

Discrimination Based on an Odds Ratio Parameterization

ANGELIKA VAN DER LINDE and GERHARD OSIUS

Universität Bremen, Germany

avdl@math.uni-bremen.de osius@math.uni-bremen.de

SUMMARY

It is argued that the association of any two random elements with positive joint probability density function is characterized by its odds ratio function. The impact of this fundamental result is explored in two applications. In Bayesian analyses it is the association between an observable random variable and a random parameter that is of primary interest. In multivariate analysis it is the association between two random vectors that is investigated. If two random elements are strongly related the corresponding conditional distributions can well be separated. A concept of dependence therefore implies an approach to solving problems of discrimination. Discrimination based on the characterization of association by the odds ratio function is exemplified in Bayesian inference and in multivariate analysis.

Keywords: ASSOCIATION; ODDS RATIO; KULLBACK-LEIBLER DISTANCE; LOGISTIC REGRESSION; MODEL CHOICE; MODEL COMPLEXITY; LOGARITHMIC SCORE FUNCTION; CANONICAL CORRELATION; DISCRIMINANT ANALYSIS.

1. ASSOCIATION AND ODDS RATIOS

1.1. Motivation

How do you present concepts of dependence in your introductory course to statistics?

The best known definitions certainly are those of a correlation coefficient and a covariance matrix. In fact, for jointly Gaussian vectors the covariance matrix does characterize the joint distribution if the marginal distributions are specified. And in multivariate analysis the investigation of structures of dependence is widely based on covariance or correlation matrices.

With respect to discrete variables, for example referring to a 2×2 -table of binary random variables X and Y , statisticians are also familiar with the odds ratio defined by

$$OR := \frac{P(Y = 1 | X = 1)}{P(Y = 0 | X = 1)} / \frac{P(Y = 1 | X = 0)}{P(Y = 0 | X = 0)}. \quad (1)$$

More generally, in a $(M + 1) \times (K + 1)$ -contingency table where (X, Y) takes values in $\{0, 1, \dots, M\} \times \{0, 1, \dots, K\}$, the odds ratio can be “moved around” yielding a matrix with entries

$$OR(m, k) := \frac{P(Y = k | X = m)}{P(Y = 0 | X = m)} / \frac{P(Y = k | X = 0)}{P(Y = 0 | X = 0)}.$$

Given the marginal distributions of X and Y , any matrix with positive entries $OR(m, k)$ defines a unique joint probability distribution. (See the proof by Plackett (1974) based on a result by Sinkhorn, (1967).) Hence the odds ratio matrix like the covariance matrix for jointly Gaussian

vectors describes that part of the joint distribution that is left if the information inherent in both marginal distributions P_X, P_Y is removed. We call that part the “association of X and Y ”. Is there a unique characterization of a joint probability distribution P_{XY} by a triple $(P_X, P_Y, \text{“association”})$ in general? How can association be formally defined?

The analysis of a “mixed” example with binary Y and arbitrary random vector X again suggests that it might be an *odds ratio function* (with reference value x_0)

$$OR^0(x) := \frac{P(Y = 1 | X = x)}{P(Y = 0 | X = x)} / \frac{P(Y = 1 | X = x_0)}{P(Y = 0 | X = x_0)} \quad (2)$$

that captures the association between X and Y . In this case

$$\log \frac{P(Y = 1 | X = x)}{P(Y = 0 | X = x)} = \log \frac{P(Y = 1 | X = x_0)}{P(Y = 0 | X = x_0)} + \log OR^0(x) = \alpha + \log OR^0(x)$$

(say). Hence

$$P(Y = 1 | X = x) = (1 + \exp\{-\alpha - \log OR^0(x)\})^{-1}, \quad (3)$$

and

$$P(Y = 1) = E_x[(1 + \exp\{-\alpha - \log OR^0(x)\})^{-1}]. \quad (4)$$

This expectation is a strictly increasing function of α and approaches the limits 1 respectively 0 as $\alpha \rightarrow \infty$ respectively $\alpha \rightarrow -\infty$. Hence for $0 < P(Y = 1) < 1$ there exists a unique $\infty < \alpha < \infty$ such that (4) holds. This shows that for fixed marginal distribution P_Y the joint distribution of (X, Y) is uniquely determined by the log-odds ratio function. Furthermore, for fixed marginal distributions and a given log-odds ratio function a joint distribution is defined by (3) with α obtained from (4).

1.2. Main Results

Consider two random elements X and Y and assume their joint probability distribution P_{XY} to have a positive density $p(x, y)$ w.r.t. a *product* measure Q . Denote the log-density by $\phi(x, y)$ and define the log-odds ratio function ψ as a function of two pairs (x, y) and (x', y') ,

$$\psi(x, y | x', y') := \phi(x, y) - \phi(x, y') - \phi(x', y) + \phi(x', y'). \quad (5)$$

Under mild regularity conditions the (log-)odds ratio function characterizes the association between X and Y , that is, P_{XY} is characterized by the triple (P_X, P_Y, ψ) . We only sketch the ideas of proof here and refer to Osius (2000) for full details. Uniqueness: Two joint distributions P_{XY} and P'_{XY} with common marginal distributions P_X, P_Y and common log-odds ratio function ψ coincide, if their log-densities are integrable. Indeed, let $p(x, y)$ and $p'(x, y)$ denote the densities of two such joint distributions. Evaluation of the Kullback-Leibler distance

$$I(p, p') = \int \log \frac{p(x, y)}{p'(x, y)} dP_{XY}(x, y)$$

yields $I(p, p') = 0$ implying $P_{XY} = P'_{XY}$. Existence: Given P_X and P_Y and a function ψ , a joint probability density exists (under regularity conditions) such that P_X and P_Y are the corresponding marginal distributions and ψ is the log-odds ratio function. More precisely, the joint density p is obtained as a limit of a sequence of densities (p_n) where the corresponding joint distributions P^n_{XY} are constructed as follows. Start with any P^0_{XY} the log-odds ratio function of which is ψ . Then given P^n_{XY} , replace the marginal distribution P^n_X of X (conditioning) by the

wanted margin P_X and obtain P_{XY}^m . Next replace the other margin P_Y^m by P_Y to obtain P_{XY}^{n+1} which has the same log-odds ratio function ψ as P_{XY}^n . The sequence (p_n) of densities converges to a density p of the required joint distribution P_{XY} (and does not depend on the starting distribution P_{XY}^0). A sufficient condition (which can be weakened) for this reasoning to hold is the integrability of the odds ratio function $\exp \psi$ which can be used as the density of P_{XY}^0 after normalization. This iterative procedure of “marginal fitting” generalizes the method used by Sinkhorn (1967) for marginals with finite support. His proof of convergence however exploits unique features of distributions with finite support which cannot be referred to in the general set-up considered here.

