Entropy **2012**, *14*, 2427-2438; doi:10.3390/e14122427

OPEN ACCESS

entropy

ISSN 1099-4300

www.mdpi.com/journal/entropy

Article

Implications of the Cressie-Read Family of Additive Divergences for Information Recovery

George G. Judge [1,†,*] and Ron C. Mittelhammer [2]

[1] 207 Gianinni Hall, University of California, Berkeley, Berkeley, CA 94720, USA

[2] School of Economic Sciences, Washington State University, Pullman, WA 99164, USA; E-Mail: mittelha@wsu.edu

† Member of the Giannini Foundation.

* Author to whom correspondence should be addressed; E-Mail: gjudge@berkeley.edu.

Received: 15 October 2012; in revised form: 20 November 2012 / Accepted: 27 November 2012 / Published: 3 December 2012

Abstract: To address the unknown nature of probability-sampling models, in this paper we use information theoretic concepts and the Cressie-Read (CR) family of information divergence measures to produce a flexible family of probability distributions, likelihood functions, estimators, and inference procedures. The usual case in statistical modeling is that the noisy indirect data are observed and known and the sampling model-error distribution-probability space, consistent with the data, is unknown. To address the unknown sampling process underlying the data, we consider a convex combination of two or more estimators derived from members of the flexible CR family of divergence measures and optimize that combination to select an estimator that minimizes expected quadratic loss. Sampling experiments are used to illustrate the finite sample properties of the resulting estimator and the nature of the recovered sampling distribution.

Keywords: conditional moment equations; Cressie-Read divergence; information theoretic methods; minimum power divergence; information functionals

PACS Codes: 89.70, 89.70.Cf

JEL Classifications: C13, C14, C25, C51

1. Introduction

Uncertainty regarding statistical models and associated estimating equations and the data sampling-probability distribution function create unsolved problems as they relate to information recovery. Although likelihood is a common loss function used in fitting statistical models, the optimality of a given likelihood method is fragile inference-wise under model uncertainty. In addition the precise functional representation of the data sampling process cannot usually be justified from physical or behavioral theory. Given this situation, a natural solution is to use estimation and inference methods that are designed to deal with systems that are fundamentally stochastic and where uncertainty and random behavior are basic to information recovery. In this context [1–2], the family of likelihood functionals permits the researcher to face the resulting stochastic inverse problem and exploit the statistical machinery of information theory to gain insights relative to the underlying causal behavior from a sample of data.

In developing an information theoretic approach to estimation and inference, the Cressie-Read (CR) family of information divergences represents a way to link the model of the process to a family of possible likelihood functions associated with the underlying sample of data. Information divergences of this type have an intuitive interpretation reflecting the uncertainty of uncertainty as it relates to a model of the process and a model of the data. These power divergences give new meaning to what is a likelihood function and what is the appropriate way to represent the possible underlying sampling distribution of statistical model.

One possibility for implementing this approach is to use estimating equations-moment conditions (prior information) to model the process and provide a link to the data. Discrete members of the CR family are then used to identify the weighting of the possible underlying density-likelihood function(s) associated with the data observations. The outcome reflects, in a probabilistic sense, what we know about the unknown parameters and possible density functions. In the case of a stochastic system *in equilibrium*, the process may be modeled as a single distribution within the CR framework. An advantage of this approach, in addition to its divergence-optimality base, is that it permits the possibility of flexible families of distributions that need not be Gaussian in nature. For discussions relative to the flexible family of distributions, under given values of moments and indirect noisy sample observations, see [3–7].

The paper is organized as follows: in Section 2 we discuss the CR family of divergence measures (DMs) and relate these DMs to the maximum likelihood (ML) principle. Given the framework developed in Section 2, Section 3 is concerned with developing a loss basis for identifying the probability space associated with data observations. In Section 4, the results of a sampling experiment are presented to illustrate finite sampling performance. Finally, in Section 5 we summarize extensions to the CR-Minimum Divergence (MD) family of estimators and provide conclusions and directions for future research.

2. Minimum Power Divergence

In identifying divergence measures that may be used as a basis for characterizing the data sampling process underlying observed data outcomes, we begin with the family of divergence measures proposed by [1–2]. In the context of a family of goodness-of-fit test statistics (see [7]), Cressie and Read (CR) proposed the following power divergence family of measures:

$$I(\mathbf{p},\mathbf{q},\gamma)=\frac{1}{\gamma(\gamma+1)}\sum_{i=1}^{n}p_i\left[\left(\frac{p_i}{q_i}\right)^{\gamma}-1\right]. \tag{1}$$

In (1), the value of indexes members of the CR family, represent the subject probability distribution, the q_i's are reference probabilities, and $\mathbf{p}$ and $\mathbf{q}$ are $n\times 1$ vectors of p_i's and q_i's, respectively. The usual probability distribution characteristics of $p_i,q_i\in[0,1]\ \forall i$, $\sum_{i=1}^{n}p_i=1$, and $\sum_{i=1}^{n}q_i=1$ are assumed to hold. The CR family of power divergences is defined through a class of additive convex functions that encompasses a broad family of test statistics, and represents *a broad family of likelihood functional relationships* within a moments-based estimation context, which will be discussed in Section 2.3. In addition, the CR measure exhibits proper convexity in **p**, for all values of and **q**, and embodies the required probability system characteristics, such as additivity and invariance with respect to a monotonic transformation of the divergence measures. In the context of extremum metrics, the general CR family of power divergence statistics represents a flexible family of pseudo-distance measures from which to derive empirical probabilities.

