

Maximum likelihood estimators uniformly minimize distribution variance among distribution unbiased estimators in exponential families

PAUL VOS* and QIANG WU**

Department of Biostatistics, East Carolina University, Greenville, NC 27834, USA.
E-mail: *vosp@ecu.edu; **wuq@ecu.edu

We employ a parameter-free distribution estimation framework where estimators are random distributions and utilize the Kullback–Leibler (KL) divergence as a loss function. Wu and Vos [*J. Statist. Plann. Inference* **142** (2012) 1525–1536] show that when an estimator obtained from an i.i.d. sample is viewed as a random distribution, the KL risk of the estimator decomposes in a fashion parallel to the mean squared error decomposition when the estimator is a real-valued random variable. In this paper, we explore how conditional versions of distribution expectation (E^\dagger) can be defined so that a distribution version of the Rao–Blackwell theorem holds. We define distributional expectation and variance (V^\dagger) that also provide a decomposition of KL risk in exponential and mixture families. For exponential families, we show that the maximum likelihood estimator (viewed as a random distribution) is distribution unbiased and is the unique uniformly minimum distribution variance unbiased (UMV † U) estimator. Furthermore, we show that the MLE is robust against model specification in that if the true distribution does not belong to the exponential family, the MLE is UMV † U for the KL projection of the true distribution onto the exponential families provided these two distribution have the same expectation for the canonical statistic. To allow for estimators taking values outside of the exponential family, we include results for KL projection and define an extended projection to accommodate the non-existence of the MLE for families having discrete sample space. Illustrative examples are provided.

Keywords: distribution unbiasedness; extended KL projection; Kullback–Leibler loss; MVUE; Pythagorean relationship; Rao–Blackwell

1. Introduction

Wu and Vos [13] introduce a parameter-free distribution estimation framework and utilize the Kullback–Leibler (KL) divergence as a loss function. They show that the KL risk of a distribution estimator obtained from an i.i.d. sample decomposes in a fashion parallel to the mean squared error decomposition for a parameter estimator, and that an estimator is distribution unbiased, or simply unbiased, if and only if its distribution mean is equal to the true distribution. Distribution unbiasedness can be defined without using any parameterization. We call this approach parameter-free even though there may be applications where it is desirable to use a particular parameterization. When the distributions are, in fact, parametrically indexed, distribution unbiasedness handles multiple parameters simultaneously and is consistent under reparameterization.

Wu and Vos [13] also show that the MLE for distributions in the exponential family is always distribution unbiased.

The KL expectation and variance functions E and V are defined by minimizing over the space of all distributions. These functions completely describe an estimator in terms of its KL divergence around any distribution. In this paper, we introduce distribution expectation and variance functions E^\dagger and V^\dagger that are defined by minimizing over a smaller space of distributions. For exponential and mixture families, the expected KL risk is a function only of these quantities.

Even though the focus of this paper is on parametric exponential families, our approach is parameter-free in that the definitions and results are provided without regard to the parameterization of the family. There are three advantages to this approach: one, the lack of invariance of bias across parameter transformations is avoided; two, we can allow for estimators taking values outside of the exponential family; three, the case where the true distribution does not belong to the family is easily addressed.

Section 2 introduces the distribution expectation and variance functions and shows how these are a generalization of the mean and expectation functions for mean square error. Exponential families and their extension are discussed in Section 3. The fundamental properties of the distribution mean and variance functions allow using the ideas of Rao–Blackwell [2] to show that the MLE is the unique uniformly minimum distribution variance unbiased estimator (UMV † UE). This result is proved in Section 4. Three examples are given in Section 5 and Section 6 contains further remarks.

2. Kullback–Leibler risk, variance, and expectation

2.1. Motivation

The parametric version of the Rao–Blackwell theorem can be proved using a Pythagorean relationship that holds for mean square error (MSE) and the expectation operator. To prove the distribution version of the Rao–Blackwell theorem, we use a similar relationship that holds for KL risk and the KL expectation along with a second Pythagorean relationship that holds in exponential families for KL divergence and the KL projection. Basic properties of the expectation operator for real-valued random variables used in the proof can be extended to distribution-valued random variables. We begin with the property that the expectation minimizes the MSE.

For (real-valued) random variable Y and $a \in \mathbb{R}$, we can define the average behavior of Y relative to a using the risk function

$$E[d(Y, a)],$$

where d is a loss function, that is, a nonnegative convex function on $\mathbb{R} \times \mathbb{R}$. When $E[d(Y, a)] < \infty$ for some a , we define

$$V_d Y \stackrel{\text{def}}{=} \inf_{b \in \mathbb{R}} E[d(Y, b)]$$

and

$$E_d Y \stackrel{\text{def}}{=} \arg \min_{b \in \mathbb{R}} E[d(Y, b)]$$

if the minimum exists, in which case,

$$V_d Y = E[d(Y, E_d Y)].$$

When $d(a, b) = L(a, b) = (a - b)^2$, that is, risk is MSE, we have

$$E_L Y \stackrel{\text{def}}{=} \arg \min_{b \in \mathbb{R}} E[L(Y, b)] = \int y \, dR_0 \stackrel{\text{def}}{=} EY, \tag{2.1}$$

$$V_L Y \stackrel{\text{def}}{=} \inf_{b \in \mathbb{R}} E[L(Y, b)] = E[L(Y, EY)] \stackrel{\text{def}}{=} VY. \tag{2.2}$$

Note that we use the loss function as subscript to indicate expectation and variance defined in terms of an argmin and infimum of the loss function, while expectations and variances without a subscript are defined in terms of an integral, or in terms of a sum if the sample space is discrete. The middle equality signs in equations (2.1) and (2.2) are well-known results for EY and VY . These two values completely characterize the risk because of the relationship

$$E[L(Y, a)] = L(E_L Y, a) + V_L Y \quad \forall a \in \mathbb{R}. \tag{2.3}$$

In particular, the MSE for a random variable Y is completely determined by knowing its expectation EY and variance VY . Note that (2.3) holds for any distribution function such that EY and VY exist. For general loss functions d , the argmin $E_d Y$ and min $V_d Y$ do not characterize the risk; that is,

$$E[d(Y, a)] - d(E_d Y, a)$$

will be a function of a .

The expectation and variance also have the following conditional properties

$$EY = EE[Y|X], \tag{2.4}$$

$$VY = VE[Y|X] + E[V(Y|X)]. \tag{2.5}$$

In the next section, we consider random variables that take values on a space of distributions \mathcal{R} and show that when the KL divergence is used to compare distributions, equations (2.1) through (2.5) hold for KL risk.

