Bernoulli **29**(2), 2023, 901–928
https://doi.org/10.3150/22-BEJ1484

Bayes-optimal prediction with frequentist coverage control

PETER HOFF[a]

Statistical Science, Duke University, Durham, USA, [a]*peter.hoff@duke.edu*

This article illustrates how indirect or prior information can be optimally used to construct a prediction region that maintains a target frequentist coverage rate. If the indirect information is accurate, the volume of the prediction region is lower on average than that of other regions with the same coverage rate. Even if the indirect information is inaccurate, the resulting region still maintains the target coverage rate. Such a prediction region can be constructed for models that have a complete sufficient statistic, which includes many widely-used parametric and nonparametric models. Particular examples include a Bayes-optimal conformal prediction procedure that maintains a constant coverage rate across distributions in a nonparametric model, as well as a prediction procedure for the normal linear regression model that can utilize a regularizing prior distribution, yet maintain a frequentist coverage rate that is constant as a function of the model parameters and explanatory variables. No results in this article rely on asymptotic approximations.

Keywords: Conformal prediction; disintegration; hypothesis testing; Neyman-Pearson lemma; tolerance region

1. Introduction

A standard statistical inference task is to construct a prediction region, that is, a set of plausible values for an unobserved random object Y having sample space $\mathcal{Y}$ based on a realization of another random object X having sample space $\mathcal{X}$. A non-randomized procedure for constructing a prediction region is a set-valued function $\mathcal{X} \to 2^{\mathcal{Y}}$, for example denoted by $x \mapsto A_x$, where $A_x \subset \mathcal{Y}$ is the set of predicted values for Y when X is observed to be x.

A prediction procedure is precise if its expected volume is small, and it is accurate if its coverage probability is high. The coverage probability of a prediction procedure $x \mapsto A_x$ is typically defined as the probability of the event $Y \in A_X$. More generally, let $\mathcal{P} = \{P_\theta : \theta \in \Theta\}$ be a collection of joint probability distributions on measurable subsets of $\mathcal{X} \times \mathcal{Y}$. The coverage probability of the set-valued function $x \mapsto A_x$ for a given θ is simply $P_\theta(A)$, where $A = \{(x,y) : y \in A_x\}$ is the graph of $x \mapsto A_x$. A prediction procedure A for which $P_\theta(A) \geq 1-\alpha$ for some target value of α and for all θ is generally referred to as having $1-\alpha$ frequentist coverage [1]. In what follows, if $P_\theta(A) = 1-\alpha$ for all θ then A will be referred to as a $1-\alpha$ constant coverage prediction region. Patel [20] reviews a variety of methods for constructing prediction regions that have frequentist coverage control, the most widely-used among them perhaps being those that are derived from pivotal quantities [17].

The precision of a prediction region can be quantified with its expected volume. Let μ be a volume measure on measurable subsets of $\mathcal{Y}$. We define the risk function of A to be the expected volume of A_X as a function of θ:

$$R_\theta(A) = \mathrm{E}_\theta[\mu(A_X)]. \tag{1}$$

As with confidence regions, precision can be increased at the expense of coverage, and so we compare among, or optimize over, regions that share a common coverage. A prediction region A is better than A' if $R_\theta(A) \leq R_\theta(A')$ and $P_\theta(A) \geq P_\theta(A')$ for all θ, with inequality for some θ. However, as with other types of statistical decision problems, typically there is not a uniformly best prediction region

1350-7265 © 2023 ISI/BS

procedure. This motivates the identification of an optimal member of a reduced class of procedures, or identification of a procedure that performs well in a particular region of the parameter space. As an example of the former approach, Evans and Fraser [8] show that for the normal linear model, the standard prediction region obtained by pivoting is, for a given coverage rate, optimal among all regions that are equivariant with respect to affine transformations. Such a prediction region performs equally well across the parameter space, in the sense that the expected volume of the prediction region (appropriately scaled by the variance) does not depend on the model parameters.

In this article, prediction regions that share a common frequentist coverage are compared to each other in terms of their average risk across the parameter space. The motivation for this criterion is that in many applications there is indirect information or prior knowledge that some distributions in $\mathcal{P}$ are more plausible than others. In such cases, it may be preferable to have a prediction region that performs well for values of θ that are most plausible, at the expense of worse performance for less plausible values. This suggests evaluating a prediction region A with a Bayes risk,

$$R(A) = \int R_\theta(A)\,\pi(d\theta), \tag{2}$$

where π is a probability distribution that gives a large weight to θ-values for which $R_\theta(A)$ is desired to be small.

A prediction region with a constant frequentist coverage rate and a small Bayes risk could be desirable in many different data analysis scenarios. First, if real prior information about θ is available, but a region with $1-\alpha$ frequentist coverage is required, then a region that minimizes (2) while maintaining $1-\alpha$ coverage is optimal from a subjective Bayesian perspective. Another scenario where specific, indirect information is available arises in the analysis of data from multiple populations, also known as multilevel data analysis or small area estimation. Many methods for multipopulation inference operate essentially by using a Bayesian procedure for each individual population based on a "prior" distribution derived from all of the populations. More details on how the procedures described in this article may be used for multipopulation inference appear in Section 5. Finally, for models where θ is a vector of parameters, the distribution π could represent beliefs that the true parameters are sparse or near zero, in the same way that commonly-used regularization penalties represent such beliefs. For example, θ could be a vector of linear regression coefficients, many of which are suspected to be close or equal to zero. In this case, we might prefer a prediction region that has particularly low expected volume when θ is sparse or close to zero, over, for example, an equivariant region that has the same expected volume for all θ-values. Calculations for and numerical examples of prediction regions for regularized regression appear in Section 4.3. In summary, while we will refer to π as a prior distribution and $R(A)$ as a Bayes risk, π may or may not be a formal prior distribution that describes subjective beliefs about the value of θ. Most generally, π can simply be viewed as a weighting function that prioritizes regions of the parameter space.

The main result of this article is that, for many commonly-used statistical models, it is possible to find a prediction region that maintains a target frequentist coverage rate and is also Bayes-optimal for its coverage. Specifically, we are able to construct a set $A^\pi \subset \mathcal{X} \times \mathcal{Y}$ such that $P_\theta(A^\pi) \geq 1-\alpha$ for all θ, and $R(A^\pi) \leq R(A)$ for all sets A having the same coverage as A^π. In particular, if A^π has a constant coverage rate of $1-\alpha$, then A^π is Bayes-optimal among all $1-\alpha$ constant coverage prediction regions. Such a prediction region A^π is Bayes-optimal among procedures that share its frequentist coverage rate. We follow Yu and Hoff [24] by referring to such a procedure as "frequentist and Bayesian", or FAB. Statistical procedures of this type go back at least to Pratt [21], who constructed a constant coverage confidence interval for the mean of a normal population that has minimum prior expected width among intervals with the same frequentist coverage. Related to this is the "Bayes-non-Bayes compromise" of Good [10], whereby test statistics are obtained from Bayes factors, but are compared to frequentist null distributions, typically via permutation. Yu and Hoff [24] and Hoff and Yu [13] extended Pratt's Bayes-optimal frequentist interval to multiparameter settings, in which the prior distribution may be empirically estimated from the data using a hierarchical model, resulting in an adaptive confidence interval procedure with guaranteed frequentist coverage, even if the hierarchical model is wrong. The confidence interval procedures developed by Yu and Hoff [24] and Hoff and Yu [13] applied only to means of univariate normal populations. This article shows how to construct analogous procedures for prediction, and in a much wider class of models - those with a complete sufficient statistic. This includes many exponential family models such as multinomial, linear regression, multivariate normal and generalized linear models, some models with varying support, nonparametric models [2], and others [18].

The results in this article are developed as follows: In order to find a risk-optimal prediction procedure, we must first characterize the set of all prediction procedures. This is done in Section 2, where we extend the observation of Faulkenberry [9] that when a sufficient statistic is available, a prediction region with a desired coverage rate may be constructed by inverting the acceptance regions of a collection of conditional point-null hypothesis tests. Faulkenberry also observed that prediction regions with constant coverage must have constant conditional coverage, given a complete sufficient statistic. Our Theorem 2.3 states that *all* $1 - \alpha$ constant coverage prediction regions may be constructed from level-α conditional tests, and Theorem 2.4 states that all prediction regions with common coverage as a function of the parameter must also have common conditional coverage as a function of the complete sufficient statistic. These results are used in Section 3, where a Bayes-optimal prediction region for a given coverage is derived from a collection of Bayes-optimal conditional tests. This is done by first obtaining a joint disintegration of the probability measures $\{P_\theta : \theta \in \Theta\}$ that define coverage, and the non-probability measure R that defines the Bayes risk. This joint disintegration allows us to find the risk-minimizing conditional test for each possible value of the sufficient statistic, using a variation of the Neyman-Pearson lemma. Theorems 3.2 and 3.4 then give the form of the risk-optimal prediction region for a given coverage function and a given constant coverage rate, respectively. Corollary 3 provides a more familiar expression for the optimal region in terms of a test statistic based on densities of $\{P_\theta : \theta \in \Theta\}$ and a prior predictive density. None of these results rely on asymptotic approximations.

While the methodology in this article applies generally to models with a complete sufficient statistic, Section 4 considers a few specific scenarios in detail, including prediction for multivariate normal and normal linear regression models, as well as nonparametric prediction using conformity scores as described in Jin et al. [14]. In particular, it is shown that the Bayes-optimal choice of a conformity score is, not surprisingly, the Bayesian posterior predictive density. The resulting FAB prediction region is Bayes optimal among nonparametric regions with constant frequentist coverage, and is of course different from a fully Bayesian posterior predictive region, as the latter does not have constant frequentist coverage. In this sense, the conformal FAB approach optimally incorporates prior information while maintaining a constant frequentist coverage rate. Additional aspects of the FAB prediction methodology, and some directions for further research are discussed in Section 5. Proofs are in an appendix.

2. Prediction regions via sufficiency

2.1. Review of Faulkenberry's construction

Let X and Y be random objects taking values in spaces $\mathcal{X}$ and $\mathcal{Y}$ respectively, with a joint distribution that is a member of a model $\mathcal{P} = \{P_\theta : \theta \in \Theta\}$ on $\mathcal{X} \times \mathcal{Y}$. For a model $\mathcal{P}$ with a sufficient statistic $Z : \mathcal{X} \times \mathcal{Y} \to \mathcal{Z}$, Faulkenberry [9] proposed constructing a $1 - \alpha$ prediction region A for Y from X as follows:

1. For each $z \in \mathcal{Z}$,
   a) identify P_z^Y, the conditional distribution of Y given $\{Z = z\}$, which by sufficiency does not depend on θ;
   b) find a set $C_z \subset \mathcal{Y}$ such that $P_z^Y(C_z) = 1 - \alpha$.
2. Let $A = \{(x,y) : y \in C_{Z(x,y)}\}$, so $y \in A_x \Leftrightarrow y \in C_{Z(x,y)}$.

The conditional probability of the event $Y \in A_X$ given $\{Z = z\}$ is $P_z^Y(C_z) = 1 - \alpha$ for all z, and so its unconditional probability is equal to $1 - \alpha$ as well.

Example 1. Let $X \sim N(\theta, k\sigma^2)$ and $Y \sim N(\theta, \sigma^2)$ be independent for some unknown value of $\theta \in \mathbb{R}$ and known value of $\sigma^2 > 0$. For example, if X were the sample mean of a sample of size n from the $N(\theta, \sigma^2)$ distribution, then $k = 1/n$. For arbitrary k, $Z = (X + kY)/(1 + k)$ is a complete sufficient statistic, with $Y|\{Z = z\} \sim N(z, \sigma^2/(k + 1))$. For any function $\delta : \mathbb{R} \to \mathbb{R}$, define for each $z \in \mathbb{R}$ the set

$$C_z = \left\{ y : \left| \frac{z - y}{\sigma/\sqrt{k+1}} + \delta(z) \right|^2 < \chi^2_{1,\delta(z)^2,1-\alpha} \right\}, \tag{3}$$

where $\chi^2_{1,\delta^2,1-\alpha}$ is the $1 - \alpha$ quantile of the χ^2_1 distribution with noncentrality parameter δ^2. Then $P_z^Y(C_z) = 1 - \alpha$ for each z. Defining A as in step 2 gives

$$A = \left\{ (x,y) : \left| \frac{x - y}{\sigma\sqrt{k+1}} + \delta(Z(x,y)) \right|^2 < \chi^2_{1,\delta(Z(x,y))^2,1-\alpha} \right\}, \tag{4}$$

which is a $1 - \alpha$ constant coverage prediction region for any choice of δ. For example, if δ is identically zero then A_x is the standard prediction interval $x \pm \sigma\sqrt{k+1} \times \Phi^{-1}(1 - \alpha/2)$ obtained via the pivotal quantity $X - Y$, where Φ^{-1} is the standard normal quantile function. In Section 4 we will obtain the function δ that minimizes the Bayes risk under a normal prior distribution for θ, and generalize the method to multivariate normal models with unknown variance.

Faulkenberry specifically considered the case that X and Y are independent and that Z is a complete sufficient statistic, but only sufficiency - not independence or completeness - is necessary to ensure that a set obtained from Faulkenberry's construction has $1 - \alpha$ constant coverage. Faulkenberry considers completeness because it provides a characterization of the set of prediction regions that have $1 - \alpha$ constant coverage, that is, regions for which $P_\theta(A) = 1 - \alpha$ for all θ. For such a region, Faulkenberry points out that $P_\theta(A|Z) = 1 - \alpha$ almost surely for each θ, because $\mathrm{E}_\theta[P_\theta(A|Z)] = P_\theta(A) = 1 - \alpha$ for all θ and Z is complete. Faulkenberry therefore concludes that, for models with a complete sufficient statistic, a prediction region has constant coverage as a function of the parameter if and only if it has constant conditional coverage (almost surely) as a function of the complete sufficient statistic.

Can all prediction procedures for models with a complete sufficient statistic be derived from Faulkenberry's construction? The answer is no, for two reasons: The first reason is that not all subsets of $\mathcal{X} \times \mathcal{Y}$ can be expressed as $A = \{(x,y) : y \in C_{Z(x,y)}\}$ for some set-valued function C mapping $\mathcal{Z}$ to subsets of $\mathcal{Y}$, unless some additional conditions on Z are met. The second reason is that, as pointed out by Dunsmore [7], there may not exist prediction procedures A with constant coverage, for example, if the P_θ's are discrete distributions. In this case, coverage above or equal to $1 - \alpha$ may still be maintained by choosing a C_z in Faulkenberry's construction so that $P_z^Y(C_z) \geq 1 - \alpha$ for all z. However, while every set with conditional coverage of at least $1 - \alpha$ also has marginal coverage of at least $1 - \alpha$, the converse is not necessarily true, even with completeness of Z: For prediction of a binomial random variable, Dunsmore provides an example of a prediction region A for which $P_\theta(A) \geq 1 - \alpha$ for all θ but $P(A|Z = z)$

falls below $1-\alpha$ for some values of z. This means that in some cases the set of prediction regions having coverage greater than $1-\alpha$ for all θ is a proper superset of those with conditional coverage greater than $1-\alpha$ for all z, and so the best procedure that can be obtained from Faulkenberry's construction may not be the best procedure obtainable.

In the remainder of this section we extend Faulkenberry's method to address these limitations. In the next subsection we generalize Faulkenberry's construction to provide a complete characterization of set-valued functions from $\mathcal{X}$ to subsets of $\mathcal{Y}$ in terms of set-valued functions from $\mathcal{Z}$ to subsets of $\mathcal{X}\times\mathcal{Y}$. In Subsection 2.3 we show how complete sufficiency allows for a characterization of all procedures that have the same coverage, in terms of collections of point-null hypothesis tests that have the same conditional size. In particular, given a candidate prediction procedure, we can characterize the class of procedures with the same coverage, and possibly find one with lower risk.

