Ann Inst Stat Math (2015) 67:541–555
DOI 10.1007/s10463-014-0464-y

On the equivariance criterion in statistical prediction

Haojin Zhou · Tapan K. Nayak

Received: 8 April 2013 / Revised: 18 February 2014 / Published online: 26 April 2014

Abstract This paper presents a general development of the basic logic of equivariance for a parametric point prediction problem. We propose a framework that allows the set of possible predictions as well as the losses to depend on the data and then explore the nature and properties of relevant transformation groups for applying the functional and formal equivariance principles. We define loss invariance and predictive equivariance appropriately and discuss their ramifications. We describe a structure of equivariant predictors in terms of maximal invariants and present a method for deriving minimum risk equivariant predictors. We explore the connections between equivariance and risk unbiasedness and show that uniquely best risk unbiased predictors are almost equivariant. We apply our theoretical results to some illustrative examples.

Keywords Loss invariance · Maximal invariant · Minimum risk · Risk unbiased · Transformation group

1 Introduction

Consider a general prediction problem in a parametric setup: predict the unobserved value of a random variable Y based on the observed value x of a random vector X, assuming that the joint density of X and Y is $f(x, y|\theta)$ (with respect to some sigma-finite measure), where $\theta \in \Theta$ is an unknown parameter, possibly vector-valued. A

H. Zhou
Department of Biostatistics and Bioinformatics, Duke University Medical Center, 2424 Erwin Road, Suite 1102, 11102E Hock Plaza, Duke Box 2721, Durham, NC 27705, USA
e-mail: haojin.zhou@duke.edu

T. K. Nayak (✉)
Department of Statistics, George Washington University, Washington, DC 20052, USA
e-mail: tapan@gwu.edu

special case that has received much attention is where X and Y are independent given θ. The books by Aitchison and Dunsmore (1975) and Geisser (1993) discuss frequentist and Bayesian methods, and Bjornstad (1990) gives a broad review of predictive likelihood approaches. Prediction is a common problem, which Pearson (1920) called "the fundamental problem of practical statistics." Also, some authors (e.g., Geisser 1993) have argued that in practice, predictions are more relevant than inferences about parameters or indexes of statistical distributions, which dominate the statistical literature. From theoretical perspective, the prediction framework is significant due to its generality. Estimation of a parametric function $g(\theta)$ is a special case, where $Y = g(\theta)$, and hence, $f(y|x, \theta)$ is degenerate and independent of x. It also covers many non-standard problems, where the quantity of inferential interest depends on x, θ and some unobserved variables, as in prediction under mixed linear models, estimation under superpopulation models, estimation after selection and loss estimation (see Yatracos 1992; Bjornstad 1996; Nayak 2000).

This paper is focused on equivariant prediction, which has been investigated by Hora and Buehler (1967), Takada (1981, 1982), Eaton and Sudderth (2001, 2004) and others, providing many interesting mathematical results. However, our primary goal is to revisit the logic of equivariance in prediction context and develop it more broadly. The equivariance argument rests on the principles of (i) functional equivariance, which says that the action taken in a decision problem should not depend on units of measurement or the coordinate system used to specify the problem and (ii) formal equivariance, which says that if two problems have the same structure, one should use the same decision rule. These principles have been developed extensively for a standard decision problem, formally described by $(\mathcal{X}, \mathcal{P}, \Theta, \mathcal{D}, L)$, where $\mathcal{X}$ is a sample space, $\mathcal{P}$ is a family of distributions with parameter θ and parameter space Θ, $\mathcal{D}$ is a decision space, and L is a loss function on $\mathcal{D} \otimes \Theta$. Some important references are as follows: Hall et al. (1965), Hora and Buehler (1966), Berger (1985), Eaton (1989), Wijsman (1990) and Lehmann and Casella (1998). To apply equivariance logic, one uses groups of transformations to change the coordinate system while preserving the formal structure of the problem.

For applying the two equivariance principles to statistical prediction, first we need a *formal description* of a prediction problem and then we need to explore *transformation groups* that are suitable for "changing the coordinates" while preserving the formal structure of the problem. For wide-ranging applicability, we try to keep the problem description as well as the transformation groups as general as possible. One new aspect of our approach is that the set of possible predictions and the loss function are allowed to depend on x, the observed data. We develop our arguments in Sect. 2, leading to a defining criterion for predictive equivariance; see (7) and (8). We motivate our framework with examples and bring out some interesting structures and properties of relevant transformation groups. One source for additional intricacies is that for each transformation of X, there may correspond multiple transformations of Y. In Sect. 3, we discuss some properties and risk optimality of equivariant predictors. Section 4 gives some illustrative examples. In Sect. 5, we discuss certain connections between equivariance and risk unbiasedness and prove that uniquely best risk unbiased predictors are almost equivariant. Section 6 is devoted to some concluding remarks.

For ease of readability, we leave out topological and measurability aspects, which are also not crucial in our discussion.

2 Equivariance in statistical prediction

2.1 Formalization of prediction problems

Hora and Buehler (1967) describe a prediction problem by $(\mathcal{X}, \mathcal{Y}, \mathcal{P}, \Theta, \mathcal{D}, L)$, where $\mathcal{X}$ and $\mathcal{Y}$ are sample spaces of X and Y, $\mathcal{P}$ is a distribution family for $Z = (X, Y)$ with θ as the parameter, Θ is the parameter space, $\mathcal{D}$ is a decision space, and L is a loss function defined on $\mathcal{D} \otimes \mathcal{Y} \otimes \Theta$. Eaton and Sudderth (2001) use a similar framework, but allow the loss function to depend on x. Both papers assume $\mathcal{D} = \mathcal{Y}$, implying that the decision set does not depend on x. They also assume that a group G acts on each of $\mathcal{X}$, $\mathcal{Y}$ and Θ, and thus, the transformations of Y do not involve X. These conditions may be overly restrictive, as the following examples show.

Example 1 Let $X_1 \leq \cdots \leq X_n$ be the order statistics of a random sample of size n from an unknown distribution. Consider predicting $Y = X_m$ based on $X = (X_1, \ldots, X_k)$, where $1 \leq k < m \leq n$ (cf., Takada 1981). Here, $Y \geq X_k$ and hence a natural decision space, when x is observed, is $\mathcal{D}_x = \{d : d \geq x_k\}$, which is data dependent.

Next, let W be a future observation from the same distribution and consider predicting $Y = I(W > X_r)$ based on $X_1, \ldots X_n$, where $1 \leq r \leq n$. Here, Y is binary, but $[0, 1]$ is a reasonable decision space, which is different from $\mathcal{Y}$. Eaton and Sudderth (2001) discuss estimation of $\psi(x, \theta) = E_\theta[Y|x]$, under an exponential model.

