# Stochastic relations and the problem of prior in the principle of maximum entropy.

Per K. Jakobsen, V.V. Lychagin

**Abstract**

In this paper we discuss the problem of prior for the maximum entropy principle. We show that stochastic relations can be used to constrain priors and in some case uniquely determine them. The principle of maximum entropy turns stochastic relations into (over)determined systems of partial difference equations for the partition function. All statistical consequences of the stochastic relations are determined by the space of solutions of the system.

## Contents

# 1 Introduction

The problem of prior has been at the center of probability theory and statistics from the very start. The general rules of probability theory tells us how to compute probabilities for derived events from probabilities of primary events. The problem of prior is concerned with the problem of assigning probabilities to primary events. The assignment is supposed to reflect an observers state of knowledge about the primary events. The assignment should be the same for different observers with the same state of knowledge but can be different for observers with different states of knowledge [4]. In this sense probability assignments are subjective [3],[1],[2]. The problem of the prior is how to turn states of knowledge into probability assignments. The first solution to this problem was used by the very founders of probability theory (Bernoulli and Laplace). If the observers only knowledge about the primary events are their number, then a uniform probability assignment should be used. This idea was later named the principle of indifference by J. M. Keynes. Generalizing this idea to countably infinite or even continuous spaces of primary events has turned out to be very problematic. Laplace himself used such a generalization is his work on probability theory. His probability distribution was uniform and not normalizable since it was defined on the whole real line. Using a uniform distribution for representing indifference about a random variable on a finite interval on the real line would seem to be more reasonable, at least it is normalizable. However even in this case serious problems arise as the well known Bertrand's paradox shows. Problems and paradoxes arising from the various generalizations of the principle of indifference to continuous random variables played no small part in the creation and for a long time complete dominance of the frequency interpretation[10] of probability theory.

The principle of maximum entropy appears first in the writings by W. Gibbs [6] on thermodynamics and statistical physics and later in the fundamental work on information theory by Shannon [7]. However it was E. T. Jaynes [5] who realized the real importance and general nature of the principle of maximum entropy. In his hands it turned into a general method for turning prior knowledge in the form of mean values for finite sets of random variables, into prior probability assignments. For a time it looked as if the problem of prior was essentially solved. However continuous valued random variables again turned out to be the Achilles heel. For finite spaces of events the principle will give a unique probability assignment, but when generalizing it to continuous random variables an unknown probability measure appears. The meaning of this measure became clear when it was realized that it is the maximum entropy distribution corresponding to no constraints. Thus it was understood that in order to apply the principle of maximum entropy one must start with a prior distribution. The principle of maximum entropy could not determine the prior, it could only tell us how to modify an already existing prior in order to satisfy constraints in the form of mean values. It seemed as if one was back at square one.

In this paper we will show that the principle of maximum entropy can be used to turn stochastic relations into constraints on the prior distribution. In some cases the relations will determine the prior uniquely but in general the prior will be constrained by a system of partial differential equations. Solutions to this system corresponds to possible priors in a many to one fashion. The space of solutions of the system of partial differential equations can either be described directly using constructive methods from the theory of ordinary and partial differential equations or, if this is impractical or impossible, the logical consequences of the original stochastic relations can be derived and classified by applying methods from the formal theory of differential equations to the prolongation hierarchy of the system of equations.

## 2 The principle of maximum entropy

In this section we give a review of the maximum entropy principle for a finite space of events. Let $\Omega = \{x_1, x_2, ...., x_n\}$ be a finite space of primary events. The algebra of possible events is the power set of $\Omega$. A probability assignment on the set of primary events is a set of numbers $\{p_i\}$ such that $0 \leq p_i \leq 1$ and $\sum_{i=1}^{n} p_i = 1$. Let $f_1, ..., f_k$ be real valued functions on $\Omega$. The principle of maximum entropy states that if the means of the functions $f_1, .., f_k$ are known $< f_i >= c_i$ one should, among all probability assignments that satisfy the constraints, pick the one that maximizes the entropy $S = -\sum_{i=1}^{n} p_i \ln p_i$. This constrained maximization problem is solved by introducing Lagrange multipliers $\lambda_0, \lambda_1, ..., \lambda_k$, one for each constraint $< f_i >= c_i$ and one for the constraint $\sum_{i=1}^{n} p_i = 1$. The well know solution is

