I²SDS
The Institute for Integrating Statistics in Decision Sciences

Technical Report TR-2007-13
September 24, 2007

Information Importance of Models and Relative Importance of Predictors: Concept, Measures, Bayes Inference and Applications

Joseph J. Retzer
Maritz Research

Ehsan S. Soofi
Sheldon B. Lubar School of Business
University of Wisconsin-Milwaukee

Refik Soyer
Department of Decision Sciences
The George Washington University

Information Importance of Models and Relative Importance of Predictors:

Concept, Measures, Bayes Inference, and Applications

J. J. Retzer[a], E. S. Soofi[b*], R. Soyer[c,]

[a]*Maritz Research, 1815 S. Meyers Rd, Suite 600, Oak brook Terrace, IL 60181, USA*
[b]*Sheldon B. Lubar School of Business, University of Wisconsin-Milwaukee, P.O. Box 742, Milwaukee, WI 53201, USA*
[c]*Department of Decision Sciences, George Washington University, Washington D.C. 20052, USA*

September 24, 2007

Abstract
Comparison of relative importance of predictors is a subject of discussion of research findings in many disciplines, as well as being input for decision-making in business practice. Relative importance methodologists have proposed measures for specific problems such as normal linear regression and logit. Some attempts have been made to set requirements for relative importance of predictors, given a measure of "importance", without characterizing the notion of "importance" itself. The main objective of this paper is to fill this gap by providing a notion of importance of predictors sufficiently general so as to be applicable to various models and data types, yet to admit a unique interpretation. The importance of predictors is characterized by the extent to which their use reduces uncertainty about predicting the response variable, namely their information importance. 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 is invariant under one-to-one transformations of the variables. Applications to various data types leads to familiar statistical quantities for various models, yet with the unified interpretation of uncertainty reduction. Bayes inference for importance and relative importance of predictors is presented. An illustrative example shows the information importance analysis and versatility of the invariance for normal regression. Two other examples show applications providing assessments of relative importance of some attributes needed for business decision-making.

Keywords: Contingency table; Entropy; Exponential family regression; Invariance; Mutual Information; Linear regression; Logit analysis.

1 Introduction

Assessment of importance and relative importance of explanatory variables is very common in reports of research studies in numerous fields, as well as in decision-making practice. In an effort to gain insight into approaches used by researchers in various fields, Kruskal and Majors (1989) conducted an extensive literature survey across diverse disciplines: economics, political science,

*Corresponding author.
Email addresses: retzerjj@maritz.com, esoofi@uwm.edu, soyer@gwu.edu

history, psychology, sociology, education, physics, chemical engineering, biology, and medicine. While they found that assessing importance was clearly of great interest, the authors expressed a disappointment:

> "We were depressed by the frequency of use of statistical significance as a measure of relative importance. Even though we had half expected that misuse, it was sad to see significance testing so often and inappropriately employed." (Kruskal and Majors 1989, p. 3)

There is a general agreement among relative importance methodologists that quantities such as significance levels, correlations and variable coefficients (standardized or not) are not appropriate for assessing relative importance of predictors. A measure of statistical significance maps the analyst's strength of confidence in making inference about an unknown parameter based on a statistic. 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 prediction of a response variable (Azen and Budescu 2003, Budescu 1993, Cox 1985, Genizi 1993, Johnson 2000, Kruskal 1984, Kruskal 1987, Kruskal and Majors 1989, Lindeman et al 1980, Pourahmadi and Soofi 2000, Pratt 1990, Schemper 1993, Soofi 1992, Soofi et al 2000, Theil and Chung 1988). Relative importance measures are defined as functions of distributional parameters such as correlation coefficients. When the parameters are known (e.g., population data, simulation study), there is no uncertainty involved and inference is irrelevant, yet assessing the relative importance of a set of variables might be of interest. When the parameters are unknown, the measures of relative importance like other parameters of interest are also subject to statistical inference.

In real-world practice, attribute relative importance assessment is a mainstay in many decision making situations. Decision makers routinely demand reporting relative importance weights of decision variables in data analysis reports, in the same vein as multiattribute decision making problems. For example, practitioners in market research rely heavily on importance assignment to drivers of customer behaviors, e.g. loyalty/re-purchase as well as customer attitudes toward the product and the company, e.g. satisfaction. Market researchers, along with practitioners in other areas, have increasingly become aware of the inappropriateness of measures such as significance levels, correlation and standardized regression coefficients for measuring importance. Many have therefore begun adopting new methodologies for importance assignment and, in doing so, have realized a competitive advantage through a greater understanding of their customers.

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, Kruskal 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), 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 importance is consequential for practice. Presently, "importance" is interpreted differently in different problems (e.g., linear regression, ANOVA, logit) and when a new measure is developed for a new problem encountered (e.g., contingency table), it may or may not be based on the criteria for the existing measures, or share a common interpretation. 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.

The relative importance literature mainly has dealt with linear regression where the importance of an explanatory variable is defined in terms of reduction of the predictive error variance (see Grömping (2007) and references therein). Conceptualizing predictive ability in terms of variance reduction leads to squared correlation measures, which do not apply much beyond normal linear regression. Variance reduction is not sufficiently general to provide appropriate measures for varieties of problems encountered in practice such as, qualitative response and/or explanatory variables, or when the random variables are not normally distributed. For a qualitative response variable the variance is meaningless. For some well known continuous distributions (e.g., Cauchy) the variance does not exist.

This paper has two objectives: (a) proposing a notion of *importance* that is sufficiently general to be applicable for various data structures and models with a common interpretation, and (b) developing Bayesian inference for relative importance of explanatory variables for commonly used models. We conceptualize importance in terms of the information provided by a predictor for reducing uncertainty about predicting the outcomes of the response variable. We will show that the concept of importance drawn from information theory is in accord with our intuitive notion of information and offers appropriate measures for diverse problems in a unified manner. We measure prediction error in terms of entropy loss instead of the squared error loss.

The information is a general probabilistic concept that provides measures of importance for categorical and discrete variables, as well as continuous variables regardless 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 is invariant under one-to-one transformations of the variable. 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, 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.

Three examples illustrate implementation of the information importance measures and the Bayesian inference for three types of data structures. The first example is purely illustrative showing versatility of the information importance and Bayesian inference about relative importance of predictors in normal linear regression. The other two examples are drawn from actual business practice. An example on the *choice of long distance provider* illustrates the relative importance of satisfaction with firm's reputation as an industry leader, the price, and long distance plan offering, for the choice of long distance phone service provider. In this case all explanatory variables are qualitative, so the data can be summarized in a contingency table. A *technology adoption* example compares the relative importance of hospital size and three product attributes (price, efficiency and quality) for the adoption of new technology in the medical industry. In this case the response is qualitative and explanatory variables are both categorical and continuous.

Section 2 presents the notion of information importance. Section 3 presents maximum entropy information importance measure for nonstochastic predictors. A subsection presents exponential regression and another subsection presents logit. Section 4 presents the expected information measure for stochastic predictors with a subsection on the normal regression model. Section 5 describes Bayesian inference about information importance and relative importance of predictors. Section 6 presents three examples. Section 7 gives the concluding remarks.

2 Notion of Information Importance

Let $\boldsymbol{x} = (x_1, \cdots, x_p)$ be a vector of predictors of a variable Y, where the prediction is probabilistic. The importance of a predictor $\boldsymbol{x}$ for Y is the extent to which the use of $\boldsymbol{x}$ reduces *uncertainty* in predicting outcomes of Y. We conceptualize uncertainty in terms of unpredictability of outcomes of Y. The most unpredictable situation is when we are unable to forecast in favor or against any values for the response variable. In most unpredictable situation we invoke Laplace's "Principle of Insufficient Reason" and assign equal probabilities to all possible values (intervals of equal width in the continuous case) of Y. This establishes uniformity of the probability distribution as the

reference point for quantifying uncertainty in terms of predictability.

The uncertainty associated with a probability distribution F having a density (mass) function f is defined by

$$\mathcal{U}(f) \leq \mathcal{U}(f^*), \tag{1}$$

such that $\mathcal{U}(f)$ is concave $\mathcal{U}[\alpha f_1 + (1-\alpha)f_2] \geq \alpha\mathcal{U}(f_1) + (1-\alpha)\mathcal{U}(f)$, $0 \leq \alpha \leq 1$ and f^* is the uniform density (improper when the support of F is unbounded). That is, $\mathcal{U}(f)$ is a measure of uniformity (lack of concentration) of probabilities under the distribution. The requirement of mapping uniformity in the definition of $\mathcal{U}(f)$ is a modification proposed by Ebrahimi et al (2007) for the uncertainty function defined by Goel and DeGroot (1981) where the only requirement for $\mathcal{U}(f)$ was to be a concave function F. Their examples include $\mathcal{U}(f) = Var_f(Y)$. However, variance is not a general measure of uniformity of the probability distribution. The variance is meaningless when outcomes of Y are categorical. For quantitative variables, variance may not be defined, and when defined, it does not necessarily map uncertainty in the sense of difficulty of predicting outcomes of the variable; e.g., under a beta distribution $Be(\alpha, \beta)$ with $\alpha, \beta < 1$, intervals of equal width at the tail are more likely than at the center, but variance is higher than the uniform distribution where all intervals of equal width are equally likely and more difficult to predict the outcomes; see Ebrahimi et al (1999).

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 $F_{Y;\boldsymbol{x}}$ which depend on $\boldsymbol{x}$ but not on the position of x_k in the vector, and has a density (mass) function $f_{Y;\boldsymbol{x}}$. For a stochastic predictor, $F_{Y;\boldsymbol{x}}$ is the conditional distribution and $f_Y = E_{\boldsymbol{x}}[f_{Y|\boldsymbol{X}}]$. For a nonstochastic predictor such a relationship is absurd.

The worth of $\boldsymbol{x}$ for the prediction of Y is mapped by the uncertainty difference

$$\Delta_{\mathcal{U}}(Y;\boldsymbol{x}) = \mathcal{U}(f_Y) - \mathcal{U}(f_{Y;\boldsymbol{x}}). \tag{2}$$

In general, $\Delta_{\mathcal{U}}(Y;\boldsymbol{x})$ can be positive, negative, or zero, mapping in turn the gain, loss, or no change of information due to the use of $\boldsymbol{x}$ for predicting the outcomes of Y. A loss of information occurs when $F_{Y;\boldsymbol{x}}$ is less concentrated and hence it makes more difficult to predict Y than using F_Y. In this case, $\boldsymbol{x}$ is useless for predicting Y.

When a predictor makes prediction more difficult, the verdict on its information importance worth is clear, hence $\Delta_{\mathcal{U}}(Y;\boldsymbol{x}) < 0$ is of no particular interest in the present context. We provide formulations that are sufficiently general satisfying (1) and give nonnegative information importance functions. We therefore proceed with the case when $\Delta_{\mathcal{U}}(Y;\boldsymbol{x}) \geq 0$.