Example 2 Suppose $X_1, \ldots, X_{n+1}$ are iid $N_p(\mu, \Sigma)$. Partition X_{n+1} into $X_{n+1}^{(1)}$ and $X_{n+1}^{(2)}$, with dimensions m and q, respectively, where $m + q = p$, and let Σ_{ij}, $i, j = 1, 2$, denote the corresponding partition of Σ. Consider predicting $Y = X_{n+1}^{(2)}$ based on $X = (X_1, \ldots, X_n, X_{n+1}^{(1)})$ under $L(d, y, \theta) = ||\Sigma_{22}^{-1}(d - y)||^2$. For this problem, Takada (1982) gives the best equivariant predictor under the transformations: $X_i \to CX_i + b$, $i = 1, \ldots, n$, $X_{n+1}^{(1)} \to C_{11}X_{n+1}^{(1)} + b_1$, $Y \to C_{22}Y + C_{21}X_{n+1}^{(1)} + b_2$, where C is a lower triangular positive definite matrix, with partition matrices $\{C_{ij}\}$ and $b \in R^p$, with partition vectors b_1 and b_2; these partitions conform with the partition of X_{n+1}. The main point here is that the transformations of Y involve both Y and X.

Recognizing that possible decisions may depend on x, we define a prediction problem as $(\mathcal{X}, \mathcal{Z}, \mathcal{P}, \Theta, \{\mathcal{D}_x\}, L)$, where $\mathcal{Z}$ is a sample space of $Z = (X, Y)$, $\mathcal{X}, \mathcal{Y}, \mathcal{P}$ and Θ are as in Hora and Buehler (1967) (described above), $\mathcal{D}_x$ is a decision space for observed x, and $L = L(d, y, x, \theta)$ is a loss function defined for all $(x, y) \in \mathcal{Z}$, $\theta \in \Theta$, and $d \in \mathcal{D}_x$. We do not require $\mathcal{D}_x = \mathcal{Y}_x$, where $\mathcal{Y}_x = \{y : (x, y) \in \mathcal{Z}\}$ is the sample space of Y when $X = x$.

Assumption 1 The marginal model for X, i.e., $\mathcal{P}_X = \{f(x|\theta), \theta \in \Theta\}$ is identifiable.

Remark 1 Assumption 1 concerns estimability of θ based on X and implies that "true θ," "true marginal distribution of X" and the "true joint distribution of (X, Y)" are equivalent, in the sense that any one implies the others. Also, the case of interest is where the true predictive distribution $f(y|x, \theta)$ depends on θ and thus unknown.

2.2 Transformations of sample spaces

To apply equivariance logic, we need to bring in transformations of Z, θ and d that preserve the mathematical problem. One of our goals is to bring out essential structures and properties of applicable transformation groups, without prejudice. In that direction, we begin with a group G of transformations of $\mathcal{Z}$. First, to preserve the problem structure G must preserve $\mathcal{P}$.

Definition 1 A group G of 1-1 transformations of $\mathcal{Z}$ onto $\mathcal{Z}$ is said to preserve $\mathcal{P}$ if for each $g \in G$ and $\theta \in \Theta$, there exists $\theta' \in \Theta$ such that

$$Z \sim f(z|\theta) \Rightarrow Z_* = g(Z) \sim f(z_*|\theta'). \tag{1}$$

If (1) holds, then for each $g \in G$, we can define $\bar{g} : \Theta \mapsto \Theta$ by $\theta' = \bar{g}(\theta)$. Next, we observe that for formal equivariance, it is not enough for $g : \mathcal{Z} \mapsto \mathcal{Z}$ to be 1-1, onto and $\mathcal{P}$ preserving. This is because like Z, $Z_* = g(Z)$ must have an observable part X_* and an unobservable component Y_*. Obviously, X_* must not depend on Y, while Y_* may depend on both X and Y. Thus, we impose the following:

Assumption 2 Each $g \in G$ consists of two functions $h_g : \mathcal{X} \mapsto \mathcal{X}$ and $k_g : \mathcal{Z} \mapsto \mathcal{Y}$ such that $g(x, y) = (h_g(x), k_g(x, y))$ for all $(x, y) \in \mathcal{Z}$.

For notational simplicity, we shall often omit the subscript g of h and k. For any fixed x, let $k_{g|x}(y) = k_g(x, y)$, viewed as a function only of y. Clearly, $k_{g|x} : \mathcal{Y}_x \mapsto \mathcal{Y}_{h_g(x)}$. While in most papers, k_g is a function only of Y, Takada (1982) allows it to depend on X, but assumes that (i) the sample space of Y is independent of x, (ii) $\mathcal{D} = \mathcal{Y}$ and (iii) $\{k_{g|x}\}$ is a group. We do not make these assumptions. Next, we ascertain certain properties of h_g, k_g and G. The following conclusions can be easily verified as consequences of just g being 1-1 and onto.

Lemma 1 *Let* $g(x, y) = (h_g(x), k_g(x, y)) : \mathcal{Z} \mapsto \mathcal{Z}$ *be* 1-1 *and onto. Then,* (i) $h_g : \mathcal{X} \mapsto \mathcal{X}$ *is onto,* (ii) $k_g(x, y) : \mathcal{Z} \mapsto \mathcal{Y}$ *is onto and* (iii) *for any fixed* x, $k_{g|x}(y) : \mathcal{Y}_x \mapsto \mathcal{Y}_{h_g(x)}$ *is* 1-1.

Example 3 This example shows that a single 1-1 and onto function $g = (h, k)$ does not imply that h and k are 1-1. Let $\mathcal{X} = \mathcal{Y} = \{0, 1, 2, \dots\}$, $\mathcal{Z} = \mathcal{X} \times \mathcal{Y}$ and

$$g(x, y) = \begin{cases} \left(\frac{x-1}{2}, 2y\right) & \text{if } x \text{ is odd,} \\ \left(\frac{x}{2}, 2y + 1\right) & \text{if } x \text{ is even.} \end{cases}$$

Here, $g : \mathcal{Z} \mapsto \mathcal{Z}$ is 1-1 and onto and has the form $g(x, y) = (h(x), k(x, y))$, but $h : \mathcal{X} \mapsto \mathcal{X}$ and $k(x, y) : \mathcal{Z} \mapsto \mathcal{Y}$ are not 1-1, and for any fixed x, $k(x, y) : \mathcal{Y} \mapsto \mathcal{Y}$ is not onto, as a function of y.