2.2. Space of all distributions \mathcal{R}

Let $(\mathbb{X}, \mathcal{X})$ be a sample space equipped with a σ -finite measure λ . When \mathbb{X} is finite or countable, λ is usually the counting measure. When $\mathbb{X} \subset \mathbb{R}^d$ and \mathbb{X} contains an open set of \mathbb{R}^d for some $d = 1, 2, \dots$, then λ is usually the Lebesgue measure on \mathbb{R}^d . Requiring \mathbb{X} to contain an open set implies that the dimension of \mathbb{X} is d . Let \mathcal{R} be the collection of all probability measures R on $(\mathbb{X}, \mathcal{X})$ that are absolutely continuous with respect to λ , that is, $\lambda(A) = 0$ implies $R(A) = 0$ for all $A \in \mathcal{X}$. This is denoted as $R \ll \lambda$. Note that we allow the support of R to be a proper subset of \mathbb{X} .

Let \mathbf{R} (in bold font) be a random quantity whose values are distributions in \mathcal{R} . The density of the distribution R with respect to λ will be denoted by r (in lower case), and the corresponding random variable by \mathbf{r} (in bold font lower case). Following Definition 2.1 in [13], \mathbf{R} is an \mathcal{R} -valued random variable if $\mathbf{R}(A)$ is a real-valued random variable for all $A \in \mathcal{X}$. We are considering the problem of estimating a distribution so for this paper $\mathbf{R} = \widehat{\mathbf{R}}_{\mathbf{X}}$ is any estimator of an unknown distribution $R_0 \in \mathcal{R}$ where \mathbf{X} is an i.i.d. sample from R_0 . A random distribution is a mapping from \mathbb{X}^n to \mathcal{R} . Let S be another random quantity that is jointly distributed with \mathbf{R} .

Theorem 2.1. *For every $S = s$, $K_s = E[\mathbf{R}|S = s]$ is a probability measure that is absolutely continuous with respect to λ , that is, $K_s \in \mathcal{R}$, is unique up to measure zero (λ), and has a density*

$$k_s(y) = E[\mathbf{r}(y)|S = s] \quad \text{for } y \in \mathbb{X}. \tag{2.6}$$

In addition, when s is replaced with the random variable S , $K_S = E[\mathbf{R}|S]$ is an \mathcal{R} -valued random variable.

Proof. For all s it is easily seen that K_s is a probability measure because K_s is countably additive and $K_s(\mathbb{X}) = 1 - K_s(\emptyset) = 1$, where \emptyset is the empty set. The remaining claims of the theorem can be established by noting that equation (2.6) can be written as

$$k_s(y) = \int_{\mathbb{X}^n} r_{\mathbf{x}}(y)r_0^n(\mathbf{x}|s) d\lambda^n(\mathbf{x}), \tag{2.7}$$

where $r_0^n(\mathbf{x}|s)$ is the conditional distribution of \mathbf{x} given s . Since

$$E[\mathbf{R}(A)|s] = \int_{\mathbb{X}^n} \int_A r_{\mathbf{x}}(y) d\lambda(y)r_0^n(\mathbf{x}|s) d\lambda^n(\mathbf{x}),$$

the set $A \in \mathcal{X}$ is arbitrary, and the integrals can be interchanged, we see that $k_s(y)$ is the density for K_s and $K_s \in \mathcal{R}$ for each s so K_S is an \mathcal{R} -valued random variable. □

For \mathcal{R} -valued random variable \mathbf{R} and $R \in \mathcal{R}$, we can define the average behavior of \mathbf{R} relative to R using the risk function

$$E[d(\mathbf{R}, R)],$$

where d is a loss function, that is, a nonnegative convex function on $\mathcal{R} \times \mathcal{R}$. Note that the expectation used to define the risk is with respect to some distribution $R_0 \in \mathcal{R}$; R_0 will be fixed but arbitrary other than constraints to ensure that the quantities in the expressions below exist and that the support of R_0 is \mathbb{X} . For any function d such that $E[d(\mathbf{R}, R)] < \infty$ for some R , we define

$$V_d \mathbf{R} \stackrel{\text{def}}{=} \inf_{R_1 \in \mathcal{R}} E[d(\mathbf{R}, R_1)]$$

and

$$E_d \mathbf{R} \stackrel{\text{def}}{=} \arg \min_{R_1 \in \mathcal{R}} E[d(\mathbf{R}, R_1)]$$

if the minimum exists, in which case,

$$V_d \mathbf{R} = E[d(\mathbf{R}, E_d \mathbf{R})].$$

For KL risk, that is, when $d(R_1, R_2) = D(R_1, R_2) \stackrel{\text{def}}{=} E_{R_1} \log(r_1/r_2)$, we have

$$E_D \mathbf{R} \stackrel{\text{def}}{=} \arg \min_{R_1 \in \mathcal{R}} E[D(\mathbf{R}, R_1)] = \int \mathbf{r}_x(y) r_0^n(\mathbf{x}) d\lambda^n(\mathbf{x}) \stackrel{\text{def}}{=} ER, \tag{2.8}$$

$$V_D \mathbf{R} \stackrel{\text{def}}{=} \inf_{R_1 \in \mathcal{R}} E[D(\mathbf{R}, R_1)] = ED(\mathbf{R}, ER) \stackrel{\text{def}}{=} VR. \tag{2.9}$$

The middle equalities in equations (2.8) and (2.9) are established in Wu and Vos [13]. Since these are equal when D is the KL divergence and we consider no other divergence functions on $\mathcal{R} \times \mathcal{R}$, we will simply write $ER \in \mathcal{R}$ and $VR \in \mathbb{R}$ for the KL mean and variance.

Furthermore, ER and VR completely characterize the average behavior of the \mathcal{R} -valued random variable \mathbf{R} relative to any distribution $R \in \mathcal{R}$ because of the relationship

$$E[D(\mathbf{R}, R)] = D(ER, R) + VR \quad \forall R \in \mathcal{R}. \tag{2.10}$$

This means the KL risk for an \mathcal{R} -valued random variable \mathbf{R} , having any distribution function, is completely determined by knowing its argmin, $ER \in \mathcal{R}$, and minimum, $VR \geq 0$. When $R = R_0$, equation (2.10) gives the decomposition of the KL risk in terms of bias and variance. The relationship in (2.10) will not hold for general nonnegative convex functions d . In this paper we only consider KL divergence $D(R_1, R_2)$. Furthermore, a conditional expectation on \mathcal{R} -valued random variables can be defined so that the following conditional properties hold

$$ER = EE[\mathbf{R}|S], \tag{2.11}$$

$$VR = VE[\mathbf{R}|S] + E[V(\mathbf{R}|S)], \tag{2.12}$$

where S could be \mathcal{R} -valued but could also be real or other valued since values of S will only be used to generate sub sigma fields.