The proper information importance of a predictor vector $\boldsymbol{x}$ is defined by the following property

$$\mathcal{I}_{\mathcal{U}}(Y;\boldsymbol{x}) = \mathcal{U}(f_Y) - \mathcal{U}(f_{Y;\boldsymbol{x}}) \geq 0. \tag{3}$$

We should note that $\mathcal{I}_{\mathcal{U}}(Y;\boldsymbol{x})$ does not depend on the position of x_k in the vector.

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

$$\mathcal{I}_{\mathcal{U}}(Y|\boldsymbol{X}) = E_{\boldsymbol{x}}[\Delta_{\mathcal{U}}(Y|\boldsymbol{X})] = \mathcal{U}(f_Y) - E_{\boldsymbol{x}}[\mathcal{U}(f_{Y|\boldsymbol{x}})] \geq 0, \tag{4}$$

where the inequality changes to equality if and only if $\boldsymbol{X}$ and Y are independent. The non-negativeness is implied by concavity of $\mathcal{U}$ and characterizes the expected gain of using the outcomes of $\boldsymbol{X}$ for the prediction. It is reasonable to require that using outcomes of $\boldsymbol{X}$, 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 outcomes of another variable (DeGroot 1961).

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

$$\mathcal{I}_{\mathcal{U}}(Y; x_{r+1}, \cdots, x_p | x_1, \cdots, x_r) = \mathcal{U}(Y; x_1, \cdots, x_r) - \mathcal{U}(Y; x_1, \cdots, x_p) \geq 0. \tag{5}$$

The first inequality is apparent (add and subtract $\mathcal{U}(F_Y)$) and inequality is implied by properness (3). We therefore have the decomposition property,

$$\mathcal{I}_{\mathcal{U}}(Y; x_1, \cdots, x_p) = \mathcal{I}_{\mathcal{U}}(Y; x_1, \cdots, x_r) + \mathcal{I}_{\mathcal{U}}(Y; x_{r+1}, \cdots, x_p | x_1, \cdots, x_r), \tag{6}$$

Successive application of (6) gives the following chain rule:

$$\mathcal{I}_{\mathcal{U}}(Y; x_1, \cdots, x_p) = \sum_{k=1}^{p} \mathcal{I}_{\mathcal{U}}(Y; x_k | x_1, \cdots, x_{k-1}), \tag{7}$$

where $\mathcal{I}_{\mathcal{U}}(Y; x_1 | x_0) \equiv \mathcal{I}_{\mathcal{U}}(Y; x_1)$, and $\mathcal{I}_{\mathcal{U}}(Y; x_k | x_1, \cdots, x_{k-1})$ is the incremental contribution of x_k to the information importance of $(x_1, \cdots, x_k)$.

2.1 Relative Importance

The incremental information function $\mathcal{I}_{\mathcal{U}}(Y; x_k | x_1, \cdots, x_{k-1})$ provides measures of relative importance of predictor x_k in the sequence $x_1, \cdots, x_p$. 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 (7) is a general representation satisfying the first property. However, in general, decomposition (7) depends on the position of x_k in $(x_1, \cdots, x_p)$, 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:

$$\overline{\mathcal{I}}_{\mathcal{U}}(Y; x_k) = \sum_{q=1}^{p!} w_q \mathcal{I}_{\mathcal{U}}(Y; x_k | x_1, \cdots, x_{k-1}; O_q), \tag{8}$$

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

2.2 Shannon Entropy

The most well-known example of an uncertainty function is Shannon entropy $\mathcal{U}(F) = H(F)$, defined by

$$\begin{aligned} H(Y) \equiv H(F) &= -\int f(y) \log f(y) dy \quad \text{(continuous case)} \\ &= -\sum f(y) \log f(y) \quad \text{(discrete \& categorical cases).} \end{aligned} \tag{9}$$

The entropy maps the concentration of probabilities under F and decreases as concentration increases. For the discrete distribution over n outcomes, $0 \leq H(F) \leq \log n$, where $H(F) = 0$ holds if and only if $f(y) = 1$ for a single point and $f(y) = 0$ for all other points; hence perfect information about Y and absence of uncertainty. The equality $H(F) = \log n$ holds if and only if the probability is uniformly distributed; hence complete absence of information about favoring an outcome of Y, and maximum uncertainty. When F is continuous, then $-\infty < H(F) < \infty$ and $H(F)$ is not invariant under one-to-one transformations of the variable. However, interpretation of continuous entropy essentially remains the same as the discrete case, i.e., $H(F)$ orders distributions according to the lack of concentration, and hence lack of information for prediction.

The entropy difference

$$\Delta_H(Y; \boldsymbol{x}) = H(Y) - H(Y; \boldsymbol{x}) \tag{10}$$

gives a measure of change in uncertainty for prediction of Y due to the use of $\boldsymbol{x}$ for predicting Y. The entropy difference (10) may be positive or negative. The outcome $\boldsymbol{x}$ is informative about Y if $F_{Y;\boldsymbol{x}}$ is more concentrated than F_Y. (In a Bayesian context where y is a parameter, $\boldsymbol{x}$ is the data, and $\Delta_{\mathcal{U}}(Y; x)$ is the difference between the prior and posterior entropies, the case of $\Delta_H(Y; \boldsymbol{x}) < 0$ is referred to as a "surprise"; Lindley 1956). In our formulation of information importance (Section 3), the entropy difference is always nonnegative, $\mathcal{I}_H(Y; \boldsymbol{x}) = \Delta_H(Y; \boldsymbol{x}) \geq 0$.

The information content of $F_{Y;\boldsymbol{x}}$ and F_Y can also be compared using an information divergence function such as the Kullback-Leibler information (relative entropy),

$$\begin{aligned} K(Y; \boldsymbol{x}) \equiv K(F_{Y;\boldsymbol{x}} : F_Y) &= \int f_{Y;\boldsymbol{x}}(y) \log \frac{f_{Y;\boldsymbol{x}}(y)}{f_Y(y)} dy \quad \text{(continuous case)} \\ &= \sum f_{Y;\boldsymbol{x}}(y) \log \frac{f_{Y;\boldsymbol{x}}(y)}{f_Y(y)} \quad \text{(discrete \& categorical cases).} \end{aligned} \tag{11}$$

This information function is always non-negative, but in general it only quantifies the change in the concentration of the distribution of Y due to $\boldsymbol{x}$ and does not indicate which of the two distributions is more concentrated. In some formulations of information importance, $K(Y;\boldsymbol{x}) = \mathcal{I}_H(Y;\boldsymbol{x})$, hence signifying more concentration of $F_{Y;\boldsymbol{x}}$.

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

$$I(Y;\boldsymbol{x}) = 1 - \frac{H(Y;\boldsymbol{x})}{H(Y)} = \frac{\mathcal{I}_H(Y;\boldsymbol{x})}{H(Y)}. \tag{12}$$

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

$$I(Y;\boldsymbol{x}) = 1 - e^{-2\mathcal{I}_H(Y;\boldsymbol{x})}. \tag{13}$$

In both cases the indices range from zero to one: $I(Y;\boldsymbol{x}) = 0$ mapping the case when the predictor does not reduce the uncertainty at all, and $I(Y;\boldsymbol{x}) = 1$ mapping the case when the predictor reduces the uncertainty completely.

3 Maximum Entropy Information

An approach that always provides nonnegative uncertainty difference (2), as well as providing a unified interpretation of predictor importance for wide varieties of applications is the Maximum Entropy (ME) information formulation. In this approach, $F_Y = F_Y^*$ is the ME model in a class of distributions subject to some constraints free from $\boldsymbol{x}$ and $F_{Y;\boldsymbol{x}} = F_{Y;\boldsymbol{x}}^*$ is the ME model in a class of distributions subject to some additional constraints involving $\boldsymbol{x}$.

In a very general set up, the ME approach begins with a class of distributions

$$\Omega_{F_Y} = \{F : E_F[T_a(Y)] = \theta_a,\ a = 1, \cdots, A\}, \tag{14}$$

where $T_a(Y)$ are real-valued integrable functions with respect to $dF(y)$ and $\theta_a,\ a_1 = 1, \cdots, A$ are specified moments. The ME model in Ω_{F_Y} is the distribution whose density maximizes (9).

The set of linearly independent moments defining Ω_{F_Y} is denoted by

$$\mathcal{T}_Y = \mathcal{T}_Y = \{T_a(Y),\ a = 1, \cdots, A\}. \tag{15}$$

For example, $T(Y) = Y,\ T(Y) = Y^2,\ T(Y) = \log Y$, and $T(Y) = \delta(\mathcal{S}_\ell)$ where $\delta(\mathcal{S}_\ell)$ is an indicator function of a subset of support of F are all legitimate, provided that they are integrable. If the expected value of elements of a moment set $\mathcal{T}_Y$ can be obtained from the expected value of elements of another moment set $\mathcal{T}_Y^*$, the two sets yield the same ME distributions.

The ME model in (14), if exists, is unique and has density in the following form:

$$f_Y^*(y) = C(\lambda)e^{\lambda_1 T_a(y) + \cdots + \lambda_A T_A(y)}, \tag{16}$$

where $\lambda = (\lambda_1, \cdots, \lambda_A)$ is the vector of Lagrange multipliers given by $\theta_a = -\dfrac{\partial}{\partial \lambda_a} \log C(\lambda)$. For all $F \in \Omega_{F_Y}$,

$$H^*(Y) \equiv H(F_Y^*) = -\log C(\lambda) - \lambda_1 \theta_1 - \cdots - \lambda_A \theta_A \geq H(F_Y); \tag{17}$$

see Ebrahimi, Soofi, and Soyer (2007).

In order to assess the information importance of a predictor $\boldsymbol{x}$ we expand the moment set $\mathcal{T}_Y$ as

$$\mathcal{T} = \mathcal{T}_Y \cup \mathcal{T}_{Y;\boldsymbol{x}} = \{T_a(Y),\ a = 1, \cdots, A\} \cup \{T_b(Y; \boldsymbol{x}),\ b = 1, \cdots, B\}, \tag{18}$$

where $\mathcal{T}_{Y;\boldsymbol{x}}$ is a set of moments of Y in terms of $\boldsymbol{x}$. For example, for single variable x, we may expand $\mathcal{T}_Y$ by $T(Y; x) = xY$ or $T(Y; x) = \log(1+x)Y$. The information moment set (18) generates a class of distributions $\Omega_{F_{Y;\boldsymbol{x}}} \subseteq \Omega_{F_Y}$. The density of ME distribution $F^*_{Y;\boldsymbol{x}} \in \Omega_{F_{Y;\boldsymbol{x}}}$ is in the form of (16) with additional parameters,

$$f^*_{Y;\boldsymbol{x}}(y) = C(\lambda(\boldsymbol{x}))e^{\lambda_1(\boldsymbol{x})T_a(y) + \cdots + \lambda_A(\boldsymbol{x})T_A(y) + \beta_1(\boldsymbol{x})T_{A+1}(y;\boldsymbol{x}) + \cdots + \beta_B T_{A+B}(\boldsymbol{x})(y;\boldsymbol{x})}, \tag{19}$$

where $\lambda(\boldsymbol{x}) = (\lambda_1(\boldsymbol{x}), \cdots, \lambda_A(\boldsymbol{x}), \beta_1(\boldsymbol{x}) \cdots, \beta_B T_{A+B}(\boldsymbol{x}))$.

