Comparing Distance Measures Between Bivariate Gamma Processes

Khadiga Arwini[1], C.T.J. Dodson[1], S. Felipussi[2] and J. Scharcanski[2]
[1]School of Mathematics, University of Manchester, Manchester M60 1QD, UK
[2]Departamento de Informatica Aplicada, Instituto de Informatica
Universidade Federal do Rio Grande do Sul, Porto Alegre, RS, Brasil

May 28, 2005

Abstract

Yue et al. [Yue et al. 2001]recently reviewed various bivariate gamma distribution models and concluded that they will be useful in hydrology. Here we contribute a detailed study of the McKay bivariate gamma distribution and demonstrate its applicability to the joint probability distribution of void and capillary sizes obtained from soil tomography. The information geometry of the space of McKay bivariate gamma distributions provides a useful mechanism for discriminating between bivariate stochastic processes with positive covariance and gamma marginal distributions. In most cases we found that the information-theoretic metric is more sensitive than the classical Bhattacharyya distance or the Kullback-Leibler divergence; this finding persisted also for data from model porous media, and for data from simulations.

Keywords: stochastic process, porous media, parameter estimation, hydrology, tomography, soil treatments, structure, data, simulations, information geometry, difference measures, gamma distribution, marginal distribution, McKay bivariate gamma, correlation, maximum-likelihood, Kullback-Leibler divergence, Bhattacharyya distance

1 Introduction

The present article adds to the study of Yue et al. [Yue et al. 2001], which reviewed various bivariate gamma distributions that are constructed from gamma marginals and concluded that bigamma distribution models will be useful in hydrology. Here we study the McKay bivariate gamma distribution, which has positive covariance, and demonstrate its applicability to the joint probability distribution of adjacent void and capillary sizes in soils; in this context we compare the discriminating power of an information theoretic metric with two classical distance functions in the space of probability distributions. We believe that similar methods may be applicable elsewhere in hydrology, to characterize stochastic structures of porous media and to model correlated flow variables. Moreover, we have elsewhere studied the information geometry of commonly used bivariate Gaussian and bivariate exponential distributions [Arwini 2004, Arwini and Dodson 2005] and these may have relevance in other stochastic hydrological processes. Phien [Phien 1993] considered the distribution of the storage capacity of reservoirs with gamma inflows that are either independent or first-order autoregressive and our methods may have relevance in modelling and quantifying correlated inflow processes. Govindaraju and Kavvas [Govindaraju and Kavvas 1992] used gamma or Gaussian distributions to model rill depth and width at different spatial locations and again an information geometric approach using a bivariate gamma or Gaussian model may be useful in further probing the joint behavior of these rill geometry variables.

2 Gamma distributions and randomness

The family of gamma probability density functions is given by

$$\{p(x;\beta,\alpha)=\left(\frac{\alpha}{\beta}\right)^{\alpha}\frac{x^{\alpha-1}}{\Gamma(\alpha)}\,e^{-\frac{\alpha}{\beta}x}|\alpha,\beta\in\mathbb{R}^{+}\},\quad x\in\mathbb{R}^{+} \tag{2.1}$$

so the space of parameters is topologically $\mathbb{R}^+ \times \mathbb{R}^+$. It is an exponential family and it includes as a special case ($\alpha = 1$) the exponential distribution itself, which complements the Poisson process on a line. It is pertinent to our interests that the property of having sample standard deviation independent of the mean actually characterizes gamma distributions, as shown recently [Hwang and Hu 1999]. They proved, for $n \geq 3$ independent positive random variables $x_1, x_2, \ldots, x_n$ with a common continuous probability density function f, that having independence of the sample mean $\bar{x}$ and sample coefficient of variation $cv = S/\bar{x}$ is equivalent to f being a gamma distribution. Of course, the exponential distribution has unit coefficient of variation.

The univariate gamma distribution is widely used to model processes involving a continuous positive random variable, for example, in hydrology the inflows of reservoirs [Phien 1993] and the depth and width of rills [Govindaraju and Kavvas 1992]. The information geometry of gamma distributions is known and has been applied recently to represent and metrize departures from randomness of, for example, the processes that allocate gaps between occurrences of each amino acid along a protein chain within the *Saccharomyces cerevisiae* genome [Cai et al. 2002], clustering of galaxies and communications, [Dodson 1999, Dodson 2000, Dodson 2001]. We have made precise and proved the statement that around every random process on the real line there is a neighborhood of processes governed by the gamma distribution, so gamma distributions can approximate any small enough departure from randomness [Arwini and Dodson 2004]. Such results are, by their topological nature, stable under small perturbations of a process, which is important in real applications. This, and their uniqueness property [Hwang and Hu 1999], gives confidence in the use of gamma distributions to model near random processes. Moreover, the information-theoretic heritage of the metric for the neighborhoods lends significance to the result.

2.1 Bivariate gamma processes

It is logical next to consider bivariate processes which may depart from independence and from randomness. Two natural choices arise for marginal distributions: Normal or log-Normal distributions and gamma distributions. For example, recently in hydrology, bivariate gamma distributions have been reviewed [Yue et al. 2001], and from [Govindaraju and Kavvas 1992] we may expect that rill depth and width admit bivariate gamma or bivariate Gaussian models with positive covariance. In this paper we concentrate on the case when the marginals are gamma and the covariance is positive, which has application to the modelling of void and capillary size in porous media like soils.

Positive covariance and gamma marginals gives rise to one of the earliest forms of the bivariate gamma distribution, due to Mckay [McKay 1934], defined by the density function

$$f(x,y) = \frac{c^{(\alpha_1+\alpha_2)} x^{\alpha_1 - 1}(y-x)^{\alpha_2 - 1} e^{-cy}}{\Gamma(\alpha_1)\Gamma(\alpha_2)} \text{ defined on } y > x > 0\ ,\ \alpha_1, c, \alpha_2 > 0 \tag{2.2}$$

One way to view this is that $f(x,y)$ is the probability density for the two random variables X and $Y = X + Z$ where X and Z both have gamma distributions. The marginal distributions of X and Y are gamma with shape parameters α_1 and $\alpha_1 + \alpha_2$, respectively. The covariance Cov and correlation coefficient ρ_M of X and Y are given by :

$$Cov(X,Y) = \frac{\alpha_1}{c^2} = \sigma_{12} \tag{2.3}$$

$$\rho_M(X,Y) = \sqrt{\frac{\alpha_1}{\alpha_1 + \alpha_2}}. \tag{2.4}$$

Observe that in this bivariate distribution the covariance, and hence correlation, tends to zero only as α_1 tends to zero.