The required joint distribution P_{XY} can also be characterized in another way. For a density $p'(x, y)$ w.r.t. a product measure Q consider the functional

$$l(p') := \int p(x)p(y) \log \frac{p'(x, y)}{p(x)p(y)} dQ(x, y) = -I(p(x)p(y), p'(x, y)).$$

The required joint density then is given as the unique density $p(x, y)$ that maximizes $l(p')$ within the space of all log-integrable densities having ψ as their log-odds ratio function. The functional $l(p')$ is strictly concave and generalizes the log-likelihood used by Haberman (1974, Th.2.6) for distributions with finite support.

1.3. Properties of the Odds Ratio Function

The log-odds ratio function has some desirable properties like the compatibility with 1:1-transformations of X and Y and the invariance under a change of the dominating (product) measure. The log-odds ratio function is already determined by any partial function ψ^0 with fixed reference values (x_0, y_0) ,

$$\psi^0(x, y) := \psi(x, y | x_0, y_0), \tag{6}$$

which can be regarded as representative of ψ . ψ^0 is also already defined using one of the conditional log-densities instead of $\phi(x, y)$ in eq.(5).

1.4. The Odds Ratio Parameterization

Using ψ^0 the log-density ϕ can be decomposed analogously to a control parameterization in an ANOVA-model,

$$\phi(x, y) = \alpha + \beta(x) + \gamma(y) + \psi^0(x, y), \tag{7}$$

where

$$\begin{aligned} \alpha &= \phi(x_0, y_0), \\ \beta(x) &= \phi(x, y_0) - \phi(x_0, y_0), \quad \beta(x_0) = 0, \\ \gamma(y) &= \phi(x_0, y) - \phi(x_0, y_0), \quad \gamma(y_0) = 0, \end{aligned}$$

and

$$\psi^0(x_0, y) = \psi^0(x, y_0) = \psi^0(x_0, y_0) = 0.$$

Thus $\psi^0(x, y)$ corresponds to an interaction term. The log-odds ratio parameterization of conditional and marginal densities then is

$$\log p(y | x) = \log p(y_0 | x) + \gamma(y) + \psi^0(x, y), \tag{8}$$

$$\log p(y) = \alpha + \gamma(y) - \log p(x_0 | y). \tag{9}$$

1.5. Bi-affine Log-Odds Ratio Functions

In standard examples the log-odds ratio function exhibits a simple structure. Let T as superscript denote the transpose of a vector or a matrix.

Example 1

For a Normal conditional distribution $Y | x \sim N(Bx, \Sigma)$, ψ^0 is in general bi-affine,

$$\psi^0(x, y) = (y - y_0)^T \Sigma^{-1} B(x - x_0). \quad (10)$$

For reference values $x_0 = 0, y_0 = 0$ this reduces to the bi-linear form

$$\psi(x, y | 0, 0) = y^T \Sigma^{-1} Bx = y^T \text{cov}(y | x)^{-1} E(y | x). \quad (11)$$

Hence if the joint distribution of X and Y is Gaussian the conditional distributions are Gaussian as well, and the association between X and Y is described by a bi-linear function ψ^0 .

Example 2

If the conditional distribution of Y given x belongs to an exponential family,

$$p(y | x) = a(y) \exp\{x^T t(y) - nM(x)\}, \quad (\text{say}), \quad (12)$$

then

$$\psi^0(x, y) = (x - x_0)^T (t(y) - t(y_0)). \quad (13)$$

Example 3

For simple random variables with finite range bi-affinity holds, too, in a general sense. For X taking values in $\{0, 1, \dots, M\}$ and Y taking values in $\{0, 1, \dots, K\}$ define transformations

$$g(m) = e_{m+1} \in \mathfrak{R}^{M+1}, \quad h(k) = e'_{k+1} \in \mathfrak{R}^{K+1}, \quad (14)$$

where e_i, e'_j denote the i -th respectively j -th unit vector. Choosing $x_0 = 0, y_0 = 0$,

$$\begin{aligned} \psi^0(m, k) &= (e_{m+1} - e_1)^T ((\log p(m | k))) (e'_{k+1} - e'_1) \\ &= (g(m) - g(0))^T ((\log p(m | k))) (h(k) - h(0)). \end{aligned} \quad (15)$$

A similar result is obtained for transformations $g_0(m) = e_m \in \mathfrak{R}^M, g_0(0) = 0_M$ and $h_0(k) = e'_k \in \mathfrak{R}^K, h_0(0) = 0_K$ corresponding to the non-degenerate representation of the multinomial distribution. The entries of the matrix in the inner product then are directly the values $\psi^0(m, k)$ instead of $\log p(m | k)$.

Motivated by these examples we call ψ^0 bi-affine if there are transformations $g : \mathcal{X} \rightarrow \mathfrak{R}^{k_x}, h : \mathcal{Y} \rightarrow \mathfrak{R}^{k_y}$ and a $k_x \times k_y$ -matrix A such that

$$\psi^0(x, y) = (g(x) - g(x_0))^T A (h(y) - h(y_0)). \quad (16)$$

1.6. Logistic Regression

If Y takes values in $\{0, 1, \dots, K\}$ which are coded using h_0 as in example 3 above and if ψ^0 is bi-affine, the log-odds ratio parameterization of the conditional density (8) induces a logistic regression model

$$\log \frac{p(k | x)}{p(k_0 | x)} = \gamma'(k) + g(x)^T a_k, \quad (\text{say}). \quad (17)$$

In this case the model is linear in the transformed values of x and A is a matrix of regression coefficients. Any modeling assumption specifying a structure of the log-odds ratio function corresponds to a (logistic) regression model, and reversely any regression model specifies a structure of ψ^0 .

1.7. Measures of Association and Dependence

So far we have described the association of two random elements X and Y by a function, ψ^0 . How can the strength of association be quantified? In general, a measure of association (or dependence) is given by a functional that assigns to a (log-)odds ratio function a real value, for example an integral. If such a functional does not involve the marginal distributions, we call it a measure of association. If it does, for example, if expectations are taken w.r.t. a marginal distribution, we call it a measure of dependence.