When F_Y^* and $F^*_{Y;\boldsymbol{x}}$ are ME distributions in Ω_{F_Y} and $\Omega_{F_{Y;\boldsymbol{x}}}$ generated by (15) and (18), the entropy difference (10) provides measures of information importance of $\boldsymbol{x}$ for predicting Y:

$$\mathcal{I}_{\boldsymbol{\Theta}}(Y; \boldsymbol{x}) = H^*(Y) - H^*(Y; \boldsymbol{x}) \geq 0, \tag{20}$$

where $\boldsymbol{\Theta}$ denotes the vector of all parameters involved. The equality in (20) is due the additional constraints reducing the maximum entropy (Jaynes 1957, Jaynes 1968, Soofi 1992, Soofi 1994). Clearly, $\mathcal{I}_{\boldsymbol{\Theta}}(Y, \boldsymbol{x})$ admits the chain rule decomposition (6).

The quantity $\mathcal{I}_{\boldsymbol{\Theta}}(Y, \boldsymbol{x})$ provides measures of information importance for various types of data and models, all with the same interpretation. For example, distributions with densities in exponential family having finite entropy are ME in appropriately defined Ω_F (Ebrahimi et al 2007). The traditional exponential family regression assumes that the underlying data distribution is known to be a particular parametric family, but its parameters are unknown. In the ME approach, no such strong distributional assumption is needed. Instead, the weaker moment assumptions (15) and (18) are formulated such that: (a) the ME models F_Y^* and $F^*_{Y;\boldsymbol{x}}$ are in the same exponential family; and (b) the constraints are formulated such that moment values are statistics $\theta_k = \hat{\theta_k}$ (see e.g., Soofi 1992). Such ME formulations gives:

$$\begin{aligned}
\mathcal{I}_{\widehat{\boldsymbol{\Theta}}}(Y, \boldsymbol{x}) &= 2n\left[H_{\widehat{\boldsymbol{\Theta}}}(F_y^*) - H_{\widehat{\boldsymbol{\Theta}}}F^*_{y|x})\right] \\
&= -2\log\left[\frac{f_{Y;\boldsymbol{x}}(y)|_{\boldsymbol{\Theta}(x)=\widehat{\boldsymbol{\Theta}}(x)}}{f_Y(y)|_{\boldsymbol{\Theta}=\widehat{\boldsymbol{\Theta}}}}\right] \\
&= 2\widehat{K}(F^*_{Y;\boldsymbol{x}} : F_Y^*).
\end{aligned}$$

The middle quantity is the likelihood ratio statistic and $\hat{K}(F^*_{Y;\boldsymbol{x}} : F^*_Y)$ is known as the deviance in the exponential family regression literature. Thus, by (20), 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.

The remainder of this section presents two examples where the ME approach lead to exponential regression and logit models. The most celebrated member of the exponential family, the normal distribution, will be presented in Section 4 to illustrate application of expected information. We should note that (20) is a general information importance measure applicable to any ME distributions, beyond the exponential family regression. Any distribution with a density in the form of (16) having finite entropy is an ME model (Ebrahimi et al 2007). Thus, (20) provides information measures for evaluating importance of types of additional moments to be included in (15) in terms of their incremental contributions to uncertainty reduction.

3.1 Exponential Regression

The ME model subject to constraint $E(Y) = \theta_1$ is the exponential distribution with density $f^*_Y(y) = \lambda e^{-\lambda y}$, where the Lagrange multiplier is given by $\lambda = \theta_1^{-1}$. The maximum entropy is $H^*_Y = 1 - \log \lambda$. The ME model subject to the additional constraint $E(xY) = \theta_2$ is the exponential distribution with density $f^*_{Y;\boldsymbol{x}} = \lambda(\boldsymbol{x}) e^{-\lambda(\boldsymbol{x})y}$, where $\lambda(\boldsymbol{x}) = \beta_0 + \beta'\boldsymbol{x}$. The maximum entropy is $H^*_{Y;\boldsymbol{x}} = 1 - \log \lambda(\boldsymbol{x})$ and $\theta(\boldsymbol{x}) = (\beta_0 + \beta'\boldsymbol{x})^{-1}$.

The information importance of predictor $\boldsymbol{x}$ is

$$\mathcal{I}_\theta(Y; \boldsymbol{x}) = H^*_Y - H^*_{Y;\boldsymbol{x}} = \log \frac{\lambda}{\lambda(\boldsymbol{x})} = \log \frac{\theta(\boldsymbol{x})}{\theta_1} \geq 0.$$

For a sample of n observations, using MLE $\hat{\lambda}$ and $\hat{\lambda}(\boldsymbol{x}) = \hat{\beta}_0 + \hat{\beta}_1 x$ we have the ME information importance in terms of the log-likelihood ratio statistic and deviance:

$$\begin{aligned}
\mathcal{I}_{\widehat{\boldsymbol{\Theta}}}(Y; \boldsymbol{x}) &= 2n \left[\hat{H}(f^*_y) - \hat{H}(f^*_{y|x}) \right] \\
&= -2 \log \left[\frac{f_{Y;\boldsymbol{x}}(y)|_{\lambda(x)=\hat{\lambda}(x)}}{f_Y(y)|_{\lambda=\hat{\lambda}}} \right] \\
&= 2n \widehat{K}(F^*_{Y;\boldsymbol{x}} : F^*_Y).
\end{aligned}$$

3.2 Logit

For qualitative and discrete variables, the ME solution (16) is a logit model. The ME formulation that leads to the logit solution equivalent to the logit model estimated by the MLE is given by Soofi (1992, 1994). Briefly, for a sample of n individuals, we have $\boldsymbol{y} = (y_1, \cdots, y_n)$ where $y_i = Y_i(\mathcal{A}_j) = y_{ij}, i = 1, \cdots, n,\ j = 1, \cdots, J$ are indicator functions of J choices with probability

distribution $f_{Y_i} = \boldsymbol{\pi_i} = (\pi_{i1}, \cdots, \pi_{iJ})$. Assuming independence among the individuals, the joint probability distribution of $\boldsymbol{Y}$ is given by $f_{\boldsymbol{Y}} = \prod_{i=1}^{n} \boldsymbol{\pi_i}$ and the joint entropy is given by $H(\boldsymbol{Y}) = H(\boldsymbol{\pi}) = \sum_{i=1}^{n} H(\boldsymbol{\pi_i})$, where the n probability vectors $\boldsymbol{\pi} = \{\boldsymbol{\pi_i} = (\pi_{i1}, \cdots, \pi_{iJ}),\ i = 1, \cdots, n\}$ to be estimated. The information constraints may be formulated in terms of predictors $\boldsymbol{x}$ representing the individual and/or the choice attributes. In order to distinguish between the two types of attributes we denote $\boldsymbol{u_i} = (u_{i1}, \cdots, u_{iA})'$ for the attributes of the ith individual and $\boldsymbol{v_{ij}} = (v_{ij1}, \cdots, v_{ijB})'$ for scores (values) given to the attributes of jth alternative by the ith decision maker.

The ME solution is the following logit model:

$$\pi^*_{ij} = \frac{e^{\boldsymbol{\alpha}_j' \boldsymbol{u}_i + \boldsymbol{\beta}' \boldsymbol{v}_{ij}}}{\sum_{\ell=1}^{J} e^{\boldsymbol{\alpha}_\ell' \boldsymbol{u}_i + \boldsymbol{\beta}' \boldsymbol{v}_{i\ell}}}, \tag{21}$$

where $\boldsymbol{\alpha}_j$ and $\boldsymbol{\beta}$ are the vectors of Lagrange multipliers, i.e., logit coefficients. The maximum uncertainty is given by the sum of the entropies of n probability distributions (21),

$$H^*_{\boldsymbol{\Theta}}(\boldsymbol{Y}; \boldsymbol{u}, \boldsymbol{v}) = H_{\boldsymbol{\alpha},\boldsymbol{\beta}}(\boldsymbol{\pi}^*; \boldsymbol{u}, \boldsymbol{v}) = \sum_{i=1}^{n} H(\boldsymbol{\pi^*}_i).$$

The constraints can be formulated in terms of sampled values such that the ME solutions that are equivalent to the maximum likelihood estimate (MLE) of the logit model (21) when assumed *a priori*; details are given in Soofi (1992, 1994). Then the information importance of predictors is given by the log-likelihood statistic

$$\begin{aligned} \mathcal{I}_{\hat{\boldsymbol{\alpha}},\hat{\boldsymbol{\beta}}}(\boldsymbol{Y}; \boldsymbol{u}, \boldsymbol{v}) &= = H^*(\boldsymbol{Y}) - H^*_{\hat{\boldsymbol{\alpha}},\hat{\boldsymbol{\beta}}}(\boldsymbol{Y}; \boldsymbol{u}, \boldsymbol{v}) \\ &= -2 \log \left[\frac{\boldsymbol{\pi}_{Y;\boldsymbol{x}}\big|_{(\boldsymbol{\alpha},\boldsymbol{\beta})=(\hat{\boldsymbol{\alpha}},\hat{\boldsymbol{\beta}})}}{\boldsymbol{\pi}_Y|_{\pi=\hat{\pi}}} \right] \\ &= 2n\widehat{K}(\boldsymbol{\pi}^*_{\boldsymbol{Y};\boldsymbol{x}} : \boldsymbol{\pi}^*_{\boldsymbol{Y}}), \end{aligned} \tag{22}$$

where $H^*(\boldsymbol{Y})$ is found as follows. When only the individuals' attributes are included, $\mathcal{T}_{\boldsymbol{Y}} = \{\delta(\mathcal{A}_j), j = 1, \cdots, J-1\}$ with $\theta_j = \hat{\theta}_j$ is the set of indicator function of $J-1$ choices under consideration. In this case, the ME solution is the sample proportions (null MLE), $\hat{\boldsymbol{\pi}}_i = \hat{\boldsymbol{\pi}} = (\hat{\pi}_1, \cdots \hat{\pi}_J),\ i = 1, \cdots n$, and the maximum uncertainty is the sum of the entropies of n identical distributions of the sample proportions, $H^*(\boldsymbol{Y}) = H(\boldsymbol{\pi}^*_{\boldsymbol{Y}}) = H(\hat{\boldsymbol{\pi}}_{\boldsymbol{Y}}) = nH(\hat{\boldsymbol{\pi}})$. When the choice attributes are included such indicator functions create singularity, so $\mathcal{T}_{\boldsymbol{Y}}$ is set as empty. In this case, the ME solution (null MLE) is the uniform distribution $\boldsymbol{\pi}^*_i = 1/J,\ i = 1, \cdots, n,\ j = 1, \cdots, J$ and the global maximum uncertainty is the sum of the entropies of n uniform distributions over J outcomes, $H^*(\boldsymbol{Y}) = H(1/J) = n \log J$.