3 Mckay bivariate gamma 3-manifold

We consider the Mckay bivariate gamma model as a 3-manifold, equipped with Fisher information as Riemannian metric. The classical family of Mckay bivariate gamma distributions is given by:

$$f(x,y;\alpha_1,\sigma_{12},\alpha_2) = \frac{(\frac{\alpha_1}{\sigma_{12}})^{\frac{(\alpha_1+\alpha_2)}{2}} x^{\alpha_1-1}(y-x)^{\alpha_2-1} e^{-\sqrt{\frac{\alpha_1}{\sigma_{12}}}y}}{\Gamma(\alpha_1)\Gamma(\alpha_2)}\ , \tag{3.5}$$

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. The correlation coefficient and marginal functions, of X and Y re given by :

$$\rho_M(X,Y) = \sqrt{\frac{\alpha_1}{\alpha_1+\alpha_2}} \tag{3.6}$$

$$f_X(x) = \frac{(\frac{\alpha_1}{\sigma_{12}})^{\frac{\alpha_1}{2}} x^{\alpha_1-1} e^{-\sqrt{\frac{\alpha_1}{\sigma_{12}}}x}}{\Gamma(\alpha_1)}, \quad x>0 \tag{3.7}$$

$$f_Y(y) = \frac{(\frac{\alpha_1}{\sigma_{12}})^{\frac{(\alpha_1+\alpha_2)}{2}} y^{(\alpha_1+\alpha_2)-1} e^{-\sqrt{\frac{\alpha_1}{\sigma_{12}}}y}}{\Gamma(\alpha_1+\alpha_2)}, \quad y>0 \tag{3.8}$$

Note that it is not possible to choose parameters such that both marginal functions are exponential.

Proposition 3.1 *Let M be the set of Mckay bivariate gamma distributions, that is*

$$\begin{aligned} M = \{&f | f(x,y;\alpha_1,\sigma_{12},\alpha_2) = \frac{(\frac{\alpha_1}{\sigma_{12}})^{\frac{(\alpha_1+\alpha_2)}{2}} x^{\alpha_1-1}(y-x)^{\alpha_2-1} e^{-\sqrt{\frac{\alpha_1}{\sigma_{12}}}y}}{\Gamma(\alpha_1)\Gamma(\alpha_2)}\ , \\ &y>x>0,\ \alpha_1,\sigma_{12},\alpha_2>0\} \end{aligned} \tag{3.9}$$

Then we have :

1. *Identifying $(\alpha_1, \sigma_{12}, \alpha_2)$ as a local coordinate system, M is a 3-manifold.*
2. *M is a Riemannian 3-manifold with Fisher information metric $G = [g_{ij}]$ given by :*

$$[g_{ij}] = \begin{bmatrix} \frac{-3\,\alpha_1+\alpha_2}{4\,{\alpha_1}^2} + \psi'(\alpha_1) & \frac{\alpha_1-\alpha_2}{4\,\alpha_1\,\sigma_{12}} & -\frac{1}{2\,\alpha_1} \\ \frac{\alpha_1-\alpha_2}{4\,\alpha_1\,\sigma_{12}} & \frac{\alpha_1+\alpha_2}{4\,{\sigma_{12}}^2} & \frac{1}{2\,\sigma_{12}} \\ -\frac{1}{2\,\alpha_1} & \frac{1}{2\,\sigma_{12}} & \psi'(\alpha_2) \end{bmatrix} \tag{3.10}$$

where $\psi(\alpha_i) = \frac{\Gamma'(\alpha_i)}{\Gamma(\alpha_i)} \quad (i=1,2)$.

4 Distance estimates in the McKay manifold

Distance between a pair of points in a Riemannian manifold is defined as the infimum of arc lengths over all curves between the points. For sufficiently nearby pairs of points there will be a unique minimizing geodesic curve that realises the infimum arc length. In general, such curves are hard to find between more distant points. However, we can obtain an upper bound on distances between two points $T_0, T_1 \in M$ by taking the sum of arc lengths along coordinate curves that triangulate the pair of points with respect to the coordinate axes. We adopted similar methods in the gamma manifold for univariate processes [Cai et al. 2002, Dodson and Scharcanski 2003]. Here we use the metric for the McKay manifold (3.10) and obtain the following upper bound on the information metric distance in M from T_0 with coordinates $(\alpha_1, \sigma_{12}, \alpha_2) = (A, B, C)$ to T_1 with coordinates $(\alpha_1, \sigma_{12}, \alpha_2) = (a, b, c)$

$$d_M(T_0,T_1) \le \left| \int_A^a \sqrt{\left| \frac{C-3x}{4x^2} + \frac{d^2\log\Gamma(x)}{dx^2} \right|}\, dx \right| + \left| \int_C^c \sqrt{|\frac{d^2\log\Gamma(y)}{dy^2}|}\, dy \right| + \left| \int_B^b \sqrt{\frac{A+C}{4\,z^2}}\, dz \right| \ . \tag{4.11}$$

