Proc. Indian Acad. Sci. (Math. Sci.) (2020) 130:42
https://doi.org/10.1007/s12044-020-00581-9

On foundation of statistical inference by C R Rao relating to information inequality

DIPAK K DEY[1,*] and GYUHYEONG GOH[2]

[1]Department of Statistics, University of Connecticut, Storrs, CT, USA
[2]Department of Statistics, Kansas State University, Manhattan, KS, USA
*Corresponding author.
E-mail: dipak.dey@uconn.edu

Abstract. This is the birth centenary year of the living legend and giant in the world of statistics, Prof. Calyampudi Radhakrishna (C R) Rao. This article is a partial reflection of Dr. Rao's contributions to statistical theory and methodology, including unbiased estimation, variance reduction by sufficiency, efficiency of estimation, information geometry, as well as the application of matrix theory in linear statistical inference. His contributions are so massive, it will be impossible to review all of them. This article will comprise mostly his major discovery, which is Cramér–Rao lower bound and related applications and extensions.

Keywords. Bregman divergence; information inequality; generalizations of Cramér–Rao lower bound.

2010 Mathematics Subject Classification. 62B10, 62F10.

1. Introduction

This article focuses on one of the major contributions of Calyampudi Radhakrishna (C R) Rao to the foundations of statistical inference. In 1945, C R Rao wrote a path breaking paper entitled "Information and accuracy attainable in the estimation of statistical parameters [11]." In the paper, he obtained a lower bound, also known as Cramér–Rao lower bound or CRLB, to the variance of an unbiased estimator. Over the years, his original work has been extended in many directions. For example, Chapman and Robbins [4] have derived a generalization of CRLB, which is free from regularity conditions. In Ghosh [5] and Gill and Levit [6], Bayesian counterparts of CRLB have been developed. In addition, sequential Cramer–Rao type integral inequalities have been derived by Prakasa Rao [9,10]. However, most of the generalizations of CRLB have been limited to lower bounds of the variance or the expected quadratic loss.

The objective of this paper is to obtain a broader and more comprehensive generalization of CRLB using a large class of loss functions, called the Bregman divergence. In statistics, divergence is defined as a measure of dissimilarity between vectors, between functions, or between probability distributions. The formal definition of divergence is similar to a metric (or the distance function) but does not require the symmetry and triangle inequality

This article is part of the "Special Issue in Honour of Professor C R Rao on His Birth Centenary".

Published online: 25 July 2020

properties. The Bregman divergence is a general class of divergences, which includes the quadratic loss, the Kullback–Leibler divergence, the Itakura–Saito distance, and the Mahalanobis distance as a special case.

The rest of this paper is organized as follows. Section 2 introduces the definition and properties of Bregman divergence. In Section 3, we propose a new generalization of CRLB using Bregman divergence. Section 4 illustrates that the proposed lower bound embraces many generalizations of CRLB as a special case. In Section 5, we conclude the paper with some remarks.

2. Bregman divergence and properties

Over the past 15 years, the application of Bregman divergence to machine learning has received increasing attention. The Bregman divergence was originally introduced by Bregman [3], while the growing popularity was triggered by the introduction of the Bregman clustering algorithm [1]. Despite its high impact on machine learning research, the Bregman divergence remains unknown to many people in statistics and mathematical sciences. In this section, we introduce a formal definition of Bregman divergence and some useful properties.

Let $\phi : \Omega \to \mathbb{R}$ be a strictly convex and differentiable function on a nonempty convex set $\Omega \subseteq \mathbb{R}^m$. Then the Bregman divergence between two vectors x and y with respect to ϕ is defined as

$$\mathrm{BD}_\phi(x, y) = \phi(x) - \phi(y) - (x - y)^\top \nabla\phi(y), \tag{1}$$

where $\nabla\phi$ is the gradient vector of ϕ. From (1), the Bregman divergence can be interpreted as the error term for the first-order Taylor approximation of $\phi(x)$ at the point y. The Bregman divergence is a *divergence* since it satisfies the following two conditions:

$$\mathrm{BD}_\phi(x, y) \geq 0 \quad \text{(non-negativity)};$$
$$\mathrm{BD}_\phi(x, y) = 0 \Leftrightarrow x = y \quad \text{(identity of indiscernibles)}.$$

However, the Bregman divergence is not a metric (or a distance) because it does not satisfies $\mathrm{BD}_\phi(x, y) = \mathrm{BD}_\phi(y, x)$ (symmetry) and $\mathrm{BD}_\phi(x, y) \leq \mathrm{BD}_\phi(x, z) + \mathrm{BD}_\phi(z, y)$ (triangle inequality). The Bregman divergence in (1) reduces to a well-known divergence measure (or a loss function) for the choice of a convex function ϕ. For example, when we set $\phi(x) = x^\top Ax$ for a positive definite matrix A, the Bregman divergence becomes the squared Mahalanobis distance as follows:

$$\mathrm{BD}_\phi(x, y) = x^\top Ax - y^\top Ay - (x - y)^\top (2Ay) = (x - y)^\top A(x - y).$$

See Table 1 for more examples. We now introduce some interesting properties of Bregman divergences discussed by Banerjee et al. [1].