$$p = \frac{e^{-\sum_{j=1}^{k} \lambda_j f_j}}{Z(\lambda_1, ..., \lambda_k)}$$

where $Z$ is the partition function and is given by

$$Z(\lambda_1, ..., \lambda_k) = \sum_{i=1}^{n} e^{-\sum_{j=1}^{k} \lambda_j f_j(i)}$$

All mean values of random variables that are (polynomial) functionals of the functions $f_1, .., f_k$ can be expressed directly in terms of partial derivatives of the the partition function. We have

$$\begin{aligned} < f_i > &= -\frac{1}{Z}\partial_{\lambda_i} Z \\ < f_i f_j > &= \frac{1}{Z}\partial_{\lambda_i \lambda_j} Z \end{aligned}$$

etc. In fact a formalism completely analogous to the classical thermodynamic formalism can be derived in this general setting. This point has been stressed by Jaynes in particular. The Lagrange multipliers, $\lambda$, are computed from the constraints values $c$ by solving the system

$$\partial_{\lambda_j} Z(\lambda_1, .., \lambda_k) = -c_j Z(\lambda_1, .., \lambda_k)$$

If there are no constraints the principle gives $Z = n$ and we get the uniform assignment

$$p_i = \frac{1}{n}$$

The maximum entropy principle can thus be viewed as a generalization of the principle of indifference. Since what we are doing is to fix the prior by making it the extremum of the functional $S$ it is natural to ask why this particular functional is used instead of some different functional. A different choice would certainly lead to a different prior distribution so it is a highly relevant question. We will not pursue this question here but just note that there are other choices of functional possible and several are used in the literature, but none has the naturality and simplicity enjoyed by the functional introduced by Gibbs, Shannon and Jaynes.

# 3 The problem of prior in the maximum entropy principle

The maximum entropy probability measure described in the previous section appears at first sight to have solved the problem of prior, at least for the cases when the observers state of knowledge consists of mean values of random variables on a finite space of events. However even in this simple setting there are problems. If there are no constraints the maximum entropy principle predicts that the correct prior is the uniform one. But what if the observer even before he is presented with constraints have some information that amounts to a nonuniform probability measure? This measure can come from previous application of the maximum entropy principle or from some other source. In the given form of the maximum entropy principle this kind of prior measure can not be taken into account. However the principle can be modified to include nonuniform prior measures simply by considering $\{p_i\}$ to be the density of the maximum entropy measure $\nu$ relative to the prior probability measure $\mu$, not the measure itself. Thus $\nu(i) = p_i\mu(i)$ The problem is now to maximize

$$S = -\sum_{i=1}^{n} p_i \ln p_i \mu(i)$$

subject to the constraints

$$\sum_{i=1}^{n} f_j(i) p_i \mu(i) = c_j$$

Maximizing the entropy now gives the following measure

$$\nu(i) = \frac{e^{-\sum_{j=1}^{k} \lambda_j f_j(i)}}{Z(\lambda_1, ..., \lambda_k)} \mu(i)$$

and no constraints gives $\nu = \mu$. This solves the problem of how to include nonuniform prior measures in the maximum entropy principle but at the same time it reveals the true nature of the principle. It is a systematic way of modifying a given prior probability assignment so that it is consistent with new, previously unknown constraints in the form of mean values of random variables. The principle in this new form has absolutely nothing to say about the prior.