Relative information importance for attributes of interest can be found by decompositions

$$\mathcal{I}_{\boldsymbol{\Theta}}(\boldsymbol{Y}; \boldsymbol{u}, \boldsymbol{v}) = \mathcal{I}_{\boldsymbol{\alpha}}(\boldsymbol{Y}; \boldsymbol{u}) + \mathcal{I}_{\boldsymbol{\beta}}(\boldsymbol{Y}; \boldsymbol{v}|\boldsymbol{u}) = \mathcal{I}_{\boldsymbol{\beta}}(\boldsymbol{Y}; \boldsymbol{v}) + \mathcal{I}_{\boldsymbol{\alpha}}(\boldsymbol{Y}; \boldsymbol{u}|\boldsymbol{v}). \tag{23}$$

4 Expected Information

For the case of stochastic predictors the expected value of entropy difference (10) and the information divergence (11) are equal, and the unique measure is referred to as the *mutual information* between Y and $\boldsymbol{X}$,

$$M(Y, \boldsymbol{X}) = E_{\boldsymbol{x}}[\Delta_H(Y|\boldsymbol{x})] = E_{\boldsymbol{x}}[K(Y|\boldsymbol{x})] \geq 0. \tag{24}$$

The interpretation and properties of mutual information are the same for discrete and continuous random variables. Other useful and insightful representations of the mutual information are:

$$\begin{aligned} M(Y, \boldsymbol{X}) &= K(F_{\boldsymbol{X},Y} : F_{\boldsymbol{X}} F_Y), & (25)\\ &= H(\boldsymbol{X}) + H(Y) - H(\boldsymbol{X}, Y) & (26)\\ &= H(Y) - H(Y|\boldsymbol{X}). & (27) \end{aligned}$$

Representations (25)-(27) provide three interpretations of the mutual information. By (25), $M(Y, \boldsymbol{X})$ is the information divergence between the actual joint distribution $F_{\boldsymbol{X},Y}$ and the distribution formed as if Y and $\boldsymbol{X}$ were stochastically independent, $G_{\boldsymbol{X},Y}(\boldsymbol{x}, y) = F_{\boldsymbol{X}}(\boldsymbol{x})F_Y(y)$. This representation shows that the mutual information is well-defined only when $F_{\boldsymbol{X},Y}$ is absolutely continuous relative to $F_{\boldsymbol{X}} F_Y$.

Representation (26) facilitates computation of mutual information using entropy expressions in terms of the distributional parameters, which are available for many multivariate distributions (Nadarajah and Zografos 2005).

In (27), $H(Y|\boldsymbol{X}) = E_{\boldsymbol{x}}[H(Y|\boldsymbol{x})]$ is the *conditional entropy* of Y given $\boldsymbol{X}$. The conditional entropy $H(Y|\boldsymbol{X})$ is the mean value of the entropies of the conditional distributions for all outcomes of the explanatory variable $\boldsymbol{X}$. Thus (27) is the representation of the mutual information as the long-run average of uncertainty reduction by the explanatory variable. By (27), $H(Y|\boldsymbol{X}) \leq H(Y)$, where the equality holds if and only if $\boldsymbol{X}$ and Y are stochastically independent.

Both indices (12) and (13) are defined in terms of (27); for this case $I(Y, \boldsymbol{X}) = 0$ if and only if the two variables are independent, and $I(Y, \boldsymbol{X}) = 1$ if and only if the two variables are functionally related in some form, linearly or nonlinearly. The mutual information admits the chain-rule decomposition of type (6); see Cover and Thomas (1991).

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

$$M(Y, \boldsymbol{X}) = M(W, \boldsymbol{X}) = M(Y, \boldsymbol{V}) = M(W, \boldsymbol{V}).$$

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. For example, correlations and other regression quantities are not invariant under nonlinear transformations.

As a final remark, we should note that in general, the equalities in (25)-(27) do not hold for other uncertainty functions and the corresponding information divergence measuring dependence. For example, for Rényi entropy and information divergence measuring dependence, the equalities in (25)-(27) do not necessarily hold.

4.1 Normal Model

When Y has a normal distribution $F_Y = N(\mu, \sigma_y^2)$, its entropy is $H_\sigma(Y) = .5 \log\left(2\pi e \sigma_y^2\right)$. If the conditional distribution $F_{Y|\boldsymbol{x}} = N(\boldsymbol{z}\boldsymbol{\gamma}, \sigma^2)$, where $\boldsymbol{z} = (1, \boldsymbol{x})$ and $\boldsymbol{\gamma} = (\beta_0, \boldsymbol{\beta})'$ are $p+1$ dimensional vectors, then its entropy is $H_{\sigma,\rho}(Y; \boldsymbol{x}) = .5 \log\left[2\pi e \sigma_y^2 \{(1 - \rho^2(Y, \boldsymbol{X})\}\right]$, where $\rho^2(Y, \boldsymbol{X}) = 1 - \sigma^2/\sigma_y^2$ is the squared multiple correlation. We note that for the normal model, $H_{\sigma,\rho}(Y; \boldsymbol{x})$ does not vary with the outcomes $\boldsymbol{x}$. Thus, the normal mutual information is equal to the entropy difference (10), and is given by

$$M_{\boldsymbol{\Theta}}(Y, \boldsymbol{X}) = \mathcal{I}_{\boldsymbol{\Theta}}(Y; \boldsymbol{x}) = -.5 \log[1 - \rho^2_{\boldsymbol{\Theta}}(Y, \boldsymbol{X})], \tag{28}$$

where $\boldsymbol{\Theta} = (\boldsymbol{\gamma}, \sigma^2)$ is the $p+2$ dimensional vector of parameters.

The normal distribution is the ME model in the class of distributions subject to the mean and variance constraints. Hence the ME information importance (20) is applicable and leads to the same result as the expected information importance (28). For the normal model the information importance is determined by the squared correlations since the form of the functional relationship can only be linear. In this case, the mutual information index is the same as the squared correlation.

Decomposition (6) for the normal regression is given by the partial mutual information

$$\begin{aligned} M_{\boldsymbol{\Theta}}(Y, X_k | X_1, \cdots X_{k-1}) &= -.5 \log[1 - \rho^2_{\boldsymbol{\Theta}}(Y, X_k | X_1, \cdots X_{k-1})] \\ &= M_{\boldsymbol{\Theta}}[Y, (X_1, \cdots X_k)] - M_{\boldsymbol{\Theta}}[Y, (X_1, \cdots X_{k-1})] \\ &= \frac{1}{2} \log\left(\frac{1 - \rho^2_{\boldsymbol{\Theta}}[Y, (X_1, \cdots X_{k-1})]}{1 - \rho^2_{\boldsymbol{\Theta}}[Y, (X_1, \cdots X_k)]}\right), \end{aligned} \tag{29}$$

where $\rho^2_{\boldsymbol{\Theta}}(Y, X_k | X_1, \cdots X_{k-1})$ is the partial correlation between Y and X_k, given $X_1, \cdots, X_{k-1}$. Successive application of (29) provides 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 (28).

Formula (28) for normal mutual information is very simple, but the normality of the distributions is crucial for its validity. For non-normal data, transformations to normality are therefore crucial. Suppose that we have data on a set of variables $W, \boldsymbol{V} = (V_1, \cdots, V_p)$ and transform the variables as $Y = S(W)$ and $X_k = T_k(V_k)$, where all transformations are one-to-one and $(Y, X_1, \cdots, X_p)$

are normal. Then, by the invariance property of the mutual information, we can compute the importance of the original explanatory variables $\boldsymbol{V}$ for prediction of W by

$$M_{\boldsymbol{\Theta}}(W, \boldsymbol{V}) = M_{\boldsymbol{\Theta}}(Y, \boldsymbol{X}) = -.5 \log[1 - \rho^2(Y, \boldsymbol{X})]. \tag{30}$$

Derivation of $M_{\boldsymbol{\Theta}}(W, \boldsymbol{V})$ directly from the joint distribution of $(W, \boldsymbol{V})$ could be difficult or impossible. However, transformation to normality often is achieved in regression analysis. Invariance is a very useful property of an importance measure. For example, Box-Cox transformations are one-to-one, but not linear. Consequently, all regression quantities must be interpreted in terms of the transformed data. However, mutual information retains its interpretation in terms of the original data.

5 Bayesian Inference

The information importance measures $\mathcal{I}_{\boldsymbol{\Theta}}(Y; \boldsymbol{x})$ and $M_{\boldsymbol{\Theta}}(Y, \boldsymbol{X})$ are functions of the model parameters $\boldsymbol{\Theta}$. Usually $\boldsymbol{\Theta}$ is unknown and induces uncertainty about the information importance measures. We present Bayesian inference for these information measures by describing prior uncertainty about the parameter vector $\boldsymbol{\Theta}$ via specifying a prior distribution $g(\boldsymbol{\Theta})$. Given the data $D = (y_i, \boldsymbol{x}_i),\ i = 1, \cdots, n$ we update the prior $g(\boldsymbol{\Theta})$ to the posterior distribution $g(\boldsymbol{\Theta}|D)$ via the Bayes rule

$$g(\boldsymbol{\Theta}|D) \propto \mathcal{L}(\boldsymbol{\Theta}; D) g(\boldsymbol{\Theta}),$$

where the form of the likelihood function $\mathcal{L}(\boldsymbol{\Theta}; D)$ is determined by the probability model. For some models, e.g, the normal regression, the posterior distribution of $\boldsymbol{\Theta}$ can be evaluated analytically using conjugate priors, but for many other cases such as logit models, the posterior distribution of $\boldsymbol{\Theta}$ is not available analytically and inference requires use of Markov Chain Monte Carlo (MCMC) methods.