A measure of the strength of the relation between X and Y defined in the same spirit as the association is the “mutual information”

$$I(X, Y) := \int p(x, y) \log \frac{p(x, y)}{p(x)p(y)} dQ(x, y). \quad (18)$$

In the log-odds ratio parameterization

$$I(X, Y) = -\alpha + E_x \log p(y_0 | x) + E_y \log p(x_0 | y) + E_{xy} \psi^0(x, y). \quad (19)$$

The measure is symmetric in X and Y , but it is not a symmetric distance between $p(x, y)$ and $p(x)p(y)$. The symmetrized distance

$$J(X, Y) := \int (p(x, y) - p(x)p(y)) \log \frac{p(x, y)}{p(x)p(y)} dQ(x, y) \quad (20)$$

can be written as

$$J(X, Y) = E_{xy} \psi^0(x, y) - E_x E_y \psi^0(x, y) \quad (21)$$

(for all reference values x_0, y_0) and thus can be expressed in terms of integrals of the log-odds ratio function ψ^0 .

2. ASSOCIATION IN BAYESIAN ANALYSIS

In this section we are interested in the association of an observed random vector Y and a random vector Θ induced by a prior distribution on a parameter θ that determines the conditional density $p(y | \theta)$. In our notation we switch from x to θ throughout this section. Details of the approach sketched here are given in (van der Linde, 2002b).

2.1. Measures of Dependence of Interest

We investigate measures of dependence defined by average Kullback-Leibler distances between probability densities q_1 and q_2 . The Kullback-Leibler distance is also called the *directed divergence* of q_1 and q_2 , and the *divergence* is the symmetrized measure

$$J(q_1, q_2) = I(q_1, q_2) + I(q_2, q_1). \quad (22)$$

The “mutual information” $I(\Theta, Y)$ is seen to be a special application to joint probability density functions of (θ, y) (cp.(18)). As

$$I(\Theta, Y) = E_y I(p(\theta | y), p(\theta)), \quad (23)$$

it is important in Bayesian inference describing the learning process about Θ using y , the transition from the prior to the posterior distribution. We do not focus on the mutual information $I(\Theta, Y)$ though, but want to draw attention to measures of discriminative information, the (directed) divergences between conditional densities.

$$I(p(y|\theta), p(y|\theta')) = E_{y|\theta} \log \frac{p(y|\theta)}{p(y|\theta')} \quad (24)$$

as an average likelihood ratio describes the difference between probability models (sampling distributions) and is of interest in hypothesis testing and model selection. The symmetrized distance $J(p(y|\theta), p(y|\theta'))$ may be interpreted as a measure of variation of sampling distributions (centred at $p(y|\theta')$), thus quantifying the range of modeling assumptions for the distribution of the observed random vector. Dually $J(p(\theta|y), p(\theta|y'))$ quantifies the sensitivity of the posterior distribution to data.

The discriminative information is related to the ‘‘global’’ measure of (symmetrized) mutual information as indicated in (23). For example, in conjugate exponential families the prior distribution and posterior distribution are of the same type, but they differ in (hyper-)parameters. For $p(\theta) = p(\theta|\lambda)$ and $p(\theta|y) = p(\theta|\lambda(y))$, say,

$$J(\Theta, Y) = E_y J(p(\theta|\lambda(y)), p(\theta|\lambda)). \quad (25)$$

Dually also

$$J(\Theta, Y) = E_\theta J(p(y|\theta), p(y)), \quad (26)$$

and in conjugate exponential families the reference density $p(y)$ can be replaced by $p(y|E\Theta)$, that is,

$$J(\Theta, Y) = E_\theta J(p(y|\theta), p(y|E\Theta)). \quad (27)$$

Example 1 (continued)

With Gaussian distributions $Y|\theta \sim N(\theta, \Sigma)$, $\Theta \sim N(0, K)$ for $\theta_0 = E\Theta$,

$$\psi^0(\theta, y) = \text{tr}(\Sigma^{-1}(\theta - E\Theta)(y - y_0)^T). \quad (28)$$

Hence

$$J(\Theta, Y) = \text{tr}(\Sigma^{-1}K), \quad (29)$$

and $E_\theta J(p(y|\theta), p(y|E\Theta))$ is given by

$$\text{tr}(E_\theta[\Sigma^{-1}(\theta - E\Theta)(\theta - E\Theta)^T]) = \text{tr}(\Sigma^{-1}K) = E_{\theta y} \psi(\theta, y|E\Theta, y_0).$$

The relation

$$J(\Theta, Y) = E_{\theta y} \psi(\theta, y|E\Theta, y_0) \quad (30)$$

observed in the example, can be proven to hold in general in exponential families.

2.2. Model Complexity

Turning from prior to posterior expectations with given data y' , it can be shown that an equation like (30) still holds approximately. Thus

$$c(\psi, y') := E_{\theta|y'} E_{y|\theta} \psi(\theta, y|E(\theta|y'), y') \quad (31)$$

can be addressed as a measure of variation of sampling distributions or as a measure of model complexity (given y').

Let \tilde{E} , \tilde{cov} denote expectation and covariance of Θ referring to either the prior or the posterior distribution. Let further $\mathcal{I}(\theta)$, $\mathcal{I}_y(\theta)$ denote an expected respectively observed Fisher information matrix and abbreviate the log-likelihood function $\log p(y | \theta) =: L_y(\theta)$. The argument to establish in general

$$\tilde{E}_\theta E_{y|\theta} \psi(\theta, y | \tilde{E}\Theta, y_0) \approx \tilde{E}_\theta J(p(y | \theta), p(y | \tilde{E}\Theta)), \quad (32)$$

which is an exact equality in the Gaussian example as well as for conjugate exponential families, is based on approximations derived from second order Taylor expansions. It can be briefly sketched as follows.

$$\begin{aligned} & \tilde{E}_\theta E_{y|\theta} \psi(\theta, y | \tilde{E}\Theta, y_0) \\ &= -\tilde{E}_\theta E_{y|\theta} [L_y(\tilde{E}\Theta) - L_y(\theta)] + (-\tilde{E}_\theta [L_{y_0}(\theta) - L_{y_0}(\tilde{E}\Theta)]) \\ &\approx \frac{1}{2} \text{tr}(\mathcal{I}(\tilde{E}\Theta) \tilde{cov}\Theta) + \frac{1}{2} \text{tr}(\mathcal{I}_{y_0}(\tilde{E}\Theta) \tilde{cov}\Theta) \\ &\approx \text{tr}(\mathcal{I}_{y_0}(\tilde{E}\Theta) \tilde{cov}\Theta) \\ &\approx 2\tilde{E}_\theta I(p(y | \theta), p(y | \tilde{E}\Theta)) \\ &\approx \tilde{E}_\theta J(p(y | \theta), p(y | \tilde{E}\Theta)). \end{aligned} \quad (33)$$

