I^2SDS
The Institute for Integrating Statistics in Decision Sciences

Technical Report TR-2009-5
April 22, 2009

On the Sample Information about Parameter and Prediction

Nader Ebrahimi
Division of Statistics
Northern Illinois University

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

Refik Soyer
Department of Decision Sciences
The George Washington University

On the Sample Information about Parameter and Prediction

Nader Ebrahimi
Division of Statistics
Northern Illinois University
DeKalb, IL 60155
nader@math.niu.edu

Ehsan S. Soofi
Sheldon B. Lubar School of Business
University of Wisconsin-Milwaukee
P.O.Box 742, Milwaukee, WI 53201
esoofi@uwm.edu

Refik Soyer
Department of Decision Sciences
George Washington University
Washington D.C. 20052
soyer@gwu.edu

May 19, 2010

Abstract

The Bayesian measure of sample information about the parameter, known as Lindley's measure, is widely used in various problems such as developing prior distributions, models for the likelihood function, and optimal designs. The predictive information is defined similarly and used for model selection and optimal designs though to a lesser extent. The parameter and predictive information measures are proper utility functions and have been also used in combination. Yet the relationship between the two measures and the effects of conditional dependence between the observable quantities on the Bayesian information measures remain unexplored. We address both issues. The relationship between the two information measures is explored through the information provided by the sample about the parameter and prediction jointly. The role of dependence is explored along with the interplay between the information measures, prior, and sampling design. For the conditionally independent sequence of observable quantities, decompositions of the joint information characterize Lindley's measure as the sample information about the parameter and prediction jointly and the predictive information as part of it. For the conditionally dependent case, the joint information about parameter and prediction exceeds Lindley's measure by an amount due to the dependence. More specific results are shown for the normal linear models and a broad subfamily of the exponential family. Conditionally independent samples provide relatively little information for prediction, and the gap between the parameter and predictive information measures grows rapidly with the sample size. Three dependence structures are studied: the intraclass (IC) and serially correlated (SC) normal models, and order statistics. For IC and SC models, the information about the mean parameter decreases and the predictive information increases with the correlation, but the joint information is not monotone and has a unique minimum. Compensation of the loss of parameter information due to dependence requires larger samples. For the order statistics, the joint information exceeds Lindley's measure by an amount which does not depend on the prior or the model for the data, but it is not monotone in the sample size and has a unique maximum.

Key Words: Bayesian predictive distribution; entropy; mutual information; optimal design; reference prior; intraclass correlation; serial correlation; order statistics.

1 Introduction

The elements of Bayesian information analysis are a set of n observations, denoted as an $n \times 1$ vector $\boldsymbol{y}$ generated from a sequence of random variables $Y_1, Y_2, \cdots$ with a joint probability model $f(\boldsymbol{y}|\theta)$ where the parameter θ has a prior probability distribution $f(\theta),\ \theta \in \Theta$, and a new outcome Y_ν. We follow the convention of using upper case letters for unknown quantities, which may be scalar or vector. Whereas the concept of prediction is usually an after thought in classical statistics, unless one deals with regression or forecasting type models, predictive inference naturally arises as a

consequence of calculus of probability and is a standard output of Bayesian analysis. Bayesians are interested in prediction of future outcomes, because eventually they will be observed which allow to settle bets in the sense of de Finetti. The predictive inference is considered as a distinguishing feature of Bayesian approach. But one can not develop predictive inference without estimation, that is, without obtaining the posterior distribution of the parameter. The parameter plays the pivotal role in prediction, and a clear perspective of the information provided by the sample about the parameter and prediction can be obtained only through viewing (Θ, Y_ν) jointly.

Information provided by the data refers to a measure that quantifies changes from a prior to a posterior distribution of an unknown quantity. Lindley (1956) framed the problem of measuring sample information about the parameter in terms of Shannon's (1948) notion of information in the noisy channel (sample) about the signal transmitted from a source (parameter). The notion is operationalized in terms of entropy and mutual information measures. Bernardo (1979a) showed that Lindley's measure of information about the parameter is the expected value of a logarithmic utility function for the decision problem of reporting a probability distribution from the space of all distributions. The information utility function belongs to a large class of utility functions discussed by Good (1971) and others which lead to the posterior distribution given by the Bayes rule as the optimal distribution. The predictive version of Lindley's measure, referred to as predictive information, quantifies the expected amount of information provided by the sample about prediction of a new outcome.

A list of articles on Lindley's measure and its methodological applications is tabulated in the Appendix. The major areas of applications is classified in terms of developing model for the likelihood function and design, and developing prior and posterior distributions. Stone (1959) was first to apply Lindley's measure to design of experiments and El-Sayyed (1969) was first to apply Lindley's measure to the exponential model. Following Bernardo (1979a,b), several authors have presented evaluation and selection of the likelihood function in terms of Lindley's measure as a Bayesian decision problem. Chaloner and Verdinelli (1995) provide an extensive review and additional references for the experimental design; see also Barlow and Hsiung (1983) and Polson (1993). Soofi (1988, 1990), Ebrahimi and Soofi (1990) examined the trade-offs between the prior and design parameters for the information about the model parameter. Carota, et al. (1996) developed an approximation for application to model elaboration. Yuan and Clarke (1999) proposed developing model for the likelihood function that maximizes Lindley's measure subject to a constraint in terms of the Bayes risk of the model. San Martini and Spezzaferri (1984) used a version of the predictive

information for model selection. Amaral and Dunsmore (1985) studied the predictive measure and applied it to the exponential parameter. Verdinelli et al. (1993) proposed use of predictive information for optimal design. Verdinelli (1992) considered a linear combination of the parameter and predictive information measures.

We explore the relationship between the parameter and predictive information measures and examine the roles of prior, design, and the dependence in the sequence $Y_i|\theta,\ i = 1, 2, \cdots$ on the informations measures and their interrelationship. This expedition integrates and expands the existing literature in three directions.

First, to this date, the relationship between the sample information about the parameter (Lindley's measure) and predictive information remains unexplored. Lindley's measure focuses on the information flow between the pair $(\boldsymbol{Y}, \Theta)$. The predictive information measure is based on the information flow between the pair $(\boldsymbol{Y}, Y_\nu)$. The key to exploring the relationship between the information provided by the sample about the parameter and for the prediction is through viewing (Θ, Y_ν) jointly as an interrelated pair. In this perspective, Θ plays an intermediary role in the information flow from the data $\boldsymbol{y}$ to the prediction quantity Y_ν. The information flow from $\boldsymbol{Y}$ to the pair $(\boldsymbol{Y}, Y_\nu)$ is different when $Y_i|\theta,\ i = 1, 2, \cdots$ are conditionally independent and conditionally dependent. Panel a) of Figure 1 depicts the conditionally independent model where the parameter θ is the only link between $\boldsymbol{Y}$ and Y_ν. As shown in the information flow diagram in Panel a) of Figure 1, in this case, the information flow from the data to the predictive distribution is solely through the parameter. This information flow from $\boldsymbol{Y}$ to Θ to Y_ν is analogous to the data processing of the information theory (Cover and Thomas 1991) where $(\boldsymbol{Y}, \Theta, Y_\nu)$ is a Markovian triplet. We will show that in this case the sample information about the parameter is in fact the entire information provided by $\boldsymbol{Y}$ about (Θ, Y_ν) jointly, and that the predictive information is only a part of it. We will further show that for some important classes of models, such as the normal linear model and a large family of lifetime models, the predictive information provided by the conditional independence sample is only a small fraction of the parameter (joint) information.

Second, thus far, the effects of dependence in the sequence $Y_i|\theta,\ i = 1, 2, \cdots$ on the Bayesian information measures remain unexplored. Panel b) of Figure 1 shows the graphical representations of the conditionally dependent model and its information flow diagram. As shown in the information flow diagram, in this case, the sample information flows from the data to predictive distribution directly due to the conditional dependence, as well as indirectly via the parameter. Consequently, the relationship between the parameter and predictive information measures is quite different than

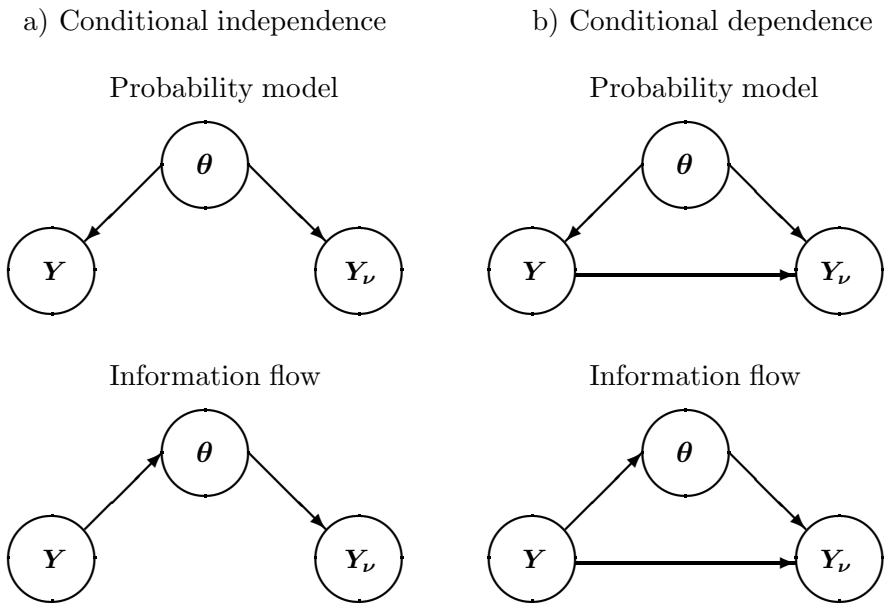

Figure 1: Graphics of conditional independent and unconditional dependent models

that for the conditionally independent case. We will show that in this case, the sample information for parameter and prediction decomposes into the parameter information (Lindley's measure) and an information measure mapping the conditional dependence. We study the role of dependence for three important models: the intraclass (IC) and serial correlation (SC) dependence structures for the normal sample, and order statistics where no particular distribution for the likelihood and prior is specified. Estimation of the normal mean and prediction under the IC and SC models are commonplace. We examine the effects of dependence on the parameter and predictive information measures drawing from Pourahmadi and Soofi (2000) who studied information measures for prediction of future outcomes in the time series context. We will show that the sample can provide substantial amount of information for prediction and the dominance of parameter information that was noted for the conditionally independent case no longer holds. Order statistics, which conditional on the parameter, form a Markovian sequence (Arnold, et al. 1992), also provide a useful context for studying the effects of dependence on information measures. The information that the first r failure times provide about the parameter as well as about the time to next failure Y_{r+1} are of interest in life testing. Here, n items are under the test and failures are observed one at a time, determining at an early stage how costly the testing is going to be and whether an action such as a redesign is warranted. Such joint parameter-predictive inferences are considered by Lawless (1971), Kaminsky and Rhodin (1985), and Ebrahimi (1992) under various sampling plans.

Third, the Bayesian information research has focused either on the design or on the prior. The past research has mainly used two types of models encompassing two different parameters: the

linear model for the normal mean parameter, and lifetime model where the scale parameter of an exponential family distribution is of interest. We consider the normal linear model with normal prior distribution for the mean and a subfamily of the exponential family under the gamma prior distribution for the scale parameter. This subfamily includes, the exponential distribution and many of parametric families such as Weibull, Pareto, and Gumbel extreme values. For each class of models, we examine the relationships between the parameter and predictive information measures. Furthermore, we explore the effects of sampling plan and prior distribution on the parameter and predictive information measures. We will show that under the optimal design for the parameter estimation, the loss of information for prediction is not nearly as severe as the loss of information about the parameter under the optimal design for prediction.

This paper is organized as follows. Section 2 presents the measures of information provided by the sample about the parameter and prediction, including results on the relationship between them for the conditionally independent model. Section 3 explores the measures of information provided by the sample about the parameter and prediction in terms of the prior and design matrix for linear models. Section 4 explores the measures of information provided by the sample about the parameter and prediction for a subfamily of the exponential family and explores the interplay between parameter and predictive information for a broad family of distributions generated by transformations of the generalized gamma family. Section 5 examines information measures for conditionally dependent samples. Section 6 gives the concluding remarks. An Appendix provides a classification of the literature on the Bayesian applications of mutual information and some technical details.

2 Information Measures

Let Q represent the unknown quantity of interest: Θ, Y_ν, individually or as a pair, or a function of them. For notational convenience we represent probability distribution with its density function $f(\cdot)$ and use subscript i for the elements of data vector $\boldsymbol{y}$ and Y_ν, $\nu \neq i$ for prediction. Information provided by the data $\boldsymbol{y}$ about Q is measured by a function that maps changes between a prior distribution $f(q)$ and the posterior distribution $f(q|\boldsymbol{y})$ obtained via the Bayes rule. Two measures of changes of the prior and posterior distributions are as follows. The uncertainty about Q is measured by the Shannon entropy

$$H(Q) = H(f) = -\int f(q) \log f(q) dq,$$

and the observed sample information about Q is measured by the entropy difference

$$\Delta H(\boldsymbol{y}; Q) = H(Q) - H(Q|\boldsymbol{y}). \tag{1}$$

The information discrepancy between the prior and posterior distributions is measured by the Kullback-Leibler divergence

$$K[f(q|\boldsymbol{y}) : f(q)] = \int f(q|\boldsymbol{y}) \log \frac{f(q|\boldsymbol{y})}{f(q)} dq \geq 0, \tag{2}$$

where the equality in (2) holds if and only if $f(q|\boldsymbol{y}) = f(q)$ almost everywhere. The observed sample information measure (1) can be positive or negative depending on which of the two distributions is more concentrated (less uniform). For a k-dimensional random vector Q, an orthonormal $k \times k$ matrix A, and a $k \times 1$ vector $\boldsymbol{c}$, $H(AQ + \boldsymbol{c}) = H(Q)$ and (1) is invariant under all linear transformations of Q. The information discrepancy (2) is a relative entropy which only detects changes between the prior and the posterior, without indicating which of the two distributions is more informative. It is invariant under all one-to-one transformations of Q and $\boldsymbol{y}$.