The square roots arise from the norms of tangent vectors to coordinate curves and it is difficult to obtain the closed form solution for information distance, d_M. However, by removing the square roots

the integrals yield information-energy values E_M, which can be evaluated analytically. Then the square root of the net information-energy differences along the coordinate curves gives a closed analytic 'energy-distance' $dE_M = \sqrt{E_M}$, which we can compare with d_M. The net information-energy differences along the coordinate curves from T_0 with coordinates $(\alpha_1, \sigma_{12}, \alpha_2) = (A, B, C)$ to T_1 with coordinates $(\alpha_1, \sigma_{12}, \alpha_2) = (a, b, c)$ is

$$\begin{aligned} E_M(T_0, T_1) &\leq \left| \int_A^a \left(\frac{C-3x}{4x^2} + \frac{d^2 \log \Gamma(x)}{dx^2} \right) dx \right| + \left| \int_C^c \frac{d^2 \log \Gamma(y)}{dy^2}\, dy \right| + \left| \int_B^b \frac{A+C}{4\,z^2}\, dz \right| \\ &\leq \left| \int_A^a \left(\frac{C-3x}{4x^2} \right) dx \right| + \left| \int_A^a \frac{d^2 \log \Gamma(x)}{dx^2}\, dx \right| + \left| \int_C^c \frac{d^2 \log \Gamma(y)}{dy^2}\, dy \right| + \frac{A+C}{4} \left| \frac{1}{b} - \frac{1}{B} \right| \\ &= \left| \frac{C}{4\,a} - \frac{C}{4\,A} + \frac{3\,\log(\frac{a}{A})}{4} \right| + |\psi(a) - \psi(A)| + |\psi(c) - \psi(C)| + \frac{A+C}{4} \left| \frac{1}{b} - \frac{1}{B} \right| \quad . \quad (4.12) \\ dE_M &= \sqrt{E_M}. \quad (4.13) \end{aligned}$$

where $\psi = \frac{\Gamma'}{\Gamma}$ is the digamma function.

Next we compare distances between bivariate gamma distributions obtained using this information metric upper bound (4.11) in the McKay manifold metric (3.10) with the classical Bhattacharyya distance [Bhattacharyya 1943] between the distributions. Some further discussion of classical distance measures can be found in Chapter 3 of Fukunga [Fukunga 1991]; the Bhattacharyya distance is actually a special case of the Chernoff distance [Fukunga 1991].

The Bhattacharyya distance from T_0 to T_1 defined on $0 < x < y < \infty$ is given by

$$\begin{aligned} d_B(T_0, T_1) &= -\log \int_{y=0}^{\infty} \int_{x=0}^{y} \sqrt{T_0\, T_1}\; dx\, dy \\ &= -\log \left(\frac{\Gamma(\frac{A+a}{2}) \Gamma(\frac{C+c}{2}) (\frac{A}{B})^{\frac{A+C}{4}} (\frac{a}{b})^{\frac{a+c}{4}} (\frac{\sqrt{a}}{2\sqrt{b}} + \frac{\sqrt{A}}{2\sqrt{B}})^{\frac{-1}{2}(A+C+a+c)}}{\sqrt{\Gamma(A)\,\Gamma(C)\,\Gamma(a)\,\Gamma(c)}} \right). \quad (4.14) \end{aligned}$$

The Kullback-Leibler 'distance' [Kullback 1959] or 'relative entropy' from T_0 to T_1 defined on $0 < x < y < \infty$ is given by

$$\begin{aligned} KL(T_0, T_1) &= \int_{y=0}^{\infty} \int_{x=0}^{y} T_0 \log \frac{T_0}{T_1}\; dx\, dy \\ &= -A + \psi(A)\; (A-a) - C + \psi(C)\; (C-c) + \log \left(\frac{\Gamma(a)\,\Gamma(c)}{\Gamma(A)\,\Gamma(C)} \right) \\ &+ \frac{(a+c)}{2}\; \log \left(\frac{A\,b}{a\,B} \right) + (A+C) \sqrt{\frac{a\,B}{A\,b}}\,. \quad (4.15) \end{aligned}$$

and we symmetrize this to give a true distance

$$d_K(T_0, T_1) = \frac{KL(T_0, T_1) + KL(T_1, T_0)}{2}. \quad (4.16)$$

5 Applications to porous media

We apply our distance measures to experimental porous media data obtained from tomographic images of soil, and to data from model porous media and to data drawn by computer from bivariate correlated gamma processes.

Structural characterization and classification of porous materials has attracted the attention of researchers in different application areas, because of its great economic importance. For example, problems related to mass transfer and retention of solids in multi-phase fluid flow through stochastic porous materials are

ubiquitous in different areas of chemical engineering. One application of gamma distributed voids to stochastic porous media has admitted a direct statistical geometric representation of stochastic fibrous networks [Dodson and Sampson 1997] and their fluid transfer properties [Dodson and Sampson 2000]. Agricultural engineering is one of the sectors that has received attention recently, mostly because of the changing practices in agriculture in developing countries, and in developed countries, with great environmental impact [Vogel and Kretzchmar 1996], volgel:03.

Phenomenologically, mass transfer in porous media depends strongly on the morphological aspects of the media—such as the shape and size of pores, and depends also on the topological attributes of these media, such as the pore network connectivity [Dodson and Sampson 2000].

Several approaches have been presented in the literature for structural characterization of porous media, involving morphological and topological aspects. [Anselmetti et al.1998] proposed the analysis of porous media sections for their pore shape and size distributions. In their work, images are obtained using a scanning electron microscope for micro-structural characterization, and an optical microscope for macro-structural characterization. However, the acquisition of samples for the analysis is destructive, and it is necessary to slice the porous media so that sections can be obtained, and then to introduce epoxy resin for contrast. These procedures influence the structure of the media solid phase and consequently the morphology and topology of the porous phase, which implies that the three-dimensional reconstruction is less reliable for soil samples. In order to overcome similar difficulties, a few years earlier [Biassusi 1996] the non-destructive testing in soil samples using tomographic images was proposed, but their goal was the evaluation of the swelling and shrinkage properties of loamy soil. Also, other researchers have concentrated on the 'fingering' phenomenon occurring during fluid flow in soils [Onody et al. 1995]. More recently, researchers have proposed geometrical and statistical approaches for porous media characterization. A skeletonization method based on the Voronoi diagram [Delrue et al. 1999] was introduced to estimate the distributions of local pore sizes in porous media, see also [Ketcham Iturrino 2005].

The statistical characterization and classification of stochastic porous media, is essential for the simulation and/or prediction of the mass transfer properties of a particular stochastic medium. Much work has been done on the characterization of porous media but the discrimination of between different models from observed data still remains a challenging issue. This is particularly true considering the tomographic images of porous media, often used in soil analysis; for two recent studies see [Al-Raoush and Willson 2005] and [Ketcham Iturrino 2005].

It turns out that tomographic images of soil structure reveal a bivariate stochastic structure of sizes of voids and their interconnecting capillaries. The information geometry of the Riemannian 3-manifold of McKay bivariate gamma distributions provides a useful mechanism for discriminating between treatments of soil. This method is more sensitive than that using the classical Bhattacharyya distance between the bivariate gamma distributions and in most cases better than the Kullback-Leibler measure for distances between such distributions.

