Information importance of predictors: Concept, measures, Bayesian inference, and applications

J.J. Retzer a, E.S. Soofi b,c,∗, R. Soyer d

a Maritz Research, 1815 S. Meyers Road, Suite 600, Oakbrook Terrace, IL 60181, USA
b Sheldon B. Lubar School of Business, University of Wisconsin-Milwaukee, P.O. Box 742, Milwaukee, WI 53201, USA
c Center for Research on International Economics, University of Wisconsin-Milwaukee, P.O. Box 742, Milwaukee, WI 53201, USA
d Department of Decision Sciences, George Washington University, Washington, DC 20052, USA

ARTICLE INFO

Article history:
Available online 13 March 2008

ABSTRACT

The importance of predictors is characterized by the extent to which their use reduces uncertainty about predicting the response variable, namely their information importance. The uncertainty associated with a probability distribution is a concave function of the density such that its global maximum is a uniform distribution reflecting the most difficult prediction situation. Shannon entropy is used to operationalize the concept. For nonstochastic predictors, maximum entropy characterization of probability distributions provides measures of information importance. For stochastic predictors, the expected entropy difference gives measures of information importance, which are invariant under one-to-one transformations of the variables. Applications to various data types lead to familiar statistical quantities for various models, yet with the unified interpretation of uncertainty reduction. Bayesian inference procedures for the importance and relative importance of predictors are developed. Three examples show applications to normal regression, contingency table, and logit analyses.

© 2008 Elsevier B.V. All rights reserved.

1. Introduction

Assessment of the relative importance of explanatory variables is very common in reports of research studies in numerous fields (Kruskal and Majors, 1989). In real-world practice, attribute relative importance assessment is a mainstay in many decision making situations. Relative importance measures, proposed by statisticians, econometricians, educational psychologists, decision scientists, and others, refer to quantities that compare the contributions of individual explanatory variables to the prediction of a response variable.

Thus far, the relative importance methodology literature has focused on developing “relative” importance measures for specific problems, mainly regression (Azen and Budescu, 2003; Genizi, 1993; Johnson, 2000; Kruskal, 1984, 1987; Lindeman et al., 1980; Pratt, 1990; Theil and Chung, 1988; Grömping, 2007). Specific measures for other problems include logit (Soofi, 1992, 1994), survival analysis (Schemper, 1993), ANOVA (Soofi et al., 2000), and time series (Pourahmadi and Soofi, 2000). Some attempts have been made to define requirements and properties of relative importance measures: game-theoretic type axioms for risk allocation (Cox, 1985; Lipovetsky and Conklin, 2001), Dominance Analysis for linear regression (Budescu, 1993), and Analysis of Importance (ANIMP) framework (Soofi et al., 2000). Little attention, however, has been given to characterizing the more general, underlying notion of “importance” itself. The lack of a unifying concept of

∗ Corresponding author at: Sheldon B. Lubar School of Business, University of Wisconsin-Milwaukee, P.O. Box 742, Milwaukee, WI 53201, USA. Tel.: +1 414 229 4281.
E-mail addresses: retzerjj@maritz.com (J.J. Retzer), esoofi@uwm.edu (E.S. Soofi), soyer@gwu.edu (R. Soyer).

0167-9473/$ – see front matter © 2008 Elsevier B.V. All rights reserved. doi:10.1016/j.csda.2008.03.010
importance is consequential for practice. At present, "importance" is interpreted differently in different problems (e.g., linear regression, ANOVA, logit). The wide spectrum of problems encountered in research and practice requires a general concept of importance which provides measures that admit a common interpretation in various applications.

We conceptualize importance in terms of the information provided by a predictor for reducing the uncertainty about predicting the outcomes of the response variable. The information as a general probabilistic concept provides measures of importance for categorical and discrete variables, as well as continuous variables regardless of whether or not their distributions are normal. For nonstochastic predictors, Maximum Entropy (ME) characterization of probability distributions provides measures of information importance. For the case of exponential family regression the ME formulation leads to the deviance measure. For stochastic predictors, the expected entropy difference gives measures of information importance, which are invariant under one-to-one transformations of the variables. Theil and Chung (1988) introduced a logarithmic function of the squared correlation in the relative importance literature, which is the information importance measure for normal regression. We will show that the invariance property of expected information makes this measure applicable to non-normal variables if normality can be achieved by one-to-one transformations of the variables.

The information measures are functions of the model parameters, and hence subject to inference. Bayesian inference about the information importance is proposed. The posterior distributions of the importance measures are computed from the posterior distributions of the parameters. The procedure is computational. The posterior outcomes of information measures are simulated from the joint posterior distribution of the model parameters. In addition, when the posterior distribution of the model parameters is not available analytically, Markov Chain Monte Carlo (MCMC) is needed.

Section 2 presents the notion of information importance. Section 3 presents the expected information measure for stochastic predictors with a subsection on the summarizes the paper and gives some concluding remarks.

2. Notion of information importance

Let \( \mathbf{x} = (x_1, \ldots, x_p)' \) be a vector of predictors of a variable \( Y \), where the prediction is probabilistic. The importance of a predictor \( X \) for \( Y \) is the extent to which the use of \( \mathbf{x} \) reduces the uncertainty in predicting outcomes of \( Y \). We conceptualize uncertainty in terms of unpredictability of outcomes of \( Y \). The most unpredictable situation is when all possible values (intervals of equal width in the continuous case) of \( Y \) are equally likely. This establishes uniformity of the probability distribution as the reference point for quantifying the uncertainty in terms of predictability.

The uncertainty associated with a probability distribution \( F \) having a density (mass) function \( f \) is defined by \( U(f) \leq U(f^*) \), such that \( U(f^*) \) is concave and \( f^* \) is the uniform density (possibly improper). That is, \( U(f) \) is a measure of uniformity (lack of concentration) of probabilities under \( F \) (Ebrahimi et al., 2007).

Without the predictors, the probabilistic prediction of the response is made based on the distribution \( F_y \) having a density (mass) function \( f_y \). With the predictors, the prediction is made based on the distribution \( F_{y|x} \), which depends on \( \mathbf{x} \) but not on the position of \( x_1 \) in the vector, and has a density (mass) function \( f_{y|x} \). For a stochastic predictor, \( F_{y|x} \) is the conditional distribution and \( f_{y|x} = E[f_y|X] \). For a nonstochastic predictor such a relationship is absurd.

The worth of \( \mathbf{x} \) for the prediction of \( Y \) is mapped by the uncertainty difference \( \Delta U(Y; \mathbf{x}) = U(f_y) - U(f_{y|x}) \), which does not depend on the position of \( x_1 \) in the vector. In general, \( \Delta U(Y; \mathbf{x}) \) can be positive, negative, or zero. When a predictor makes prediction more difficult, the verdict on its information importance worth is clear, hence \( \Delta U(Y; \mathbf{x}) < 0 \) is of no particular interest in the present context. We provide formulations that are sufficiently general satisfying \( \Delta U(Y; \mathbf{x}) \geq 0 \) and give non-negative information importance functions. The proper information importance of a predictor vector \( \mathbf{x} \) is defined by the following property:

\[
I_U(Y; \mathbf{x}) = U(f_y) - U(f_{y|x}) \geq 0.\]

For a stochastic predictor the information importance of outcomes \( \mathbf{x} \) of \( \mathbf{X} \) for predicting \( Y \) is given by the expected uncertainty change

\[
I_U(Y|X) = E_x[\Delta U(Y|X)] = U(f_y) - E_x[U(f_{y|x})] \geq 0,
\]

where the inequality changes to equality if and only if \( \mathbf{X} \) and \( Y \) are independent. The non-negativeness is implied by concavity of \( U \) and it characterizes the expected gain of using the outcomes of \( \mathbf{X} \) for the prediction. It is reasonable to require that using the outcomes of \( \mathbf{X} \), on average, will yield some information useful for making predictions about \( Y \). At worst, the long-run use of a variable has no information importance for predicting the outcomes of another variable (DeGroot, 1962).

