Lecture Notes in Mathematics

Khadiga Arwini
C. T. J. Dodson

Information Geometry

1953

Near Randomness and Near Independence

Springer

Khadiga A. Arwini · Christopher T.J. Dodson

Information Geometry

Near Randomness and Near Independence

Springer

Authors

Khadiga A. Arwini
Al-Fateh University
Faculty of Sciences
Mathematics Department
Box 13496
Tripoli, Libya
arwini2001@yahoo.com

Christopher T.J. Dodson
University of Manchester
School of Mathematics
Manchester M13 9PL, United Kingdom
ctdodson@manchester.ac.uk

ISBN: 978-3-540-69391-8 e-ISBN: 978-3-540-69393-2
DOI: 10.1007/978-3-540-69393-2

Lecture Notes in Mathematics ISSN print edition: 0075-8434
ISSN electronic edition: 1617-9692

Library of Congress Control Number: 2008930087

Mathematics Subject Classification (2000): 53B50, 60D05, 62B10, 62P35, 74E35, 92D20

Cover design: SPi Publishing Services

Printed on acid-free paper

9 8 7 6 5 4 3 2 1

springer.com

Preface

The main motivation for this book lies in the breadth of applications in which a statistical model is used to represent small departures from, for example, a Poisson process. Our approach uses information geometry to provide a common context but we need only rather elementary material from differential geometry, information theory and mathematical statistics. Introductory sections serve together to help those interested from the applications side in making use of our methods and results. We have available *Mathematica* notebooks to perform many of the computations for those who wish to pursue their own calculations or developments.

Some 44 years ago, the second author first encountered, at about the same time, differential geometry via relativity from Weyl's book [209] during undergraduate studies and information theory from Tribus [200, 201] via spatial statistical processes while working on research projects at Wiggins Teape Research and Development Ltd—cf. the Foreword in [196] and [170, 47, 58]. Having started work there as a student laboratory assistant in 1959, this research environment engendered a recognition of the importance of international collaboration, and a lifelong research interest in randomness and near-Poisson statistical geometric processes, persisting at various rates through a career mainly involved with global differential geometry. From correspondence in the 1960s with Gabriel Kron [4, 124, 125] on his Diakoptics, and with Kazuo Kondo who influenced the post-war Japanese schools of differential geometry and supervised Shun-ichi Amari's doctorate [6], it was clear that both had a much wider remit than traditionally pursued elsewhere. Indeed, on moving to Lancaster University in 1969, receipt of the latest *RAAG Memoirs Volume 4 1968* [121] provided one of Amari's early articles on information geometry [7], which subsequently led to his greatly influential 1985 Lecture Note volume [8] and our 1987 *Geometrization of Statistical Theory Workshop* at Lancaster University [10, 59].

Reported in this monograph is a body of results, and computer-algebraic methods that seem to have quite general applicability to statistical models admitting representation through parametric families of probability density functions. Some illustrations are given from a variety of contexts for geometric characterization of statistical states near to the three important standard basic reference states: (Poisson) randomness, uniformity, independence. The individual applications are somewhat heuristic models from various fields and we incline more to terminology and notation from the applications rather than from formal statistics. However, a common thread is a geometrical representation for statistical perturbations of the basic standard states, and hence results gain qualitative stability. Moreover, the geometry is controlled by a metric structure that owes its heritage through maximum likelihood to information theory so the quantitative features—lengths of curves, geodesics, scalar curvatures etc.—have some respectable authority. We see in the applications simple models for galactic void distributions and galaxy clustering, amino acid clustering along protein chains, cryptographic protection, stochastic fibre networks, coupled geometric features in hydrology and quantum chaotic behaviour. An ambition since the publication by Richard Dawkins of *The Selfish Gene* [51] has been to provide a suitable differential geometric framework for dynamics of natural evolutionary processes, but it remains elusive. On the other hand, in application to the statistics of amino acid spacing sequences along protein chains, we describe in Chapter 7 a stable statistical qualitative property that may have evolutionary significance. Namely, to widely varying extents, all twenty amino acids exhibit greater clustering than expected from Poisson processes. Chapter 11 considers eigenvalue spacings of infinite random matrices and near-Poisson quantum chaotic processes.