The expected sample information measures are obtained by viewing the observed information measures (1) and (2) as functions of the data and averaging them with respect to the marginal distribution of $\boldsymbol{Y}$. The expected entropy difference and expected Kullback-Leibler provide the same measure, known as the *mutual information*

$$M(\boldsymbol{Y}; Q) = E_{\boldsymbol{y}}\{\Delta H(\boldsymbol{y}; Q)\} = E_{\boldsymbol{y}}\left\{K[f(q|\boldsymbol{y}) : f(q)]\right\}, \tag{3}$$

where $E_{\boldsymbol{y}}$ denotes averaging with respect to

$$f(\boldsymbol{y}) = \int f(\theta) f(\boldsymbol{y}|\theta) d\theta. \tag{4}$$

Other representations of $M(\boldsymbol{Y}; Q)$ are:

$$\begin{aligned} M(\boldsymbol{Y}; Q) &= H(Q) - \mathcal{H}(Q|\boldsymbol{Y}) \\ &= K[f(q, \boldsymbol{y}) : f(q) f(\boldsymbol{y})], \end{aligned} \tag{5}$$

where

$$\mathcal{H}(Q|\boldsymbol{Y}) = E_{\boldsymbol{y}}\{H(Q|\boldsymbol{y})\} = \int H(Q|\boldsymbol{y}) f(\boldsymbol{y}) d\boldsymbol{y}$$

is referred to as the *conditional entropy* in the information theory literature. The first representations in (3) and (5) are in terms of the expected uncertainty reduction, and the second representation in (5) shows that the mutual information is symmetric in Q and $\boldsymbol{Y}$. It is noteworthy to mention that the equalities in (3) and (5) do not hold, in general, for generalizations of Shannon entropy and Kullback-Leibler information divergence, such as Rényi measures.

Some useful properties of the mutual information are as follows.

1. $M(\boldsymbol{Y}; Q) \geq 0$, where the equality holds if an only if Q and $\boldsymbol{Y}$ are independent.

2. The conditional mutual information is defined by $M(\boldsymbol{Y}; Q|S) = E_s[M(\boldsymbol{Y}; Q|s)] \geq 0$, where the equality holds if an only if Q and $\boldsymbol{Y}$ are conditionally independent.

3. Given $f(q)$, $M(\boldsymbol{Y}; Q)$ is convex in $f(q|\boldsymbol{y})$ and given $f(q|\boldsymbol{y})$, $M(\boldsymbol{Y}; Q)$ is concave in $f(q)$.

4. Let $\boldsymbol{Y_n}$ denote a vector of dimension n, $Y_j \in \boldsymbol{Y_n}$ and $Y_j \notin \boldsymbol{Y_{n-1}}$. Then

$$M(\boldsymbol{Y_n}; Q) = M(\boldsymbol{Y_{n-1}}; Q) + M(Q; Y_j|\boldsymbol{Y_{n-1}}) \geq M(\boldsymbol{Y_{n-1}}; Q), \tag{6}$$

   thus $M(\boldsymbol{Y_n}; Q)$ is increasing in n.

5. $M(\boldsymbol{Y}; Q)$ is invariant under one-to-one transformations of Q and $\boldsymbol{Y}$.

2.1 Marginal information

For $Q = \Theta$, the observation $\boldsymbol{y}$ provides the likelihood function, $\mathcal{L}(\theta) \propto f(\boldsymbol{y}|\theta)$ and updates the prior to the posterior distribution

$$f(\theta|\boldsymbol{y}) \propto f(\theta) f(\boldsymbol{y}|\theta). \tag{7}$$

The expected sample information about the parameter, $M(\boldsymbol{Y}; \Theta)$ is known as Lindley's measure (Lindley 1956) and is referred to as the parameter information.

The following properties are also well-known.

1. Let $S_n = S(\boldsymbol{Y})$ be a general transformation. Then $M(\boldsymbol{Y}; \Theta) \geq M(S_n; \Theta)$, where the equality holds if and only if S_n is a sufficient statistic for θ.

2. $M(\boldsymbol{Y_n}; \Theta)$ is concave in n, which implies that $M(Y_j; \Theta|\boldsymbol{Y_{n-1}}) \leq M(Y_j; \Theta)$.

3. Ignorance between two neighboring values in the parameter space, $P(\theta) = P(\theta + \delta(\theta)) = .5$, implies that $M(\boldsymbol{Y}; \Theta) \approx 2\delta^2(\theta)\mathcal{I}_F(\theta)$ as $\delta\theta \to 0$, where $\mathcal{I}_F(\theta)$ is Fisher information (Lindley 1961, p. 467). Similar approximation holds for $M(\boldsymbol{Y}; Q)$; also see Kullback (1959).

For $Q = Y_\nu$, the prior and posterior predictive distributions, respectively, are given by

$$f(y_\nu) = \int f(y_\nu|\theta) f(\theta) d\theta$$

and

$$f(y_\nu|\boldsymbol{y}) = \int f(y_\nu|\theta) f(\theta|\boldsymbol{y}) d\theta. \tag{8}$$

The expected information $M(\boldsymbol{Y}; Y_\nu)$ is referred to as the predictive information (San Martini and Spezzaferri 1984, Amaral and Dunsmore 1985).

In some problems both, the parameter and the prediction, are of interest (Chaloner and Verdinelli 1995). Verdinelli (1992) proposed the linear combination of marginal utilities

$$U(\boldsymbol{Y}; \boldsymbol{\Theta}, Y_\nu) = w_1 M(\boldsymbol{Y}; \boldsymbol{\Theta}) + w_2 M(\boldsymbol{Y}; Y_\nu), \tag{9}$$

where $w_k \geq 0,\ k = 1, 2$ are weights that reflect the relative importance of the parameter and prediction for the experimenter. Since $\boldsymbol{\Theta}$ and Y_ν are not independent quantities, $M(\boldsymbol{Y}; \boldsymbol{\Theta})$ and $M(\boldsymbol{Y}; Y_\nu)$ are not additively separable. The weights in (9) do not take into account the dependence between the prediction and the parameter.

2.2 Joint information

Taking the dependence between the parameter and prediction into account requires considering the joint information for the vector of parameter and prediction. The observed and expected information measures are defined by (1) and (3) where $Q = (\Theta, Y_\nu)$, and will be denoted as $\Delta H[\boldsymbol{y}; (\Theta, Y_\nu)]$ and $M[\boldsymbol{Y}; (\Theta, Y_\nu)]$. Next Theorem encapsulates the relationships between the joint, parameter, and predictive information measures for the conditionally independent samples.

Theorem 1 . *If $Y_1|\theta, Y_2|\theta, \cdots$ are conditionally independent. Then:*

(a) $\Delta H(\boldsymbol{y}; \Theta) = \Delta H[\boldsymbol{y}; (\Theta, Y_\nu)]$;

(b) $M(\boldsymbol{Y}; \Theta) = M[\boldsymbol{Y}; (\Theta, Y_\nu)]$;

(c) $M(\boldsymbol{Y}; Y_\nu) \leq M(\boldsymbol{Y}; \Theta)$.

Proof. The proof of (a) is as follows. The joint entropy decomposes additively as

$$H(\Theta, Y_\nu) = H(\Theta) + \mathcal{H}(Y_\nu|\Theta),$$

where $\mathcal{H}(Y_\nu|\Theta) = E_\theta\{H(Y_\nu|\theta)\}$ is the conditional entropy. Letting $Q = (\Theta, Y_\nu)$ in (1) and applying the entropy decomposition to each entropy, we have

$$\Delta H[\boldsymbol{y}; (\Theta, Y_\nu)] = H(\Theta) + \mathcal{H}(Y_\nu|\Theta) - \{H(\Theta|\boldsymbol{y}) + \mathcal{H}(Y_\nu|\Theta, \boldsymbol{y})\},$$

where $\mathcal{H}(Y_\nu|\Theta, \boldsymbol{y}) = E_\theta\{H(Y_\nu|\theta, \boldsymbol{y})\}$. The first and third terms give $\Delta H(\boldsymbol{y}; \Theta)$. The conditional independence implies for each $\theta,\ H[f(y_\nu|\theta, \boldsymbol{y})] = H[f(y_\nu|\theta)]$, thus $E_\theta\{H(Y_\nu|\theta, \boldsymbol{y})\} = E_\theta\{H(Y_\nu|\theta)\}$,

and the second and fourth terms cancel out, which gives (a). Since $\boldsymbol{Y} \to \Theta \to Y_\nu$ is a Markovian triplet, Parts (b) and (c) are implied by properties of the mutual information functions of Markovian sequences (see, e.g., Cover and Thomas 1991, pp. 27, 32-33).

By part (a) of Theorem 1, under the conditionally independent model, the information provided by each and every sample about the parameter is the same as the joint information for the parameter and prediction.

Part (b) of Theorem 1 provides a broader interpretation of Lindley's information, namely expected information provided by the data about the parameter and for the prediction. An immediate implication is that the prior distribution (Bernardo 1979a,b), the design (Chaloner and Verdinelli 1995 and Polson 1993), and the likelihood model (Yuan and Clarke 1999) that maximize $M(\boldsymbol{Y}; \Theta)$ also maximizes sample information about the parameter and prediction jointly. However, by part (c) of Theorem 1, such optimal prior, design, and model may not be optimal according to $M(\boldsymbol{Y}; Y_\nu)$. Similarly, the optimal design of Verdinelli et al. (1993) and the optimal model of San Martini and Spezzaferri (1984) which maximize $M(\boldsymbol{Y}; Y_\nu)$ may not be optimal according to $M(\boldsymbol{Y}; \Theta)$.

The inequality in (c) is the Bayesian version of the information processing inequality of information theory, and can be referred to as the *Bayesian data processing inequality* mapping the information flow $\boldsymbol{Y} \to \Theta \to Y_\nu$ through the equations (7) and (8), as shown in Figure 1a.

By Parts b) and c) of Theorem 1 we have

$$M(\boldsymbol{Y}; \Theta) = M(\boldsymbol{Y}; Y_\nu) + M(\boldsymbol{Y}; \Theta | Y_\nu), \tag{10}$$

where $M[(\boldsymbol{Y}; \Theta)|Y_\nu] = E_{y_\nu} \{K[f(\boldsymbol{y}, \theta)|y_\nu) : f(\theta|y_\nu) f(\boldsymbol{y}|y_\nu)]\}$ is the conditional mutual information between Θ and $\boldsymbol{Y}$, given Y_ν. This measure is the link between the parameter and predictive information measures and is key for studying their relationship. Applying (10) to the utility function (9) gives the weights for the additive information measures in (10) as

$$U(\boldsymbol{Y}; \boldsymbol{\Theta}, Y_\nu) = w_1 M(\boldsymbol{Y}; \boldsymbol{\Theta}|Y_\nu) + (w_1 + w_2) M(\boldsymbol{Y}; Y_\nu). \tag{11}$$

3 Linear Models

Consider the normal linear model

$$\boldsymbol{y} = X\boldsymbol{\beta} + \boldsymbol{\epsilon}, \tag{12}$$

where $\boldsymbol{y}$ is an $n \times 1$ vector of observations, X is an $n \times p$ design matrix, $\boldsymbol{\beta}$ is the $p \times 1$ parameter vector, $\boldsymbol{\epsilon}$ is the error vector. Under the conditionally independent model $f(\boldsymbol{\epsilon}|\boldsymbol{\beta}) = N(\mathbf{0}, \sigma_1^2 I_n)$, $\sigma_1^2 > 0$ is known and I_n is identity matrix of dimension n.

It will be more insightful to use the orthonormal rotation $Z = XG$ and $\boldsymbol{\theta} = G'\boldsymbol{\beta}$, where G is the matrix of eigenvectors of $X'X$, and $\Lambda = Z'Z = \text{diag}[\lambda_1, \cdots, \lambda_p]$ where $\lambda_j > 0,\ j = 1 \cdots, p$ are the eigenvalues of $X'X$. By the invariance of entropy under orthonormal transformations $\Delta H(\boldsymbol{y}; \boldsymbol{\Theta}) = \Delta H(\boldsymbol{y}; \boldsymbol{\beta})$ and by invariance of mutual information under all one-to-one transformations, $M(\boldsymbol{Y}; \boldsymbol{\Theta}) = M(\boldsymbol{Y}; \boldsymbol{\beta})$.