2.2. Characterizing set-valued functions

Let $\mathcal{X}$ and $\mathcal{Y}$ be spaces. For a subset A of $\mathcal{X}\times\mathcal{Y}$ and element $x\in\mathcal{X}$, the *section* of A at x is the set $A_x=\{y:(x,y)\in A\}$, a subset of $\mathcal{Y}$. The sections of A define a set-valued function $\mathcal{X}\to 2^{\mathcal{Y}}$ given by $x\mapsto A_x$. Conversely, for each $x\in\mathcal{X}$, let A_x be a subset of $\mathcal{Y}$. Every such set-valued function has a *graph* $A=\{(x,y):y\in A_x\}$, a subset of $\mathcal{X}\times\mathcal{Y}$. The operations of calculating the graph of a set-valued function $\mathcal{X}\to 2^{\mathcal{Y}}$, and calculating the sections of a subset of $\mathcal{X}\times\mathcal{Y}$, are inverses of each other, and so there is a bijection between set-valued functions from $\mathcal{X}\to 2^{\mathcal{Y}}$ and subsets of $\mathcal{X}\times\mathcal{Y}$. As such, in what follows we will use the same symbol (e.g. "A") for a set-valued function and its graph, and whether or not the symbol represents a function $\mathcal{X}\to 2^{\mathcal{Y}}$ or a subset of $\mathcal{X}\times\mathcal{Y}$ will be clear from context.

Another representation of a subset of $\mathcal{X}\times\mathcal{Y}$ is given by any surjective mapping $Z:\mathcal{X}\times\mathcal{Y}\to\mathcal{Z}$. Overloading notation somewhat, for each $z\in\mathcal{Z}$ let $A_z=A\cap Z^{-1}\{z\}$. Then the mapping $z\mapsto A_z$ is a set-valued function from $\mathcal{Z}$ to $2^{\mathcal{X}\times\mathcal{Y}}$. Conversely, if A_z is a subset of the fiber $Z^{-1}\{z\}$ for each $z\in\mathcal{Z}$, then $\cup_{z\in\mathcal{Z}}A_z$ is a subset of $\mathcal{X}\times\mathcal{Y}$. We summarize this observations and those of the preceding paragraph as follows:

Lemma 2.1. *Let $Z:\mathcal{X}\times\mathcal{Y}\to\mathcal{Z}$ be a surjection. Then there is a bijection between each pair of the following sets:*

1. $2^{\mathcal{X}\times\mathcal{Y}}$*;*
2. *set-valued functions $\mathcal{X}\to 2^{\mathcal{Y}}$;*
3. *set-valued functions $A:\mathcal{Z}\to\mathcal{X}\times\mathcal{Y}$ for which $A_z\subset Z^{-1}\{z\}$.*

In the next subsection, we will show how the bijection between items 2 and 3 can be used to represent a prediction procedure (item 2) as a collection of acceptance regions of hypothesis tests (item 3).

Now recall that Faulkenberry's method is to construct a prediction procedure $A:\mathcal{X}\to 2^{\mathcal{Y}}$ from a set-valued function $C:\mathcal{Z}\to 2^{\mathcal{Y}}$ by setting $A=\{(x,y):y\in C_{Z(x,y)}\}$. Can all prediction procedures be represented in this way? Since Lemma 2.1 shows there is a bijection between functions $\mathcal{X}\to 2^{\mathcal{Y}}$ and functions $\mathcal{Z}\to 2^{\mathcal{X}\times\mathcal{Y}}$, intuitively Faulkenberry's representation can only be complete if somehow y and $Z(x,y)$ determine x, for example if $Z(x,y)=x+y$. In fact, such a condition is necessary and sufficient for Faulkenberry's representation to be a complete characterization of the set-valued functions from $\mathcal{X}$ to $2^{\mathcal{Y}}$, or equivalently, subsets of $\mathcal{X}\times\mathcal{Y}$:

Lemma 2.2. *Let $Z:\mathcal{X}\times\mathcal{Y}\to\mathcal{Z}$. A set $A\subset\mathcal{X}\times\mathcal{Y}$ can be written as $A=\{(x,y):y\in C_{Z(x,y)}\}$ for some $C:\mathcal{Z}\to 2^{\mathcal{Y}}$ if and only if the function $\mathcal{X}\to\mathcal{Z}$ given by $Z_y(x)=Z(x,y)$ is injective for each y.*

Similarly, if x and $Z(x,y)$ determine y, then the subsets of $\mathcal{X}\times\mathcal{Y}$ can be characterized in terms of set-valued functions from $\mathcal{Z}$ to $2^{\mathcal{X}}$:

Corollary 1. *A set $A\subset\mathcal{X}\times\mathcal{Y}$ can be written as $A=\{(x,y): x\in B_{Z(x,y)}\}$ for some $B:\mathcal{Z}\to 2^{\mathcal{X}}$ if and only if the function $\mathcal{Y}\to\mathcal{Z}$ given by $Z_x(y)=Z(x,y)$ is injective for each x.*

Example 1 (Continued). For the statistic $Z(x,y)=(x+ky)/(1+k)$, both Z_x and Z_y are injective for each x and y. Let A be a subset of $\mathcal{X}\times\mathcal{Y}$, and let $C_z=\{y:(z\times(1+k)-ky,y)\in A\}$ and $B_z=\{x:(x,(z\times(1+k)-x)/k)\in A\}$. Then $A=\{(x,y):y\in C_{Z(x,y)}\}=\{(x,y):x\in B_{Z(x,y)}\}$.

Example 2. Let $X\sim N(\theta,1)$ and $Y|\{X=x\}\sim N(\theta+x,1)$ with unknown $\theta\in\mathbb{R}$. This is a simplified first-order autoregressive model, for which $Z(x,y)=y$ is a complete sufficient statistic. In this case, Z_y is not injective, but Z_x is, since $Z_x(y)=y$. Faulkenberry's construction does not characterize the prediction regions in this case, but a modification does: For any $A\subset\mathcal{X}\times\mathcal{Y}$, let $B:\mathcal{Y}\to 2^{\mathcal{X}}$ be defined as $B_y=\{x:(x,y)\in A\}$, the section of the reflection of A at y. Then clearly $A=\{(x,y):x\in B_y\}$. A $1-\alpha$ prediction region A for Y can be constructed by choosing B_y so that $P^X_y(B_y)=1-\alpha$ for each y and then letting $A=\{(x,y):x\in B_y\}$.

Example 3. Let $X\sim N_p(0,\sigma^2 I)$, and let $Y|\{X=x\}\sim N_p(\theta x,I)$, where $\theta\in\mathbb{R}$ and $\sigma^2\in\mathbb{R}^+$ are both unknown. This model can be seen as a simplified linear regression model with a random explanatory variable. For this model, $Z(x,y)=(\|x\|,x^\top y)$ is a complete sufficient statistic, but Z_y is not injective for any y, nor is Z_x injective for any x. A subset A of $\mathcal{X}\times\mathcal{Y}$ cannot in general be represented as $\{(x,y): x\in B_{Z(x,y)}\}$ or $\{(x,y):y\in C_{Z(x,y)}\}$ for set-valued functions $B:\mathcal{Z}\to 2^{\mathcal{X}}$ or $C:\mathcal{Z}\to 2^{\mathcal{Y}}$. However, any $A\subset\mathcal{X}\times\mathcal{Y}$ can be expressed as $A=\{(x,y):(x,y)\in A_{Z(x,y)}\}$ simply by setting $A_z=A\cap Z^{-1}\{z\}$.

To summarize, a prediction region is a set-valued function $\mathcal{X}\to 2^{\mathcal{Y}}$, or equivalently, a subset of $\mathcal{X}\times\mathcal{Y}$. For any surjective function $Z:\mathcal{X}\times\mathcal{Y}\to\mathcal{Z}$, there is a bijection between such set-valued functions and set-valued functions $A:\mathcal{Z}\to 2^{\mathcal{X}\times\mathcal{Y}}$ for which $A_z\subset Z^{-1}\{z\}$ for all z in $\mathcal{Z}$. Furthermore, if the function $y\mapsto Z(x,y)$ is injective for each x, then there is a bijection between prediction regions and functions that map $\mathcal{Z}$ to $2^{\mathcal{X}}$. Similarly, if the function $x\mapsto Z(x,y)$ is injective for each y, then there is a bijection between prediction regions and functions that map $\mathcal{Z}$ to $2^{\mathcal{Y}}$.

2.3. Characterizing regions as hypothesis tests

Let $(\mathcal{X},\mathcal{F})$ and $(\mathcal{Y},\mathcal{G})$ be measurable spaces, and let $(\mathcal{X}\times\mathcal{Y},\mathcal{A})$ be their product space, so that $\mathcal{A}=\mathcal{F}\otimes\mathcal{G}$ is the smallest σ-algebra containing the rectangles $\{F\times G: F\in\mathcal{F}, G\in\mathcal{G}\}$. As described previously, a prediction procedure for a $\mathcal{Y}$-valued random object Y from an $\mathcal{X}$-valued random object X is a set-valued function $A:\mathcal{X}\to 2^{\mathcal{Y}}$, or alternatively, a subset A of $\mathcal{X}\times\mathcal{Y}$. The coverage of A under any probability measure P on $(\mathcal{X}\times\mathcal{Y},\mathcal{A})$ is simply $P(A)$, which in order to be well-defined, requires A to be $\mathcal{A}$-measurable. Conversely, if $A\in\mathcal{A}$ then $A_x\equiv\{y:(x,y)\in A\}$ is a measurable subset of $\mathcal{Y}$ for each $x\in\mathcal{X}$, that is, $A:\mathcal{X}\to\mathcal{G}$ [11, Theorem 34.A]. However, not all set-valued functions $\mathcal{X}\to\mathcal{G}$ have $\mathcal{A}$-measurable graphs. [22, Chapter 14].

Let $\mathcal{P}=\{P_\theta:\theta\in\Theta\}$ be a statistical model for $(\mathcal{X}\times\mathcal{Y},\mathcal{A})$ that has a boundedly complete regular sufficient statistic $Z:(\mathcal{X}\times\mathcal{Y},\mathcal{A})\to(\mathcal{Z},\mathcal{H})$. Regular sufficiency means that there is a function $\mathcal{Z}\times\mathcal{A}\to[0,1]$, denoted $(z,A)\mapsto P_z(A)$, for which

1. P_z is a probability measure on $(\mathcal{X}\times\mathcal{Y},\mathcal{A})$ for all $z\in\mathcal{Z}$;
2. $P_Z(A)$ is a version of $P_\theta(A|Z)$ for all $\theta\in\Theta$ and $A\in\mathcal{A}$;

3. $P_z(Z^{-1}\{z\}) = 1$ for $\mathcal{P}$-almost all z.

Recall that Z is boundedly complete if for all bounded measurable real-valued functions f, $\mathrm{E}_\theta[f(Z)] = c$ for all $\theta \in \Theta$ implies $f(z) = c$ for $\mathcal{P}$-almost all z. As noted by Faulkenberry [9], completeness of Z can be used to relate coverage to conditional coverage. In particular, suppose A is a $1-\alpha$ constant coverage prediction region, so that $P_\theta(A) = 1-\alpha$ for all θ. Since $P_\theta(A) = \mathrm{E}_\theta[P_Z(A)]$, completeness of Z implies $P_z(A) = 1-\alpha$ for almost all z. Combining this with Lemma 2.1 gives the following characterization of all $1-\alpha$ constant coverage prediction regions:

Theorem 2.3. *Let Z be a boundedly complete sufficient statistic for the model $\mathcal{P} = \{P_\theta : \theta \in \Theta\}$ on $(\mathcal{X}\times\mathcal{Y},\mathcal{A})$. A prediction region $A \in \mathcal{A}$ has constant coverage $P_\theta(A) = 1-\alpha$ for all $\theta \in \Theta$ if and only if $A = \{(x,y) : (x,y) \in A_{Z(x,y)}\}$ for a set-valued function $A : \mathcal{Z} \to 2^{\mathcal{X}\times\mathcal{Y}}$ for which*

1. *$\cup_{z\in\mathcal{Z}} A_z$ is measurable;*
2. *A_z is a subset of $Z^{-1}\{z\}$;*
3. *$P_z(A_z) = 1-\alpha$ for $\mathcal{P}$-almost all z.*

Note that each A_z can be thought of as the acceptance region of a non-randomized size-α test of $H_z : (X,Y) \sim P_z$. If such a test exists for each z, then *all* $1-\alpha$ constant coverage prediction regions may be expressed as inversions of such tests. If the distributions $\{P_z : z \in \mathcal{Z}\}$ are non-atomic, then these size-α tests will exist for any choice of α, and so constant coverage regions will exist for any choice of α. Even for some cases where the P_z's are discrete, there will exist constant coverage prediction regions for certain choices of α, as illustrated by the following example:

Example 4. Suppose $Y_1,\dots,Y_{n+1}$ is an independent and identically distributed (i.i.d.) random sample and we wish to predict Y_{n+1} from a realization of $X = (Y_1,\dots,Y_n)$. Let the model $\mathcal{P}$ on $\mathcal{Y}^{n+1}$ be such that the probability of ties among the Y_i's is zero and that the set of unordered values of $Y_1,\dots,Y_{n+1}$ constitute a boundedly complete sufficient statistic. This includes several nonparametric families [2] as well as many parametric families [18]. Let $z = \{y_1,\dots,y_{n+1}\} \subset \mathcal{Y}$ be a possible value of the sufficient statistic. Then the conditional distribution of (X,Y_{n+1}) given $Z = z$ has mass $1/(n+1)!$ on each permutation of the vector $(y_1,\dots,y_{n+1})$. It is therefore possible to find a set A_z for $\mathcal{P}$-almost all z such that $P_z(A_z) = 1-\alpha$ for any $\alpha = k/(n+1)!$ with integer k between zero and $(n+1)!$. However, symmetry suggests that if $(y_1,\dots,y_n,y_{n+1}) \in A_z$ then $(y_{p_1},\dots,y_{p_n},y_{n+1})$ should be in A_z as well, for any permutation $p_1,\dots,p_n$ of $1,\dots,n$. Prediction regions with this kind of symmetry can only have constant $1-\alpha$ coverage for values $\alpha = k/(n+1)$ with $k \in \{0,1,\dots,n+1\}$.

In some cases for which the P_z's are discrete there will not exist constant coverage prediction regions. However, completeness can still be used to characterize prediction regions that have the same coverage. Specifically, let A and A' be two prediction regions for which $P_\theta(A) = P_\theta(A')$ for all $\theta \in \Theta$. Then

$$0 = P_\theta(A) - P_\theta(A') \tag{5}$$

$$= \mathrm{E}_\theta[P_Z(A) - P_Z(A')] \tag{6}$$

for all θ, which implies that $P_z(A) = P_z(A')$ for $\mathcal{P}$-almost all z by the bounded completeness of Z. Combining this observation with Lemma 2.1 gives the following characterization of prediction regions with the same coverage:

Theorem 2.4. *Let Z be a boundedly complete sufficient statistic for the model $\mathcal{P} = \{P_\theta : \theta \in \Theta\}$ on $(\mathcal{X}\times\mathcal{Y},\mathcal{A})$. Sets A and A' in $\mathcal{A}$ satisfy $P_\theta(A) = P_\theta(A')$ for all θ if and only if $A = \{(x,y) : (x,y) \in$*

$A_{Z(x,y)}\}$ and $A' = \{(x,y) : (x,y) \in A'_{Z(x,y)}\}$ for set-valued functions A and A' mapping $\mathcal{Z} \to 2^{\mathcal{X}\times\mathcal{Y}}$ for which

1. *$\cup_{z\in\mathcal{Z}} A_z$ and $\cup_{z\in\mathcal{Z}} A'_z$ are measurable;*
2. *A_z and A'_z are subsets of $Z^{-1}\{z\}$;*
3. *$P_z(A_z) = P_z(A'_z)$ for $\mathcal{P}$-almost all z.*

This result allows us to characterize all prediction regions that have coverage equal to that of a given prediction region. In the next section, this result will be used to construct regions that are Bayes-optimal for their coverage.