As we show next, more can be said about h and k when g is a member of a group G satisfying Assumption 2, where the inverse of g also has the structure $g^{-1}(x, y) = (h_{g^{-1}}(x), k_{g^{-1}}(x, y))$. This is not true for the g in Example 3. We shall denote composition of transformations by $\circ$. Note that for any $g_1 = (h_1, k_1)$ and $g_2 = (h_2, k_2)$ in G,

$$g_1 \circ g_2(x, y) = (h_1 \circ h_2(x), k_1(h_2(x), k_2(x, y))). \tag{2}$$

Lemma 2 *Let G be a group of* 1-1 *transformations of $\mathcal{Z}$ onto itself, satisfying Assumption* 2. *Then, for each* $g = (h_g, k_g) \in G$, (i) h_g *is* 1-1 *and* (ii) *for each* x, $k_{g|x}(y)$ *is onto* $\mathcal{Y}_{h_g(x)}$.

Proof For any $g = (h_g, k_g) \in G$, using its inverse $g^{-1} = (h_{g^{-1}}, k_{g^{-1}})$ in (2), we get

$$(x, y) = g^{-1} \circ g(x, y) = (h_{g^{-1}} \circ h_g(x), k_{g^{-1}}(h_g(x), k_g(x, y))), \tag{3}$$

for all $(x, y) \in \mathcal{Z}$. Thus, $h_{g^{-1}}$ is the inverse of h_g, and hence, h_g must be 1-1.

As $(x, y) = g^{-1} \circ g(x, y) = g \circ g^{-1}(x, y)$, it follows from (3) that $k_{g|x} : \mathcal{Y}_x \mapsto \mathcal{Y}_{h_g(x)}$ and $k_{g^{-1}|h_g(x)} : \mathcal{Y}_{h_g(x)} \mapsto \mathcal{Y}_x$ are inverse of each other and hence those must be onto. □

Let $H = \{h : (h, k) \in G$ for some $k\}$, and for each $h \in H$, let $G_h = \{g \in G : g = (h, k)$ for some $k\}$ and $K_h = \{k : (h, k) \in G\}$. In most applications, G_h (and correspondingly K_h) is a singleton, but that need not be true in general, as we see next.

Example 4 Let $X_1, \ldots, X_n, X_{n+1}$ be i.i.d. $N(0, \sigma^2)$, $\sigma > 0$. Consider predicting $Y = \sum_{i=1}^{n+1} X_i$ based on $X = (X_1, \ldots, X_n)$. Here, $(X, Y) \sim N_{n+1}(0, \sigma^2 D)$ with $D = ((d_{ij}))$, where $d_{i,i} = d_{i,n+1} = d_{n+1,i} = 1, i = 1, \ldots, n, d_{n+1,n+1} = n + 1$ and $d_{i,j} = 0$ in all other cases. This model is invariant under $G = \{g_c, c > 0\}$, where $g_c(x, y) = (cx, cy)$.

However, a larger transformation group also preserves the model. For each $c > 0$, let $h_c(x) = cx$, $k_{1c}(x, y) = cy$ and $k_{2c}(x, y) = c(2\sum_{i=1}^{n} x_i - y)$. Let $g_{ic}(x, y) = (h_c(x), k_{ic}(x, y))$, $G_i = \{g_{ic}, c > 0\}, i = 1, 2$ and $G_* = G_1 \cup G_2$ (note that $G_1 = G$). It is easy to verify that G_* is a group (under composition) and it preserves the distribution family of (X, Y). Note that here for any $c > 0$, $G_{h_c} = \{g_{1c}, g_{2c}\}$ has two elements.

Proposition 1 *If G is a group, then H and* $G_e = \{g \in G : g = (e, k)$ *for some* $k\}$ *are also groups, where e is the identity transformation of* $\mathcal{X}$.

This result is easy to verify, and hence, we omit its proof. For $h \neq e$, G_h is not closed under composition and so it is not a group. However, we note that if $g = (e, k)$ and $g_1 = (h, k_1)$, then $g_1 \circ g = (h, k_1(e, k)) \in G_h$, and for any $g_1 = (h, k_1) \in G_h$ and $g_2 = (h^{-1}, k_2) \in G_{h^{-1}}$, $g_2 \circ g_1 = (e, k_2(h, k_1)) \in G_e$. Also, if $g = (h, k) \in G_h$ and $g_1 = (h, k_1) \in G_h$, then $g_0 = g^{-1} \circ g_1 \in G_e$ and $g \circ g_0 = g_1$. Thus, G_h can be generated from G_e and any member of G_h. In summary, we have:

Proposition 2 *If $g = (h, k) \in G$, then $gG_e = G_e g = G_h$, and $g_* G_h = G_h g_* = G_e$ for any $g_* \in G_{h^{-1}}$.*

Recall that $\bar{g}$ is defined via $(X, Y) \sim f(x, y|\theta) \Rightarrow g(X, Y) = (h(X), k(X, Y)) = (X_*, Y_*) \sim f(x_*, y_*|\bar{g}(\theta))$. Also, by Assumption 1, the marginal distributions of X and X_* determine the joint distributions of (X, Y) and (X_*, Y_*), respectively (see Remark 1). Then, since h determines the distribution of X_*, from $X \sim f(x|\theta)$, we get the following:

Proposition 3 *If G preserves $\mathcal{P}$, then for any $g = (h, k) \in G$, the induced transformation $\bar{g}$ of Θ is determined solely by h.*

In view of this result, the transformation of Θ induced by $g = (h, k)$ shall often be denoted by $\bar{h}$ instead of $\bar{g}$. In fact, $\bar{h}$ can be defined using $X \sim f(x|\theta) \Rightarrow h(X) = X_* \sim f(x_*|\bar{h}(\theta))$. It is easy to see $\bar{h} : \Theta \mapsto \Theta$ is 1-1 and onto and $\bar{H} = \{\bar{h} : h \in H\}$ is a group.

2.3 Invariant loss functions

If G preserves $\mathcal{P}$, then each $g \in G$ gives a new coordinate system for X and Y and also for θ, via $\bar{h}$. Next, to apply equivariance logic, for each $g \in G$, we need a matching transformation of possible decisions that preserves the loss function $L(d, y, x, \theta)$. This means that for each $g = (h, k) \in G$, $x \in \mathcal{X}$ and $d \in \mathcal{D}_x$, there exists $d_* \in \mathcal{D}_{h(x)}$ such that

$$L(d, y, x, \theta) = L(d_*, k(x, y), h(x), \bar{h}(\theta)) \quad \text{for all } y \in \mathcal{Y}_x, \theta \in \Theta. \tag{4}$$

Assumption 3 For each $x \in \mathcal{X}$, $\mathcal{D}_x$ is identifiable with respect to the loss function, i.e., if $d_1, d_2 \in \mathcal{D}_x$ and $L(d_1, y, x, \theta) = L(d_2, y, x, \theta)$ for all $y \in \mathcal{Y}_x$ and $\theta \in \Theta$, then $d_1 = d_2$.