We use the normal conjugate prior $f(\boldsymbol{\theta}) = N(\boldsymbol{m}_0, \sigma_0^2 V_0)$, where $V_0 = \text{diag}[v_{01}, \cdots, v_{0p}]$. The posterior distribution is $f(\boldsymbol{\theta}|\boldsymbol{y}) = N(\boldsymbol{m}_1, \sigma_1^2 V_1)$ where $\boldsymbol{m}_1 = V_1^{-1}\left(\eta V_0^{-1}\boldsymbol{m}_0 + Z'\boldsymbol{y}\right), V_1 = \left(\eta V_0^{-1} + Z'Z\right)^{-1}$ and $\eta = \dfrac{\sigma_1^2}{\sigma_0^2}$. All distributions and information measures are conditional on Z and σ_1^2 which are assumed to be given and will be suppressed. The prior and posterior entropies are $H(\boldsymbol{\Theta}|\sigma_k^2 V_k) = \dfrac{p}{2}\log(2\pi e) + \dfrac{1}{2}\log|\sigma_k^2 V_k|,\ k = 0, 1$, where $|\cdot|$ denotes the determinant. Since entropy is location invariant, $\boldsymbol{m}_k$ does not matter. Also since V_1 does not depend on data $\boldsymbol{y}$, the conditional entropy and posterior entropies are equal, $\mathcal{H}(\boldsymbol{\Theta}|\boldsymbol{Y}, Z, \eta, V_0) = H(\boldsymbol{\Theta}|\boldsymbol{y}, Z, \eta, V_0)$. Thus, the observed and expected sample information measures are the same, given by

$$\begin{aligned} M(\boldsymbol{Y}; \boldsymbol{\Theta}|Z, \eta, V_0) = \Delta H(\boldsymbol{y}; \boldsymbol{\Theta}|Z, \eta, V_0) &= \frac{1}{2}\log|I_p + \eta^{-1}V_0 Z'Z| \\ &= \frac{1}{2}\sum_{j=1}^{p}\log\left(1 + \eta^{-1}v_{0j}\lambda_j\right). \end{aligned} \tag{13}$$

From (13) it is clear that the parameter (joint) information is decreasing in η and increasing in v_{0j}, λ_j and σ_0^2. Thus, given the prior, the information can be optimized through the choices of design parameters $\lambda_j,\ j = 1, \cdots, p$ and for given data (design), the information can be optimized through the prior parameters σ_0^2 and $v_{0j},\ j = 1, \cdots, p$.

The prior and posterior predictive distributions of a future outcome Y_ν to be taken at a point $\boldsymbol{z}_\nu$ are normal $N\left(\boldsymbol{z}_\nu'\boldsymbol{\mu}_{\boldsymbol{k}},\ \sigma_k^2\boldsymbol{z}_\nu' V_k \boldsymbol{z}_\nu + \sigma_1^2\right),\ k = 0, 1$, and

$$M(\boldsymbol{Y}; Y_\nu|\boldsymbol{z}_\nu, Z, \eta, V_0) = \Delta H(\boldsymbol{y}; Y_\nu|\boldsymbol{z}_\nu, Z, \eta, V_0) = \frac{1}{2}\log\left(\frac{\eta^{-1}\boldsymbol{z}_\nu' V_0 \boldsymbol{z}_\nu + 1}{\boldsymbol{z}_\nu' V_1 \boldsymbol{z}_\nu + 1}\right). \tag{14}$$

Parts (a) and (b) of Theorem 1 give $\Delta H[\boldsymbol{y}; (\boldsymbol{\Theta}, Y_\nu)|\boldsymbol{z}_\nu, Z, \eta, V_0] = \Delta H(\boldsymbol{y}; \boldsymbol{\Theta}|Z, \eta, V_0)$, and $M[\boldsymbol{Y}; (\boldsymbol{\Theta}, Y_\nu)|\boldsymbol{z}_\nu, Z, \eta, V_0] = M(\boldsymbol{Y}; \boldsymbol{\Theta}|Z, \eta, V_0)$. Therefore all existing results for $M(\boldsymbol{Y}; \boldsymbol{\Theta}|Z, \eta, V_0)$ apply to the joint parameter and predictive information, as well. Part (c) of Theorem 1 provides an additional insight: $M(\boldsymbol{Y}; Y_\nu|\boldsymbol{z}_\nu, Z, \eta, V_0) \leq M(\boldsymbol{Y}; \boldsymbol{\Theta}|Z, \eta, V_0)$. These relationships hold for multiple predictions, as well.

3.1 Optimal Designs

Several authors have studied parameter information in the context of experimental design. It is clear from (13) that given $V_0 = I_p$ and the trace $\mathrm{Tr}(Z'Z) = \sum_{j=1}^{p} \lambda_j$, the optimal parameter information design is obtained when all eigenvalues are equal, $\lambda_j = \bar{\lambda} = \frac{1}{p}\sum_{k=1}^{p} \lambda_k$, which gives the Bayesian D-optimal design (see Chaloner and Verdinelli (1995) for references). That is, with the uncorrelated prior the information optimal design is orthogonal. For the case of weak prior information, $\sigma_0^2 \to \infty$, maximizing the expected parameter information gain is equivalent to the classical criterion of $D-$optimality. If the experimental information is weak then the Bayesian criterion reduces to classical criterion of $A-$optimality when $V_0 = I_p$ (Polson, 1993). Verdinelli et al. (1993) used the predictive information optimal design for accelerated life testing.

For illustrating implications of Theorem 1 for design we consider the simple case when $x_{ij} \in \{0, 1\}$. This is a one-way ANOVA structure, when the averages (parameters) as well as contrasts between the individual outcomes are of interest. In this case, $\mathrm{Tr}(\Lambda) = \sum_{j=1}^{p} n_j = n$ and the design parameters are $\lambda_j = n_j$. The following proposition gives the optimal designs according to the parameter (joint) information $M(\boldsymbol{Y}; \boldsymbol{\Theta})$ and predictive information $M(\boldsymbol{Y}; Y_\nu)$.

Proposition 1 *Given V_0 and $\sum_{j=1}^{p} n_j = n$.*

(a) The optimal sample allocation scheme according to the parameter (joint) information $M(\boldsymbol{Y}; \boldsymbol{\Theta})$ is

$$\begin{cases} n_1^* = \dfrac{n}{p} - \dfrac{\eta}{p}\displaystyle\sum_{j=2}^{p}\left(v_{0j}^{-1} - v_{01}^{-1}\right) \\ \\ n_j^* = n_1^* - \eta\left(v_{0j}^{-1} - v_{01}^{-1}\right), \quad j = 2, \cdots, p. \end{cases} \tag{15}$$

and the minimum sample size is determined by $n_1^* > \max\left\{\left(v_{0j}^{-1} - v_{01}^{-1}\right)\eta,\ j = 2, \cdots, p\right\}$.

(b) The information optimal sample allocation scheme according to the predictive information $M(\boldsymbol{Y}; Y_\nu)$ for prediction at $\boldsymbol{z}_\nu$ is

$$\begin{cases} n_1^* = \dfrac{|z_{\nu 1}|n}{\sum_{j=1}^{p}|z_{\nu j}|} + \dfrac{\eta}{\sum_{j=1}^{p}|z_{\nu j}|}\displaystyle\sum_{j=2}^{p}\left(v_{0j}^{-1} - v_{01}^{-1}\right) \\ \\ n_j^* = \dfrac{|z_{\nu j}|}{|z_{\nu 1}|}n_1^* - \dfrac{\eta}{|z_{\nu 1}|}\left(|z_{\nu 1}|v_{0j}^{-1} - |z_{\nu j}|v_{01}^{-1}\right), \quad j = 2, \cdots, p. \end{cases} \tag{16}$$

and the minimum sample size is determined by $n_1^* > \max\left\{\dfrac{\eta}{|z_{\nu j}|}\left(|z_{\nu 1}|v_{0j}^{-1} - |z_{\nu j}|v_{01}^{-1}\right),\ j = 2, \cdots, p\right\}$.

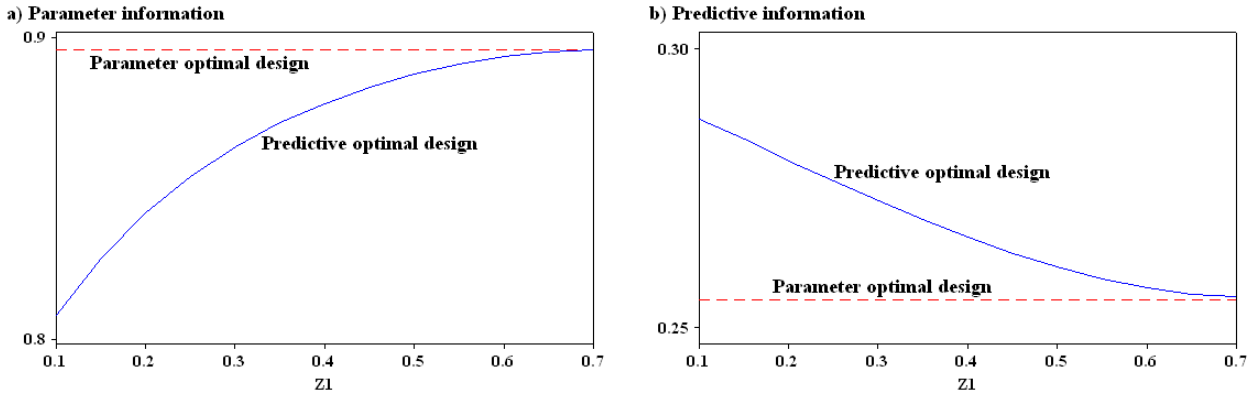

Figure 2: Parameter information per dimension $M(\boldsymbol{Y};\boldsymbol{\Theta}|Z,\eta)/p$ and predictive information $M(\boldsymbol{Y};Y_\nu|Z,\boldsymbol{z},\eta)$ under the predictive and parameter optimal designs against $z_{\nu 1},\ \boldsymbol{z}_\nu'\boldsymbol{z}_\nu = 1$ for $p = 2,\ n = 10, \eta = 1, v_{01} = v_{02} = 1$.

Proof. See Appendix.

Note that by Theorem 1, the maximum predictive information attained with optimal design (16) is dominated by the parameter information:

$$M(\boldsymbol{Y};\boldsymbol{\Theta}|n_i^*,\boldsymbol{z}_\nu) = M[\boldsymbol{Y};(\boldsymbol{\Theta},Y_\nu)|n_i^*,\boldsymbol{z}_\nu] \geq M(\boldsymbol{Y};Y_\nu|n_i^*,\boldsymbol{z}_\nu).$$

Figure 2a shows the plots of the parameter information measures under the parameter and predictive optimal designs against $z_{\nu 1},\ \boldsymbol{z}_\nu'\boldsymbol{z}_\nu = 1$ for $p = 2, n = 10, v_{01} = v_{02} = 1$, and $\eta = 1$. In order to make the two information measures dimensionally comparable, we have plotted information per parameter $\bar{M}(\boldsymbol{Y};\Theta|n_i^*) = M(\boldsymbol{Y};\Theta|n_i^*)/p$. Figure 2b shows the plots of the predictive information measures under the parameter and predictive optimal designs. Note that the vertical axes of the two panels are different. These plots show that the joint information per dimension is much higher than the predictive information even though the design is optimal for prediction and not for the parameter. The dashed lines show the information quantities for the D-optimal design, which is optimal for the parameter (joint) and for prediction at the diagonal $z_1 = z_2 = 1/\sqrt{2} \approx .707$. The sample is least informative for prediction in this direction. We note that the loss of information for prediction is not nearly as severe as the loss of information about the parameter. This is due to the fact that by Theorem 1, the parameter information measures the joint information about the parameter and prediction and inclusive of the predictive information. Thus, use of D-optimal design would be preferable if the experimenter has interest in inference about the parameter as well as about a prediction.

3.2 Optimal Prior Variance

Next we illustrate application to developing prior as the Bayesian solution to the collinearity problem. When the regression matrix X is ill-conditioned, posterior inference about individual parameters are unreliable. The effects of collinearity on the posterior distribution and compensating for the collinearity effects by using $V_0 = I_p$ are discussed in Soofi (1990). In the orthogonal prior variance case $\sum_{j=1}^{p} v_{0j} = p$ is distributed uniformly among the components of V_0. The following proposition gives an optimal prior variance allocation according the parameter (joint) information $M(\boldsymbol{Y}; \boldsymbol{\Theta})$ that will be useful when $X'X$ is nearly singular.

Proposition 2 *Let* $\lambda_1 \geq \cdots, \geq \lambda_p$, $\sum_{j=1}^{p} \lambda_j = p$, *and given* $\sum_{j=1}^{p} v_{0j} = c$. *The optimal prior variance allocation according the parameter (joint) information* $M(\boldsymbol{Y}; \boldsymbol{\Theta})$ *is*

$$\begin{cases} v_{01}^* = \dfrac{c}{p} + \dfrac{\eta}{p} \displaystyle\sum_{j=2}^{p} \left(\lambda_j^{-1} - \lambda_1^{-1}\right) \\ \\ v_{0j}^* = v_{01}^* - \eta \left(\lambda_j^{-1} - \lambda_1^{-1}\right), \quad j = 2, \cdots, p. \end{cases} \tag{17}$$

and the minimum prior variance is determined by $v_{01}^* > \left(\lambda_p^{-1} - \lambda_1^{-1}\right) \eta$.