3. Bayes-optimal prediction regions

3.1. Existence and uniqueness of optimal regions

As described in the Introduction, we define the risk of a prediction procedure $A \in \mathcal{A}$ under P_θ as its expected $\mathcal{Y}$-volume, so that $R_\theta(A) = \mathrm{E}_\theta[\mu(A_X)]$ where μ is a σ-finite measure on $(\mathcal{Y},\mathcal{G})$. Letting P_θ^X be the marginal distribution of X under P_θ, for any $A \in \mathcal{A}$ we have

$$\begin{aligned} R_\theta(A) &= \int \mu(A_x)\, P_\theta^X(dx) && (7)\\ &= \int\int 1(y \in A_x)\, \mu(dy)\, P_\theta^X(dx) && (8)\\ &= \int\int 1((x,y) \in A)\, P_\theta^X(dx)\, \mu(dy) && (9) \end{aligned}$$

by Tonelli's theorem, and so R_θ is simply the product measure $P_\theta^X \times \mu$ on $(\mathcal{X}\times\mathcal{Y},\mathcal{A})$. Now let π be a prior probability measure on the measurable space $(\Theta,\mathcal{T})$ for which $F \mapsto P_\theta^X(F)$ is a $\mathcal{T}$-measurable function of θ for each $F \in \mathcal{F}$. Then P_π^X, defined by $P_\pi^X(F) = \int P_\theta^X(F)\,\pi(d\theta)$, is a probability measure on $(\mathcal{X},\mathcal{F})$. The Bayes risk $R(A)$ of $A \in \mathcal{A}$ under π is then

$$\begin{aligned} R(A) &= \int R_\theta(A)\,\pi(d\theta) && (10)\\ &= \int\int 1((x,y) \in A)\, P_\pi^X(dx)\, \mu(dy), && (11) \end{aligned}$$

and so R is the product measure $P_\pi^X \times \mu$ on $(\mathcal{X}\times\mathcal{Y},\mathcal{A})$. Note that P_π^X is finite and μ is σ-finite, so R is σ-finite.

Recall that by the regular sufficiency of Z, the coverage of a prediction region A can be written as

$$P_\theta(A) = \int P_z(A_z)\,\nu_\theta(dz), \qquad (12)$$

where ν_θ is the marginal probability measure of Z under P_θ, defined as $\nu_\theta(H) = P_\theta(Z^{-1}H)$ for $H \in \mathcal{H}$. In other words, the coverage probability of a region A can be represented as an average of conditional probabilities of the sets $\{A_z = A \cap Z^{-1}\{z\} : z \in \mathcal{Z}\}$. We now show how the Bayes risk can be similarly represented. Let ν_R be the image measure of R under Z, so that $\nu_R(H) = R(Z^{-1}H)$ for $H \in \mathcal{H}$. A collection of σ-finite measures $\{R_z : z \in \mathcal{Z}\}$ on $(\mathcal{X}\times\mathcal{Y},\mathcal{A})$ is a (Z,ν_R)-*disintegration* of R if

1. $z \mapsto R_z(A)$ is measurable for each $A \in \mathcal{A}$;
2. $R_z(Z^{-1}\{z\}) = R_z(\mathcal{X} \times \mathcal{Y})$ for ν_R-almost all z;
3. $R(A) = \int R_z(A)\, \nu_R(dz)$ for each $A \in \mathcal{A}$.

Additionally, R_z is a probability measure for ν_R-almost all z if ν_R is σ-finite [6, Theorem 2]. As the third item in the list indicates, a disintegration is a generalization of a regular conditional probability distribution to σ-finite measures that are not necessarily probability measures, such as R for some choices of μ. Just as a conditional probability distribution of P_θ allows us to write $P_\theta(A)$ as an average over conditional probabilities $\{P_z(A) : z \in \mathcal{Z}\}$, a disintegration of R allows us to write $R(A)$ as an average over $\{R_z(A) : z \in \mathcal{Z}\}$, where each R_z is itself a measure with support on $Z^{-1}\{z\} \subset \mathcal{X} \times \mathcal{Y}$: If a (Z, ν_R)-disintegration of R exists, we can write

$$R(A) = \int R_z(A)\, \nu_R(dz) = \int R_z(A_z)\, \nu_R(dz) \tag{13}$$

where the last equality holds because R_z has mass only on $Z^{-1}\{z\}$ for almost all z.

The disintegration of the risk R given by (13) together with the “disintegration” of the coverage given by (12) provide a means of constructing a $1-\alpha$ constant coverage prediction region that is optimal in terms of the Bayes risk R, assuming the set of regions with this coverage is not empty. We first describe the idea informally: By completeness of Z, a region A for which $P_\theta(A) = 1-\alpha$ for all θ must also satisfy $P_z(A_z) = 1-\alpha$ for $\mathcal{P}$-almost all z. To construct a set A^π that is optimal among such regions, simply let A^π_z be the subset of $Z^{-1}\{z\}$ that has minimal R_z-measure among sets with P_z-measure equal to $1-\alpha$. By (13), the region $A^\pi = \cup_z A^\pi_z$ will then have minimal R-measure among regions with $1-\alpha$ constant coverage. This line of reasoning can also be used to construct regions that lack constant coverage (as will typically be the case for models having discrete distributions), but are nevertheless Bayes-optimal in the sense that they minimize the Bayes risk among regions having the same coverage as each other, as a function of θ.

Generally speaking, a region A^π will be Bayes-optimal for its own coverage function if for each z, $R_z(A_z) \geq R_z(A^\pi_z)$ for any A_z such that $P_z(A_z) = P_z(A^\pi_z)$. In other words, a subset A^π_z of $Z^{-1}\{z\}$ is optimal if no sets having greater or equal P_z-coverage have smaller R_z-risk. The form of such a minimizing set is given by the following variant of the Neyman-Pearson lemma:

Lemma 3.1. *Let P and R be probability measures on $(\Omega, \mathcal{A})$, and let $\tilde{A}$ have the form*

$$\tilde{A} = \{\omega : p(\omega) > kr(\omega)\} \tag{14}$$

where $k > 0$ and p and r are densities of P and R with respect to a common dominating measure. Then $R(A) \geq R(\tilde{A})$ for all $A \in \mathcal{A}$ such that $P(A) \geq P(\tilde{A})$, with equality only if $P(A \Delta \tilde{A}) = 0$.

Applying this lemma to P_z and R_z for each z gives an R_z-optimal set A^π_z for each z. Combining these gives an R-optimal set $A^\pi = \cup A^\pi_z$:

Theorem 3.2. *Let $\mathcal{P} = \{P_\theta : \theta \in \Theta\}$ be a model on $\{\mathcal{X} \times \mathcal{Y}, \mathcal{A}\}$ with a boundedly complete regular sufficient statistic $Z : (\mathcal{X} \times \mathcal{Y}, \mathcal{A}) \to (\mathcal{Z}, \mathcal{H})$ having marginal distributions $\{\nu_\theta : \theta \in \Theta\}$ on $(\mathcal{Z}, \mathcal{H})$. Let R be a σ-finite measure on $(\mathcal{X} \times \mathcal{Y}, \mathcal{A})$ with a σ-finite image measure ν_R on $(\mathcal{Z}, \mathcal{H})$ and a (Z, ν_R)-disintegration $\{R_z : z \in \mathcal{Z}\}$. Assume that*

1. *$\nu_\theta(H) = 0$ for all θ implies $\nu_R(H) = 0$;*
2. *$\nu_\theta \ll \nu_R$ for all θ.*

Let $A^{\pi} \in \mathcal{A}$ have the form

$$A^{\pi} \cap Z^{-1}\{z\} \equiv A_z^{\pi} = \{(x,y) \in Z^{-1}\{z\} : p_z(x,y) > k_z r_z(x,y)\} \tag{15}$$

where $k_z > 0$ and p_z and r_z are densities of P_z and R_z with respect to a common dominating measure. Let $A \in \mathcal{A}$ be such that $P_\theta(A) = P_\theta(A^{\pi})$ for all $\theta \in \Theta$. Then $R(A) \geq R(A^{\pi})$, with equality only if $R(A^{\pi}) = \infty$ or $P_\theta(A \Delta A^{\pi}) = 0$ for all θ.

The prediction region A^{π} is a Bayes procedure in the sense that it minimizes a Bayes risk, but it is also frequentist in the sense that its optimality is among procedures having the same frequentist coverage rate. Following Yu and Hoff [24], we refer to such a procedure as being "frequentist and Bayesian", or FAB.

Before obtaining a less abstract form for the optimal sets $\{A_z^{\pi} : z \in \mathcal{Z}\}$ and the resulting FAB prediction region A^{π}, we first discuss some of the conditions of the theorem. The starting point is the existence of a disintegration of R. As discussed in Chang and Pollard [6], the existence of a disintegration typically requires some topological assumptions about the underlying spaces. While not as general as it could be, the following corollary of their Theorem 1 is sufficient for many applications. In particular, the conditions on $\mathcal{X}$, $\mathcal{Y}$ and $\mathcal{Z}$ are met by Euclidean spaces, most countable spaces, many topological manifolds, and products of these spaces.

Corollary 2 (Corollary of Chang and Pollard [6, Theorem 1]). *Let $\mathcal{X}$ and $\mathcal{Y}$ be complete, separable and locally compact metric spaces, with $\mathcal{F}$ and $\mathcal{G}$ being the Borel sets. Let P_π^X be a probability measure on $(\mathcal{X},\mathcal{F})$, and let μ be a σ-finite Radon measure on $(\mathcal{Y},\mathcal{G})$. Let $\mathcal{A} = \mathcal{F} \otimes \mathcal{G}$ and define the product measure $R = P_\pi^X \times \mu$ on $(\mathcal{X} \times \mathcal{Y},\mathcal{A})$. Then R is a σ-finite Radon measure. Additionally, let $Z : (\mathcal{X} \times \mathcal{Y},\mathcal{A}) \to (\mathcal{Z},\mathcal{H})$ where $\mathcal{Z}$ is a separable metric space and $\mathcal{H}$ is the Borel σ-algebra. If the image measure ν_R of R under Z is σ-finite, then R has a (Z,ν_R)-disintegration.*

We caution that the image measure ν_R can fail to be σ-finite if P_π^X is not a proper probability measure, even if R is σ-finite. For example, consider the case that $\mathcal{X} = \mathcal{Y} = \mathbb{R}$, P_π^X and μ are both Lebesgue measure, and $Z = X + Y$. Then $\nu_R([a,b]) = \infty$ for all $a < b$. However, if $P_\pi^X(\mathbb{R}) = 1$ (or is finite) then ν_R is σ-finite.

Conditions 1 and 2 of Theorem 3.2 concern the relative absolute continuity of ν_R and $\{\nu_\theta : \theta \in \Theta\}$. Condition 1 roughly means that z-values that are impossible under the model should not contribute to the risk of a prediction procedure. More specifically, let $N \in \mathcal{H}$ be such that $\nu_\theta(N) = 0$ for all θ. A competitor A to A^{π} does not need to maintain $P_z(A_z) = P_z(A_z^{\pi})$ for $z \in N$ in order to maintain $P_\theta(A) = P_\theta(A^{\pi})$ for all θ, and so it could be that $R_z(A_z) < R_z(A_z^{\pi})$ for $z \in N$. Without the condition, if $\nu_R(N) > 0$ it is possible that $R(A) < R(A^{\pi})$. Conversely, with the condition we have $\nu_R(N) = 0$, and so $R(A) \geq R(A^{\pi})$. Condition 2 of the theorem is not completely necessary, but it does imply that A^{π} is essentially unique, in that any other set with the same coverage and risk as A^{π} can only differ from A^{π} by a set of measure zero.

3.2. Optimal regions for a given level

Recall that by Theorem 2.3, every prediction region with constant coverage also has constant conditional coverage. Theorem 3.2 then implies that if k_z in (15) is chosen to yield a conditional coverage rate of $1 - \alpha$ for all $z \in \mathcal{Z}$, then a Bayes-optimal $1 - \alpha$ constant coverage prediction region may be obtained. Specifically, if A_z^{π} is defined as in (15), with k_z chosen to satisfy $P_z(A_z^{\pi}) = 1 - \alpha$ for all z, then the FAB region A^{π} has minimum Bayes risk among all $1 - \alpha$ constant coverage prediction regions.

This result does not by itself imply that A^π is optimal among regions with non-constant coverage of $1-\alpha$ or greater, that is, regions A such that $P_\theta(A) \geq 1-\alpha$ for all θ with inequality for some θ. More generally, consider the risk optimality of A^π among prediction regions A for which $P_\theta(A) \geq P_\theta(A^\pi)$ for all $\theta \in \Theta$, with inequality for some θ. Intuitively we expect $R(A)$ to be larger than $R(A^\pi)$, since the larger coverage probability of A should correspond to a larger volume, and hence a larger risk. However, as discussed in Section 2.1, completeness of Z does not rule out the possibility that $P_z(A) < P_z(A^\pi)$ for some z even if $P_\theta(A) \geq P_\theta(A^\pi)$ for all θ, and so it is possible that $R_z(A) < R_z(A^\pi)$ for these values of z. As a result, $R(A)$ could be smaller than $R(A^\pi)$, depending on where the image measure ν_R on $(\mathcal{Z},\mathcal{H})$ places mass.

However, some models $\mathcal{P}$ are rich enough so that $P_\theta(A) \geq P_\theta(A^\pi)$ for all $\theta \in \Theta$ *does* imply that $P_z(A) \geq P_z(A^\pi)$ for $\mathcal{P}$-almost all z, in which case A^π is risk-optimal among all procedures with equal or greater coverage. Models for which this is the case are those for which the set of image probability measures $\{\nu_\theta : \theta \in \Theta\}$ on $(\mathcal{Z},\mathcal{H})$ has elements with arbitrarily high concentration on subsets of $\mathcal{Z}$:

Lemma 3.3. *Suppose for every $\epsilon > 0$ and $\{\nu_\theta : \theta \in \Theta\}$-non-null set $H \in \mathcal{H}$ there exists a $\theta_\epsilon \in \Theta$ such that $\nu_{\theta_\epsilon}(H) > 1-\epsilon$. Then if $h : (\mathcal{Z},\mathcal{H}) \to (\mathbb{R},\mathcal{B}(\mathbb{R}))$ is a bounded function for which $\int h(z)\,\nu_\theta(dz) \geq 0$ for all $\theta \in \Theta$, then $h(z) \geq 0$ for ν_θ-almost all z, for every $\theta \in \Theta$.*

From this lemma we have the following result on the risk optimality of A^π:

Theorem 3.4. *Under the conditions of Theorem 3.2 and Lemma 3.3, if $A \in \mathcal{A}$ satisfies $P_\theta(A) \geq P_\theta(A^\pi)$ for all $\theta \in \Theta$, then $R(A) \geq R(A^\pi)$, that is, A^π is risk-optimal among prediction regions with equal or greater coverage.*

Models for which the conditions of the theorem hold include those for which the location and scale of the distribution of Z can be set arbitrarily, such as multiparameter exponential families.

3.3. Expressions for optimal regions

We now obtain some less abstract expressions for the optimal sets $\{A^\pi_z : z \in \mathcal{Z}\}$ given by (15), in the case that the risk measure R dominates the model $\{P_\theta : \theta \in \Theta\}$. In this case, it will be shown that each R_z dominates the corresponding conditional distribution P_z, and so the optimal sets given by (15) may equivalently be expressed as subsets of $Z^{-1}\{z\}$ such that dP_z/dR_z exceeds some threshold. Each A^π_z may therefore be recognized as the acceptance region of a test of $H_z : (X,Y) \sim P_z$ versus $K_z : (X,Y) \sim R_z$ using the most powerful test statistic dP_z/dR_z.