It has been well documented in the literature that the maximum entropy principle formulated directly in terms of probability measures as in the previous section can not be generalized to the case of continuous random variables. For the continuous case the maximum entropy measure $\nu$ has to be absolutely continuous with respect to the prior measure $\mu$. This means that there exists a probability density $\varphi$ such that

$$\nu(V) = \int_V \varphi d\mu$$

The maximum entropy principle will now maximize the entropy functional

$$S(\varphi) = -\int \varphi \ln \varphi d\mu$$

subject to the constraints

$$\int f_j \varphi d\mu = c_j \quad j = 1, .., k$$

The method of Lagrange multipliers gives

$$\varphi = \frac{e^{-\sum_{j=1}^{k} \lambda_j f_j}}{Z(\lambda_1, ..., \lambda_k)}$$

where the partition function is

$$Z(\lambda_1, .., \lambda_k) = \int_\Omega e^{-\sum_{j=1}^{k} \lambda_j f_j} d\mu$$

All formal rules for computing means of random variables using partial derivatives of the partition function are the same as for the case of finite probability spaces. The maximum entropy principle can be generalized in many directions, it can even be applied to the case when the random quantities correspond to noncommuting observables as in quantum mechanics [9].

In most applications of probability theory in statistics there is no underlying abstract probability space $\Omega$ and the random variables are not some functions defined on this space. What one typically has is a finite number of real valued random variables. Thus in the typical case $\Omega = \mathbb{R}^n$ and the random variables are just the coordinate function on $\mathbb{R}^n$. The prior probability measure is a measure, $\mu$ on $\mathbb{R}^n$. For this case the formula for the partition function is

$$Z(\lambda_1, .., \lambda_n) = \int_{\mathbb{R}^n} e^{-\sum_{j=1}^{n} \lambda_j x_j} d\mu(x_1, .., x_n)$$

The partition function is thus nothing else than the Laplace transform of the prior measure. This relation can be inverted using the Fourier transform when the prior measure has a density, $\rho_0$ with respect to the standard measure on $\mathbb{R}^n$.

$$\rho_0(x_1, ..x_n) = \frac{1}{(2\pi)^n} \int_{\mathbb{R}^n} Z(i\lambda_1, .., i\lambda_n) e^{i\sum_{j=1}^{n} \lambda_j x_j} d\lambda_1 ..d\lambda_n$$

From these formulas it is clear that the maximum entropy principle has nothing to say about the prior. It merely defines a transform between prior measures and partition functions. However we will see in the next section that the transform can be used to turn algebraic relations between statistical quantities into constraints on the prior measure.

# 4 Stochastic relations

In probability theory and statistics random variables are studied by computing statistical quantities. These are certain algebraic combinations of means of functions of the random variables. A large set of such statistical quantities are in use, some simple examples are (angle brackets signify the mean)

| | |
|---|---|
| $< x >$ | The mean of x. |
| $< x^2 > - < x >^2$ | The variance of x. |
| $< x^3 > -3 < x >< x^2 > +2 < x >^3$ | The third cumulant |
| $< xy > - < x >< y >$ | The cross variance of x and y. |

All such quantities can systematically be expressed as functions of the form $F(q_1, .., q_k)$ where the variables $q_j$ are means of monomials in the random variables. We will define stochastic relations to be systems of equations for the quantities $q_j$.

$$F_i(q_1, .., q_k) = 0 \qquad i = 1, ..., s$$

Such relations are common in probability and statistics. Examples are zero mean, fixed variance, uncorrelated variables and identities expressing higher order cumulants in terms of lower ones. In the previous section we have seen that the maximum entropy principle defines a Laplace transform that map the prior measure to a partition function. As a direct consequence of this transformation we can express means of monomials in the random variables in terms of partial derivatives of the partition function

$$< x_{i_1} x_{i_2} .. x_{i_r} >= \frac{(-1)^r}{Z} \partial^r_{i_1 i_2 .. i_r} Z$$