Proof. See Appendix.

The optimal information prior (17) allocates prior variances to the components $\theta_j, j = 1, \cdots, p$ based on the eigenvalues $\lambda_1 \geq \cdots, \geq \lambda_p$ of $X'X$. So it is in the same spirit as Zellner's g prior where $v_{0j} \propto \lambda_j^{-1}, j = 1, \cdots, p$. In the same spirit, West (2003) and Maruyama and George (2010) have defined generalized g priors that are applicable when X is singular. Our information optimal allocation scheme is another generalization of the g prior tailored for the collinearity problem where X is full-rank, but nearly singular.

The optimal allocation scheme (17) can be represented in terms of the condition indices $\kappa_j = \sqrt{\lambda_1/\lambda_j}, j = 1, \cdots, p$ of $X'X$ as

$$\begin{cases} \dfrac{\lambda_1 v_{01}^* + \eta}{\lambda_j v_{0j}^* + \eta} = \kappa_j^2, \quad j = 2, \cdots, p, \\ \\ \displaystyle\sum_{j=1}^{p} v_{0j}^* = c. \end{cases}$$

The largest portion of the total prior variance v_{0p}^* is allocated to the component θ_p that corresponds to smallest eigenvalue λ_p such that $\dfrac{\lambda_1 v_{01}^* + \eta}{\lambda_p v_{0p}^* + \eta} = \kappa^2$, where $\kappa = \kappa(X'X) = \sqrt{\lambda_1/\lambda_p}$ is the *condition number* of $X'X$ which is used for collinearity diagnostics (Stewart 1987, Soofi, 1990, Belsley 1991).

In some prediction problems, the prediction point $\boldsymbol{z}_\nu$ is given. For example, in the accelerated life testing, $\boldsymbol{z}_\nu$ is the environmental condition and the experiment must be designed such that prediction at $\boldsymbol{z}_\nu$ is optimal. The information decomposition (13) provides the clue when the quantity of interest is the mean response $Q = E(Y|\boldsymbol{z}_\nu)$. The components of $\boldsymbol{\theta} = (\theta_1, \cdots, \theta_p)'$ are independent, *a'priori* and *a'posteriori*, and from (13), $M(\theta_j, \boldsymbol{Y}|Z, \eta, V_0) = .5 \log(1 + \eta^{-1} v_{0j} \lambda_j)$. Under the orthogonal prior, the sample is most informative about the linear combination of the regression coefficients $\theta_1 = \boldsymbol{g}_1' \boldsymbol{\beta}$ where g_1 is the first eigenvector of $X'X$. Thus the optimal design for the expected response at a covariate vector $\boldsymbol{z}_\nu$ is X^* such that $\boldsymbol{z}_\nu$ is the first eigenvector of $I_p + \eta^{-1} V_0 X^{*\prime} X^*$. Under the uncorrelated prior or weak prior, X^* is frequentist *E-optimal* design, which can be different than the designs that are optimal with respect to parameter (joint) information. The optimal allocation scheme (17) provides improvement to the orthogonal prior for prediction of the expected response when $\boldsymbol{z}_\nu$ is in the space of the eigenvectors corresponding to the large eigenvalues.

Figure 3 compares information measures for the optimal scheme, the orthogonal prior, and $V_0 \propto \Lambda^{-1}$ which is used in some priors such as the g-prior with $p = 2,\ c = 100, \eta = 1$. Figure 3a shows the plots of parameter information $M(\boldsymbol{Y}; \boldsymbol{\Theta}|Z, \eta)$ against the condition number $\kappa = \sqrt{\lambda_1/\lambda_2}$ of $X'X$. Under all three priors, the parameter information $M(\boldsymbol{Y}; \boldsymbol{\Theta}|Z, \eta)$ decrease with κ, i.e., as the regression matrix descents toward singularity. The parameter information under the optimal scheme slightly dominates the measure under the orthogonal prior, and both dominate the information under the g-prior which deteriorates quickly with collinearity. By Theorem 1, the parameter information measure is the joint information about the parameter and prediction and is inclusive of the predictive information. Figure 3b shows $M(\boldsymbol{Y}; \theta_1|Z, \eta)$ for the direction of the first eigenvector $\theta_1 = G_1' \boldsymbol{\beta}$, i.e., the most informative direction for prediction of the expected response. The optimal and orthogonal priors improve the information under collinearity, but the measure for the g-prior deteriorates quickly.

4 Exponential Family

Consider distributions in the exponential family that provide likelihood functions in the form of

$$\mathcal{L}(\theta) \propto \theta^n e^{-\theta s_n}, \quad \theta > 0, \tag{18}$$

where s_n is a sufficient statistic for θ. An important class of models whose likelihood functions are in the form of (18) is the time-transformed exponential (TTE) (Barlow and Hsiung 1983). The TTE models are usually defined in terms of the survival function $\bar{F}(y|\theta) = \exp\{-\theta\phi(y)\},\ y \geq 0$

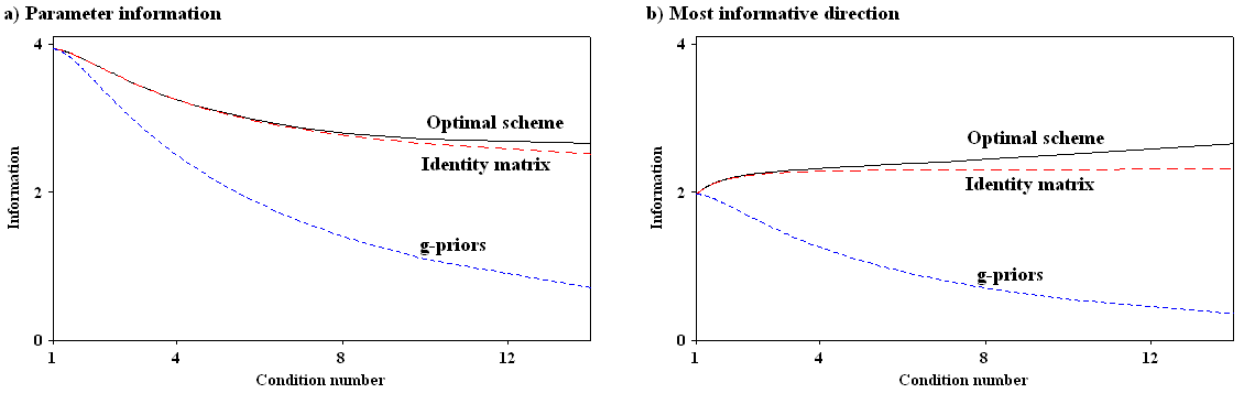

Figure 3: Parameter information $M(\boldsymbol{Y};\boldsymbol{\Theta}|Z,\eta)$ and information for the most informative direction for prediction of the expected response $M(\boldsymbol{Y};\theta_1|Z,\eta)$ for three types of prior variance allocations $(p=2,\ c=100,\eta=1)$.

where $\phi(y) = -\log \bar{F}_0$ and θ is the "proportional hazard". The density functions of the TTE models are in the form of

$$f\left(\phi(y)|\theta\right) = \theta\phi'(y)e^{-\theta\phi(y)}, \tag{19}$$

where $\phi(y)$ is a one-to-one transformation of Y with the exponential distribution $f(y|\theta) = \theta e^{-\theta y}$. For TTE models $s_n = \sum_{i=1}^{n} \phi(y_i)$. Examples include the exponential $\phi(y) = y,\ y \geq 0$, Weibull $\phi(y) = y^q,\ y \geq 0$, Pareto Type I $\phi(y) = \log(y/a),\ y \geq a > 0$, Pareto Type II $\phi(y) = \log(1+y),\ y \geq 0$, Pareto Type VI $\phi(y) = \log(1+y^a),\ y \geq 0,\ a > 0$, and the extreme value $\phi(y) = e^y$.

The family of conjugate priors for (18) is gamma $\mathcal{G}(\alpha,\beta)$ with density function

$$f(\theta) = \frac{\beta^\alpha}{\Gamma(\alpha)}\theta^{\alpha-1}e^{-\beta\theta}. \tag{20}$$

The posterior distribution is $\mathcal{G}(\alpha+n, \beta+s_n)$.

The information in the observed sample is given by

$$\Delta H(\boldsymbol{y};\Theta) = H_{\mathcal{G}}(\alpha) - H_{\mathcal{G}}(\alpha+n) + \log\left(1+\frac{s_n}{\beta}\right),$$

where $H_{\mathcal{G}}(\alpha)$ is the entropy of $\mathcal{G}(\alpha,1)$ given by

$$H_{\mathcal{G}}(\alpha) = \log\Gamma(\alpha) - (\alpha-1)\psi(\alpha) + \alpha,$$

and $\psi(\alpha) = \dfrac{d\log\Gamma(\alpha)}{d\alpha}$ is digamma function.

For the TTE family (19), the marginal distribution of s_n is inverted beta (beta prime) distribution with density

$$f(s_n) = \frac{1/\beta}{B(\alpha,n)}\frac{(s_n/\beta)^{n-1}}{(1+s_n/\beta)^{\alpha+n}}, \quad s_n \geq 0,$$

where $B(\alpha, n)$ is the beta function. Using $E_{s_n}\left\{\log\left(1+\frac{s_n}{\beta}\right)\right\} = \psi(\alpha+n)-\psi(\alpha)$, the expected information for all models with likelihood functions in form of (18) is:

$$M[\boldsymbol{Y};(\Theta;Y_\nu)] = M(\boldsymbol{Y};\Theta) = H_{\mathcal{G}}(\alpha) - H_{\mathcal{G}}(\alpha+n) + \psi(\alpha+n) - \psi(\alpha). \tag{21}$$

An interesting property of (21) is the following recursion:

$$M(\boldsymbol{Y_n};\Theta|\alpha) = M(\boldsymbol{Y_{n-1}};\Theta|\alpha) + K_{\mathcal{G}}(\alpha+n-1), \tag{22}$$

where $\boldsymbol{Y_n}$ and $\boldsymbol{Y_{n-1}}$ are vectors of dimensions n and $n-1$, and

$$K_{\mathcal{G}}(v) = K(\mathcal{G}_v : \mathcal{G}_{v+1}) = \frac{1}{v} + \psi(v) - \log v \tag{23}$$

is the Kullback-Leibler information between $\mathcal{G}_v = \mathcal{G}(\nu,\beta)$ and $\mathcal{G}_{v+1} = \mathcal{G}(\nu+1,\beta)$. The recursion (22) is found using $\psi(\alpha+1) = \psi(\alpha) + \frac{1}{\alpha}$. By (6), $M(\boldsymbol{Y_n};\Theta|\boldsymbol{Y_{n-1}}) = K_{\mathcal{G}}(\alpha+n-1)$. That is, on average, the incremental contribution of an additional observation is equivalent to the information divergence due to one unit increase of the prior variance.

The prior predictive distribution for TTE family (19) is Pareto $\mathcal{P}(\alpha,\beta)$ with density function

$$f(y_\nu) = \frac{\alpha\beta^\alpha}{(\beta+y_\nu)^{\alpha+1}},, \quad y_\nu \geq 0.$$

The posterior predictive distribution $f(y_\nu|\boldsymbol{y})$ is also Pareto with the updated parameters $\mathcal{P}(\alpha+n, \beta+s_n)$. The predictive information measures are given by

$$\begin{aligned} \Delta H(\boldsymbol{y};Y_\nu) &= H_{\mathcal{P}}(\alpha) - H_{\mathcal{P}}(\alpha+n) - \log\left(1+\frac{s_n}{\beta}\right) \\ M(\boldsymbol{Y};Y_\nu) &= H_{\mathcal{P}}(\alpha) - H_{\mathcal{P}}(\alpha+n) - \psi(\alpha+n) + \psi(\alpha), \end{aligned} \tag{24}$$

where $H_{\mathcal{P}}(\alpha) = \frac{1}{\alpha} - \log\alpha + 1$ is the entropy of $\mathcal{P}(\alpha,1)$.

By Theorem 1, $\Delta H[s_n;(\Theta,Y_\nu)] = \Delta H(\boldsymbol{y};\Theta)$, $M[\boldsymbol{Y};(\Theta,Y_\nu)] = M(\boldsymbol{Y};\Theta)$, and $M(\boldsymbol{Y};Y_\nu) \leq M(\boldsymbol{Y};\Theta)$. The following Theorem gives more specific pattern of relationships.