For any subvector of length \( r < p \), the incremental (partial) contribution of \( x_{r+1}, \ldots, x_p \) to the information importance of \( (x_1, \ldots, x_r) \) is given by

\[
I_U(Y; x_{r+1}, \ldots, x_p | x_1, \ldots, x_r) = U(Y; x_1, \ldots, x_r) - U(Y; x_1, \ldots, x_p) \geq 0, \quad r < p.
\]

The equality is apparent (add and subtract \( U(f_y) \)) and inequality is implied by the properness \( \Delta U(Y; \mathbf{x}) \geq 0 \). We therefore have the decomposition property,

\[
I_U(Y; x_1, \ldots, x_p) = I_U(Y; x_1, \ldots, x_r) + I_U(Y; x_{r+1}, \ldots, x_p | x_1, \ldots, x_r).
\]
Successive application of (2) gives the following chain rule:

$$I_U(Y; x_1, \ldots, x_p) = \sum_{k=1}^p I_U(Y; x_k|x_1, \ldots, x_{k-1}),$$

(3)

where $I_U(Y; x_k|x_0) = I_U(Y; x_1)$, and $I_U(Y; x_k|x_1, \ldots, x_{k-1})$ is the incremental contribution of $x_k$ to the information importance of $(x_1, \ldots, x_k)$.

The incremental information function $I_U(Y; x_k|x_1, \ldots, x_{k-1})$ provides measures of the relative importance of predictor $x_k$ in the sequence $x_1, \ldots, x_p$. The Analysis of Importance (ANIMP) framework proposed by Soofi et al. (2000) encapsulates two properties found to be desirable by many researchers in the relative importance literature: additively separable, and order-independence in the absence of a natural ordering. The additive decomposition (3) is a general representation satisfying the first property. However, in general, decomposition (3) depends on the position of $x_k$ in $(x_1, \ldots, x_p)$, so it does not satisfy order-independence. For satisfying the order-independence condition of ANIMP, the relative information importance can be computed by an averaging over all orderings of the explanatory variables:

$$I_U(Y; x_k) = \sum_{q=1}^p w_q I_U(Y; x_k|x_1, \ldots, x_{k-1}; O_q),$$

(4)

where $w_q$ is the weight attached to the importance of $x_k$ in the arrangement of the $n$ predictors $O_q$, $q = 1, \ldots, p!$. The most commonly used weights are uniform, justified on various grounds, including "tradition in statistics" (Kruskal, 1987), game theoretic axioms (Cox, 1985), mathematical argument (Chevan and Sutherland, 1991), and the maximum entropy principle (Soofi et al., 2000). The use of unequal weights is equally plausible.

3. Maximum entropy information

An uncertainty function is Shannon entropy $H(F) = H(F)$, defined by

$$H(Y) = H(F) = -\int \log f(y) dF(y),$$

(5)

where $dF(y) = f(y) dy$ for the continuous case and $dF(y) = f(y)$ for the discrete case. The entropy maps the concentration of probabilities under $F$ and decreases as concentration increases, thus $-H(F)$ is a measure of informativeness of $F$ about $y$ (Zellner, 1971, 1997).

In order to assess the information importance of a predictor $x$, we consider a vector containing a set of linearly independent information moments,

$$T^*(Y; \mathbf{x}) = [T_1^*(Y), T_2^*(Y; \mathbf{x})] = [T_1(Y), T_2(Y), T_{A+1}(Y; \mathbf{x}), \ldots, T_{A+B}(Y; \mathbf{x})],$$

where $T_k(Y)$, $k = 1, \ldots, A$ and $T_k(Y; \mathbf{x})$, $A + 1, \ldots, A + B$ are real-valued integrable functions with respect to $dF_y$ and $dF_{T_k; \mathbf{x}}$, respectively. Examples include $T(Y) = Y$, $T(Y) = Y^2$, $T(Y) = \log Y$, and $T(Y) = \delta(z)$, where $\delta(z)$ is an indicator function of a subset of support of $F$, and for a single predictor, $T(Y; x) = x f$ and $T(Y; x) = \log(1+x) Y$, provided that they are all integrable.

The information moment set $T$ generates a class of distributions:

$$\Omega_{T; \mathbf{x}} = \{F : E_{T; \mathbf{x}}[T_k(Y; \mathbf{x})] = 0, k = 1, \ldots, A + B\},$$

where $\theta_k(\mathbf{x})$, $k = 1, \ldots, A$, $A + 1, \ldots, A + B$ are specified moments in terms of $\mathbf{x}$.

The Maximum Entropy (ME) model in $\Omega_{T; \mathbf{x}}$ is the distribution $f_{T; \mathbf{x}}$, whose density maximizes (5). The ME model, if it exists, is unique and has density in the following form:

$$f_{T; \mathbf{x}}(y) = C(\lambda(\mathbf{x}), \beta(\mathbf{x})) \exp \left[ -\lambda(\mathbf{x}) T(Y) - \beta(\mathbf{x}) T(Y; \mathbf{x}) \right],$$

(6)

where $[\lambda(\mathbf{x}), \beta(\mathbf{x})] = [\lambda_1(\mathbf{x}), \ldots, \lambda_A(\mathbf{x}), \beta_1(\mathbf{x}), \ldots, \beta_B(\mathbf{x})]$ is the vector of Lagrange multipliers and $C(\lambda(\mathbf{x}), \beta(\mathbf{x}))$ is the normalizing factor. When $\theta_k(\mathbf{x}) = 0$, $k = 1, \ldots, A$ are free from $\mathbf{x}$, (6) gives $f_{T; \mathbf{x}}(y) = C(\lambda) \exp \left[ -\lambda T(Y) \right]$, with $\lambda = (\lambda_1, \ldots, \lambda_A)$ free from $\mathbf{x}$, which is the ME model in the class of distributions $\Omega_{T} \supset \Omega_{T; \mathbf{x}}$ generated by $T(Y)$.

Let $H^*(Y) = H(F_1^*)$ and $H^*(Y; \mathbf{x}) = H(F_{T; \mathbf{x}})$.

$$I_{\theta}(Y; \mathbf{x}) = I_{H^*(Y; \mathbf{x})} = H^*(Y) - H^*(Y; \mathbf{x}) \geq 0,$$

(7)

where $\theta$ denotes the vector of all parameters involved. The equality in (7) is due to the additional constraints reducing the maximum entropy (Jaynes, 1957, 1968; Soofi, 1992, 1994). Clearly, $I_{\theta}(Y; \mathbf{x})$ admits the chain rule decomposition (2).

The quantity $I_{\theta}(Y; \mathbf{x})$ provides measures of information importance for various types of data and models, all with the same interpretation. For example, Ebrahimi et al. (2007b) have shown that distributions with densities in the exponential
family having finite entropy are ME in appropriately defined $\Omega$. The constraints can be formulated such that moment values are statistics $\theta_i = \hat{\theta}_i$ (see e.g., Soofi (1992)). Then for the exponential family regression, we obtain

$$I_\theta(Y; x) = 2 \ln \left[ H_{\theta_0}(F_y) - H_{\theta}(F_{y,x}) \right]$$

$$= -2 \ln \left[ \frac{f_{y,X}(Y)|_{\theta=\hat{\theta}}}{f_{y}(Y)|_{\theta=\hat{\theta}}} \right]$$

$$= 2\tilde{K}(F_{y,X}: F_y).$$

The middle quantity is the likelihood ratio statistic and $\tilde{K}(F_{y,X}: F_y)$ is an estimate of the Kullback–Leibler information (relative entropy),

$$K(Y; x) = K(F_{y,X}: F_y) = \int \log \frac{f_{y,X}(y)}{f_y(y)} dF_{y,X}(y),$$

known as the deviance in the exponential family regression literature. Thus, by (7), the deviance is also an estimate of the ME difference providing a measure of the information importance of predictors in terms of uncertainty reduction for the exponential family regression. We should note that (7) is a general information importance measure applicable to any ME distribution, beyond the exponential family regression. Any distribution with a density in the form of (6) having finite entropy is an ME model (Ebrahimi et al., 2007b).

