Transformation Group Approach To Applications Of The Beta Distribution

Robert W. Johnson

Alphawave Research, Jonesboro, GA, USA

[Received XXXXX, 2013. Revised XXXXX, 2013]

Summary. A transformation group approach to the prior for the parameters of the beta distribution is suggested. The relationship between the beta distribution and the Poisson and gamma distributions in the continuum is derived. Comparison to the maximum likelihood estimate of the parameters from an empirical distribution of network links is performed. The effect of the prior on the analysis of some well known examples from statistical genetics also is examined. Use of the beta distribution in the classification problem is discussed.

Keywords: Bayesian inference; Inductive reasoning; Beta distribution

1. Introduction

The beta distribution of the first kind, usually written in terms of the incomplete beta function, can be used to model the distribution of measurements whose values all lie between zero and one. The most widely known technique for estimating the parameters, the method of moments, simply selects that beta distribution with the same first and second moments as found empirically from the data. However, such procedure is not well-justified from the perspective of probability theory. To evaluate the reliability of the estimate of a model’s parameters, as well as to determine the net evidence for a particular model relative to some other, one needs to follow the mathematical procedure which has come to be known as Bayesian data analysis.

Use of the beta distribution can be found in a variety of applications. One common use is as a model for an input process within a stochastic simulation (Kuhl et al., 2010). Another is in the calculation of costs expected from a civil or industrial engineering project (Bektowski and Pownuk, 2004). It also has widespread use in the study of population genomics (Balding and Nichols, 1995; Price et al., 2006). This paper concerns itself not so much with the choice of application but rather focuses on the methodology used to evaluate the parameters of the model given a set of measurements. Various methods have been suggested for the estimation of its parameters, including the method of moments (AbouRizk et al., 1991) and variants of the Kolmogorov-Smirnov test (Press et al., 1992), but here we will follow the Bayesian approach expressed in terms of conditional probabilities for the chance, evidence, prior, and likelihood, as related by the joint distribution of the quantities appearing in the model.

This paper is organized as follows. After a brief description of Bayesian data analysis, we demonstrate the relation between the beta distribution and the Poisson and gamma distributions in the continuum. The joint density can be expressed in alternate coordinate systems through geometric transformations which preserve the volume. An analysis of network link data grouped by search engine rank is performed to determine whether highly ranked sites have a link distribution which is significantly different than sites of lower rank. The model is then applied to some well known examples of genomic inference from population statistics of an observable locus. After that, we examine the use of the beta distribution in the classification problem, where one predicts the type of some new object from comparison to a set of known objects. We will conclude with a discussion of our findings and a summary of our results.

Some readers may find our use of the transformation group approach reactionary, archaic, or even naive, in light of the voluminous literature discussing other, more complicated strategies for deriving the form of the prior given some model for the likelihood of the observations. Our response to such criticisms is, why fix what is not broken? The analysis of similarity transformations has a long history in physics, leading one to statements of conservation of energy and momentum respective to translations in time and space. When discussing the analysis of data, one should never forget that real measurements carry an index for location on the universal manifold and are subject to the laws of nature; how much use is made of that information depends upon the application and the investigator.
2. Brief Description of Bayesian Data Analysis

The Bayesian approach to data analysis is best discussed using the language of conditional probability theory (Bretthorst, 1988; Durrett, 1994; Sivia, 1996). The expression for “the probability of A given B” can be written as

\[ p(A \mid B) \equiv p^A_B, \quad (1) \]

where \( A \) and \( B \) can have arbitrary dimensionality; for example, \( A \) could be a vector of measurements, and \( B \) could include both the vector of parameters associated with some model as well as any other conditioning statements such as the model index. The sum and product rules of probability theory yield the expressions for marginalization and Bayes’ theorem,

\[ p_A = \int_{\{B\}} dB \ p_{A,B}, \quad (2) \]
\[ p_A p_B p^B_A = p^B_A p_A p_B \equiv p^A_B \quad (3) \]

where marginalization follows from the requirement of unit normalization, and Bayes’ theorem follows from requiring logical consistency of the joint density \( p_{A,B} \). Let us write as the vector \( m \) the parameters for some model \( M \), and let the data be written as \( x \). Bayes’ theorem then relates the evidence for the parameters given the data \( p^m_x \) to the likelihood of the data given the parameters \( p^x_m \) through the expression

\[ p^m_x \propto p^x_m p^m \quad (4) \]

where the factor \( p^m \) describes the prior expectation over the parameter manifold in the absence of data, and the constant of proportionality \( p_x \) represents the chance of measuring the data, which is usually recovered from the normalization requirement of the evidence density \( \int_{\{m\}} dm \ p^m_x = 1 \).

The essential feature of Bayesian data analysis which takes it beyond maximum likelihood analysis is the inclusion of the prior density \( p^m \). The selection of the appropriate form of the prior for some coordinate mapping of the parameter manifold is guided by the principle of indifference applied to the behavior of the model under similarity transformations (Jaynes, 1968; Sivia, 1996; Dose, 2003). Here, indifference is realized by examining the transformation group of the parameter manifold given by \( m \). Having found the prior measure for one coordinate system, the prior measure for alternate coordinate systems can be found through the use of a Jacobian transformation.

When only one model is in play, its quality of fit is irrelevant. If no other description of the data is available, the most one can do is fit the parameters for the model at hand. In order to accomplish the task of hypothesis testing, Bayesian data analysis forces one to specify explicitly the alternatives. For a set of models indexed by \( M \), the factors in Equation (4) must be conditioned on the choice of \( M \). For two models \( M \in \{1, 2\} \), the relative evidence is given by the ratio of the net evidence for each model,

\[ p_1 | x \propto p_1^x p_1 / p_2^x p_2, \quad (5) \]

where the factor \( p_1 / p_2 \) describes any prior preference between the models and usually is identified as unity. The factors in the likelihood ratio are given by the marginalization of the joint density over the parameter manifold for each model,

\[ p_M^x = \int_{\{m\}} dm \ p_{M,m}^x = \int_{\{m\}} dm \ p_{m,M}^x p_M^m, \quad (6) \]

where the use of properly normalized densities for the likelihood and prior is required. In particular, the prior \( p^m \) is normalized to unity over the parameter manifold while the likelihood \( p^x_m \) retains its physical normalization.

An interesting feature of Bayesian model selection is that it accounts naturally for Occam’s principle of efficiency. Assuming model 1 has some parameter \( a \) with uniform prior of extent \( \Delta a \), and taking the quadratic (Gaussian) approximation of its likelihood, without prior preference for either model the evidence ratio becomes

\[ \frac{p(1 \mid x)}{p(2 \mid x)} = \frac{p(x \mid a_0, 1)}{p(x \mid 2)} \left( \frac{2 \pi \delta^2}{\Delta a^2} \right)^{1/2}, \quad (7) \]
where \( a_0 \) is the optimum value of the parameter and \( \delta_a^2 \) is its variance. With an adjustable parameter, model 1 very likely provides a better quality of fit as measured by the ratio \( p(x \mid a_0, 1) / p(x \mid 2) \); however, that is not the only factor in the net evidence ratio. The improved fit to the data comes at the cost of the Occam factor \((2\pi \delta_a^2 / \Delta_a^2)^{1/2}\) which measures the distribution of the evidence density relative to the parameter domain. One requirement for the Gaussian approximation is that the prior not severely restrict the likelihood \( \Delta_a \gg \delta_a \), thus the Occam factor works against the peak likelihood in the net evidence ratio in Equation (7). Another interesting feature is that, all else being equal, the model whose parameters have the larger variance is the one preferred by probability theory, as more of its parameter space is compatible with the measurements. Suppose model 2 has its own parameter \( b \) with comparable domain \( \Delta_b \approx \Delta_a \) and provides a comparable fit to the data \( p(x \mid b_0, 2) \approx p(x \mid a_0, 1) \). In this case, the net evidence ratio reduces to

\[
p_\chi^2 / p_\delta^2 \approx \delta_a / \delta_b ,
\]  

so that the net evidence for model 1 relative to 2 is given by the ratio of the deviation of their parameters.

