FAB Prediction

Peter Hoff
Duke University

Prediction

Statistical prediction

- input into decision-making procedures;
- quantification of uncertainties about unobserved observables;
- the foundation of all meaningful inferences.

Prediction regions

- Bayes posterior: $\Pr(Y \in A(x)|X = x) = 1 - \alpha \ \forall x$.
- Frequentist: $\Pr(Y \in A(X)|\theta) = 1 - \alpha \ \forall \theta$.
- FAB: Find a set-valued function $A(x)$ that minimizes

$$R(A, \pi) = \int |A(x)| \, p(x|\theta)\pi(\theta) \, d\theta \quad \text{subject to}$$

$$\Pr(Y \in A(X)|\theta) = 1 - \alpha \ \forall \theta.$$

$\pi(\theta)$ represents prior or *indirect* information.

Prediction

Statistical prediction

- input into decision-making procedures;
- quantification of uncertainties about unobserved observables;
- the foundation of all meaningful inferences.

Prediction regions

- Bayes posterior: $\Pr(Y \in A(x)|X = x) = 1 - \alpha \ \forall x$.
- Frequentist: $\Pr(Y \in A(X)|\theta) = 1 - \alpha \ \forall \theta$.
- FAB: Find a set-valued function $A(x)$ that minimizes

$$R(A, \pi) = \int |A(x)| \, p(x|\theta)\pi(\theta) \, d\theta \quad \text{subject to}$$

$$\Pr(Y \in A(X)|\theta) = 1 - \alpha \ \forall \theta.$$

$\pi(\theta)$ represents prior or *indirect* information.

Prediction

Statistical prediction

- input into decision-making procedures;
- quantification of uncertainties about unobserved observables;
- the foundation of all meaningful inferences.

Prediction regions

- Bayes posterior: $\Pr(Y \in A(x)|X = x) = 1 - \alpha \ \forall x$.
- Frequentist: $\Pr(Y \in A(X)|\theta) = 1 - \alpha \ \forall \theta$.
- FAB: Find a set-valued function $A(x)$ that minimizes

$$R(A, \pi) = \int |A(x)| \, p(x|\theta)\pi(\theta) \, d\theta \quad \text{subject to}$$

$$\Pr(Y \in A(X)|\theta) = 1 - \alpha \ \forall \theta.$$

$\pi(\theta)$ represents prior or *indirect* information.

Prediction

Statistical prediction

- input into decision-making procedures;
- quantification of uncertainties about unobserved observables;
- the foundation of all meaningful inferences.

Prediction regions

- Bayes posterior: $\Pr(Y \in A(x)|X = x) = 1 - \alpha \ \forall x$.
- Frequentist: $\Pr(Y \in A(X)|\theta) = 1 - \alpha \ \forall \theta$.
- FAB: Find a set-valued function $A(x)$ that minimizes

$$R(A, \pi) = \int |A(x)| \, p(x|\theta)\pi(\theta) \, d\theta \quad \text{subject to}$$

$$\Pr(Y \in A(X)|\theta) = 1 - \alpha \ \forall \theta.$$

$\pi(\theta)$ represents prior or *indirect* information.

Prediction

Statistical prediction

- input into decision-making procedures;
- quantification of uncertainties about unobserved observables;
- the foundation of all meaningful inferences.

Prediction regions

- Bayes posterior: $\Pr(Y \in A(x)|X = x) = 1 - \alpha \ \forall x$.
- Frequentist: $\Pr(Y \in A(X)|\theta) = 1 - \alpha \ \forall \theta$.
- FAB: Find a set-valued function $A(x)$ that minimizes