This assumption prevents redundancy in $\mathcal{D}_x$ and ensures that d_* in (4) is unique, when it exists, in which case we can define a function: $\varphi_{g|x}(d) : \mathcal{D}_x \mapsto \mathcal{D}_{h(x)}$ via $\varphi_{g|x}(d) = d_*$.

Definition 2 A loss function $L(d, y, x, \theta)$ is said to be invariant under G if for each $g = (h, k) \in G$ and $x \in \mathcal{X}$, there exists $\varphi_{g|x} : \mathcal{D}_x \mapsto \mathcal{D}_{h(x)}$ such that

$$L(d, y, x, \theta) = L(\varphi_{g|x}(d), k(x, y), h(x), \bar{h}(\theta)) \quad \text{for all } d \in \mathcal{D}_x, y \in \mathcal{Y}_x, \theta \in \Theta. \tag{5}$$

The transformations of decisions, induced by G and loss invariance, have a group structure, if we consider the augmented space $\mathcal{D}_* = \{(x, d) : x \in \mathcal{X}, d \in \mathcal{D}_x\}$. For any $g \in G$, let $\tilde{g} : \mathcal{D}_* \mapsto \mathcal{D}_*$ be defined by $\tilde{g}(x, d) = (h(x), \varphi_{g|x}(d))$. Also, let $\tilde{G} = \{\tilde{g} : g \in G\}$.

Lemma 3 *Suppose the loss function is invariant under G and Assumption 3 holds. Then,* (i) *for any g_1 and g_2 in G, $\tilde{g}_1 \circ \tilde{g}_2 = \widetilde{g_1 \circ g_2}$ and* (ii) $(\tilde{g})^{-1} = \widetilde{g^{-1}}$ *for all $g \in G$.*

Proof Take any $g_1 = (h_1, k_1)$ and $g_2 = (h_2, k_2)$. Then, using (2) and (5), we get

$$\begin{aligned}
&L(\varphi_{g_1 \circ g_2|x}(d), k_1(h_2(x), k_2(x, y)), h_1 \circ h_2(x), \overline{h_1 \circ h_2}(\theta)) \\
&\quad = L(d, y, x, \theta) \\
&\quad = L(\varphi_{g_2|x}(d), k_2(x, y), h_2(x), \overline{h_2}(\theta)) \\
&\quad = L(\varphi_{g_1|h_2(x)}(\varphi_{g_2|x}(d)), k_1(h_2(x), k_2(x, y)), h_1 \circ h_2(x), \overline{h_1 \circ h_2}(\theta)),
\end{aligned}$$

and hence,

$$\varphi_{g_1 \circ g_2|x}(d) = \varphi_{g_1|h_2(x)}(\varphi_{g_2|x}(d)) \quad \text{for all } (x, d) \in \mathcal{D}, \tag{6}$$

by Assumption 3. The lemma now follows easily from the definition of $\tilde{g}$. □

Theorem 1 *If L is invariant under G, then $\tilde{G} = \{\tilde{g} : g \in G\}$ is a group of* 1-1 *transformations of $\mathcal{D}_*$ onto $\mathcal{D}_*$.*

This theorem follows easily from Lemma 3. Takada (1982) assumed that $\varphi_{g|x} = k_{g|x}$ and hence $\tilde{g} = g$, which actually holds in the following special case.

Lemma 4 *Suppose $\mathcal{D}_x = \mathcal{Y}_x$ for all $x \in \mathcal{X}$, L is invariant and $L(d, y, x, \theta) = 0$ if and only if $d = y$. Then, for all $g \in G$ and $x \in \mathcal{X}$, $\varphi_{g|x} = k_{g|x}$ and hence $\tilde{g} = g$.*

Proof Under the assumptions of the lemma, for any $g = (h, k) \in G$, we get

$$0 = L(d, d, x, \theta) = L(\varphi_{g|x}(d), k_{g|x}(d), h(x), \bar{h}(\theta))$$

for all $\theta \in \Theta$, $x \in \mathcal{X}$ and $d \in \mathcal{Y}_x$. This implies $\varphi_{g|x} = k_{g|x}$, from the condition for $L = 0$. □

2.4 Equivariance criterion

Definition 3 A prediction problem $(\mathcal{X}, \mathcal{Z}, \mathcal{P}, \Theta, \{\mathcal{D}_x\}, L)$ is said to be invariant under a transformation group G, satisfying Assumption 2, if Definitions 1 and 2 hold.

Suppose a prediction problem is invariant and the original problem is transformed using some $g = (h, k) \in G$ and corresponding $\bar{h}$ and $\{\varphi_{g|x}\}$. The logic of functional equivariance says that if we choose decision d after observing x in the original problem, then in the transformed problem we should choose the decision $\varphi_{g|x}(d)$ when $h(x)$ is observed. On the other hand, the formal equivariance principle tells us to use the same decision rule δ in the two problems as they have the same formal structure. These arguments lead to the following:

Definition 4 Suppose a prediction problem is invariant under G. Then, a predictor $\delta(X)$ is said to be equivariant under G if for all $g = (h, k) \in G$,

$$\delta(h(x)) = \varphi_{g|x}(\delta(x)) \quad \text{for all } x \in \mathcal{X}. \tag{7}$$

We should note a subtle feature of the condition in (7). Specifically, the left side of (7) depends on g only through h and thus the right side must also be so, i.e., for each $x \in \mathcal{X}$ and $h \in H$, $\varphi_{g|x}(\delta(x))$ must be the same for all $g \in G_h$. Recalling the definition of $\tilde{g}$, the condition for δ to be equivariant can also be stated as:

$$\tilde{g}(x, \delta(x)) = (h(x), \delta(h(x)) \quad \text{for all } g = (h, k) \in G \text{ and } x \in \mathcal{X}. \tag{8}$$

Lemma 5 (i) *A necessary condition for $\delta(X)$ to be an equivariant predictor is that*

$$\delta(x) = \varphi_{g|x}(\delta(x)) \quad \textit{for all } g \in G_e \textit{ and } x \in \mathcal{X}. \tag{9}$$

(ii) *If $\delta(X)$ satisfies* (9), *then for all $x \in \mathcal{X}$ and $h \in H$, $\varphi_{g|x}(\delta(x))$ is the same for all $g \in G_h$.*