(1) *Non-negativity*. $\mathrm{BD}_\phi(x, y) \geq 0$, where the equality holds if and only if $x = y$.
(2) *Convexity*. $\mathrm{BD}_\phi(x, y)$ is always convex in x, but not necessarily in y.
(3) *Linearity*. $\mathrm{BD}_{c_1\phi_1 + c_2\phi_2}(x, y) = c_1\mathrm{BD}_{\phi_1}(x, y) + c_2\mathrm{BD}_{\phi_2}(x, y)$ for $c_1, c_2 \geq 0$, where ϕ_1 and ϕ_2 are strictly convex and differentiable.

Table 1. Examples of Bregman divergence

$\phi(x)$	Bregman divergence	Loss function
$\|\|x\|\|^2$	$\|\|x-y\|\|^2$	Squared error loss
$\sum_{i=1}^m x_i \log x_i$	$\sum_{i=1}^m \left\{x_i \log\left(\frac{x_i}{y_i}\right) - x_i + y_i\right\}$	Kullback–Leibler divergence
$\sum_{i=1}^m -\log x_i$	$\sum_{i=1}^m \left\{\frac{x_i}{y_i} - \log\left(\frac{x_i}{y_i}\right) - 1\right\}$	Itakura–Saito distance
$\sum_{i=1}^m \mathrm{e}^{cx_i}$	$\sum_{i=1}^m \{\mathrm{e}^{cy_i}(\mathrm{e}^{c(x_i-y_i)} - c(x_i - y_i) - 1)\}$	Weighted Linex loss

(4) *Equivalence classes*. If $\phi(x) = \phi^*(x) + b^\top x + c$ for $b \in \mathbb{R}^m$ and $c \in \mathbb{R}$, then $\mathrm{BD}_\phi(x, y) = \mathrm{BD}_{\phi^*}(x, y)$. Therefore, the set of strictly convex functions can be partitioned into equivalence classes such that $[\phi^*] = \{\phi : \mathrm{BD}_\phi(x, y) = \mathrm{BD}_{\phi^*}(x, y)\}$.
(5) *Dual divergences.* Let ϕ be a Legendre function and $\tilde{\phi}$ be its conjugate, then $\mathrm{BD}_\phi(x, y) = \mathrm{BD}_{\tilde{\phi}}(\tilde{y}, \tilde{x})$, where $\tilde{x}$ and $\tilde{y}$ are respectively obtained by the Legendre transformation of x and y.
(6) *Generalized Pythagorean theorem.* The Bregman divergence satisfies the following equation: $\mathrm{BD}_\phi(x, z) = \mathrm{BD}_\phi(x, y) + \mathrm{BD}_\phi(y, z) - (x-y)^\top \{\nabla\phi(z) - \nabla\phi(y)\}$ for any $x, y, z \in \mathbb{R}^m$.

3. Generalized information inequality with Bregman divergence

Let X be a random variable with probability measure $P_\theta, \theta \in \Theta$. For notational simplicity, throughout this paper we assume that X is a real-valued random variable, but our result can be generalized to multivariate random variables without any additional assumptions. Define $T = T(X)$ to be an estimator such that $E(T) = \mu(\theta) := \mu_\theta$, where $\mu(\cdot)$ is a differentiable parametric function. Let $S = S(X, \theta)$ be a random variable such that $E(S) = a_0$ for a constant a_0 (not depending on θ). The following theorem provides a general lower bound for the expected Bregman divergence between T and μ_θ.

Theorem 1. *Suppose that $S \neq a_0$ almost surely. Then, for any strictly convex function ϕ,*

$$E\{BD_\phi(T, \mu_\theta)\} \geq \frac{[E\{(S-a_0)\mathrm{sgn}(T-\mu_\theta)\sqrt{BD_\phi(T, \mu_\theta)}\}]^2}{E\{(S-a_0)^2\}}, \tag{2}$$

where $\mathrm{sgn}(\cdot)$ *denotes the sign function and the equality holds if and only if*

$$\sqrt{BD_\phi(T, \mu_\theta)} = c(\theta)(S - a_0)$$

for some constant $c(\theta) \in \mathbb{R}$.