This means that the maximum entropy principle turns stochastic relations into systems of partial differential equations for the partition function and therefore indirectly imposes constraints on the prior measure. The problem is now to describe the space of solutions of the system of partial differential equations. In general not all solutions to the equations can correspond to prior measures. From the definition of the partition function it is clear that

$$\begin{aligned} Z(0) &= 1 \\ D^2 Z(\lambda_1, .., \lambda_n) &\geq 0 \end{aligned}$$

must hold for any acceptable solution. Finding sufficient and necessary condition for partition functions to be the Laplace transform of a probability measure is not a simple matter, but some results are known [8]. We will not discuss this problem but rather try to explicitly construct the solution space or apply methods from the formal theory of differential equations. Typically, the solution space is not a linear space and even when it is, the dimension could easily be infinite. However, depending on the number and types of stochastic relations the solution space can end up being parametrized by a finite set of parameters or even be a single point. In this last situation the stochastic relations determine the prior uniquely. Note that in ordinary (parametric) statistics finite parameter families of probability distributions (Gaussian, Poisson, Bernoulli, t-distribution, etc) are assumed to apply in given situations. From the point of view discussed in this paper this means that in ordinary statistics stochastic relations constrain the solution space enough for it to be parameterized in terms of a finite number of parameters. Nonparametric statistics correspond to the situation when the solution space is so weakly constrained that it can not be parameterized in terms of a finite number of parameters. Methods from the theory of partial differential equations can in some cases parameterize such weakly constrained solution spaces, not in terms of real numbers, but in terms of arbitrary functions. However for such weakly constrained solution spaces there is another powerful tool available. This is the formal theory of partial differential equations. The main object of study in this theory is the infinite prolonged hierarchy of the given systems of differential equations. Thus one studies the infinite set of all differential consequences of a given system of equations. Each such differential consequence can be converted back to a stochastic relation by using the relation between mean of monomials and partial derivatives in reverse. One therefore gets the corresponding hierarchy of stochastic relations that are consequences of the original relations induced by the maximum entropy principle and implemented through the Laplace transform.

In the remaining part of the paper we will discuss several examples that illustrate the method that has been outlined in this and previous sections.

## 4.1 Stochastic relations for one random variable

Essentially all families of distribution in use in parametric statistics can be derived from simple stochastic relations involving the mean, variance and skewness. In this section we show some examples that support this statement.

### 4.1.1 Delta distribution

Let us consider the stochastic relation corresponding to a fixed mean. It is

$$< x > -q = 0.$$

The Laplace transform converts this into the ordinary differential equation

$$Z_\lambda = -qZ.$$

For this simple stochastic relation our system of partial differential equations is a single linear ordinary differential equation. The solution space is linear and parameterized by a single parameter

$$Z(\lambda) = ae^{-q\lambda}.$$

The condition $Z(0) = 1$ fixes the parameter $a$ to be one and we have a unique solution. It is a simple matter to apply the inverse transform and show that the corresponding prior measure is

$$\mu = \delta(x - q).$$

### 4.1.2 Normal distribution

The stochastic relation corresponding to constant variance is

$$var(x) = q$$

and the corresponding differential equation is

$$ZZ_{\lambda\lambda} - Z_\lambda^2 - qZ^2 = 0.$$

This is a second order nonlinear ordinary differential equation. The general solution of the nonlinear equation that satisfies the requirement $Z(0) = 1$ is

$$Z(\lambda) = e^{a\lambda + \frac{1}{2}q\lambda^2}, \qquad a \in \mathbb{R}.$$

and the corresponding measure has a density with respect to the standard measure on $\mathbb{R}$ given by

$$\rho(x) = \frac{1}{\sqrt{2\pi q}} e^{-\frac{(x+a)^2}{2q}}.$$

which is the normal distribution.