The second author has benefited from collaboration (cf. [34]) with the group headed by Andrew Doig of the Manchester Interdisciplinary Biocentre, the University of Manchester, and has had long-standing collaborations with groups headed by Bill Sampson of the School of Materials, the University of Manchester (cf.eg. [73]) and Jacob Scharcanski of the Instituto de Informatica, Universidade Federal do Rio Grande do Sul, Porto Alegre, Brasil (cf.eg. [76]) on stochastic modelling. We are pleased therefore to have co-authored with these colleagues three chapters: titled respectively, Amino Acid Clustering, Stochastic Fibre Networks, Stochastic Porous Media and Hydrology.

The original draft of the present monograph was prepared as notes for short Workshops given by the second author at Centro de Investigaciones de Matematica (CIMAT), Guanajuato, Mexico in May 2004 and also in the Departamento de Xeometra e Topoloxa, Facultade de Matemáticas, Universidade de Santiago de Compostela, Spain in February 2005.

The authors have benefited at different times from discussions with many people but we mention in particular Shun-ichi Amari, Peter Jupp, Patrick Laycock, Hiroshi Matsuzoe, T. Subba Rao and anonymous referees. However, any overstatements in this monograph will indicate that good advice may have been missed or ignored, but actual errors are due to the authors alone.

Khadiga Arwini, Department of Mathematics, Al-Fateh University, Libya
Kit Dodson, School of Mathematics, the University of Manchester, England

Contents

1

Mathematical Statistics and Information Theory

There are many easily found good books on probability theory and mathematical statistics (eg [84, 85, 87, 117, 120, 122, 196]), stochastic processes (eg [31, 161]) and information theory (eg [175, 176]); here we just outline some topics to help make the sequel more self contained. For those who have access to the computer algebra package *Mathematica* [215], the approach to mathematical statistics and accompanying software in Rose and Smith [177] will be particularly helpful.

The word stochastic comes from the Greek *stochastikos*, meaning skillful in aiming and *stochazesthai* to aim at or guess at, and *stochos* means target or aim. In our context, stochastic colloquially means involving chance variations around some event—rather like the variation in positions of strikes aimed at a target. In its turn, the later word statistics comes through eighteenth century German from the Latin root *status* meaning state; originally it meant the study of political facts and figures. The noun random was used in the sixteenth century to mean a haphazard course, from the Germanic randir to run, and as an adjective to mean without a definite aim, rule or method, the opposite of purposive. From the middle of the last century, the concept of a random variable has been used to describe a variable that is a function of the result of a well-defined statistical experiment in which each possible outcome has a definite probability of occurrence. The organization of probabilities of outcomes is achieved by means of a probability function for discrete random variables and by means of a probability density function for continuous random variables. The result of throwing two fair dice and summing what they show is a discrete random variable.

Mainly, we are concerned with continuous random variables (here measurable functions defined on some $\mathbb{R}^n$) with smoothly differentiable probability density measure functions, but we do need also to mention the Poisson distribution for the discrete case. However, since the Poisson is a limiting approximation to the Binomial distribution which arises from the Bernoulli distribution (which everyone encountered in school!) we mention also those examples.

K. Arwini, C.T.J. Dodson, *Information Geometry*.
Lecture Notes in Mathematics 1953,

1.1 Probability Functions for Discrete Variables

For discrete random variables we take the domain set to be $\mathbb{N} \cup \{0\}$. We may view a probability function as a subadditive measure function of unit weight on $\mathbb{N} \cup \{0\}$

$$p : \mathbb{N} \cup \{0\} \to [0,1) \qquad \text{(nonnegativity)} \tag{1.1}$$

$$\sum_{k=0}^{\infty} p(k) = 1 \qquad \text{(unit weight)} \tag{1.2}$$

$$p(A \cup B) \leq p(A) + p(B), \ \forall A, B \subset \mathbb{N} \cup \{0\}, \qquad \text{(subadditivity)} \tag{1.3}$$
$$\text{with equality} \iff A \cap B = \emptyset.$$