Proof Part (i) follows directly from (7). To prove part (ii), let $g_1, g_2 \in G_h$ for some h and $g_1 \neq g_2$. By Proposition 2, $g_1 = g_2 \circ g_0$ for some $g_0 \in G_e$. Now, (9) and Lemma 3 yield

$$\tilde{g}_1(x, \delta(x)) = \tilde{g}_2 \circ \tilde{g}_0(x, \delta(x)) = \tilde{g}_2(x, \delta(x)),$$

which shows that $\varphi_{g_i|x}(\delta(x))$, $i = 1, 2$, are the same. □

3 Risk comparison and optimality

For a predictor δ, let $R(\delta, \theta) = E_\theta[L(\delta(X), Y, X, \theta)]$ denote its risk function, where the expectation is with respect to both X and Y. The transformation groups H and $\bar{H}$ define partitions of $\mathcal{X}$ and Θ. For any $\theta_0 \in \Theta$, the θ_0-orbit, to be denoted $\text{Orb}(\theta_0)$, is defined as $\text{Orb}(\theta_0) = \{\bar{h}(\theta_0) : \bar{h} \in \bar{H}\}$. Two points $\theta_1, \theta_2 \in \Theta$ are said to be equivalent if they belong to the same orbit, i.e., $\theta_2 = \bar{h}(\theta_1)$ for some $\bar{h} \in \bar{H}$. Similarly, for any $x_0 \in \mathcal{X}$, the x_0-orbit is defined as $\text{Orb}(x_0) = \{h(x_0) : h \in H\}$. The orbits define partitions of Θ and $\mathcal{X}$. The risk function of any equivariant decision rule is known to be constant on each orbit of Θ (see Berger 1985, p. 396). Similarly, one can verify the following.

Theorem 2 *If δ is equivariant, then $R(\delta, \theta) = R(\delta, \bar{h}(\theta))$ for all $\bar{h} \in \bar{H}, \theta \in \Theta$. If Θ is transitive, i.e., Θ has a single orbit under $\bar{H}$, then any equivariant predictor has a constant risk.*

A structure of equivariant predictors can be revealed via maximal invariants. A statistic $T(X)$ is invariant with respect to H if $T(h(x)) = T(x)$ for all $x \in \mathcal{X}$ and $h \in H$ and $T(X)$ is a maximal invariant if it is invariant and $T(x_1) = T(x_2)$ implies $x_1 = h(x_2)$ for some $h \in H$. A maximal invariant is not unique but all maximal invariants induce the same partition of $\mathcal{X}$ as the one induced by H. Also, any invariant statistic is a function of a maximal invariant, and its distribution depends on θ only through a maximal invariant on Θ (see Berger 1985, Sec. 6.5).

Lemma 6 *If δ is equivariant, then on each orbit A of $\mathcal{X}$, $\delta(x)$ is fully determined by its value at any one point in A.*

Proof Consider any orbit A and any fixed point x_0 in it. Then, for any other point x in A, $x = h(x_0)$ for some $h \in H$. Now, using any $g \in G_h$ and equivariance of δ, we obtain $\delta(x) = \delta(h(x_0)) = \varphi_{g|x_0}(\delta(x_0))$, which shows that $\delta(x)$ can be obtained from $\delta(x_0)$. □

For each orbit A of $\mathcal{X}$, fix a (any) point $x(A) \in A$. Assume that H acts freely on $\mathcal{X}$, i.e., for any $x_1, x_2 \in \mathcal{X}$, there is at most one $h \in H$ such that $x_1 = h(x_2)$. Then, all equivariant predictors can be constructed as follows: (i) For each fixed point $x(A)$, take a decision $d_{x(A)}$ in $\mathcal{D}^e_{x(A)} = \{d \in \mathcal{D}_{x(A)} : d = \varphi_{g|x(A)}(d)$ for all $g \in G_e\}$ and assign it to $x(A)$, i.e., define $\delta(x(A)) = d_{x(A)}$. (ii) For any other $x \in \mathcal{X}$, first find the orbit A that contains x, then find the unique $h \in H$ such that $x = h(x(A))$ and finally take any $g \in G_h$ and define $\delta(x) = \varphi_{g|x(A)}(d_{x(A)})$.

As in the proof of Lemma 5(ii), it can be seen that δ as constructed above is well defined, i.e., it does not depend on the choice of g in G_h. Note that if $\mathcal{D}^e_{x(A)}$ is empty for some A, then an equivariant predictor does not exist, by Lemma 5(i). This is a new feature of predictive equivariance. In contrast, equivariant rules always exist in a standard decision problem.

To verify that δ as defined above is equivariant, take any $x \in \mathcal{X}$ and any $g = (h, k) \in G$. Let x_0 be the fixed point of the orbit that contains x. Then, there exists $g_1 = (h_1, k_1) \in G$ such that $h_1(x_0) = x$ and by construction, $(x, \delta(x)) = \tilde{g}_1(x_0, \delta(x_0))$. Let $g_2 = g \circ g_1 = (h_2, k_2)$ and note that $h_2(x_0) = h(x)$ and hence $\tilde{g}_2(x_0, \delta(x_0)) = (h(x), \delta(h(x))$, by construction. Thus, $\tilde{g}(x, \delta(x)) = \tilde{g} \circ \tilde{g}_1(x_0, \delta(x_0)) = (h(x), \delta(h(x))$ and hence δ satisfies (8).

The preceding characterization of equivariant predictors is useful for obtaining a minimum risk equivariant predictor when H acts freely on $\mathcal{X}$ and Θ is transitive under $\bar{H}$. By Theorem 2, the risk function of any equivariant predictor is independent of θ, so it can be calculated under any fixed $\theta_0 \in \Theta$. Now, if $T = T(X)$ is a maximal invariant statistics, then

$$R(\delta, \theta) = E^Z_{\theta_0} L(\delta(X), Y, X, \theta) = E^T_{\theta_0} E^{Z|T=t}_{\theta_0}[L(\delta(X), Y, X, \theta)|T(X) = t]. \quad (10)$$

Note that $T(X) = t$ defines a maximal invariant partition set, say A_t, and the conditional expectation depends on δ only through $\delta(x)$ for $x \in A_t$, which in turn depends only on $\delta(x(A_t))$. So, finding a best equivariant predictor reduces to minimizing the conditional expectation in (10), for each t, with respect to $\delta(x(A_t))$, taking a value in $\mathcal{D}^e_{x(A_t)}$.