Theorem 2.2 (Characterization theorem for expected KL divergence on \mathcal{R}). *Let $R_0 \in \mathcal{R}$ have support \mathbb{X} and let \mathbf{R} be an \mathcal{R} -valued random variable such that the KL mean ER and the KL variance VR exist and are finite. Then for any $R \in \mathcal{R}$ the mean divergence between \mathbf{R} and R depends only on the KL mean ER and KL variance VR . Furthermore, the KL mean and KL variance satisfy the classical conditional equalities (2.11) and (2.12).*

Proof. Equation (2.10) follows from the definition of KL variance and Theorem 5.2 in [13] who show that the expected KL loss $E[D(\mathbf{R}, R)]$ from an \mathcal{R} -valued random variable \mathbf{R} to a distribution $R \in \mathcal{R}$ decomposes as

$$E[D(\mathbf{R}, R)] = E[D(\mathbf{R}, ER)] + D(ER, R). \tag{2.13}$$

Equation (2.11) follows from the fact that the KL means $E\mathbf{R}$ and $E[\mathbf{R}|S]$ have densities with respect to λ and the order of integration can be interchanged. The steps are the same as those that establish $EX = EE[X|Y]$ for \mathbb{R} -valued random variables X and Y . We rewrite (2.10) as

$$E[D(\mathbf{R}, R)] - D(E\mathbf{R}, R) = V\mathbf{R}. \quad (2.14)$$

Note that both expectations (with domain \mathbb{R} -valued random variables and with domain \mathcal{R} -valued random variables) and the variance depend on the data generation distribution R_0 , which can be any point in \mathcal{R} with support \mathbb{X} . If this equation holds for random sample X_1, \dots, X_n then it also applies to the conditional distribution of X_1, \dots, X_n given $S = s$

$$E[D(\mathbf{R}, R)|s] - D(E[\mathbf{R}|s], R) = V(\mathbf{R}|s).$$

Substituting S into the equation above and taking expectation gives

$$E[D(\mathbf{R}, R)] - E[D(E[\mathbf{R}|S], R)] = E[V(\mathbf{R}|S)]. \quad (2.15)$$

Substituting $E[\mathbf{R}|S]$ into \mathbf{R} in (2.14) and using $EE[\mathbf{R}|S] = E\mathbf{R}$ gives

$$E[D(E[\mathbf{R}|S], R)] - D(E\mathbf{R}, R) = V(E[\mathbf{R}|S]). \quad (2.16)$$

Adding (2.15) to (2.16) and substituting from (2.14) proves (2.12). \square

The random variable \mathbf{R} is a distribution function defined on the sample space and it will be useful to relate \mathbf{R} to a statistic T . We define $\mu_T(R) = E_RT \in \mathbb{R}^d$ and when we consider only one statistic we write $\mu(R) = \mu_T(R)$. The \mathbb{R}^d -valued random variable $\mu(\mathbf{R})$ describes the behavior of the \mathcal{R} -valued random variable \mathbf{R} and the mean of $\mu(\mathbf{R})$ can be obtained from the KL mean.

Theorem 2.3 (Expectation property on \mathcal{R}). *For any statistic T such that $\mu(\mathbf{R}) < \infty$ a.e., the mean of T under $E\mathbf{R}$ equals the mean of \mathbb{R}^d -valued random variable $\mu(\mathbf{R})$*

$$\mu(E\mathbf{R}) = E[\mu(\mathbf{R})]. \quad (2.17)$$

Proof. The density for $E\mathbf{R}$ can be written as $\int r_{\mathbf{x}}(y)r_0^n(\mathbf{x})d\lambda^n(\mathbf{x})$ so that

$$\begin{aligned} \mu(E\mathbf{R}) &= \int T(y) \int r_{\mathbf{x}}(y)r_0^n(\mathbf{x})d\lambda^n(\mathbf{x})d\lambda(y) \\ &= \int r_0^n(\mathbf{x}) \int T(y)r_{\mathbf{x}}(y)d\lambda(y)d\lambda^n(\mathbf{x}) = E[\mu(\mathbf{R})] \end{aligned}$$

because the order of integration can be switched. \square

2.3. General subspace \mathcal{P}

We typically are interested in a subfamily of distributions $\mathcal{P} \subset \mathcal{R}$ and we describe a distribution in terms of the KL risk $E[D(\mathbf{R}, P)]$ for $P \in \mathcal{P}$. We add the regularity condition that the support

of each distribution in \mathcal{P} is \mathbb{X} . Equation (2.10) shows that ER and VR give the KL risk for any $P \in \mathcal{P}$. However, generally $ER \notin \mathcal{P}$ even if R takes values only in \mathcal{P} . We consider whether an expectation can be defined that takes values in \mathcal{P} and so that (2.10) holds. We will define this expectation as a minimum over \mathcal{P} . We define

$$V^\dagger R = \inf_{P \in \mathcal{P}} E[D(R, P)]$$

and

$$E^\dagger R = \arg \min_{P \in \mathcal{P}} E[D(R, P)]$$

if the minimum exists, in which case

$$V^\dagger R = E[D(R, E^\dagger R)].$$

Equation (2.10) now becomes

$$E[D(R, P)] = D(E^\dagger R, P) + V^\dagger R + \Delta(ER, E^\dagger R, P) \quad \forall P \in \mathcal{P}, \tag{2.18}$$

where

$$\Delta(ER, E^\dagger R, P) = D(ER, P) - D(ER, E^\dagger R) - D(E^\dagger R, P). \tag{2.19}$$

If Δ vanishes for all $P \in \mathcal{P}$ then the argmin $E^\dagger R$ and the min $V^\dagger R$ completely characterize R in terms of KL risk. When Δ is small these functions can be used to approximate the KL risk of R . We will show the term Δ vanishes when \mathcal{P} is an exponential family. The relationship between the expectations ER and $E^\dagger R$ can be expressed by using the KL projection onto \mathcal{P}

$$\Pi R = \arg \min_{P \in \mathcal{P}} D(R, P).$$

By equation (2.10),

$$E^\dagger R = \Pi ER. \tag{2.20}$$

For any \mathcal{P} , we have that $VR \leq V^\dagger R$ since $\mathcal{P} \subset \mathcal{R}$. These results are summarized in the following theorem.

Theorem 2.4. *Let $R_0 \in \mathcal{R}$ such that the support of R_0 is \mathbb{X} and let R be an \mathcal{R} -valued random variable such that the distribution mean $E^\dagger R$ and the distribution variance $V^\dagger R$ exist and are finite. Then for any $P \in \mathcal{P}$ the mean divergence between R and P is given by (2.18). The term Δ measures the extent to which the KL mean, distribution mean, and P depart from forming a dual Pythagorean triangle. The KL variance is less than or equal to the distribution variance, $VR \leq V^\dagger R$, and the distribution mean is the KL projection of the KL mean onto \mathcal{P} , $E^\dagger R = \Pi ER$.*

Wu and Vos [13] show that $\Delta = 0$ for all $P \in \mathcal{P}$ an exponential family. For mixture families $ER = E^\dagger R$. Hence, Δ vanishes when \mathcal{P} is either an exponential family or mixture family.

While we don't know how to write E^\dagger as an integral and the expectation property (2.17) does not hold for E^\dagger in general, we show equations (2.11) and (2.12) hold with E replaced with E^\dagger and V replaced with V^\dagger when \mathcal{P} is either an exponential or mixture family. Furthermore, the expectation property will hold for E^\dagger when \mathcal{P} is an exponential family and T is the canonical statistic.