Without loss of generality, assume P^X_π is dominated by a σ-finite measure λ on $(\mathcal{X},\mathcal{F})$, with density $p^X_\pi(x) > 0$ for all $x \in \mathcal{X}$ (we could take λ to be P^X_π). Then $R \equiv P_\pi \times \mu \ll \lambda \times \mu$ with density p^X_π. Since $P_\theta \ll R \ll \lambda \times \mu$ for each θ, by the factorization theorem [12], there exist functions g and $\{h_\theta : \theta \in \Theta\}$ such that

$$p_\theta(x,y) \equiv [dP_\theta/d(\lambda \times \mu)](x,y) = h_\theta(Z(x,y)) \times g(x,y). \tag{16}$$

Therefore,

$$dP_\theta/dR = h_\theta(Z(x,y)) \times g(x,y)/p^X_\pi(x). \tag{17}$$

The following theorem shows that the density of each dP_z/dR_z has a similar form:

Theorem 3.5. *Let λ and μ be σ-finite measures on $(\mathcal{X},\mathcal{F})$ and $(\mathcal{Y},\mathcal{G})$ respectively, with product measure $\lambda\times\mu$ on $(\mathcal{X}\times\mathcal{Y},\mathcal{F}\otimes\mathcal{G})$. Let R be a measure and $\{P_\theta:\theta\in\Theta\}$ be a model on $(\mathcal{X}\times\mathcal{Y},\mathcal{F}\otimes\mathcal{G})$, the latter having regular sufficient statistic $Z:(\mathcal{X}\times\mathcal{Y},\mathcal{F}\otimes\mathcal{G})\to(\mathcal{Z},\mathcal{H})$. If*

1. *$\{P_\theta:\theta\in\Theta\}\ll R\ll\lambda\times\mu$;*
2. *$dR/d(\lambda\times\mu)\equiv p_\pi^X>0$ with $\int p_\pi^X(x)\,\lambda(dx)=1$;*
3. *the image measure ν_R of R under Z is σ-finite;*
4. *R has (Z,ν_R)-disintegration $\{R_z:z\in\mathcal{Z}\}$,*

then there is a common regular conditional distribution $\{P_z:z\in\mathcal{Z}\}$ for each element of $\{P_\theta:\theta\in\Theta\}$ such that $P_z\ll R_z$ for all z with density

$$dP_z/dR_z=c(z)\times g(x,y)/p_\pi^X(x), \tag{18}$$

where $c(z)$ is a positive function and $g(x,y)$ is defined by the factorization $dP_\theta/d(\lambda\times\mu)=h_\theta(Z(x,y))\times g(x,y)$.

These results are related to Theorem 3 of Chang and Pollard [6], which describes how the disintegration of one measure can be related to the disintegration of another measure that dominates it. We note that completeness is not used in this lemma, and that the existence of a common regular conditional distribution for the model is simply the definition of Z being a regular sufficient statistic. What the lemma provides is the existence of a conditional distribution that is dominated by the disintegration of the risk measure, and an expression for the corresponding densities $\{dP_z/dR_z:z\in\mathcal{Z}\}$. This expression provides the following version of Theorem 3.2, written in terms of the somewhat familiar quantities $g(x,y)$ and $p_\pi^X(x)$:

Corollary 3. *Under the conditions of Theorem 3.2 and Theorem 3.5, let $A^\pi\in\mathcal{F}\otimes\mathcal{G}$ satisfy*

$$A_z^\pi=\{(x,y)\in Z^{-1}\{z\}:g(x,y)/p_\pi^X(x)>k_z\}. \tag{19}$$

Let $A\in\mathcal{A}$ be such that $P_\theta(A)=P_\theta(A^\pi)$ for all $\theta\in\Theta$. Then $R(A)\geq R(A^\pi)$, with equality only if $P_\theta(A\Delta A^\pi)=0$ for all θ.

Practical details concerning the constructing of the FAB region A^π for a few specific models are described in Section 4, but we make some comments here. Having observed $X=x$, construction of a $1-\alpha$ FAB prediction region amounts to determining the values $y\in\mathcal{Y}$ for which $g(x,y)/p_\pi^X(x)$ meets or exceeds the α quantile of $g(X,Y)/p_\pi^X(X)$ under P_z, the conditional distribution of (X,Y) given $Z=z$, where $z=Z(x,y)$. Thus the region is constructed by inverting tests of $(X,Y)\sim P_z$ using the test statistic $t_\pi(x,y)=g(x,y)/p_\pi^X(x)$, and so we refer to $t_\pi(x,y)$ as being a Bayes-optimal prediction statistic under the prior distribution π. In some problems it will be more convenient or familiar to work with an alternative statistic that corresponds to the same Bayes-optimal tests. For example, any statistic that is a strictly increasing function of $t_\pi(x,y)$ is also a Bayes-optimal prediction statistic, as is any statistic that can be expressed as $h(z)\times t_\pi(x,y)$ for some positive function $h(z)$. This implies, for example, that $p_{\theta_0}(x,y)/p_\pi^X(x)$ is a Bayes-optimal prediction statistic for any choice of $\theta_0\in\Theta$.

Construction of the FAB prediction region A^π may be simplified in the case that the function $Z_x:\mathcal{Y}\to\mathcal{Z}$ defined by $Z_x(y)=Z(x,y)$ is injective. As described in Corollary 1, if Z_x is injective for each x then any prediction region A can be expressed as $A=\{(x,y):x\in B_{Z(x,y)}\}$ for some $B:\mathcal{Z}\to 2^X$, and therefore derived by inverting tests of $H_z:X\sim P_z^X$. In some cases, the optimal tests are equivalent to most powerful tests of $H_z:X\sim P_z^X$ versus $K_z:X\sim P_\pi^X$, and so are those that accept H_z for large values of $p_z^X(x)/p_\pi^X(x)$, where p_z^X is the density of the conditional distribution of X given $Z=z$. We first illustrate this with an example, and then give some general results.

Example 5. Consider again the model in Example 1 where $X \sim N(\theta, k\sigma^2)$ and $Y \sim N(\theta, \sigma^2)$ are independent, with σ^2 known and $\theta \in \mathbb{R}$ unknown. Taking the volume measure μ to be Lebesgue measure on $\mathbb{R}$, $t_\pi(x,y)$ can be written as

$$t_\pi(x,y) = \exp(-\tfrac{1}{2\sigma^2}[x^2/k + y^2])/p_\pi^X(x). \tag{20}$$

Further simplification is possible in this case. For $z = (x + yk)/(1+k)$, $t_\pi(x,y)$ can be written as

$$t_\pi(x,y) = h(z)\exp(-\tfrac{k+1}{2k^2\sigma^2}(x-z)^2)/p_\pi^X(x), \tag{21}$$

which has a numerator that is proportional to the density of the $N(z, \sigma^2k^2/(k+1))$ distribution - the distribution of X given $Z = z$. Therefore, having observed $X = x$, the Bayes-optimal $1-\alpha$ constant coverage prediction region includes y if $p_z^X(x)/p_\pi^X(x)$ exceeds its α quantile under $X \sim P_z$, where $z = Z(x,y)$ and p_z^X is the conditional density of X given $Z = z$. In other words, a Bayes-optimal prediction region with constant $1-\alpha$ coverage is $A^\pi = \{(x,y) : x \in B^\pi_{Z(x,y)}\}$ where B_z^π can be expressed as

$$B_z^\pi = \{x : p_z^X(x)/p_\pi^X(x) > k_z\}, \tag{22}$$

with k_z being the α quantile of $p_z^X(X)/p_\pi^X(X)$ under $X \sim N(z, \sigma^2k^2/(k+1))$. Thus B_z^π is the acceptance region of the most powerful level-α test of $H_z : X \sim P_z^X$ versus $K : X \sim P_\pi^X$. A reasonably simple formula for the prediction region under a conjugate prior distribution for π is given in the next section.

A similar result holds for other models where Z_x is injective. For example, if the distributions $\{P_\theta : \theta \in \Theta\}$ are all dominated by counting measure with densities $p_\theta(x,y) = h_\theta(Z(x,y))g(x,y)$, then the conditional density of X given $Z = z$ is

$$p_z^X(x) = \frac{\sum_{y':Z(x,y')=z} h_\theta(Z(x,y'))g(x,y')}{\sum_{(x',y'):Z(x',y')=z} h_\theta(Z(x',y'))g(x',y')} \tag{23}$$

$$= \sum_{y':Z(x,y')=z} g(x,y')/c_z \tag{24}$$

$$= g(x, Z_x^{-1}(z))/c_z, \tag{25}$$

where the last equality holds by the injectivity of Z_x. This means that, on $\{(x,y) : Z(x,y) = z\}$, we have $g(x,y) = c_z p_z^X(x)$. By plugging this into (19) we may express the FAB prediction region as $A^\pi = \{(x,y) : x \in B^\pi_{Z(x,y)}\}$, where

$$B_z^\pi = \{x : p_z^X(x)/p_\pi^X(x) > k_z\}. \tag{26}$$

Alternatively, if $\{P_\theta : \theta \in \Theta\}$ has densities with respect to Lebesgue measure and $Z(x,y)$ is differentiable, then via the usual change of variables formula the density of (X,Z) under P_θ is

$$p_\theta^{X,Z}(x,z) = p_\theta(x, Z_x^{-1}(z))/J(x,z) \tag{27}$$

$$= h_\theta(z) \times g(x, Z_x^{-1}(z))/J(x,z), \tag{28}$$

where $J(x,z)$ is $|dZ(x,y)/dy|$ evaluated at (x, Z_x^{-1}), and so the conditional density of X given $Z = z$ is

$$p_z^X(x) = c_z \times g(x, Z_x^{-1}(z))/J(x,z). \tag{29}$$

Therefore, on $Z(x,y) = z$, the Bayes-optimal test statistic (18) is equal to $J(x,z) \times p_z^X(x)/p_\pi^X(x)$, and the FAB region may be expressed as $A^\pi = \{(x,y) : x \in B^\pi_{Z(x,y)}\}$, where

$$B_z^\pi = \{x : J(x,z) \times p_z^X(x)/p_\pi^X(x) > \tilde{k}_z\}. \tag{30}$$

If $J(x,z)$ is constant in x, then B_z^π can be written $B_z^\pi = \{x : p_z^X(x)/p_\pi^X(x) > k_z\}$, just as with discrete models. This will be the case if Z has the form $Z(x,y) = s^X(x) + s^Y(y)$, as with many exponential family models for independent X and Y.

To summarize, when Z_x is injective the optimal test statistic (18) may be written as a function of x and z, and so the optimal acceptance regions may be written as acceptance regions of tests of $H_z : X \sim P_z^X$. The inclusion of a value y into a prediction region based on observing $X = x$ can be determined as follows:

1. Set $z = Z(x,y)$.
2. Test $H_z : X \sim P_z^X$ based on observing $X = x$.
3. Include y in the region if H_z is accepted, otherwise exclude y from the region.

The coverage of such a region as a function of θ is $1 - \mathrm{E}_\theta[a(Z)]$, where $a(z)$ is the size of the test in Step 2. The region will be Bayes-optimal for its coverage function if the statistic used in Step 2 is one that is equivalent to (18), which in some cases includes $p_z^X(x)/p_\pi^X(x)$ or $J(x,z) \times p_z^X(x)/p_\pi^X(x)$.

4. Examples

4.1. Nonparametric prediction

Conformal prediction [14,23] is a method of constructing nonparametric prediction regions for a random object Y_{n+1} based on $X = (Y_1,\dots,Y_n)$, in models for which $Y_1,\dots,Y_{n+1}$ are exchangeable. A generic conformal prediction region is constructed as follows: Let $c : \mathcal{Y}^{n+1} \to \mathbb{R}$ be a function that is invariant to permutations of its first n elements. Typically, $c(y_1,\dots,y_{n+1})$ is chosen to be some numerical measure of "conformity" between y_{n+1} and the (multi)set $\{y_1,\dots,y_n\}$. Having observed $(Y_1,\dots,Y_n) = (y_1,\dots,y_n)$, a value y_{n+1} is included in the prediction region if the conformity between y_{n+1} and $\{y_1,\dots,y_n\}$ is comparable to the conformity between y_i and $\{y_1,\dots,y_{i-1},y_{n+1},y_{i+1},\dots,y_n\}$ for some minimal fraction of indices $i \in \{1,\dots,n\}$. Specifically, a value y_{n+1} is included in the $1-\alpha$ conformal prediction region if c_{n+1} is greater than the α sample quantile of $\{c_1,\dots,c_{n+1}\}$, where $c_i = c(y_1,\dots,y_{i-1},y_{n+1},y_{i+1},\dots,y_n,y_i)$. By exchangeability, the coverage probability of the resulting prediction region is greater than or equal to $1-\alpha$.

Conformal prediction is a special case of Faulkenberry's method, as applied to exchangeable models. For notational simplicity, we consider the i.i.d. case where $Y_1,\dots,Y_{n+1} \sim$ i.i.d. $P \in \mathcal{P}$ where $\mathcal{P}$ is a model on $(\mathcal{Y},\mathcal{G})$, so that the $(n+1)$-fold product measure $P^{n+1} \equiv \times_1^{n+1} P$ is the joint distribution of $(Y_1,\dots,Y_{n+1})$. Then the multiset $\{Y_1,\dots,Y_{n+1}\}$ of unordered observed values (the "order statistics") is a sufficient statistic. Faulkenberry's prediction method in this case is that, having observed $(Y_1,\dots,Y_n) = (y_1,\dots,y_n)$, a value y_{n+1} is included in the prediction region if y_{n+1} is in the acceptance region of a level-α test of Y_{n+1} being uniformly distributed on $\{y_1,\dots,y_{n+1}\}$, as this is the conditional distribution of Y_{n+1} given $\{Y_1,\dots,Y_{n+1}\} = \{y_1,\dots,y_{n+1}\}$, for any member P of $\mathcal{P}$. A test that accepts values y_{n+1} for which c_{n+1} is greater than the α sample quantile of $\{c_1,\dots,c_{n+1}\}$, is equivalent to the conformal prediction procedure.

For many models the unordered values $\{Y_1,\dots,Y_{n+1}\}$ are not only a sufficient statistic, but a complete sufficient statistic [2,18]. In particular, this will be the case if $\mathcal{P}$ is the space of probability measures dominated by a common measure μ, where μ is non-atomic (e.g. Lebesgue measure) or $\mathcal{Y}$ is countable. In these cases, the class of all prediction regions with a given coverage function may be characterized in terms of conditional coverage given the order statistics, and under some regularity conditions a Bayes-optimal FAB prediction region may be found. We first derive an expression for the region assuming that the conditions of Corollary 3 are met, and then we discuss the conditions.

Recall that Corollary 3 says that a value y is accepted into the FAB prediction region if $g(x,y)/p_\pi^X(x)$ is large compared to its conditional distribution given the sufficient statistic. In the i.i.d. sampling case being considered here where $X = (Y_1,\dots,Y_n)$ and $Y = Y_{n+1}$, the function $g(x,y)$ from the factorization theorem is constant, and so a risk-optimal prediction region is one that accepts a value y_{n+1} if $p_\pi(y_1,\dots,y_n)$ is small compared to the conditional distribution of $p_\pi(Y_1,\dots,Y_n)$ given $\{Y_1,\dots,Y_{n+1}\} = \{y_1,\dots,y_{n+1}\}$, where $p_\pi(y_1,\dots,y_n)$ is the prior predictive density of $(Y_1,\dots,Y_n)$ under the prior distribution π on $\mathcal{P}$. Recall from the discussion in Section 3.3 that an equivalent criterion is to accept y_{n+1} if $h(z)/p_\pi(y_1,\dots,y_n)$ is large, where $h(z)$ is any function of the the sufficient statistic $z = \{y_1,\dots,y_{n+1}\}$. One such function is the prior predictive density of $Y_1,\dots,Y_{n+1}$, denoted by $p_\pi(y_1,\dots,y_{n+1})$, which is constant on $\{Y_1,\dots,Y_{n+1}\} = \{y_1,\dots,y_{n+1}\}$ because prior predictive distributions under i.i.d. sampling are exchangeable. Therefore, the FAB prediction region is one that accepts values y_{n+1} for which the posterior predictive density $p_\pi(y_{n+1}|y_1,\dots,y_n) = p_\pi(y_1,\dots,y_{n+1})/p_\pi(y_1,\dots,y_n)$ is large.