4 Examples

Example 1 (continued). As before, let $X_1 \leq \cdots \leq X_n$ be the order statistics of a random sample of size n. Now, assume that the population distribution is Exp(α, σ) with pdf

$$f_{\alpha,\sigma}(x) = \frac{1}{\sigma}\mathrm{e}^{-\frac{x-\alpha}{\sigma}}, \quad x > \alpha,\ \alpha \in \mathbb{R},\ \sigma > 0.$$

(a) Consider predicting $Y = X_m$ based on $X = (X_1, \ldots, X_k)$, where $1 \le k < m \le n$. Here, the model is invariant under location-scale transformations

$$\begin{aligned} g_{a,b}(x_1, \ldots, x_k, y) &= (ax_1 + b, \ldots, ax_k + b, ay + b) \\ &= (ax + b\vec{1}, ay + b), \quad a > 0,\ b \in \mathbb{R}. \end{aligned}$$

Suppose the loss function is $L(d, y, \sigma) = (\frac{d-y}{\sigma})^2$. Then, the problem is invariant with $\bar{g}_{a,b}(\mu, \sigma) = (a\alpha + b, a\sigma)$ and $\varphi_{g_{a,b}|x}(d) = ad + b$. By (7), $\delta(x)$ is an equivariant predictor if

$$\delta(ax + b\vec{1}) = a\delta(x) + b \quad \text{for all } x, a > 0, b \in \mathbb{R}. \tag{11}$$

Let $V_i = (n - i + 1)(X_i - X_{i-1})$, $i = 2, \ldots, n$, and $S = \sum_{i=k+1}^{m} \frac{1}{n-i+1} V_i$. Note that $(X_1, Q = \sum_{i=2}^{k} V_i)$ is complete sufficient and $Y = X_k + S$. It follows that $T(X) = (\frac{X_3 - X_1}{X_2 - X_1}, \ldots, \frac{X_k - X_1}{X_2 - X_1})$ is a maximal invariant and $\delta(x)$ satisfies (11) if and only if

$$\delta(x) = x_k + w(t)Q(x) \quad \text{for all } x \tag{12}$$

and some function w (see Lehmann and Casella 1998, Sec. 3.3). As Q and T are independent, by Basu's theorem, and the parameter space is transitive, for any δ of the form (12), we get

$$\begin{aligned} R(\delta, \alpha, \sigma) &= E_{0,1}[L(\delta(X), Y, 0, 1)|T = t] \\ &= w^2(t)E_{0,1}Q^2 - 2w(t)E_{0,1}[QS] + E_{0,1}[S^2]. \end{aligned} \tag{13}$$

Minimizing (13) w.r.t. $w(t)$ for each t, we obtain the best equivariant predictor of Y as

$$\delta_0(X) = X_k + \frac{Q}{k} \sum_{i=k+1}^{m} \frac{1}{n - i + 1}.$$

We may remark that the preceding approach may be used to obtain a best equivariant predictor when more generally the loss is a function of $\frac{d-\mu}{\sigma}$ and $\frac{y-\mu}{\sigma}$.

(b) Consider predicting $Y = I(W > X_r)$ based on $X = (X_1, \ldots, X_k)$, where W is a future observation from the same $\mathrm{Exp}(\alpha, \sigma)$ distribution and $r \le k \le n$. Note that $P_{\alpha,\sigma}[Y = 1|x] = \mathrm{e}^{-\frac{1}{\sigma}(x_r - \alpha)}$ and the distribution family for (X, Y) is invariant under the transformations

$$g_{a,b}(x_1, \ldots, x_k, y) = (ax_1 + b, \ldots, ax_k + b, y), \quad a > 0,\ b \in \mathbb{R}.$$

Suppose $\mathcal{D}_x = [0, 1]$ for all x. Then, for any loss function that depends only on d and y, the problem is invariant with $\bar{g}_{a,b}(\alpha, \sigma) = (a\alpha + b, a\sigma)$ and $\varphi_{g_{a,b}|x}(d) = d$. So, a predictor δ is equivariant if $\delta(ax + b\vec{1}) = \delta(x)$ for all $x, a > 0, b \in \mathbb{R}$, i.e., δ is invariant and hence must be a function of T (as defined before and a maximal invariant). By Theorem 2, the risk function of any $\delta(T)$ is a constant and under $\alpha = 0, \sigma = 1$, the conditional risk of $\delta(T)$ given $T = t$ is

$$\begin{aligned} E^{Y|t}[L(\delta(T), Y)] &= E^{X|t}\{E^{Y|X,t}[L(\delta(T), Y)]\} \\ &= E^{X|t}[L(\delta(t), 1)\mathrm{e}^{-X_r} + L(\delta(t), 0)(1 - \mathrm{e}^{-X_r})] \\ &= L(\delta(t), 1)\eta(t) + L(\delta(t), 0)(1 - \eta(t)), \end{aligned} \tag{14}$$

where $\eta(t) = E_{(0,1)}[\mathrm{e}^{-X_r}|t]$. Now, a best equivariant predictor can be obtained by minimizing (14) w.r.t. $\delta(t)$ for each t. For $L(d, y) = (d - y)^2$, the best equivariant predictor of Y is $\eta(T)$. For $L(d, y) = |d - y|$, (14) reduces to $(1 - \delta(t))\eta(t) + \delta(t)[1 - \eta(t)]$ and minimizing this w.r.t. $\delta(t)$ we obtain the best predictor as $\delta_1(T) = 0$ if $\eta(T) \leq 1/2$ and $\delta_1(T) = 1$ if $\eta(T) > 1/2$.

Now, suppose $\mathcal{D}_x = \{0, 1\}$ for all x, which is the sample space of Y, and suppose $L(0, 0) = L(1, 1) = 0$, $L(0, 1) = c_1$ and $L(1, 0) = c_2$. Here, the predictor that minimizes (14) is given by $\delta_2(T) = 0$ if $\eta(T) \leq c_1/(c_1 + c_2)$ and $\delta_2(T) = 1$ if $\eta(T) > c_1/(c_1 + c_2)$. Note that $\delta_2(t)$ coincides with $\delta_1(t)$ when $c_1 = c_2$. The arguments used in this example are also applicable for predicting $Y_* = I[W > \psi(X)]$ for any ψ satisfying $\psi(ax + b\vec{1}) = a\psi(x) + b$ for all $x, a > 0, b \in \mathbb{R}$.

Example 5 Here, we consider a more general version of Example 4. Suppose $X = (X_1, \ldots, X_n)$ and Y have the following location-scale family of distributions:

$$X_1, \ldots, X_n|\mu, \sigma \sim \frac{1}{\sigma^n} f\left(\frac{x_1 - \mu}{\sigma}, \ldots, \frac{x_n - \mu}{\sigma}\right), \quad x_1, \ldots, x_n,\ \mu \in \mathbb{R},\ \sigma > 0,$$
$$Y|X = x, \sigma \sim \frac{1}{\sigma} q\left(\frac{y - \psi(x)}{\sigma}\right), \quad y \in \mathbb{R},$$

where f, q and ψ are known functions.