Theorem 2 . *The following results hold for the TTE family (19) and gamma prior (20).*

(a) $M(\boldsymbol{Y};\Theta|\alpha)$ *and* $M(\boldsymbol{Y};Y_\nu|\alpha)$ *are decreasing functions of* α*, increasing functions of* n*, and as* $n\to\infty$*,* $M(\boldsymbol{Y_{n+1}};\Theta|\alpha) - M(\boldsymbol{Y_n};\Theta|\alpha) \to 0$ *and* $M(\boldsymbol{Y};Y_\nu|\alpha) \to K_{\mathcal{G}}(\alpha)$*.*

(b) $M(\boldsymbol{Y};\Theta|\alpha) = M(\boldsymbol{Y};Y_\nu|\alpha) + M(\boldsymbol{Y};\Theta|\alpha+1)$*, where* $M(\boldsymbol{Y};\Theta|\alpha+1)$ *is the sample information with gamma prior* $\mathcal{G}(\alpha+1,\beta)$*.*

(c) $M(\boldsymbol{Y};\Theta|\alpha) - M(\boldsymbol{Y};Y_\nu|\alpha)$ *increases with* α *and with* n*.*

Proof. For (a), it is known that the expected parameter and predictive measures are increasing functions of n. It is shown in Ebrahimi and Soofi (1990) that for the exponential model, $M(\boldsymbol{Y};\Theta|\alpha)$ is decreasing in α. By the invariance of the mutual information the same result holds for the TTE family. The limits are found by noting that $K_{\mathcal{G}}(v) \to 0$ as $v \to \infty$. The expected predictive measure is decreasing in α is found by taking the derivative, using series expansion of the trigamma function $\psi'(u) = \sum_{k=1}^{\infty} \frac{1}{(u+k)^2}$ (Abramowitz and Stegun 1970), and an induction on n that shows the derivative is negative. Parts (b) is found using recursion $\psi(\alpha+1) = \psi(\alpha) + \frac{1}{\alpha}$. Part (c) is implied by (a) and (b). The difference is $M(\boldsymbol{Y};\Theta|\alpha+1)$ which is increasing (decreasing) in n (α).

By part a) of Theorem 2, the parameter and predictive information both increase with n. Part b) of Theorem 2 gives the relationship between the parameter (joint) and the predictive information measures. Part c) indicates that under conditional independence, the parameter (joint) information grows faster than the predictive information with the sample size.

As an application, consider Type II censoring where observing the number of failure is a design parameter. Let $y_1 \le y_2 \le \cdots \le y_n$ be the order statistics of a sample of size n. Then, for example, for the exponential likelihood (18) the sufficient statistic for θ is

$$t_r = y_1 + \cdots + y_{r-1} + (n-r+1)y_r, \quad r \le n.$$

The parameter information $M(T_r;\Theta|n)$ is given by (21). Ebrahimi and Soofi (1990) examined the loss of information about the parameter. The predictive information is given by (24) with $n = r$. By part a) of Theorem 2, censoring also results in loss of predictive information. As in the case of parameter information, the loss of predictive information can be compensated by the prior parameter α.

Figure 4 shows plots of the expected parameter and predictive information measures. Figure 4a illustrates the information decomposition part (Theorem 2 part b) for $\alpha = 1$ as function of n. The parameter information and predictive information are both increasing in n. The parameter information increases at a faster rate than the predictive information. In this case, the difference between the parameter and predictive information is $M(\boldsymbol{y};\Theta|\alpha+1)$ also shown in Figure 4a. These information measures are decreasing in α. Figure 4b shows the plots of loss of information due to Type II censoring for $n = 25$ and $\alpha = 1, 2$. We note that the predictive information loss is not as severe as the parameter information loss. As seen in the figure, the information losses can be recovered by increase in prior precision.

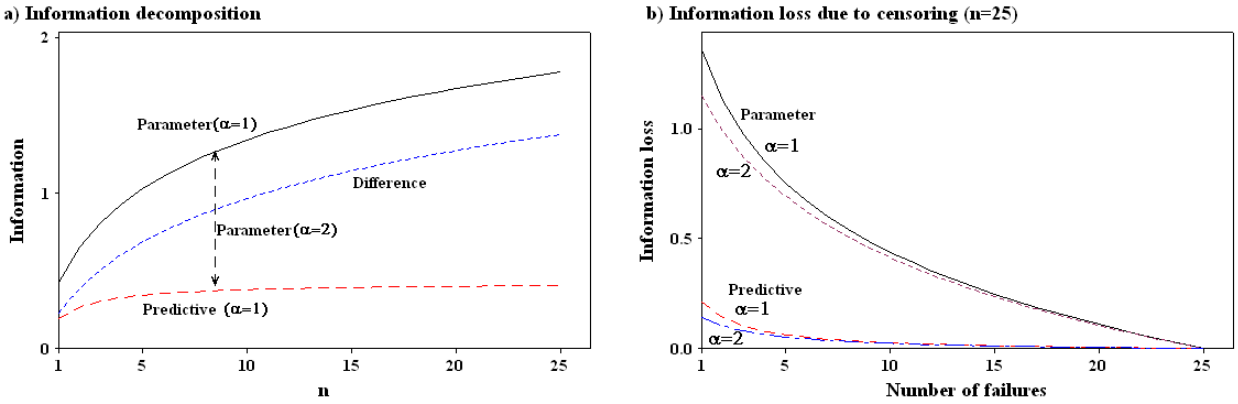

Figure 4: Decomposition of the joint (parameter) information $M(T_r;\Theta|\alpha,n)$ into predictive information $M(T_r;Y_\nu|\alpha,n)$ and $M(T_r;\Theta|\alpha+1,n)$ and loss of information $M(T_n;\Theta|\alpha)-M(T_r;\Theta|\alpha)$ due to Type II censoring of exponential data.

By Part a) of Theorem 2, $M(\boldsymbol{Y};\Theta)$ and $M(\boldsymbol{Y};Y_\nu)$ are maximized by choosing α as small as possible. It is natural to expect that the limiting case, which is the Jeffreys prior $f(\theta)\propto\frac{1}{\theta}$, to be optimal with respect to both the parameter and prediction information. But its use is consequential. Since Jeffreys prior is improper, the expected parameter information is given by the negative conditional entropy of the posterior distribution which is proper. However, unlike the mutual information, the entropy is not invariant under one-to-one transformations and the result depends on the parametric function of interest. For example, for $p=1$ so the model is exponential with failure rate θ. The posterior distribution of θ is gamma $f(\theta|s_n)=\mathcal{G}(n,s_n)$ and its entropy is $H[f(\theta|s_n)]=H_{\mathcal{G}}(n)-\log s_n$. The distribution of S_n is Pareto $f(s_n)=\mathcal{P}(n-1)$ which is proper for $n>1$ and $E(\log S_n)=\frac{n}{n-1}$. The expected parameter information is $\mathcal{I}(\Theta|S_n)=-\mathcal{H}[f(\theta|s_n)]$, which is a decreasing function of n. But the posterior distribution of the mean parameter $\mu=\frac{1}{\theta}$ is inverse-gamma and information about the mean is increasing in n. With Jeffreys prior, the prior predictive distribution is also improper. The posterior predictive is Pareto $\mathcal{P}(n,s_n)$ and its entropy is $H[f(Y_\nu|s_n)]=H_{\mathcal{P}}(n)+\log s_n$. The expected predictive information is $\mathcal{I}(Y_\nu|S_n)=-\mathcal{H}[f(Y_\nu|s_n)]$, which is an increasing function of n.

5 Dependent Sequences

When the sequence of random variables $Y_i,\ i=1,2,\cdots$ are not conditionally independent, the information provided by the sample about the parameter and prediction jointly decomposes as:

$$M[\boldsymbol{Y};(\Theta,Y_\nu)]=M(\boldsymbol{Y};\Theta)+M(\boldsymbol{Y},Y_\nu|\theta)\geq M(\boldsymbol{Y};\Theta), \tag{25}$$

where $M(\boldsymbol{Y}, Y_\nu|\theta) \geq 0$ is the measure of conditional dependence, and the inequality becomes equality for the case of conditional independence. From (10) and (25), we find that

$$M(\boldsymbol{Y}, Y_\nu) \leq M(\boldsymbol{Y}; \Theta) \quad \text{if and only if} \quad M(\boldsymbol{Y}; \Theta|\boldsymbol{y}_\nu) \geq M(\boldsymbol{Y}, Y_\nu|\theta).$$

Thus, under strong conditional dependence the predictive information $M(\boldsymbol{Y}, Y_\nu)$ can dominate $M(\boldsymbol{Y}; \Theta)$ the parameter information.

In this section we first examine the effects of correlation between observations on the information about the mean parameter and prediction where the data are normally distributed. We then consider order statistics where no prior and model for the likelihood function are assumed.

5.1 Intraclass and Serially Correlated Models

We consider the intercept linear model $f(\boldsymbol{y}|\theta) = N(\theta \boldsymbol{z}, \sigma_1^2 R)$, where $\boldsymbol{z}$ is an $n \times 1$ vector of ones and $R = R|\theta = [\rho_{ij|\theta}]$ is a known correlation matrix. By invariance of the mutual information, the results hold for all distributions of variables that are one-to-one transformation of $\boldsymbol{y}$, e.g., log-normal model. As before, $\sigma_1^2 > 0$ is known and $f(\theta) = N(\mu_0, \sigma_0^2)$. The posterior variance is given by $\sigma^2_{\theta|\boldsymbol{y}} = \sigma_0^2 \left[1 + T_n(R)\eta^{-1}\right]^{-1}$, where $T_n(R) = \boldsymbol{z}' R^{-1} \boldsymbol{z}$ is the sum of all elements of R^{-1}. The parameter information is given by

$$M(\boldsymbol{Y}; \Theta|R) = .5 \log\left(1 + \eta^{-1} T_n(R)\right). \tag{26}$$

The following representations facilitate computation and study of the predictive and joint information measures. If Y_ν and $Y_\nu|\boldsymbol{y}$ are normal, then the predictive information is given by

$$M(\boldsymbol{Y}; Y_\nu) = -.5 \log\left(1 - \rho^2_{y_\nu, \boldsymbol{y}}\right) = .5 \log[C^{-1}]_{\nu\nu}, \tag{27}$$

where $\rho^2_{y_\nu, \boldsymbol{y}}$ is the square of unconditional multiple correlation coefficient of the regression of Y_ν on $\boldsymbol{y}$, $C = [c_{ij}]$, $i, j = 1, \cdots, n+1$ denotes the correlation matrix of the $(n+1)$-dimensional vector $(\boldsymbol{Y}, Y_\nu)$, and $[C^{-1}]_{\nu\nu}$ denotes the (ν, ν) element of C^{-1}.

The joint information about the parameter and prediction is given by

$$M[\boldsymbol{Y}; (\Theta, Y_\nu)] = M(\boldsymbol{Y}; \Theta) + M(\boldsymbol{Y}, Y_\nu|\Theta) \geq M(\boldsymbol{Y}; \Theta). \tag{28}$$

Thus, due to the conditional dependence, the joint information about parameter and prediction exceeds $M(\boldsymbol{Y}; \Theta)$ by an amount $M(\boldsymbol{Y}; Y_\nu|\Theta) \geq 0$ which is the information measure of the conditional dependence between Y_ν and the data. This quantity can be computed similar to (27) with conditional correlation matrix

$$M(\boldsymbol{Y}; Y_\nu|\Theta) = -.5 \log\left(1 - \rho^2_{y_\nu, \boldsymbol{y}|\theta}\right) = .5 \log[C^{-1}|\theta]_{\nu\nu} \geq 0, \tag{29}$$

Table 1. Formulas for uncorrelated, intraclass, and serial correlation models.

	Uncorrelated (UC)	Intraclass (IC)	Serial Correlation (SC)
$\|R\|\theta\|$	1	$[1+(n-1)\rho](1-\rho)^{n-1}$	$1-\rho^2$
$T_n(R\|\theta)$	n	$\frac{n}{1+(n-1)\rho}$	$\frac{n-(n-2)\rho}{1+\rho}$
$\rho^2_{y_\nu,\boldsymbol{y}\|\theta}$	0	$\frac{n\rho^2}{1+(n-1)\rho}$	ρ^2
ρ^2_p	$\frac{1}{1+\eta}$	$\frac{1+\eta\rho}{1+\eta}$	$\frac{1+\eta\rho^k}{1+\eta}$
$\rho^2_{y_\nu,\boldsymbol{y}}$	$\frac{n}{(1+\eta)(n+\eta)}$	$\frac{n\rho_p^2}{1+(n-1)\rho_p}$	Immediate future ρ^2_p

where $\rho^2_{y_\nu,\boldsymbol{y}|\theta}$ is the square of conditional multiple correlation coefficient and $C|\theta = [c_{ij}|\theta],\ i,j = 1,\cdots,n+1$ is the correlation matrix of conditional distribution of $(\boldsymbol{Y}, Y_\nu)$, given θ. Note that $C|\theta$ includes R and an additional row/column for Y_ν.

Measures such as the determinant $|R|$ and condition number $\kappa(R) = \sqrt{\lambda_1/\lambda_n}$, where $\lambda_1 < \cdots < \lambda_n$ are eigenvalues of R, can be used to rank dependence of the normal samples. However, in general, these measures do not provide a unique ranking. In order to rank the dependence uniquely as well as for ranking the predictive information in terms of sample dependence, we assume some structures for R. We consider two important models: the intraclass (IC) model with $\rho_{ij|\theta} = \rho$ for all $i \neq j$, and the serial correlation (SC) model with $\rho_{i,i\pm k|\theta} = \rho^k \geq 0,\ k > 0$. Dependence within each model and between the two models are ranked uniquely by $|R|$ and $\kappa(R)$.