More concretely, construction of a risk-optimal $1-\alpha$ prediction region proceeds as follows: The conditional distribution of $p_\pi(Y_{n+1}|Y_1,\dots,Y_n)$ given $\{Y_1,\dots,Y_{n+1}\} = \{y_1,\dots,y_{n+1}\}$ has mass $1/(n+1)$ on each of the values $c_1,\dots,c_{n+1}$, where $c_i = p_\pi(y_i|y_1,\dots,y_{i-1},y_{n+1},y_{i+1},\dots,y_n)$ for $i = 1,\dots,n$ and $c_{n+1} = p_\pi(y_{n+1}|y_1,\dots,y_n)$. If $\alpha = k/(n+1)$ for some integer k between zero and $n+1$ and there are no ties, then y_{n+1} is accepted into the $1-\alpha$ FAB prediction region if c_{n+1} is greater than the kth order statistic of $c_1,\dots,c_{n+1}$. This is equivalent to implementing the conformal prediction procedure using the posterior predictive density $p_\pi(y_{n+1}|y_1,\dots,y_n)$ as the conformity function.

We now provide some conditions under which the procedure described above provides a risk-optimal prediction region. In doing so, we use the fact that the sufficient statistic may be equivalently be expressed as the empirical distribution Z of $Y_1,\dots,Y_{n+1}$, so $[Z(y_1,\dots,y_{n+1})](G) \equiv \sum_{i=1}^{n+1} 1(y_i \in G)/(n+1)$ for $G \in \mathcal{G}$.

Theorem 4.1. *Let μ be a σ-finite Radon measure on $(\mathcal{Y},\mathcal{G})$, where $\mathcal{Y}$ is a complete, separable and locally compact metric space and $\mathcal{G}$ is the Borel σ-algebra. Let $(\mathcal{P},\sigma(w))$ be the measurable space of probability measures on $(\mathcal{Y},\mathcal{G})$ that are dominated by μ with $\sigma(w)$ being the Borel σ-algebra under the weak topology. Let π be a probability measure over $(\mathcal{P},\sigma(w))$ and let P_π^X be the probability measure on the n-fold product space $(\mathcal{Y}^n,\mathcal{G}^n)$ of $(\mathcal{Y},\mathcal{G})$, defined by $P_\pi^X(G_1\times\cdots\times G_n) = \int\prod_{i=1}^n P(G_i)\pi(dP)$. Then*

1. *$R = P_\pi^X \times \mu$ is a σ-finite Radon measure;*
2. *the image measure ν_R of R under Z is σ-finite;*
3. *R has a (Z,ν_R)-disintegration.*

If additionally

- A1. *μ is non-atomic, or $\mathcal{Y}$ is discrete and μ is counting measure, and*
- A2. *P_π^X and $\times_{i=1}^n\mu$ are mutually absolutely continuous,*

then Z is a complete regular sufficient statistic and a prediction region given by

$$A^\pi_{(y_1,\dots,y_n)} = \{y_{n+1} : p_\pi(y_{n+1}|y_1,\dots,y_n) > k_{Z(y_1,\dots,y_{n+1})}\}. \tag{31}$$

is risk-optimal in that if $P^{n+1}(A) = P^{n+1}(A^\pi)$ for all $P \in \mathcal{P}$ then $R(A) \ge R(A^\pi)$, with equality only if $P^{n+1}(A\Delta A^\pi) = 0$ for all $P \in \mathcal{P}$.

Some of these conditions may be relaxed. As discussed in Chang and Pollard [6], disintegrations of R may exist under other conditions on $(\mathcal{Y},\mathcal{G})$. Also, Z is a complete sufficient statistic for models other than the ones mentioned in assumption A1, as described in Bell, Blackwell and Breiman [2] and Mattner [18].

We comment that Theorem 3.4, regarding optimality of A^π among procedures with coverage that is bounded below by $1-\alpha$, does not apply to these nonparametric models because under i.i.d. sampling the distributions $\{\nu_P : P \in \mathcal{P}\}$ of Z cannot put mass on arbitrarily small subsets of $\mathcal{Z}$. However, a constant coverage conformal prediction procedure using $p(y_{n+1}|y_1,\dots,y_n)$ as the conformity score will be optimal among conformal procedures with equal or greater constant coverage, because all constant coverage conformal procedures have constant conditional coverage given Z.

Finally, we note that just because the model $\mathcal{P}$ is nonparametric, the prior distribution π does not need to be nonparametric for the conditions of Theorem 4.1 to hold. For example, if $\mathcal{Y}$ is a Euclidean space and μ is Lebesgue measure, then a prior distribution π that has mass only on the normal distributions yields a P^X_π that satisfies assumption A2 of the theorem. A $1-\alpha$ prediction region constructed using such a P^X_π still has exact $1-\alpha$ coverage under *all* probability distributions dominated by Lebesgue measure, and is expected to have a small volume if the true distribution P is a normal distribution, but might have a large volume if P is far from normality.

4.2. Normal populations

Let $X \sim N_p(\theta, k\Sigma)$ and $Y \sim N_p(\theta,\Sigma)$ be independent. We first consider predicting Y from X in the case that Σ is known. In this case $Z = (X + kY)/(1+k)$ is a complete sufficient statistic, with P^X_z being the $N_p(z, \Sigma k^2/(k+1))$ distribution. Since Z_x is injective for each x, a Bayes-optimal $1-\alpha$ prediction region can be constructed from Equation (26), so in particular, a value of y is accepted into the prediction region if $p^X_z(x)/p^X_\pi(x)$ exceeds the α quantile of $p^X_z(X)/p^X_\pi(X)$ under $X \sim P^X_z$, where $z = Z(x,y)$. If the prior information for θ is represented by $\theta \sim N_p(\mu, \lambda\Sigma)$ for some $\mu \in \mathbb{R}^p$ and $\lambda > 0$, then P^X_π is the $N_p(\mu,(k+\lambda)\Sigma)$ distribution. After some manipulation of $\log p^X_z(x)/p^X_\pi(x)$, a Bayes-optimal test statistic is obtained:

$$t_z(x) = ||\Sigma^{-1/2}(x-z)/v^{1/2} + \delta_z||^2 \tag{32}$$

where $\delta_z = \Sigma^{-1/2}(\mu - z)v^{1/2}/(\lambda + k/(k+1))$ with $v = k^2/(k+1)$ and $\Sigma^{1/2}$ is any matrix for which $\Sigma^{1/2}(\Sigma^{1/2})^\top = \Sigma$. Since $\Sigma^{-1/2}(x-z)/v^{1/2} \sim N(0,I)$ under $H_z : X \sim P^X_z$, the null distribution of the statistic is a non-central χ^2_p distribution with noncentrality parameter $||\delta_z||^2$. Thus, a value y is included in the prediction region if $t_{Z(x,y)}(x)$ is less than $\chi^2_{p,||\delta_z||^2,1-\alpha}$, the $1-\alpha$ quantile of this distribution.

Some intuition for this statistic may be obtained by expressing it slightly differently. One useful reëxpression is obtained by recalling that $z = (x+ky)/(1+k)$ on $Z(x,y) = z$, which gives

$$t_{Z(x,y)}(x) = ||\Sigma^{-1/2}(x-y)/\sqrt{k+1} + \delta_{Z(x,y)}||^2, \tag{33}$$

and so the Bayes-optimal $1-\alpha$ prediction region can be written

$$A^\pi_x = \{y : ||\Sigma^{-1/2}(x-y)/\sqrt{k+1} + \delta_{Z(x,y)}||^2 < \chi^2_{p,||\delta_{Z(x,y)}||^2,1-\alpha}\}. \tag{34}$$

For comparison, the usual equivariant prediction region is

$$A^E_x = \{y : ||\Sigma^{-1/2}(x-y)/\sqrt{k+1}||^2 < \chi^2_{p,0,1-\alpha}\}. \tag{35}$$

Since $\delta_z \to 0$ as $\lambda \to \infty$, this standard region can be viewed as a limit of regions of the form (34) under a sequence of increasingly non-informative prior distributions.

The optimal statistic (32) and region (34) can also be related to a fully Bayesian prediction procedure as well. Rearranging terms in (32) gives

$$t_{Z(x,y)} = \|\Sigma^{-1/2}(y - \hat{\theta}^{\pi})/v_{\lambda}^{1/2}\|^2 \times (k+1)/v_{\lambda} \tag{36}$$

$$A_x^{\pi} = \{y : \|\Sigma^{-1/2}(y - \hat{\theta}^{\pi})/v_{\lambda}^{1/2}\|^2 \times (k+1)/v_{\lambda} < \chi^2_{p,\|\delta_{Z(x,y)}\|^2,1-\alpha}\}, \tag{37}$$

where $\hat{\theta}^{\pi} = (X/k + \mu/\lambda)/(1/k + 1/\lambda)$ is the posterior mean estimator of θ given X, and $v_{\lambda} = (\lambda(k+1) + k)/(k+\lambda)$ relates to the prior predictive variance of $Y - \hat{\theta}$, which is $v_{\lambda}\Sigma$. For comparison, the posterior predictive distribution for Y under the prior $\theta \sim N_p(\mu, \lambda I)$ is $Y|\{X = x\} \sim N_p(\hat{\theta}^{\pi}, v_{\lambda}\Sigma)$, and so the fully Bayesian $1 - \alpha$ prediction region with highest posterior predictive density is given by

$$A_x^B = \{y : \|\Sigma^{-1/2}(y - \hat{\theta}^{\pi})/v_{\lambda}^{1/2}\|^2 < \chi^2_{p,0,1-\alpha}\}. \tag{38}$$

However, since the distribution of $\|\Sigma^{-1/2}(y - \hat{\theta}^{\pi})/v_{\lambda}^{1/2}\|^2$ (unconditionally on Z) depends on the unknown value of θ, this fully Bayesian posterior predictive region will have frequentist coverage that varies as a function of θ. In contrast, the FAB region is obtained by comparing (36) to its $1 - \alpha$ conditional quantile given $Z = z$ in order to maintain constant frequentist coverage as a function of θ. In particular, unlike the posterior predictive region, the region (37) is *not* centered around $\hat{\theta}^{\pi}$ because the conditional quantile depends on the noncentrality parameter $\|\delta_z\|^2$, which varies as a function of $z = Z(x, y)$ and hence varies with y.

Some numerical comparisons of FAB regions to the standard equivariant regions are given in Figures 1, 2 and 3. The first figure displays 90% prediction intervals and widths as a function of x for the case that $p = 1$, $\sigma^2 = 1$, $k = 1$ and $\mu = 0$. As shown in the left panel for the case that $\lambda = 1$, when x is close to zero (as is expected under the prior distribution), the FAB interval is narrower than the equivariant interval ($\lambda = \infty$), at the cost of being wider for values of x that are less likely under the prior distribution. The right panel of the figure summarizes the widths for a range of λ values. The biggest differences between the FAB and equivariant intervals occur for highly informative prior distributions, that is, when λ is small. In contrast, for large values of λ the FAB interval can be better than the usual interval over a wide range of x values, but the improvement is smaller. Figure 2 displays analogous results for the case that $p = 2$, $\Sigma = I$, $k = 1$ and $\mu = (0,0)$. When x is in accord with the prior distribution the FAB region can be substantially smaller than the usual region - close to half the area in this case. As x moves away from μ the area increases in order to accommodate both the prior distribution and the requirement of 90% frequentist coverage. Figure 3 compares the frequentist risk of the FAB procedures as a function of θ and λ for $p = 1$ and $p = 2$. The risk differences are smaller than the volume differences as the former are obtained by averaging the latter over the values of X, with respect to the distribution $X \sim N_p(\theta, k\Sigma)$.

In practice the covariance matrix Σ will be unknown, but estimable from available data. For example, suppose we want to predict $Y \sim N_p(\theta, \Sigma)$ from $X = (\hat{\theta}, \hat{\Sigma})$, where $\hat{\theta} \sim N_p(\theta, k\Sigma)$ and $\nu\hat{\Sigma}$ ~Wishart(ν, Σ) are independent of each other. A complete sufficient statistic for the joint distribution of Y and X is $Z = \{(\hat{\theta} + kY)/(1+k), \nu\hat{\Sigma} + YY^{\top} + \hat{\theta}\hat{\theta}^{\top}/k\}$. Furthermore, Z_x is injective for each x, and so given a prior distribution on (θ, Σ) one could construct a FAB prediction region from (26). However, without going into too many details, this approach will be quite cumbersome as it involves singular conditional distributions and an optimal test statistic that must be numerically approximated, at least for any prior distribution on Σ of which I am aware. As an alternative, a simpler FAB statistic may be constructed

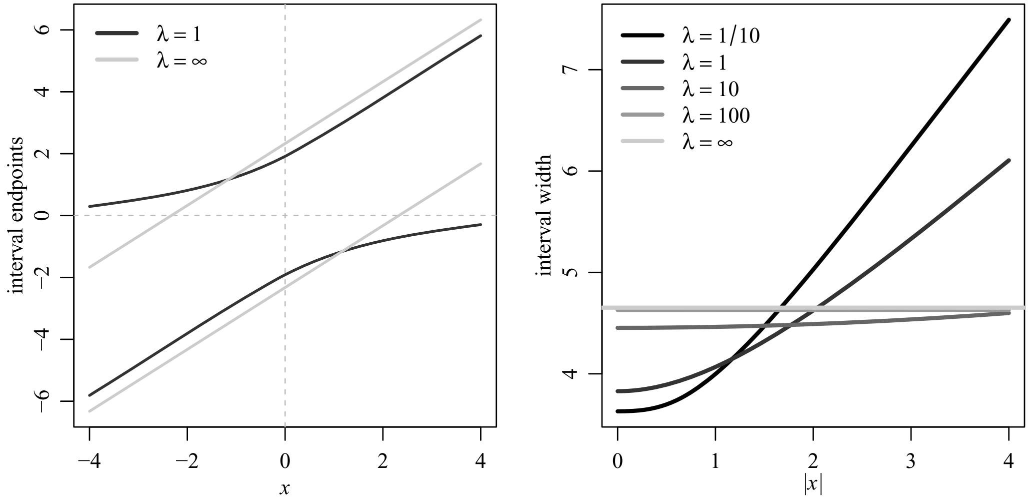

Figure 1. Normal 90% prediction intervals for $p = 1$. Left panel: Intervals as a function of x for $\lambda = 1$ and $\lambda = \infty$. Right panel: Interval widths as a function of $|x|$ for $\lambda \in \{.1, 1, 10, 100, \infty\}$. Finite values of λ correspond to FAB procedures, $\lambda = \infty$ corresponds to the standard equivariant procedure.

by replacing each appearance of Σ in (33) with estimates: Changing notation slightly, let $\hat{\theta} = X$, $Z = (X + kY)/(1 + k)$ and consider the statistic

$$t_{Z(x,y)}(x) = \|\hat{\Sigma}^{-1/2}(x - y)/\sqrt{k+1} + \tilde{\delta}_{Z(x,y)}\|^2 \tag{39}$$

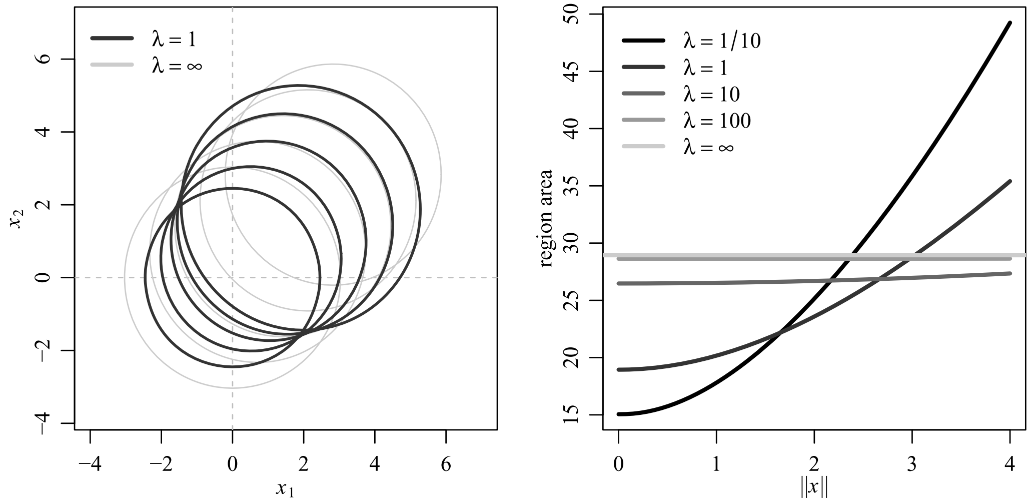

Figure 2. Normal 90% prediction regions for $p = 2$. Left panel: Regions as a function of x for $\lambda = 1$ and $\lambda = \infty$. Right panel: Region areas as a function of $\|x\|$ for $\lambda \in \{.1, 1, 10, 100, \infty\}$. Finite values of λ correspond to FAB procedures, $\lambda = \infty$ corresponds to the standard equivariant procedure.