Case 1. Suppose ψ is location-scale equivariant, i.e., $\psi(ax + b\vec{1}) = a\psi(x) + b$ for all $x, a > 0$ and $b \in \mathbb{R}$. Then, the problem of predicting Y based on X under any loss function of the form $L(\frac{d-\mu}{\sigma}, \frac{y-\mu}{\sigma}, \frac{1}{\sigma}(x - \mu\vec{1}))$ is invariant under the transformations: $g_{a,b}(x, y) = (ax + b\vec{1}, ay + b)$, $\bar{h}_{a,b}(\mu, \sigma) = (a\mu + b, a\sigma)$, $\varphi_{a,b}(d) = ad + b$, $a > 0, b \in \mathbb{R}$. Here, a predictor $\delta(X)$ is equivariant if it satisfies (11) and hence has the representation $\delta(x) = \alpha(x) + w(T(x))Q(x)$ for all x, where α is any function that satisfies (11), Q is location invariant and scale equivariant, and T is a maximal invariant (see Lehmann and Casella 1998, Sec. 3.3). Thus, one may use the approach outlined in Sect. 3 for finding a best equivariant predictor, for given f and L.

Now, suppose $q(.)$ is also symmetric around 0. Then, as in Example 4, the distribution family of (X, Y) is invariant under a larger group G_*, which contains two transformations of Y, viz., $Y \to aY + b$ and $Y \to a[2\psi(X) - Y] + b$, corresponding to each $h_{a,b}$. Then, invoking Lemma 4, it can be seen that a loss function is invariant if and only if it is a function only of $|\frac{d-y}{\sigma}|$ and $\frac{1}{\sigma}(x-\mu\vec{1})$. In such cases, a predictor δ is equivariant under G_* if it satisfies both (11) and

$$\delta(ax+b\vec{1}) = a[2\psi(x)-\delta(x)]+b \quad \text{for all } x,\ a>0,\ b\in\mathbb{R}. \tag{15}$$

Comparing the right sides of (11) and (15), we see that $\delta(X)=\psi(X)$ is the *only* (and hence best) equivariant predictor of Y.

Case 2. Suppose ψ is location invariant and scale equivariant, i.e., $\psi(ax+b\vec{1}) = a\psi(x)$ for all $x, a>0$ and $b\in\mathbb{R}$. Then, the model for (X,Y) is invariant under the transformations $X\to h_{a,b}(X)$, $Y\to aY$, $a>0, b\in\mathbb{R}$, with $\bar{h}_{a,b}$ being the same as in Case 1. Applying Lemma 5, along with its assumptions, we get $\varphi_{a,b}(d)=ad$. This has two important implications. First, $\delta(X)$ is equivariant if

$$\delta(ax+b\vec{1}) = a\delta(x) \quad \text{for all } x,\ a>0,\ b\in\mathbb{R}. \tag{16}$$

Any such δ is location invariant and must be of the form $\delta(x)=\delta_0(x)w(T(x))$, where T is a maximal invariant, and δ_0 is any given predictor satisfying (16) (see Lehmann and Casella 1998, Sec. 3.3). Actually, one may take $\psi(X)$ for δ_0. Second, L satisfies (5) if and only if

$$L(d,y,x,\mu,\sigma) = L\left(\frac{d}{\sigma},\frac{y}{\sigma},\frac{1}{\sigma}(x-\mu\vec{1}),0,1\right) \quad \text{for all } x,y,d,\mu \text{ and } \sigma.$$

If $q(.)$ is also symmetric around 0 and L is a function of $|\frac{d-y}{\sigma}|$ and $\frac{1}{\sigma}(x-\mu\vec{1})$, the problem is invariant under a larger group, which for each $h_{a,b}(X)$ contains two transformations of Y, viz., $Y\to aY$ and $Y\to a[2\psi(X)-Y]$. Corresponding transformations of d are $d\to ad$ and $d\to a[2\psi(X)-d]$. Here, $\delta(X)$ is equivariant if and only if $a\delta(x)=\delta(ax+b\vec{1})=a[2\psi(x)-\delta(x)]$ for all $a>0, b\in\mathbb{R}$ and x. Form this, it follows that $\psi(X)$ is the only (and hence best) equivariant predictor of Y.

5 Equivariance and risk unbiasedness

In decision theory, Lehmann (1951) introduced risk unbiasedness as a general concept that includes mean and median unbiasedness as special cases and proved that if among all risk unbiased decision rules there exists a unique rule with minimum risk, then it is almost equivariant, and under certain conditions, best equivariant estimators are risk unbiased. Xiao (2000), Deshpande and Fareed (1995) and Nayak and Qin (2010) have generalized the concept for predictors. For a given loss function L, let

$$L_1(d,x,\theta) = E[L(d,Y,x,\theta)|x,\theta],$$

and assume that for each x, $\min_d L_1(d,x,\theta)$ is independent of θ. Then, from Nayak and Qin (2010), we obtain:

Definition 5 A predictor $\delta(X)$ is said to be risk unbiased if for each $\theta \in \Theta$,

$$E_\theta[L_1(\delta(X), X, \theta)] \leq E_\theta[L_1(\delta(X), X, \theta')] \quad \text{for all } \theta' \neq \theta.$$

For squared error loss, risk unbiasedness reduces to mean unbiasedness if $E_\theta(Y|X)] = h(X) + \eta(\theta)$ for some h and η. This holds for location-scale models, for which Takada (1981) showed that best unbiased predictors are equivariant. Our main results are as follows.

Theorem 3 *Consider a prediction problem that is invariant under a group G and for any given predictor δ and $g = (h, k) \in G$, let $\delta_g(x) = \varphi_{g^{-1}|h(x)}(\delta(h(x)))$ for all $x \in \mathcal{X}$. Then,*

(i) $R(\delta_g(X), \theta) = R(\delta(X), \bar{h}(\theta))$ *for all* $\theta \in \Theta$.
(ii) *If* $\delta(X)$ *is risk unbiased, then so is* $\delta_g(X)$.

Proof First, note that definition of δ_g, invariance of L and (6) yield

$$L(\delta_g(x), y, x, \theta) = L(\delta(h(x)), k(x, y), h(x), \bar{h}(\theta)). \tag{17}$$

Now, part (i) can be established by noting that invariance of the model gives

$$E_\theta L(\delta(h(X)), k(X, Y), h(X), \bar{h}(\theta)) = E_{\bar{h}(\theta)} L(\delta(X), Y, X, \bar{h}(\theta)).$$