Table 1 shows $|R|$ and $T_n(R)$ for the IC, SC models along with the independent model (UC). The determinants and inverses of the IC and SC matrices which are well known. Using $T_n(R)$ in (26) gives the parameter information. The third row of Table 1 shows $\rho^2_{y_\nu,\boldsymbol{y}|\theta}$ which is computed using (29) with $(n+1)$-dimensional IC and SC matrices for C. Table 1 also shows the square of unconditional (predictive) correlation $\rho^2_p = c_{ij}$, which are used in (27) for computing the predictive information measures. Computation of $\rho^2_p = c_{ij}$ is shown in the Appendix. The last row of Table 1 shows the square of unconditional multiple correlation coefficient $\rho^2_{y_\nu,\boldsymbol{y}}$ computed from (27). The predictive measure for SC model is for the one-step prediction.

The following Theorem summarizes the effects of the IC and SC correlation structures on the

normal information measures (26)-(28).

Theorem 3

(a) For all three models, $M(\boldsymbol{Y};\Theta|\rho)$, $M(\boldsymbol{Y};Y_\nu|\rho)$, and $M[\boldsymbol{Y};(\Theta,Y_\nu)|\rho)]$ increase with n and decrease with η.

(b) For both IC and SC models, $M(\boldsymbol{Y};\Theta|\rho)$ decreases with ρ, and

$$M^{IC}(\boldsymbol{Y};\Theta|\rho) \leq M^{SC}(\boldsymbol{Y};\Theta|\rho) \leq M^{UC}(\boldsymbol{Y};\Theta),$$

where the last equality holds if and only if $\rho = 0$.

(c) For both IC and SC models, $M(\boldsymbol{Y};Y_\nu|\rho)$ increases with ρ, and

$$M^{IC}(\boldsymbol{Y};Y_\nu|\rho) \geq M^{SC}(\boldsymbol{Y};Y_\nu|\rho) \geq M^{UC}(\boldsymbol{Y};Y_\nu),$$

where the last equality holds if and only if $\rho = 0$.

(d) For both IC and SC models, $M[\boldsymbol{Y};(\Theta,Y_\nu)|\rho)]$ decreases in ρ for $\rho \leq \rho_0(n,\eta)$ and increases in ρ for $\rho > \rho_0(n,\eta)$, where $\rho_0^{IC}(n,\eta)$ and $\rho_0^{SC}(n,\eta)$ are roots of quadratic equations and both are increasing in n and decreasing in η.

Proof. (a) Can be easily seen by taking derivatives. (b) It is also easy to see that for the correlated models $T_n(R)$ are decreasing functions of ρ and that $T_n^{IC}(R) \leq T_n^{SC}(R) \leq T_n^{UC}(R) = n$. (c) This is implied by the facts that $\rho_p^{IC} > \rho_p^{SC}$ and the predictive information increases with ρ, as expected. (d) Taking the derivative, $\rho_0(n,\eta)$ is given by the root of $A_{n,\eta}\rho^2 + B_{n,\eta}\rho + C_{n,\eta} = 0$, where $A_{n,\eta}^{IC} = n-1,\ B_{n,\eta}^{IC} = 2(1+n\eta^{-1}), A_{n,\eta}^{SC} = 1+(2n-1)\eta^{-1},\ B_{n,\eta}^{SC} = 1+(2n-1)\eta^{-1}$, and $C_{n,\eta} = (1-n)\eta^{-1}$. For each model there is only a unique positive solution.

Theorem 3 formalizes the intuition that samples with stronger dependence are less informative about the parameter and more informative about prediction. Figure 5 shows plots of $M(\boldsymbol{Y};\Theta|\rho)$ and $M(\boldsymbol{Y};Y_\nu|\rho)$ against sample size for the UC model and the correlated models IC and SC with $\rho = .50, .75$. Some noteworthy features are as follows. All information measures are increasing in n. For the UC model, the parameter information is the highest and has the fastest rate of increase with n, and the predictive information is the lowest with the slowest (almost flat) rate of increase. For the SC model, the parameter information is higher and increases much faster than the predictive information. For the IC model, the parameter information is lower than the predictive information

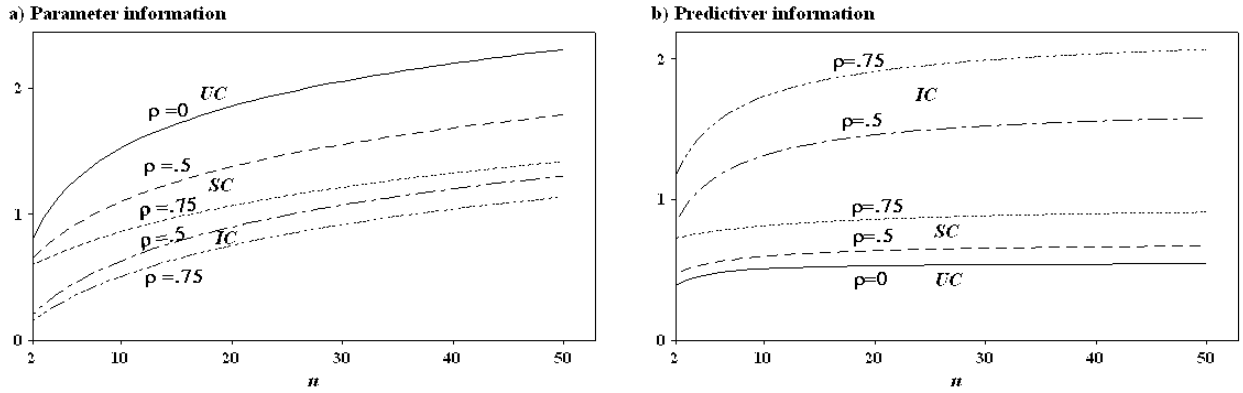

Figure 5: The parameter information $M(\boldsymbol{Y};\Theta|\rho)$ and predictive information $M(\boldsymbol{Y};Y_\nu|\rho)$ for the independent, IC, and SC normal models as functions of the sample size ($\eta = .5$).

while both measures have about the same rates of increase. Interestingly, as seen in Figures 5a and 5b, for the UC and SC models, the differences between the parameter and predictive information measures grow with n much faster than the predictive information measures. That is, the share of predictive information decreases with the sample size.

Since $M(\boldsymbol{Y};\Theta|\rho)$ is increasing in n, one can compensate the loss of parameter information due to the dependence by increasing the sample size. For example, as can be seen in Figure 5a, to gain about one unit (nit) of information, we need $n = 3$ from the UC, and with $\rho = .50, .75$, we need $n = 8, 16$ observations under SC, and $n = 26, 37$ observations under IC models, respectively.

The effects of prior on the information quantities are induced through η which is proportional to prior precision. It is clear that (26) and (27), and their difference are decreasing in η. Thus, the optimal prior for inference about the parameter and prediction is to choose the prior variance as large as possible.

Figure 6a shows the plots of the joint information measures for the SC and IC models as functions of ρ^2 for $n = 5, 10$ and $\eta = .5$. Note that the joint information of SC model dominates the joint information of IC model when dependence is weak. After the minimum point, the of rate of growth of joint information for the IC model is steep and the IC information measure dominates the SC information measure for when the dependence is rather strong. Figure 6b shows the plots of the minimum joint information measures for the SC and IC families as functions of n for $\eta = .25, .5, .75$. These plots are useful for determining sample size for each family such that the minimum information exceeds a given value. For example, to gain about 1.5 units (nits) of information from an SC sample with unknown ρ, we need $n = 9, 25, 37$ with $\eta = .25, .50, .75$,

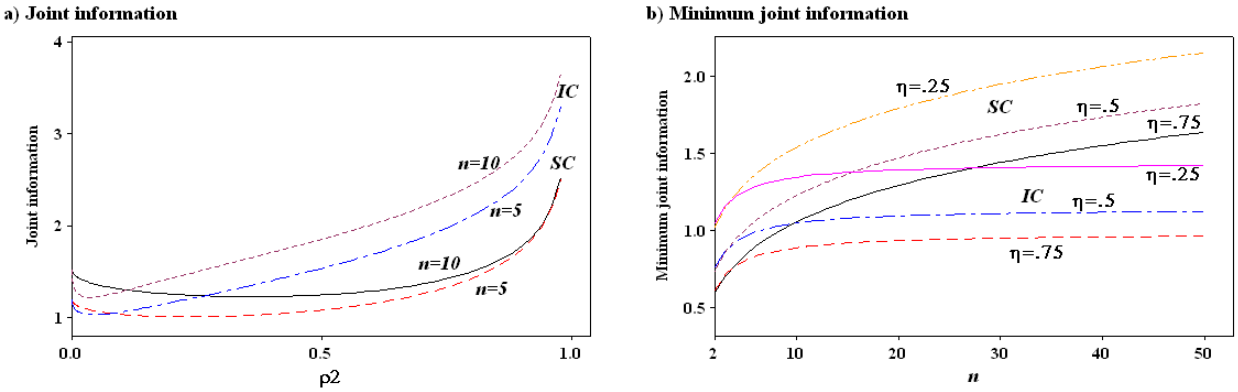

Figure 6: The joint parameter and predictive information $M[\boldsymbol{Y};(\Theta,Y_\nu)|\rho]$ and minima of the joint information $\min_\rho M[\boldsymbol{Y};(\Theta,Y_\nu)|\rho]$ for SC and IC normal models.

respectively. The plots show that

$$M_0^{IC}[\boldsymbol{Y};(\Theta,Y_\nu)|n,\eta] \leq M_0^{SC}[\boldsymbol{Y};(\Theta,Y_\nu)|n,\eta],$$

where $M_0[\boldsymbol{Y};(\Theta,Y_\nu)|n,\eta] = \min_\rho M[\boldsymbol{Y};(\Theta,Y_\nu)|\rho]$. This inequality can be proved by substituting $\rho_0^{IC}(n,\eta)$ and $\rho_0^{SC}(n,\eta)$ in the expressions for $T_n(R)$ and $\rho^2_{y_\nu,\boldsymbol{y}|\theta}$.

5.2 Order Statistics

Let $Y_1 \leq Y_2 \leq \cdots \leq Y_n$ be the order statistics of conditionally independent sample $X_1,\cdots,X_n$ from a continuous distribution with density function $g(x|\theta)$, and let $\boldsymbol{y}_r = (y_1,\cdots,y_r)$, $r \leq n$. Conditional on θ, the order statistics have a Markovian dependence (Arnold 1992). The mutual information between consecutive order statistics given by

$$\begin{aligned} M(Y_r;Y_{r+1}|\theta) &= M_n(r) \\ &= \log B(r+1,n-r+1) + \log(n+1) - 1 \\ &\quad -r\{\psi(r)-\psi(n)\} - (n-r)\{\psi(n-r)-\psi(n)\}; \end{aligned} \tag{30}$$

see Ebrahimi et al.(2004). That is, $M_n(r)$ is the measure of Markovian dependence between order statistics of the independent sample conditional on θ. It is shown in Ebrahimi et al.(2004) that $M_n(r)$ is increasing in n, and for a given n, the information is symmetric in r and $n-r$, and attains its maximum at the median (see Figure 7). Next lemma gives generalizations of (30). All information functions are conditional on r and n, which will be suppressed when unnecessary.

Lemma 1 *Let $Y_1 \leq \cdots \leq Y_n$ denote the order statistics of random variables $X_1, \cdots, X_n$ which, given θ, are independent and have identical distribution $g(x|\theta)$ and $\boldsymbol{Y}_r$ and $\boldsymbol{Y}_q$ denote the disjoint subvectors of order statistics. Then:*

(a) $M(\boldsymbol{Y}_r; \boldsymbol{Y}_q|\theta)$ is free from the parent distribution $f(x|\theta)$ and the prior distribution $f(\theta)$,

(b) For any two consecutive subvectors $\boldsymbol{Y_r} = (Y_{k+1}, \cdots, Y_{k+r})$ and $\boldsymbol{Y}_q = (Y_{k+r+1}, \cdots, Y_{k+r+q})$, $M(\boldsymbol{Y}_r; \boldsymbol{Y}_q|\theta) = M_n(r)$.

Proof. Let $U = G(X)$. Then U is uniform and its order statistics $W_1 \leq W_2 \leq \cdots \leq W_n$ are given by $W_i = G(Y_i)$, and $\boldsymbol{W}_r$ and $\boldsymbol{W}_q$ be the subvectors corresponding to $\boldsymbol{Y}_r$ and $\boldsymbol{Y}_q$. Since $W_i = G(Y_i)$ is one-to-one, we have $M(\boldsymbol{Y}_r; \boldsymbol{Y}_q) = M(\boldsymbol{W}_r; \boldsymbol{W}_q)$. Furthermore the distribution of any subset of order statistics is ordered Dirichlet with parameters n and the indices of the order statistics contained in the subset, hence $M(\boldsymbol{Y}_r; \boldsymbol{Y}_q) = M(\boldsymbol{W}_r; \boldsymbol{W}_q)$ is free from the parent distribution $f(x|\theta)$. Part (b) Follows from $Y_1|\theta, \cdots, Y_n|\theta$ being a Markovian sequence.

It can easily be shown that information provided by the first r order statistics about the parameter $M(\boldsymbol{Y}_r, \Theta)$ satisfies (6). The predictive distributions of order statistics are given by $f(y_i) = \int f(y_i|\theta) f(\theta) d\theta,\ i = 1, \cdots, n$. Note that $y_1 \leq y_2 \leq \cdots \leq y_n$ are the order statistics of a sample of the exchangeable sequence $X_1, \cdots, X_n$, unconditionally. The following results provide some insight about the parameter and predictive information for order statistics.