Normalized information indices map $I_\theta(Y; x)$ into the unit interval. For the discrete case, the information importance index is defined by the fraction of uncertainty reduction due to $x$:

$$I(Y; x) = 1 - \frac{H(Y; x)}{H(Y)} = \frac{I_\theta(Y; x)}{I_\theta(Y; \emptyset)},$$

For the continuous case the entropy reduction index (12) is not meaningful, and the information index is computed by exponential transformation:

$$I(Y; x) = 1 - e^{-2I_\theta(Y; x)}.$$  

In both cases the indices range from zero to one: $I(Y; x) = 0$ mapping the case when the predictor does not reduce the uncertainty at all, and $I(Y; x) = 1$ mapping the case when the predictor reduces the uncertainty completely. The entropy reduction index (12) does but the exponential transformation index does not satisfy the additive decompositions (2) and (3).

3.1. Exponential regression

The ME model subject to constraint $E(Y) = \theta_1$ is the exponential distribution with density $f_\theta(y) = \lambda e^{-\lambda y}$, where the Lagrange multiplier is given by $\lambda = \theta_1^{-1}$. The maximum entropy is $H^*_{\theta_1} = 1 - \log \lambda$. The ME model subject to the additional constraint $E(x) = \theta_2$ is the exponential distribution with density $f_{\theta_2}(x) = \lambda(x) = \theta_0 + \beta x$. The maximum entropy is $H^*_{\theta_2} = 1 - \log \lambda(x)$ and $\theta(x) = (\theta_0 + \beta x)^{-1}$.

The information importance of predictor $x$ is

$$I_\theta(Y; x) = H_{\theta_1}^* - H_{\theta_2}^* = -\log \frac{\lambda}{\lambda(x)} = -\log \frac{\theta(x)}{\theta_1} \geq 0.$$  

For a sample of $n$ observations, using maximum likelihood estimate (MLE) $\hat{\lambda}$ and $\hat{\lambda}(x)$ we have the ME information importance in terms of the log-likelihood ratio statistic (9) and deviance (10).

3.2. Log-linear and logit models

In a $d_y \times d_{x_1} \times \cdots \times d_{x_p}$ contingency table, the response and predictors are all categorical variables with $d_y$, $d_{x_1}$, . . . , $d_{x_p}$ categories, respectively. The ME problem pertains to estimation of a single probability vector $\pi = (\pi_1, \ldots, \pi_n)$ for a vector of indicator functions $Y = (y_1, \ldots, y_d)$, where $y_i \in \{0, 1\}$, $d = d_y \times d_{x_1} \times \cdots \times d_{x_p}$, and $\sum_{j=1}^{d} \pi_j = \sum_{j=1}^{d} y_j = 1$. The information moments are $T(Y; x) = [T_1(x), \ldots, T_d(x)]$, where $T_i(x) \in \{0, 1\}$ is a cell indicator function. For $B < d$ the ME solution is unique; it can be written as a logit or a log-linear model: $\pi_{y}^* = g(C\beta) - B^T (Y; x)$. (Note that the log-linear representation is not unique; sets of linearly equivalent constraints lead to different log-linear representations all providing the same ME solution for the probabilities.)

In general, logit analysis pertains to the prediction of $n$ vectors of indicator functions $Y_i = (y_{i1}, \ldots, y_{id})$, $i = 1, \ldots, n$. The following ME formulation solve a solution corresponding to the standard ecometric specification of a general logit (Soofi, 1992, 1994). Suppose that $y_{ij} \in \{0, 1\}$ is the indicator of the choice of an individual $i$ among a set of alternatives and the predictors are $x_i = (u_i, v_j)$, where $u_i = (u_{i1}, \ldots, u_{ia})$ is a set of individual's attributes and $v_j = (v_{j1}, \ldots, v_{jb})$, $j = 1, \ldots, j$.
is a set of scores (values) assigned to the attributes of the jth alternative. The ME constraints in terms of these predictors for logit are

\[
T_{a1}(\mathbf{x}) = [T_{a11}(\mathbf{x}), \ldots, T_{a1n}(\mathbf{x})], \\
\vdots \\
T_{a(j-1)1}(\mathbf{x}) = [T_{a(j-1)1}(\mathbf{x}), \ldots, T_{a(j-1)n}(\mathbf{x})], \quad a = 1, \ldots, A
\]

\[
T_b(x) = [T_{b1}(\mathbf{x}), \ldots, T_{bn}(\mathbf{x})], \quad b = 1, \ldots, B,
\]

where \( T_{a(i)}(\mathbf{x}) = (u_{ai}, 0, \ldots, 0), \) \( T_{a(j-1)i}(\mathbf{x}) = (0, \ldots, 0, u_{ai}, 0), i = 1, \ldots, n, a = 1, \ldots, A, \) are \((j-1)A\) constraints for the individual’s attributes and \( T_b(x) = (v_{1b}, \ldots, v_{nb}), i = 1, \ldots, n, b = 1, \ldots, B. \) Since for each individual, \( u_{ai} \) remains constant across the alternatives, each \( u_{ai} \) requires \( j - 1 \) constraints. However, since the choice attributes vary across the alternatives, each requires one constraint. For a display of the constraint matrix see Soofi (1994).

Then the ME solution is the following logit model:

\[
\pi^*_y = \frac{e^{\alpha_i + \beta^t \mathbf{v}_i}}{\sum_{i=1}^{n} e^{\alpha_i + \beta^t \mathbf{v}_i}},
\]

where \( \alpha_j, j = 1, \ldots, J - 1 \) and \( \beta \) are the vectors of \((J - 1)A + B\) Lagrange multipliers; i.e., logit coefficients Soofi (1992, 1994). In econometric terminology the coefficients of an individual’s attributes are not identifiable for all alternatives. The coefficient for one of the alternatives is found by a side condition, e.g., \( \alpha_j = 0 \) or \( \alpha_j = - \sum_{j=1}^{J-1} \alpha_j. \)

When the constraints’ values are set equal to the sample statistics, the ME results are the MLE of the logit model (15) assumed a priori; details are given in Soofi (1992, 1994). Then the information importance of predictors is given by the log-likelihood statistic (9) and deviance (10).

The maximum uncertainty for the joint distribution of \( Y_i = (y_{i1}, \ldots, y_{ij}), i = 1, \ldots, n, \) is given by

\[
H(Y, x) = \sum_{i=1}^{n} H(Y, x_i) = - \sum_{i=1}^{n} \sum_{j=1}^{n} \pi^*_y \log \pi^*_y.
\]

For the MLE formulation, the log-likelihood function is \(-H(Y, x)\). When the choice attributes are present in the problem, (15) cannot include an intercept term because it creates a singularity. In this case the null probabilities are all uniform \( \pi^*_y = \frac{1}{J} \) for all \( i, j, \) and \( H(Y) = n \log J \) is the entropy of \( n \) uniform distributions and is the global maximum under no constraint. In this case the null log-likelihood function equals \(-H(Y)\). When there is no choice attribute in the problem one may include \( J - 1 \) constraints so that (15) includes an intercept term for each alternative and null gives sample proportions \( \pi^*_y = \hat{\pi} = (\hat{\pi}_1, \ldots, \hat{\pi}_J) \) for all \( i = 1, \ldots, n. \) In this case the null log-likelihood function equals \(-H(Y) = -nH(\hat{\pi}).\)

4. Expected information

For the case of stochastic predictors, \( F_{Y|x} \) is the conditional distribution \( F_{Y|x} \) and all information quantities are conditional on \( x \) and, as functions of \( X \) are stochastic. The expected value of \( \Delta_{ij}(Y|X) \) and the expected value of (11) are equal, and the unique measure is referred to as the mutual information between \( Y \) and \( X \).

\[
M(Y, X) = E_X[\Delta_{ij}(Y|X)] = E_X[K(Y|X)].
\]

Other useful and insightful representations of the mutual information are:

\[
M(Y, X) = K(F_{X,Y} : F_{X}F_{Y})
\]

\[
= H(X) + H(Y) - H(X, Y)
\]

\[
= H(Y) - H(Y|X),
\]

where \( H(Y|X) = E_X[H(Y|X)] \) is referred to as the conditional entropy. By (18), \( M(Y, X) \geq 0 \) is the information divergence between the joint distribution \( F_{X,Y} \) and the product of marginals \( F_XF_Y. \) Thus, \( M(Y, X) = 0 \) if and only if \( Y \) and \( X \) are stochastically independent. Representation (19) facilitates the computation of mutual information and (20) depicts the expected uncertainty reduction interpretation.