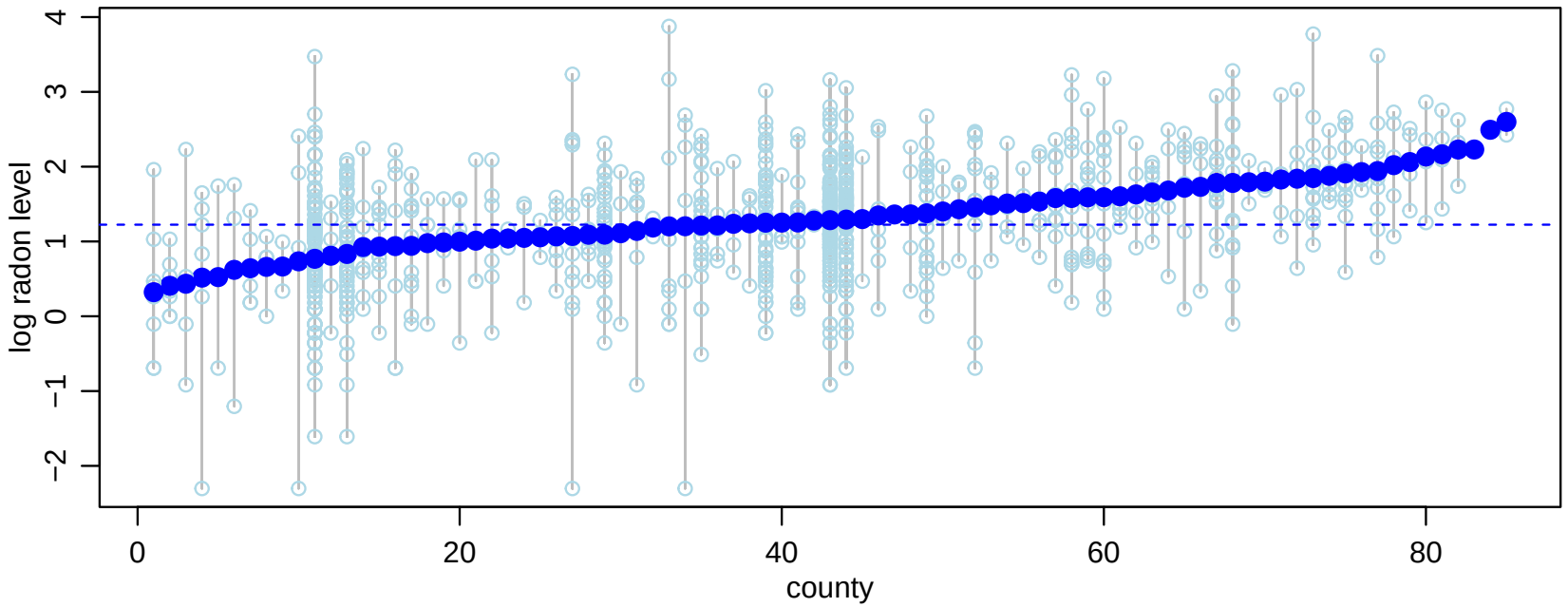
$$R(A, \pi) = \int |A(x)| \, p(x|\theta)\pi(\theta) \, d\theta \quad \text{subject to}$$

$$\Pr(Y \in A(X)|\theta) = 1 - \alpha \ \forall \theta.$$

$\pi(\theta)$ represents prior or *indirect* information.

Radon data

Radon data



Prediction region procedures

Goal: Prediction region for Y_j^* based on $\{y_{i,j} : i = 1, \ldots, n_j, j = 1, \ldots, p\}$.

Direct methods: t-pivot interval, nonparametric conformal prediction interval.

$$\mathbf{Y}_j \to \theta_j \to Y_j^*$$

Maintain frequentist coverage control,
don't use indirect information.

Indirect methods: posterior predictive distributions, random effects models.

$$\begin{array}{rcl} & \mathbf{Y}_j \to \theta_j \to Y_j^* \\ \mathbf{Y}_{-j} \to \boldsymbol{\theta}_{-j} & \nearrow & \end{array}$$

Lack frequentist coverage control,
do use indirect information.

FAB methods:
Maintain frequentist coverage control,
do use indirect information.

Prediction region procedures

Goal: Prediction region for Y_j^* based on $\{y_{i,j} : i = 1, \ldots, n_j, j = 1, \ldots, p\}$.

Direct methods: t-pivot interval, nonparametric conformal prediction interval.

$$\mathbf{Y}_j \to \theta_j \to Y_j^*$$

Maintain frequentist coverage control,
don't use indirect information.

Indirect methods: posterior predictive distributions, random effects models.

$$\begin{array}{rcl} & \mathbf{Y}_j \to \theta_j \to Y_j^* \\ \mathbf{Y}_{-j} \to \boldsymbol{\theta}_{-j} & \nearrow & \end{array}$$

Lack frequentist coverage control,
do use indirect information.

FAB methods:
Maintain frequentist coverage control,
do use indirect information.

Prediction region procedures

Goal: Prediction region for Y_j^* based on $\{y_{i,j} : i = 1, \ldots, n_j, j = 1, \ldots, p\}$.

Direct methods: t-pivot interval, nonparametric conformal prediction interval.

$$\mathbf{Y}_j \to \theta_j \to Y_j^*$$

Maintain frequentist coverage control,
don't use indirect information.

Indirect methods: posterior predictive distributions, random effects models.

$$\begin{array}{rl} & \mathbf{Y}_j \to \theta_j \to Y_j^* \\ \mathbf{Y}_{-j} \to & \boldsymbol{\theta}_{-j} \nearrow \end{array}$$

Lack frequentist coverage control,
do use indirect information.

FAB methods:
Maintain frequentist coverage control,
do use indirect information.

Prediction region procedures

Goal: Prediction region for Y_j^* based on $\{y_{i,j} : i = 1, \ldots, n_j, j = 1, \ldots, p\}$.

Direct methods: t-pivot interval, nonparametric conformal prediction interval.

$$\mathbf{Y}_j \to \theta_j \to Y_j^*$$

Maintain frequentist coverage control,
don't use indirect information.

Indirect methods: posterior predictive distributions, random effects models.

$$\begin{array}{rl} & \mathbf{Y}_j \to \theta_j \to Y_j^* \\ \mathbf{Y}_{-j} \to & \boldsymbol{\theta}_{-j} \nearrow \end{array}$$

Lack frequentist coverage control,
do use indirect information.