The CR statistic is a single index family of divergence measures that can be interpreted as encompassing a wide array of empirical goodness-of-fit and estimation criteria. As γ varies, the resulting estimators that minimize power divergence exhibit qualitatively different sampling behavior. Using data consistent empirical sample moments-constraints such as $\mathbf{h}(\mathbf{Y},\mathbf{X},\mathbf{Z};\boldsymbol{\beta})=n^{-1}\left[\mathbf{Z}'(\mathbf{Y}-\mathbf{X}\boldsymbol{\beta})\right]\xrightarrow{p}\mathbf{0}$, where $\mathbf{Y},\mathbf{X}$, and $\mathbf{Z}$ are respectively a $n\times 1$, $n\times k$, $n\times m$ vector/matrix of dependent variables, explanatory variables, and instruments, and the parameter vector $\boldsymbol{\beta}$ is the objective of estimation, a solution to the stochastic inverse problem, based on the optimized value of $I(\mathbf{p},\mathbf{q},\gamma)$, is one basis for representing a range of data sampling processes and likelihood function values.

To place the CR family of power divergence statistics in an entropy perspective, we note that there are corresponding [8–10] families of entropy functionals-divergence measures. As demonstrated by [6], over defined ranges of the divergence measures, the CR and entropy families are equivalent. Relative to [8–10], the CR family has a more convenient normalization factor $1/(\gamma(\gamma+1))$ and has proper convexity for all powers, both positive and negative. The CR family allows for separation of variables in optimization, over the range of $\gamma\in\mathbb{R}$, when the underlying variables belong to stochastically independent subsystems, called the *independent subsystems property* [6]. This separation of variables permits the partitioning of the state space and is valid for divergences in the form of a convex function.

2.1. The CR Family and Minimum Power Divergence Estimation

In a linear model context, if we use (1) as the goodness-of-fit criterion, along with moment-estimating function information, the estimation problem based on the CR divergence measure (CRDM) may, for any given choice of γ, be formulated as the following extremum-type estimator for $\boldsymbol{\beta}$**:**

$$\hat{\boldsymbol{\beta}}(\gamma)=\underset{\boldsymbol{\beta}\in\mathbf{B}}{\arg\min}\left[\min_{\mathbf{p}}\left\{I(\mathbf{p},\mathbf{q},\gamma)\mid\sum_{i=1}^{n}p_i\mathbf{Z}'_{i.}\left(Y_i-\mathbf{X}_{i.}\boldsymbol{\beta}\right)=\mathbf{0},\sum_{i=1}^{n}p_i=1,p_i\geq 0\,\forall i\right\}\right], \tag{2}$$

where $\mathbf{q}$ is taken as given, and $\mathbf{B}$ denotes the appropriate parameter space for $\boldsymbol{\beta}$. Note in (2) that $\mathbf{X}_{i,}$ and $\mathbf{Z}_{i.}$ denote the i^{th} rows of $\mathbf{X}$ and $\mathbf{Z}$, respectively. This class of estimation procedures is referred to as *Minimum Power Divergence (MPD)* estimation and additional details of the solution to this stochastic inverse problem are provided in the sections ahead. The MPD optimization problem may be represented as a two-step process. In particular, one can first optimize with respect to the choice of the sample probabilities, $\mathbf{p}$, and then optimize with respect to the structural parameters $\boldsymbol{\beta}$, for any choice of the CR family of divergence measures identified by the choice of γ, given $\mathbf{q}$.

It is important to note that the *family* of power divergence statistics defined by (1), is symmetric in the choice of which set of probabilities are considered as the subject and reference distribution arguments of the function (2). In particular, as noted by [11–12], whether the statistic is designated as $I(\mathbf{p},\mathbf{q},\gamma)$ or $I(\mathbf{q},\mathbf{p},\gamma)$, *the same collection* of members of the family of divergence measures are ultimately spanned, when considering all of the possibilities for $\gamma\in(-\infty,\infty)$

2.2. Popular Variants of $I(\mathbf{p},\mathbf{q},\gamma)$

Three discrete CR alternatives for $I(\mathbf{p},\mathbf{q},\gamma)$, where $\gamma\in\{-1,0,1\}$, have received the most attention in the literature, and to our knowledge these are the only variants that have been utilized empirically to date. In reviewing these, we adopt the notation $\text{CR}(\gamma)\equiv I(\mathbf{p},\mathbf{q},\gamma)$, where the arguments $\mathbf{p}$ and $\mathbf{q}$ are tacitly understood to be evaluated at relevant vector values. In the two special cases where $\gamma=0$ or -1, $\text{CR}(0)$ and $\text{CR}(-1)$ are to be interpreted as the continuous limits, $\lim_{\gamma\to 0}\text{CR}(\gamma)$, and $\lim_{\gamma\to -1}\text{CR}(\gamma)$, respectively.