3. Exponential family \mathcal{P}

For a general subspace $\mathcal{P} \subset \mathcal{R}$ the distribution mean $E^\dagger \mathbf{R}$ and distribution variance $V^\dagger \mathbf{R}$ do not characterize $E[D(\mathbf{R}, P)]$ for $P \in \mathcal{P}$. However, when \mathcal{P} is an exponential family these quantities do characterize $E[D(\mathbf{R}, P)]$ and the classical equalities relating conditional mean and variance hold. A standard reference for exponential families is Brown [3], but the approach we take here is slightly different since our emphasis is on the distributions without regard to any particular parameterization. An exponential family \mathcal{P} will be defined by selecting a point $P_0 \in \mathcal{R}$ and statistic $T(x)$ taking values in \mathbb{R}^d . The defining property of an exponential family is that for any $P \in \mathcal{P}$ the log of the density of P with respect to P_0 is a linear combination of $T(x)$ and the constant function. We start with some definitions and basic properties.

3.1. Definitions and the projection property

Definition 3.1. \mathcal{P} is an exponential family on \mathbb{X} if there exists $P_0 \in \mathcal{R}$ such that the support of P_0 is \mathbb{X} and a function $T : \mathbb{X} \mapsto \mathbb{R}^d$ such that for any $P \in \mathcal{P}$

$$dP \propto e^{\theta' T(x)} dP_0 \quad \text{for some } \theta \in \mathbb{R}^d.$$

The distribution P_0 is called a base point and T is called the canonical statistic of \mathcal{P} . The canonical parameter space is

$$\theta(\mathcal{P}) = \{ \theta \in \mathbb{R}^d : \text{for some } P \in \mathcal{P}, dP \propto e^{\theta' T(x)} dP_0 \}.$$

Without loss of generality, we can choose a base point P_0 such that $P_0 \in \mathcal{P}$. We'll refer to exponential families using base points that belong to the family.

Definition 3.2. Let \mathcal{P} be an exponential family with base point P_0 , canonical statistic T , and set $\Theta = \{ \theta \in \mathbb{R}^d : \int e^{\theta' T(x)} dP_0 < \infty \}$. The cumulant function has domain Θ and is defined as

$$\psi(\theta) = \log \int e^{\theta' T(x)} dP_0.$$

The density with respect to P_0 for any $P \in \mathcal{P}$ is

$$\frac{dP}{dP_0} = \exp\{ \theta' T(x) - \psi(\theta) \} \quad \text{for some } \theta \in \theta(\mathcal{P}).$$

The family \mathcal{P} is regular if $\theta(\mathcal{P})$ is open and \mathcal{P} is full if $\theta(\mathcal{P}) = \Theta$.

By the factorization theorem, T is sufficient. It will often be useful to restrict the choice of T so that it is complete for the full exponential family \mathcal{P} .

Definition 3.3. A statistic T is complete for \mathcal{P} if

$$E_P h(T) = 0 \quad \forall P \in \mathcal{P} \implies h(T) = 0 \quad \text{a.e. } \mathcal{P}.$$

The following theorem shows that the projection operator on \mathcal{P} behaves like the expectation operator on \mathcal{R} (Theorem 2.3) and will be used to show that the classical conditional expectation equation holds for E^\dagger .

Theorem 3.1 (Projection property on \mathcal{P}). If Π is the KL projection onto \mathcal{P} , where \mathcal{P} is an exponential family having canonical statistic T and $\mu(R) = E_R T$, then for any $R \in \mathcal{R}$ such that $\mu(R) \in \mu(\mathcal{P})$,

$$\mu(\Pi R) = \mu(R), \tag{3.1}$$

where $\mu(\mathcal{P}) = \{\mu \in \mathbb{R}^d : \text{for some } P \in \mathcal{P}, \mu = E_P T\}$ is the mean parameter space of \mathcal{P} .

Proof. This result follows from the relationship between the natural and expectation parameters for an exponential family \mathcal{P} . Let $\mu_1 = \mu(P_1)$ for some $P_1 \in \mathcal{P}$. Then the natural parameter $\theta(P_1)$ of this distribution satisfies

$$\theta(P_1) = \arg \max_{\theta \in \Theta} [\theta' \mu_1 - \psi(\theta)] \tag{3.2}$$

and since θ parameterizes \mathcal{P} ,

$$P_1 = \arg \max_{P \in \mathcal{P}} [\theta(P)' \mu_1 - \psi(\theta(P))]. \tag{3.3}$$

The result now follows for exponential family \mathcal{P} by simple calculation

$$\begin{aligned} \Pi R_1 &= \arg \min_{P \in \mathcal{P}} D(R_1, P) \\ &= \arg \min_{P \in \mathcal{P}} (E_{R_1} \log r_1 - E_{R_1} \log p) \\ &= \arg \min_{P \in \mathcal{P}} E_{R_1} \log p \\ &= \arg \min_{P \in \mathcal{P}} (\theta(P)' \mu(R_1) - \psi(\theta(P))) \\ &= P_1, \end{aligned}$$

where $\mu(P_1) = \mu(R_1)$ by (3.3). □

Corollary 3.1 (Pythagorean property for exponential families). Let \mathcal{P} be an exponential family and let $R \in \mathcal{R}$ such that ΠR exists. For all $P \in \mathcal{P}$

$$D(R, P) = D(R, \Pi R) + D(\Pi R, P). \tag{3.4}$$

This is a well-known result. See, for example, [4] or [6].

We define an extended projection $\bar{\Pi}R$ to be any distribution in \mathcal{R} such that expectation and Pythagorean properties hold and it belongs to the “boundary” of \mathcal{P} ; that is,

$$\mu(R) = \mu(\bar{\Pi}R), \tag{3.5}$$

$$D(R, P) = D(R, \bar{\Pi}R) + D(\bar{\Pi}R, P) \quad \forall P \in \mathcal{P}, \tag{3.6}$$

$$\inf_{P \in \mathcal{P}} D(\bar{\Pi}R, P) = 0.$$

Note that ΠR satisfies these three equalities, and that the last two equalities imply

$$D(R, \bar{\Pi}R) = \inf_{P \in \mathcal{P}} D(R, P).$$

The extended projection allows us to define the extended MLE in the next section.

The Pythagorean property allows us to improve \mathcal{R} -valued random variables by the projection Π or, more generally, by $\bar{\Pi}$.

Corollary 3.2 (Projection property for \mathcal{R} -valued random variables). *If $\bar{\Pi}\mathbf{R}$ exists a.e., then*

$$E[D(\mathbf{R}, P)] \geq E[D(\bar{\Pi}\mathbf{R}, P)]$$

with equality holding if and only if $\bar{\Pi}\mathbf{R} = \mathbf{R}$ a.e.