Proof of Theorem 1. Define two random variables Z_1 and Z_2 by $Z_1 = \mathrm{sgn}(T - \mu_\theta)\sqrt{\mathrm{BD}_\phi(T, \mu_\theta)}$ and $Z_2 = S - a_0$. By the Cauchy–Schwarz inequality, we have

$$\sqrt{E(Z_1^2)E(Z_2^2)} \geq |E(Z_1 Z_2)|,$$

where the equality holds if and only if $Z_1 = cZ_2$ for some constant c. Since $Z_2 \neq 0$ almost surely, by the assumption, it follows that

$$E(Z_1^2) \geq \frac{\{E(Z_1 Z_2)\}^2}{E(Z_2^2)},$$

which immediately completes our proof. □

Since many loss functions including the quadratic loss belong to the class of Bregman divergence, the expected lower bound of the Bregman divergence in (2) provides broader and more comprehensive generalizations of CRLB. Suppose that X belongs to a natural exponential family with parameter θ. That is, the probability density function of X can be written as

$$p_\theta(x) = h(x) \exp\{\theta x - \psi(\theta)\} \tag{3}$$

for a differential and strictly convex function $\psi(\cdot)$. Let ϕ be the conjugate of ψ. Due to the fact that $E(X) = \psi'(\theta) := \mu$, it follows from the Legendre transformation that $\psi(\theta) = \mu\phi'(\mu) - \psi(\mu)$ and $\theta = \phi(\mu)$. This implies that the exponential density function in (3) can be re-written as

$$p_\mu(x) = h(x) \exp\left\{(x-\mu)\phi'(\mu) + \phi(\mu)\right\} = g(x) \exp\{-\mathrm{BD}_\phi(x, \mu)\},$$

where $g(x) = h(x) \exp\{\phi(x)\}$. For two probability densities q and p, the Kullback–Leibler divergence from q to p is defined by

$$\mathrm{KL}(p\|q) = \int \log\left(\frac{p(x)}{q(x)}\right) p(x)\mathrm{d}x.$$

Then, it can be shown that

$$\begin{aligned}\mathrm{KL}(p_\mu \| p_\xi) &= \int \left\{(x-\mu)\phi'(\mu) + \phi(\mu) - (x-\xi)\phi'(\xi) - \phi(\xi)\right\} p_\mu(x)\mathrm{d}x \\ &= \mathrm{BD}(\mu, \xi),\end{aligned}$$

which implies that, when p_μ and p_ξ belong to an exponential family, the Kullback–Leibler divergence from p_ξ to p_μ is the Bregman divergence between μ and ξ. By letting $\xi = T$, a lower bound of the expected Kullback–Leibler divergence from p_T to p_μ can be directly obtained from Theorem 1 as follows:

$$E\{\mathrm{KL}(p_\mu \| p_T)\} \geq \frac{[E\{(S-a_0)\mathrm{sgn}(T-\mu)\sqrt{\mathrm{KL}(p_\mu \| p_T)}\}]^2}{E\left\{(S-a_0)^2\right\}}.$$

In the next section, we will show that the proposed lower bound includes many existing generalizations of CRLB as a special case.

4. Lower bounds on variance

When the convex function ϕ is defined to be $\phi(x) = x^2$, the Bregman divergence reduces to the quadratic loss,

$$\mathrm{BD}_\phi(x, y) = x^2 - y^2 - (x - y)2y = (x - y)^2.$$

Recall that $E(T) = \mu_\theta$. Hence, for $\phi(x) = x^2$, it can be easily shown that the expected Bregman divergence between T and μ_θ is the same as the variance of T, i.e.,

$$E\{\mathrm{BD}_\phi(T, \mu_\theta)\} = \mathrm{Var}(T).$$

As a result, when we set $\phi(x) = x^2$, we can obtain a general lower bound on the variance as follows.

Theorem 2. *Suppose that $S = S(X, \theta) \neq a_0$ almost surely. Then*

$$\mathrm{Var}_\theta(T) \geq \frac{[E\{(S - a_0)(T - \mu_\theta)\}]^2}{E\{(S - a_0)^2\}}, \tag{4}$$

where the equality holds if and only if

$$T = b(\theta) + c(\theta)S$$

for some constant $b(\theta), c(\theta) \in \mathbb{R}$.

Proof of Theorem 2. For $\phi(x) = x^2$, we have $\mathrm{BD}_\phi(x, y) = (x - y)^2$. This implies that

$$\mathrm{sgn}(x - y)\sqrt{\mathrm{BD}_\phi(x, y)} = \mathrm{sgn}(x - y)|x - y| = x - y.$$

By Theorem 1, we thus obtain the lower bound in (4). □

We now show that our result in Theorem 2 embraces existing generalizations of CRBL.