To prove part (ii), let $(U, V) = (h(X), k(X, Y))$ and note that since h is 1-1, the distribution of U depends only on the distribution of X and the conditional of V given U is determined by the conditional distribution of Y given X. So, invariance of the model imply that if $X \sim f_\theta(x)$ and $Y|x \sim f_{\theta'}(y|x)$, then $U \sim f_{\bar{h}(\theta)}(u)$ and $V|u \sim f_{\bar{h}(\theta')}(v|u)$. Now, for any $\theta, \theta' \in \Theta$,

$$\begin{aligned}
R(\delta_g, \theta) &= \int\int L(\delta(x), y, x, \bar{h}(\theta)) f(x|\bar{h}(\theta)) f(y|x, \bar{h}(\theta))\, \mathrm{d}y\, \mathrm{d}x \quad \text{by part (i)}\\
&\leq \int\int L(\delta(x), y, x, \bar{h}(\theta)) f(x|\bar{h}(\theta)) f(y|x, \bar{h}(\theta'))\, \mathrm{d}y\, \mathrm{d}x \quad \text{as } \delta \text{ is risk unbiasedness}\\
&= \int\int L(\delta(h(x)), k(x, y), h(x), \bar{h}(\theta)) f(x|\theta) f(y|x, \theta')\, \mathrm{d}y\, \mathrm{d}x \quad \text{by model invariance}\\
&= \int\int L(\delta_g(x), y, x, \theta) f(x|\theta) f(y|x, \theta')\, \mathrm{d}y\, \mathrm{d}x, \quad \text{by Eq. (17)}
\end{aligned}$$

which proves the result. □

Theorem 4 *If among all risk unbiased predictors a unique* $(w.p.1)$ *uniformly minimum risk predictor exists, then it must be almost equivariant, i.e., equivariant w.p.* 1.

Proof Suppose δ is uniquely best among all risk unbiased predictors. Then, for all $g \in G$, $\delta_g(X)$ and $\delta_{g_{-1}}(X)$ are also risk unbiased, by Theorem 3, and

$$R(\delta, \theta) \leq R(\delta_g, \theta) = R(\delta, \bar{h}(\theta)) \leq R(\delta_{g^{-1}}, \bar{h}(\theta)) = R(\delta, \theta),$$

where the two inequalities follow as δ is best risk unbiased and the two equalities hold by Theorem 3, part (i). In the above, all quantities must be equal, which implies that for all $g \in G$, δ_g and $\delta_{g^{-1}}$ are also best risk unbiased predictors. Now, by uniqueness, for each $\theta \in \Theta$ and $g \in G$, $\delta(X) = \delta_{g^{-1}}(X)$ w.p. 1 and hence $\delta(h(x)) = \varphi_{g|x}(\delta(x))$ for almost all x. □

6 Discussion

In this paper, we formalized the basic ideas of equivariance for prediction problems more generally than previous approaches. We allowed the prediction space and losses to depend on the data and imposed only some logically necessary restrictions on the transformations for changing the "coordinates" of the problem. The resulting transformation groups have novel features and properties. In particular, G_e may not be a singleton, which can also be useful for selecting a best predictor, as seen in Example 5. We showed that some well-known results in equivariant decision theory continue to hold for prediction. For example, maximal invariants depict a structure of equivariant predictors that is useful for finding a best predictor, and a best risk unbiased predictor is almost equivariant.

In this paper, we only considered point prediction. We believe that investigating equivariance for obtaining a prediction interval or a predictive distribution (see Eaton and Sudderth 2001, 2004; Lawless and Fredette 2005) is an interesting topic for future research. Under suitable conditions, best equivariant decision rules are known to be Bayes rules with respect to right invariant Haar measures. Hora and Buehler (1967), Takada (1982) and Eaton and Sudderth (2001) gave similar results for prediction. Generalizing those results in our framework is also a topic for future research. Finally, applications of our results to specific problems should be explored.

Acknowledgments We thank Professor J.K. Ghosh for a helpful and encouraging discussion and two referees for several helpful comments on the original manuscript. Nayak's research was partially supported by National Science Foundation Grant DMS-0915156.

References

Aitchison, J., Dunsmore, I. R. (1975). *Statistical prediction analysis*. New York: Cambridge University Press.

Berger, J. O. (1985). *Statistical decision theory and Bayesian analysis* (2nd ed.). New York: Springer.

Bjornstad, J. F. (1990). Predictive likelihood: A review (with discussion). *Statistical Science*, *5*, 242–265.

Bjornstad, J. F. (1996). On the generalization of the likelihood function and the likelihood principle. *Journal of the American Statistical Association*, *91*, 791–806.

Deshpande, J. V., Fareed, T. P. M. (1995). A note on conditionally unbiased estimation after selection. *Statistics & Probability Letters*, *22*, 17–23.

Eaton, M. L. (1989). *Group invariance applications in statistics*. Hayward: Institute of Mathematical Statistics.

Eaton, M. L., Sudderth, W. D. (2001). Best invariant predictive distributions. In M. Viana, D. Richards (Eds.), *Algebraic methods in statistics and probability* (pp. 49–62). Providence: American Mathematical Society.

Eaton, M. L., Sudderth, W. D. (2004). Properties of right Haar predictive inference. *Sankhya*, *66*, 487–512.

Geisser, S. (1993). *Predictive inference: An introduction*. New York: Chapman & Hall.

Hall, W. J., Wijsman, R. A., Ghosh, J. K. (1965). The relationship between sufficiency and invariance with applications in sequential analysis. *Annals of Mathematical Statistics*, *36*, 575–614.

Hora, R. B., Buehler, R. J. (1966). Fiducial theory and invariant estimation. *Annals of Mathematical Statistics*, *37*, 643–656.

Hora, R. B., Buehler, R. J. (1967). Fiducial theory and invariant prediction. *Annals of Mathematical Statistics*, *38*, 795–801.

Lawless, J. F., Fredette, M. (2005). Frequentist prediction intervals and predictive distributions. *Biometrika*, *92*, 529–542.

Lehmann, E. L. (1951). A general concept of unbiasedness. *Annals of Mathematical Statistics*, *22*, 587–592.

Lehmann, E. L., Casella, G. (1998). *Theory of point estimation* (2nd ed.). New York: Springer.

Nayak, T. K. (2000). On best unbiased prediction and its relationships to unbiased estimation. *Journal of Statistical Planning and Inference*, *84*, 171–189.

Nayak, T. K., Qin, M. (2010). The concept of risk unbiasedness in statistical prediction. *Journal of Statistical Planning and Inference*, *140*, 1923–1938.

Pearson, K. (1920). The fundamental problem of practical statistics. *Biometrika*, *13*, 1–16.

Takada, Y. (1981). Relation of the best invariant predictor and the best unbiased predictor in location and scale families. *Annals of Statistics*, *9*, 917–921.

Takada, Y. (1982). A comment on best invariant predictors. *Annals of Statistics*, *10*, 971–978.

Wijsman, R. A. (1990). *Invariant measures on groups and their use in statistics*. Hayward: Institute of Mathematical Statistics.

Xiao, Y. (2000). Linex unbiasedness in a prediction problem. *Annals of the Institute of Statistical Mathematics*, *52*, 712–721.

Yatracos, Y. G. (1992). On prediction and mean squared error. *Canadian Journal of Statistics*, *20*, 187–200.