A derivation of (33) is given in (Kullback, 1959/1968, pp.26f). Spiegelhalter et al. (2002) suggested

$$p_D(y') := 2E_{\theta|y'} [L_{y'}(E(\theta | y')) - L_{y'}(\theta)] \quad (34)$$

as measure of model complexity and studied its properties and interpretation as “effective degrees of freedom” in many examples. The derivation above shows that $c(\psi, y') \approx p_D(y')$ and thus theoretically substantiates the interpretation of $c(\psi, y')$ respectively $p_D(y')$ as measures of model complexity. The dual structure of the association between Θ and Y and the corresponding duality of measures of dependence furthermore links the Bayesian measure $c(\psi, y')$ to the frequentist approach to model complexity in terms of sensitivity (e.g. Efron, 1986; Ye, 1998). $c(\psi, y')$ and $p_D(y')$ may be regarded as estimates of the measure of model complexity

$$c(\psi) := E_{y'} c(\psi, y') \quad (35)$$

which does not depend on the data and therefore is more appropriate as general definition. For computational purposes though, $p_D(y')$ is most advantageous.

Example 1 (continued) In this case for all y'

$$\begin{aligned} c(\psi) &= c(\psi, y') = p_D(y') \\ &= \text{tr}(\Sigma^{-1} \text{cov}(\theta | y')) = \text{tr}(I + K^{-1}\Sigma)^{-1}. \end{aligned} \quad (36)$$

If $K = \tau^2 K'$ and $\tau^2 \rightarrow \infty$, then $c(\psi) \rightarrow \text{tr} I_q$ where q is the dimensionality of θ ($\theta \in \mathbb{R}^q$), that is, the number of unknown parameters in the sampling model.

Example 2 (continued)

For exponential families we have

$$\begin{aligned} c(\psi, y') &= nE_{\theta|y'}[(\theta - E(\theta | y'))^T \nabla M(\theta)] \\ &= n(\text{tr}[\text{cov}_{\theta|y'}(\theta, E_{\theta|y'}[t(y)])]), \end{aligned}$$

and

$$p_D(y') = 2nE_{\theta|y'}[M(\theta) - M(E(\theta | y'))].$$

The approximation suggests

$$c(\psi, y') \approx p_D(y') \approx \text{tr}(\mathcal{I}_{y_0}(E(\theta | y')) \text{cov}(\theta | y')). \quad (37)$$

2.3. Expected Utilities

The relation (cp. (31) and (33))

$$2E_{\theta|y'} I(p(y | \theta), p(y | E(\theta | y'))) \approx c(\psi, y')$$

and the very definition of $\psi(\theta, y | E(\theta | y'), y_0)$ indicate that $c(\psi, y')$ may serve as a correction term for imputing a representative value like $E(\theta | y')$ for θ in $\log p(y' | \theta)$ and similarly in expected utilities based on the logarithmic score function for probability densities as belief functions for a quantity of interest. In particular in criteria for model choice the decomposition into "model fit" and "model complexity" results from such a corrected imputation.

For example, the Deviance Information Criterion (DIC) introduced by Spiegelhalter et al. (2002) is based on

$$-2E_{\theta|y'} E_{y|\theta} \log p(y | E(\theta | y')) \approx -2 \log p(y' | E(\theta | y')) + 2p_D(y'). \quad (38)$$

DIC is defined by

$$DIC = D(\bar{\theta}) + 2p_D(y'),$$

where D denotes the deviance and the bar indicates the posterior expected value. Consider for illustration the expected loss $-2E_{\theta|y'} E_{y|\theta} \log p(y | \theta)$. Using this shorthand notation a first imputation step

$$-2E_{\theta|y'} E_{y|\theta} \log p(y | \theta) \approx -2E_{\theta|y'} E_{y|\theta} \log p(y | \bar{\theta}) - c(\psi, y') \quad (39)$$

can be derived directly from (33), and a second imputation step also used in the derivation of DIC is

$$\begin{aligned} & -2E_{y|\theta} \log \frac{p(y | \bar{\theta})}{p(y' | \bar{\theta})} \\ &= -2E_{y|\theta} \log \frac{p(y | \bar{\theta})}{p(y | \theta)} - 2E_{y|\theta} \log \frac{p(y | \theta)}{p(y' | \theta)} - 2E_{y|\theta} \log \frac{p(y' | \theta)}{p(y' | \bar{\theta})} \\ &= 2E_{y|\theta} \psi(\theta, y | \bar{\theta}, y') - 2E_{y|\theta} \log \frac{p(y | \theta)}{p(y' | \theta)}. \end{aligned} \quad (40)$$

Neglecting the second term of expectation zero (under $E_{y'|\theta}$), we obtain

$$-2E_{\theta|y'} E_{y|\theta} \log p(y | \bar{\theta}) \approx -2 \log p(y' | \bar{\theta}) + 2c(\psi, y'). \quad (41)$$

Thus $c(\psi, y')$ serves as correction term for imputing y' in $\log p(y | \bar{\theta})$ which amounts to imputing θ in $\log p(\tilde{y} | \bar{\theta})$ twice for $\tilde{y} \in \{y, y'\}$. The imputation is corrected by $c(\psi, y') \approx E_{\theta|y'} J(p(y | \theta), p(y | \bar{\theta}))$, the posterior average symmetrized distance between the corresponding densities. Combining (39) and (41),

$$-2E_{\theta|y'} E_{y|\theta} \log p(y | \theta) \approx -2 \log p(y' | \bar{\theta}) + c(\psi, y'). \tag{42}$$

The argument shows that it is the odds ratio parameterization that provides the formal access to useful approximations of expected utilities. It may lead to special definitions adapted to a particular application like that of $c(\psi, y')$ or $p_D(y')$ in model choice or it may help to link and to compare different approximations.

3. ASSOCIATION IN MULTIVARIATE ANALYSIS

We consider random vectors X and Y and measure the strength of their relation by $J(X, Y)$. We now assume ψ^0 to be bi-affine, referring to the discussion in section 1.5. From (16) and (21) we obtain

$$J(X, Y) = \text{tr}(A\Sigma_{HG}), \tag{43}$$

where $\Sigma_{HG} = \text{cov}_{xy}(h(y), g(x))$. In order to simplify notations we will denote mean vectors by μ and covariance matrices by Σ with appended subscripts in capital letters indicating the random variables. In particular the random vectors resulting from (coding) transformations (cp. 1.5, example 3) will be denoted by $(G_0) G$ and $(H_0) H$ respectively. The matrix $A\Sigma_{HG}$ is a quadratic $(k_x \times k_x)$ -matrix but not necessarily symmetric.