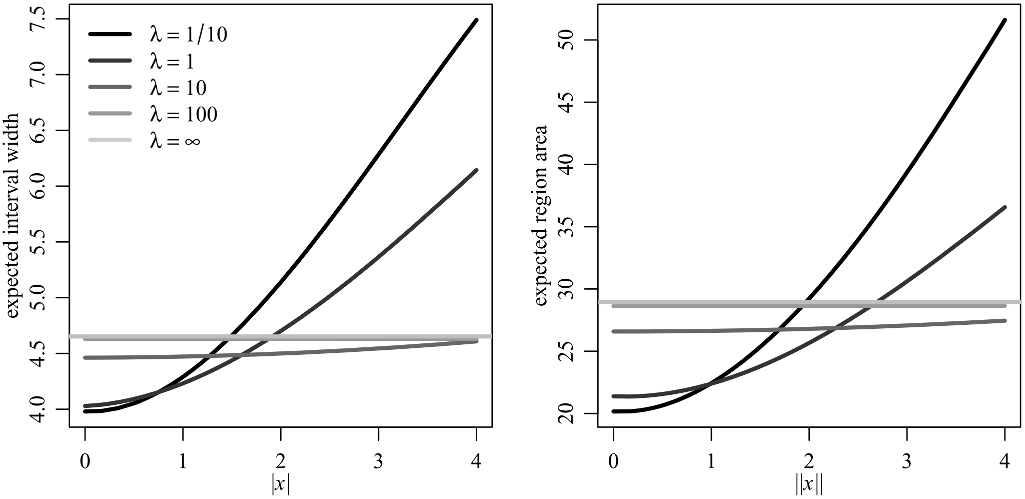

Figure 3. Expected 90% prediction region volumes as a function of x and λ for $p = 1$ (left panel) and $p = 2$ (right panel). Finite values of λ correspond to FAB procedures, $\lambda = \infty$ corresponds to the standard equivariant procedure.

where $\tilde{\delta}_z = \tilde{\Sigma}^{-1/2}(\mu - z)v^{1/2}/(v_\lambda - v)$, and $\hat{\Sigma}^{-1/2}$ and $\tilde{\Sigma}^{-1/2}$ are estimates of of $\Sigma^{-1/2}$. Specifically, let $\hat{\Sigma}^{-1/2}$ be the Cholesky factorization of $\hat{\Sigma}^{-1}$, and let $\tilde{\Sigma}^{-1/2}$ be any other estimate of $\Sigma^{-1/2}$ that is deterministic (e.g. based on prior information) or statistically independent of $\hat{\theta}$ and $\hat{\Sigma}$. It is straightforward to show that the conditional distribution of $(X - Y)/\sqrt{k+1}$ given $Z = z$ is $N_p(0, \Sigma)$, and further that the conditional distribution of $\hat{\Sigma}^{-1/2}(X - Y)/\sqrt{k+1}$ given $Z = z$ does not depend on the unknown parameters (μ, Σ). Therefore, $t_{Z(x,y)}$ can be used to evaluate each hypothesis $H_z : X \sim P_z^X$, and these tests may be inverted to obtain the approximately optimal prediction region

$$\tilde{A}_x^\pi = \{y : \|\hat{\Sigma}^{-1/2}(x - y)/\sqrt{k+1} + \tilde{\delta}_{Z(x,y)}\|^2 < q_{\tilde{\delta}_{Z(x,y)},1-\alpha}\} \tag{40}$$

where $q_{b,1-\alpha}$ is the $1 - \alpha$ quantile of the distribution of $\|T + b\|^2$, where $T \overset{d}{=} \hat{\Sigma}^{-1/2}(X - Y)/\sqrt{k+1}$. If $p = 1$ then T has the t-distribution with v degrees of freedom.

4.3. Linear regression

Suppose we wish to predict $Y \sim N(v^\top \beta, \sigma^2)$ from $X \sim N_n(U\beta, \sigma^2 I)$ where X and Y are independent, and $v \in \mathbb{R}^p$ and $U \in \mathbb{R}^{n \times p}$ are fixed, non-stochastic explanatory variables. We first consider the case that σ^2 is known. In this case, $Z = U^\top X + vY$ is a complete sufficient statistic, and the joint density of (X, Y) factorizes as $p_\beta(x, y) = h_\beta(z) \times g(x, y)$ where $g(x, y) = \exp\{-(x^\top x + y^2)/(2\sigma^2)\}$. Under the prior distribution $\beta \sim N_p(0, \sigma^2 \Psi^{-1})$, the prior predictive distribution of X is $N_p(0, \sigma^2(I + U\Psi^{-1}U^\top))$. By Corollary 3, the FAB prediction region is formed by inverting tests that accept $H_z : (X, Y) \sim P_z$ when $g(x, y)/p_\pi^X(x)$ is large. After some manipulation, we have

$$-2\sigma^2 \log g(x, y)/p_\pi^X(x) = y^2 + x^\top U(U^\top U + \Psi)^{-1} U^\top x. \tag{41}$$

On $Z(x, y) = z$ we have $U^\top x = z - vy$ and so (41) can be written

$$-2\sigma^2 \log g(x, y)/p_\pi^X(x) = y^2[1 + v^\top (U^\top U + \Psi)^{-1} v] - 2y v^\top (U^\top U + \Psi)^{-1} z + c_z \tag{42}$$

where c_z does not depend on y. By completing the square and rearranging, we see that an optimal test is one that accepts H_z for small values of the statistic

$$t_z(y) = \tfrac{1}{\sigma}\sqrt{1 + v^\top(U^\top U)^{-1}v} \times \left| y - \frac{v^\top(U^\top U + \Psi)^{-1}z}{1 + v^\top(U^\top U + \Psi)^{-1}v} \right|$$

$$= \tfrac{\sqrt{w_0}}{\sigma}|y - v^\top[S_\psi/w_\psi]z| \tag{43}$$

where $S_\psi = (U^\top U + \Psi)^{-1}$ and $w_\psi = 1 + v^\top S_\psi v$, and so in particular $S_0 = (U^\top U)^{-1}$ and $w_0 = 1 + v^\top(U^\top U)^{-1}v$. To find the critical value for this test, rewrite (43) as

$$t_z(y) = \left| \tfrac{\sqrt{w_0}}{\sigma}(y - v^\top[S_0/w_0]z) + \delta_z \right|, \tag{44}$$

where $\delta_z = v^\top(S_0/w_0 - S_\psi/w_\psi)z\sqrt{w_0}/\sigma$. Since $Y|\{Z = z\} \sim N(v^\top[S_0/w_0]z, \sigma^2/w_0)$, the critical value q_z for the level-α test using this statistic is the solution in q to the equation $\Phi(q - \delta_z) - \Phi(-q - \delta_z) = 1 - \alpha$, which may be obtained numerically. The $1 - \alpha$ FAB prediction region therefore consists of y-values for which $t_{Z(x,y)}(y) \le q_{Z(x,y)}$.

The prediction region generated by the FAB test statistic (44) bears some resemblance to the standard equivariant $1 - \alpha$ prediction region: The first term in (44) may be rearranged as follows:

$$y - v^\top[S_0/w_0]z = y - v^\top(U^\top U)^{-1}(U^\top x + vy)/w_0 \tag{45}$$

$$= y - v^\top\hat{\beta}/w_0 - (w_0 - 1)y/w_0 \tag{46}$$

$$= (y - \hat{\beta}^\top v)/w_0 \tag{47}$$

where $\hat{\beta}$ is the ordinary least-squares estimate. Therefore, the statistic may be written $t_z(y) = |(y - \hat{\beta}^\top v)/(\sigma\sqrt{w_0}) + \delta_z|$, and the FAB prediction region is

$$A_x^\pi = \{y : |(y - \hat{\beta}^\top v)/(\sigma\sqrt{w_0}) + \delta_{Z(x,y)}| \le q_{Z(x,y)}\}. \tag{48}$$

For comparison, the equivariant region based on the pivotal quantity $y - \hat{\beta}^\top v$ is $\{y : |(y - \hat{\beta}^\top v)/(\sigma\sqrt{w_0})| \le \Phi^{-1}(1 - \alpha/2)\}$. This interval is a limit of FAB intervals of the form (48): As the prior distribution becomes increasingly non-informative as the eigenvalues of Ψ go to zero, we have $S_\psi \longrightarrow S_0$ and $w_\psi \longrightarrow w_0$, and so $\delta_z \longrightarrow 0$ and $q_z \longrightarrow \Phi^{-1}(1 - \alpha/2)$.

To relate the FAB region to the fully Bayesian region derived from the posterior predictive distribution of Y given $X = x$, note that

$$v^\top(U^\top U + \Psi)^{-1}z = v^\top(U^\top U + \Psi)^{-1}(U^\top x + vy) \tag{49}$$

$$= v^\top\hat{\beta}_\psi + (w_\psi - 1)y \tag{50}$$

where $\hat{\beta}_\psi = (U^\top U + \Psi)^{-1}U^\top x$ is the posterior mean of β given $X = x$ under the prior distribution $\beta \sim N_p(0, \sigma^2\Psi^{-1})$. The statistic (44) may therefore be written

$$t_{Z(x,y)} = |y - \hat{\beta}_\psi^\top v| \times \sqrt{w_0}/(\sigma w_\psi), \tag{51}$$

and the prediction region may be expressed as

$$A_x^\pi = \{y : \hat{\beta}_\psi^\top v - q_{Z(x,y)}\sigma w_\psi/\sqrt{w_0} < y < \hat{\beta}_\psi^\top v + q_{Z(x,y)}\sigma w_\psi/\sqrt{w_0}\}. \tag{52}$$

This can be compared to the fully Bayesian interval as follows: The posterior distribution of β given $X = x$ is $\beta|\{X = x\} \sim N_p(\hat{\beta}_\psi, \sigma^2(U^\top U + \Psi)^{-1})$, which gives the posterior predictive distribution $Y|\{X = x\} \sim N(\hat{\beta}_\psi^\top v, \sigma^2 w_\psi)$. The $1-\alpha$ highest posterior predictive density region for Y given $X = x$ is therefore

$$A_x = \{y : \hat{\beta}_\psi^\top v - \Phi^{-1}(1-\alpha/2)\sigma\sqrt{w_\psi} < y < \hat{\beta}_\psi^\top v + \Phi^{-1}(1-\alpha/2)\sigma\sqrt{w_\psi}\}. \tag{53}$$

This posterior predictive interval is centered around the biased estimator $\hat{\beta}_\psi^\top v$ of $\beta^\top v$, and as a result will have a frequentist coverage rate that varies in β. In contrast, the FAB interval (52) is not centered around $\hat{\beta}_\psi^\top v$, as the acceptable deviation from this estimator varies in y through $q_{Z(x,y)}$, which is set conditionally in order to maintain $1-\alpha$ frequentist coverage for all values of β.

Typically the value of σ^2 is unknown and must be estimated from the data. As in the previous subsection, we forego formulating a prior distribution for σ^2 and instead describe a FAB prediction region that maintains exact $1-\alpha$ frequentist coverage for every value of β but is only approximately risk-optimal. As an alternative to (44) consider the statistic

$$t_z(y) = |\tfrac{\sqrt{w_0}}{\hat{\sigma}}(y - v^\top[S_0/w_0]z) + \tilde{\delta}_z|, \tag{54}$$

where $\tilde{\delta}_z = v^\top(S_0/w_0 - S_\psi/w_\psi)z\sqrt{w_0}/\tilde{\sigma}$, with $\hat{\sigma}^2$ and $\tilde{\sigma}^2$ being two estimates of σ^2 that are statistically independent of each other and of $U^\top X$. In particular, assume that $\nu\hat{\sigma}^2/\sigma^2 \sim \chi^2_\nu$, in which case $(y - v^\top[S_0/w_0]z)\sqrt{w_0}/\hat{\sigma} \sim t_\nu$ under $H_z : Y \sim P_z^Y$. The critical value q_z for a level-α test of H_z therefore satisfies $F_\nu(q_z - \tilde{\delta}_z) - F_\nu(-q_z - \tilde{\delta}_z) = 1-\alpha$, where F_ν is the cumulative distribution function of the t_ν distribution. Rearranging terms as before, the resulting FAB prediction region can be expressed

$$A_x^\pi = \{y : |(y - \hat{\beta}^\top v)/(\hat{\sigma}\sqrt{w_0}) + \tilde{\delta}_z| \le q_{Z(x,y)}\}. \tag{55}$$

Appropriate values of $\hat{\sigma}^2$ and $\tilde{\sigma}^2$ can often be obtained from the data at hand: For example, both of these quantities may be obtained from a decomposition of the residual sum of squares from the regression of X on U. Details on this approach are available from the replication files for this article at my website. Alternatively, $\hat{\sigma}^2$ could be the usual unbiased estimate of σ^2 from this regression, with $\tilde{\sigma}^2$ coming from non-stochastic prior information or other data that are independent of X.

Some risk comparisons are displayed in Figure 4. A single 100 × 75 matrix U was randomly generated, with columns that were correlated but having zero mean and unit variance. Expected widths of regions for predicting $Y \sim N(v^\top\beta, \sigma^2)$ were computed, where v ranged over the rows of U. The expected interval widths were averaged over these values of v to obtain an average Bayes risk. This was done for $\sigma^2 = 1$ and under prior distributions $\beta \sim N_{75}(0, \tau^2 I)$, where τ^2 ranged from 1/10 to 10. The horizontal light gray line in the figure gives the average width of the standard equivariant interval (labeled with $\tau_\pi = \infty$), which is constant as a function of τ^2. The medium gray line gives the average Bayes risk of the approximately optimal FAB procedure, given by (55), as a function of τ (labeled $\tau_\pi = \tau$). For the lowest value $\tau^2 = 1/10$, the FAB interval has an average expected width that is about 10% smaller than the equivariant interval. As τ^2 increases the risk remains below that of the equivariant interval but the improvement decreases, in accord with our understanding of the equivariant interval as a limit of FAB intervals.

We briefly consider the performance of the FAB procedure when the prior distribution is not in accord with the actual value of β. We expect that the FAB procedure will improve upon the equivariant interval as long as the prior distribution π is sufficiently diffuse, i.e. the prior variance τ^2 used to construct the FAB region is not much smaller than the true magnitude of β. However, using a very diffuse prior will not provide much of a risk improvement over the equivariant interval. Since the risk improvements are achieved when τ^2 is small, it is more useful to consider the potential downside to

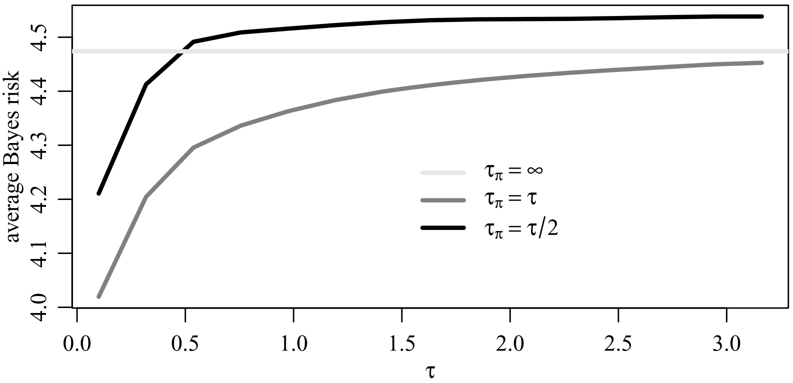

Figure 4. Bayes risk comparisons of 90% FAB and equivariant prediction intervals for the normal linear model, averaged across vectors of explanatory variables. The horizontal gray line is the risk of the equivariant interval. The medium gray line is the average Bayes risk of the FAB interval. The black line is the average Bayes risk of a FAB interval using an overly-concentrated prior distribution.

using a prior distribution that is overly concentrated around zero. An example of this behavior is given by the black line in Figure 4, which gives the Bayes risk of a FAB procedure for which the "true" τ^2 (the one used to compute the Bayes risk) is four times that of τ_π^2, the value used to construct the FAB region. The figure indicates that, for these values of U and σ^2, an overly-concentrated prior distribution still yields a FAB region that improves upon the equivariant interval for small values of τ^2, but can be worse for larger values. However, we reiterate that even if the prior distribution is misspecified, the FAB procedure will still maintain $1-\alpha$ frequentist coverage for each value of β, no matter how large or small.