### 4.1.3 Poisson distribution

Let us consider the stochastic relation

$$var(x) =< x > .$$

The corresponding differential equation is

$$ZZ_{\lambda\lambda} - Z_\lambda^2 + ZZ_\lambda = 0.$$

This equation and most equations derived from stochastic relations simplify considerably if we introduce a new function $\varphi$ through $Z = e^\varphi$. The equation for $\varphi$ is

$$\varphi_{\lambda\lambda} = -\varphi_\lambda.$$

This equation is easy to solve and the corresponding family of partition functions satisfying, as always, the constraint $Z(0) = 1$ is

$$Z(\lambda) = e^{a(e^{-\lambda}-1)}.$$

The corresponding distribution is supported on $\Omega = \{0, 1, 2, ....\}$ and is of the form

$$\rho(k) = \frac{e^{-a}a^k}{k!}.$$

This is the Poisson distribution.

### 4.1.4 Gamma distribution

Let us consider a stochastic relation

$$var(x) = \frac{1}{k} < x >^2 \quad k > 0.$$

Expressed in terms of $\varphi$ the corresponding differential equation is

$$\varphi_{\lambda\lambda} = \frac{1}{k}\varphi_\lambda^2.$$

The general solution of this equation gives the following family of partition functions

$$Z(\lambda) = (1 - a\lambda)^{-k} \quad a > 0.$$

The corresponding distribution is supported on $\Omega = (0, \infty)$ and is

$$\rho(x) = x^{k-1}\frac{e^{-\frac{x}{a}}}{a^k\Gamma(k)}.$$

This is the Gamma distribution

### 4.1.5 Bernoulli and Binomial distribution

Let the variance be the following quadratic function of the mean

$$var(x) =< x > (1- < x >).$$

The corresponding differential equation for $\varphi$ is

$$\varphi_{\lambda\lambda} = -\varphi_\lambda(1 + \varphi_\lambda).$$

The solution of the equation gives the following family of partition functions

$$Z(\lambda) = p + qe^{-\lambda} \qquad p + q = 1.$$

The corresponding distribution is supported on $\Omega = \{0, 1\}$ and is given by $\rho(0) = p$, $\rho(1) = q$. This is the Bernoulli distribution. If we generalize the stochastic relation to

$$var(x) =< x > (1 - \frac{1}{n} < x >).$$

where $n$ is a natural number we get the differential equation

$$\varphi_{\lambda\lambda} = -\varphi_\lambda(1 + \frac{1}{n}\varphi_\lambda).$$

The partition function is found to be

$$Z(\lambda) = (p + qe^{-\lambda}\ )^n \qquad p + q = 1.$$

The corresponding density is supported on $\Omega = \{0, 1, ...n\}$ and is given by

$$\rho(k) = \binom{n}{k} p^k q^{n-k}.$$

This is the Binomial distribution.

## 4.2 Stochastic relations for more than one random variable

When the number of random variables becomes larger than one, stochastic relations in general lead to systems of nonlinear partial differential equations. Unless the number and type of relations is right, it is impossible to describe the solution space in terms of a finite number of parameters. This lead us into the domain of nonparametric statistics. This is the domain where the methods from the formal theory of differential equations comes into play. It is not possible to give nontrivial applications of the theory in this short communication, we will limit ourselves to two simple examples.

### 4.2.1 The Multinomial distribution

Let $x_1, ...x_n$ be $n$ random variables and consider the following system of stochastic relations

$$\begin{aligned} var(x_i) &= < x_i > (1 - \frac{1}{n} < x_i >) \;\; i = 1, ..n \\ cov(x_i, x_j) &= -\frac{1}{n} < x_i >< x_j > \quad i, j = 1, ...n, \quad i \neq j \end{aligned}$$

The corresponding system of partial differential equations is

$$\begin{aligned} \varphi_{\lambda_i\lambda_i} &= -\varphi_{\lambda_i}(1 + \frac{1}{n}\varphi_{\lambda_i}) \\ \varphi_{\lambda_i\lambda_j} &= -\frac{1}{n}\varphi_{\lambda_i}\varphi_{\lambda_j} \end{aligned}$$

The second part of the system of equations has general solutions of the form $\varphi = n\ln(\theta)$ where $\theta(\lambda_1, .., \lambda_n) = \sum_{i=1}^n \theta_i(\lambda_i)$. Inserted into the first part of the system this form of $\varphi$ easily gives the partition function corresponding to the multinomial distribution. This system of relations thus constrained the space of solutions so much that it could be describes in terms of a finite number of parameters.