If we let $\mathbf{q}=n^{-1}\mathbf{1}_n$, the reference distribution is the empirical distribution function (EDF) associated with the observed sample data, and also the nonparametric maximum likelihood estimate of the data sampling distribution. Minimizing $\text{CR}(-1)$ is then equivalent to maximizing $\sum_{i=1}^{n}\ln(p_i)$ and leads to the traditional maximum empirical log-likelihood (MEL) objective function. Minimizing $\text{CR}(0)$ is equivalent to maximizing $-\sum_{i=1}^{n}p_i\ln(p_i)$, and leads to the maximum empirical exponential likelihood (MEEL) objective function, which is also equivalent to [13] entropy. Finally, minimizing $\text{CR}(1)$ is equivalent to maximizing $-\frac{n}{2}\sum_{i=1}^{n}\left(p_i^2-\frac{1}{n}\right)$, and leads to the maximum log-Euclidean likelihood (MLEL) objective function. Note the latter objective function is also equivalent to minimizing the sum of squares function $\left(\mathbf{p}-n^{-1}\mathbf{1}_n\right)'\left(\mathbf{p}-n^{-1}\mathbf{1}_n\right)$.

With regard to MPD ($CR(\gamma)$ family) estimators, under the usual assumed regularity conditions, all of the MPD estimators of $\boldsymbol{\beta}$ obtained by optimizing the are consistent and asymptotically normally distributed. They are also asymptotically efficient, relative to the optimal estimating function (OptEF) estimator [14], when a uniform distribution, or equivalently the empirical distribution function (EDF), is used for the reference distribution. The solution to the constrained optimization problem yields optimal estimates, $\hat{\mathbf{p}}(\gamma)$ and $\hat{\boldsymbol{\beta}}(\gamma)$, that cannot, in general, be expressed in closed form, and thus must be obtained using numerical methods.

2.3. Relating Minimum Power Divergence to Maximum Likelihood

The objectives of minimizing power divergence and maximizing likelihood are generally not equivalent. However, two of the historical variants presented in the preceding section have direct conceptual linkages to maximum likelihood concepts, and the third is an analog to least squares. The traditional MEL criterion $\mathrm{CR}(-1)$ coincides with the estimation objective of maximizing the joint empirical log likelihood, $\sum_{i=1}^{n}\ln(p_i)$, conditional on moment constraints, $E_{\mathbf{p}}(\mathbf{h}(\mathbf{X},\theta))=\mathbf{0}$, where $E_{\mathbf{p}}(\mathbb{R})$ denotes an expectation taken with respect to the empirical probability distribution defined by $\mathbf{p}$[15], [7]. In the sense of objective function analogies, the choice of $\gamma=-1$ defines an empirical analog to the classical maximum likelihood approach, except that no explicit functional form for the likelihood function is assumed known or specified at the outset of the estimation problem.

The $\mathrm{CR}(0)$ criterion of minimizing $\sum_{i=1}^{n}p_i\ln(p_i)$ is equivalent to minimizing the Kullback-Leibler (KL) information criterion defined by $\sum_{i=1}^{n}p_i\ln(p_i/n^{-1})=\sum_{i=1}^{n}p_i\ln(p_i)+\ln(n)$, where the reference distribution, $\mathbf{q}$, is specified to be the EDF, or uniform distribution, supported on the data observations [16]. Interpreting the estimation problem in the KL context, the estimation objective is to find the feasible probability distribution, $\mathbf{p}$, that defines the minimum value of all possible *expected log-likelihood ratios*, $E_{\mathbf{p}}\ln\left(\frac{p}{n^{-1}}\right)$, subject to any imposed moment constraints. The expectation of the log-likelihood ratio has the restricted (by any moment constraints) likelihood in the numerator (i.e., the solved p_i's), and the unrestricted empirical distribution function (i.e., the uniform distribution) likelihood in the denominator.

The $\mathrm{CR}(1)$ solution seeks the empirical probability distribution, $\mathbf{p}$, that minimizes the Euclidean distance of $\mathbf{p}$ from the EDF (uniform distribution), or equivalently, that minimizes the square of the Euclidean distance, $(\mathbf{p}-n^{-1}\mathbf{1}_n)'(\mathbf{p}-n^{-1}\mathbf{1}_n)$. This estimation objective is effectively the least squares fit of the probability weights, $\mathbf{p}$, to the empirical distribution function, $n^{-1}\mathbf{1}_n$, subject to the moment constraints $E_{\mathbf{p}}(\mathbf{h}(\mathbf{X},\theta))=\mathbf{0}$, where θ denotes whatever vector of parameters the moment conditions depend on.