5.1 Comparison of analyses of tomographic images of porous media

Three-dimensional tomographic images were obtained from thin slices of soil and model samples and new algorithms were used to reveal features of the pore size distribution, and its connectivity. These features are relevant for the quantitative analysis of samples in several applications of economic importance. The methodology is applicable generally to stochastic porous media but here we focus on the analysis of soil samples, in terms of the soil compaction resulting from different soil preparation techniques. The interconnectivity of the pore network is analyzed through a fast algorithm that simulates flow. The image analysis methods employed to extract features from these images are beyond the scope of this paper, and will be discussed elsewhere.

The two variables $0 < x < y < \infty$ correspond as follows: y represents the cross-section area of a pore in the soil and x represents the corresponding cross-sectional area of the throats or capillaries that connect it to neighboring voids. It turns out that these two variables have a positive covariance and can be fitted quite well to the bivariate gamma distribution (3.5). The maximum likelihood parameters $(\alpha_1, \sigma_{12}, \alpha_2)$ for the data are shown in Table 1, together with the McKay correlation coefficients, ρ_M and the measured correlation ρ_{Data}. In these experiments, we used tomographic images of soil samples, and packings of spheres of different sizes as phantoms. The soil samples were selected from untreated (i.e. forest soil type), and treated (i.e. conventional mechanized cultivated soil, and direct plantation cultivated soil).

The image analysis methods employed to measure the pore and throat size distributions in these images are out of the scope of this paper, and will be discussed elsewhere.

Typical scatterplots of the throat area and pore area data are shown in Figure 1 for the untreated soil forest A, which shows strong correlation, and in Figure 2 for the model structure 1 made from beds of spheres, which shows weak correlation.

We see from Table 1 that the theoretical McKay correlation agrees well with that found experimentally. The four distance functions d_M, dE_M, d_B, d_K are given for the four soil treatments. The information metric d_M is the most discriminating and the Bhattacharyya distance d_B is the least discriminating. Over all treatments the grand means for distances from untreated (forest) soils are respectively 2.423, 1.302, 0.474, 2.113, for d_M, dE_M, d_B, d_K. The distance measures are found also for the model structures of spheres and the simulation, Table 2, but here the experimental correlation is much less than that expected for the McKay distribution.

The soil results from Table 1 are shown in Figure 3. The first two plots use the information distance d_M and energy-distance dE_M bounds (4.11, 4.13 respectively) for the McKay manifold metric (3.10), the other two plots use the Bhattacharyya distance d_B (4.14) and the Kullback-Leibler distance d_K (4.14) between the corresponding bivariate gamma distributions. The base plane $d = 0$ represents the natural or forest soil; data is in pairs, two points of the same size correspond to the same soils with two treatments. Importantly, the information metric, d_M, is mainly the most discriminating between the treated and untreated soils—the points being generally highest in the graphic for d_M, though Soil D direct treatment has a particularly high d_K value. Except for Sample A, all distances agree on the ranking of effects of treatments.

The sphere packing model results and the simulation results from Table 2 are shown in Figure 4. Note that the McKay bivariate gamma distribution does not contain the case of both marginal distributions being exponential nor the case of zero covariance—both of these are limiting cases and so cannot serve as base points for distances. Thus, there is no natural reference structure from which to base distances in these model results so here the distances shown are measured from the mean values of the three parameters taken over the spheres. We note that the distances obtained are in each case ordered: $d_B < dE_M < d_K < d_M$; they all agree on ranking of the effects of conditions.

5.2 Computer simulation of bivariate gamma processes

Four sets of 5000 pairs (x_i, y_i) with $0 < x_i < y_i$ were drawn by computer from gamma distributions, with different parameters and with varying positive covariance between the two variables, Figure 5. This data was analyzed and from it maximum likelihood fits were made of McKay bivariate gamma distributions. Table 3 summarizes the parameters, and the distances measured between the corresponding points are shown in Tables 4, 5, 6 and 7. These experiments confirm that data sets 1 and 2 are more similar to each other, than to the data sets 3 and 4, which is verified by visual inspection of the scatterplots shown in Figure 5. If we consider that data sets 1 and 2 form one cluster, and that data sets 3 and 4 form another cluster, it is important to verify how the distance measures we are comparing perform in terms of data discrimination. Table 8 shows the ratios between the mean inter and intra cluster distances, indicating that the best data separability is obtained by dE_M^{sym} and d_M^{sym} [1].

6 Conclusions

The recently derived Riemannian geometry of the 3-manifold of McKay bivariate gamma distributions provides a useful mechanism for quantitative discrimination between bivariate processes with random variables $0 < x < y$ and positive covariance. We applied this to some tomographic images of soil structure and to similar images of model beds of spheres; these media exhibit a bivariate positively correlated stochastic spatial process for the sizes of voids and their interconnecting capillaries. We further illustrated the methodology using some simulation data.

In most cases, the information geometry, which uses a maximum likelihood metric, is more discriminating than the classical Bhattacharyya distance, or the Kullback-Leibler divergence, between pairs of bivariate

[1] Given a pair of data sets A and B, the distance bounds dE_M and d_M are made symmetric through the average over both directions, i.e. $d^{sym} = \frac{1}{2}(d(A,B) + d(B,A))$, which are denoted here by dE_M^{sym} and d_M^{sym}.

gamma distributions. We have also available the information geometry of bivariate Gaussian and bivariate exponential distributions and we expect that our methodology may have other applications in the modelling of bivariate stochastic processes in hydrology.

Acknowledgement

The authors wish to thank University of Manchester, UFRGS, The British Council and CNPq for hospitality and support of Dodson and Scharcanski during exchange visits; thanks are due also to the Libyan Ministry of Education for a scholarship for Arwini.

References

[Al-Raoush and Willson 2005] Al-Raoush, R.I. and Willson, C. S. 2005. Extraction of physically realistic pore network properties from three-dimensional synchrotron X-ray microtomography images of unconsolidated porous media systems. Moving through scales of flow and transport in soil. *Journal of Hydrology*, 300, 1-4, 44-64.