5. Discussion

The FAB prediction procedure introduced in this article allows for incorporation of prior or indirect information while still maintaining a target frequentist coverage rate. In many cases, the FAB procedure is Bayes-risk optimal among procedures with a given frequentist coverage rate. In practice, this means that a FAB prediction region will have a smaller expected volume than other regions with the same coverage rate, if there is not a large discrepancy between the prior distribution and the population from which the data are to be sampled. If there is a large discrepancy, the FAB region will still maintain the target frequentist coverage rate but could have a volume that is large compared to other procedures. This raises the question of when, in practice, would one be confident enough in their prior distribution to use a FAB procedure? One possibility is with multipopulation scenarios, where the "prior distribution" for a population may be obtained using data from the others, perhaps via a hierarchical model. Specifically, suppose $(X_j, Y_j) \sim P_{\theta_j}$ independently for $j \in \{1,\dots,p\}$. For prediction of Y_j from X_j, one could fit a hierarchical model for the parameters $\theta_1,\dots,\theta_p$ using data $X_1,\dots,X_{j-1},X_{j+1},\dots,X_p$. The hierarchical model provides indirect information about θ_j that can be used to construct a FAB prediction region. The resulting region for Y_j will still maintain frequentist coverage because the prior distribution for θ_j is obtained from data that are statistically independent of (Y_j, X_j). This type of approach has been used before for frequentist confidence interval construction for univariate normal populations [5,24]. In those applications, FAB confidence intervals maintain population-specific frequentist coverage guarantees, while being narrower than standard frequentist procedures on average across populations.

The FAB prediction regions in this article are constructed with a conditional representation of coverage via the conditional distributions $\{P_z : z \in \mathcal{Z}\}$. If $P_z(A_z) = 1 - \alpha$ for $\mathcal{P}$-almost all z, then $P_\theta(A)$ must have $1 - \alpha$ coverage for all θ. A reviewer has suggested that this conditional representation might also be useful for investigating robustness of procedures to deviations from modeling assumptions. For example, if we have only $P_z(A_z) \approx 1 - \alpha$, or that this holds only for some z, then the actual coverage could be computed as $\int P_z(A_z)\,\nu_\theta(dz)$. This representation might also provide a way to develop procedures that have approximate $1 - \alpha$ coverage.

In addition to the aforementioned topics, other areas of further research include development of methods for specific models, such as general and generalized linear models, exponentially parameterized random graph models for networks, and the contamination models described in Mattner [18]. Nontrivial details to be worked out for each model include identification of optimal test statistics as well as efficient computational methods for inverting the tests to construct a prediction region. Additionally, in models for which the conditions of Theorem 3.4 do not hold, there remains the interesting open question of whether or not a FAB prediction region can have larger Bayes risk than a region with equal or greater coverage.

Computer code to construct FAB prediction regions for the multivariate normal and normal linear regression models is available at https://github.com/pdhoff/FABPrediction.

Proofs

Proof of Lemma 2.1. First we show that the function that maps each set-valued function to its graphs is a bijection. To see that this function is surjective, note that for any subset A of $\mathcal{X} \times \mathcal{Y}$, the graph of the set-valued function $x \mapsto A_x$, where $A_x = \{y : (x, y) \in A\}$, is A. To see that it is injective, suppose that two set-valued functions, say $x \mapsto A_x$ and $x \mapsto A'_x$, are not the same. Then there exists an $\tilde{x}$ for which $A_{\tilde{x}} \neq A'_{\tilde{x}}$, and so there is a $\tilde{y}$ that is an element of one but not the other. Suppose $\tilde{y}$ is in $A_{\tilde{x}}$ but is not in $A'_{\tilde{x}}$. Then $(\tilde{x}, \tilde{y})$ is in the graph of $x \mapsto A_x$ but not that of $x \mapsto A'_x$.

Now we show that the function that maps a set-valued function $z \mapsto A_z$ to $\cup_{z\in\mathcal{Z}} A_z$ is a bijection from item 3 to item 1. To see that it is surjective, note that for a given subset A of $\mathcal{X} \times \mathcal{Y}$ the set-valued function $z \mapsto A_z$ with $A_z = A \cap Z^{-1}\{z\}$ satisfies $\cup_{z\in\mathcal{Z}} A_z = A$. To see that it is injective, suppose the functions $z \mapsto A_z$ and $z \mapsto A'_z$, are not the same, and that A_z and A'_z are subsets of $Z^{-1}\{z\}$ for each z. Then there exists a $\tilde{z}$ for which $A_{\tilde{z}} \neq A'_{\tilde{z}}$, and so there is a point $(\tilde{x}, \tilde{y})$ with $\tilde{z} = Z(\tilde{x}, \tilde{y})$ that is an element of one but not the other. Suppose $(\tilde{x}, \tilde{y})$ is in $A_{\tilde{z}}$ but is not in $A'_{\tilde{z}}$. Then $(\tilde{x}, \tilde{y})$ must be in $\cup_{z\in\mathcal{Z}} A_z$ because it is in $A_{\tilde{z}}$. But $(\tilde{x}, \tilde{y})$ cannot be in $\cup_{z\in\mathcal{Z}} A'_z$, because it is not in $A'_{\tilde{z}}$, and the condition that $A'_z \subset Z^{-1}\{z\}$ for each z precludes $(\tilde{x}, \tilde{y})$ from being in A'_z for any z other than $\tilde{z}$. □

Proof of Lemma 2.2. Let Z_y be injective. Then any $A \subset \mathcal{X} \times \mathcal{Y}$ can be written as $A = \{(x, y) : y \in C_{Z(x,y)}\}$ where $C_z = \{y \in f(Z^{-1}(z)) : (Z_y^{-1}(z), y) \in A\}$, where f is the canonical projection map from $2^{\mathcal{X}\times\mathcal{Y}} \to 2^{\mathcal{Y}}$. Conversely, suppose Z_y is not injective for a particular $y \in \mathcal{Y}$. Then there exists $\{x_a, x_b\} \subset \mathcal{X}$ with $x_a \neq x_b$ but for which $Z(x_a, y) = Z(x_b, y)$. Now let A include (x_a, y) but not (x_b, y). Then if C is to represent A, we must have $y \in C_{Z(x_a,y)}$. This would imply $y \in C_{Z(x_b,y)}$, since $Z(x_a, y) = Z(x_b, y)$. So C cannot represent A, since $(y, x_b) \notin A$ by assumption. □

Proof of Corollary 1. Apply Lemma 2.2 with the roles of $\mathcal{X}$ and $\mathcal{Y}$ interchanged. □

Proof of Theorem 2.3. Suppose $P_\theta(A) = 1 - \alpha$ for all θ. For each $z \in \mathcal{Z}$ define $A_z = A \cap Z^{-1}\{z\}$. Then $\cup A_z = A \in \mathcal{A}$ and $A_z \subset Z^{-1}\{z\}$ so conditions 1 and 2 are satisfied. By regular sufficiency, the set $H_1 = \{z : P_z(Z^{-1}\{z\}) = 1\}$ has P_θ-probability one for each θ, and by completeness, the set $H_2 = \{z : P_z(A) = 1 - \alpha\}$ also has P_θ-probability one for each θ. Therefore, the set $H = H_1 \cap H_2$ has P_θ-probability one for each θ. For each z in H, we therefore have $P_z(A_z) = P_z(A) = 1 - \alpha$, and so condition 3 is satisfied. Conversely, suppose $\{A_z : z \in \mathcal{Z}\}$ defines a set-valued function that satisfies conditions 1, 2 and 3. Let $A = \cup_{z \in \mathcal{Z}} A_z$ and note that $A \cap Z^{-1}\{z\} = A_z$ by condition 2. Let $H = \{z : P_z(A_z) = 1 - \alpha\} \cap \{z : P_z(Z^{-1}\{z\}) = 1\}$. By condition 3 and regular sufficiency of Z, $\nu_\theta(H) = 1$ for all θ, where ν_θ is the marginal probability measure of Z under P_θ. For all θ we therefore have

$$P_\theta(A) = \int P_z(A)\,\nu_\theta(dz) = \int_H P_z(A \cap Z^{-1}\{z\})\,\nu_\theta(dz) \tag{56}$$

$$= \int_H P_z(A_z)\,\nu_\theta(dz) \tag{57}$$

$$= \int_H (1 - \alpha)\,\nu_\theta(dz) = 1 - \alpha. \tag{58}$$

□

Proof of Theorem 2.4. The proof is essentially the same as that of Theorem 2.3 and so is omitted. □

Proof of Lemma 3.1. For sets A and $\tilde{A}$ described in the lemma,

$$0 \leq P(A) - P(\tilde{A}) = P(A \cap \tilde{A}^c) - P(A^c \cap \tilde{A}). \tag{59}$$

Similarly, $R(A) - R(\tilde{A}) = R(A \cap \tilde{A}^c) - R(A^c \cap \tilde{A})$. The first term in this difference is

$$R(A \cap \tilde{A}^c) = \int_{A \cap \tilde{A}^c} r(\omega)\,\lambda(d\omega) \tag{60}$$

$$\geq \int_{A \cap \tilde{A}^c} [p(\omega)/k]\,\lambda(d\omega) = P(A \cap \tilde{A}^c)/k. \tag{61}$$

As for the second term in the risk difference, in the case that $P(A^c \cap \tilde{A}) > 0$ we have

$$P(A^c \cap \tilde{A})/k = \int_{A^c \cap \tilde{A}} [p(\omega)/k]\,\lambda(d\omega) > \int_{A^c \cap \tilde{A}} r(\omega)\lambda(d\omega) = R(A^c \cap \tilde{A}) \tag{62}$$

because $r(\omega) < p(\omega)/k$ on $\tilde{A}$. Therefore, if $P(A^c \cap \tilde{A}) > 0$, we have $R(A \cap \tilde{A}^c) \geq P(A \cap \tilde{A}^c)/k$, $R(A^c \cap \tilde{A}) < P(A^c \cap \tilde{A})/k$, and so

$$R(A) - R(\tilde{A}) = R(A \cap \tilde{A}^c) - R(A^c \cap \tilde{A}) \tag{63}$$

$$> [P(A \cap \tilde{A}^c) - P(A^c \cap \tilde{A})]/k \geq 0. \tag{64}$$

In the case that $P(A^c \cap \tilde{A}) = 0$ we must also have $R(A^c \cap \tilde{A}) = 0$ since $r(\omega) < p(\omega)/k$ on $\tilde{A}$. The risk difference in this case is then just $R(A \cap \tilde{A}^c)$, which was already shown to be greater than or equal to $P(A \cap \tilde{A}^c)$, which is greater than or equal to $P(A \cap \tilde{A}^c)$, which is zero in this case. Therefore, the risk difference is positive and can only be zero if $P(A \Delta \tilde{A}) = 0$. □

Proof of Theorem 3.2. Let $A_z = A \cap Z^{-1}\{z\}$ for each z. By regular sufficiency and completeness, there exists a set $H_P \in \mathcal{H}$ such that $\nu_\theta(H_P) = 1$ for all θ and $P_z(A_z^\pi) = P_z(A_z)$ for all $z \in H_P$. Let

$H_R = \{z : R_z(Z^{-1}\{z\}) = 1 - R_z(Z^{-1}\{z\}^c) = 1\}$, that is, the subset of $\mathcal{Z}$ for which R_z is a probability measure concentrated on $Z^{-1}\{z\}$. Now $\nu_R(H_P^c) = 0$ by assumption 1, and $\nu_R(H_R^c) = 0$ by Theorem 2 of Chang and Pollard [6], and so $\nu_R((H_P \cap H_R)^c) = 0$ as well. This implies that

$$R(A) = \int R_z(A)\,\nu_R(dz) = \int_{H_P \cap H_R} R_z(A_z)\,\nu_R(dz) \tag{65}$$

$$\geq \int_{H_P \cap H_R} R_z(A_z^\pi)\,\nu_R(dz) = R(A^\pi), \tag{66}$$

where the second line holds because $R_z(Z^{-1}\{z\}^c) = 0$ for all $z \in H_R$, and the third line holds because for $z \in H_P \cap H_R$, R_z is a probability measure ($z \in H_R$), $P_z(A_z) = P_z(A_z^\pi)$ ($z \in H_P$), and so by Lemma 3.1 and the form of A_z^π, we have $R_z(A_z) \geq R_z(A_z^\pi)$.

Condition 2 implies almost sure uniqueness of A^π as a minimizer of risk, as long as the risk is finite: Suppose $R(A^\pi) < \infty$ and $R(A) = R(A^\pi)$. Then the set $\{z : R_z(A_z) > R_z(A_z^\pi)\}$ has ν_R-measure zero. This implies that this set also has ν_θ-measure zero for every θ by condition 2, and so $R_z(A_z) = R_z(A_z^\pi)$ with probability one under each ν_θ. But by Lemma 3.1, $R_z(A_z) = R_z(A_z^\pi)$ is only possible for a given z if $P_z(A_z \Delta A_z^\pi) = 0$. Therefore, we must have $P_z(A_z \Delta A_z^\pi) = 0$ with ν_θ-probability one for each θ. Therefore, equality of finite risks implies

$$P_\theta(A \Delta A^\pi) = \int P_z(A \Delta A^\pi)\,\nu_\theta(dZ) \tag{67}$$

$$= \int P_z(A_z \Delta A_z^\pi)\,\nu_\theta(dZ) = 0. \tag{68}$$

□

Proof of Corollary 2. Let γ be a σ-finite Radon measure on a metric space Ω with Borel sets $\mathcal{A}$, and let $Z : (\Omega, \mathcal{A}) \to (\mathcal{Z}, \mathcal{H})$ be a measurable function. Theorem 1 of Chang and Pollard [6] says that if $\mathcal{H}$ is countably generated and contains the singleton sets, and the image measure ν of γ under $\mathcal{Z}$ is σ-finite, then γ has a (Z, ν) disintegration. Corollary 2 is a corollary of this theorem if, under the assumptions of the corollary,

1. R is a σ-finite Radon measure on the product σ-algebra $\mathcal{A} = \mathcal{F} \otimes \mathcal{G}$, and
2. $\mathcal{H}$ is countably generated and contains the singleton sets.