Proof. Replacing R with \mathbf{R} in equation (3.6) and taking expectations shows

$$E[D(\mathbf{R}, P)] = E[D(\mathbf{R}, \bar{\Pi}\mathbf{R})] + E[D(\bar{\Pi}\mathbf{R}, P)] \quad \forall P \in \mathcal{P}$$

and the result follows from the fact that $E[D(\mathbf{R}, \bar{\Pi}\mathbf{R})] \geq 0$ with equality holding if and only if $\mathbf{R} = \bar{\Pi}\mathbf{R}$ a.e. □

3.2. Fundamental equations for distribution mean and variance

For exponential families, the distribution expectation and variance have the same properties as the KL expectation and variance. One distinction is that the expectation property of E holds for any statistic while for E^\dagger the expectation property holds only for the canonical statistic T .

Theorem 3.2 (Characterization of expected KL divergence on \mathcal{P}). *Let $R_0 \in \mathcal{R}$ have support \mathbb{X} and let \mathbf{R} be an \mathcal{R} -valued random variable such that the distribution mean $E^\dagger\mathbf{R}$ exists and the distribution variance $V^\dagger\mathbf{R}$ is finite. Then for any $P \in \mathcal{P}$, where \mathcal{P} is an exponential family, the mean KL divergence between \mathbf{R} and P depends only on the distribution mean and distribution variance*

$$E[D(\mathbf{R}, P)] = D(E^\dagger\mathbf{R}, P) + V^\dagger\mathbf{R} \quad \forall P \in \mathcal{P}. \tag{3.7}$$

Assuming the conditional expectations and variances exist, the distribution mean and distribution variance satisfy the classical conditional equalities

$$E^\dagger \mathbf{R} = E^\dagger E^\dagger[\mathbf{R}|S], \tag{3.8}$$

$$V^\dagger \mathbf{R} = V^\dagger E^\dagger[\mathbf{R}|S] + E[V^\dagger(\mathbf{R}|S)], \tag{3.9}$$

where S is a real-valued random vector. Furthermore, the expectation property holds for the canonical statistic T

$$\mu(E^\dagger \mathbf{R}) = E[\mu(\mathbf{R})]. \tag{3.10}$$

Proof. By Corollary 3.1 and equation (3.1) the correction term (2.19) vanishes showing that equation (3.7) holds. Equation (3.10) follows from

$$\mu(E^\dagger \mathbf{R}) = \mu(E\mathbf{R}) \tag{3.11}$$

and the expectation property on \mathcal{R} (2.17). Equation (3.11) follows from the (extended) projection property for exponential families (3.1) and (3.5) and the relationship between E and E^\dagger (2.20). Now equation (3.8) follows from

$$\begin{aligned} \mu(E^\dagger E^\dagger[\mathbf{R}|S]) &= E[\mu(E^\dagger[\mathbf{R}|S])] \\ &= E[\mu(E[\mathbf{R}|S])] \\ &= \mu(EE[\mathbf{R}|S]) \\ &= \mu(E\mathbf{R}) \\ &= \mu(E^\dagger \mathbf{R}), \end{aligned}$$

where the first equality follows from (3.10), the second and fifth equalities follow from (3.11), the third equality follows from the expectation property of the KL mean on \mathcal{R} , and the fourth equality follows from the conditional expectation property that holds on \mathcal{R} (2.11). Equation (3.9) follows again the same steps that justified (2.12). We rewrite (3.7) as

$$E[D(\mathbf{R}, R)] - D(E^\dagger \mathbf{R}, R) = V^\dagger \mathbf{R}. \tag{3.12}$$

If this equation holds for random sample X_1, \dots, X_n then it also applies to the conditional distribution of X_1, \dots, X_n given $S = s$

$$E[D(\mathbf{R}, R)|s] - D(E^\dagger[\mathbf{R}|s], R) = V^\dagger(\mathbf{R}|s).$$

Substituting S into the equation above and taking expectation gives

$$E[D(\mathbf{R}, R)] - E[D(E^\dagger[\mathbf{R}|S], R)] = E[V^\dagger(\mathbf{R}|S)]. \tag{3.13}$$

Substituting $E^\dagger[\mathbf{R}|S]$ into \mathbf{R} in (3.12) and using $E^\dagger E^\dagger[\mathbf{R}|S] = E^\dagger \mathbf{R}$ gives

$$E[D(E^\dagger[\mathbf{R}|S], R)] - D(E^\dagger \mathbf{R}, R) = V^\dagger E^\dagger[\mathbf{R}|S]. \tag{3.14}$$

Adding (3.13) to (3.14) and substituting from (3.12) proves (3.9). □

4. Rao–Blackwell and the MLE as the unique UMV[†]U distribution estimator

An immediate corollary to the characterization theorem on \mathcal{P} (equations (3.7), (3.8), and (3.9)) is that for any random distribution \mathbf{R} and any statistic S , the random distribution $E^\dagger[\mathbf{R}|S]$ will have the same distribution mean and have distribution variance less than or equal to that of \mathbf{R} . If $S = T$ is sufficient then $E^\dagger[\mathbf{R}|T]$ is an estimator and if T is also complete $E^\dagger[\mathbf{R}|T]$ will have smaller variance than \mathbf{R} unless they are equal with probability one. This conditional expectation is enough to establish a Rao–Blackwell result for distribution estimators if these were restricted to \mathcal{P} . However, since we are allowing \mathcal{R} -valued estimators we also need to project the distributions onto \mathcal{P} using $\bar{\Pi}$.

For an exponential family $\{P(y; \tau)\}$ having mean parameter $\tau \in \mu(\mathcal{P}) = M$ and discrete sample space we typically have that $\Pr(T \in M) < 1$ while $\Pr(T \in \bar{M}) = 1$ where \bar{M} is the closure of M . In this case, the MLE does not always exist. However, the characterization theorem applies to \mathcal{R} -valued estimators so we can define an estimator that equals the MLE $P(y; t)$ when it exists and as a distribution $\bar{P}(y; t)$ such that $\mu(\bar{P}(y; t)) = t$ and $\inf_{P \in \mathcal{P}} D(\bar{P}, P) = 0$ if $t \notin M$. The *extended MLE* as distribution estimator is

$$\widehat{P}^*(y; t) = \begin{cases} P(y; t) & \text{if } t \in M, \\ \bar{P}(y; t) & \text{if } t \notin M. \end{cases}$$

Unbiasedness of \widehat{P}^* follows from the following theorem.