### 3. Beta, Poisson, and Gamma Distributions

The beta distribution can be derived from consideration of the Poisson and gamma distributions in the continuum (Press et al., 1992). Let us begin by supposing the amount for some quantity measured per unit time \( A \) is given by a Poisson process with rate parameter \( a \) expressed in the same units \( u_a = u_A \), thus the likelihood can be written

\[
p_a^A = e^{-a} a^A / \Gamma(A + 1) = e^{-a} A^{-1} / \Gamma(A) ,
\]

in terms of the gamma function. According to Jaynes (1968), the parameter for a Poisson process must satisfy

\[
\pi \delta = \pi_a / \pi_\alpha \equiv \pi_\delta / \pi_\alpha = 1.000 \pm 0.0001 ,
\]

which defines the infinite constant \( \pi \).

The integral of the likelihood with respect to the parameter is unity, \( \int_0^\infty d a p_a^A = 1 \). The integral over \( a \) of the joint density \( p^{a,A} \) evaluates to

\[
\int_0^\infty d a p_a^A = \int_0^\infty d a e^{-a} a^{A-1} / \Gamma(A) = A^{-1} / C_0 \equiv p^A ,
\]

which is recognized as the chance of measuring \( A \). Having equivalent physical units, the quantities \( a \) and \( A \) possess the same transformation group, thus their intrinsic densities must be functionally identical. The integral of the evidence density,

\[
\int_0^\infty d A p_A^a = \int_0^\infty d a p_a^A p^A / p^A = \int_0^\infty d a e^{-a} a^{A-1} / \Gamma(A) = 1 ,
\]

verifies its normalization. The joint density can be written explicitly as

\[
p^{a,A} = a^A / C_0 e^a a A \Gamma(A)
\]

and shown to have unit normalization,

\[
\int_0^\infty d A \int_0^\infty d a p^{a,A} = \int_0^\infty d A p^A = 1 .
\]

The remaining integral \( \int_0^\infty d A p_a^A \) cannot be easily evaluated; however, an heuristic argument (given in Appendix A) indicates that its value also is unity.

Note that the joint density \( p^{a,A} \) does not care whether \( a \) and \( A \) are identified as parameter and observable, respectively, or vice versa. The identification of evidence, chance, likelihood, and prior similarly is arbitrary.
as long as one is consistent [\text{sivia}, 1996]. The decomposition through Bayes’ theorem of the joint density allows one to write
\begin{equation}
    a^{-1}\text{Poisson}(A \mid a) = A^{-1}\text{Gamma}(a \mid A),
\end{equation}
thus the gamma distribution is the evidence for a Poisson process likelihood, and \textit{vice versa}. With the reinterpretation as the evidence integral, the statement \( \int_0^\infty dA p_A^2 = 1 \) must hold. The second parameter commonly associated with the gamma distribution can be identified as the ratio of the units for the parameter and observable, which here is specified as unity.

Now let us consider the joint density \( p_{a,b}^{A,B} \), which can be written as
\begin{equation}
    p_{a,b}^{A,B} = \left[ \frac{a^A b^B}{e^{A+B} A! B!} \right] / abC_0^2.
\end{equation}
Under a change of coordinate mapping \((a, b) \rightarrow (x, y)\) such that
\begin{equation}
    \begin{bmatrix} x \\ y \end{bmatrix} = \begin{bmatrix} a/(a+b) \\ a+b \end{bmatrix} \Longleftrightarrow \begin{bmatrix} a \\ b \end{bmatrix} = \begin{bmatrix} xy \\ (1-x)y \end{bmatrix},
\end{equation}
with domain \( x \in [0, 1] \) and \( y \in [0, \infty] \), the Jacobian matrix is given by
\begin{equation}
    J_{x,y}^{a,b} = \frac{\partial(a, b)}{\partial(x, y)} = \begin{bmatrix} y & x \\ -y & 1-x \end{bmatrix},
\end{equation}
whose determinant is \( |J_{x,y}^{a,b}| = y \). The intrinsic density in the new coordinates is thus
\begin{equation}
    p_{x,y}^{x,y} = p_{x,y}^{a,b} |J_{x,y}^{a,b}| = x^{-1}(1-x)^{-1}y^{-1}/C_0^2,
\end{equation}
and the conditional density is
\begin{equation}
    p_{x,y}^{A,B} = x^{A-1}(1-x)^{B-1} y^{A+B-1} e^{-y} / A\Gamma(A)\Gamma(B).
\end{equation}
One can then write
\begin{equation}
    p_{A,B}^{x,y} = p_{x,y}^{A,B} p_{x,y}^{x,y} / p_{A,B}^{A,B} = x^{A-1}(1-x)^{B-1} y^{A+B-1} e^{-y} / \Gamma(A)\Gamma(B),
\end{equation}
since \( p_{A,B}^{A,B} = A^{-1}B^{-1}/C_0^2 \), which integrates to unity,
\begin{equation}
    \int_0^1 dx \int_0^\infty dy p_{A,B}^{x,y} = 1,
\end{equation}
using the evaluations
\begin{equation}
    \int_0^\infty dy y^{A+B-1} e^{-y} = \Gamma(A+B),
\end{equation}
\begin{equation}
    \int_0^1 dx x^{A-1}(1-x)^{B-1} = \beta(A,B).
\end{equation}
Marginalization then yields
\begin{equation}
    p_{x}^{x,A,B} = \int_0^\infty dy p_{x,y}^{x,A,B} = x^{A-1}(1-x)^{B-1} / \beta(A,B)ABC_0^2
\end{equation}
\begin{equation}
    = p_{A,B}^{x} p_{A,B}^{A,B},
\end{equation}
which is the main result of this section. With the interpretation of \( x = a/(a+b) \) as a normalized frequency (rate of observance), one can state that the intrinsic density for an absolute probability is \( p_x = x^{-1}(1-x)^{-1}/C_0 \), while that for a relative probability \( z = a/b \in [0, \infty] \) is \( p_z = z^{-1}/C_0 \).
4. **Application to an Empirical Distribution**

The first application we will look at is the use of the beta distribution to model an empirical distribution of observables \( x \equiv x_n \) for \( n \in [1, N] \). The evidence for the parameters \( A \) and \( B \) in this case can be written as

\[
p^A_B \propto p^A_B p^{A,B} = p^A_B \prod_n p^{x_n}_{A,B} = p^A_B \prod_n x_n^{-1}(1-x_n)^{B-1}/\beta(A,B),
\]

(26a)

when the observables are independent with uniform weight. Using the notation \( q \equiv -\log p \) for the information density, one has

\[
q^{A,B} = \log A + \log B + 2L_0, \\
q^{A,B}_{A,B} = N \log \beta(A,B) + \sum_n [(1-A)\log x_n + (1-B)\log(1-x_n)],
\]

(27, 28)

where \( L_0 = \log C_0 \) is an infinite constant.