3.1. Linear Discriminant Functions

We aim at a decomposition of $J(X, Y)$ analogously to a principal component analysis (PCA), that is, we want to decompose the trace into components of decreasing importance and would like to relate these components to functions that capture the relation between X and Y in decreasing amounts. To this end we use a singular value decomposition (SVD) of $A\Sigma_{HG}$ which provides rank optimal approximations to $A\Sigma_{HG}$. The eigenvectors of $\Sigma_{GH} A^T A \Sigma_{HG}$ occurring in this SVD then are used as coefficients to form linear combinations of the random variables in G , which we call "linear discriminant functions". The procedure is slightly modified due to the requirement that (the variance of) such discriminant functions should be standardized. Formally therefore we use the following definition.

Definition

Let $\tilde{\Sigma}$ be a positive definite $k_x \times k_x$ -matrix and $\tilde{\Sigma}^{-\frac{1}{2}}$ its (e.g. Gramian) square root. Let r_x denote the rank of $A\Sigma_{HG}$, $r_x \leq \min\{k_x, k_y\}$. Let the SVD of

$$\tilde{C}_x := \tilde{\Sigma}^{\frac{1}{2}} A \Sigma_{HG} \tilde{\Sigma}^{-\frac{1}{2}} \tag{44}$$

be given by $\tilde{C}_x = \tilde{U}_x \Lambda_x \tilde{V}_x^T$, where Λ_x is a $r_x \times r_x$ diagonal matrix with entries λ_j^x , $j = 1, \dots, r_x$ which are the square roots of the eigenvalues of $\tilde{C}_x^T \tilde{C}_x$ in decreasing order. Finally denote the columns of $V_x := \tilde{\Sigma}^{-\frac{1}{2}} \tilde{V}_x$ by $v_{x,j}$, $j = 1, \dots, r_x$. Then the j -th linear discriminant function of X w.r.t. $\tilde{\Sigma}$ is defined by

$$L_j^x(x) := v_{x,j}^T g(x). \tag{45}$$

The term "linear discriminant function" is chosen based on the intuition that functions, say of X , which capture the relation between X and Y can be expected to have a potential for

discriminating y 's given their values. For example, if Y is a group indicator, then L_1^x is supposed to be useful in grouping observed units for which X has been recorded. Kullback (1959/1968) used the term in a similar spirit but under special distributional assumptions, and his work motivated our approach (cp. Kullback, 1959/1968, ch.9.4). Kullback's interest was in testing equality of (Gaussian) sampling distributions in different groups, and he suggested divergences as measures of variation to be used as test statistics. His linear discriminant functions can be shown to result from a special application of our more general definition.

3.2. Standard Situations

In standard situations where X and Y are (conditionally) Gaussian or multinomial, bi-affinity of the log-odds ratio function can be checked directly as exemplified in section 1.5. The matrices A and Σ_{HG} in these cases are obtained explicitly and so are the linear discriminant functions. We summarize some results for these classical set-ups. Proofs and further details are given in (van der Linde, 2002a).

(a) Joint Gaussian distribution of X and Y .

Under the assumption of a joint Normal distribution of X and Y ,

$$J(X, Y) = E_y[(\mu_{X|y} - \mu_X)^T \Sigma_{X|y}^{-1} (\mu_{X|y} - \mu_X)]. \quad (46)$$

The j -th linear discriminant function w.r.t. Σ_X equals the j -th canonical variate in X . Thus in this case the decomposition of $J(X, Y)$ is equivalent to that in canonical correlation analysis (CCA).

(b) Y a grouping indicator, X Gaussian in each group.

Assume Y to take values k in $\{0, 1, \dots, K\}$ and $X|k \sim N(\mu_{X|k}, \Sigma_k)$. If the group specific distributions of X have equal covariance matrices, $\Sigma_k = \Sigma$, say, then

$$J(X, Y) = \text{tr}(\Sigma^{-1} B), \quad (47)$$

where $B = \sum_{k=0}^K p_Y(k) (\mu_{X|k} - \mu_X) (\mu_{X|k} - \mu_X)^T$. The linear discriminant functions in this case coincide with Fisher's discriminant functions. Thus the classical linear discriminant analysis (LDA) proves to be a special case of our approach.

If the conditional covariance matrices are heterogeneous we have the set-up leading to classical quadratic discriminant functions. In fact, ψ^0 is bi-affine but w.r.t. a transformation $g(x)$ that involves not only components X_i but also products $X_i X_j$ of components of X . The bi-affine form is given explicitly, and the use and performance of the linear discriminant functions is illustrated in a numerical example in (van der Linde, 2002a). The heterogeneous set-up is of particular interest in applications and used for instance in model based clustering (Yeung et al, 2001). A common recommendation to check a cluster analysis (e.g. Seber, 1988, p.390) is to use a biplot based on PCA. Linear discriminant functions instead are optimally targeted to grouping, and we suggest to use a biplot based on them for model checking.

(c) X and Y taking finitely many values.

In example 3 of section 1.5, in particular in (15) the log-odds ratio function ψ^0 for simple random functions was shown to be bi-affine, and the matrix A for the non-degenerate transformation h_0 to have entries $\psi^0(m, k)$. The use of linear discriminant functions in this case is illustrated in a problem of seriation in (van der Linde, 2002a).

(d) *Use of linear discriminant functions.*

As indicated in the discussion of standard situations we suggest to use linear discriminant functions for dimension reduction rather than allocation. Their derivation requires the matrix A which is a matrix of regression coefficients under the (linear logistic) regression model corresponding to the assumption of bi-affinity of ψ^0 . This assumption is frequently made for general distributions of X when Y is a grouping variable, and allocation can then be based on the logistic regression model. Dimension reduction is useful e.g. to obtain graphical displays that help to solve problems of seriation, identification and interpretation of groups or in model checking.

3.3. *Duality*

Interchanging X and Y directly yields dual linear discriminant functions, and the special case of CCA illustrates this feature of duality. Qualitatively, however, X and Y can be rather different. In discriminant analysis, for example, Y as a grouping variable typically is discrete and of low dimension whereas X as a vector of many feature variables is continuous or mixed and of high dimension. Hastie et al. (1994, 1995) discussed this problem in the context of LDA and suggested solutions in classical terms.