Once the posterior distribution $g(\boldsymbol{\Theta}|D)$ is available either in analytical form or via simulation, the posterior distribution of information importance measures $\mathcal{I}_{\boldsymbol{\Theta}}(Y; \boldsymbol{x})$ and $M_{\boldsymbol{\Theta}}(Y, \boldsymbol{X})$ can be obtained. If the posterior distribution $g(\boldsymbol{\Theta}|D)$ is analytically available, it may be possible to evaluate the posterior distributions of information importance measures analytically for some simple cases. However, in general, the information measures are not simple functions of the model parameters and must be approximated using standard Monte Carlo methods by generating samples from $g(\boldsymbol{\Theta}|D)$. Posterior means and other moments can also be approximated using a Monte Carlo procedure. For example, the posterior mean of $\mathcal{I}_{\boldsymbol{\Theta}}(Y; \boldsymbol{x})$ is approximated as

$$E[\mathcal{I}_{\boldsymbol{\Theta}}(Y; \boldsymbol{x}|D)] \approx \frac{1}{S} \sum_{s=1}^{S} \left[\mathcal{I}_{\boldsymbol{\Theta}_{(s)}}(Y; \boldsymbol{x})\right]$$

based on posterior samples $\boldsymbol{\Theta}_{(s)}$, $s = 1, \cdots, S$, from $g(\boldsymbol{\Theta}|D)$. Posterior distributions for other importance measures, such as the mutual information index $M_{\boldsymbol{\Theta}}(Y, \boldsymbol{X})$, can also be obtained in a similar manner. Some special cases will be discussed next.

5.1 Bayesian Inference for the Normal Model

For the normal regression we can specify a conjugate multivariate normal-inverse gamma prior for $\boldsymbol{\Theta} = (\boldsymbol{\gamma}, \sigma^2)$ as

$$g(\boldsymbol{\Theta}) = g(\boldsymbol{\gamma}, \sigma^2) = g(\boldsymbol{\gamma}|\sigma^2)g(\sigma^2)$$

where $g(\boldsymbol{\gamma}|\sigma^2)$ is a multivariate normal and $g(\sigma^2)$ is an inverse gamma density. Standard Bayesian updating yields a multivariate normal-inverse gamma posterior density; see for example Zellner (1971). If the precision matrix of the multivariate normal $g(\boldsymbol{\gamma}|\sigma^2)$ is specified as the zero matrix and the scale parameter of the inverse gamma density $g(\sigma^2)$ is set to zero then an improper joint prior is obtained. A commonly used form is given by

$$g(\boldsymbol{\gamma}, \sigma^2) \propto \frac{1}{\sigma}.$$

Let $\boldsymbol{Z} = [\mathbf{1} : \tilde{\boldsymbol{X}}]$ where $\mathbf{1}$ is an $n \times 1$ vector of ones and $\tilde{\boldsymbol{X}}$ denotes the data matrix on $\boldsymbol{X}$. It is well known that if $\boldsymbol{Z'Z}$ is nonsingular, then the posterior distribution $g(\boldsymbol{\gamma}, \sigma^2|D)$ is a proper multivariate normal-inverse gamma density. Then the posterior mean of $\boldsymbol{\gamma}$ given σ^2 is given by the least squares estimator $\widehat{\boldsymbol{\gamma}}$ and the posterior variance is $\sigma^2(\boldsymbol{Z'Z})^{-1}$. In the regression model application of Section 6, we will use the improper prior.

Following Press and Zellner (1978) we can write the squared multiple correlation as

$$\rho^2_{\boldsymbol{\Theta}}(Y, \boldsymbol{X})) = \frac{\boldsymbol{\beta' Q \beta}}{\boldsymbol{\beta' Q \beta} + n\sigma^2}, \tag{31}$$

where n is the number of observations and

$$\boldsymbol{Q} = \tilde{\boldsymbol{X}}'\tilde{\boldsymbol{X}} - \frac{1}{n}\left(\tilde{\boldsymbol{X}}'\mathbf{1}\mathbf{1}'\tilde{\boldsymbol{X}}\right).$$

It is important to note that in the above representation $\boldsymbol{\beta}$ excludes the intercept term and σ^2 is the conditional variance of Y.

The posterior distribution of $\rho^2_{\boldsymbol{\Theta}}(Y, \boldsymbol{X})$ can be approximated by Monte Carlo simulation drawing samples from $g(\boldsymbol{\gamma}, \sigma^2|D)$ and computing (31). Then the posterior distribution of mutual information $M_{\boldsymbol{\Theta}}(Y, \boldsymbol{X})$ is approximated using (28). Posterior distributions of the components of the decomposition (29) can be evaluated in a similar manner using regression models with different numbers of predictors. By using the chain rule, we can also evaluate the relative information importance based on all orderings of the predictors as given by (7). More specifically, we can evaluate partial mutual information (29) for $k = 1, \cdots, p$ and for all p permutations and take an average. Note that

by (29) for any set of the variables $\rho^2_{\boldsymbol{\Theta}}$ of the set must be more than $\rho^2_{\boldsymbol{\Theta}}$ of all of its subsets. The posterior sample values not satisfying such constraints will be rejected. If we define the average relative importance for X_k as $\overline{M}_{\boldsymbol{\Theta}}(Y, X_k)$ then we can obtain its posterior distribution based on the samples from the posteriors.

5.2 Bayesian Inference for the Logit Model

For the logit model (21), for any choice of prior distribution $g(\boldsymbol{\Theta}) = g(\boldsymbol{\alpha}, \boldsymbol{\beta})$, the posterior distribution can not be obtained in analytical form. However, Bayesian analysis for the logit model can be developed using MCMC techniques such as Gibbs sampling or Metropolis-Hastings algorithm; see for example Casella and George (1992) and Chib and Greenberg (1995). Such analysis can be easily performed in an environment such as WinBUGS.

It is common to use diffused but proper normal priors for components of $g(\boldsymbol{\alpha}, \boldsymbol{\beta})$. Once the samples from the posterior distribution $g(\boldsymbol{\alpha}, \boldsymbol{\beta}|D)$ are generated via MCMC methods the logit probabilities are evaluated via $\pi_{ij}(\boldsymbol{\alpha}, \boldsymbol{\beta})$. This provides a posterior distribution $g(\pi_{ij}|D)$ for the choice probabilities π_{ij}'s. Then the entropy of the logit model for n individuals given by the posterior distribution of $H^*_{\boldsymbol{\alpha}, \boldsymbol{\beta}}(\boldsymbol{\pi}; \boldsymbol{u}, \boldsymbol{v})$ and can be obtained from (21) by using samples from $g(\boldsymbol{\alpha}, \boldsymbol{\beta}|D)$. The normalized ME information index and the chain rule for the logit can be obtained similar to case of normal regression model.

6 Applications

6.1 Financial Data

This example demonstrates the versatility of mutual information for measuring importance and shows Bayesian inference about relative importance of predictors in linear regression. We use a subset of variables chosen from the Stock Liquidity data described in Frees (1996, p. 263). The variables are: the trading volume for a three month period in millions shares (Volume W), 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). We have chosen these variables for the purpose of illustration. Figure 1 shows the residual plots of the linear regression for these variables and for their log-transformations, $Y = \log W$, $X_k = \log X_k, k = 1, 2, 3$. Clearly the normality assumption is violated for the original data, but the conditional normality seems plausible for the transformed data.

Table 1 shows the regression results and information analysis for the data. Panel (a) of Table 1 shows the least squares regression results and joint information importance of three predictors. The regression results for original and transformed variables are different. The regression results for the original data are not valid, but for the transformed data are valid. We do not need the

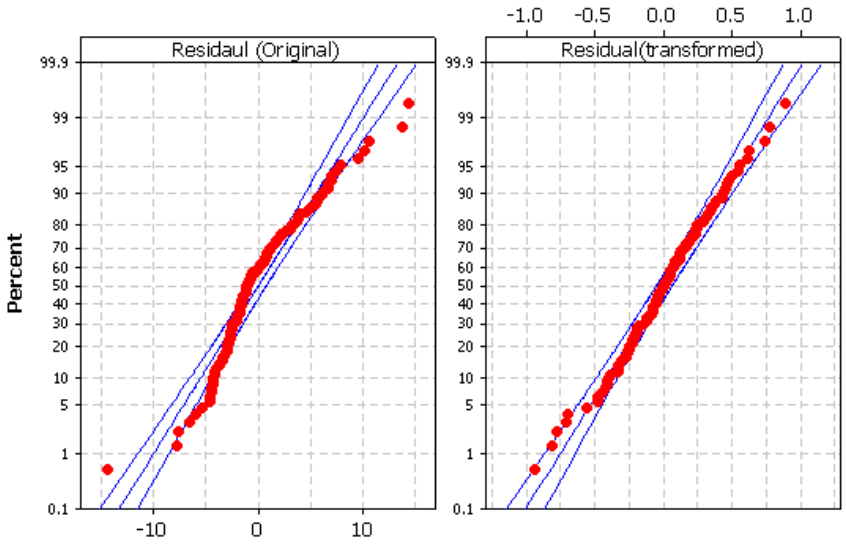

Figure 1: Residual plots of linear regression of financial variables and their log-transformations.

regression results for the original data for computing the information importance for Transaction, Share, and Value for predicting the Volume. Since we were able to transform the data to normality, we can use the R^2 of the of transformed variables ($R^2 = 84.3\%$) for computing and comparing the information importance of Transaction, Share, and Value for the Volume. By (30), our information importance analysis is applicable to the original as well as the transformed variables.

Panel (b) of Table 1 shows the F-ratio, R^2, mutual information and posterior results for the information importance of each subset of the predictors. Again note that F and R^2 are interpretable in terms of the transformed variables, yet the results for information importance are interpretable in terms of the original as well as the transformed variables. The posterior intervals of information importance are useful for inference about subset models. We note that the posterior intervals for the single variable models do not intersect, so we can infer that the importance of model containing X_1 is the highest, X_3 is the lowest, and X_2 is in between. For the two variable models we can infer that the importance of the models containing X_1 are not different from one another, but are different from the model not containing X_1. Furthermore, we can infer that the importance of the models containing X_1 are not different, but are different from the models that not containing X_1. This analysis establishes that for predicting Volume, the models containing Transaction are of higher importance those not containing it. 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.

Panel (c) of the Table 1 gives the decompositions of the joint information importance into the relative importance of each variable (7) for all six orderings of the variables. The entries are

Table 1. Regression results and information importance for log-normal and transformed data.