The maximum likelihood estimate of the optimal parameter values can be found by solving the equations for \( \nabla q^{A,B}_{A,B} = 0 \), yielding

\[
N^{-1} \sum_n \log x_n = \Lambda_1(A) - \Lambda_1(A + B), \\
N^{-1} \sum_n \log(1-x_n) = \Lambda_1(B) - \Lambda_1(A + B),
\]

(29, 30)

using the notation \( \Lambda_k(r) \equiv (\partial_r)^k \log \Gamma(r) \) for the polygamma functions with integer order \( k \) and real argument \( r \). The LHS of the equations above are the empirical expectation values, while the RHS are recognized as \( \langle \log x \rangle_{x|A,B} \) and \( \langle \log(1-x) \rangle_{x|A,B} \) taken with respect to the beta distribution,

\[
\langle f(x) \rangle_{x|A,B} \equiv \int_0^1 dx f(x) p^A_B,
\]

(31)

and the solution to the system cannot be expressed analytically but is easily found numerically. For comparison, the method of moments solves the equations

\[
\langle x_n \rangle_n = A/(A+B), \\
\langle (x_n - \langle x_n \rangle_n)^2 \rangle_n = AB/(A+B+1)(A+B)^2.
\]

(32, 33)

The maximum evidence analysis includes the effect of the intrinsic terms,

\[
\langle \log x_n \rangle_n = (NA)^{-1} + \Lambda_1(A) - \Lambda_1(A + B), \\
\langle \log(1-x_n) \rangle_n = (NB)^{-1} + \Lambda_1(B) - \Lambda_1(A + B),
\]

(34, 35)

such that \( \nabla q^{A,B}_{A,B} = -\nabla q^{A,B} \) at the optimal estimate. In the limit \( N \to \infty \) the prior becomes irrelevant, but when the amount of data is small \( N \gtrsim 1 \) it can be significant.

Under a change of coordinates \( (A, B) \to (\lambda, \mu) \), defined by

\[
\left[ \begin{array}{c} \lambda \\ \mu \end{array} \right] = \left[ \begin{array}{c} 1/(A+B+1) \\ A/(A+B) \end{array} \right] \iff \left[ \begin{array}{c} A \\ B \end{array} \right] = \left[ \begin{array}{c} \lambda(1-\lambda)/\mu \\ (1-\mu)(1-\lambda)/\lambda \end{array} \right],
\]

(36)

the location of the mode of the evidence is not simply the mapping of the mode in \((A, B)\) on account of the Jacobian transformation of the measure factor \(|J_{A,B}^{A,B}| = (1-\lambda)/\lambda^3\). To find the mode in the new coordinates, let the merit function \( F(A, B) \) be the non-constant part of the evidence information \( q^{A,B} \). The gradient
and Hessian of \( F \) in \((A,B)\) are then given by \( G(A,B) = \nabla F(A,B) \) and \( H(A,B) = \nabla^T G(A,B) \), with \( \nabla \equiv (\partial_A, \partial_B) \) a covariant (row) vector. The sequence of transformations then proceeds geometrically,

\[
F(\lambda, \mu) = F(A,B) - \log |J| ,
\]

\[
G(\lambda, \mu) = J^T G(A,B) - \nabla \log |J| ,
\]

\[
H(\lambda, \mu) = J^T H(A,B) J - \nabla^T \nabla \log |J| ,
\]

where the operator \( \nabla \) is evaluated in the new \((\lambda, \mu)\) coordinate system, as are the parameter values \( A \) and \( B \), and \( J \equiv J^{A,B} \).

What does not change under a remapping of the parameter manifold is the expectation value of an observable,

\[
\langle f(A,B) \rangle_{A,B|x} = \int_0^\infty dA \int_0^\infty dB \ f(A,B) p^{A,B}_{x|\psi}
\]

\[
= \int_0^1 d\lambda \int_0^1 d\mu \ f(\lambda, \mu) p^{A,B}_{x|\psi} = \langle f(\lambda, \mu) \rangle_{\lambda,\mu|x} ,
\]

as the Jacobian transformation is a volume preserving operation such that \( p^{A,B} = \lambda^{-1}(1 - \lambda)^{-1} \mu^{-1}(1 - \mu)^{-1}/C_0^2 \). In particular, the expectation value of the likelihood \( \langle p^{A,B}_{x|A,B} \rangle_{A,B} \) appearing in the model comparison ratio can be evaluated in any coordinate system. The reason for using the expected likelihood rather than the peak likelihood when assessing which model is best supported by the data becomes clear when one considers the effect of coordinate transformations. When one of the models has more parameters than the other, care must be taken to exclude the boundary of the manifold, as described following; an alternative method which cancels the infinite constants formally by assigning one of the models a subset of the data is beyond the scope of this article.

When the infinite constant \( C_0 \) does not cancel out of the model comparison ratio (Bayes factor), it must be replaced by a finite number. The extent to which the boundary can be excluded can be surmised by asking only very general questions about the nature of the data, such as its scale. The existence of measurements presumes the existence of a measuring apparatus, whose finite resolution ultimately determines what one can say \textit{a priori} about the extent of a sensible region in the parameter manifold. Suppose we know that the data \( x_k \) are bounded such that \( \epsilon < x_k < 1 - \epsilon \) for all \( k \) and \( 0 < \epsilon < 1/2 \). The finite normalization can be evaluated as

\[
C_\epsilon = \int_{\epsilon}^{1-\epsilon} dx x^{-1}(1-x)^{-1} = \int_{z(\epsilon)}^{z(1-\epsilon)} dz z^{-1} = 2 \log(\epsilon^{-1} - 1)
\]

in terms of absolute probability \( x = z/(z + 1) \) or relative probability \( z = x/(1 - x) \). Sensible bounds for \( \mu \) are then given by those of \( x \), and for symmetry let us use the same limits for \( \lambda \). The parameter \( \mu \) has the interpretation as the expected value of \( x \), while the interpretation of \( \lambda \) in this context is not so clear. If one considers a set of objects with \( A \) members of type 1 and \( B \) members of type 2, with one object whose type is not determined, then \( \lambda \) has the interpretation as the chance of selecting the undetermined object out of the set, and \( \mu \) has the interpretation as the chance of that object being of type 1. Values for \( C_\epsilon \) in terms of the boundary exclusion \( \epsilon \) are given in Table 1.

Let us now turn to the analysis of some real data. The data set considered here is a collection of measurements \( N_{k,r} \) of subnet links to web pages indexed by keyword \( k \in [1,K] \) and search engine rank \( r \in [1,R] \). The density \( x_{k,r} = N_{k,r}/N_k \) is defined to be the number of subnet links to the \( r \)th page for keyword \( k \) normalized by the total number of links to pages for \( k \) such that \( \sum_r x_{k,r} = 1 \). This measurement process by construction excludes the values of 0 and 1 from the data set; if a page has no links, it cannot be found, while if it has all the links, other pages must have no links. This data set has \( K = 100 \) and \( R = 10 \) and is summarized in Table 2.
are now not values for $x$ at some locus has a dominant allele $G$ next let us look at how the beta distribution is used in the analysis of genetic profiles. Suppose the gene at some locus has a dominant allele $G$ and a recessive allele $g$ such that the genotypes $GG$, $Gg$, and $gg$ are distinguishable. According to Balding and Nichols (1995), the allele frequency $x$ for finding $G$ at the locus follows a beta distribution with parameters $A = \mu(1-\lambda)/\lambda$ and $B = (1-\mu)(1-\lambda)/\lambda$. The measurements are now not values for $x_k$ but rather the number of members of each genotype observed within a sampling

The practical question we are interested in is whether the distribution of subnet links bears any significant relation to search engine rank. To answer that question, we divide the data into two (or more) subgroups according to rank then compare the expected likelihoods for the multi-model fits $M_m$ indexed by $m \in [1, 10]$ to that for the single model $M_0$ describing all the data. We have devised two ways to perform the subdivision into groups indexed by $g$, which we will call method 1 and method 2. For method 1, the ranks $r \leq m$ are assigned to one group $g = 0$ while the remaining ranks $r > m$ are assigned to another $g = \infty$, such that $M_{10} \equiv M_0$ with only one set of parameters. For method 2, the ranks $r \leq m$ are treated independently $g = m$, and the remaining ranks $r > m$ are lumped together as before; in this case $M_{10} \equiv M_9$ as either model has ten sets of parameters.