The normalized indices (12) and (13) are applicable to \( M(Y, X) \); in this case \( l(Y, X) = 0 \) if and only if the two variables are independent, and \( l(Y, X) = 1 \) if and only if the two variables are functionally related in some form, linearly or non-linearly. The mutual information admits the chain-rule decomposition of type (2); see Cover and Thomas (1991).

An important property of the mutual information, apparent from (18), is invariance under one-to-one transformations of the variables. For example, let \( Y = S(W) \) and \( X_j = T_j(V_j), j = 1, \ldots, k, \) where \( S \) and \( T_j \) are one-to-one transformations. Then,

\[
\]

The invariance is a powerful property in the present context in that the importance of an explanatory variable is independent of the functional form of the relationship between the variables. This feature of the mutual information distinguishes it from all other measures thus far proposed in the relative importance literature.
4.1. Normal model

The entropy of a random variable $Y$ having normal distribution $F_Y = N(\mu, \sigma_Y^2)$ is $H_Y = .5 \log(2\pi e\sigma_Y^2)$. If the conditional distribution $F_{Y|X} = N(\beta^T x', \sigma^2)$, where $x' = (1, x')$ and $\beta$ are $p + 1$ dimensional vectors, then its entropy is $H(Y|X) = .5 \log [2\pi e\sigma^2(1 - \rho^2(Y, X))]$, where $\rho^2(Y, X) = 1 - \sigma^2/\sigma_Y^2$ is the squared multiple correlation between $Y$ and $X$.

The normal mutual information is given by

$$I_{\Phi}(Y, X) = -5 \log[1 - \rho^2(Y, X)] = .5 \log \phi^{-1},$$

where $\Phi$ is the correlation matrix of $(Y, X)$ and $\phi^{-1}$ denotes the first element of $\phi^{-1}$. The first equality in (21) is due to the fact that, for the normal model, $H(Y|X)$ does not vary with the outcomes $x$. The normal distribution is the ME model in the class of distributions subject to the mean and variance constraints, and the ME information importance (7) gives the same result as (21). The information index (13) gives $I_{\Phi}(Y; X) = \rho^2(Y, X)$.

Decomposition (2) for the normal regression is given by the partial mutual information

$$M_{\rho}(Y, X) = I_{\Phi}(Y|X) = -5 \log[1 - \rho^2(Y, X)],$$

where $\rho(Y, X)$ is the partial correlation between $Y$ and $X$, given $X_1, \ldots, X_{k-1}$. Successive application of (23) provides a chain rule for normal mutual information. Theil and Chung (1988) proposed measuring the relative importance of variables in univariate and multivariate regression models based on transforming the regression $R^2$ as in (21).

Formula (21) for normal mutual information is very simple, but normality of the distributions is crucial for its validity. For non-normal data, transformations to normality are therefore also crucial. Suppose that we have data on a set of variables $W, V = (V_1, \ldots, V_p)$ and we transform the variables as $Y = S(W)$ and $X_k = T_k(V_k)$ such that all transformations are one-to-one and $Y$ and $Y_1, \ldots, Y_p$ are normal. Then, by the invariance property of the mutual information, we can compute the importance of the original explanatory variables $V$ for the prediction of $W$ by

$$I_{\Phi}(W, V) = M_{\rho}(Y; X) = -5 \log[1 - \rho^2(Y, X)].$$

Numerous tests of normality are available; see, e.g., Coin (2008). When normality fails in a regression analysis, transformation to normality often can be achieved, for example, by Box–Cox transformations. Since Box–Cox transformations are non-linear, all regression quantities must be interpreted in terms of the transformed data. Yet Box–Cox transformations are one-to-one and the mutual information (24) retains its interpretation in terms of the original data. Thus invariance is a very useful property for an importance measure.

5. Bayesian inference

The information importance measures $I_{\Phi}(Y; X)$ and $M_{\rho}(Y, X)$ are functions of the model parameters $\Theta$. The likelihood function $L(\Theta|Y, x)$ is determined by the probability model and contains the sample information for $\Theta$. For the formulations discussed in the previous sections, the maximum likelihood estimates of the information quantities satisfy non-negativity condition (1). For Bayesian inference, (1) must hold stochastically. This allows inference about the relative information importance of predictors based on all orderings of the predictors using the chain rule (3). The posterior mean of each measure is its Bayes estimate under quadratic loss.

We provide Bayesian inference for the information measures of normal regression, contingency tables, and general logit analysis.

5.1. Normal regression

For normal regression, representation (22) allows computing the posterior distribution of $M_{\rho}(Y; X)$ from the posterior distribution of the simple correlation coefficients. Representation (22) ensures that (1) holds stochastically.

Inference about $M_{\rho}(Y; X)$ can be obtained in one of two ways. Under the stochastic regressors formulation, one can use the multivariate normal Bayesian inference for the correlation matrix and compute the posterior distribution of $M_{\rho}(Y; X)$ based on the posterior distribution of $\Phi$. Alternatively, one can use the nonstochastic regressors formulation where $\rho^2(Y, X) = R^2(Y, x)$, the usual $R^2$ of the regression, and

$$\Phi = \begin{pmatrix} 1 & r_{1,x} \\ r_{1,x} & R_k \end{pmatrix}$$

defined as follows: $R_k = [r_{k,1}, \ldots, r_{k,p}]$, $k \neq 1, \ldots, p$ is the given correlation matrix of the predictors; and $r_{j,x} = (r_1, \ldots, r_j)$, where

$$r_k^2 = r_{j,x}^2 = \frac{SS_j \rho_k^2}{SS_j \rho_j^2 + n\sigma_k^2}.$$
is the coefficient of determination for the simple regression $y = \alpha_k + \beta_k x_k + \epsilon_k$, $SS_k = \sum_{i=1}^{n}(x_{ki} - \bar{x}_k)^2$, and $n$ is the number of observations.

We use the nonstochastic regressors formulation and apply the inference method of Press and Zellner (1978) to (26). The algorithm for computing the posterior of $M(Y; X)$ is as follows (codes for implementation in R are available).

1. For the simple regression $y = \alpha_k + \beta_k x_k + \epsilon_k$, specify a prior for $\Theta_k = (\alpha_k, \beta_k, \sigma_k^2)$. We will use the non-informative prior $g(\alpha_k, \beta_k, \sigma_k^2) \propto \frac{1}{\sigma_k^2}$.

2. Update the prior to the joint posterior $g(\alpha_k, \beta_k, \sigma_k^2 | y)$ and compute the conditional posterior distribution of $\beta_k | \sigma_k^2$ and the marginal posterior distribution of $\sigma_k^2$. For the case of the above prior, the posterior distribution of $\beta_k | \sigma_k^2$ is normal with mean $b_k$, the least squares estimate, and variance $\frac{\sigma_k^2}{n}$. The posterior distribution of $\eta = \frac{(n-2)\sigma_k^2}{\sigma_k^2}$, where $s_k^2$ is the mean squares error of the least squares regression, is Chi-square with $n - 2$ degrees of freedom.

3. Simulate outcomes $(\beta_k^s, | \sigma_k^2 | f^s)$, $s = 1, \ldots, S$ from the posterior distributions and compute the correlation coefficient $r_k^s = \beta_k^s \sqrt{\frac{SS_k}{SS_k|f^s + n|\sigma_k^2 | f^s}}$, $k = 1, \ldots, p$.