(a) Regression results

	Original data			Transformed data		
	V_1	V_2	V_3	X_1	X_2	X_3
Coefficient	.002	.009	- 0.040	.750	.236	.037
S.E.	.0001	.0071	.0894	.0682	.0786	.0581
F-ratio, d.f.'s (2,97)	203.10			213.30		
R^2	83.7%			84.3%		
Mutual information	.926			.926		

(b) Information importance of all subsets of variables

	Data (Likelihood)				Bayes (Posterior)		
Subset	F	d.f.'s	R^2	Information	Mean	SD	95% Interval
X_1	493.21	$122, 1$	.803	.812	.743	.032	$(.704, .809)$
X_2	259.50	$122, 1$	.682	.573	.536	.032	$(.498, .601)$
X_3	163.04	$122, 1$	.574	.427	.408	.030	$(.371, .470)$
X_1, X_2	322.17	$122, 2$	.843	.926	.847	.037	$(.801, .922)$
X_1, X_3	295.03	$122, 2$	.831	.889	.815	.036	$(.770, .886)$
X_2, X_3	129.87	$122, 2$	.684	.576	.572	.053	$(.506, .678)$
X_1, X_2, X_3	213.30	$122, 3$	.843	.926	.855	.043	$(.801, .942)$

(c) Information importance of three variables for all orderings

Ordering	X_1	X_2	X_3	(X_1, X_2, X_3)
$X_1X_2X_3$	.812	.113	.000	.926
$X_1X_3X_2$	.812	.037	.077	.926
$X_2X_1X_3$	.353	.573	.000	.926
$X_2X_3X_1$	.350	.573	.003	.926
$X_3X_1X_2$	.462	.037	.427	.926
$X_3X_2X_1$	.350	.149	.427	.926
Average	.523	.247	.156	.926

Posterior results for averages

Mean	.462	.237	.157
S.D.	.009	.018	.016
95% Interval	$(.450, .480)$	$(.215, .273)$	$(.136, .189)$

Pairwise differences (Column - Row)

	X_1	X_2
X_2	$(.204, .235)$	
X_3	$(.287, .314)$	$(.079, .083)$

computed using the subset MLE information measures. 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 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 is computed using equal weights $w_q = 1/6$ in (8). These results indicate the overall average relative importance of the three variables.

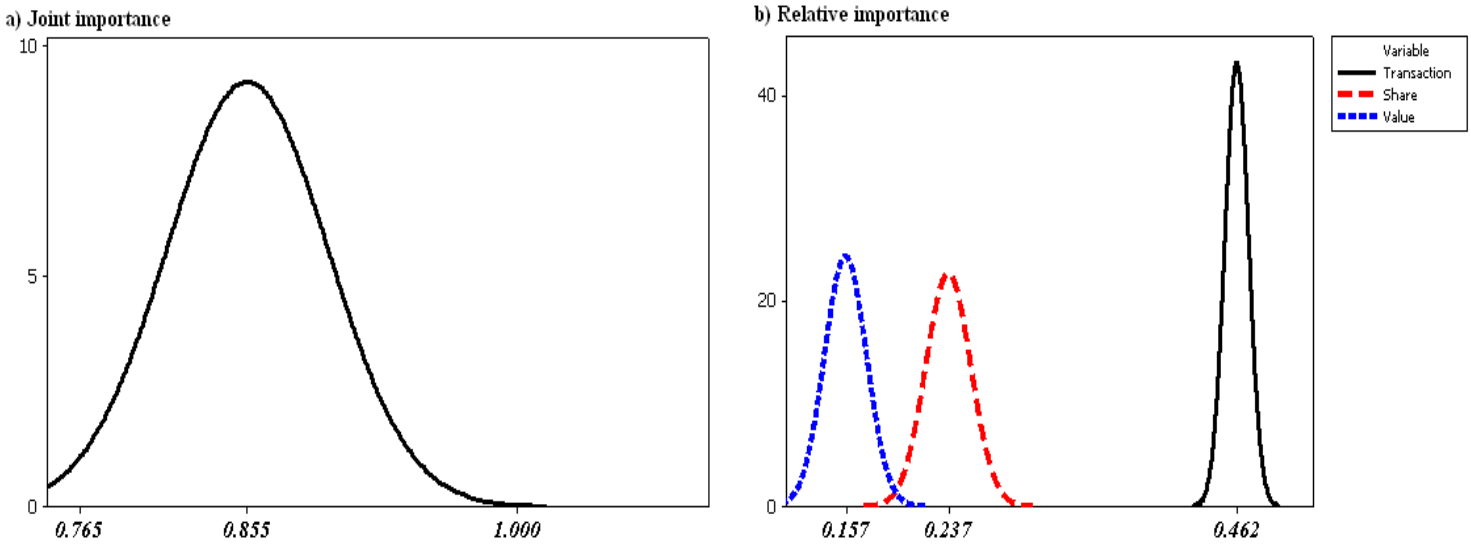

Figure 2: Posterior distributions of importance measures of predictors for Financial Data.

The average information importance of Transaction is more than twice that of Share and is more than three times that of Value.

Posterior intervals for the average importance of each variable and the pairwise differences between them are also shown in Panel (c) of Table 1. We can infer that overall, Transaction (X_1) is the most important variable, followed by Share (X_2), followed by Value (X_3), and that the importance of three variables are different. Posterior distributions for the joint importance and overall average importance of the three variables are shown in Figure 2. We note that the posterior distribution for the joint importance is skewed. The posterior distributions for the average relative importance of variables are close to normal, due to central tendency.

6.2 Choice of 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 the 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 (X_1), price (X_2), and offering calling plans that meet the customers needs (X_3). Each explanatory variable has two categorical outcomes: low and high. Assessment of 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 for this data.

Table 2. Information importance of three explanatory variables for long distance providers.

(a) Data

Reputation (X_1)	Low				High				
Price (X_2)	Low		High		Low		High		
Plans (X_3)	Low	High	Low	High	Low	High	Low	High	Total
Service Provider (Y)									
Sprint	113	35	19	36	21	27	8	74	333
AT&T	98	18	8	9	60	66	8	113	380
MCI	73	17	7	15	5	9	6	32	164
Total	284	70	34	60	86	102	22	219	877

(b) Information importance of all subsets of variables

	Data (Likelihood)			Bayes (Posterior)		
Subset	Information	Chi-sq.	d.f.	Mean	SD	95% Interval
X_1	.041	72.79	2	.041	.004	(.034, .049)
X_2	.001	1.93	2	.003	.001	(.002, .004)
X_3	.002	4.21	2	.004	.001	(.002, .004)
X_1, X_2	.050	86.82	6	.055	.005	(.046, .063)
X_1, X_3	.048	83.31	6	.049	.004	(.039, .057)
X_2, X_3	.006	9.82	6	.009	.001	(.007, .010)
X_1, X_2, X_3	.060	104.91	14	.063	.004	(.054, .071)

(c) Information importance of three variables for all orderings

Ordering	X_1	X_2	X_3	(X_1, X_2, X_3)
$X_1X_2X_3$	.041	.008	.011	.060
$X_1X_3X_2$	.041	.013	.006	.060
$X_2X_1X_3$	.048	.001	.011	.060
$X_2X_3X_1$	.055	.001	.004	.060
$X_3X_1X_2$	.045	.013	.002	.060
$X_3X_2X_1$	.055	.003	.002	.060
Average	.047	.007	.006	.060

Posterior results for averages

Mean	.048	.009	.006
S.D.	.0037	.0004	.0004
95% Interval	(.041, .055)	(.008, .009)	(.005, .007)

Pairwise differences (Column - Row)

	X_1	X_2
X_2	(.032, .046)	
X_3	(.036, .049)	(.002, .003)

Table 2 shows the data and importance analysis. Panel (a) of Table 2 shows the data in a $3\times2\times2\times2$ contingency table. Panel (b) of Table 2 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 using the cell frequencies. These

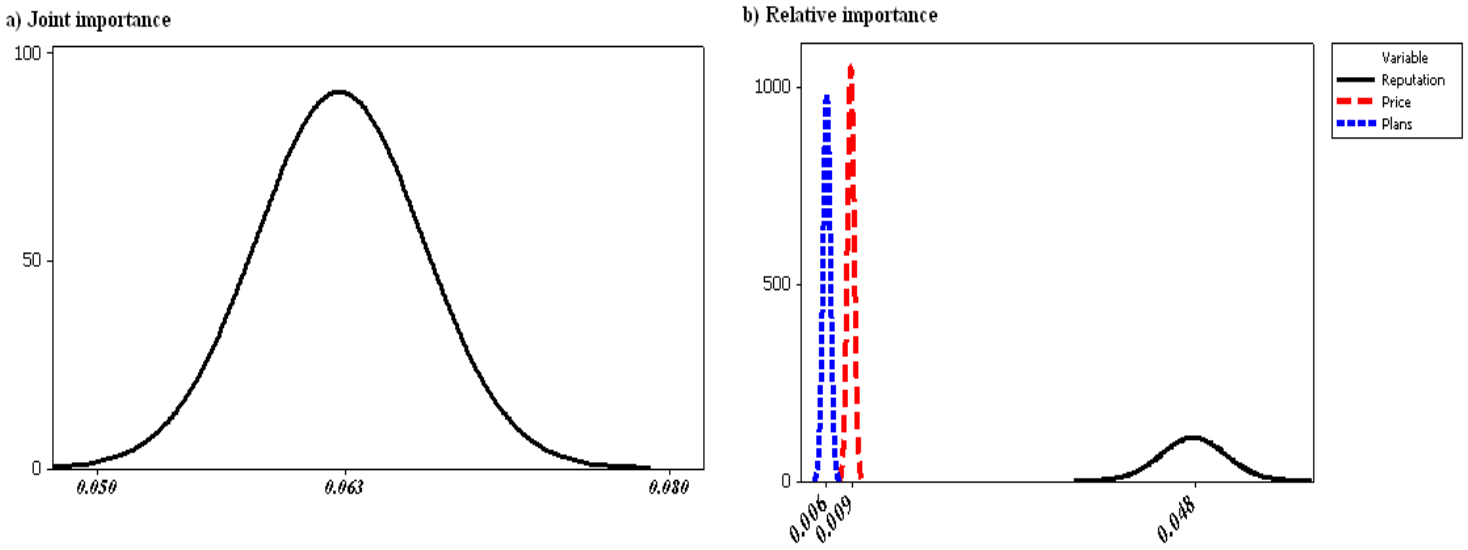

Figure 3: Posterior distributions of Importance and Relative Importance of Predictors for Long Distance Provider Data.

measures can also be obtained by ME formulations subject to marginal constraints (Gokhale and Kullback 1978, Soofi and Retzer 2002). The information chi-square statistics are found by $\Xi^2 = 2n\mathcal{I}(Y, \boldsymbol{X})$. 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 results are obtained via the logit formulation.

The posterior intervals of information importance shown in Panel (b) suggest the following inferences for subset models. The posterior intervals for the X_1 model does not intersect the posterior intervals for the models including each of the other two variables singly, so we can infer among these models, the model containing X_1 is of the highest importance,and the other two single variable models are of equally low importance since their posterior intervals intersect. For the two-variable models we can infer that the importance of the models containing X_1 are not different from one another, but are different from the model not containing X_1. We can also infer that the importance of the models containing a single variable are less than the importance of the full model containing all three variables. However, the importance of the full model is not different from any of the two variable models containing X_1, but is different from the two variable model that does not contain X_1. This analysis suggests that for predicting the choice of long distance provider, the models containing Reputation and any of the other two variables are of importance. 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.

Panel (c) of the Table 2 gives the decompositions of the joint information in terms of all six orderings of the variables. We note that the information importance measures are order-dependent,

particularly for Price (X_2) and Plan offering (X_3). The average information importance of Reputation is about seven times that of Price and is about eight times that of Plan offering.