The expectation value $\langle O_g \rangle_{\lambda, \mu|x}$ of the observables $O \in [\lambda, \mu]$ indexed by group $g$ and model $m$ are displayed in Table 3. The boundary of the parameter manifold is excluded by $\epsilon = 10^{-5}$, which sufficiently encompasses the entire set of data. The log evidence $\log_{10} \rho_0^m$ for each model $M_m$ relative to the defaul model $M_0$ is given in Table 4 using base 10 for easier comparison. For method 1, the relative evidence is

$$
\rho_0^m = \langle p(x_0 | \lambda_0, \mu_0) \rangle \langle p(x_{\infty} | \lambda_{\infty}, \mu_{\infty}) \rangle / \langle p(x_{\text{all}} | \lambda_{\text{all}}, \mu_{\text{all}}) \rangle,
$$

while for method 2 it is

$$
\rho_0^m = \prod_{l=1}^{m} \langle p(x_l | \lambda_l, \mu_l) \rangle \langle p(x_{\infty} | \lambda_{\infty}, \mu_{\infty}) \rangle / \langle p(x_{\text{all}} | \lambda_{\text{all}}, \mu_{\text{all}}) \rangle.
$$

For either method, the model $M_1$ isolates the first rank from the remainder, thus their expected likelihoods are equal, and the correspondence of $M_{10}$ indicated above is apparent in the table. Our interpretation of these results is that the first rank is significantly distinguished from the others, while the second and third ranks are marginally distinguished. After the third, there is nothing to be gained by distinguishing the ranks. What we have done is use mathematics to confirm one’s intuition that the pages with the lion’s share of the ranks are marginally distinguished. After the third, there is nothing to be gained by distinguishing the ranks.

5. Application to the Balding-Nichols Model

Next let us look at how the beta distribution is used in the analysis of genetic profiles. Suppose the gene at some locus has a dominant allele $G$ and a recessive allele $g$ such that the genotypes $GG$, $Gg$, and $gg$ are distinguishable. According to Balding and Nichols (1995), the allele frequency $x$ for finding $G$ at the locus follows a beta distribution with parameters $A = \mu(1-\lambda)/\lambda$ and $B = (1-\mu)(1-\lambda)/\lambda$. The measurements are now not values for $x_k$ but rather the number of members of each genotype observed within a sampling

| $r$ | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 | 10 |
|-----|---|---|---|---|---|---|---|---|---|----|
| $\langle \log x_{k,r} \rangle_k$ | -2.68 | -3.64 | -3.54 | -3.60 | -3.82 | -3.95 | -3.76 | -4.08 | -4.18 | -4.03 |
| $\langle \log(1 - x_{k,r}) \rangle_k$ | -0.28 | -0.11 | -0.16 | -0.11 | -0.15 | -0.11 | -0.10 | -0.09 | -0.13 | -0.11 |

### Table 3. Expectation value of observables $O_g$ for model $M_m$, indexed by $g$ and $m$

| $m$ | 0 | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 | 10 |
|-----|---|---|---|---|---|---|---|---|---|---|----|
| $\langle \lambda_0 \rangle$ | 0.29 | 0.25 | 0.25 | 0.23 | 0.23 | 0.23 | 0.22 | 0.21 | 0.22 | 0.22 |
| $\langle \mu_0 \rangle$ | 0.20 | 0.15 | 0.14 | 0.13 | 0.13 | 0.12 | 0.11 | 0.11 | 0.11 | 0.11 |
| $\langle \lambda_m \rangle$ | 0.29 | 0.17 | 0.24 | 0.17 | 0.25 | 0.19 | 0.16 | 0.16 | 0.25 | 0.21 |
| $\langle \mu_m \rangle$ | 0.20 | 0.10 | 0.13 | 0.09 | 0.12 | 0.09 | 0.09 | 0.08 | 0.11 | 0.10 |
| $\langle \lambda_\infty \rangle$ | 0.22 | 0.20 | 0.20 | 0.20 | 0.20 | 0.19 | 0.19 | 0.20 | 0.23 | 0.21 |
| $\langle \mu_\infty \rangle$ | 0.11 | 0.10 | 0.10 | 0.10 | 0.10 | 0.09 | 0.09 | 0.09 | 0.10 | 0.10 |

### Table 4. Log evidence $\log_{10} \rho_0^m$ for model $M_m$, relative to $M_0$

| $m$ | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 | 10 |
|-----|---|---|---|---|---|---|---|---|---|----|
| method 1 | 4.21 | 0.30 | 0.68 | -0.51 | -0.21 | -1.38 | -2.04 | -2.79 | -3.21 | 0.00 |
| method 2 | 4.21 | 1.02 | -1.74 | -4.59 | -7.32 | -10.90 | -13.75 | -16.49 | -19.63 | -19.63 |
of the $k$th population, $N_k = N_{k,GG} + N_{k,Gg} + N_{k,gg}$. In terms of the parameters, the probability for an individual to be a member of the genotype is given by

$$p_{A,B}^{GG} \equiv \langle x^2 \rangle_{x \mid A,B} = A(A + 1)/(A + B)(A + B + 1)$$  \hspace{1cm} (44a)

$$= \lambda \mu + (1 - \lambda)\mu^2$$  \hspace{1cm} (44b)

for the dominant homozygote, and by

$$p_{A,B}^{gg} \equiv \langle (1-x)^2 \rangle_{x \mid A,B} = B(B + 1)/(A + B)(A + B + 1)$$  \hspace{1cm} (45a)

$$= \lambda(1 - \mu) + (1 - \lambda)(1 - \mu)^2$$  \hspace{1cm} (45b)

for the recessive homozygote, while the heterozygote appears with probability

$$p_{A,B}^{Gg} \equiv 2\langle x(1-x) \rangle_{x \mid A,B} = 2AB/(A + B)(A + B + 1)$$  \hspace{1cm} (46a)

$$= 2(1 - \mu)\mu(1 - \mu)$$  \hspace{1cm} (46b)

where the factor of 2 accounts for the indistinguishability of the order of the alleles. In matrix form with unit 1-norm, the joint distribution of the genotypes can be written

$$\begin{bmatrix} p_{A,B}^{GG} & p_{A,B}^{Gg}/\sqrt{2} \\ p_{A,B}^{Gg}/\sqrt{2} & p_{A,B}^{gg} \end{bmatrix} = \lambda \begin{bmatrix} \mu & 0 \\ 0 & 1 - \mu \end{bmatrix} + (1 - \lambda) \begin{bmatrix} \mu^2 & \mu(1 - \mu) \\ \mu(1 - \mu) & (1 - \mu)^2 \end{bmatrix},$$  \hspace{1cm} (47)

yielding the interpretation of $\mu = (1 + B/A)^{-1}$ as the mean dominant allele frequency and of $\lambda = (1 + A + B)^{-1}$ as a measure of heterozygote suppression.