4. Construct the $\Phi^s$ using $r_k^s$, $k = 1, \ldots, p$, compute the inverse matrix $[\Phi^s]^{-1}$, and the information function $M_{\phi^s}(Y; x) = .5 \log [\Phi^s]^{-1}$.

5. For a subset of the predictors, use the corresponding submatrix of $\Phi^s$, compute its inverse, and compute the information function using its first element. This implies (1) stochastically.

5.2. Contingency table

Bayesian inference for the information importance of predictors in a contingency table analysis is obtained by specifying a Dirichlet prior for the cell probabilities $\pi = (\pi_1, \ldots, \pi_J)$. The Maximum Entropy Dirichlet (MED) algorithm of Mazzuchi et al. (2000) is applicable. However, for the importance analysis, a simpler approach is to formulate the ME constraints in terms of the marginal fitting approach of Gokhale and Kullback (1978), Soofi and Retzer (2002) such that the ME model corresponds to an independence structure, and use the mutual information $M(Y; X)$.

The algorithm for computing the posterior of $M(Y; X)$ is as follows (codes for implementation in MINITAB and R are available).

1. Specify a Dirichlet prior $\pi \sim D(\beta, \pi_0)$, where $\pi_0 = E(\pi)$ is the prior expected distribution and $\beta$ is the strength of belief parameter.

2. Update the prior vector using the sample proportions $\hat{\pi}$ and obtain the posterior Dirichlet distribution for $\pi$.

$$\pi | \hat{\pi} \sim D \left( \beta + n, \frac{\beta \pi_0 + n \hat{\pi}}{\beta + n} \right).$$

(27)

3. Simulate $\pi^s$, $s = 1, \ldots, S$ from Dirichlet (27) and compute the marginal distribution $f^s$ and $H^s(Y)$.

4. Compute the entropy of the joint distribution $H^s(Y, X)$.

5. Compute the marginal distribution $f^s$ and marginal entropy $H^s(X)$.

6. Compute the mutual information $M^s(Y, X)$ using (19).

7. For a subset of the predictors collapse the corresponding dimension of the table in step 3 and compute steps 4–6. This implies (1) stochastically.

5.3. General logit

For the logit model (15), inference about the information quantities can be implemented by viewing them as functions of the logit parameters $\Theta = (\alpha_1, \ldots, \alpha_{J-1}, \beta')$. For any choice of prior distribution $g(\Theta)$, the posterior distribution can not be obtained in analytical form. However, Bayesian analysis for the logit model has been developed using MCMC techniques such as Gibbs sampling or the Metropolis–Hastings algorithm; see for example Chib and Greenberg (1995). Such analysis can be easily performed in an environment such as WinBUGS; see Spiegelhalter et al. (1996). Once the samples from the posterior distribution $g(\Theta | D)$ are generated via MCMC, posterior distributions of entropies and information indices can be easily computed. We note that in the logit model we have nonstochastic predictors and thus the mutual information is not meaningful.

The algorithm for computing the posterior distribution of $l(Y; x)$ is as follows.

1. Specify diffused but proper normal priors for components of $\Theta$.

2. Update the prior using MCMC and obtain the posterior samples $(\alpha_i^1, \ldots, \alpha_i^{J-1}, \beta')$, $s = 1, \ldots, S$.

3. Use each posterior sample $\alpha_i^1, \ldots, \alpha_i^{J-1}, \beta'$ and $\alpha_i^0$ (found by a side condition, e.g., $\alpha_i^0 = 0$) in (15) and compute the probability vectors $\pi_i^s = (\pi_1^s, \ldots, \pi_J^s)$ for all individuals $i = 1, \ldots, n$.

4. Compute the entropy $H^i(Y; x) = -\sum_{j=1}^{J} \pi_j^s \log \pi_j^s$ for all individuals $i = 1, \ldots, n$, and the joint (overall) entropy $H(Y; x) = \sum_{i=1}^{n} H^i(Y; x)$.

5. For all subsets of the predictors compute steps 1–4 and check (1) by the following inequalities:
Fig. 1 shows the results of the information analysis. Panel (a) shows the correlation matrices for the Retain posterior samples that satisfy (28) for all permutations of the subscripts of the predictors so that through the functional relationships through the specification holds et al. (12)

3.2 stochastically, is applicable to the general and the chain rule

3.2

stochastically. Then compute between parameters of a model and parameters of all of its submodels. The above algorithm for logit does not have such a built-in property. The MED algorithm of of

2000

between parameters of a model and parameters of all of its submodels. The above algorithm for logit does not have such a built-in property. The MED algorithm of Mazzuchi et al. (2000), which satisfies (1) stochastically, is applicable to the general logit. However, rejection sampling is a simpler approach. Thus, it is important to choose the posterior sample size \( S \) large enough to have sufficient number of realizations that satisfy (28).

6. Applications

6.1. Financial data

This example uses a subset of variables chosen from the Stock Liquidity data described in Frees (1996, p. 263). The variables chosen for the purpose of illustration are: the trading volume for a three month period in millions of shares (Volume \( V \)), total number of transactions for the three months (Transaction \( V_1 \)), number of shares outstanding at the end of the three month period in millions (Share \( V_2 \)), and market value in billion dollars (Value \( V_3 \)).

Table 1 shows the results of the information analysis. Panel (a) of Table 1 shows the correlation matrices for the original variables and their log-transformations. The normal probability plots of the residuals of the linear regressions for all seven subsets of these variables clearly showed violation of the normality assumption. Box–Cox transformation parameters for all variables are shown in Panel (a). The parameter values are near zero, suggesting a log-transformation. The normal probability plots of the residuals of the linear regressions for all seven subsets of the log-transformed variables, \( Y = \log W, X_k = \log V_k, k = 1, 2, 3 \), also confirmed the plausibility of conditional normality. By (24), the information importance analysis of the transformed variables is applicable to the original variables.

Panel (b) of Table 1 shows the \( R^2 \), mutual information and posterior results for the information importance of each subset of the predictors. Note that \( R^2 \) is interpretable in terms of the reduction of variance for the transformed variables, yet the results for information importance are interpretable in terms of the original as well as the transformed variables. The posterior quantities are computed by application of the nonstochastic regressors algorithm described in the Section 5.1 to the correlation matrix of the transformed data.

The posterior intervals of information importance are shown in Panel (b) of Table 1. Fig. 1(a) depicts these posterior intervals, graphically ordered from low to high importance. The middle dot on each interval depicts the posterior mean. We note that the posterior interval for the full model indicates skewness of the posterior distribution. We also note that the posterior interval for \( X_1 \) intersects with the interval for \( X_2 \), which in turn intersects with the interval for \( X_3 \). Based on these intervals we can infer that \( X_1 \) singly is more important than \( X_3 \). All intervals for models containing \( X_1 \) intersect. Thus, we can infer that the importance of the models containing \( X_1 \) does not differ. The interval for \( X_2X_3 \) does not intersect with the intervals for other models of size two and three. So we may infer that the model with \( X_2X_3 \) is different from the models of size...
Table 2
Information importance analysis of individual predictors of the financial data

<table>
<thead>
<tr>
<th>Ordering</th>
<th>$X_1$</th>
<th>$X_2$</th>
<th>$X_3$</th>
<th>$(X_1, X_2, X_3)$</th>
</tr>
</thead>
<tbody>
<tr>
<td>$X_1X_2X_3$</td>
<td>.802</td>
<td>.112</td>
<td>.028</td>
<td>.942</td>
</tr>
<tr>
<td>$X_1X_3X_2$</td>
<td>.802</td>
<td>.061</td>
<td>.079</td>
<td>.942</td>
</tr>
<tr>
<td>$X_2X_1X_3$</td>
<td>.348</td>
<td>.566</td>
<td>.028</td>
<td>.942</td>
</tr>
<tr>
<td>$X_2X_3X_1$</td>
<td>.362</td>
<td>.566</td>
<td>.014</td>
<td>.942</td>
</tr>
<tr>
<td>$X_3X_1X_2$</td>
<td>.461</td>
<td>.061</td>
<td>.420</td>
<td>.942</td>
</tr>
<tr>
<td>$X_3X_2X_1$</td>
<td>.362</td>
<td>.160</td>
<td>.420</td>
<td>.942</td>
</tr>
<tr>
<td><strong>Average</strong></td>
<td>.523</td>
<td>.254</td>
<td>.165</td>
<td>.942</td>
</tr>
<tr>
<td><strong>95% interval</strong></td>
<td>(.399, .643)</td>
<td>(.167, .388)</td>
<td>(.121, .248)</td>
<td></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Difference (Column $X_k$ - Row $X_\ell$)</th>
<th>$X_2$</th>
<th>$X_3$</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Average</strong></td>
<td>(.089, .436)</td>
<td>(.219, .487)</td>
</tr>
</tbody>
</table>

Fig. 1. Posterior 95% intervals for information importance of models and posterior distributions of relative importance of predictors for the financial data.

two and three. These inferences are based on the 95% probability intervals for each model. An adjustment (Bonferroni type) is needed for the probability of the inference about model comparison. The last column of Panel (b) shows the information index (13), which for the normal regression is $R^2$.