### 4.2.2 Stochastic relations for the mean

For a single random variable, stochastic relations involving only the mean gives distributions located on a finite set of points. For more than one random variable such relations gives rise to nonparametric statistics, or solution spaces parameterized by functions. The theory of partial differential equations can be used to give a full description of these solution spaces. As an example of such a relation consider the case of two random variables whose means are constrained to be on a circle of radius $r$.

$$< x >^2 + < y >^2 = r^2$$

The corresponding partial differential equation is in terms of $\varphi$

$$\varphi_\lambda^2 + \varphi_\mu^2 = r^2$$

The following $Z$ is in the solution space

$$Z = e^{r\sqrt{\lambda^2+\mu^2}}$$

This partition functions predicts that the following stochastic relation should hold

$$var(x) = \left(\frac{< y >}{< x >}\right)^2 var(y)$$

The partial differential equation has, however, infinitely many solutions. The method of characteristics can be used to describe the complete solution space. In order to derive stochastic relations that holds for all $Z$ in the solution space, these are the ones that can be said to be consequences of the of the circle constrain, we should consider differential prolongations of the original differential equation. The first prolongation is the system

$$\begin{aligned}
\varphi_\lambda^2 + \varphi_\mu^2 &= r^2 \\
\varphi_\lambda \varphi_{\lambda\lambda} + \varphi_\mu \varphi_{\mu\lambda} &= 0 \\
\varphi_\lambda \varphi_{\lambda\mu} + \varphi_\mu \varphi_{\mu\mu} &= 0
\end{aligned}$$

this system implies that

$$\varphi_{\lambda\lambda} = \left(\frac{\varphi_\mu}{\varphi_\lambda}\right)^2 \varphi_{\mu\mu}$$

Translated into stochastic relations this is exactly the one we derived for the special solution $\varphi = r\sqrt{\lambda^2 + \mu^2}$ and it thus holds for all solutions. It is of considerable interest to find a finite set of basic stochastic relations that through some construction procedure implies all consequences of some given system of stochastic relations. This is exactly the kind of question addressed in the formal theory of partial differential equations and the tools developed there can now through the maximum entropy principle be brought into the area of nonparametric statistics.

# References

[1] "Probability,Frequency and Reasonable Expectation", American Journal of Physics, **14** (1946), 1–13.

[2] "The algebra of probable inference", Johns Hopkins University Press (1961).

[3] H. Jeffreys, "Theory of Probability", Oxford Uuniversity Press (1961).

[4] E.T. Jaynes, "Probability Theory: The Logic of Science", Cambridge University press (2003).

[5] E.T. Jaynes, "Information Theory and Statistical Mechanics", The Physical Review, **106**, no.4, (May 15, 1957), 620–630.

[6] W.J. Gibbs "Elementary principles in statistical mechanics", Schribner's Sons (1902).

[7] C.E. Shannon, "A mathematical theory of communications", Bell System Technical Journal, **27** (July-October 1948), 379-423, 623–656.

[8] O.S.Rothaus, "Some properties of Laplace transforms of measures", Transactions of the American Mathematical Society, **131**, no. 1 (Apr. 1968), 163–169.

[9] P. Jakobsen, V.L. Lychagin, "Maximum Entropy Wavefunctions", The Lobachevskii Journal of Mathematics, **23** (2006), 29–56

[10] R. Fisher, "On the mathematical foundation of theoretical statistics", Philosophical Transactions of the Royal Society, A, **222** (1922), 309–368.