[Anselmetti et al.1998] Flavio S. Anselmetti, Luthi, Stefan and Eberli, Gregor P. 1998. Quantitative Characterization of Carbonate Pore Systems by Digital Image Analysis. *AAPG Bulletin*, 82, 10, 1815-1836.

[Arwini 2004] Arwini, Khadiga. 2004. *Differential geometry in neighbourhoods of randomness and independence Phd Thesis*, Department of Mathematics, University of Manchester, Institute of Science and Technology, Manchester, U.K.

[Arwini and Dodson 2004] Arwini, Khadiga and Dodson, C.T.J. 2004. Information geometric neighbourhoods of randomness and geometry of the McKay bivariate gamma 3-manifold. In press *Sankhya: Indian Journal of Statistics*. 66, 2, 211-231.

[Arwini and Dodson 2005] Arwini, Khadiga and Dodson, C.T.J. 2005. Neighbourhoods of independence and associated geometry. Preprint 2005.
http://www.ma.umist.ac.uk/kd/PREPRINTS/nhdindep.pdf

[Bhattacharyya 1943] Bhattacharyya, A. 1943. On a measure of divergence between two statistical populations defined by their distributions. *Bull. Calcutta Math. Soc.* 35, 99-110.

[Biassusi 1996] Biassusi, Marcelo. 1996. *Estudo da Deformacao de um Vertissolo Atraves da Tomografia Computadorizada de Dupla Energia Simultanea. Phd Thesis*, UFRGS - Federal University of Rio Grande do Sul, Porto Alegre, Brazil.

[Cai et al. 2002] Cai, Y, Dodson, C.T.J., Wolkenhauer, O. and Doig, A.J. 2002. Gamma Distribution Analysis of Protein Sequences shows that Amino Acids Self Cluster. *J. Theoretical Biology* 218, 4, 409-418.

[Delrue et al. 1999] Delrue, J. F., Perrier, E., Yu, Z. Y. and Velde, B. 1999. New Algorithms in 3D Image Analysis and Their Application to the Measurement of a Spatialized Pore Size Distribution in Soils. *Phys. Chem. Earth*, 24, 7, 639-644.

[Dodson 1999] Dodson, C.T.J. 1999. Spatial statistics and information geometry for parametric statistical models of galaxy clustering. *Int. J. Theor. Phys.*, 38, 10, 2585-2597.

[Dodson 2000] Dodson, C.T.J. 2000. Information geodesics for communication clustering. *J. Statistical Computation and Simulation* 65, 133-146.

[Dodson 2001] Dodson, C.T.J. 2001. Geometry for stochastically inhomogeneous spacetimes. *Nonlinear Analysis*, 47, 2951-2958.

[Dodson and Sampson 1997] Dodson, C.T.J. and W.W. Sampson, W.W. 1997. Modeling a class of stochastic porous media. *Applied Mathematics Letters* 10, 2, 87-89.

[Dodson and Sampson 2000] Dodson, C.T.J. and W.W. Sampson, W.W. 2000. Flow simulation in stochastic porous media. *Simulation*, 74:6, 351-358.

[Dodson and Scharcanski 2003] Dodson, C.T.J. and Scharcanski, J. 2003. Information Geometric Similarity Measurement for Near-Random Stochastic Processes. *IEEE Transactions on Systems, Man and Cybernetics - Part A,* 33, 4, 435-440.

[Fukunga 1991] Fukunga, K. 1991. *Introduction to Statistical Pattern Recognition*, 2^{nd} Edition, Academic Press, Boston.

[Govindaraju and Kavvas 1992] Govindaraju, Rao S. and Kavvas, M. Levent. 1992. Characterization of the rill geometry over straight hillslopes through spatial scales. *Journal of Hydrology*, 130, 1, 339-365.

[Hwang and Hu 1999] Hwang, T-Y. and Hu, C-Y. 1999. On a characterization of the gamma distribution: The independence of the sample mean and the sample coefficient of variation. *Annals Inst. Statist. Math.* 51, 4, 749-753.

[Ketcham Iturrino 2005] Ketcham, R.A and Iturrino, Gerardo J. 2005. Nondestructive high-resolution visualization and measurement of anisotropic effective porosity in complex lithologies using high-resolution X-ray computed tomography. *Journal of Hydrology*, 302, 1-4, 92-106.

[Kullback 1959] Kullback, S. 1959. *Information and Statistics*, J. Wiley, New York, 1959.

[McKay 1934] McKay, A.T. 1934. Sampling from batches. *J. Royal Statist. Soc.* 2, 207-216.

[Onody et al. 1995] Onody, R. N., Posadas, A.N.D. and Crestana, S. 1995. Experimental Studies of the Fingering Phenomena in Two Dimensions and Simulation Using a Modified Invasion Percolation Model. *Journal of Applied Physics*, v. 78, n. 5, 2970-2976.

[Phien 1993] Phien, Huynh Ngoc. 1993. Reservoir storage capacity with gamma inflows. *Journal of Hydrology*, 146, 1, 383-389.

[Vogel and Kretzchmar 1996] Vogel, H. J. and Kretzchmar, A. 1996. *Topological Characterization of Pore Space in Soil-Sample Preparation and Digital Image-Processing. Geoderma*, 73, 23-18.

[Vogel and Roth 2003] Vogel, H. J. and Roth, K. 2003. Moving through scales of flow and transport in soil. *Journal of Hydrology*, 272, 1-4, 95-106.

[Yue et al. 2001] Yue, S., Ouarda, T.B.M.J. and Bobée, B. 2001. A review of bivariate gamma distributions for hydrological application. *Journal of Hydrology*, 246, 1-4, 1-18.