More generally, minimizing power divergence (1), with $q_i=n^{-1}$, can be interpreted as minimizing the empirical expectation of the γ-power of the likelihood ratio. Given the adding up condition, $\sum_{i=1}^{n}p_i=1$, the objective function is equivalent to $E_{\mathbf{p}}\left(\left(\frac{p}{n^{-1}}\right)^{\gamma}\right)\equiv\sum_{i=1}^{n}p_i\left(\frac{p_i}{n^{-1}}\right)^{\gamma}$.

Because the likelihood function and the sample space are inexplicably linked, it would be useful, given a sample of indirect noisy observations and corresponding moment conditions, to have an optimum choice of a member of the CR family. It is typical in applied statistics, given a sample of data and corresponding moment conditions, that there is ambiguity-uncertainty regarding the choice of likelihood function.

3. Identifying the Probability Space

Given the CR family of divergence measures (1), indirect noisy data and linear functionals in the form of estimating equations-moments, the next question concerns how to go about identifying the underlying probability distribution function-probability space of a system or process. Since the data and the moments are directly linked, the divergence- measures permits us to exploit the statistical machinery of information theory to gain insights into the PDF behavior of stochastic systems and processes. The likelihood functionals-divergences have a natural interpretation in terms of uncertainty and measures of distance. Many formulations have been proposed for a proper selection of the probability space, but their applicability depends on characteristics of the data, such as stationarity of the noise process. In the sections ahead we make use of the CR family of divergence measures to choose the optimal probability system under quadratic loss.

3.1. Distance–Divergence Measures

In Section 2, we used the CR power divergence measure (1) to define, as γ takes on different values, a family of likelihood function relationships. Given this family, we follow [17–18], and consider a parametric family of concave entropy-likelihood functions, which satisfy additivity and trace conditions. Using the CR divergence measures, this parametric family is essentially the linear convex combination of the cases where $\gamma = 0$ and $\gamma = -1$. This family is tractable analytically and provides a basis for joining (combining) statistically independent subsystems. When the base measure of the reference distribution **q** is taken to be a uniform probability density function (PDF), we arrive at a family of additive convex functions. In this context, one is effectively considering the convex combination of the MEL and maximum empirical exponential likelihood (MEEL) measures. From the standpoint of extremum-minimization with respect to **p**, the generalized divergence family reduces to:

$$S_{\alpha}^{*}(q) = \sum_{i=1}^{n}\left((1-\alpha)p_i \ln\left(p_i / q_i\right) - \alpha q_i \ln p_i\right). \tag{3}$$

In the limit, as $\alpha \to 0$, the minimum *KL* divergence *I*(**p** ||**q**) of the probability mass function **p,** with respect to **q**, is recovered. As $\alpha \to 1$, the **q**-weighted MEL stochastic inverse problem *I*(**q**||**p)** results. *This generalized family of divergence measures permits a broadening of the canonical distribution functions and provides a framework for developing a loss-minimizing estimation rule.* In an extremum estimation context, when α = 1/2, this results in what is known in the literature as Jeffrey's *J*-divergence [19]. In this case, the full objective function, *J*-divergence *J*(**p**||**q**)=*I*(**p**||**q**) + *I*(**q**||**p**), is a convex combination of *KL* divergence *I*(**p**||**q**) and the reverse *KL* divergence *I*(**q**||**p**). In line with the complex nature of the problem, in the sections to follow, we demonstrate a convex estimation rule, which seeks to choose among MPD-type estimators to minimize quadratic risk (QR).

3.2. A Minimum Quadratic Risk (QR) Estimation Rule

To choose an estimation rule, we use the well-known squared error-quadratic loss criterion and associated QR function to make optimal use of a given set of discrete alternatives for the CR goodness-of-fit measures and associated estimators for β. In choosing an estimation rule, the objective is to define the convex combination of a set of estimators for β that minimizes QR, where each estimator is defined by the solution to the extremum problem:

$$\hat{\beta}(\gamma) = \arg\max_{\beta \in \mathbf{B}} \left[\max_{\mathbf{p}} \left\{ -I(\mathbf{p},\mathbf{q},\gamma) \mid \sum_{i=1}^{n} p_i \mathbf{Z}'_{i.} (Y_i - \mathbf{X}_{i.}\beta) = \mathbf{0}, \sum_{i=1}^{n} p_i = 1, p_i \geq 0 \, \forall i \right\} \right]. \tag{4}$$

The squared error loss function is defined by $\ell(\hat{\beta},\beta) = (\hat{\beta}-\beta)'(\hat{\beta}-\beta)$ and has the corresponding QR function given by:

$$\rho(\hat{\beta},\beta) = \mathrm{E}\left[\ell(\hat{\beta},\beta)\right] = \mathrm{E}\left[(\hat{\beta}-\beta)'(\hat{\beta}-\beta)\right]. \tag{5}$$

The convex combination of estimators is defined by:

$$\bar{\beta}(\alpha) = \sum_{j=1}^{J} \alpha_j \hat{\beta}(\gamma_j), \text{ where } \alpha_j \geq 0 \ \forall j, \text{ and } \sum_{j=1}^{J} \alpha_j = 1. \tag{6}$$

Given (6) the optimum use of the discrete alternatives under QR is determined by choosing the particular convex combination of the estimators that minimizes QR, as:

$$\bar{\beta}(\hat{\alpha}) = \sum_{j=1}^{J} \hat{\alpha}_j \hat{\beta}(\gamma_j), \text{ where } \hat{\alpha} = \arg\min_{\alpha \in CH} \left\{ \rho(\bar{\beta}(\alpha),\beta) \right\}, \tag{7}$$

and *CH* denotes the *J*-dimensional convex hull of possibilities for the $J \times 1$ α vector, defined by the non-negativity and adding-up conditions represented in (6). This represents, in a loss context, an appropriate choice of the γ value in the definition of the CR power divergence criterion.