Posterior intervals for the average importance of each variable and the pairwise differences between them are shown in Panel (c) of Table 2. We can infer that overall, Reputation (X_1) is the most important variable, far distant Price (X_2) and Plan offering (X_3), and that the importance of the three variables are different. The posterior distributions for the joint importance and overall average importance of the three variables are shown in Figure 3. In this case, all four distributions resemble normality.

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 (price, efficiency, and quality) was needed for inputs to a business decision by the technology provider.

The hospital size categories are small, medium, and large. The technology attributes are scores. The ME procedure (logit analysis) is implemented as follows. Hospital size is represented by two indicator variables as: U_1 for small and U_2 for medium; the large size is the base category $(U_1, U_2) = (0, 0)$.

For the assessment of the importance of the explanatory variables we compute the ME information indices. The ME solutions for the choice distributions are given by logit (21). Thus, the exponential family regression results can be used for the information importance computations. The information analysis results are obtained using of exponential family regression via logit formulation.

Table 3 shows the results. Panel (a) of the table gives the logit coefficients (Lagrange multipliers for the ME derivation). These are obtained using SAS PROC PHREG (Allison 1999) for the model containing both types of variables. The log-likelihood chi-square statistics for the variables are related to information measures $\Xi^2 = 2n\mathcal{I}_{\hat{\boldsymbol{\Theta}}}(Y; x_k) = H^*_{\hat{\boldsymbol{\Theta}}}(Y; \boldsymbol{x}_{(k)}) - H^*_{\hat{\boldsymbol{\Theta}}}(Y; \boldsymbol{x})$, where $\boldsymbol{x}_{(k)}$ is the vector excluding x_k, $k = 1, \cdots, 7$.

Panel (b) of Table 3 shows the information importance, the chi-square statistics, their degrees of freedom, and posterior results for the hospital size (two variables), for technology attributes (three variables), and for both groups (full model). The information chi-square statistics are given by (22). Since the choice attributes are included, the global ME model (null model) is the uniform distribution over the three choices $\boldsymbol{\pi_i} = (1/3, 1/3, 1/3)$, $i = 1, \cdots, 121$ and $H^*(Y) = 121 \log 3 = 132.93$. The log-likelihood function without variables (null) is $-2H^*(Y)$. The log-likelihood function with

Table 3. Information importance for choice of medical technology.

(a) MLE Logit

	Organization Size (S)				Technology (T)		
	Small		Medium		Price	Efficiency	Quality
	$j=1$	$j=2$	$j=1$	$j=2$	P	E	Q
Logit coefficient	2.24	3.29	1.71	2.23	.81	.63	1.11
Standard Error	1.21	1.16	.51	.52	.20	.22	.26
Chi-square (df=1)	3.46	7.98	11.32	18.68	16.16	8.47	18.60

(b) Subset of types of attributes

	Data (Likelihood)			Bayes (Posterior)		
Subset	Information	Chi-sq.	d.f.	Mean	SD	95% Interval
Hospital size S	.148	39.38	4	.164	.035	$(.095, .232)$
Technology $T=(P,E,Q)$	.289	76.84	3	.299	.049	$(.200, .397)$
Both types (S,T)	.447	118.94	7	.472	.052	$(.367, .567)$

(c) Information importance of types of attributes

Ordering	S	T	(S,T)
ST	.148	.299	.447
TS	.158	.289	.447
Average	.153	.294	.447

Posterior results for averages			
Mean	.168	.304	
S.D.	.019	.033	Difference $T-S$
95% Interval	$(.131, .201)$	$(.236, .366)$	$(.106, .165)$

(d) Information importance of technology variables over orderings beyond the size

Ordering	$P\|S$	$E\|S$	$Q\|S$	$(P,E,Q)\|S$
PEQ	.123	.089	.087	.299
PQE	.123	.036	.140	.299
EPQ	.091	.120	.087	.299
EQP	.070	.120	.109	.299
QPE	.091	.036	.172	.299
QEP	.070	.057	.172	.299
Average	.094	.077	.128	.299

Posterior results for averages				
Mean	.099	.078	.132	.308
S.D.	.005	.005	.007	.017
95% Interval	$(.089, .107)$	$(.069, .087)$	$(.114, .142)$	$(.272, .335)$

Pairwise differences (Column - Row)		
$E\|S$	$(.019, .023)$	
$Q\|S$	$(-.035, -.026)$	$(-.056, -.045)$

variables (model) is $-2H^*_{\hat{\boldsymbol{\Theta}}}(Y;\boldsymbol{x})$, where $\boldsymbol{x}=\boldsymbol{v}$ for the hospital size, $\boldsymbol{x}=\boldsymbol{u}$ for technology, and $\boldsymbol{x}=\boldsymbol{u},\boldsymbol{v}$ for the two sets combined.

The posterior intervals of information importance shown in Panel (b) suggest the following inferences for hospital size and technology attributes. The posterior intervals for the submodels

intersect, so we do not infer the importance of one is higher than the other. The intervals for the model containing the size variables and the full model do not intersect, leading to inference that the importance of the full model is higher than the model containing the size variables. But the intervals for the model containing technology attributes and the full model intersect, leading to the inference that the importance 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.

Panel (c) of Table 3 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 information importance measures are not strongly order-dependent. The average relative information importance of the hospital size is about half of the technology attributes. Posterior results for the average relative importance for each group of variables and the difference between the averages of the two groups are also shown in Panel (c). We can infer that the average importance of technology attributes is higher than the hospital size.

Panel (d) of Table 3 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 dependency of the information importance. When each technology variable is in the first position in the sequence (i.e., is singly added to the size model), the incremental contributions of price and technology efficiency are almost equally important ($\mathcal{I}_{\hat{\mathbf{\Theta}}}(P|S) = .123$ and $I\mathcal{I}_{\hat{\mathbf{\Theta}}}(E|S) = .120$), but the importance of quality is higher ($\mathcal{I}_{\hat{\mathbf{\Theta}}}(Q|S) = .172$). The importance of each variable when last in a sequence deteriorates to about half for price and quality ($\mathcal{I}_{\hat{\mathbf{\Theta}}}(P|EQS) = .070$ and $\mathcal{I}_{\hat{\mathbf{\Theta}}}(Q|EPS) = .087$), and to almost one-third for the efficiency ($\mathcal{I}_{\hat{\mathbf{\Theta}}}(E|PQS) = .036$). The average importance over all orderings gives ratios of about 9:8:13 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 (d). We can infer that overall, Quality $(Q|S)$ is the most important variable, followed by Efficiency $(E|S)$, followed by Price $(P|S)$, and that the importance of three variables are different. The posterior distributions for the joint importance and overall average importance of the three variables are shown in Figure 4.

7 Conclusions

The importance methodology research has been mainly concerned with providing warnings against the use of usual statistical quantities such as P-value and standardized coefficients, suggesting certain measures, and developing some frameworks for the properties of the importance measures.

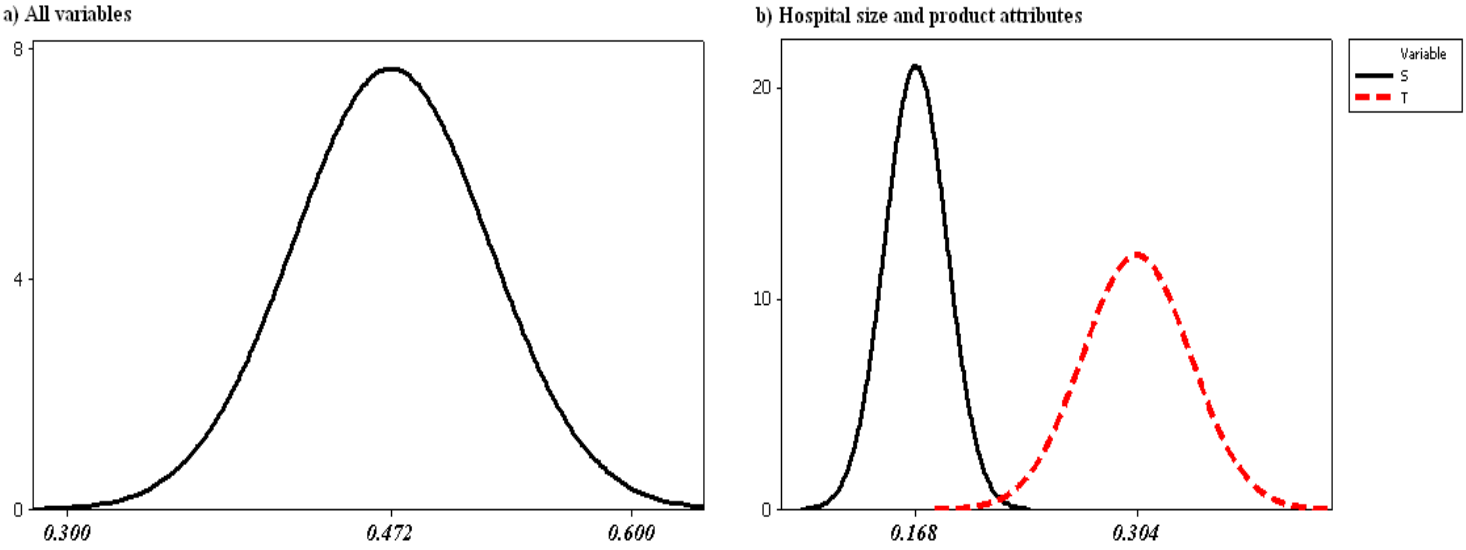

Figure 4: Posterior distributions of Importance and Relative Importance of Hospital size and Product Attributes for Medical Technology.