The probability of obtaining the measurements given knowledge of the parameter values is the product of the genotype likelihoods weighted by the number of members. For a single population,

$$q_{A,B}^{NGG,NGg,Ngg} = (p_{A,B}^{GG})^{NGG}(p_{A,B}^{Gg})^{NGg}(p_{A,B}^{gg})^{Ngg},$$  \hspace{1cm} (48)

thus the information content of the data (negative log likelihood) is

$$L(A, B) \equiv q_{A,B}^{NGG,NGg,Ngg} = N_{GG} q_{A,B}^{GG} + N_{Gg} q_{A,B}^{Gg} + N_{gg} q_{A,B}^{gg},$$  \hspace{1cm} (49)

recalling $q = -\log p$. The nontrivial solution of $\nabla L(A, B) = 0$ yields the maximum likelihood estimate of the optimal parameter values

$$\begin{bmatrix} a_L \\ b_L \end{bmatrix} = \begin{bmatrix} (2N_{GG} - N_{GG}^2)/(4N_{GG}N_{gg} - N_{Gg}^2) \\ (2N_{GG}N_{gg}^2 + N_{gg}^2)/(4N_{GG}N_{gg}^2 - N_{Gg}^2) \end{bmatrix},$$  \hspace{1cm} (50)

which corresponds to the location

$$\begin{bmatrix} \lambda_L \\ \mu_L \end{bmatrix} = \begin{bmatrix} (4N_{GG}N_{gg} - N_{Gg}^2)/(2N_{GG}N_{gg} + N_{Gg}^2) \\ 2N_{GG} + N_{Gg} \end{bmatrix}/2(N_{GG} + N_{Gg} + N_{gg})$$  \hspace{1cm} (51)

on the $(\lambda, \mu)$ manifold. Suppose now instead of the genotype observations our data consists of the raw allele counts for $G$ and $g$, given by $N_G = 2N_{GG} + N_{Gg}$ and $N_g = 2N_{gg} + N_{Gg}$ such that $2N = N_G + N_g$. The log likelihood in this case becomes

$$q_{A,B}^{NGG,Ng} = N_{GG} \log(1 + B/A) + N_g \log(1 + A/B)$$  \hspace{1cm} (52a)

$$= -N_G \log \mu - N_g \log(1 - \mu),$$  \hspace{1cm} (52b)

whose optimal estimate is the same $\mu_L = (1 + N_g/N_G)^{-1}$ with $\lambda$ undetermined. From the raw allele counts one can resolve only the dominant allele frequency for a single population.

The merit function for the evidence density $p_{NGG,NGg,Ngg}^{\lambda,\mu}$ in terms of the parameters $(\lambda, \mu)$ can be written as

$$F(\lambda, \mu) = L(\lambda, \mu) + \log[\lambda(1 - \lambda)\mu(1 - \mu)],$$  \hspace{1cm} (53)
using an unnormalized prior. When $\lambda = 0$, the population is said to be in Hardy–Weinberg equilibrium with a single parameter $\mu$ for the dominant allele frequency; however, one should observe that $\lambda = 1$ is also an equilibrium solution with a single parameter $\mu$. Those two cases correspond to the peaks in the prior for $\lambda$ when the boundary is not excluded. In the limit $\epsilon \to 0$, the normalized prior $p(\lambda)$ has the value $1/2$ at $\lambda$ equal to 0 or 1 and the value 0 everywhere else. Similarly, when $\mu$ equals 0 or 1, one finds that $\lambda$ is undetermined by the likelihood, thus those models have zero free parameters. The five models under consideration (for a single population) can thus be labeled $M_{\lambda,\mu}$, $M_{0,\mu}$, $M_{1,\mu}$, $M_0$, and $M_1$, where the first is a two parameter model, the next two are one parameter models, and the last two zero parameter models, all of which are conditioned on the value of the boundary exclusion $\epsilon$ determined in principle by the nature of the measurement apparatus. A similar approach is suggested by Johnson and Rossell (2010). See Figure 1 for a depiction of the mapping from the parameter manifold to the model labels using a large value of $\epsilon$ for clarity.

It is instructive to look at the information content of the data with respect to the various models. For $M_{\lambda,\mu}$ with two parameters, $L_{\lambda,\mu}$ is given by Equation (49), whose mode provides a good starting point for the numerical optimization of $F_{\lambda,\mu}$; four other points to consider are the projections of the likelihood mode onto the boundaries of the manifold. The model $M_{0,\mu}$ has an information density of

$$L_{0,\mu}(\mu) = -N_{Gg} \log 2 - (2N_{GG} + N_{Gg}) \log \mu - (2N_{gg} + N_{Gg}) \log(1 - \mu),$$

(54)

retaining the constant term with $N_{Gg}$, and $M_{1,\mu}$ has

$$L_{1,\mu}(\mu) = -N_{Gg} \log 0 - N_{GG} \log \mu - N_{gg} \log(1 - \mu),$$

(55)

supported only when $N_{Gg} = 0$ such that $N_{Gg} \log p_{\lambda=1} = \log 0^0 = 0$; otherwise, $L_{1,\mu} = \infty$. For either one parameter model, it is possible for certain values of the input data to yield an evidence density which is uniform in $\mu$; in those cases, the mode is undetermined and the unnormalized evidence density is equal to 1. For the zero parameter models,

$$L_0 = -N_{GG} \log 0 - N_{Gg} \log 0 - N_{gg} \log 1,$$

(56)

which equals 0 when only $N_{gg} > 0$ else is infinite, and by symmetry

$$L_1 = -N_{GG} \log 1 - N_{Gg} \log 0 - N_{gg} \log 0.$$

(57)
favored by a factor of approximately $7/4$ by the likelihood analysis.

The mean values of the parameters. Some small variation to those numbers is not visible when truncated. The selected, consistent with the amount of data. The location of the mode, when it exists, is displayed, as are the maximum likelihood values comparing the other models.

Either 0 or 1 according to whether they are supported by the data, which sets the unit of evidence when truncated. Since the zero parameter models have a manifold of a single point, the net evidence (mean likelihood) is normalized to yield the quality factors $Q_M$. The conventional interpretation is to state that the equilibrium model is not rejected on account of the small $P$ value of $\chi^2$ statistic for 1 degree of freedom (3 from the data less 2 used in the model) gives the significance of the deviation from equilibrium; in the table $P(\chi^2)$ is also given as a percent, and for comparison $P(3.84) \approx 95\%$. The conventional interpretation is to state that the equilibrium model is not rejected on account of the small value of $\chi^2$; however, a formal interpretation of the $P$ value requires one to state that the observations have a chance of $P$ not to be described by the chosen model. That interpretation yields a preference for the equilibrium model of 1 $- P(\chi^2)$. The relative likelihood of the non-equilibrium model is then estimated as $\rho(\chi^2) = P(\chi^2)/[1 - P(\chi^2)]$. From the given data, one sees that the non-equilibrium model actually is favored by a factor of approximately $7/4$ by the likelihood analysis.

The evidence analysis of the same data is shown in Table 6 for all five models. A value of $\epsilon = 10^{-5}$ is selected, consistent with the amount of data. The location of the mode, when it exists, is displayed, as are the mean values of the parameters. Some small variation to those numbers is not visible when truncated. The net evidence for each model $M$ is given in terms of its negative logarithm $q^M_N \equiv - \log (p^M_N/m)$ for parameter vector $m$ and data vector $N \equiv (N_{GG}, N_{Gg}, N_{gg})$. The $Q$ value for each model, interpreted as the probability that the model describes the data, is determined from

$$Q^M_N \equiv \exp(-q^M_N) / \sum_M \exp(-q^M_N),$$

such that $\sum_M Q^M_N = 1$. Of the two models supported by the data, that for Hardy–Weinberg equilibrium $M_{0,\mu}$ is assigned a probability close to 70%.

Next let us look at some data from James et al. (1983), displayed in Table 7. This time the data is broken down into that for subpopulations indexed by $k$ according to the geographic region of the observations. Since none of the populations have only $N_{gg} > 0$, the model $M_0$ can be discarded immediately. The practical question we are interested in is whether any single population is significantly different than the remainder. To answer that question, the net evidence (expected likelihood) for the models applied to the entire population $N_0 \equiv \sum_k N_k$ is compared to the product of the evidence for the subdivision into $N_k$ and $N_{\sim k} \equiv N_0 - N_k$. The results of this analysis are shown in Table 8 using a value of $\epsilon = 10^{-5}$. Values of 0 or 1 for the parameter mode appearing in the table are understood to be on the boundary given by $\epsilon$.