Sample	α_1	α_2	σ_{12}	ρ_M	ρ_{Data}	d_M	dE_M	d_B	d_K
A forest	7.6249	3.581	19199.4	0.8249	0.8555	(4.1904)	(1.4368)	(0.1292)	(0.5775)
A conv	4.7931	7.4816	22631.2	0.6249	0.7725	2.8424	1.2643	0.7920	3.2275
A direct	1.8692	3.4911	33442.6	0.5905	0.5791	3.181	1.5336	0.5855	2.7886
B forest	1.2396	2.2965	41402.8	0.5920	0.5245	(4.1611)	(1.7835)	(0.3948)	(1.9502)
B conv	5.6754	4.8053	34612.2	0.7359	0.5500	3.7034	1.8784	0.3214	1.6816
B direct	1.3622	2.0074	30283.5	0.6358	0.5215	0.6322	0.5820	0.0346	0.1390
C forest	1.6920	2.6801	37538.3	0.6221	0.5582	(2.7858)	(1.8896)	(0.2140)	(1.0462)
C conv	0.7736	1.2488	30697.5	0.6185	0.5466	2.3931	1.5372	0.1727	0.7403
C direct	2.8476	2.6413	25840.8	0.7203	0.7975	1.1146	0.9518	0.0910	0.3741
D forest	1.8439	1.7499	15818.7	0.7163	0.6237	(6.1671)	(2.2155)	(1.6124)	(8.6557)
D conv	0.963	1.2533	26929.7	0.6592	0.5324	1.7529	1.3028	0.0526	0.2221
D direct	3.4262	9.9762	39626.4	0.5056	0.3777	5.5669	1.7660	2.3136	10.3993
E forest	2.7587	1.4647	26501.5	0.8082	0.7952	(1.4423)	(1.1962)	(0.0830)	(0.3519)
E conv	3.0761	2.4388	38516.1	0.7468	0.6799	1.3283	0.9314	0.1388	0.5609
E direct	1.4987	2.107	47630.9	0.6447	0.6400	1.8977	1.2728	0.2401	0.9932

Table 1: Maximum likelihood parameters of McKay bivariate gamma fitted to hydrological survey data extracted from tomograghic images of five soil samples, each with three treatments. For each soil, the natural state is untreated forest; two treatments are compared : conventional and direct. The distance functions d_M, dE_M, d_B, d_K are used to measure effects of treatments—values given are distances from the forest untreated case, except that values in brackets give the distances between the conventional and direct treatments. In most cases, d_M is most discriminating and dE_M is second best. Except for Sample A, all distances agree on the ranking of effects of treatments compared with untreated soils.

Sample	α_1	α_2	σ_{12}	ρ_M	ρ_{Data}	d_M	dE_M	d_B	d_K
1(2.4 a 3.3 mm)	1.0249	0.1469	54.8050	0.9341	0.3033	3.6369	1.9071	0.4477	2.9319
2(1.4 a 2.0 mm)	1.6789	0.5117	4.3863	0.8755	0.1714	2.3057	1.5185	0.4195	1.7888
3(1.0 a 1.7 mm)	2.1416	2.8396	0.7103	0.6557	0.1275	4.5013	2.1216	0.6751	4.1202
Simulation	0.1514	0.3185	4137	0.5676	0.1118	8.5318	2.9209	0.8539	11.4460

Table 2: Maximum likelihood parameters of McKay bivariate gamma fitted to data extracted from tomograghic images of model beds of spheres and a simulation. There is no reference structure from which to base distances so here the distances shown are measured from the mean values $(\overline{\alpha_1}, \overline{\sigma_{12}}, \overline{\alpha_2}) = (1.6151, 19.9672, 1.1661)$ of the three parameters taken over the spheres. We note that the distances obtained for the sphere models are in every case ordered: $d_B < dE_M < d_K < d_M$ and they agree on the ranking of effects of conditions. The simulation data, having $\alpha_1 << 1$, seems very different from the model sphere beds.

#	$\bar{x}$	$\bar{y}$	α_1	α_2	σ_{12}	ρ_M	ρ_{data}
1	0.9931	1.5288	1.0383	0.9174	0.9607	0.7286	0.9017
2	1.1924	1.7275	1.0176	0.7934	1.4117	0.7443	0.9304
3	2.9793	3.5151	1.0383	0.3598	8.5813	0.8618	0.9873
4	3.2072	3.7429	1.0165	0.3432	10.1006	0.8646	0.9892

Table 3: For the four simulated bivariate gamma data sets 1,2,3,4 we give the maximum likelihood McKay parameters.

dE_M^{sym}	sample set 2	sample set 3	sample set 4
sample set 1	1.2505	2.5489	2.2506
sample set 2	0.0000	2.1854	2.2988
sample set 3	0.0000	0.0000	0.2067

Table 4: Pairwise dE_M^{sym} distances for the four simulated bivariate gamma data sets 1,2,3,4.

d_M^{sym}	sample set 2	sample set 3	sample set 4
sample set 1	1.2625	2.4836	2.6511
sample set 2	0.0000	2.0532	2.1788
sample set 3	0.0000	0.0000	0.1653

Table 5: Pairwise d_M^{sym} distances for the four simulated bivariate gamma data sets 1,2,3,4.

d_K	sample set 2	sample set 3	sample set 4
sample set 1	0.6804	1.0368	1.1793
sample set 2	0.0000	0.7013	0.1828
sample set 3	0.0000	0.0000	0.0040

Table 6: Pairwise d_K distances for the four simulated bivariate gamma data sets 1,2,3,4.

d_B	sample set 2	sample set 3	sample set 4
sample set 1	0.1661	0.2254	0.2506
sample set 2	0.0000	0.1596	0.1815
sample set 3	0.0000	0.0000	0.0010

Table 7: Pairwise d_B distances for the four simulated bivariate gamma data sets 1,2,3,4.

dE_M^{sym}	d_M^{sym}	d_K	d_B
1.6734	1.6401	1.2225	1.1323

Table 8: Expected ratio between inter/intra cluster distances for the four simulated bivariate gamma data sets 1,2,3,4.