3.3. The Case of Two CR Alternatives

As an example, consider the case where there are two discrete alternative CR measures of interest. In this context, the objective is to make optimal use of the information contained in the two associated estimators of β, $\hat{\beta}(\gamma_1)$ and $\hat{\beta}(\gamma_2)$. The corresponding QR function may be written as:

$$\rho(\bar{\beta}(\alpha),\beta) = \\ \mathrm{E}\left[\left[\alpha(\hat{\beta}(\gamma_1)-\beta) + (1-\alpha)(\hat{\beta}(\gamma_2)-\beta) \right]' \left[\alpha(\hat{\beta}(\gamma_1)-\beta) + (1-\alpha)(\hat{\beta}(\gamma_2)-\beta) \right] \right], \tag{8}$$

and can be represented in terms of the QR functions of $\hat{\beta}(\gamma_1)$ and $\hat{\beta}(\gamma_2)$ as:

$$\rho(\bar{\beta}(\alpha),\beta) = \\ \alpha^2 \rho(\hat{\beta}(\gamma_1),\beta) + (1-\alpha)^2 \rho(\hat{\beta}(\gamma_2),\beta) + 2\alpha(1-\alpha)\mathrm{E}\left[(\hat{\beta}(\gamma_1)-\beta)' (\hat{\beta}(\gamma_2)-\beta) \right] \tag{9}$$

To minimize $\rho(\bar{\beta}(\alpha),\beta)$, the first-order condition, with respect to α, is given by:

$$\frac{d\rho\left(\bar{\beta}(\alpha),\beta\right)}{d\alpha}=$$
$$2\alpha\rho\left(\hat{\beta}(\gamma_1),\beta\right)-2(1-\alpha)\rho\left(\hat{\beta}(\gamma_2),\beta\right)+2(1-2\alpha)\mathrm{E}\left[\left(\hat{\beta}(\gamma_1)-\beta\right)'\left(\hat{\beta}(\gamma_2)-\beta\right)\right]=0. \tag{10}$$

Solving for the optimal value of α yields:

$$\hat{\alpha}=\frac{\rho\left(\hat{\beta}(\gamma_2),\beta\right)-\mathrm{E}\left[\left(\hat{\beta}(\gamma_1)-\beta\right)'\left(\hat{\beta}(\gamma_2)-\beta\right)\right]}{\rho\left(\hat{\beta}(\gamma_1),\beta\right)+\rho\left(\hat{\beta}(\gamma_2),\beta\right)-2\mathrm{E}\left[\left(\hat{\beta}(\gamma_1)-\beta\right)'\left(\hat{\beta}(\gamma_2)-\beta\right)\right]}, \tag{11}$$

and the optimal convex-combined estimator is defined as:

$$\bar{\beta}(\hat{\alpha})=\hat{\alpha}\hat{\beta}(\gamma_1)+(1-\hat{\alpha})\hat{\beta}(\gamma_2). \tag{12}$$

By construction, $\bar{\beta}(\hat{\alpha})$ is QR superior to either $\hat{\beta}(\gamma_1)$ or $\hat{\beta}(\gamma_2)$, unless the optimal convex combination resides at one of the boundaries for α, or the two estimators have identical risks and $\mathrm{E}\left[\left(\hat{\beta}(\gamma_1)-\beta\right)'\left(\hat{\beta}(\gamma_2)-\beta\right)\right]=0$. In any case, QR-wise, the resulting estimator $\bar{\beta}(\hat{\alpha})$ is no worse than either $\hat{\beta}(\gamma_1)$ or $\hat{\beta}(\gamma_2)$.