A switch of variables in our approach yields

$$\psi^0(x, y) = (h(y) - h(y_0))A^T(g(x) - g(x_0)), \tag{48}$$

and hence $J(X, Y) = \text{tr } A\Sigma_{HG} = \text{tr } A^T\Sigma_{GH}$, where now $A^T\Sigma_{GH}$ is a $k_y \times k_y$ -matrix and typically $k_y \ll k_x$. A (standardized) SVD of $A^T\Sigma_{GH}$ yields dual linear discriminant functions,

$$L_j^y := w_j^T H \quad (\text{say}). \tag{49}$$

If H is a coding transformation, then L_j^y represents an "optimal scoring variable" assigning real values w_{jk} to realizations k of Y . The key idea for applications is to define and use approximate dual functions

$$L_j^{y|x} := w_j^T E(H | X = x) \tag{50}$$

which are functions of x again. In general, $E(H | X = x)$ is not linear in x but according to the related regression model $\log E(H | X = x)$ is linearly related to $g(x)$.

4. DISCUSSION

In this concluding section we want to (re-)emphasize some prominent features of the odds ratio parameterization that determine its applications and to point to some experiences using it that suggest further work.

We introduced the odds ratio parameterization as a *universal* formal language which can be useful in any investigation of association or dependence of two random elements. We illustrated its potential in two major fields, both analyses basically referring to the symmetrized mutual information as measure of dependence. Beyond these applications though, its impact has hardly been elaborated. For example, different measures of association and dependence may be of interest, the ideas of sufficiency and ancillarity or for instance of "copulas" should be related. In multivariate analysis there appears to be a dominance of "Normal thinking" in as much as analyses are based on covariance matrices. Alternative approaches in discrete multivariate analysis might be interpreted in terms of association. Therefore our presentation is to be understood as an invitation and stimulation to work out the concept of association, in particular the benefits of an odds ratio parameterization in other fields of research.

The association between two random elements X and Y is defined *symmetrically* and thus induces a *dual* theory. Hence parameterizing with the log-odds ratio function ψ^0 is particularly compatible with a symmetric concept like $J(X, Y)$ but less so with asymmetric or directed analyses like those based on $I(X, Y)$, although these can be expressed in terms of the odds ratio parameterization. Similarly the dual theory can be very helpful in a symmetric setting as demonstrated in discriminant analysis, or it may reveal different features of probability distributions as for example in sensitivity analysis as compared to hypothesis testing. Yet it is certainly worth elaborating deliberately the dual theories to explore their potential whenever an odds ratio parameterization is used.

We emphasized issues of parameterization and the identification of quantities of interest like measures of dependence. We did not even mention how to eventually *estimate* these quantities, and we would like to make some points in this respect. Association is characterized probabilistically, it is inherent in both conditional distributions. Thus the association between Θ and Y is best accessible in a Bayesian approach. The association between X and Y is estimable already using sampling schemes based on one conditional distribution only, a fact that has been much discussed in epidemiology (see the famous paper by Prentice and Pyke, 1978). Measures of dependence like $J(X, Y)$ may require knowledge of the joint distribution, however. The odds ratio parameterization can be helpful in sorting out problems of estimability in restricted sampling schemes. The relation is one-to-one for a third component in the triple (P_X, P_Y, ψ^0) if the two remaining components are fixed, but this relation is unfortunately not explicit in terms of the parameterization. Yet - as is well-known in epidemiology and classical discriminant analysis - the key to estimation is the induced regression model.

REFERENCES

- Efron, B. (1986). How biased is the apparent error rate of a prediction rule? *J. Amer. Statist. Assoc.* **81**, 461–470.
- Haberman, S. J. (1974). *The Analysis of Frequency Data*. Chicago: The University of Chicago Press.
- Hastie, T., Tibshirani, R. and Buja, A. (1994). Flexible discriminant analysis by optimal scoring. *J. Amer. Statist. Assoc.* **89**, 1255–1270.
- Hastie, T., Buja, A. and Tibshirani, R. (1995). Penalized discriminant analysis. *Ann. Statist.* **23**, 73–102.
- Kullback, S. (1959). *Information Theory and Statistics*. Reprinted New York: Dover, 1968.
- van der Linde, A. (2002a). Dimension reduction and linear discriminant functions based on an odds ratio parameterization. *Tech. Rep.*, Univ. of Bremen, Germany, www.math.uni-bremen.de/~avdl.
- van der Linde, A. (2002b). On the association between a random parameter and an observable. *Tech. Rep.*, Univ. of Bremen, Germany, www.math.uni-bremen.de/~avdl.
- Osius, G. (2000). The association between two random elements: A Complete Characterization in terms of Odds Ratios. *Tech. Rep.*, Univ. of Bremen, Germany, www.math.uni-bremen.de/~osius.
- Plackett, R. L. (1974). *The Analysis of Categorical Data*. London: Griffin.
- Prentice, R. L. and Pyke, R. (1979). Logistic disease incidence models and case-control studies. *Biometrika* **66**, 403–412.
- Seber, G. A. F. (1984). *Multivariate Observations*. New York: Wiley
- Sinkhorn, R. (1967). Diagonal equivalence to matrices with prescribed row and column sums. *Amer. Math. Mon.*, **74**, 402–405.
- Spiegelhalter, D. J., Best, N. G., Carlin, B. P. and van der Linde, A. (2002). Bayesian measures of model complexity and fit. *J. Roy. Statist. Soc. B* **64**, 1–34, (with discussion).
- Ye, J. (1998). On measuring and correcting the effects of data mining and model selection. *J. Amer. Statist. Assoc.* **93**, 120–131.
- Yeung, K.Y., Fraley, C., Murua, A., Raftery, A.E., Ruzzo, W.L. (2001). Model-based clustering and data transformations for gene expression data. *Tech. Rep.*, University of Washington, USA.

DISCUSSION

ROBERT E. KASS (*Carnegie Mellon University, USA*)

I would like to preface my discussion by saying how pleased I am to be contributing to this volume in honor of Dennis Lindley. Professor Lindley played a crucial role in the development of Bayesian methods, serving as their chief champion for many years, and gave great encouragement to many aspiring young Bayesians, myself included.

Now, concerning the paper, I must admit I find it difficult. It combines a seemingly technical result, that (P_X, P_Y, ψ) characterizes P_{XY} under weak conditions, with a sweeping conceptual vision, of “[the] odds ratio parameterization as a universal formal language.” On the one hand, it is easy to agree that the log-odds ratio is important, indeed, fundamental; but that by itself is hardly new. There are, here, a series of potentially interesting observations, including remarks about a formal duality between some aspects of frequentist and Bayesian inference (in certain cases), and about some aspects of multivariate analysis. But it is not easy to appreciate the importance of these observations in the absence of some interesting new consequences. Reinterpretation, by itself, is at best tantalizing, and does not necessarily constitute progress.