Formally, we have a discrete measure space of total measure 1 with σ-algebra the power set and measure function induced by p

$$sub(\mathbb{N} \cup \{0\}) \to [0,1) : A \mapsto \sum_{k \in A} p(k)$$

and as we have anticipated above, we usually abbreviate $\sum_{k \in A} p(k) = p(A)$.

We have the following expected values of the random variable and its square

$$\mathcal{E}(k) = \overline{k} = \sum_{k=0}^{\infty} k\, p(k) \tag{1.4}$$

$$\mathcal{E}(k^2) = \overline{k^2} = \sum_{k=0}^{\infty} k^2\, p(k). \tag{1.5}$$

Formally, statisticians are careful to distinguish between a property of the whole population—such as these expected values—and the observed values of samples from the population. In practical applications it is quite common to use the bar notation for expectations and we shall be clear when we are handling sample quantities. With slight but common abuse of notation, we call $\overline{k}$ the mean, $\overline{k^2} - (\overline{k})^2$ the variance, $\sigma_k = +\sqrt{\overline{k^2} - (\overline{k})^2}$ the standard deviation and $\sigma_k / \overline{k}$ the coefficient of variation, respectively, of the random variable k. The variance is the square of the standard deviation.

The moment generating function $\Psi(t) = \mathcal{E}(e^{tX}),\ t \in \mathbb{R}$ of a distribution generates the r^{th} moment as the value of the r^{th} derivative of Ψ evaluated at $t = 0$. Hence, in particular, the mean and variance are given by:

$$\mathcal{E}(X) = \Psi'(0) \tag{1.6}$$

$$Var(X) = \Psi''(0) - (\Psi'(0))^2, \tag{1.7}$$

which can provide an easier method for their computation in some cases.

1.1.1 Bernoulli Distribution

It is said that a random variable X has a Bernoulli distribution with parameter p $(0 \leq p \leq 1)$ if X can take only the values 0 and 1 and the probabilities are

$$P_r(X = 1) = p \tag{1.8}$$

$$P_r(X = 0) = 1 - p \tag{1.9}$$

Then the probability function of X can be written as follows:

$$f(x|p) = \begin{cases} p^x(1-p)^{1-x} & \text{if } x = 0, 1 \\ 0 & \text{otherwise} \end{cases} \tag{1.10}$$

If X has a Bernoulli distribution with parameter p, then we can find its expectation or mean value $\mathcal{E}(X)$ and variance $Var(X)$ as follows.

$$\mathcal{E}(X) = 1 \cdot p + 0 \cdot (1 - p) = p \tag{1.11}$$

$$Var(X) = \mathcal{E}(X^2) - (\mathcal{E}(X))^2 = p - p^2 \tag{1.12}$$

The moment generating function of X is the expectation of e^{tX},

$$\Psi(t) = \mathcal{E}(e^{tX}) = pe^t + q \tag{1.13}$$

which is finite for all real t.

1.1.2 Binomial Distribution

If n random variables $X_1, X_2, \ldots, X_n$ are independently identically distributed, and each has a Bernoulli distribution with parameter p, then it is said that the variables $X_1, X_2, \ldots, X_n$ form n Bernoulli trials with parameter p.

If the random variables $X_1, X_2, \ldots, X_n$ form n Bernoulli trials with parameter p and if $X = X_1 + X_2 + \ldots + X_n$, then X has a binomial distribution with parameters n and p.

The binomial distribution is of fundamental importance in probability and statistics because of the following result for any experiment which can have outcome only either success or failure. The experiment is performed n times independently and the probability of the success of any given performance is p. If X denotes the total number of successes in the n performances, then X has a binomial distribution with parameters n and p. The probability function of X is:

$$P(X = r) = P(\sum_{i=1}^{n} X_i = r) = \binom{n}{r} p^r(1-p)^{n-r} \tag{1.14}$$

where $r = 0, 1, 2, \ldots, n$.