3.4. Empirical Calculation of α

To implement the optimal convex combination of estimators empirically, a value for $\hat{\alpha}$ in (12) is needed. The calculation of the exact $\hat{\alpha}$ value in (12) requires unknown parameters as well as unknown probability distributions. Thus, one must seek an estimate of $\hat{\alpha}$ based on sample observations. Working toward a useful estimate for $\hat{\alpha}$, note that:

$$\begin{aligned}&\mathrm{E}\left[\left(\hat{\beta}(\gamma_1)-\hat{\beta}(\gamma_2)\right)'\left(\hat{\beta}(\gamma_1)-\hat{\beta}(\gamma_2)\right)\right]\\&=\mathrm{E}\left[\left(\hat{\beta}(\gamma_1)-\beta\right)'\left(\hat{\beta}(\gamma_1)-\beta\right)\right]+\mathrm{E}\left[\left(\hat{\beta}(\gamma_2)-\beta\right)'\left(\hat{\beta}(\gamma_2)-\beta\right)\right]-2\mathrm{E}\left[\left(\hat{\beta}(\gamma_1)-\beta\right)'\left(\hat{\beta}(\gamma_2)-\beta\right)\right]\\&=\rho\left(\hat{\beta}(\gamma_1),\beta\right)+\rho\left(\hat{\beta}(\gamma_2),\beta\right)-2\mathrm{E}\left[\left(\hat{\beta}(\gamma_1)-\beta\right)'\left(\hat{\beta}(\gamma_2)-\beta\right)\right].\end{aligned} \tag{13}$$

Thus, an unbiased estimate of the denominator term in (11) is given directly by calculating $\left(\hat{\beta}(\gamma_1)-\hat{\beta}(\gamma_2)\right)'\left(\hat{\beta}(\gamma_1)-\hat{\beta}(\gamma_2)\right)$. Given the consistency of both estimators, this value would consistently estimate the denominator of (11) as well.

Deriving an estimate of the numerator in (11) is challenging since in general, the estimators $\hat{\beta}(\gamma)$ are biased. Thus, neither the risk term nor the subtracted expectation of the cross-product term in (11) can be simplified. This complication persists, even if the estimators are calculated from independent samples. Under independence, the risk function $\rho\left(\hat{\beta}(\gamma_2),\beta\right)$ does not merely simplify to a function of variances as in [7]. The term $E\left[\left(\hat{\beta}(\gamma_1)-\beta\right)'\left(\hat{\beta}(\gamma_2)-\beta\right)\right]$ remains nonzero and, in fact, is equal to a cross product of bias vectors, $bias\left(\hat{\beta}(\gamma_1)\right)' bias\left(\hat{\beta}(\gamma_2)\right)$. However, making the usual assumption that the moment conditions are correctly specified, the $\hat{\beta}(\gamma)$ estimators are consistent under regularity conditions no more stringent than the usual conditions imposed to obtain consistency in the generalized methods of moments context or in classical linear models. Thus, as an approximation, one might ignore the bias terms because they converge to zero as *n* increases.

Ignoring the bias terms, and assuming the estimators $\hat{\beta}(\gamma_1)$ and $\hat{\beta}(\gamma_2)$ are based on two independent samples of data, the expression for the optimal α simplifies to the following:

$$\hat{\alpha} = \frac{tr\left(Cov\left(\hat{\beta}(\gamma_2)\right)\right)}{tr\left(Cov\left(\hat{\beta}(\gamma_1)\right)\right) + tr\left(Cov\left(\hat{\beta}(\gamma_2)\right)\right)}. \tag{14}$$

In effect, the use of this $\hat{\alpha}$ in forming a convex combination of the two estimators can be viewed as pursuing an objective of minimizing the variation in the resultant estimator (12). If we make use of the optimum α in the optimal convex estimator in (12), the result comes out in the form of a Stein-like estimator [20-21], where for a given samples of data, shrinkage is from $\hat{\beta}(\gamma_2)$ to $\hat{\beta}(\gamma_1)$.The level of shrinkage is determined by the relative bias-variance tradeoff.

A question that remains is the finite sampling performance of the estimators based on the estimated value $\hat{\alpha}$. To provide some perspective on the answer to this question, in the next section, we present the results of a sampling experiment that implements (14), in choosing a convex combination of the estimators $\hat{\beta}(-1)$ and $\hat{\beta}(0)$. The objective is to define a new estimator via a combination of both estimators that is superior to either in terms of quadratic loss.

4. Finite Sample Performance

To illustrate finite sample performance of a convex combination of $\hat{\beta}(-1)$ and $\hat{\beta}(0)$, we follow [7] and consider a simple data sampling process involving an instrumental variable model similar to that used by [22]. The sampling model is:

$$\begin{aligned} y_i &= x_i\beta + \varepsilon_i \\ x_i &= \mathbf{z}_{i.}\delta + v_i, \quad i = 1,\ldots,n \end{aligned} \tag{15}$$

where y_i denotes outcomes of the variable of interest, x_i denotes outcomes of a scalar endogenous regressor, $\mathbf{z}_{i.}$ denotes a 1×2 row vector of instrumental variable outcomes, and *n* denotes sample size. In the sampling experiment, the value of β is set equal to 1, and $\mathbf{z}_{i.}$ and $(\varepsilon_i, v_i)'$ are independent and *iid* outcomes with probability distributions, $N(\mathbf{0},\mathbf{I}_2)$, and, $N\left(\begin{bmatrix}0\\0\end{bmatrix},\begin{bmatrix}1 & \tau\\ \tau & 1\end{bmatrix}\right)$, respectively. The theoretical first stage R^2 is given by $R^2 = \frac{\delta'\delta}{\delta'\delta+1}$, and we let, $\delta = \begin{bmatrix}\xi\\ \xi\end{bmatrix}$, so that $R^2 = \frac{2\xi^2}{2\xi^2+1}$. In this sample design the value of τ determines the degree of endogeneity, and R^2 determines the strength of the instruments $\mathbf{z}_{i.}$ for x_i, with $\xi = \left(\frac{R^2}{2\left(1-R^2\right)}\right)^{1/2}$. We examine sample of sizes $n = 100$ *and* 250, with $\tau = 0.5$ and $R^2 = 0.5$.