Item 2 holds under the assumption that $\mathcal{Z}$ is a separable metric space and $\mathcal{H}$ the Borel σ-algebra. Regarding item 1, since P_π^X and μ are both σ-finite, so is their product measure R. Also, since $\mathcal{X}$ is complete and separable then any probability measure on $(\mathcal{X}, \mathcal{F})$, including P_π^X, is a Radon measure [3, Theorem 1.4]. By assumption, μ is also a σ-finite Radon measure. Now let $\mathcal{B}$ be the Borel sets of $\mathcal{X} \times \mathcal{Y}$ under the product topology, and recall that $\mathcal{A} \subset \mathcal{B}$ in general. By Lang [16, Theorem 9.6.3], assuming $\mathcal{X}$ and $\mathcal{Y}$ are σ-compact, there exists a unique Radon measure $\tilde{R}$ on $\mathcal{B}$ that matches R on $\mathcal{A}$. But since $\mathcal{X}$ and $\mathcal{Y}$ are separable spaces, $\mathcal{A} = \mathcal{B}$ [4, Theorem 6.4.2]. Therefore, $R = \tilde{R}$ and so R is a Radon measure. □

Proof of Lemma 3.3 . Define $N_\delta = \{z : h(z) < -\delta\}$ for $\delta > 0$. We will first show that $\nu_\theta(N_\delta) = 0$ for all θ by contradiction. Suppose N_δ is non-null. Then by the assumption, for any $\epsilon > 0$ there exists a θ_ϵ

such that $\nu_{\theta_\epsilon}(N_\delta) > 1 - \epsilon$. This implies

$$\int h(z)\,\nu_{\theta_\epsilon}(dz) = \int_{N_\delta} h(z)\,\nu_{\theta_\epsilon}(dz) + \int_{N_\delta^c} h(z)\,\nu_{\theta_\epsilon}(dz) \tag{69}$$

$$< -\delta(1-\epsilon) + \|h\|_\infty \epsilon \tag{70}$$

which is less than zero for $\epsilon < \delta/(\delta + \|h\|_\infty)$. Therefore, if $\int h(z)\,\nu_\theta(dz) \geq 0$ for all θ, then $\nu_\theta(N_\delta) = 0$ for all $\delta > 0$ and θ. Now note that $\{z : h(z) < 0\} = \cup_{k=1}^\infty N_{1/k}$, and so

$$\nu_\theta(\{z : h(z) < 0\}) = \nu_\theta(\cup_{k=1}^\infty N_{1/k}) = \lim_{k\to\infty} \nu_\theta(N_{1/k}) = 0 \tag{71}$$

because $N_{1/k} \subset N_{1/(k+1)}$ and $\nu_\theta(N_{1/k}) = 0$ for all positive integers k. □

Proof of Theorem 3.4. Let $A \in \mathcal{A}$ satisfy $P_\theta(A) \geq P_\theta(A^\pi)$ for all $\theta \in \Theta$. Then $\int (P_z(A) - P_z(A^\pi)) \times \nu_\theta(dz)$ is non-negative for all θ, and so by Lemma 3.3, $P_z(A) \geq P_z(A^\pi)$ almost surely ν_θ for each θ. By Lemma 3.1 we then also have $R_z(A) \geq R_z(A^\pi)$ almost surely ν_θ for each θ, which implies $R_z(A) \geq R_z(A^\pi)$ almost everywhere ν_R by Assumption 1 of Theorem 3.2. Finally, this gives

$$R(A) - R(A^\pi) = \int (R_z(A) - R_z(A^\pi))\,\nu_R(dz) \geq 0. \tag{72}$$

□

Proof of Theorem 3.5. The density of P_θ with respect to R is $h_\theta(Z(x,y))g(x,y)/p_\pi^X(x)$, and so

$$\begin{aligned}
P_\theta(A) &= \int_A h_\theta(Z(x,y))[g(x,y)/p_\pi^X(x)]\,R(dx\,dy) \\
&= \int \left(\int_A h_\theta(Z(x,y))[g(x,y)/p_\pi^X(x)]\,R_z(dx\,dy)\right) \nu_R(dz) \\
&= \int h_\theta(z)\left(\int_A [g(x,y)/p_\pi^X(x)]\,R_z(dx\,dy)\right) \nu_R(dz) \equiv \int h_\theta(z) Q_z(A)\,\nu_R(dz),
\end{aligned}$$

where we have defined for each z the measure Q_z on $(\mathcal{X} \times \mathcal{Y}, \mathcal{F} \otimes \mathcal{G})$ such that $[dQ_z/dR_z](x,y) = g(x,y)/p_\pi^X(x)$. By Fubini's theorem, $Q_z(\mathcal{X} \times \mathcal{Y})$ is finite for ν_R-almost all z since $P_\theta(\mathcal{X} \times \mathcal{Y})$ is finite. Additionally, since R_z concentrates on $Z^{-1}\{z\}$ for ν_R-almost all z and $Q_z \ll R_z$ for each z, we have that Q_z concentrates on $Z^{-1}\{z\}$ for ν_R-almost all z as well.

The measures $\{Q_z : z \in \mathcal{Z}\}$ can be related to a conditional distribution $\{P_z : z \in \mathcal{Z}\}$ for the family $\{P_\theta : \theta \in \Theta\}$, which then gives an expression for dP_z/dR_z. First, note that $\{h_\theta(z) Q_z : z \in \mathcal{Z}\}$ gives a $\{Z, \nu_R\}$-disintegration of P_θ. For $H \in \mathcal{H}$,

$$\nu_\theta(H) \equiv P_\theta(Z^{-1}H) = \int h_\theta(z) Q_z(Z^{-1}H)\,\nu_R(dz) \tag{73}$$

$$= \int h_\theta(z)\left(\int 1(Z(x,y) \in H)\,Q_z(dx\,dy)\right) \nu_R(dz) \tag{74}$$

$$= \int_H h_\theta(z) \int Q_z(dx\,dy)\,\nu_R(dz) \tag{75}$$

$$= \int_H h_\theta(z) Q_z(\mathcal{X} \times \mathcal{Y})\,\nu_R(dz), \tag{76}$$

and so $[d\nu_\theta/d\nu_R](z) = h_\theta(z)Q_z(\mathcal{X}\times\mathcal{Y})$. Therefore, continuing from (73) we have

$$\begin{aligned} P_\theta(A) &= \int h_\theta(z)Q_z(A)\,\nu_R(dz) \\ &= \int [Q_z(A)/Q_z(\mathcal{X}\times\mathcal{Y})]h_\theta(z)Q_z(\mathcal{X}\times\mathcal{Y})\,\nu_R(dz) \\ &= \int [Q_z(A)/Q_z(\mathcal{X}\times\mathcal{Y})]\,\nu_\theta(dz) \equiv \int P_z(A)\,\nu_\theta(dz) \end{aligned} \tag{77}$$

so that P_z is a probability measure on $(\mathcal{X}\times\mathcal{Y},\mathcal{A})$. We now show that $\{P_z : z\in\mathcal{Z}\}$ is a regular conditional probability distribution of P_θ given Z for each θ. First, $P_z(Z^{-1}\{z\}) = 1$ for ν_R-almost all z and so also for ν_θ-almost all z for each θ, since $P_\theta \ll R$ for each θ. Additionally, the map $z \to P_z(A)$ is measurable for each $A\in\mathcal{A}$. To see this, recall that $z\to R_z(A)$ is measurable by assumption, and so $z\to\int f(x,y)\,R_z(dxdy)$ is measurable for any simple function f. Letting $f_k(x,y)\uparrow g(x,y)/p^X_\pi(x)$ as $k\to\infty$ we have $Q_z(A) = \lim_{k\to\infty}\int f_k(x,y)\,R_z(dxdy)$, so $z\mapsto Q_z(A)$ is a limit of measurable functions and so is measurable. Thus, $z\mapsto P_z(A) = Q_z(A)/Q_z(\mathcal{X}\times\mathcal{Y})$ is measurable as well. Finally,

$$dP_z/dR_z = [Q_z(\mathcal{X}\times\mathcal{Y})]^{-1}dQ_z/dR_z = [Q_z(\mathcal{X}\times\mathcal{Y})]^{-1}g(x,y)/p^X_\pi(x). \tag{78}$$

□

Proof of Corollary 3. The corollary results from plugging in the form of dP_z/dR_z from Theorem 3.5 into the expression for the optimal set in Theorem 3.2. □

Proof of Theorem 4.1. By Corollary 2, item 3 will be true if items 1 and 2 are true. Regarding item 1, the risk measure $R = P^X_\pi\times\mu$ is a σ-finite Radon measure by the same arguments as in the proof of Corollary 2. Regarding item 2, first we show that the sufficient statistic Z is a measurable map from $(\mathcal{Y}^{n+1},\mathcal{G}^{n+1})$ to $(\mathcal{Z},\mathcal{H})$, where $\mathcal{Z}$ is the subset of $\mathcal{P}$ consisting of measures with $n+1$ or fewer support points and $\mathcal{H}$ is the σ-algebra generated by the weak topology on $\mathcal{Z}$. Because $\mathcal{Y}$ is separable, $\mathcal{Z}$ with this topology can be separably metrized [19, Theorem 6.2]. In this case, $\mathcal{H}$ is the same as the σ-algebra generated by sets of the form $\{P\in\mathcal{P} : P(G) < c\}$ for $c\in[0,1]$, $G\in\mathcal{G}$ [15]. Let H be such a set. Then $Z^{-1}H = \{(y_1,\dots,y_{n+1})\in\mathcal{Y}^{n+1} : \sum 1(y_i\in G)/(n+1) < c\}$, which is a measurable subset of $\mathcal{G}^{n+1}$.

Let ν_R be the image measure of R under Z, defined by $\nu_R(H) = R(Z^{-1}H)$ for $H\in\mathcal{H}$. For any $G\in\mathcal{G}$, define $H_G = \{P\in\mathcal{Z} : P(G) = 1\}$. Then $Z^{-1}H_G = \{(y_1,\dots,y_{n+1})\in\mathcal{Y}^{n+1} : \{y_1,\dots,y_{n+1}\}\subset G\} = G^n$. Therefore, $\nu_R(H_G) = P^X_\pi(G^n)\times\mu(G)$, which is less than infinity if $\mu(G) < \infty$. Now by σ-finiteness of μ, there exists a countable set $\{G_k : k\in\mathbb{N}\}\subset\mathcal{G}$ such that $\cup_1^\infty G_k = \mathcal{Y}$, $G_k\subset G_{k+1}$ and $\mu(G_k) < \infty$ for each $k\in\mathbb{N}$. Therefore, $\nu_R(H_{G_k}) < \infty$ for each $k\in\mathbb{N}$ as well. We now show $\cup_k H_{G_k} = \mathcal{Z}$: Because $\cup_k G_k = \mathcal{Y}$, for each $y\in\mathcal{Y}$ there is some k_y such that $y\in G_l$ for all $l\geq k_y$. Let $z\in\mathcal{Z}$, and let $y_1,\dots,y_m$ be the support points of z (m could be $n+1$ or smaller). Then $z\in H_{G^*_k}$ where $k^* = \max\{k_{y_1},\dots,k_{y_m}\}$. Thus $\cup_k H_{G_k} = \mathcal{Z}$, and so ν_R is σ-finite. The assumptions of Corollary 2 are met and so R has a (Z,ν_R)-disintegration.

Complete sufficiency of Z under assumption A1 follows from Bell, Blackwell and Breiman [2]. The remaining conditions for A^π to be risk-optimal for its coverage function are conditions 1 and 2 of Theorem 3.2 and condition 1 of Theorem 3.5. Assumption A2 immediately implies the last two of these. The remaining condition is that if $P^{n+1}(Z^{-1}H) = 0$ for each $P\in\mathcal{P}$, we must have $\nu_R(H) = 0$. Assumption A2 implies something stronger, that for $E\in\mathcal{G}^{n+1}$, $P^{n+1}(E) = 0$ for all $P\in\mathcal{P}$ implies $R(E) = 0$, or conversely, that $R(E) > 0$ implies there exists a $P\in\mathcal{P}$ such that $P^{n+1}(E) > 0$. To see this, recall that σ-finiteness of μ implies the existence of a strictly positive function f on $\mathcal{Y}$ for which $\int f(y)\mu(dy) = 1$. Let P be the measure on $(\mathcal{Y},\mathcal{G})$ with density f with respect to μ, and P^{n+1} its $n+1$-fold product measure. Then $P^{n+1}(E) > 0$ if $\mu^{n+1}(E) > 0$ or equivalently, if $R(E) > 0$. □

References

[1] Aitchison, J. and Dunsmore, I.R. (1975). *Statistical Prediction Analysis*. Cambridge-Melbourne: Cambridge Univ. Press. MR0408097

[2] Bell, C.B., Blackwell, D. and Breiman, L. (1960). On the completeness of order statistics. *Ann. Math. Statist.* **31** 794–797. MR0116427 https://doi.org/10.1214/aoms/1177705808

[3] Billingsley, P. (1999). *Convergence of Probability Measures*, 2nd ed. *Wiley Series in Probability and Statistics: Probability and Statistics*. New York: Wiley. A Wiley-Interscience Publication. MR1700749 https://doi.org/10.1002/9780470316962

[4] Bogachev, V.I. (2007). *Measure Theory. Vol. I, II*. Berlin: Springer. MR2267655 https://doi.org/10.1007/978-3-540-34514-5

[5] Burris, K. and Hoff, P.D. (2019). Exact adaptive confidence intervals for small areas. *Journal of Survey Statistics and Methodology* **8** 206–230.

[6] Chang, J.T. and Pollard, D. (1997). Conditioning as disintegration. *Statist. Neerlandica* **51** 287–317. MR1484954 https://doi.org/10.1111/1467-9574.00056

[7] Dunsmore, I.R. (1976). A note on Faulkenberry's method of obtaining prediction intervals. *J. Amer. Statist. Assoc.* **71** 193–194.

[8] Evans, M. and Fraser, D.A.S. (1980). An optimum tolerance region for multivariate regression. *J. Multivariate Anal.* **10** 268–272. MR0575929 https://doi.org/10.1016/0047-259X(80)90018-4

[9] Faulkenberry, G.D. (1973). A method of obtaining prediction intervals. *J. Amer. Statist. Assoc.* **68** 433–435.

[10] Good, I.J. (1992). The Bayes/non-Bayes compromise: A brief review. *J. Amer. Statist. Assoc.* **87** 597–606. MR1185188

[11] Halmos, P.R. (1950). *Measure Theory*. New York, N. Y.: D. Van Nostrand Co., Inc. MR0033869

[12] Halmos, P.R. and Savage, L.J. (1949). Application of the Radon-Nikodym theorem to the theory of sufficient statistics. *Ann. Math. Statistics* **20** 225–241. MR0030730 https://doi.org/10.1214/aoms/1177730032

[13] Hoff, P. and Yu, C. (2019). Exact adaptive confidence intervals for linear regression coefficients. *Electron. J. Stat.* **13** 94–119. MR3896147 https://doi.org/10.1214/18-ejs1517

[14] Jin, C., Zhu, J., Steen-Adams, M.M., Sain, S.R. and Gangnon, R.E. (2013). Spatial multinomial regression models for nominal categorical data: A study of land cover in Northern Wisconsin, USA. *Environmetrics* **24** 98–108. MR3067334 https://doi.org/10.1002/env.2189

[15] Karr, A.F. (1991). *Point Processes and Their Statistical Inference*, 2nd ed. *Probability: Pure and Applied* **7**. New York: Dekker. MR1113698

[16] Lang, S. (1993). *Real and Functional Analysis*, 3rd ed. *Graduate Texts in Mathematics* **142**. New York: Springer. MR1216137 https://doi.org/10.1007/978-1-4612-0897-6

[17] Lawless, J.F. and Fredette, M. (2005). Frequentist prediction intervals and predictive distributions. *Biometrika* **92** 529–542. MR2202644 https://doi.org/10.1093/biomet/92.3.529

[18] Mattner, L. (1996). Complete order statistics in parametric models. *Ann. Statist.* **24** 1265–1282. MR1401849 https://doi.org/10.1214/aos/1032526968

[19] Parthasarathy, K.R. (2005). *Probability Measures on Metric Spaces*. Providence, RI: AMS Chelsea Publishing. Reprint of the 1967 original. MR2169627 https://doi.org/10.1090/chel/352

[20] Patel, J.K. (1989). Prediction intervals—a review. *Comm. Statist. Theory Methods* **18** 2393–2465. MR1029127 https://doi.org/10.1080/03610928908830043

[21] Pratt, J.W. (1963). Shorter confidence intervals for the mean of a normal distribution with known variance. *Ann. Math. Statist.* **34** 574–586. MR0148150 https://doi.org/10.1214/aoms/1177704170

[22] Rockafellar, R.T. and Wets, R.J.-B. (1998). *Variational Analysis*. *Grundlehren der Mathematischen Wissenschaften [Fundamental Principles of Mathematical Sciences]* **317**. Berlin: Springer. MR1491362 https://doi.org/10.1007/978-3-642-02431-3

[23] Shafer, G. and Vovk, V. (2008). A tutorial on conformal prediction. *J. Mach. Learn. Res.* **9** 371–421. MR2417240

[24] Yu, C. and Hoff, P.D. (2018). Adaptive multigroup confidence intervals with constant coverage. *Biometrika* **105** 319–335. MR3804405 https://doi.org/10.1093/biomet/asy009

Received August 2021 and revised February 2022