We write

$$f(r|p) = \begin{cases} \binom{n}{r} p^r (1-p)^{n-r} & \text{if r=0, 1, 2, } \ldots \text{, n} \\ 0 & \text{otherwise} \end{cases} \tag{1.15}$$

In this distribution n must be a positive integer and p must lie in the interval $0 \le p \le 1$. If X is represented by the sum of n Bernoulli trials, then it is easy to get its expectation, variance and moment generating function by using the properties of sums of independent random variables—cf. §1.3.

$$\mathcal{E}(X) = \sum_{i=1}^{n} \mathcal{E}(X_i) = np \tag{1.16}$$

$$Var(X) = \sum_{i=1}^{n} Var(X_i) = np(1-p) \tag{1.17}$$

$$\Psi(t) = \mathcal{E}(e^{tX}) = \prod_{i=1}^{n} \mathcal{E}(e^{tX_i}) = (pe^t + q)^n. \tag{1.18}$$

1.1.3 Poisson Distribution

The Poisson distribution is widely discussed in the statistical literature; one monograph devoted to it and its applications is Haight [102].

Take $t, \tau \in (0, \infty)$

$$p \,:\, \mathbb{N} \cup \{0\} \to [0,1) : k \mapsto \left(\frac{t}{\tau}\right)^k \frac{1}{k!} e^{-t/\tau} \tag{1.19}$$

$$\bar{k} = t/\tau \tag{1.20}$$

$$\sigma_k = t/\tau. \tag{1.21}$$

This probability function is used to model the number k of events in a region of measure t when the mean number of events per unit region is τ and the probability of an event occurring in a region depends only on the measure of the region, not its shape or location. Colloquially, in applications it is very common to encounter the usage of 'random' to mean the specific case of a Poisson process; formally in statistics the term random has a more general meaning: probabilistic, that is dependent on random variables. Figure 1.1 depicts a simulation of a 'random' array of 2000 line segments in a plane; the centres of the lines follow a Poisson process and the orientations of the lines follow a uniform distribution, cf. §1.2.1. So, in an intuitive sense, this is the result of the least choice, or maximum uncertainty, in the disposition of these line segments: the centre of each line segment is equally likely to fall in every region of given area and its angle of axis orientation is equally likely to fall in every interval of angles of fixed size. This kind of situation is representative