The covariance matrices used to implement the optimal convex combination weight in (14) (which are variances in this case because the $\hat{\beta}(\gamma)s$ are scalars), are of the form [15]:

$$\hat{\text{var}}\left(\hat{\beta}(\gamma)\right) = \left(\left[\sum_{i=1}^{n} \hat{p}_i(\gamma) x_i \mathbf{z}_{i.}\right]\left[\sum_{i=1}^{n} \hat{p}_i(\gamma)\left(y_i - x_i\hat{\beta}(\gamma)\right)^2 \mathbf{z}'_{i.}\mathbf{z}_{i.}\right]^{-1}\left[\sum_{i=1}^{n} \hat{p}_i(\gamma) x_i \mathbf{z}'_{i.}\right]\right)^{-1} \tag{16}$$

where the $\hat{p}_i(\gamma)s$ are the data or probability weights calculated in the solution to the estimation problem, when either $\gamma = -1$ *or* $\gamma = 0$. The calculated convex weight (14) simplifies to $\hat{\alpha} = \frac{\text{var}\left(\hat{\beta}(0)\right)}{\text{var}\left(\hat{\beta}(-1)\right) + \text{var}\left(\hat{\beta}(0)\right)}$, and the convex combination estimator is given by $\overline{\beta}(\alpha) = \alpha\beta(-1) + (1-\alpha)\beta(0)$.

The results for the sampling experiment are presented in Table 1. It is evident that, across all scenarios, the convex combination of $\hat{\beta}(-1)$ and $\hat{\beta}(0)$ estimators is substantially superior, under quadratic loss, to either of the individual estimators. It is also evident that the risks of the individual estimators are quite close in magnitude to one another across all scenarios. As expected the MEL $(\gamma = -1)$ estimator is generally slightly better than the estimator based on the Kullback-Leibler$(\gamma = 0)$ distance measure for the larger sample size of 250, but not uniformly for the smaller sample size of 100. Given the similarity in mean squared error (MSE) performance, it is not surprising that the optimal $\alpha's$ used in forming the convex combination had an average value of .5, which is consistent with the Kullback-Leibler balanced J-divergence. As the degree of endogeneity increases (*i.e.*, when τ increases) and the effectiveness of the instruments decreases (*i.e.*, when R^2 decreases), the QR of all of the estimators increases, but the overall performance of the estimators, and especially the convex combination estimator, remains very good.

Table1. MSE Results for Convex Combinations of $\hat{\beta}(-1)$ and $\hat{\beta}(0)$.

Scenario $\{\text{n}, \tau, \text{R}^2\}$	$\text{MSE}\left(\hat{\beta}(-1)\right)$	$\text{MSE}\left(\hat{\beta}(0)\right)$	$\bar{\hat{\alpha}}(\gamma = -1)$	$\text{std}(\hat{\alpha})$	$\text{MSE}\left(\overline{\beta}(\hat{\alpha})\right)$
100,0.25,0.75	0.00343	0.00364	0.49712	0.29082	0.00180
100,0.5,0.5	0.01129	0.01113	0.49996	0.07340	0.00528
100,0.75,0.25	0.03801	0.03159	0.48670	0.30753	0.02105
250,0.25,0.75	0.00122	0.00136	0.49978	0.01591	0.00062
250,0.5,0.5	0.00437	0.00452	0.50158	0.02813	0.00219
250,0.75,0.25	0.01309	0.01323	0.50031	0.07018	0.00639

5. Concluding Remarks

In this paper, we have suggested estimation possibilities not accessible by considering individual members of the CR family. This was achieved by taking a convex combination of estimators associated with two members of the CR family, under minimum expected quadratic loss. The sampling experiments reported illustrate superior finite sample performance of the resulting convex estimation rules. In particular, we recognize that relevant statistical distributions underlying data sampling processes that result from solving MPD-estimation problems may not always be well described by popular integer choices for the index value in the CR divergence measure family. Building on the problem of identifying the probability space that is noted in Section 3, we demonstrate that it is possible to derive a one-parameter family of appropriate likelihood function relationships to describe statistical distributions. One possibility for the one-parameter family is essentially a convex combination of the CR integer functionals, and $\gamma \to -1$. This one-parameter family of additive-trace form of CR divergence functions leads to an additional rich set of possibly non-Gaussian distributions that broadens the set of probability distributions that can be derived from the CR power divergence family. With this new flexible family, one can develop a new family of estimators and probability distributions.