Theorem 4.1 (Distribution unbiased estimators in exponential families). *Let \mathcal{P} be an exponential family with complete sufficient statistic T and let \mathbf{R} be a \mathcal{R} -valued random variable. The estimator \mathbf{R} is distribution unbiased for $P_0 = \Pi R_0$ if and only if $\mu(E(\mathbf{R}|T)) = T$ a.e.*

Proof. We must show $\Pi E\mathbf{R} = P_0$ for all $P_0 \in \mathcal{P}$ if and only if $\mu(E(\mathbf{R}|T)) = T$ a.e. for all $P_0 \in \mathcal{P}$. Consider the following equivalencies each of which holds for all $P_0 \in \mathcal{P}$:

$$\begin{aligned} \Pi E\mathbf{R} &= P_0 \\ \iff \mu(\Pi E\mathbf{R}) &= \mu(P_0) \\ \iff \mu(E\mathbf{R}) &= \mu(P_0) \\ \iff \mu(E E(\mathbf{R}|T)) &= \mu(P_0) \\ \iff E[\mu(E(\mathbf{R}|T))] &= \mu(P_0) \\ \iff E[\mu(E(\mathbf{R}|T))] &= E(T). \end{aligned}$$

The first equivalence follows because the expectation of T parameterizes \mathcal{P} , the second equivalence follows from the projection property for exponential families, the third equivalence follows from the conditional expectation defined for the KL mean, the fourth equivalence follows from the expectation property for the KL mean, and the fifth equivalence follows from the definition

of the function μ . Clearly, $\mu(E[\mathbf{R}|T]) = T$ a.e. implies the last equality. Since T is complete and the last equality holds for all $P_0 \in \mathcal{P}$, this implies

$$\mu(E[\mathbf{R}|T]) = T \quad \text{a.e.} \quad \square$$

Theorem 4.2 (Optimality of the MLE for exponential families). *Let X_1, \dots, X_n be i.i.d. from a distribution $R_0 \in \mathcal{R}$ such that the support of R_0 is \mathbb{X} . Let \mathcal{P} be an exponential family with complete sufficient statistic T such that $\mu(R_0) \in \mu(\mathcal{P})$. If $\hat{\mathbf{P}}$ is the MLE or an extended MLE that exists a.e., then $\hat{\mathbf{P}}$ is distribution unbiased for the ΠR_0 and it is the unique uniformly minimum distribution variance estimator among all \mathcal{R} -valued estimators that are distribution unbiased for ΠR_0 and for which the extended projection $\bar{\Pi}\mathbf{R}$ exists a.e.*

Proof. Uniqueness and uniform minimum distribution variance follow from the projection property for \mathcal{R} -valued random variables, the characterization theorem on \mathcal{P} described above, and the unbiasedness from Theorem 4.1. □

5. Examples

5.1. Binomial distribution

We consider the number of events or “successes” in n trials. The sample space is

$$\mathbb{X} = \{0, 1, 2, \dots, n\}.$$

Under the assumptions that these trials are independent and each trial has the same success probability $0 < \theta < 1$, the distribution of X belongs to the n -binomial family

$$\mathcal{P} = \left\{ P \in \mathcal{R} : P(x) = P_\theta(x) = \binom{n}{x} \theta^x (1 - \theta)^{n-x} \text{ for some } 0 < \theta < 1 \right\}.$$

The MLE for the parameter θ is $\hat{\theta} = x/n$ for $x \notin \{0, n\}$ but is undefined otherwise. The extended MLE (it will correspond in a natural way to the extended MLE distribution estimator) is $\hat{\theta} = x/n$ for all $x \in \mathbb{X}$ and it is unbiased for θ . However, it is not unbiased for other parameterizations such as the odds $v = \theta/(1 - \theta)$, or the log odds $\gamma = \log v$. When viewed as a distribution, that is, $P_{\hat{\theta}}(x)$, equivalently, $P_{\hat{v}}(x)$ or $P_{\hat{\gamma}}(x)$ (where we allow the odds v and log odds γ to take values in the extended reals), the MLE is the unique uniformly minimum distribution variance unbiased estimator. As is common practice, we have used the same notation $\hat{\theta}$ for both the MLE and the extended MLE.

Estimators, whether real-valued or distribution-valued, are functions with domain \mathbb{X} . For the n -binomial family an estimator is given by a sequence of $n + 1$ values, real numbers for $\hat{\theta}$ and probability distributions for $P_{\hat{\theta}}$. For $\hat{\theta}$, we have the sequence

$$\frac{0}{n}, \frac{1}{n}, \frac{2}{n}, \dots, \frac{n-1}{n}, \frac{n}{n}. \tag{5.1}$$

Let P_{θ_0} be a distribution in \mathcal{P} . If probabilities of P_{θ_0} are used to assign weights to the values in (5.1), then the real number that is closest to the weighted values of (5.1) is θ_0 . That is,

$$\theta_0 = \arg \min_{\theta \in (0,1)} E \left(\frac{X}{n} - \theta \right)^2.$$

By the Rao–Blackwell theorem, for any other sequence of $n + 1$ real numbers

$$y(0), y(1), y(2), \dots, y(n - 1), y(n) \tag{5.2}$$

that satisfy

$$\theta_0 = \arg \min_{\theta \in (0,1)} E (y(X) - \theta)^2,$$

the realized minimum will be greater than the minimum obtained using the values in (5.1) unless the sequences are equal, $y(x) = x/n$ for $x \in \{0, 1, 2, \dots, n\}$.

A distribution estimator $P_{\hat{\theta}}$ obtained from the real valued estimator given in (5.1) can be defined as

$$I_0(x), P_{1/n}(x), P_{2/n}(x), \dots, P_{(n-1)/n}(x), I_1(x), \tag{5.3}$$

where I_a is the indicator function for its subscript; that is, the degenerate distribution putting all mass on 0 or 1. Since $\inf_{P \in \mathcal{P}} D(I_a, P) = 0$ it is easily checked that $\overline{\Pi} I_a = I_a$ which means that the sequence in (5.3) is the extended MLE $\widehat{\mathbf{P}}^*$. Hence, $\widehat{\mathbf{P}}^* = P_{\hat{\theta}}$. Again, we let P_{θ_0} be any distribution in \mathcal{P} . If P_{θ_0} is used to assign weights to the distributions in (5.3), then the distribution in \mathcal{P} that is closest to the weighted average of the distributions in (5.3) is P_{θ_0} . That is,

$$P_{\theta_0} = \arg \min_{P \in \mathcal{P}} E [D(P_{\hat{\theta}}, P)].$$

By the distribution version of the Rao–Blackwell theorem (Theorem 4.2) for any estimator $\tilde{\theta}$, expressed as a distribution estimator,

$$P_{\tilde{\theta}(0)}, P_{\tilde{\theta}(1)}, \dots, P_{\tilde{\theta}(n)} \tag{5.4}$$

that satisfies

$$P_{\theta_0} = \arg \min_{P \in \mathcal{P}} E [D(P_{\tilde{\theta}}, P)],$$

the realized minimum will be greater than that of the MLE (5.3) unless the two sequences of functions (5.3) and (5.4) are equal. Theorem 4.2 provides a stronger result than this since the distributions need not belong to \mathcal{P} . In the class of all distribution unbiased estimators of the form

$$R_0(x), R_1(x), R_2(x), \dots, R_{n-1}(x), R_n(x)$$

for which the extended projections $\overline{\Pi}$ exists, the MLE (5.3) has smallest distribution variance. In the Hardy–Weinberg model estimators that do not belong to the family \mathcal{P} have been suggested. We consider the details in Section 5.2.