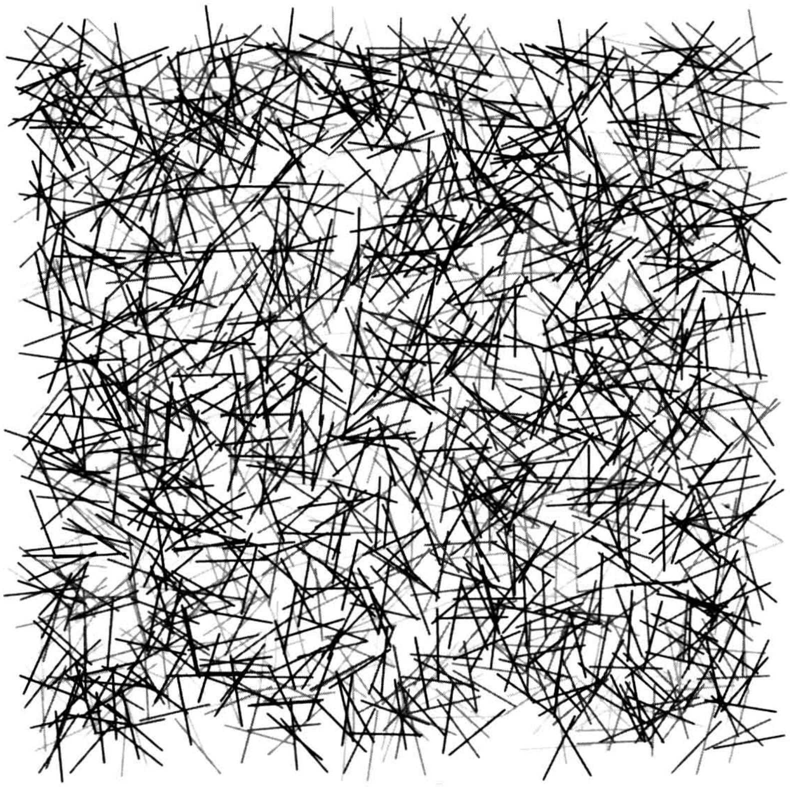

Fig. 1.1. Simulation of a random array of 2000 line segments in a plane; the centres of the lines follow a Poisson process and the orientations of the lines follow a uniform distribution. The grey tones correspond to order of deposition.

of common usage of the term 'random process' to mean subordinate to a Poisson process. A 'non-random' processes departs from Poisson by having constraints on the probabilities of placing of events or objects, typically as a result of external influence or of interactions among events or objects.

Importantly, the Poisson distribution can give a good approximation to the binomial distribution when n is large and p is close to 0. This is easy to see by making the correspondences:

$$e^{-pn} \longrightarrow (1-(n-r)p) \tag{1.22}$$

$$n!/(n-r)! \longrightarrow n^r . \tag{1.23}$$

Much of this monograph is concerned with the representation and classification of deviations from processes subordinate to a Poisson random variable, for example for a line process via the distribution of inter-event (nearest neighbour, or inter-incident) spacings. Such processes arise in statistics under the term renewal process [150].

We shall see in Chapter 9 that, for physical realisations of stochastic fibre networks, typical deviations from Poisson behaviour arise when the centres of the fibres tend to cluster, Figure 9.1, or when the orientations of their axes have preferential directions, Figure 9.15. Radiographs of real stochastic fibre networks are shown in Figure 9.3 from Oba [156]; the top network consists of fibres deposited approximately according to a Poisson planar process whereas in the lower networks the fibres have tended to cluster to differing extents.

1.2 Probability Density Functions for Continuous Variables

We are usually concerned with the case of continuous random variables defined on some $\Omega \subseteq \mathbb{R}^m$. For our present purposes we may view a probability density function (pdf) on $\Omega \subseteq \mathbb{R}^m$ as a subadditive measure function of unit weight, namely, a nonnegative map on Ω

$$f : \Omega \to [0, \infty) \qquad \text{(nonnegativity)} \tag{1.24}$$

$$\int_\Omega f = f(\Omega) = 1 \qquad \text{(unit weight)} \tag{1.25}$$

$$f(A \cup B) \leq f(A) + f(B), \ \forall A, B \subset \Omega, \qquad \text{(subadditivity)} \tag{1.26}$$
$$\text{with equality} \iff A \cap B = \emptyset.$$

Formally, we have a measure space of total measure 1 with σ-algebra typically the Borel sets or the power set and the measure function induced by f

$$sub(\Omega) \to [0,1] : A \mapsto \int_A f = \text{integral of } f \text{ over } A$$

and as we have anticipated above, we usually abbreviate $\int_A f = f(A)$. Given an integrable (ie measurable in the σ-algebra) function $u : \Omega \to \mathbb{R}$, the expectation or mean value of u is defined to be

$$\mathcal{E}(u) = \overline{u} = \int_\Omega uf.$$

We say that f is the joint pdf for the random variables $x_1, x_2, \ldots, x_m$, being the coordinates of points in Ω, or that these random variables have the joint probability distribution f. If x is one of these random variables, and in particular for the important case of a single random variable x, we have the following

$$\overline{x} = \int_\Omega xf \tag{1.27}$$

$$\overline{x^2} = \int_\Omega x^2 f. \tag{1.28}$$

Again with slight abuse of notation, we call $\overline{x}$ the mean and the variance is the mean square deviation

$$\sigma_x^2 = \overline{(x - \overline{x})^2} = \overline{x^2} - (\overline{x})^2.$$

Its square root is the standard deviation $\sigma_x = +\sqrt{\overline{x^2} - (\overline{x})^2}$ and the ratio $\sigma_x/\overline{x}$ is the coefficient of variation, of the random variable x. Some inequalities for the probability of a random variable exceeding a given value are worth mentioning.