Table 2 gives the decompositions (3) of the joint information importance for all six orderings of the variables. The entries are computed using the posterior means of the mutual information $M$. The orderings are shown in the first column. Each of the middle three columns shows the relative importance for the position of the variable in the sequence shown in the first column. The last column gives the joint importance, which is the row sum. The relative information importance of each variable is strongly order dependent. The average information importance measures shown in the last row are computed using equal weights $w_k = 1/6$ in (4). These results indicate the overall average relative importance of the three variables. The average information importance of Transaction is more than twice that of Share and is more than three times that of Value.

Table 2 also shows the posterior intervals for the average importance of each variable and the pairwise differences between them. We can infer that, overall, Transaction ($X_1$) is more important than each of the other two variables, but the importance of Share ($X_2$) and Value ($X_3$) does not differ. Posterior distributions overall average importance of the three variables are shown in Fig. 1(b). The posterior distributions for the average relative importance of variables are close to normal, due to central tendency.

6.2. Long distance provider

This example uses a subset of data collected for Sprint by Maritz Research via non-sponsored telephone interviews. The respondents were asked to evaluate their current long distance provider and at least one alternative company based on past usage and/or current consideration. The questions were reflective of the respondents’ satisfaction with the company’s attributes. The response variable is long distance provider ($Y$) with three outcomes: Sprint, AT&T, and MCI. The explanatory variables are overall satisfaction with the company’s reputation as an industry leader, price, and a number of other attributes. Each explanatory variable has two categorical outcomes: low and high. Assessment of the relative importance of these variables was needed for inputs to a business decision. Soofi and Retzer (2002) reported derivations and assessments of some information theoretic models using three of the attributes. Mazzuchi et al. (2000) used this data to illustrate the Maximum Entropy Dirichlet (MED) inference procedure for the marginal fitting.
Table 3
Information importance analysis of subsets of variables for long distance providers

(a) Data

<table>
<thead>
<tr>
<th>Reputation X₁</th>
<th>Price X₂</th>
<th>Plans X₃</th>
<th>Service provider Y</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td></td>
<td>Sprint</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>AT&amp;T</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>MCI</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>Total</td>
</tr>
<tr>
<td>Low</td>
<td>Low</td>
<td></td>
<td></td>
</tr>
<tr>
<td>High</td>
<td>Low</td>
<td></td>
<td></td>
</tr>
<tr>
<td>High</td>
<td>Low</td>
<td></td>
<td></td>
</tr>
<tr>
<td>High</td>
<td>Low</td>
<td></td>
<td></td>
</tr>
<tr>
<td>High</td>
<td>Low</td>
<td></td>
<td></td>
</tr>
<tr>
<td>High</td>
<td>Low</td>
<td></td>
<td></td>
</tr>
<tr>
<td>High</td>
<td>Low</td>
<td></td>
<td></td>
</tr>
<tr>
<td>High</td>
<td>Low</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

(b) Information importance of three subsets of variables

<table>
<thead>
<tr>
<th>Subset</th>
<th>Data Information</th>
<th>Chi-sq.</th>
<th>d.f.</th>
<th>Mean</th>
<th>SD</th>
<th>95% interval</th>
<th>95% interval</th>
</tr>
</thead>
<tbody>
<tr>
<td>X₁</td>
<td>.040</td>
<td>72.92</td>
<td>2</td>
<td>.040</td>
<td>.009</td>
<td>(.024, .060)</td>
<td>(.023, .057)</td>
</tr>
<tr>
<td>X₂</td>
<td>.001</td>
<td>1.87</td>
<td>2</td>
<td>.002</td>
<td>.002</td>
<td>(.000, .007)</td>
<td>(.000, .007)</td>
</tr>
<tr>
<td>X₃</td>
<td>.002</td>
<td>4.21</td>
<td>2</td>
<td>.003</td>
<td>.002</td>
<td>(.000, .010)</td>
<td>(.000, .009)</td>
</tr>
<tr>
<td>X₁, X₂</td>
<td>.056</td>
<td>97.44</td>
<td>6</td>
<td>.055</td>
<td>.010</td>
<td>(.036, .077)</td>
<td>(.034, .074)</td>
</tr>
<tr>
<td>X₁, X₃</td>
<td>.048</td>
<td>83.31</td>
<td>6</td>
<td>.048</td>
<td>.010</td>
<td>(.030, .068)</td>
<td>(.029, .066)</td>
</tr>
<tr>
<td>X₂, X₃</td>
<td>.006</td>
<td>9.82</td>
<td>6</td>
<td>.009</td>
<td>.004</td>
<td>(.003, .017)</td>
<td>(.002, .017)</td>
</tr>
<tr>
<td>X₁, X₂, X₃</td>
<td>.060</td>
<td>104.91</td>
<td>14</td>
<td>.064</td>
<td>.011</td>
<td>(.044, .086)</td>
<td>(.042, .083)</td>
</tr>
</tbody>
</table>

Fig. 2. Posterior 95% intervals for information importance of models and posterior distributions of relative importance of predictors for long distance data.

Table 3 shows the data and importance analysis. Panel (a) of Table 3 shows the data in a $2 \times 2 \times 2 \times 3$ contingency table. Panel (b) of Table 3 shows the information importance, the information chi-square, their degrees of freedom, and posterior results for all subsets of the explanatory variables. The information importance is mutual information computed by (19). The information chi-square statistics are found by $\chi^2 = 2nM(Y, X)$. The information measure and chi-square can also be obtained using outputs of the exponential family regression by log-linear or logit models that include all the interactions between the variables.

The posterior intervals of information importance are shown in Panel (b) of Table 3. These results are obtained using a Dirichlet prior with $B = 24$ and a uniform distribution for $\pi_0$. Fig. 2(a) depicts the posterior intervals graphically for the information importance of the three models, ordered from low to high importance. Two clusters of models become apparent. The posterior intervals for all model containing $X_1$ intersect, but they do not intersect with intervals for models in which $X_1$ is not present. The posterior intervals for the models with $X_2$ and/or $X_3$ intersect. Based on these intervals we can infer that $X_1$ singly is more important than any combination of $X_2$ and $X_3$, whose importances do not differ. Neither do the importances of the models containing $X_1$. Again, these inferences are based on the 95% probability intervals for each model. An adjustment (Bonferroni type) is needed for the probability of the inference about model comparison. The last column of Panel (b) shows the posterior 95% intervals for the information index (12).