The authors move from the log-odds ratio to some discussion of its expectation, and the expectation of various log densities. This raises the general question, When is it Bayesianly interesting to consider expectations over the sample space?

I have come across this question, and puzzled over it a bit, in the context of the use of mutual information in the analysis of neuronal data. The Figure displays the firing times of a single neuron under two different experimental conditions (see Olson *et al.* 2000, and Ventura *et al.* 2001). The differential firing rates under the two conditions was the subject of the experiment, and it may be seen that the firing rate is somewhat elevated in the “pattern” condition compared to that in the “spatial” condition toward the end of the given epoch. To be specific, the neuron appears to discriminate between the two conditions in the 200 millisecond time interval (400,600) but not in the interval (0,200). In widely-cited work, Optican and Richmond (1987) suggested that mutual information provides a useful measure for quantifying such temporal contrasts. In the context of this application, let θ be a dichotomous indicator of experimental condition and Y be the data collected over a particular interval, which, for reasons that will become clear in a moment, I would like to denote by e . Optican and Richmond’s suggestion is that we evaluate the amount of information (about the condition) provided by the neuron during the interval e using the mutual information

$$I(\theta, Y | e) = \text{Entropy}(\theta | e) - \text{Entropy}(\theta | Y, e). \quad (1)$$

Many neurophysiologists like this idea, and having thought about it, I do too. But this begs the question, In what sense is mutual information interesting from a Bayesian point of view?

An answer was suggested by Lindley (1956) and Bernardo (1979), who showed that mutual information could be regarded as a Bayesian experimental design criterion. Specifically, for an experimental design e they proposed and studied the criterion of choosing the design to maximize $I(\theta, Y | e)$ given in (1). Thus, evaluating informativeness of data according to $I(\theta, Y | e)$ has a well-established Bayesian interpretation when we consider the alternatives e to amount, essentially, to alternative experimental designs. In the neurophysiological context the analogy with experimental design works well: when we choose alternative intervals of time, we are effectively choosing alternative data to examine. That is, time plays the role of the design variable.

Returning to the more general question, experimental design is also a leading example of a situation in which it is Bayesianly interesting to examine an expectation over the sample space.

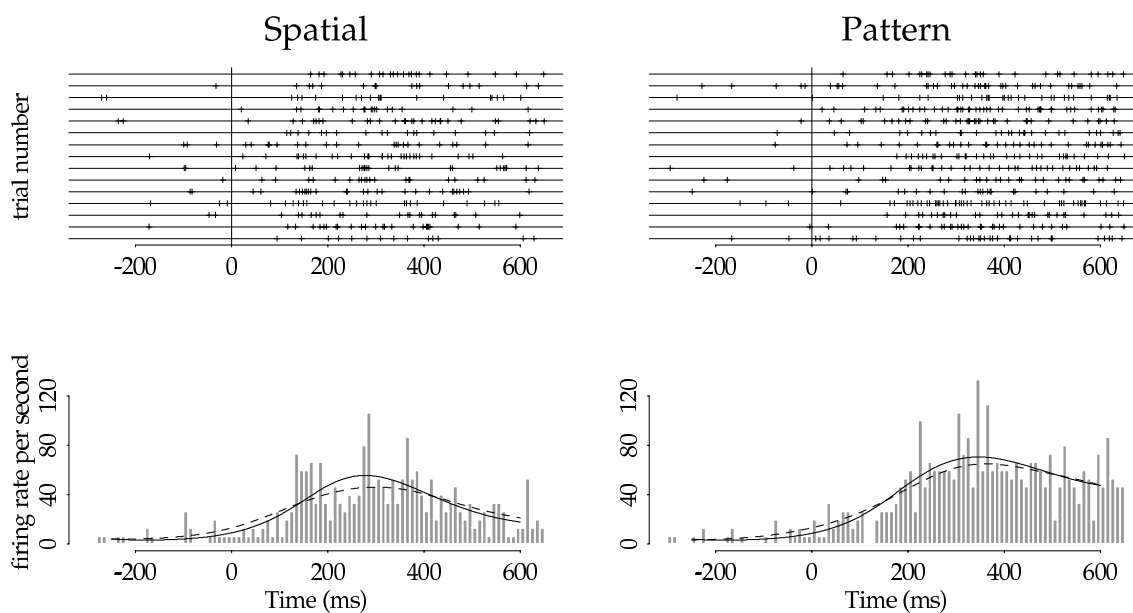


Figure 1. Firing times of a neuron in repeated experimental trials under two experimental conditions. The data are shown in the top portions of the figure as hash marks along lines, each line representing a distinct trial (experimental replication). The bottom portions of the figure display results of pooling the data into 10 millisecond (ms) time bins, pooling across the trials. Smooth curves (using two alternative smoothing methods) are overlaid on the binned-data histograms

A second leading example involves the evaluation of Bayes risk. This brings me to my final substantive comment.

My impression was that a major motivation for this paper was the decision-theoretic interpretation of the DIC criterion, as discussed by Spiegelhalter *et al.* (2002). The essential result, as I understand it, is

$$\text{DIC} \approx E_{\Theta|y} E_{Y|\theta} (-2 \log p(y|\bar{\theta})) \quad (2)$$

where $\bar{\theta} = E(\theta|y)$ is the posterior mean, here used as a “plug-in” estimator. The approximation is not asymptotic, but rather uses some asymptotics while also invoking the *possibility* that a particular term (having zero expectation) is small. I am grateful to Professor van der Linde for emphasizing to me, in personal communication, that the right-hand side of (2) should be considered the risk (under the logarithmic loss) in using the sampling distribution $p(y|\bar{\theta})$ for the future data Y . In connecting this with the frequentist view, it is worth noting that the right-hand side of (2) has the form $E_{\Theta|y} R(\Theta)$ where $R(\theta) = E_{Y|\theta} (-2 \log p(y|\theta))$ is the usual frequentist risk, so that the predictive criterion on the right-hand side of (2) is the posterior mean of the risk.

I found the Spiegelhalter *et al.* paper stimulating and provocative, and think that (2) provides a potentially very interesting interpretation of DIC. But if we take this predictive risk seriously as a model selection criterion, we should ask

- How close is DIC to the predictive risk in (2)? From the interpretation of predictive risk as a posterior mean together with the analysis of Efron (1986) one sees immediately (as Spiegelhalter *et al.* pointed out) that DIC will behave asymptotically like AIC. What more can one say?
- How close is model selection using DIC to model selection based on risk?
- How can we numerically approximate model selection based on risk?