The choice of the n -binomial model \mathcal{P} was based on the assumptions that the data represented independent and identical trials. If either of these assumptions were grossly violated, the binomial model would not be appropriate. However, this model can be used when these assumptions hold approximately in the sense that there is a distribution $P_0 = \Pi R_0$ in \mathcal{P} that is close to the data generation distribution R_0 , that is, $D(R_0, P_0)$ is small. In this case, the MLE is the unique UMV[†]U estimator for P_0 .

5.2. Hardy–Weinberg model

For a single pair of alleles A and a, which occur with probabilities θ and $(1 - \theta)$ for $\theta \in (0, 1)$, the Hardy–Weinberg (HW) model defines the relative frequency of genotypes AA, Aa, and aa to be $\pi_1(\theta) = \theta^2$, $\pi_2(\theta) = 2\theta(1 - \theta)$, and $\pi_3(\theta) = (1 - \theta)^2$. For this example, we can take \mathcal{R} to be the collection of trinomial models with probabilities (π_1, π_2, π_3) for $\pi_1 + \pi_2 + \pi_3 = 1$ which can be represented by the simplex in 2-dimensional space. See Figure 1 for the simplex. The open circles in Figure 1 are the extended MLE $(\hat{\pi}_1, \hat{\pi}_2, \hat{\pi}_3) = (Y_1, Y_2, Y_3)/n$ for the trinomial with $n = 6$ trials, where Y_1 and Y_2 are the counts for AA and Aa. The solid curve in the simplex is the HW model

$$\mathcal{P} = \{(\pi_1, \pi_2, \pi_3) : \pi_1 = \theta^2, \pi_2 = 2\theta(1 - \theta), \pi_3 = (1 - \theta)^2\}$$

which is a one dimensional exponential family with canonical sufficient statistic $T = 2Y_1 + Y_2$ and canonical natural parameter $\log(\theta/(1 - \theta))$. Chow and Fong [5] find the UMVU for π_1 and π_3 using

$$E_\theta[(\hat{\pi}_1 - \theta^2)^2] + E_\theta[(\hat{\pi}_3 - (1 - \theta)^2)^2]$$

as squared-error loss. They show the UMVU is inadmissible by exhibiting a dominating estimator. Both the UMVU and the dominating estimator take values outside the HW model. In terms of distribution estimators, these are \mathcal{R} -valued estimators.

The extended MLE for the HW model is $\hat{\theta} = (2Y_1 + Y_2)/2n$ while the extended distribution MLE is $P_{\hat{\theta}}$ where P_0 is the degenerate distribution putting all its mass on $(0, 0, 6)$ (the lower left vertex) and P_1 is the degenerate distribution putting all its mass on $(6, 0, 0)$ (the lower right vertex). The extended HW MLE is represented by the solid dots in Figure 1.

Among the difficulties with the UMVU estimator and the dominating estimator is that there are other ways to define squared-error loss (using one bin or two other bins). These are avoided by using KL divergence. Since \mathcal{P} is an exponential family the extended MLE is the UMV[†]U for all \mathcal{P} -valued estimators but also for all \mathcal{R} -valued estimators since the projection exists for all points in the simplex other than the two lower vertices which satisfy the extended projection. As a comparison, the KL mean, represented by the dashed curve in Figure 1, lives outside the model so the extended MLE isn't KL unbiased. This is due to the curvature in the exponential family.

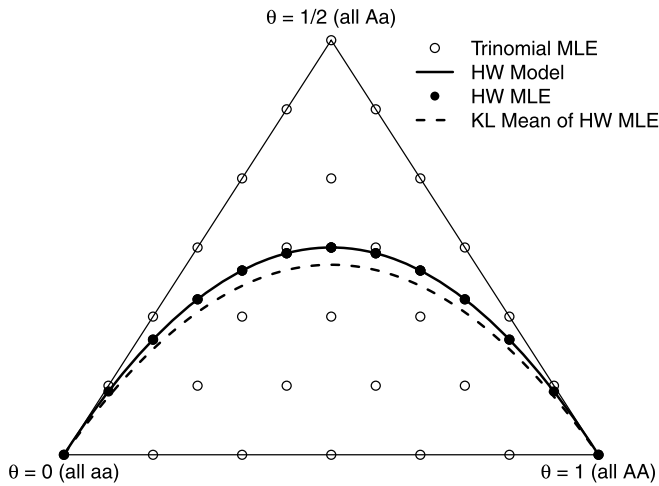


Figure 1. A Hardy–Weinberg (HW) model with $n = 6$ trials. The simplex represents the trinomial model space on (π_1, π_2, π_3) for $\pi_1 + \pi_2 + \pi_3 = 1$, while the solid curve is the HW model space on $\pi_1(\theta) = \theta^2$, $\pi_2(\theta) = 2\theta(1 - \theta)$, and $\pi_3(\theta) = (1 - \theta)^2$ for $0 < \theta < 1$. The open circles represent the (extended) MLE under the trinomial model $(\hat{\pi}_1, \hat{\pi}_2, \hat{\pi}_3) = (Y_1, Y_2, Y_3)/n$, and the solid dots are the (extended) MLE under the HW model $\hat{\theta} = (2Y_1 + Y_2)/2n$. The dashed curve shows the KL mean of the HW MLE for each value of θ .

5.3. Poisson distribution

The Poisson family of distributions is

$$\mathcal{P} = \left\{ P \in \mathcal{R} : P_\lambda(x) = e^{-\lambda} \frac{\lambda^x}{x!} \text{ for some } \lambda > 0 \right\},$$

where $x \in \mathbb{X} = \{0, 1, 2, \dots\}$.

Let X_1, \dots, X_n be a simple random sample from a Poisson distribution P_{λ_0} . The sum $S_n = X_1 + \dots + X_n$ is a complete sufficient statistic of the family. Although the Poisson family is typically parametrized by a single parameter, we consider estimates for the probability $\Pr(X_1 = i) = \lambda_0^i e^{-\lambda_0} / i!$ for some $i = 0, 1, \dots$. A crude but unbiased estimator is

$$\delta_{0i} = \begin{cases} 1 & \text{if } X_1 = i, \\ 0 & \text{otherwise.} \end{cases}$$

Given the sum S_n , X_1 is distributed as a binomial($S_n, 1/n$) random variable, the Rao–Blackwell theorem shows that

$$\delta_{1i} = E[\delta_{0i} | S_n] = \begin{cases} \binom{S_n}{i} \left(\frac{1}{n}\right)^i \left(1 - \frac{1}{n}\right)^{S_n - i} & \text{if } i \leq S_n, \\ 0 & \text{otherwise,} \end{cases}$$

is an unbiased estimator of $\Pr(X_1 = i)$. Since δ_{1i} depends on the complete sufficient statistic S_n only, it must be the unique MVUE of $\Pr(X_1 = i)$. Using the criterion of distribution unbiasedness, these anomalous estimators do not arise. Since S_n is the canonical statistic, the MLE $\bar{X} = S_n/n$ is the unique UMVU estimator for λ and the extended distribution MLE $P_{\bar{X}}$ is the UMVU estimator for P_λ where $P_{\bar{X}}$ is I_0 when $\bar{X} = 0$.