COROLLARY 3 (Cramér-Rao lower bound [11])

Define $S = S(X, \theta) = \frac{\partial}{\partial\theta} \log p_\theta(X)$, where p_θ is the density function with respect to P_θ. Assume that for a function $h(x)$ (not depending on θ),

$$\int \frac{\partial^k}{\partial\theta^k} h(x) p_\theta(x) \mathrm{d}x = \frac{\partial^k}{\partial\theta^k} \int h(x) p_\theta(x) \mathrm{d}x \tag{5}$$

for $k = 1, 2$. Then, the proposed lower bound in (4) reduces to the Cramér–Rao lower bound, that is,

$$\mathrm{Var}(T) \geq \frac{\{\mu'(\theta)\}^2}{I(\theta)},$$

where $\mu'(\theta) = \frac{\partial}{\partial\theta}\mu(\theta)$ and $I(\theta) = E\left\{\frac{\partial^2}{\partial\theta^2} \log p_\theta(X)\right\}$ is the Fisher information.

Proof of Corollary 3. Under the assumption (5), we have

$$a_0 = \int \left\{ \frac{\partial}{\partial\theta} \log p_\theta(x) \right\} p_\theta(x)\mathrm{d}x = \int \frac{\partial}{\partial\theta} p_\theta(x)\mathrm{d}x = \frac{\partial}{\partial\theta} \int p_\theta(x)\mathrm{d}x = 0.$$

This implies that

$$\begin{aligned} E\{(S-a_0)(T-\mu_\theta)\} &= \int (T(x)-\mu_\theta) \left(\frac{\partial}{\partial\theta} \log p_\theta(x) \right) p_\theta(x)\mathrm{d}x \\ &= \frac{\partial}{\partial\theta} \int T(x) p_\theta(x)\mathrm{d}x = \mu'(\theta). \end{aligned}$$

Similarly, it can be shown that

$$E\{(S-a_0)^2\} = I(\theta),$$

which completes our proof. □

COROLLARY 4 (Chapman–Robbins lower bound [4])

Define $S = S(X,\theta) = p_{\theta+\delta}(X)/p_\theta(X)$ *for* δ $(\neq 0)$. *Then, the proposed lower bound in* (4) *reduces to the Chapman–Robbins lower bound, that is,*

$$\mathrm{Var}(T) \geq \sup_{\delta\neq 0} \frac{(\mu_{\theta+\delta}-\mu_\theta)^2}{E\left[\left\{ \frac{p_{\theta+\delta}(X)}{p_\theta(X)} - 1 \right\}^2 \right]},$$

where $\mu_{\theta+\delta} = \int T(x) p_{\theta+\delta}(x)\mathrm{d}x$.

Proof of Corollary 4. Note that $a_0 = 1$ for any δ since

$$E(S) = \int \frac{p_{\theta+\delta}(x)}{p_\theta(x)} p_\theta(x)\mathrm{d}x = 1.$$

Hence, we have

$$\begin{aligned} E\{(S-a_0)(T-\mu_\theta)\} &= \int (T(x)-\mu_\theta) \left(\frac{p_{\theta+\delta}(x)}{p_\theta(x)} - 1 \right) p_\theta(x)\mathrm{d}x \\ &= \mu(\theta+\delta) - \mu(\theta). \end{aligned}$$

Then, our result immediately follows from Theorem 2. □

COROLLARY 5 (Bhattacharyya lower bound [2])

Let $K = (k_{ij})_{r\times r}$ *be an* $(r \times r)$ *matrix such that*

$$k_{ij} = E\left\{ \frac{\partial^i}{\partial\theta^i} p_\theta(X) \left(\frac{\partial^j}{\partial\theta^j} p_\theta(X) \right) \right\}.$$