To identify which single population displays the most significant deviation from the remainder, for each $k$ the minimum $q^M_{N_k}$ is added to the minimum $q^M_{N_{\sim k}}$, then the minimum $q^M_{N_0}$ is subtracted to yield the (negative) log evidence for the subdivision relative to the net population $q^k$. Those values are then exponentiated and normalized to yield the quality factors $Q^k_N$. From Table 5 one sees that $M_{\lambda,\mu}$ is the model best supported by

| $N_{GG}$ | $N_{Gg}$ | $N_{gg}$ | $\lambda_L$ | $\mu_L$ | $\chi^2_\mu$ | $P(\chi^2)$ | $\rho(\chi^2)$ |
|---------|---------|---------|------------|--------|-----------|-----------|-----------|
| 1469    | 138     | 5       | 2.270      | 95.409 | 0.831     | 63.800    | 1.762     |

Table 6. Evidence analysis of the data from Table 5

| mode | $\lambda, \mu$ | $0, \mu$ | $1, \mu$ | 0 | 1 |
|------|---------------|---------|---------|---|---|
| mean | (0.00001, 0.954) | 0.954 | NaN | NaN | NaN |
| mean | (0.00958, 0.954) | 0.954 | NaN | NaN | NaN |
| $q^M_N$ | 510.4 | 509.6 | Inf | Inf | Inf |
| $Q^M_N$ | 0.302 | 0.698 | 0.000 | 0.000 | 0.000 |

Table 7. Genotype observations from James et al. (1983)

| $k$ | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 | 10 | 11 | 12 |
|-----|---|---|---|---|---|---|---|---|---|----|----|----|
| $N_{GG}$ | 29 | 14 | 15 | 9 | 9 | 23 | 23 | 29 | 5 | 1 | 0 | 1 |
| $N_{Gg}$ | 0 | 3 | 2 | 0 | 0 | 5 | 3 | 3 | 0 | 0 | 1 | 0 |
| $N_{gg}$ | 0 | 3 | 3 | 0 | 2 | 4 | 1 | 0 | 0 | 0 | 0 | 0 |
the net population and all the remainder populations, but all four models can be supported by some of the subpopulations \( N_k \). In Table 9 we display which model \( M_k \) best fits population \( N_k \) as well as the relative log evidence \( q_k \) and the quality factors \( Q_k \) in units of percent. Of the thirteen models under comparison, the most significant is the subdivision of the first population \( k = 1 \) from the remainder, whose \( Q \) is close to 94%. The suppression of the recessive allele in that population would appear to be significant, while that for the other populations displaying only \( N_{GG} > 0 \) is less so. A thorough analysis would consider all possible groupings of the subpopulations to determine the most statistically significant division of the net population from the given data. A more thorough analysis would make use of knowledge of the geographical regions sampled to consider only those groupings of populations in physical contact. The possibilities are endless and left as an exercise for the reader.

### 6. Application to Prediction and Classification

Let us begin this section by talking about baseball. Specifically, let us consider the use of the seasonal batting average as a predictor for whether a player will reach base on his next appearance. Let each appearance be indexed by time given by integer \( t \in [1, T] \), and let us identify a successful appearance as an event of type \( A \), while outs are of type \( B \). The record of successful appearances can be notated by \( A^t \), and similarly for \( B^t \). The evidence for the value of the batting average \( x \) is the product of the prior and likelihood factors, yielding the beta distribution \( p_{J,K}^x \propto x^{J-1}(1-x)^{K-1} \) with mode \( x_B = (J-1)/(J+K-2) \) and expectation value \( \langle x \rangle_{J,K} = J/(J+K) \), which coincides with
the likelihood mode $x_L$ and gives the predicted rate of success for subsequent appearances.

One can get fancy with the form of the prior $p_x$ on the basis of additional information pertinent to the problem at hand. In particular, one can use knowledge of the seasonal nature of the sport to impose sensible limits on the domain $x \in [\epsilon, 1-\epsilon]$. If our player’s season is not yet over, then there must be at least one more at bat scheduled. A sensible limit is thus given by $\epsilon = 1/(T+1)$, which incorporates the notions that nobody is perfect (1 is excluded) and of the benefit of the doubt (0 is excluded): assuming our player is a professional at least one event of each type should be observed per season, even for pitchers. One effect of such a prior is that it does not allow observations of only one type of event to pull the evidence mode all the way to the hypothetical limits of 0 and 1. Another effect is that early in the season $T \gtrsim 1$ the domain of $x$ requires an observation of the batter before starting to make predictions; once we are certain the batter is playing this season $T = 1$, we can state our expected chance of success is equal to 1/2, the only allowed point, with further observations expanding the domain until at the end of a long season $T \gg 1$ the prior is wide open.

Let us now turn to consideration of classifying some new event as type $A$ or $B$ on the basis of its location relative to those for $T$ observations whose classification is assigned. The elements of the measurement vectors $A$ and $B$ are now locations along some axis $\tau$, with a measurement uncertainty expressed by the Gaussian deviation $\sigma$. If the chance an event is of type $A$ is independent of location, one can write $p_{\tau}^{x, A, B} \propto p_{\tau}^{x, K} p_{\tau}^{x, A, B}$ where $p_{\tau}^{x, A, B}$ is a Gaussian centered on the mean location of all the events and each margin is normalized independently. That is obviously not the solution we are looking for, which should give an expectation of the form $x(\tau)$ based on a joint density that can be factored as $p_{\tau}^{x, A, B} = p_{\tau}^{x} A, B^{\tau}$ for $p^{\tau} \propto 1$.

Another way to express the notion that location has become irrelevant is by taking the limit $\sigma \rightarrow \infty$. In that case, one should require $p_{\tau}^{x, A, B} \rightarrow p_{\tau}^{x, K}$ for all $\tau$, which corresponds to neglecting the stadium of appearance in the batting average problem above. In doing so, we have not said that location does not exist, but rather that location does not matter. For finite $\sigma$, we should write $p_{\tau}^{x, A, B} \propto p^{x} A, B$, whose limit for $\tau \rightarrow \infty$ is $p^{x}$; observations nearby should not significantly affect our prediction for a galaxy far, far away. The problem now is one of assigning the appropriate form for the likelihood factor. For inspiration, we have looked at various approaches suggested in the literature (Terrell and Scott, 1992; Hall et al., 2008; Kim and Scott, 2012; Eberts and Steinwart, 2013).

At this stage the discussion becomes a bit heuristic. When the observations are independent, we can
factor the likelihood into the form

\[ p_{\sigma, \tau, x}^{A, B} = \prod_j p_{\sigma, \tau, x}^j \prod_k p_{\sigma, \tau, x}^k, \]  

(59)

where \( p_{\sigma, \tau, x}^j \) represents the chance datum \( j \) is of type \( A \), and similarly for \( p_{\sigma, \tau, x}^k \). What, then, is the form of \( p_{\sigma, \tau, x}^j \) that yields sensible results for all \( \sigma \) and irrespective of the underlying spatial distributions of the two types of events? A form which suggests itself is more clearly notated in terms of its logarithm

\[ q_{\sigma, \tau, x}^A = -r_\tau^A \log x, \]

\[ q_{\sigma, \tau, x}^B = -r_\tau^B \log(1-x), \]

(60)

whose limits are \( J \log x + K \log(1-x) \) for \( \sigma \to \infty \) and 0 for \( \tau \to \infty \), in accord with our requirements for the evidence density. Let us identify \( A(\tau) = \sum_j r_\tau^A \), and similarly for \( B(\tau) \); then the likelihood can be written as \( x^{A(\tau)}(1-x)^{B(\tau)} \), and the evidence for the value \( x \) at \( \tau \) is given by

\[ p_{\sigma, \tau, x}^{x, A, B} \propto x^{A(\tau)-1}(1-x)^{B(\tau)-1}, \]

(61)

which has the form of a beta distribution at all locations. An example of \( A(\tau) \) and \( B(\tau) \) for an arbitrary distribution of \( A \) and \( B \) in units of the deviation \( \sigma = 1 \) is shown in panel (a) of Figure 2. The values \( A_j \) are drawn uniformly over two disjoint regions each with a span of 2 units, and the values \( B_k \) are selected from a region spanning 2 units which overlaps partially one of the type \( A \) regions.