Theorem 4 *Let $M[\boldsymbol{Y}_r; (\Theta, Y_{r+1})]$ denote the information provided by the first r order statistics about the parameter and for prediction of the next order statistic jointly. Then:*

(a) $M[(\boldsymbol{Y}_r; (\Theta, Y_{r+1})] = M(\boldsymbol{Y}_r; \Theta) + M_n(r) \geq M(\boldsymbol{Y}_r; \Theta)$.

(b) The following statements are equivalent:

(i) $M(\boldsymbol{Y}_r; Y_{r+1}) \geq (\leq)\ M_n(r)$.

(ii) $M(\Theta; Y_{r+1}) \geq (\leq)\ M(\boldsymbol{Y_{r+1}}; \Theta) - M(\boldsymbol{Y}_r; \Theta)$, where $\boldsymbol{Y_{r+1}} = (Y_1, \cdots, Y_r, Y_{r+1})$.

Proof. Using the following decompositions of mutual information we have,

$$M[(\Theta, Y_{r+1}); \boldsymbol{Y}_r] = M(\boldsymbol{Y}_r; \Theta) + M(\boldsymbol{Y}_r; Y_{r+1}|\Theta).$$

Applying part (b) of Lemma 1 to the second term gives the result (a). For (b) we use the following decompositions of mutual information $M[(\boldsymbol{Y}_r, \Theta); Y_{r+1}]$ decompositions:

$$M[(\boldsymbol{Y}_r, \Theta); Y_{r+1}] \quad = \quad M(\boldsymbol{Y}_r; Y_{r+1}) + M(\Theta; Y_{r+1}|\boldsymbol{Y}_r)$$

$$= \quad M(\Theta; Y_{r+1}) + M(\boldsymbol{Y}_r; Y_{r+1}|\Theta).$$

Equating the two decomposition with $M(Y_{r+1}; \boldsymbol{Y}_r|\Theta) = M_n(r)$ gives equivalence of (i) and

$$M(\Theta; Y_{r+1}) \geq (\leq)\ M(\Theta; Y_{r+1}|\boldsymbol{Y}_r). \tag{31}$$

The equivalence with (ii) is obtained by solving

$$M(\boldsymbol{Y_{r+1}}; \Theta) = M(\boldsymbol{Y}_r; \Theta) + M(\Theta; Y_{r+1}|\boldsymbol{Y}_r),$$

for $M(\Theta; Y_{r+1}|\boldsymbol{Y}_r)$ and substituting in (31).

Part (a) of Theorem 4 shows $M[\boldsymbol{Y}_r; (\Theta, Y_{r+1})]$ is inclusive of Lindley's measure reflecting the fact that conditional on θ, order statistics are dependent. So the information provided by the first r order statistics about the parameter and for prediction of the next order statistic is more than the information provided about the parameter. However, the excess amount of information measures the Markovian dependence between order statistics of the independent sample and does not depend on $f_{x|\theta}$ and f_θ. An implication of this result is that reference posterior corresponding to the prior that maximizes the parameter information $M(\boldsymbol{Y}_r; \Theta)$ also remains optimal with respect to $M[\boldsymbol{Y}_r; (\Theta, Y_{r+1})]$.

Part (b) of Theorem 4 gives the equivalence of the orders of information in terms of (i) the predictive and sample order statistics and (ii) the expected information about the parameter provided by an order statistic in terms of the incremental amount of information provided about the parameter.

Example 1 For the case of exponential model with the gamma prior, the conditional distribution of $(r+1)$st order statistic given θ and the first r order statistics is exponential with density

$$f(y_{r+1}|y_r, \theta) = (n-r)\theta e^{-\theta(n-r)(y_{r+1}-y_r)}, \quad y_{r+1} > y_r. \tag{32}$$

The posterior predictive distribution of $(r+1)$st order statistic given first r order statistics is Pareto with parameters $\alpha+r$, $b_r = \dfrac{\beta + t_r}{n-r}$, and a location parameter y_r. Since entropy is location-invariant, $H(Y_{r+1}|\boldsymbol{y}_r)$ is $H(Y_{r+1}|t_r, y_r, r, n) = H(Y|t_r, r) - \log(n-r)$.

Figure 7a shows plots of $M(\boldsymbol{Y}_r; \Theta|\alpha, r, n)$ for the exponential sample with $n = 26$ and $\alpha = .5, 1, 2, 4$, superimposed by the Markovian dependence information measure $M_n(r)$ for the order statistics. Since $M(T_r; \Theta|\alpha, r, n)$ is increasing in r, censoring results in loss of information about the parameter. Thus, without consideration of cost the experiment, $r^* = n = 26$. Since $M_n(r)$ is

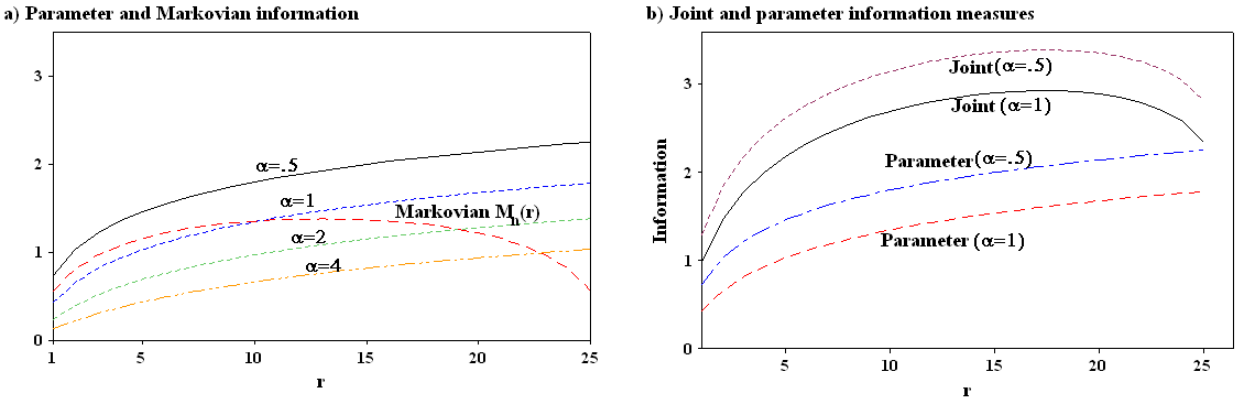

Figure 7: Expected information about the parameter $M(\boldsymbol{Y}_r, \Theta)$ and the joint information about the parameter and prediction of the $(r+1)$st order statistic $M[\boldsymbol{Y}_r; (\Theta, Y_{r+1})]$ provided by the vector of preceding order statistics $\boldsymbol{Y}_r$, and the information due to the Markovian dependence between order statistics ($n = 26$).

decreasing for r larger than the median, censoring beyond the median results in gain of information about the next outcome. Figure 7b shows the plots of the parameter information and joint information $M[\boldsymbol{Y}_r; (\Theta, Y_{r+1})|\alpha, r, n]$ computed using part (a) of Theorem 4 for $\alpha = 0.5, 1$. We note that $M[\boldsymbol{Y}_r; (\Theta, Y_{r+1})|\alpha, r, n]$ is not monotone because the Markovian dependence information measure $M_n(r)$ decreases for the order statistics above the median. The optimal r for the joint parameter and predictive information, without consideration of cost the experiment is $r^* = 17 < n$. Thus, unlike the case of conditionally independent model, the parameter information utility and the joint parameter-predictive information utility lead to different sampling plans.

In Section 4 we noted that under Jeffreys prior, at least one observation is needed for obtaining a proper posterior. Following this idea more generally, we compare the expected uncertainty change due to the first r order statistics with the first order statistic $r = 1$, give by

$$\mathcal{B}[\boldsymbol{Y}_r; (\Theta, Y_{r+1})] = H[(\Theta, Y_1)] - \mathcal{H}[(\Theta, Y_{r+1})|\boldsymbol{Y}_r],$$

where $\mathcal{H}[(\Theta, Y_{r+1})|\boldsymbol{Y}_r] = E_{s_r}\{H[(\Theta, Y_{r+1})|\boldsymbol{Y}_r]\}$ is the conditional joint entropy of (Θ, Y_{r+1}) given the first order statistic, averaged with respect to $f(\boldsymbol{y}_r)$. The expected uncertainty change $\mathcal{B}(\boldsymbol{Y}_r; Y_{r+1})$ for prediction of $(r+1)$st order statistic is defined similarly. These measures, which can be referred to as the information bridge between the first and $(r+1)$st order statistics, are invariant under linear transformations, but can be negative. It can be shown that for any parent distribution $g(x|\theta)$

where θ is the scale parameter and any prior $f(\theta)$,

$$M(\boldsymbol{Y}_r;\Theta) = \mathcal{B}[\boldsymbol{Y}_r;(\Theta,Y_{r+1})] + \log\left(\frac{n}{n-r}\right)$$
$$M(\boldsymbol{Y}_r;Y_{r+1}) = \mathcal{B}(\boldsymbol{Y}_r;Y_{r+1}) + \log\left(\frac{n}{n-r}\right).$$

Clearly, $\mathcal{B}(\cdot,\cdot|r,n) \to M(\cdot,\cdot|r,n)$ as $\dfrac{r}{n} \to 0$. So, the quantity $\log\left(\dfrac{n}{n-r}\right)$ can be interpreted as the finite sample correction factor for the information.

6 Concluding Remarks

This paper is the first attempt to study the relationship between the parameter and predictive information measures, the analytical behavior of the predictive information in terms of prior parameter, and the effects of conditional dependence between the observable quantities on the Bayesian information measures. We provided analytical results and showed applications in some statistical and modeling problems.

The measure of information that sample provides about the parameter and prediction jointly led to some new insights about the marginal parameter and predictive information measures. For the case of conditionally independent observations, decompositions of the joint information revealed that the parameter information is in fact the measure of information about the parameter and prediction jointly. This finding implies that all existing results about Lindley's information are applicable to the joint measure of parameter and predictive information. In particular, the reference posterior and the optimal design that maximize the sample information about the parameter are also optimal solutions for the sample information about the parameter and prediction jointly. Yet another information decomposition revealed that predictive information is a part of the information that sample provides about the parameter.

We examined interplay between the information measures and the prior and design parameters for two general classes of models: the linear models for the normal mean, and a broad subfamily of the exponential family. A few applications showed the usefulness of the information measures and some insights were developed. A proposition provided the optimal designs with respect to the parameter (joint) information and predictive information measures for an ANOVA type model. The results include the minimum sample sizes required in terms of the given prior variances and the covariate vector for the prediction. Another proposition provided the optimal prior variance allocation scheme with respect to the parameter (joint) information for collinear regression, which includes the minimum prior variance required for the problem. Examples for the linear and the

exponential family models revealed that the predictive information provided by the conditional independence sample is only a small fraction of the parameter (joint) information and the gap between the parameter and predictive information measures grows rapidly with the sample size. This finding indicates that despite of the importance of prediction in the Bayesian paradigm, the parameter takes the major share of the information provided by conditionally independent samples. An example examined the parameter information when the parameter of interest is the vector of means of two treatments and the predictive information of interest is for the weighted average (or contrast) between outcomes of the two treatments. This example revealed that the loss of information about the parameter under the optimal design for predictive information is much higher than the loss of predictive information under the optimal design for the parameter information. The parameter is the major share holder of the sample information so its loss is more severe than the loss of predictive information under suboptimal designs.

We have examined, for the first time, the role of conditional dependence between observable quantities on the sample information about the parameter and prediction. For a dependent sequence, the joint parameter and predictive information decomposes into the parameter information (Lindley's measure) and an information measure mapping the conditional dependence. We provided more specific results for correlated variables whose distributions can be transformed to normal and for the order statistics without any distributional assumption. For the normal sample, we compared the information measures for the independent, the intraclass correlation, and serial correlation models. We showed that the parameter information decreases and predictive information increases with the correlation. However, the joint information decreases in the correlation to a minimum point, which is determined by the prior precision and sample size, and then increases. For conditionally dependent sequences, the dominance of parameter information that was noted for the conditionally independent samples does not hold. Since all information measures increase with the sample size, loss of parameter information due to dependence can be compensated by taking larger samples.

Order statistics also provided a context for information analysis of conditionally Markovian sequences. Extension of a result on information properties of order statistics was needed to show that the Markovian dependence measure depends neither on the model for the data, nor on the prior distribution for the parameter. By this finding, the reference posterior that maximizes the sample information about the parameter, retains its optimality according to the joint parameter and predictive information measure of the order statistics. An example illustrated implication in terms of the optimal number of failures to be observed under Type II censoring.

Several authors have used information in various Bayesian contexts, which are not listed in Table A of the Appendix; see for example, Aitchison (1975), Zellner (1977, 1988, 1997), Geisser (1993), Keyes and Levy (1996), Ibrahim and Chen (2000), Brown, George and Xu (2008), and references therein. Nicolae et al. (2008) defined some measures of fraction of missing information and have pointed out connection between their measures and the entropy, stating that "essentially all measures we presented have entropy flavor". Measures of information for nonparametric Bayesian data analysis is also available (Müller and Quintana 2004). Since our focus is on the mutual information, e.g. Lindley's measure and its predictive version, we did not discuss other information measures.