Assume that K is positive-definite. Let $\eta = (\mu_\theta^{(1)}, \ldots, \mu_\theta^{(r)})^\top$, where $\mu_\theta^{(j)} = \frac{\partial^j}{\partial\theta^j}\mu(\theta)$ for $j = 1, \ldots, r$. Define $S = S(X, \theta) = \sum_{j=1}^r \alpha_j \frac{1}{p_\theta(X)} \frac{\partial^j}{\partial\theta^j} p_\theta(X)$, where α_j is the j-th element of $\alpha = K^{-1}\eta$. Assume that

$$E\left\{\frac{1}{p_\theta(X)}\frac{\partial^j}{\partial\theta^j}p_\theta(X)\right\} = 0$$

for $j = 1, \ldots, r$. Then, the proposed lower bound in (4) reduces to the Bhattacharyya lower bound as follows:

$$\mathrm{Var}(T) \geq \eta^\top K^{-1}\eta.$$

Proof of Corollary 5. Note that $E\left\{\frac{1}{p_\theta(X)}\frac{\partial^j}{\partial\theta^j}p_\theta(X)\right\} = 0$ for $j = 1, \ldots, r$ by the assumption. This implies that

$$a_0 = E(S) = \sum_{j=1}^r \alpha_j E\left\{\frac{1}{p_\theta(X)}\frac{\partial^j}{\partial\theta^j}p_\theta(X)\right\} = 0.$$

Hence, we have

$$E\left\{(S-a_0)^2\right\} = E\left[\left\{\sum_{j=1}^r \alpha_j \frac{1}{p_\theta(X)}\frac{\partial^j}{\partial\theta^j}p_\theta(X)\right\}^2\right] = \alpha^\top K\alpha. \tag{6}$$

Similarly, we have

$$\begin{aligned} E\left\{(T-\mu_\theta)(S-a_0)\right\} &= \int (T(x)-\mu_\theta)\left(\sum_{j=1}^r \alpha_j \frac{1}{p_\theta(X)}\frac{\partial^j}{\partial\theta^j}p_\theta(X)\right)p_\theta(x)\mathrm{d}x \\ &= \sum_{j=1}^r \alpha_j \frac{\partial^j}{\partial\theta^j}\int T(x)p_\theta(x)\mathrm{d}x = \alpha^\top\eta. \end{aligned} \tag{7}$$

Taking $\alpha = K^{-1}\eta$, it follows from Equations (6) and (7) that

$$\frac{[E\{(S-a_0)(T-\mu_\theta)\}]^2}{E\left\{(S-a_0)^2\right\}} = \frac{(\alpha^\top\eta)^2}{\alpha^\top K\alpha} = \eta^\top K^{-1}\eta.$$

Therefore, Theorem 2 leads to our result. □

5. Concluding remarks

As Hodges and Lehmann [7] used CRLB to show the admissibility for the quadratic loss, our result can be used to achieve the admissibility for a general class of loss functions that belong to the Bregman divergence. Our result can also be extended to multi-parameter cases using Bregman matrix divergences [8]. A Bayesian version of Theorem 1 can be developed easily, which will extend results obtained by Ghosh [5] and Gill and Levit [6].

Acknowledgements

The first author acknowledges Professors Partha P Majumder and B L S Prakasa Rao for the invitation to present this work at the Indian Academy of Sciences in honor of the birth centenary conference of the living legend of statistics, Prof. Calyampudi Radhakrishna (C R) Rao.

References

[1] Banerjee A, Merugu S, Dhillon I S and Ghosh J, Clustering with Bregman divergences, *J. Mach. Learn. Res*. **6** (2005) 1705–1749

[2] Bhattacharyya A, On some analogues of the amount of information and their use in statistical estimation, *Sankhyā Indian J. Stat.*. **8** (1946) 1–14

[3] Bregman L M, The relaxation method of finding the common points of convex sets and its application to the solution of problems in convex programming, *USSR Comput. Math. Math. Phys*. **7** (1967) 200–217

[4] Chapman D G and Robbins H, Minimum variance estimation without regularity assumptions, *Ann. Math. Stat.*. **22** (1951) 581–586

[5] Ghosh M, Cramér-Rao bounds for posterior variances, *Stat. Probab. Lett*. **17** (1993) 173–178

[6] Gill R D and Levit B Y, Applications of the van Trees inequality: a Bayesian Cramér–Rao bound, *Bernoulli* **1** (1995) 59–79

[7] Hodges J L and Lehmann E L, Some applications of the Cramér–Rao inequality, Proceedings of the Second Berkeley Symposium on Mathematical Statistics and Probability (1951) (University of California Press) pp. 13–22

[8] Kulis B, Sustik M A and Dhillon I S, Low-rank kernel learning with Bregman matrix divergences, *J. Mach. Learn. Res*. **10** (2009) 341–376

[9] Prakasa Rao B L S, Sequential Cramér–Rao type integral inequality, *Proc. Andhra Pradesh Akademi Sci*. **5** (2000) 23–38

[10] Prakasa Rao B L S, Improved sequential Cramér–Rao type integral inequality, *Sequential Anal*. **37** (2018) 59–68

[11] Rao C R, Information and the accuracy attainable in the estimation of statistical parameters, *Bull. Calcutta Math. Soc*. **37** (1945) 81–91