Out of respect for our heuristic argument, we should consider some alternative definitions for the likelihood. If instead of the relative probabilities \( r_\tau^j \) one defines \( A(\tau) \) as the sum of the absolute probabilities

\[ p_{\sigma, \tau}^j = (2\pi \sigma^2)^{-1/2} r_\tau^j \]  

such that \( \int d\tau \sum_j p_{\sigma, \tau}^j = J \), one has in the limit \( \sigma \to \infty \) the result \( A(\tau) \to 0 \), which does not recover the beta distribution in terms of \( J \) and \( K \). If one uses the product of the datum likelihoods to define

\[ A(\tau) = J(2\pi \sigma^2)^{-1/2} \exp^{-1/2[(\tau - \mu_A)^2J/\sigma^2]} \]  

for \( \mu_A = \langle A_j \rangle \), which also integrates over \( \tau \) to \( J \), one’s estimate for the evidence depends upon only the first moments of the event distributions, a procedure which is easily foiled when the underlying location distribution are not Gaussian. Finally, if one uses \( p_{\sigma, \tau, x}^j = x p_{\sigma, \tau}^j \), one recovers simply the independent distributions over \( x \) and \( \tau \). Examples of these definitions of \( A(\tau) \) and \( B(\tau) \) are displayed in panels (b) through (d) respectively of Figure 2 for the same \( A_j \) and \( B_k \).

A maximum likelihood predictor can be formed from the expression

\[ x_{ML}(\tau) = [1 + B(\tau)/A(\tau)]^{-1}, \]

(62)

which is evaluated from the measurements \( A \) and \( B \) with respect to \( \sigma \). The expectation value \( x_{EV}(\tau) = \langle x \rangle_{x|\sigma, A, B} \) however, takes into account the full domain of \( x \) as measured by the evidence density. In Figure 3 we display the maximum likelihood and expected value predictors for the distributions \( A(\tau) \) and \( B(\tau) \) shown in Figure 2. The likelihood estimate \( x_{ML} \) is the same in panels (a) and (b), since the ratio \( B(\tau)/A(\tau) \) in terms of the summed likelihoods does not depend on their normalization. The expectation value in panel (b) is more conservative, in that it more quickly approaches the expectation value of the prior, compared to panel (a). The likelihood estimate for the method of panel (c) gives a prediction for the region \( \tau \in [-2, -1] \) that is contrary to the observations, while its expectation value is very quickly drawn to that of the prior, even in the region \( \tau > 1.5 \) where only type \( A \) events are observed. The likelihood and expectation value predictors are identical when the location information is ignored, as seen in panel (d).

Let us now repeat the evaluation of the evidence densities \( p_{\sigma, \tau, x}^x \) for the various definitions of \( A(\tau) \) and \( B(\tau) \), but this time let us suppose that \( \sigma = 10 \) for the same locations \( A \) and \( B \). Let us also inspect the evidence densities directly, to see which one best encodes a reasonable estimate of the solution to our problem. In Figure 3 we display the evidence density for \( x \) as a function of \( \tau \) for the various likelihood models. We can see that panel (a) is the one most like panel (d), which evaluates the beta distribution without regard to location. The other models, panels (b) and (c), are not in accord with the conclusions a reasonable observer would draw intuitively from the presented data; surely with close to 100 observations the relative rate of production should be fairly well determined over the common region of the events. While our justification of Equation (60) is heuristic, its form is the same as that of a Gaussian with unequal weights, where each datum factor in the likelihood is an absolute probability to the power of a relative probability.
Fig. 3. Prediction values $x(\tau)$ from $A(\tau)$ and $B(\tau)$ as described in the text. The maximum likelihood predictor $x_{\text{ML}}$ is shown as □, and the expectation value $x_{\text{EV}}$ is shown as ◆.

Fig. 4. Evidence densities $p_{\sigma, \tau, A, B}$ for $\sigma = 10$ as described in the text.
What can we say about the limit $\sigma \to 0$, which indicates that observations are relevant only to predictions at the same location? With respect to the finite resolution of whatever apparatus is used to take the location measurements, what we really mean in that limit is that locations are resolved over a set of discrete channels which have no influence or bearing on events in other channels. Returning to the baseball analogy, that model asserts that batting averages for each stadium should be evaluated independently, which is not an unreasonable procedure, given by $A(\tau) = \sum_{A(j) = \tau} r^j_1$ and similarly for $B(\tau)$. The parameter $\tau$ can in fact be an abstract location, not just a physical one, with the interpretation of $r^j_1$ as the relevance of observations in one channel to predictions in another. We should also point out that we have been treating the location $\tau$ of the predicted classification as a quantity known exactly; if the location of the unclassified event $\tau^j$ is itself subject to measurement deviation $\sigma$, then one must convolute the evidence density with its normalized distribution, $p_{\tau^j,\tau^j} = \int d\tau p_{\tau^j,\tau} p_{\tau^j,\tau^j}$. Furthermore, if the value of $\sigma$ is unknown, it can be integrated out by treating it as as a nuisance parameter, $p_{\tau^j,\tau^j} = \int d\sigma p_{\tau^j,\tau} p_{\tau^j,\tau^j}$ for $p^\sigma \propto \sigma^{-1}$.

7. Discussion and Conclusion

Those who use Bayesian methods are often asked to explain the significance of the prior. On its own, Bayes’ theorem does not tell one how to assign the intrinsic probability density for the parameter manifold. For that task, one must turn to some other maxim. The principle of indifference is essentially a geometric argument that posits the existence of some coordinate mapping of the parameter manifold for which the information content is uniform. That mapping might not be the one most convenient for the investigator, thus the appearance of the prior may be nonuniform in one’s chosen coordinates. The main effect of the prior is to prevent one from overestimating structure in the model not supported by imperfect data. If the prior is neglected, one may unintentionally introduce a bias into one’s results.

With respect to the beta distribution, use of the transformation group prior is implicit in its functional form. In the absence of observations, what remains is the Haldane prior $p^x \propto x^{-1}(1-x)^{-1}$ expressing complete indifference to the value of an absolute probability. If the observations $A$ and $B$ are restricted to integer counts of class membership, then the effect of the prior is to require an observation of each type of event before one is certain both types are present within the population; until both types have been observed, the evidence density is infinite on the boundary at either 0 or 1. If one of each type has been observed, we are then certain that the production rate $x$ is between 0 and 1 with uniform distribution. Further observations then refine the estimate until the likelihood and evidence modes converge in the limit of infinite data.