In summary, while I find many of the results, including those on the formal “duality” of

frequentist and Bayesian inference, intriguing, I feel I would need to see more in order to be convinced of the value of these interpretations.

I would like to close by mentioning my own view that model selection is hard when we have limited data and we don't know how (or are unwilling) to follow the subjectivist prescription, i.e.,

- We use default priors *that matter* (they matter much more than in estimation problems);
- And/or we use default utilities *that matter*.

The second item makes it particularly hard to do numerical comparisons: we are continually facing the tautology that a model selection method will perform well according to the criterion that defines it. Thus, we continue to be presented with competing model selection criteria, with each appearing sensible to its proponents, and we are unable to find any basis for reaching a consensus.

My guess is that this is an inherently insoluble problem. Is there any way forward? Perhaps case studies might help. There, one would have to present a specific scientific problem with well-justified statistical goals that require some kind of evaluation of alternatives models. In such a specific context it may be possible to argue convincingly that a particular method of model selection is more helpful than others in achieving those goals. An attempt of this sort was made in Viele *et al.* (2002), but I hope others will present different, and more informative case studies in the future.

REPLY TO THE DISCUSSION

We agree that the paper is difficult because it briefly indicates rather than spells out the impact of our results. We did elaborate on our findings in subsequent papers but take the opportunity to point again to some applications.

First of all we do not think that the characterization of P_{XY} by (P_X, P_Y, ψ^0) is technical. The log-odds ratio is indeed fundamental, and the result shows to what extent. In contrast to (previously) wide spread beliefs it turns out that the odds ratio is *the* parameter of interest whenever the association between two random elements is to be studied in terms of (conditional) densities. Furthermore we demonstrate under which modelling assumptions (*logistic* regression models) and sampling schemes (*conditional* and *joint* sampling) it is estimable. These results are valid for rather general distributions but have been known and used in restricted set-ups only, namely for (simple) random variables X and Y with finite range (i.e. contingency tables). For a random vector X and simple Y (finite range) log-odds ratios have been used in logistic regression to model the conditional distribution $P_{Y|X}$. But even in this special case our characterization of the joint distribution P_{XY} in terms of odds ratios has not been given so far. And in the general situation with arbitrary random vectors (or even arbitrary random elements, cp. Osius (2000)) X and Y the characterization of P_{XY} and the resulting odds ratio models like e.g. bi-affine models have not been given before.

In consequence we see the benefit of the odds ratio parameterization in its potential for *distributionally* adequate generalizations and the reinterpretation of known results as guidance to such generalizations. For example, the definition of linear discriminant functions is validated by the identification of CCA and LDA as special cases, but it provides a general approach for arbitrary distributions with bi-affine log-odds ratio functions, which was exemplified for multinomial distributions. Hence, e.g. new diagnostic (bi-)plots are suggested which take into account the distributional assumptions of the model. Similarly, identifying $J(X, Y)$ as a (transformed) coefficient of determination in Gaussian multiple regression guides its generalization to non-Gaussian regression and hence induces procedures of variable selection that are based on the contribution of each variable X_i to $J(X, Y)$.

Also, the discussion of model complexity offers clarity and mathematical foundation that was missing in the introduction of $p_D(y')$. In this way some issues of the related discussion can be settled. $p_D(y')$ turns out to be an *estimate* of a well interpretable quantity (posterior version of symmetrized mutual information) which can be derived independently and is invariant under one-to-one transformations of the parameter of interest θ . Related quantities like the one suggested by M. Plummer (in the discussion of Spiegelhalter et al., 2002) based on intuition can easily be qualified as being equal to $c(\psi, y')$ in exponential families.

Turning to the comparison of DIC with the predictive risk in (2) of the discussion we can only give a partial answer. Two “errors” can cause a difference: (i) the neglected term

$$-2E_{\theta|y'}E_{y|\theta} \log \frac{p(y|\theta)}{p(y'|\theta)}$$

(cp. eg.(40)) and (ii) the estimation error $c(\psi, y') - p_D(y')$. It is certainly worth studying for which type of distribution these terms may effect the decision.

Although we agree that to a pragmatic statistician ‘progress’ may not be obvious, we insist on and claim mathematical progress. Establishing an “economy of thought” derived from general mathematical structures has always been a genuine task for mathematicians. Foundational mathematical views allow for structurally well justified results and procedures in future work. For example, establishing links between reference priors on model (hyper-)parameters and model priors as decreasing functions of model complexity requires a formalization in which parameters of interest and their approximations respectively estimates can be well separated. We do believe that the log-odds ratio parameterization provides such a tool for proofs.

We agree that this point of view aims at objective Bayesian procedures (targeting parsimony of a model in model choice for example) rather than subjectivist specifications ‘that matter’. While we prefer substantial prior specifications whenever possible, we also see a need for well founded default priors (e.g. on model (hyper-)parameters).

Finally we would like to (re-)emphasize that the duality of the frequentist and Bayesian approach results in coherent inference just about the association between Θ and Y respectively about ψ^0 obtainable from both conditional distributions - i.e. from the likelihood as well as the posterior distribution. Discarding P_Y and P_Θ means not to refer to sampling expectations and not to invoke a(n informative) prior.

ADDITIONAL REFERENCES IN THE DISCUSSION

- Bernardo, J. M. (1979). Expected information as expected utility. *Ann. Statist.* **7**, 686–690.
- Lindley, D. V. (1956). On the measure of information provided by an experiment *Ann. Statist.* **27**, 986–1005.
- Olson, C. R., Gettner, S. N., Ventura, V., Carta, R. and Kass, R. E. (2000). Neuronal activity in macaque supplementary eye field during planning of saccades in response to pattern and spatial cues. *J. Neurophysiology* **84** 1369–1384.
- Optican, L. M. and Richmond, B. J. (1987). Temporal encoding of two-dimensional patterns by single units in primate inferior temporal cortex. III: Information-theoretic analysis. *J. Neurophysiology*, **57** 162–178.
- Ventura, V., Carta, R., Kass, R. E., Gettner, S. N. and Olson, C. R. (2002). Statistical analysis of temporal evolution in single-neuron firing rates. *Biostatistics*, **3** 1–20.
- Viele, K., Kass, R. E., Tarr, M. J., Behrmann, M., and Gauthier, I. (2002). Recognition of faces versus Greebles: A case study in model selection *Case Studies in Bayesian Statistics VI*, (Gatsonis, C., Kass, R.E., Carriquiry, A., Gelman, A., Higdon, D., Pauler, D., and Verdinelli, I., eds.), New York: Springer, 91–136, (with discussion).