributions and relation to information geometry. *J. Statistical Inference for Stochastic Processes* 147, (2001) 307-318.

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

[8] Y.Sato, K.Sugawa and M.Kawaguchi. The geometrical structure of the parameter space of the two-dimensional normal distribution. Division of information engineering, Hokkaido University, Sapporo, Japan, 1977.

[9] H. Wang. Neural network based control for output probability density functions for nonlinear stochastic systems and its applications. In Proc. $\mathbf{6}^{th}$ **European Congress on Intelligent Techniques and Soft Computing** Aachen, 1998.

[10] H. Wang. Control of the output probability density functions for a class of nonlinear stochastic systems. In Proc. **IFAC Workshop on Algorithms and Architectures for Real Time Control** 1998, pp 113-117.

[11] H. Wang. Detect unexpected changes of particle size distribution in paper-making white water system. In Proc. **IFAC Fault Detection and Diagnosis** Lyon, 1998, pp. 78-84.

[12] H. Wang **Bounded Dynamic Stochastic Systems: Modelling and Control**, Springer-Verlag London, 2000.

[13] H. Wang. A model approximation and control of output probability density functions for bounded dynamic stochastic systems, *Trans. Inst. Measu., and Contr*, in press 2003.