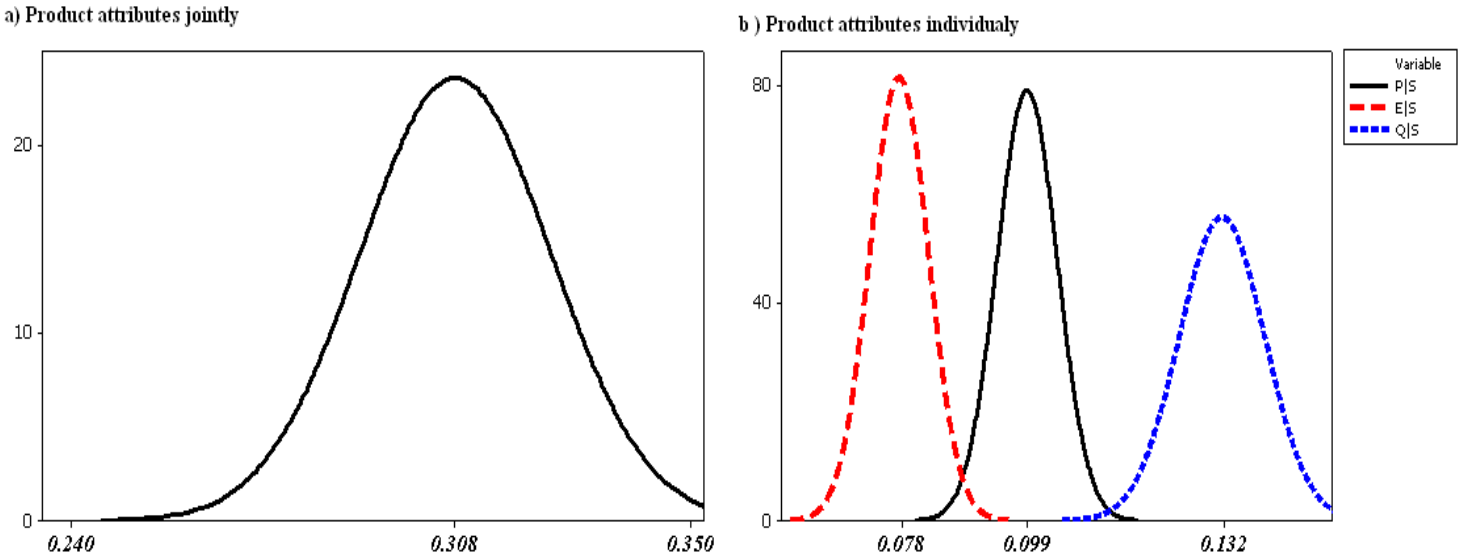

Figure 5: Posterior distributions of Importance and Relative Importance of Product Attributes for Medical Technology in Addition to the Hospital Size.

Little attention has been given to the more general notion of "importance" itself. This paper 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. Information measures of importance are applicable to qualitative as well as continuous random variables. Within the framework of information theory, importance measures for qualitative, categorical, discrete, and continuous explanatory and response variables are provided in a

unified manner.

In the information theoretic approach uncertainty is mapped by a concave function of probability density with global maximum at the uniform distribution reflecting the most unpredictable situation. We conceptualize information importance of predictors in terms of the difference between the uncertainty associated with the probability distributions of the response variable when the predictors are absent and present. We operationalized the uncertainty reduction in terms of Shannon entropy.

For nonstochastic predictors, the ME formulation provides importance measures. The ME information measure is a versatile measure for quantifying the importance of explanatory variables. The ME measures are particularly useful when the explanatory variables are continuous or have several levels, and the response variable is qualitative. In these situations a model is needed for relating the probabilities of the response outcomes to the explanatory variables. The ME procedure derives the model along with the importance measures. However, 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. The conceptual implication is that the importance of an explanatory variable X does not depend on the form of its functional relationship with the response variable Y. If X decreases uncertainty for predicting Y in the long-run, then it does not matter whether we use X or any function of it, e.g., $g(X)$ to predict a response Y or a different function, e.g., $q(Y)$, as long as each variable is identifiable from its transformation. This is reflective of the probabilistic nature of dependence/independence between two variables. When a distinction between functional forms of relationship is needed, one may use the entropy difference (10) for a given outcome of the explanatory variable or use a measure that is based on a more specific notion of dependence.

The practical implication of the invariance property of the mutual information is that if the distributions of variables are not normal, but can be transformed to normality, then the multivariate normal mutual information formula can be used for computing the actual mutual information between the original variables. This was illustrated in the normal regression model using the usual log transformation of variables that have skewed distributions. With transformation to normality, the regression results change and must be interpreted in terms of the transformed variables. However, the information importance of the transformed variables remains unchanged.

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 information analysis. The notion of information importance and the Bayesian in-

ference methods we present here have potential applications in Bayesian networks that deal with assessment of conditional independence based on high dimensional data. 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 considered also in *deviance information criterion (DIC)* proposed by Spiegelhalter et al. (2002). Our current work investigates this connection and explores information importance in terms of Bayes factors; see Kass and Raftery (1995) and Bayesian model averaging.

Three examples illustrated 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 regression. In this example using textbook data, we assessed the relative importance of the total number of transactions, number of outstanding shares and market value of the firm in predicting its trading volume. The distributions of all four variables were highly skewed, so normal regression analysis was deemed inappropriate. However, normality could be achieved by a log transformation. The normal regression of the transformed variables allowed assessments of the importance and relative importance of the explanatory variables.

Two other examples illustrated real world applications. In the choice of long distance provider example, we assessed the relative importance of long distance company's reputation, price, and plan offering for the customer's choice among three providers. In this example, all variables were qualitative. 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 how the logit output of a statistical package (e.g. SAS) can be used to derive an ME logit model and compute the ME information indices for the attributes. The ME measures are applicable to various logit models and can be easily computed from the logit outputs.

Furthermore, the examples also illustrated how Bayesian information analysis can be developed using MCMC methods and what additional insights about information importance can be obtained from the Bayesian analysis. Such analysis can be easily performed using WinBUGS which is publicly available (www.mrc-bsu.cam.ac.uk/bugs). The Bayesian approach gave us some additional insights in the analysis. For example, we noted that posterior intervals of information importance of models with subset variables provided us insights about subset selection.

References

Ahmad N. A., Gokhale D. V., 1989. Entropy expressions and their estimators for multivariate distributions. IEEE Transactions on Information Theory IT-35(3), 688-692.

Allison, P. D., 1999. Logistic Regression Using the SAS System: Theory and Application. North Carolina: SAS Institute Inc.

Azen R., Budescu, D. V., 2003. The dominance analysis approach for comparing predictors in multiple regression. Psychological Methods, 8, 129-148.

Budescu, D. V., 1993. Dominance analysis: a new approach to the problem of relative importance of predictors in multiple regression", Psychological Bulletin, 114, 542-551.

Casella, G., George, E. I., 1992. Explaining the Gibbs sampler. The American Statistician, 46, 167-174.

Chevan, A., Sutherland, M., 1991. Hierarchical Partioning. The American Statistician 45, 90-96.

Chib, S., Greenberg, E., 1995. Understanding the Metropolis-Hastings algorithm. The American Statistician, 49, 327-335.

Cover, T. M., Thomas, J. A., 1991. Elements of Information Theory. John Wiley, New York.

Cox, L. A., 1985. A new measure of attributable risk for public health applications. Management Science, 31, 800- 813.

DeGroot, M. H., 1962. Uncertainty, information, and sequential experiments. Annals of Mathematical Statistics 33, 404-419.

Dempster, A. P.,1997. The direct use of likelihood for significance testing, Statistics and Computing, 7, 247-252.

Ebrahimi, N., Maasoumi, E., Soofi, E. S., 1999. 'Ordering univariate distributions by entropy and variance. Journal of Econometrics 90, 317-336.

Ebrahimi, N., Kirmani, S.N.U.A., Soofi, E.S., 2007. Dynamic information about parameters of lifetime distribution. 56th Session of International Statistical Institute, Lisbon, Portugal.

Ebrahimi, N., Soofi, E.S., Soyer, R., 2007. Multivariate maximum entropy identification, transformation, and dependence. Journal of Multivariate Analysis. Available at Science Direct.

Frees, E. W., (1996). Data Analysis Using Regression Models: The Business Perspective. Prentice-Hall, Englewood Cliffs, NJ.

Genizi, A., 1993. Decomposition of R^2 in multiple regression with correlated variables. Statistica Sinia, 3, 407-420.

Goel, P.K. and DeGroot, M.H., 1981. Information about hyperparameters in hierarchical models. Journal of the American Statistical Association 76, 140-147.

Gokhale, D. V., Kullback, S., 1978. The Information in Contingency Tables. Marcel Dekker, New York.

Grömping, U., 2007. Estimators of relative importance in linear regression based on variance decomposition. The American Statistician, 61, 139-147.

Jaynes, E. T., 1957. Information theory and statistical mechanics. Physics Review, 106, 620-630.

Jaynes, E. T., 1968. On the rationale of maximum-entropy methods. Proceedings of IEEE, 70, 939-952.

Johnson, J. W., 2000. A heuristic method for estimating the relative weight of predictor variables in multiple regression. Applied Behavioral Research, 35, 1-19.

Jones, R. M., Miller, K. S., 1966. On the multivariate log-normal distribution. Journal of Industrial Mathematics, 16, 63-76.

Kass, R. E., Raftery., A. E., 1995. Bayes factors, Journal of the American Statistical Association, 90, 773-795.

Kruskal, W., 1984. Concepts of relative importance. Qüestiiò, 8, 39-45.

Kruskal, W., 1987. Relative importance by averaging over orderings. The American Statistician, 41, 6-1.

Kruskal, W., Majors, R., 1989. Concepts of relative importance in scientific literature. The American Statistician, 43, 2-6.

Lindeman, R. H., Merenda, P. F., Gold, R. Z., 1980. Introduction to Bivariate and Multivariate Analysis. Glenview, IL: Scott, Foresman, and Company.

Lindley, D., 1956. On a measure of information provided by an experiment. Ann. Math. Statist. 27, 986-1005.

Lipovetsky, H., Conklin, M., 2001. Analysis of regression in game theory approach. Applied Stochastic Models in Business and Industry, 17, 319-330.

Nadarajah, S., Zografos, K., 2005. Expressions for Rényi and Shannon entropies for bivariate distributions. Information Sciences 170, 173-189.

Pourahmadi, M., Soofi, E. S., 2000. Predictive variance and information worth of observations in time series. Journal of Time Series Analysis, 21, 413-434.

Pratt, J. W., 1990. Measuring relative variable importance. ASA Proceedings of the Business and Economic Statistics Section, American Statistical Association.

Press, S.J., Zellner, A., 1978. Posterior distribution for the multiple correlation coefficient with fixed regressors. Journal of Econometrics, 8, 307-321.

Schemper, M., 1993. The relative importance of prognostic factors in studies of survival. Statistics in Medicine, 12, 2377-2382.

Soofi, E. S., 1992. A generalizable formulation of conditional logit with diagnostics. Journal of the American Statistical Association, 87, 812-816.

Soofi, E. S., 1994. Capturing the intangible concept of information. Journal of the American Statistical Association, 89, 1243-1254.

Soofi, E. S., Retzer, J. J., 2002. Information indices: unification and applications. Journal of Econometrics, 107, 17-40.

Soofi, E. S., Retzer, J. J., Yasai-Ardekani, M., 2000. A framework for measuring the importance of variables with applications to management research and decision models. Decision Sciences, 31, 595-625.

Spiegelhalter, D. J., Best, N. G, Carlin, B. P., van der Linde, A., 2002. Bayesian measures of model complexity and fit, Journal of the Royal Statistical Society, Series B, 64, 1-34.

Theil, H., Chung, C., 1988. Information-theoretic measures of fit for univariate and multivariate linear regressions. The American Statistician, 42, 249-252.

Zellner, A., 1997. Bayesian Analysis in Econometrics and Statistics: The Zellner View and Papers. Cheltenham UK: Edward Elgar.

Zellner, A. (1971). An Introduction to Bayesian Inference in Econometrics. Wiley, New York (reprinted in 1996 by Wiley).