Table 4 gives the decompositions of the joint information in terms of the six orderings of the variables. We note that the information importance measures are highly order dependent. On average, the information importance of reputation ($X_1$) is about five times that of price ($X_2$) and about eight times that of the plan offering ($X_3$). Posterior intervals for the average
Table 4
Information importance analysis of variables for long distance providers

<table>
<thead>
<tr>
<th>Ordering</th>
<th>X₁</th>
<th>X₂</th>
<th>X₃</th>
<th>(X₁, X₂, X₃)</th>
</tr>
</thead>
<tbody>
<tr>
<td>X₁X₂X₃</td>
<td>.040</td>
<td>.015</td>
<td>.009</td>
<td>.064</td>
</tr>
<tr>
<td>X₁X₃X₂</td>
<td>.040</td>
<td>.016</td>
<td>.007</td>
<td>.064</td>
</tr>
<tr>
<td>X₂X₁X₃</td>
<td>.053</td>
<td>.002</td>
<td>.009</td>
<td>.064</td>
</tr>
<tr>
<td>X₃X₁X₂</td>
<td>.055</td>
<td>.002</td>
<td>.007</td>
<td>.064</td>
</tr>
<tr>
<td>X₃X₂X₁</td>
<td>.044</td>
<td>.016</td>
<td>.003</td>
<td>.064</td>
</tr>
<tr>
<td>X₁X₃X₂</td>
<td>.055</td>
<td>.005</td>
<td>.003</td>
<td>.064</td>
</tr>
<tr>
<td>Average</td>
<td>.048</td>
<td>.009</td>
<td>.006</td>
<td>.064</td>
</tr>
<tr>
<td>95% posterior (X₂ - X₁)</td>
<td>(.031, .067)</td>
<td>(.004, .017)</td>
<td>(.003, .011)</td>
<td></td>
</tr>
<tr>
<td>Difference (Column X₃ - Row X₂)</td>
<td>(.021, .058)</td>
<td>(.025, .066)</td>
<td>(-.003, .011)</td>
<td></td>
</tr>
</tbody>
</table>

Importance of each variable and the differences between them are shown in the last row of Table 4. Posterior distributions for the overall average importance of the three variables are shown in Fig. 2(b).

Finally, we note that the variables have higher predictive information when their positions in the sequence are second and third. Such variables in the linear regression context are referred to as “suppressors” in psychometric literature (see, e.g., Azen and Budescu (2003)). The more general probabilistic representation of “suppressor” variable is as follows. A variable $X_2$ is a “suppressor” if $(Y, X_2)$ are independent but not conditionally independent, given $X_1$; i.e., $f(y, x_2) = f(y|x_2)f(x_2|x_1)$ $\neq f(y|x_1)f(x_2|x_1)$. Noting that $E_x f(x, x_2|x_1) = f(y, x_2) = f(y|f(x_2))$, the independence of $(Y, X_2)$ is in fact due to averaging. Hence, the lack of predictive power of $X_2$, alone, is a loss of information due to aggregation (i.e., a priori averaging).

6.3. Adoption of new technology

This example uses data on revealed choices amongst three types of diagnostic equipment by 121 hospitals. Hospital diagnostic equipment purchasing agents evaluated each technology on the basis of various attributes. The variables selected for this example are hospital size and three technology attributes: price, efficiency, and quality of the equipment. Assessment of the importance of the hospital size and technology attributes singly and as a group was needed for the technology provider’s marketing strategy.

The information importance analysis is implemented using the ME logit (15). The hospital size categories are small, medium, and large. The size is represented by two indicator variables: $u_1$ for small and $u_2$ for medium; the large size is the base category ($u_1, u_2) = (0, 0)$. The technology attributes are scores price $(v_1 = P)$, efficiency $(v_2 = E)$, and quality $(v_3 = Q)$. The hospital size variables $u_{1i}, u_{2i}$ remain constant across the alternatives, so by (14), each variable requires two constraints, leading to two Lagrange multipliers $\alpha_j, j = 1, 2$ for two of the three alternatives; the coefficient for the third alternative is found by a side condition such as $\alpha_3 = 0$. Since the technology attributes vary across the alternatives $(v_{1i} = P_{1i}, v_{2i} = E_{2i}, v_{3i} = Q_{3i}, j = 1, 2, 3)$, each requires one constraint.

Table 5 shows the results. Panel (a) gives the MLE logit coefficients (Lagrange multipliers for the ME) obtained using SAS PROC PHREG. The log-likelihood chi-square statistics for the variables are related to information measures $\chi^2 = 2ln_p(Y; x_0) = 2[H^*_p(Y, x_{(1)}) - H^*_p(Y, x)]$, where $x_{(k)}$ is the vector excluding $x_k, k = 1, \ldots, 7$.

Panel (b) of Table 5 shows the information importance, the chi-square statistics, their degrees of freedom, and posterior results for hospital size $S = (u_1, u_2)$, technology attributes $T = (P, E, Q)$ and its subsets, and both groups of variables combined (full model). The information importance is the normalized index (12). Since the choice attributes are included, the global ME model (null model) is the uniform distribution over the three choices $\pi_i = (1/3, 1/3, 1/3), i = 1, \ldots, 121$ and $H^*(Y) = 121 \log 3 = 132.93$. The information chi-square statistics are given by (10). The log-likelihood function without variables (null) is $-2H^*(Y)$. The log-likelihood function with variables (model) is $-2H^*_p(Y, x)$, where $x = v$ for the hospital size, $x = u$ for technology, and $x = u, v$ for the two sets combined.

The Bayesian results are obtained using WinBUGS as described in Section 5.3. We used differenced normal priors with means zero and variances 100 for all seven parameters. Twenty-five thousand posterior samples were generated and 9406 samples satisfying (1) were obtained by rejection sampling. The posterior 95% intervals of information importance for the models are shown in Panel (b) of Table 5 and Fig. 3(a). The posterior intervals for the submodels containing technology variables intersect, so we cannot infer that the importance of one is higher than that of the other. The intervals for the model containing only the size variables and the full model do not intersect, leading to the inference that the importance of the full model is higher than that of the model containing only the size variables. But the intervals for the model containing the size and two technology attributes and the full model intersect, leading to the inference that the importances of these two models are about equal. These inferences are based on the 95% probability intervals for each model. An adjustment (Bonferroni type) is needed for the probability of the inference about model comparison.

Table 6 shows the importance analysis of the variables. Panel (a) of Table 6 gives the decompositions of the joint information in terms of two orderings of the size $S = (u_1, u_2)$ and technology attributes $T = (P, E, Q)$. We note that the
Table 5  
Information importance of hospital size and technology for choice of medical technology

(a) MLE Logit

<table>
<thead>
<tr>
<th>Organization Size (S)</th>
<th>Technology (T)</th>
<th>Price</th>
<th>Efficiency</th>
<th>Quality</th>
</tr>
</thead>
<tbody>
<tr>
<td>j = 1</td>
<td>j = 2</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Small</td>
<td>Medium</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>j = 1</td>
<td>j = 2</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Logit coefficient</td>
<td>2.24</td>
<td>3.29</td>
<td>1.71</td>
<td>2.23</td>
</tr>
<tr>
<td>Standard Error</td>
<td>1.21</td>
<td>1.16</td>
<td>.51</td>
<td>.52</td>
</tr>
<tr>
<td>Chi-square (df = 1)</td>
<td>3.46</td>
<td>7.98</td>
<td>11.32</td>
<td>18.68</td>
</tr>
</tbody>
</table>

(b) Subset of types of attributes

<table>
<thead>
<tr>
<th>Data (Likelihood)</th>
<th>Bayes (Posterior)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Subset</td>
<td>Information</td>
</tr>
<tr>
<td>Hospital size S</td>
<td>.148</td>
</tr>
<tr>
<td>Technology T = (P, E, Q)</td>
<td>.289</td>
</tr>
<tr>
<td>S, P</td>
<td>.271</td>
</tr>
<tr>
<td>S, E</td>
<td>.269</td>
</tr>
<tr>
<td>S, Q</td>
<td>.321</td>
</tr>
<tr>
<td>S, P, E</td>
<td>.359</td>
</tr>
<tr>
<td>S, P, Q</td>
<td>.411</td>
</tr>
<tr>
<td>S, E, Q</td>
<td>.378</td>
</tr>
<tr>
<td>Both types (S, T)</td>
<td>.447</td>
</tr>
</tbody>
</table>

Fig. 3. Posterior 95% intervals for information importance of models and posterior distributions of relative importance of predictors for technology data.

information importance measures are not strongly order dependent. The average relative information importance of the hospital size is about half that of the technology attributes. Posterior results for the average relative importance of each group of variables and the difference between the averages of the two groups are also shown in Panel (a). We can infer that the average importance of technology attributes is higher than that of hospital size.