The methodology introduced in this paper is very general, and there is no reason to focus exclusively on combinations of only the estimators associated with $\gamma \to 0$ and $\gamma \to -1$. Other choices of γ can be considered in forming combinations, and there is the interesting question for future work regarding the most useful intial choices of γ to consider when combining estimators from the CR family. Moreover, there is also no reason to limit combinations to only two alternative CR estimators, and combinations of three or more estimators could be considered and possibly lead to even greater gains in estimating efficiency.

Looking ahead we note that physical and behavioral processes and systems are seldom in equilibrium, and new methods of modeling and information recovery are needed to explain the hidden dynamic world of interest and understand the dynamic systems that produce the indirect noisy effects data that we observe. The information theoretic methods presented in this paper represent a basis for modeling and information recovery for systems in disequilibrium and provide a framework for capturing temporal-causal information.

Acknowledgements

It is a pleasure to acknowledge the helpful comments of M. George, J. Geweke, A. Gorban, M. Grendar, J. Lee, H. Pesaran, P. Phillips, A. Plastino, S. Berto Villas-Boas and A. Zaman. We also thank two anonymous journal reviewers for their helpful comments and insights. Of course they are not to be held responsible for the current set of words and symbols.

References

1. Cressie, N.; Read, T.; Multinomial goodness of fit tests. *J. R. Stat. Soc.* **1984**, *B46*, 440–464.
2. Read, T.R.; Cressie, N.A. *Goodness of Fit Statistics for Discrete Multivariate Data*; Springer Verlag: New York, NY, USA, 1988.
3. Bjelakovic, I.; Dueschel, J.; Kruger, T.; Seiler, R.; Schultze, R.; Szkola, A. Typical Support and Sanov Large Deviations of Correlated States. *Commun. Math. Phys.* **2008**, *279*, 559–584.
4. Ojima, I.; Okamura, K. *Large Deviation Strategy for Inverse Problems*; Kyoto Institute, Kyoto University: Kyoto, Japan, 2011.

5. Hanel, R.; Thurner, S. *A Comprehensive Classification of Complex Statistical System and Distribution Functions*; Sante Fe Institute: Sante Fe, NM, USA, 2007.
6. Gorban, A.; Gorban, P.; Judge, G. Entropy: The Markov Ordering Approach. *Entropy* **2010**, *5*, 1145–1193.
7. Judge, G.G.; Mittelhammer, R.C. *An Information Theoretic Approach to Econometrics*; Cambridge University Press: Cambridge, UK, 2012.
8. Renyi, A. On measures of entropy and information. In *Proceedings of the 4th Berkeley Symposium on Mathematics, Statistics and Probability*, Berkeley, CA, USA, 20 June–30 July 1960; University of California Press: Berkeley, CA, USA, 1961; pp. 547–561.
9. Renyi, A. *Probability Theory*; North-Holland: Amsterdam, The Netherlands, 1970.
10. Tsallis, C. Possible Generalization of Boltzmann-Gibbs Statistics. *J. Stat. Phys.* **1988**, *52*, 479–487.
11. Osterreicher, F.; *Csiszar's f-Divergencies-Basic Properties*. Institute of Mathematics, University of Salzburg: Salzburg, Austria, 2002. Available online: http://www.unisalzburg.at/pls/portal/docs/1/246178.PDF, accessed on 21 October 2012.
12. Osterreicher, F.; Vajda, I. A New Class of Metric Divergences on Probability Spaces and its Applicability in Statistics. *Ann. Inst. Stat. Math.* **2003**, *55*, 639–653.
13. Shannon, C.E. A mathematical theory of communication. *Bell System Tech. J.* **1948**, *27*, 379–423.
14. Baggerly, K.A. Empirical likelihood as a goodness of fit measure. *Biometrika* **1998**, *85*, 535–547.
15. Mittelhammer, R.M.; Judge, G.G.; Miller, D.J. *Econometrics Foundations*; Cambridge University Press: New York, NY, USA, 2000.
16. Kullback, S.; Leibler, R.A. On information and sufficiency. *Ann. Math. Stat.* **1951**, *22*, 79–86.
17. Gorban, A. *Equilibrium Encircling Equations of Chemical Kinetics and Their Thermodynamic Limit.* Nauka: Novosibirsk, Russia, 1984.
18. Gorban, A.; Karlin, I.V. Family of Additive Entropy Functions Out of the Thermodynamic Limit. *Phys. Rev. E* **2003**, *67*, 016104.
19. Grendar, M.; Grendar, M. On the Probabilistic Rationale of *I* Divergence and *J* Divergence Minimization. **2000**, arXiv:math/0008037.
20. James, W.; Stein, C. Estimation with Quadratic Loss. In *Proceedings of Fourth Berkeley Symposium on Statistics and Probability*, University of California Press: Berkeley, CA, USA, 1961; pp. 361–379.
21. Judge, G.; Bock, M.E. *The Statistical Implication of Pre-Test and Stein-Rule Estimators*; North Holland: Amsterdam, The Netherlands, 1978.
22. Hahn, J.; Hausman, J. Notes on Bias in Estimators for Simultaneous Equation Models. *Econ. Lett.* **2002**, *75*, 237–241.