The transformation group approach leads one to specify $q^{A,B} = \log A + \log B$ as the logarithm of the unnormalized prior measure over the $(A,B)$ manifold. In the course of this project we investigated use of the entropic prior $p^{A,B} \propto \exp\left( q_{A,B}^{A,B} \right)$, where

$$-q^{A,B} = \log \beta(A,B) + (A + B - 2)\Lambda_1(A + B) + (1 - A)\Lambda_1(A) + (1 - B)\Lambda_1(B)$$  \hspace{1cm} (63)$$

is evaluated from the Shannon-Jaynes expression [Lazo and Rathie, 1978]. The entropic expression for the prior was discarded after finding in the context of the Balding-Nichols genotype analysis that it did not lead to a hierarchy of models. The Jeffreys invariant prior

$$-q^{A,B} = 2^{-1} \log \left\{ \Lambda_2(A)\Lambda_2(B) - [\Lambda_2(A) + \Lambda_2(B)]\Lambda_2(A + B) \right\} ,$$  \hspace{1cm} (64)$$

proportional to the square root of the determinant of the Fisher matrix, likewise was considered. Its prior density is very similar to that given by the transformation group, thus results based on that prior should be close to the results presented here. Finally, the conjugate prior approach is discounted because there is no physical reason to suppose that the evidence and prior should be of the same algebraic form, mathematical convenience notwithstanding. Note that the appearance of the beta function in the beta distribution results from the normalization over the axis $x \in [0,1]$: if the domain of $x$ is more restrictive, the expression for the normalization as a function of the parameters $A$ and $B$ is more complicated. In that case, neither the entropic nor the Jeffreys prior is appropriate without severe modification, whereas the transformation group prior is unaltered.

Many investigators are troubled by the use of an improper prior, leading to an entire industry devoted to the generation of ever more complicated functions to be used as priors for statistical analysis of data. One
should think very carefully before deciding to employ any of those alternative strategies. The transformation group approach is based on the physical properties of the objects under consideration, with respect to the nature of the universe that we live in. The prior it yields represents a measure of uniform information content over the parameter manifold. The one dimensional improper transformation group priors are in fact just the uniform prior under a change of coordinates,
\[
\int_0^1 dx/x(1-x) = \int_0^\infty dz/z = \int_{-\infty}^\infty du \quad \text{for} \quad u = \log z \quad \text{and} \quad z = x/(1-x).
\]
The appearance of infinite densities on the boundary of the prior indicate where simpler models with fewer parameters exist; these models can be addressed by evaluating their Bayes factor relative to the model with the most complexity.

In summary, we have derived the beta distribution from the relation between the Poisson and gamma distributions with respect to the transformation group prior and have demonstrated its use in a variety of applications. The reason the beta distribution is so ubiquitous is that it is the natural distribution for a quantity that is an absolute probability as measured by the frequency of events. It can be used to model an empirical distribution of normalized quantities, as a model for the allele frequency in a breeding population, and to predict the classification of an event according to its location relative to other classified events. Its study has a long history in the literature, and it continues to be quite useful in the modern day.

A. Appendix section

In Section 3 we encountered an integral that could not be put into closed form analytically. This integral also appears in the maximum entropy formulation of the Fourier transform (Johnson, 2012). In this appendix we present an heuristic argument for its evaluation. Let
\[
I(a) \equiv \int_0^\infty dA \frac{a^A}{\Gamma(A+1)}
\]
represent the integral in question, and what we want to show is that \( I(a) = \exp a \). The exponential function is defined to be that function whose derivative is equal to itself, \( \partial_a \exp a = \exp a \), and so let us look at the derivative of the integral,
\[
\partial_a I(a) = \int_0^\infty dA \frac{Aa^{A-1}}{\Gamma(A+1)} = \int_0^\infty dA \frac{a^{A-1}}{\Gamma(A)}.
\]
Shifting the limits of integration down by one unit, we have
\[
\partial_a I(a) = \int_{-1}^\infty dA \frac{a^A}{\Gamma(A+1)} = \int_{-1}^0 dA \frac{a^A}{\Gamma(A+1)} + \int_0^\infty dA \frac{a^A}{\Gamma(A+1)},
\]
whereupon arguing that the measure vanishes \( dA = 0 \) when \( A < 0 \), the first term on the RHS evaluates to 0, and one is left with the result \( \partial_a I(a) = I(a) \). Note that the discrete analogue of \( I(a) \) is simply the familiar Poisson sum \( \sum_{A=0}^\infty \frac{a^A}{A!} = \exp a \) which holds for arbitrary units \( u_a = u_A \).

Acknowledgements

The author would like to thank Calum MacLeod for providing the network link data that initiated this investigation.

References

AbouRizk, S. M., D. W. Halpin, and J. R. Wilson (1991). Visual interactive fitting of beta distributions. Journal of Construction Engineering and Management 117(4), 589–605.

Balding, D. and R. Nichols (1995). A method for quantifying differentiation between populations at multi-allelic loci and its implications for investigating identity and paternity. Genetica 96, 3–12. 10.1007/BF01441146.

Bętkowski, M. and A. Pownuk (2004, September). Calculating risk of cost using Monte Carlo simulation with fuzzy parameters in civil engineering. In Proceeding of NSF workshop on Reliable Engineering Computing, pp. 179–192. Center for Reliable Engineering Computing, Georgia Tech Savannah.
Bretthorst, G. L. (1988). *Bayesian Spectrum Analysis and Parameter Estimation*. Berlin, Germany: Springer-Verlag.

Dose, V. (2003). Hyperplane priors. *AIP Conference Proceedings 659*(1), 350–360.

Durrett, R. (1994). *The Essentials of Probability*. Belmont, CA: Duxbury Press, A Division of Wadsworth, Inc.

Eberts, M. and I. Steinwart (2013). Optimal regression rates for SVMs using Gaussian kernels. *Electron. J. Statist.* 7, 1–42.

Ford, E. B. (1971). *Ecological genetics, [by] E. B. Ford* (3rd ed. ed.). London, UK: Chapman and Hall.

Hall, P., B. U. Park, and R. J. Samworth (2008). Choice of neighbor order in nearest-neighbor classification. *Annals of Statistics 36*(5), 2135–2152.

James, S. H., A. P. Wylie, M. S. Johnson, S. A. Carstairs, and G. A. Simpson (1983, Dec). Complex hybridity in *Isotoma petraea* v. allozyme variation and the pursuit of hybridity. *Heredity 51*(3), 653–663.

Jaynes, E. T. (1968). Prior probabilities. *IEEE Transactions On Systems Science and Cybernetics 4*(3), 227–241.

Johnson, R. W. (2012). MaxEnt power spectrum estimation using the Fourier transform for irregularly sampled data applied to a record of stellar luminosity. *Astrophysics and Space Science 338*, 35–48.

Johnson, V. E. and D. Rossell (2010). On the use of non-local prior densities in Bayesian hypothesis tests. *Journal of the Royal Statistical Society: Series B (Statistical Methodology) 72*(2), 143–170.

Kim, J. and C. D. Scott (2012, Sep). Robust kernel density estimation. *Journal of Machine Learning Research 13*, 2529–2565.

Kuhl, M. E., J. S. Ivy, E. K. Lada, N. M. Steiger, M. A. Wagner, and J. R. Wilson (2010). Univariate input models for stochastic simulation. *Journal of Simulation 4*(2), 81–97.

Lazo, A. V. and P. Rathie (1978, January). On the entropy of continuous probability distributions (corresp.). *IEEE Trans. Inf. Theor. 24*(1), 120–122.

Press, W., S. Teukolsky, W. Vetterling, and B. Flannery (1992). *Numerical Recipes in C* (2nd ed.). Cambridge, England: Cambridge University Press.

Price, A. L., N. J. Patterson, R. M. Plenge, M. E. Weinblatt, N. A. Shadick, and D. Reich (2006, Aug). Principal components analysis corrects for stratification in genome-wide association studies. *Nature Genetics 38*(8), 904–909.

Sivia, D. S. (1996, July). *Data Analysis: A Bayesian Tutorial*. Oxford Science Publications. Oxford, UK: Oxford University Press.

Terrell, G. R. and D. W. Scott (1992). Variable kernel density estimation. *Annals of Statistics 20*(3), 1236–1265.