Panel (b) of Table 6 shows the decomposition of the partial information of the technology variables $P$, $E$, and $Q$, in addition to the size for all six orderings of $P$, $E$, and $Q$. The results show rather strong order dependence of the information importance. The average importance over all orderings gives ratios of about 11:10:14 to price, efficiency, and quality, respectively. Posterior intervals for the average incremental importance of each variable to the size and the pairwise differences between them are also shown in Panel (b). The intervals for the averages intersect and the intervals for their differences include zero. We can infer that the importance of product attributes over and above the hospital size does not differ significantly. Fig. 3(b) shows the posterior distributions of the averages.

7. Conclusions

This paper has characterized the concept of importance of an explanatory variable as its contribution to the reduction of uncertainty about predicting outcomes of the response variable, namely, its information importance. The uncertainty is mapped by a concave function of probability density with a global maximum at the uniform distribution reflecting the most unpredictable situation. We conceptualized the information importance of predictors in terms of the difference between the uncertainty associated with the probability distributions of the response variable when specific predictors are absent and when they are present. We operationalized the uncertainty reduction in terms of Shannon entropy.
Information measures of importance are applicable to categorical as well as continuous random variables. Within the framework of information theory, importance measures for categorical, discrete, and continuous explanatory and response variables are provided in a unified manner. Such unification is attainable due to the fact that the probabilistic notion of information is general and axiomatic. However, the statistical measures of fit are usually problem specific, and do not necessarily admit a common interpretation, nor have an axiomatic basis. Some, but not all, of the statistical fit measures may be explicated in terms of information.

For nonstochastic predictors, the ME formulation provides importance measures. The ME procedure derives the model along with the importance measures. For the exponential family regression, the ME measures can be obtained using log-likelihood statistics. For stochastic predictors, the information importance is defined by the expected uncertainty reduction. The expected difference of Shannon entropies of the response variable’s distributions without and with use of predictors is the mutual information. We elaborated on conceptual and practical implications of the invariance property of the mutual information for measuring importance.

An additional contribution of our work is the development of Bayesian inference for the information importance measures and illustration of the additional insights that the Bayesian approach brings into the importance analysis. As shown in Section 3, in the exponential family regression, the information importance of predictors is given by the log-likelihood ratio or the deviance. Thus, the Bayesian estimation of information importance provides a Bayesian posterior analysis of the likelihood ratio, as suggested by Dempster (1997). The concept of Bayesian deviance is also considered in the deviance information criterion (DIC) proposed by Spiegelhalter et al. (2002). We are currently studying this connection and exploring information importance in terms of Bayes factors and Bayesian model averaging; see Kass and Raftery (1995).

The notion of information importance and the Bayesian inference methods presented here have potential applications in Bayesian networks that deal with the assessment of conditional independence for applications problems such as the one considered by Quali et al. (2006).

Three examples illustrated the implementation and applications of the information importance concept and measures. The first example, serving purely an illustrative purpose, showed the versatility of the invariance property of mutual information in linear regression. Two other examples illustrated real-world applications. In the choice of long distance provider example, we assessed the relative importance of the long distance company’s reputation, price, and plan offering for customer choice among three providers. In this example, all variables are categorical. In the technology adoption example, we applied the ME procedure to assess the importance of hospital size and three technology attributes for the prediction of choice of medical diagnostic equipment. This example demonstrated a comparison of information importance of choice and decision-maker attributes in the logit analysis. The Bayesian approach provided additional insights about differences between the information importance of models and relative importance of predictors in these examples.

Acknowledgements

We thank the Co-editor Professor Belsley and two referees for their comments and suggestions leading to substantial improvement of the presentation of this paper.
References


Joseph J. Retzer is the Director of Marketing Sciences for Maritz Research Chicago. Prior to joining Maritz, he was the Director of Product Development at MarketProbe, Market Research Inc., Milwaukee. He has taught statistics, economics, and management science courses at the University of Wisconsin-Milwaukee. His research interests are in information-theoretic and Bayesian statistics and their applications in economics, management science, and decision problems. He has published numerous articles in statistics, econometrics, and management science fields. His research has appeared in Journal of the American Statistical Association, Journal of Royal Statistical Society, Biometrika, Journal of Multivariate Analysis, Computational Statistics & Data Analysis, Journal of Econometrics, Journal of Applied Probability, Operations Research, Marketing Science, and IEEE transactions on Information Theory. He served as an associate editor of Journal of the American Statistical Association (1990–2005) and as an Associate Editor of Entropy, An International and Interdisciplinary Journal of Entropy and Information Studies (1991–2006). He also served as the Chair of the Evaluation Committee of the Savage Thesis Award, sponsored by the International Society for Bayesian Analysis (ISBA) and the American Statistical Association, as a Vice President of the International Association for Statistical Computing (IASC), and as the chair of the IASC Publication Committee. He is an elected member of International Statistical Institute and a Fellow of the American Statistical Association. He holds a Bachelors degree in Mathematics from U.C.L.A., a Masters degree in Statistics from the University of California, Berkeley, and a Ph.D. in Applied Statistics from the University of California, Riverside.

Ehsan S. Soofi is a professor of management science and statistics and a Roger L. Fitzsimonds Distinguished Scholar at Sheldon B. Lubar School of Business and a Research Associate of the Center for Research on International Economics at the University of Wisconsin-Milwaukee. His research interest is in information-theoretic and Bayesian statistics and their applications in economics, management science, and decision problems. He has published numerous articles in statistics, econometrics, and management science fields. His research has appeared in Journal of the American Statistical Association, Journal of Royal Statistical Society, Biometrika, Journal of Multivariate Analysis, Computational Statistics & Data Analysis, Journal of Econometrics, Journal of Applied Probability, Operations Research, Marketing Science, and IEEE transactions on Information Theory. He served as an associate editor of Journal of the American Statistical Association (1990–2005) and as an Associate Editor of Entropy, An International and Interdisciplinary Journal of Entropy and Information Studies (1991–2006). He also served as the Chair of the Evaluation Committee of the Savage Thesis Award, sponsored by the International Society for Bayesian Analysis (ISBA) and the American Statistical Association, as a Vice President of the International Association for Statistical Computing (IASC), and as the chair of the IASC Publication Committee. He is an elected member of International Statistical Institute and a Fellow of the American Statistical Association. He holds a Bachelors degree in Mathematics from U.C.L.A., a Masters degree in Statistics from the University of California, Berkeley, and a Ph.D. in Applied Statistics from the University of California, Riverside.
Refik Soyer is a professor of Decision Sciences and Statistics and Director of Institute for Integrating Statistics in Decision Sciences at the George Washington University. His areas of interests are Bayesian statistics and decision analysis, stochastic modeling, statistical aspects of reliability analysis, and time series analysis. His research focuses on modeling and methodology development and applications to problems such as portfolio selection models, maintenance practices for railroad tracks, modeling call center arrivals, analyzing pulse trains, and modeling mental health data. He has published in the leading journals in statistics, econometrics, and management sciences such as Journal of the American Statistical Association, Journal of Royal Statistical Society, Technometrics, Biometrics, Computational Statistics & Data Analysis, Journal of Econometrics, Management Science, Naval Research Logistics, IEEE transactions on Reliability, and IEEE Transactions on Software Engineering. He is the lead co-editor of a volume titled Mathematical Reliability: An Expository Perspective. He is an elected member of International Statistical Institute and a Fellow of the American Statistical Association, and has served as an Associate Editor of Journal of the American Statistical Association. He holds a B.A. in Economics from Bogazici University, Turkey, an M.Sc. in Operations Research from Sussex University, England, and a D.Sc. degree in Operational Research from the George Washington University.