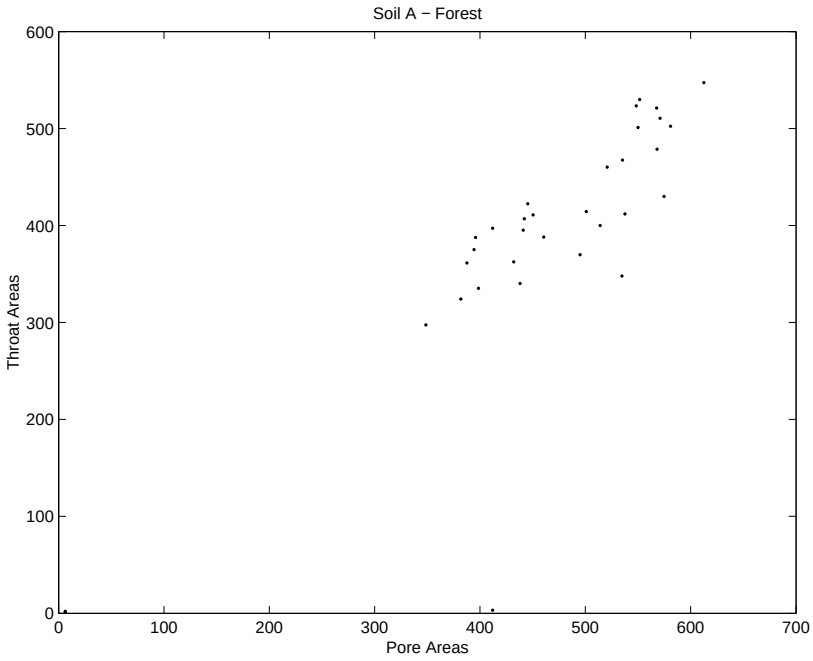

Figure 1: *Scatterplot of the data for untreated forest A soil sample, Table 1.*

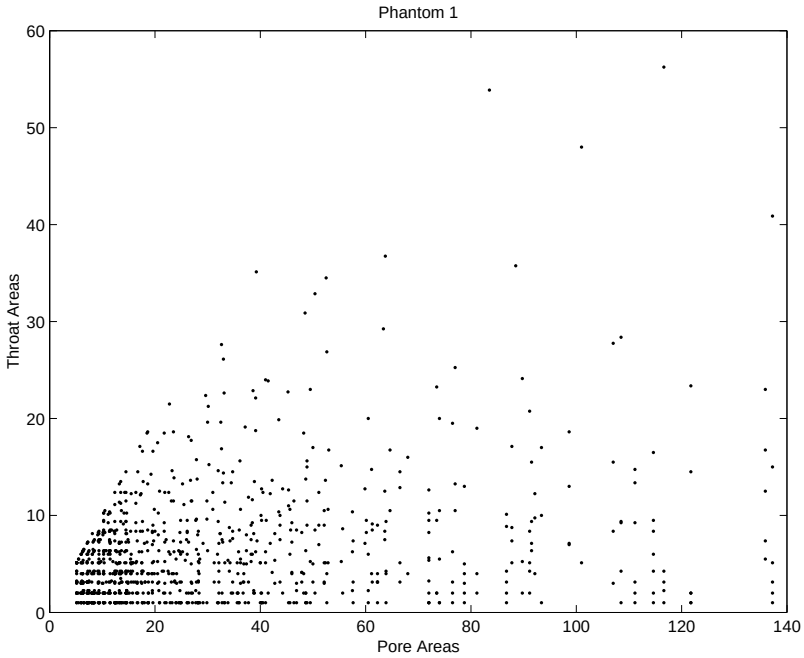

Figure 2: *Scatterplot of the data for model sphere bed structure Sample 1 of Table 2.*