To show how the UMVU estimator can fail completely, Lehmann [11] considers the parameter $\delta = (P(X = 0))^3$ for $n = 1$. In this case, the unique UMVU estimator is $(-2)^x$. Since the sample consists of nonnegative integers this estimator is represented by the following sequence of real numbers

$$1, -2, 4, -8, 16, \dots$$

Parametric unbiasedness means that if the Poisson distribution that assigns probability $\delta^{1/3}$ to $P(X = 0)$ is used to assign probability to the terms in the sequence then $\delta = \arg \min_{a \in \mathbb{R}} E((-2)^X - a)^2$. That is, the parameter is the real number that is closest to this sequence in terms of mean square error. In addition, the weighted average of the above sequence is δ .

By focusing on distributions rather than the parameters that name the distributions these problems are avoided. The MLE, as a distribution estimator, is represented by the following sequence of probability distributions

$$I_0(x), e^{-1} \frac{1^x}{x!}, e^{-2} \frac{2^x}{x!}, e^{-3} \frac{3^x}{x!}, \dots$$

Distribution unbiasedness means that if the Poisson distribution P_λ is used to assign probability to the terms in the sequence then

$$P_\lambda = \arg \min_{P \in \mathcal{P}} E[D(P_\lambda, P)].$$

That is, the distribution that generates the data is the distribution in the exponential family that is closest to this sequence in terms of KL risk. Any other sequence of distributions with this property will have greater distribution variance.

6. Discussion

The distribution version of the Rao–Blackwell theorem 4.2 has been developed by analogy with important properties of mean square error for the parametric version. In particular, we have used a Pythagorean-type property for two asymmetric distribution-like functions: the KL divergence $D(\cdot, \cdot)$ and its expectation $E[D(\cdot, \cdot)]$. For exponential family \mathcal{P} , we have

$$D(\mathbf{R}, P) = D(\mathbf{R}, \Pi \mathbf{R}) + D(\Pi \mathbf{R}, P) \quad \forall P \in \mathcal{P}$$

while for all \mathcal{R}

$$E[D(\mathbf{R}, R)] = E[D(\mathbf{R}, ER)] + E[D(ER, R)]$$

so that the expectation operator E defined on \mathcal{R} -valued random variables for the KL risk plays the role of the projection operator Π for the KL divergence. Each operator is a map from a more complicated space to a simpler space, E from \mathcal{R} -valued random variables to a distribution in \mathcal{R} and Π from distributions in \mathcal{R} to a distribution in \mathcal{P} , that preserve the KL risk and KL divergence, respectively.

The restriction to exponential families is essentially required by the criterion of having a sufficient statistic of fixed dimension for all sample sizes n . Specifically, the Darrois–Koopman–Pitman theorem which follows from independent works of Darrois [7], Koopman [10] and Pitman [12] shows that when only continuous distributions are considered, the family of distributions of the sample has a sufficient statistic of dimension less than n if and only if the population distribution belong to the exponential family. Denny [8] shows that for a family of discrete distributions, if there is a sufficient statistic for the sample, then either the family is an exponential family or the sufficient statistic is equivalent to the order statistics.

The MLE is parameter-invariant which means that the same distribution is named by the parametric ML estimate regardless of the parameter chosen to index the family. One approach to studying parameter-invariant quantities is to use differential geometry (e.g., Amari [1] or Kass and Vos [9]). The parameter-invariant approach does not work well for parameter-dependent quantities such as bias and variance of parametric estimators. Our approach allows for the definition of parameter-free versions of bias and variance. Furthermore, the distribution version of the Rao–Blackwell provides two extensions: (1) minimum variance is taken over a larger class of estimators that includes estimators that are not required to take values in the model space \mathcal{P} , (2) the true distribution need not belong to \mathcal{P} .

The fact that the MLE is the unique uniformly minimum distribution variance unbiased estimator for exponential families distinguishes the MLE from other estimators. This is in contrast to asymptotic methods applied to MSE that can be used to show superior properties of the MLE but, being asymptotic results, do not apply uniquely to the MLE.

Asymptotically, MSE and KL risk are the same and the MSE can be viewed as an approximation to KL risk for large n . The distribution version of the Rao–Blackwell Theorem 4.2 provides support for Fisher’s claim of the superiority of the MLE even in small samples.

Acknowledgements

We thank the associate editor and external reviewers for their insightful comments and suggestions which have made great improvement on this paper.

References

- [1] Amari, S.-i. (1990). *Differential-Geometrical Methods in Statistics*. New York: Springer.
- [2] Blackwell, D. (1947). Conditional expectation and unbiased sequential estimation. *Ann. Math. Statist.* **18** 105–110. [MR0019903](#)
- [3] Brown, L.D. (1986). Fundamentals of statistical exponential families with applications in statistical decision theory. In *Institute of Mathematical Statistics Lecture Notes – Monograph Series* **9**. Hayward, CA: IMS. [MR0882001](#)

- [4] Čencov, N.N. (1982). *Statistical Decision Rules and Optimal Inference. Translations of Mathematical Monographs* **53**. Providence, RI: Amer. Math. Soc. Translation from the Russian edited by Lev J. Leifman. [MR0645898](#)
- [5] Chow, M.S. and Fong, D.K.H. (1992). Simultaneous estimation of the Hardy–Weinberg proportions. *Canad. J. Statist.* **20** 291–296. [MR1190573](#)
- [6] Csiszár, I. (1975). I -divergence geometry of probability distributions and minimization problems. *Ann. Probab.* **3** 146–158. [MR0365798](#)
- [7] Darmois, G. (1935). Sur les lois de probabilité à estimation exhaustive. *C. R. Math. Acad. Sci. Paris* **200** 1265–1266.
- [8] Denny, J.L. (1972). Sufficient statistics and discrete exponential families. *Ann. Math. Statist.* **43** 1320–1322. [MR0339366](#)
- [9] Kass, R.E. and Vos, P.W. (1997). *Geometrical Foundations of Asymptotic Inference. Wiley Series in Probability and Statistics: Probability and Statistics*. New York: Wiley. [MR1461540](#)
- [10] Koopman, B.O. (1936). On distributions admitting a sufficient statistic. *Trans. Amer. Math. Soc.* **39** 399–409. [MR1501854](#)
- [11] Lehmann, E.L. (1983). *Theory of Point Estimation. Wiley Series in Probability and Mathematical Statistics: Probability and Mathematical Statistics*. New York: Wiley. [MR0702834](#)
- [12] Pitman, E.J.G. (1936). Sufficient statistics and intrinsic accuracy. *Math. Proc. Cambridge Philos. Soc.* **32** 567–579.
- [13] Wu, Q. and Vos, P. (2012). Decomposition of Kullback–Leibler risk and unbiasedness for parameter-free estimators. *J. Statist. Plann. Inference* **142** 1525–1536. [MR2891504](#)

Received May 2013 and revised April 2014