Acknowledgments

We thank the reviewers for their comments and suggestions which led us to improve the exposition. Ehsan Soofi's research was partially supported by a Lubar School's Business Advisory Council Summer Research Fellowship.

Appendix

Classification of literature

Table A gives a classification of literature on the Bayesian applications of mutual information.

Table A. Classification of articles on Lindley's measure of Sample Information about the Parameter and its Predictive Version.

Parameter information:	
Likelihood model & design:	Lindley (1956, 1957, 1961), Stone (1959), El-Sayyed (1969), Brooks (1980, 1982), Smith & Verdinelli (1980), Turrero (1982), Barlow & Hsiung (1983), Soofi (1988, 1990), Ebrahimi & Soofi (1990), Carlin & Polson (1991), Verdinelli & Kadane (1992), Polson (1992), Verdinelli (1993), Parmigiani & Berry (1994), Chaloner & Verdinelli (1995), Carota, et al. (1996), Singpurwalla (1996), Yuan & Clarke (1999)
Prior & posterior distributions:	Bernardo (1979a,b), Soofi (1988, 1990), Ebrahimi & Soofi (1990), Bernardo & Rueda (2002), Bernardo (2004)
Predictive information	
Likelihood model & Design:	San Martini & Spezzaferri (1984), Amaral & Dunsmore (1985), Verdinelli (1993), Verdinelli, et al. (1993), Chaloner & Verdinelli (1995), Singpurwalla (1996)

Proof of Proposition 1

(a) Noting that $\lambda_j = n_j, j = 1, \cdots, p$ and letting $n_1 = n - \sum_{j=2}^{p} n_j$ in (13) gives the first order conditions

$$\frac{\partial M(\boldsymbol{Y}; \boldsymbol{\Theta}|Z, \eta, V_0)}{\partial n_j} = \frac{1}{2}\left[\frac{v_{0j}}{\eta + v_{0j}n_j} - \frac{v_{01}}{\eta + v_{01}n_1}\right] = 0,\ j = 2, \cdots, p.$$

Solutions to this system give $n_j^*, j = 2, \cdots, p$ in (15) and n_1^* is found from $n_1^* = n - \sum_{j=2}^{p} n_j^*$. It can be verified by the second order conditions that the solutions give the maximum.

(b) Using $V_1 = \left(\eta V_0^{-1} + Z'Z\right)^{-1}$ in (14) gives

$$M(\boldsymbol{Y}; Y_\nu|\boldsymbol{z}_\nu, Z, \eta, V_0) = \frac{1}{2}\log\left(\eta^{-1}\boldsymbol{z}_\nu' V_0 \boldsymbol{z}_\nu + 1\right) - \frac{1}{2}\log\left(\boldsymbol{z}_\nu'\left(\eta V_0^{-1} + Z'Z\right)^{-1}\boldsymbol{z}_\nu + 1\right).$$

The first term does not depend on the design, so it is sufficient to minimize

$$h(n_1, \cdots, n_p) = \boldsymbol{z}_\nu'\left(\eta V_0^{-1} + Z'Z\right)^{-1}\boldsymbol{z}_\nu = \sum_{j=1}^{p} \frac{v_{0j} z_j^2}{\eta + v_{0j}n_j}$$

subject to the constraint $\sum_{j=1}^{p} n_j = n$. Letting $n_1 = n - \sum_{j=2}^{p} n_j$ gives the first order conditions

$$\frac{\partial h(n_1, \cdots, n_p)}{\partial n_j} = -\frac{v_{0j}^2 z_j^2}{(\eta + v_{0j}n_j)^2} + \frac{v_{01}^2 z_1^2}{(\eta + v_{01}n_1)^2} = 0,\ j = 2, \cdots, p.$$

Solutions to this system give $n_j^*, j = 2, \cdots, p$ in (16) and n_1^* is found from $n_1^* = n - \sum_{j=2}^{p} n_j^*$. It can be verified by the second order conditions that the solutions give the maximum.

Proof of Proposition 2

The solutions are found similarly to Part a) of Proposition 1 by taking the derivative of (13) with respect to v_{0j} subject to $\sum_{j=1}^{p} v_{0j} = c$.

Computation of Normal Predictive Correlation

We compute the predictive correlation ρ_p through the well-known formula for partial correlation:

$$\rho_{ij|k} = \frac{\rho_{ij} - \rho_{ik}\rho_{jk}}{(1 - \rho_{ik}^2)^{1/2}(1 - \rho_{jk}^2)^{1/2}}. \tag{33}$$

In our case, i, j, k represent Y_i, Y_ν, and θ, respectively. Note that

$$\rho_{i\theta}^2 = 1 - \frac{\sigma_{\theta|y_i}^2}{\sigma_0^2} = \frac{1}{1+\eta},\ \text{for all } i = 1, 2, \cdots.$$

Letting $\rho_{ik}^2 = \rho_{jk}^2 = \rho_{i\theta}^2$ in (33) gives the unconditional (predictive) correlation as

$$\rho_{i\nu} = \rho_{i\theta}^2 + (1 - \rho_{i\theta}^2)\rho_{i\nu|\theta} = \frac{1 + \eta\rho_{i\nu|\theta}}{1 + \eta}.$$

Letting $\rho_{i\nu|\theta} = 0, \rho, \rho^{\nu-i},\ \nu > i$ respectively for UC, IC and SC models, we obtain the entries of Table 1 for the three models.

References

Abramowitz, M. and Stegun, I. A. (1970). *Handbook of Mathematical Functions, with Formulas, and Mathematical Tables*, New York: Dover.

Abel, P. S. and Singpurwalla, N. D. (1994). To survive or to fail: that is the question. *Ameri. Statist.* **48** 18-21.

Aitchison, J. (1975). Goodness of prediction fit. *Biometrika* **62** 547-554.

Amaral-Turkman, M. A. and Dunsmore, I. (1985). Measures of information in the predictive distribution. In *Bayesian Statist.* **2** (eds J.M. Bernardo, M.H. DeGroot, D.V. Lindley and A.F.M. Smith), 603-612.

Arnold, B., Balakrishnan, N. and Nagaraja, H. N. (1992). *First Course in Order Statistics*, New York: Wiley.

Barlow, R. E. and Hsiung, J. H. (1983). Expected information from a life test experiment. *Statistician* **48** 18-21.

Belsley, D. A. (1991). *Conditioning Diagnostics: Collinearity and Weak Data in Regression*, New York: Wiley.

Bernardo, J. M. (1979a). Expected information as expected utility. *Ann. Statist.* **7** 686-690.

Bernardo, J. M. (1979b). Reference posterior distribution for Bayesian inference (with discussion). *J. Roy. Statist. Soc. Ser. B* **41** 605-647.

Brooks, R. J. (1980). On the relative efficiency of two paired-data experiment. *J. Roy. Statist. Soc. Ser. B* **42** 186-191.

Brooks, R. J. (1982). On loss of information through censoring. *Biometrika* **69** 137-144.

Brown, L. D., George, E. I. and Xu, X. (2008). Admissible predictive density estimation. *Ann. Statist.* **36** 1156-1170.

Chaloner, K. and Verdinelli, I. (1995). Bayesian experimental design: A review. *Statist. Sci.* **10** 273-304.

Cover, T. M. and Thomas, J. A. (1991). *Elements of Information Theory*, New York: Wiley.

Carlin, B. P. and Polson, N. G. (1991). An expected utility approach to influence diagnostics. *J. Amer. Statist. Assoc.* **87,** 1013-1021.

Carota, C., Parmigiani, G. and Polson, N. G. (1996). Diagnostic measures for model criticism. *J. Amer. Statist. Assoc.* **91** 753-762.

Ebrahimi, N. (1992) Prediction intervals for future failures in the exponential distribution under hybrid censoring. *IEEE Trans. Reliability* **41** 127-132.

Ebrahimi, N. and Soofi, E. S. (1990). Relative information loss under type II censored exponential data. *Biometrika* **77** 429-435.

Ebrahimi, N., Maasoumi, E. and Soofi, E. S. (1999). Ordering univariate distributions by entropy and variance. *J. Econometrics* **90** 317-336.

Ebrahimi, N., Soofi, E. S. and Zahedi, H. (2004). Information properties of order statistics and spacings. *IEEE Trans. Information Theory* **50** 177-183.

El-Sayyed, G. M. (1969). Information and sampling from exponential distribution. *Technometrics* **11** 41-45.

Geisser, S. (1993). *Predictive Inference: An Introduction*, New York: Chapman-Hall.

Good, I. J. (1971). Discussion of article by R.J. Buehler, in *Foundations of Statistical Inference*, Eds. V.P. Godambe, and D.A. Sprott, pp. 337-339. Toronto: Holt, Rinehart and Winston.

Ibrahim, J. G. and Chen, M. H. (2000). Power prior distributions for regression models. *Statist. Sci.* **15** 46-60.

Kaminsky, K. S. and Rhodin, L.S. (1985). Maximum likelihood prediction. *Ann. Inst. Statist. Math.* **37** 505-517.

Keyes, T. K. and Levy, M.S. (1996). Goodness of prediction fit for multivariate linear models. *J. Amer. Statist. Assoc.* **91** 191-197.

Kullback, S. (1959). *Information Theory and Statistics*, New York: Wiley (reprinted in 1968 by Dover).

Lawless, J. L. (1971). A prediction problem concerning samples from the exponential distribution with application in life testing. *Technometrics* **13** 725-730.

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

Lindley, D. V. (1957). Binomial sampling schemes and the concept of information. *Biometrika* **44** 179-186.

Lindley, D. V. (1961). The use of prior probability distributions in statistical inference and decision. *Proceedings of Fourth Berkeley Symposium* **1** 436-468, Berkeley: UC Press.

Nicolae, D.L., Meng, X-L. and Kong, A. (2008). Quantifying the fraction of missing information for hypothesis testing in statistical and genetic studies (with discussion). *Statist. Sci.* **19** 95-110.

Maruyama, Y. and George, E. I. (2010). Fully Bayes Model Selection with a Generalized g-Prior. Working Paper. University of Pennsylvania.

Müller, P. and Quintana, F. A. (2004). Nonparametric Bayesian data analysis. *Statist. Sci.* **23** 287-331.

Parmigiani, G. and Berry D. A. (1994). Applications of Lindley information measures to the design of clinical experiments. In *Aspects of Uncertainty: Attribute to D.V. Lindley*, eds. P.R. Freeman and A.F.M. Smith, pp. 329-348, West Sussex, UK: Wiley.

Polson, N. G. (1992). On the expected amount of information from a nonlinear model. *J. Roy. Statist. Soc. Ser. B* **54** 889-895.

Polson, N. G. (1993). A Bayesian perspective on the design of accelerated life tests. *Advances in Reliability*, Ed. A. P. Basu, pp. 321-330. North Holland.

Pourahmadi, M. Soofi, E. S. (2000). Predictive variance and information worth of observations in time series. *J. Time Ser. Anal.* **21** 413-434.

San Martini, A. and Spezzaferri, F. (1984). A predictive model selection criteria. *J. Roy. Statist. Soc. Ser. B* **46** 296-303.

Singpurwalla, N.D. (1996). Entropy and Information in Reliability. in *Bayesian Analysis of Statistics and Econometrics: Essays in Honor of Arnold Zellner*, Eds. D. Berry, K. Chaloner, and J. Geweke, 459-469, New York: Wiley.

Soofi, E. S. (1988). Principal component regression under exchangeability. *Comm. Statist. Theory and Methods* **A17** 1717-1733.

Soofi, E. S. (1990). Effects of collinearity on information about regression coefficients. *J. Econometrics* **43** 255-274.

Stewart, G. W. (1987). On collinearity and least squares regression (with discussion). *Statist. Sci.* **2** 68-100.

Stone, M. (1959). Application of a measure of information to the design and comparison of regression experiments. *Ann. Math. Statist.* **29** 55-70.

Yuan, A. and Clarke, B. (1999). An information criterion for likelihood selection. *IEEE Trans. Information Theory* **45** 562-571.

Verdinelli, I. (1992). Advances in Bayesian experimental design. In Bayesian Statistics **4** (J. O. Berger, J. M. Bernardo, A. P. Dawid and A. F. M. Smith, Eds.) 467-481, Wiley.

Verdinelli, I. and Kadane, J. B. (1992). Bayesian designs for maximizing information and outcome. *J. Amer. Statist. Assoc.* **87** 510-515.

Verdinelli, I., Polson, N. G. and Singpurwalla, N.D. (1993). Shannon information and Bayesian design for prediction in accelerated life-testing. *Reliability and Decision Making*, Eds. R.E. Barlow, C.A. Clarotti, and F. Spizzichino, 247-256. London: Chapman Hall.

West, M. (2003). Bayesian factor regression models in the "large p, small n" paradigm. In J. M. Bernardo, M. J. Bayarri, J.O. Berger, A.P. Dawid, D. Heckerman, A.F.M. Smith, and M. West (eds.) *Bayesian Statist.* **7** 723732. Oxford University Press.

Zellner, A. (1986). On assessing prior distributions and Bayesian regression analysis with g-prior distributions. *Bayesian Inference and Decision Techniques*, Eds. P. Goel and A. Zellner, 233-243, Amsterdam: North-Holland.

Zellner, A. (1988). Optimal information processing and Bayes' theorem (with discussion). *Amer. Statist.* **42** 278-284.