Markov's Inequality: If x is a nonnegative random variable with probability density function f then for all $a > 0$, the probability that $x > a$ is

$$\int_a^\infty f \le \frac{\overline{x}}{a}. \tag{1.29}$$

Chebyshev's Inequality: If x is a random variable having probability density function f with zero mean and finite variance σ^2, then for all $a > 0$, the probability that $x > a$ is

$$\int_a^\infty f \le \frac{\sigma^2}{\sigma^2 + a^2}. \tag{1.30}$$

Bienaymé-Chebyshev's Inequality: If x is a random variable having probability density function f and u is a nonnegative non-decreasing function on $(0, \infty)$, then for all $a > 0$ the probability that $|x| > a$ is

$$1 - \int_{-a}^a f \le \frac{\bar{u}}{u(a)}. \tag{1.31}$$

The cumulative distribution function (cdf) of a nonnegative random variable x with probability density function f is the function defined by

$$F : [0, \infty) \to [0, 1] : x \mapsto \int_0^x f(t)\, dt. \tag{1.32}$$

It is easily seen that if we wish to change from random variable x with density function f to a new random variable ξ when x is given as an invertible function of ξ, then the probability density function for ξ is represented by

$$g(\xi) = f(x(\xi)) \left| \frac{dx}{d\xi} \right|. \tag{1.33}$$

If independent real random variables x and y have probability density functions f, g respectively, then the probability density function h of their sum $z = x + y$ is given by

$$h(z) = \int_{-\infty}^\infty f(x)\, g(z - x)\, dx \tag{1.34}$$

and the probability density function p of their product $r = xy$ is given by

$$p(r) = \int_{-\infty}^{\infty} f(x)\, g\left(\frac{r}{x}\right) \frac{1}{|x|} dx. \tag{1.35}$$

Usually, a probability density function depends on a set of parameters, $\theta_1, \theta_2, \ldots, \theta_n$ and we say that we have an n-dimensional family. Then the corresponding change of variables formula involves the $n \times n$ Jacobian determinant for the multiple integrals, so generalizing (1.33).

1.2.1 Uniform Distribution

This is the simplest continuous distribution, with constant probability density function for a bounded random variable:

$$u : [a, b] \to [0, \infty) : x \mapsto \frac{1}{b-a} \tag{1.36}$$

$$\overline{x} = \frac{a+b}{2} \tag{1.37}$$

$$\sigma_x = \frac{b-a}{2\sqrt{3}}. \tag{1.38}$$

The probability of an event occurring in an interval $[\alpha, \beta] \subsetneqq [a, b]$ is simply proportional to the length of the interval:

$$P(x \in [\alpha, \beta]) = \frac{\beta - \alpha}{b - a}.$$

1.2.2 Exponential Distribution

Take $\lambda \in \mathbb{R}^+$; this is called the parameter of the exponential probability density function

$$f : [0, \infty) \to [0, \infty) : [a, b] \mapsto \int_{[a,b]} \frac{1}{\lambda} e^{-x/\lambda} \tag{1.39}$$

$$\overline{x} = \lambda \tag{1.40}$$

$$\sigma_x = \lambda. \tag{1.41}$$

The parameter space of the exponential distribution is $\mathbb{R}^+$, so exponential distributions form a 1-parameter family. In the sequel we shall see that quite generally we may provide a Riemannian structure to the parameter space of a family of distributions. Sometimes we call a family of pdfs a parametric statistical model.

Observe that, in the Poisson probability function (1.19) for events on the real line, the probability of zero zero events in an interval t is

$$p(0) = e^{-t/\tau}$$