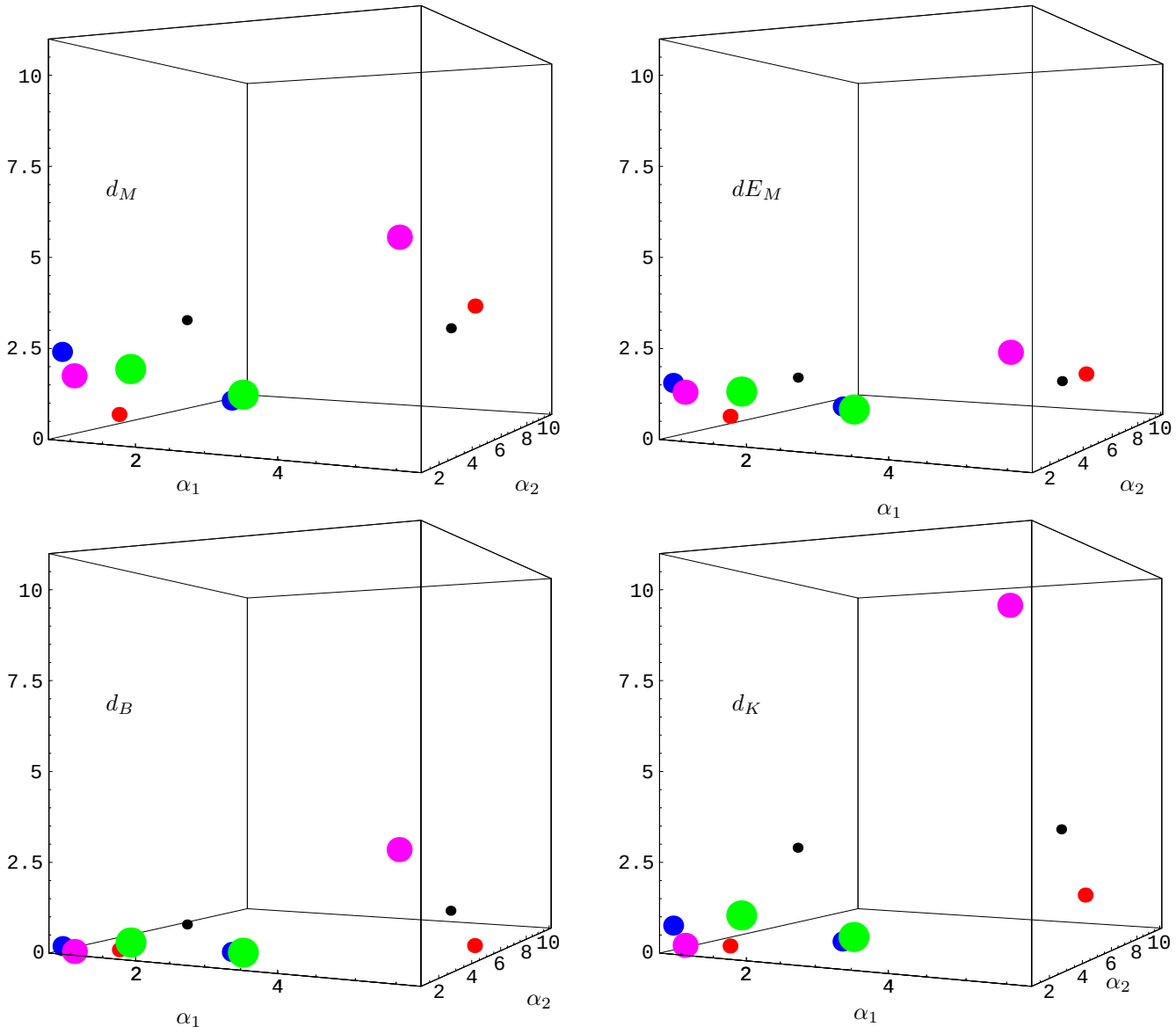

Figure 3: *Distances of 5 pairs of treated soil samples from the untreated forest soil, in terms of their porous structure from tomographic images, using the data from Table 1. Clockwise from the top left the plots use: the information theoretic bounds (4.11) d_M and root energy dE_M for the McKay manifold metric (3.10), the the Bhattacharyya distance (4.14) d_B and the Kullback-Leibler distance (4.14) d_K between the corresponding bivariate gamma distributions. The plane $d = 0$ represents the natural or forest soil; data is in pairs, two points of the same size correspond to the same soils with two treatments. The information metric, top left, is most discriminating.*

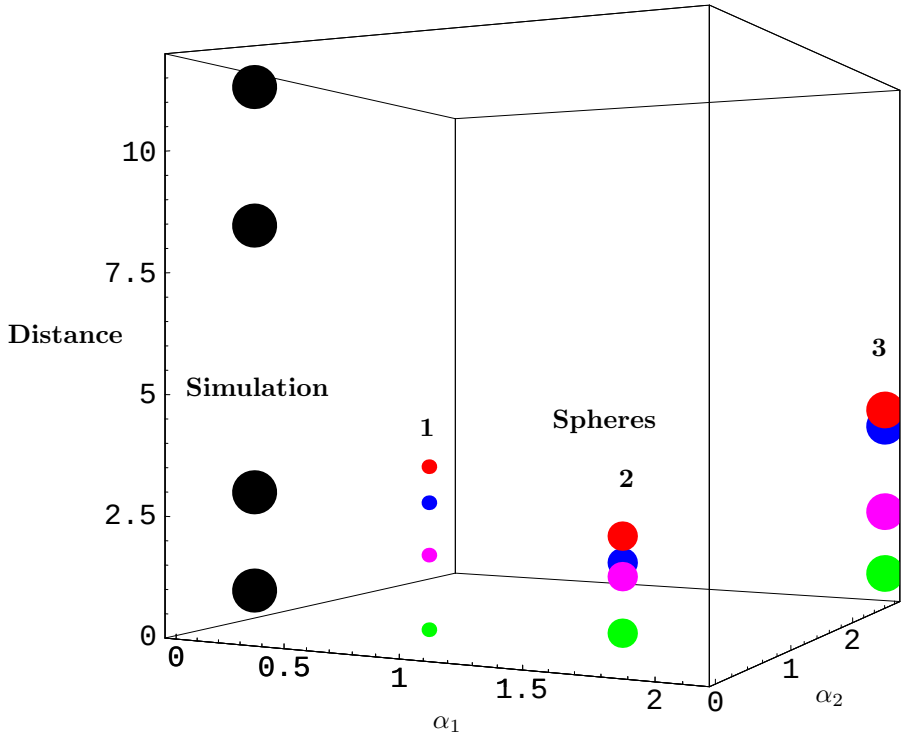

Figure 4: *Distances of model sphere samples and a simulation, measured from the average parameter values for the three sphere samples, using the data from Table 2. For the model sphere structures, the distances are in each case ordered:* $d_B < dE_M < d_K < d_M$*, with the exception of the simulation. The increasing point sizes refer to: the model sphere beds 1, 2, 3, respectively.*

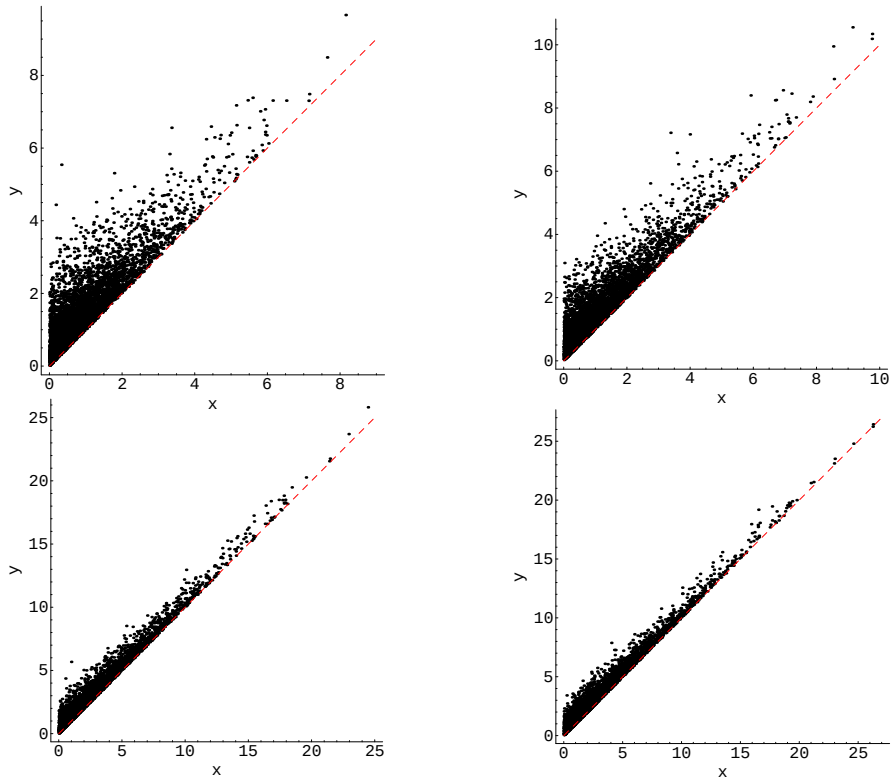

Figure 5: *Scatterplot of the data for computer simulations of the four positively correlated gamma processes with* $x < y$ *: top row # 1,2, second row # 3,4. Maximum likelihood McKay parameters are given in Table 3.*