FAB methods:
Maintain frequentist coverage control,
do use indirect information.

Prediction procedures, sections and graphs

Model:

- Random variables X and Y taking values in $\mathcal{X}$ and $\mathcal{Y}$.
- $\Pr((X, Y) \in A) = P(A)$ for some $P \in \{P_\theta : \theta \in \Theta\}$.

Objective: Predict values of Y based on observing $X = x$.

Prediction procedure: A set-valued function $\mathcal{X} \to 2^{\mathcal{Y}}$, written $x \mapsto A_x \subset \mathcal{Y}$.

- The *graph* of $\{A_x : x \in \mathcal{X}\}$ is $A = \{(x, y) : y \in A_x\} \subset \mathcal{X} \times \mathcal{Y}$.
- The *x-section* of $A \subset \mathcal{X} \times \mathcal{Y}$ is $A_x = \{y : (x, y) \in A\} \subset \mathcal{Y}$.

There is a bijection between functions $\mathcal{X} \to 2^{\mathcal{Y}}$ and $2^{\mathcal{X} \times \mathcal{Y}}$.

$$y \in A_x \Leftrightarrow (x, y) \in A$$

Coverage rate: If $A \subset \mathcal{X} \times \mathcal{Y}$ satisfies

$$P_\theta(A) \equiv P_\theta(\{(x, y) : y \in A_x\}) = 1 - \alpha \ \forall \theta \in \Theta$$

then it is a $1 - \alpha$ constant coverage prediction region.

Prediction procedures, sections and graphs

Model:

- Random variables X and Y taking values in $\mathcal{X}$ and $\mathcal{Y}$.
- $\Pr((X, Y) \in A) = P(A)$ for some $P \in \{P_\theta : \theta \in \Theta\}$.

Objective: Predict values of Y based on observing $X = x$.

Prediction procedure: A set-valued function $\mathcal{X} \to 2^{\mathcal{Y}}$, written $x \mapsto A_x \subset \mathcal{Y}$.

- The *graph* of $\{A_x : x \in \mathcal{X}\}$ is $A = \{(x, y) : y \in A_x\} \subset \mathcal{X} \times \mathcal{Y}$.
- The *x-section* of $A \subset \mathcal{X} \times \mathcal{Y}$ is $A_x = \{y : (x, y) \in A\} \subset \mathcal{Y}$.

There is a bijection between functions $\mathcal{X} \to 2^{\mathcal{Y}}$ and $2^{\mathcal{X} \times \mathcal{Y}}$.

$$y \in A_x \Leftrightarrow (x, y) \in A$$

Coverage rate: If $A \subset \mathcal{X} \times \mathcal{Y}$ satisfies

$$P_\theta(A) \equiv P_\theta(\{(x, y) : y \in A_x\}) = 1 - \alpha \; \forall \theta \in \Theta$$

then it is a $1 - \alpha$ constant coverage prediction region.

Prediction procedures, sections and graphs

Model:

- Random variables X and Y taking values in $\mathcal{X}$ and $\mathcal{Y}$.
- $\Pr((X, Y) \in A) = P(A)$ for some $P \in \{P_\theta : \theta \in \Theta\}$.

Objective: Predict values of Y based on observing $X = x$.

Prediction procedure: A set-valued function $\mathcal{X} \to 2^{\mathcal{Y}}$, written $x \mapsto A_x \subset \mathcal{Y}$.

- The *graph* of $\{A_x : x \in \mathcal{X}\}$ is $A = \{(x, y) : y \in A_x\} \subset \mathcal{X} \times \mathcal{Y}$.
- The *x-section* of $A \subset \mathcal{X} \times \mathcal{Y}$ is $A_x = \{y : (x, y) \in A\} \subset \mathcal{Y}$.

There is a bijection between functions $\mathcal{X} \to 2^{\mathcal{Y}}$ and $2^{\mathcal{X} \times \mathcal{Y}}$.

$$y \in A_x \Leftrightarrow (x, y) \in A$$

Coverage rate: If $A \subset \mathcal{X} \times \mathcal{Y}$ satisfies

$$P_\theta(A) \equiv P_\theta(\{(x, y) : y \in A_x\}) = 1 - \alpha \; \forall \theta \in \Theta$$

then it is a $1 - \alpha$ constant coverage prediction region.

Prediction procedures, sections and graphs

Model:

- Random variables X and Y taking values in $\mathcal{X}$ and $\mathcal{Y}$.
- $\Pr((X, Y) \in A) = P(A)$ for some $P \in \{P_\theta : \theta \in \Theta\}$.

Objective: Predict values of Y based on observing $X = x$.

Prediction procedure: A set-valued function $\mathcal{X} \to 2^{\mathcal{Y}}$, written $x \mapsto A_x \subset \mathcal{Y}$.

- The *graph* of $\{A_x : x \in \mathcal{X}\}$ is $A = \{(x, y) : y \in A_x\} \subset \mathcal{X} \times \mathcal{Y}$.
- The *x-section* of $A \subset \mathcal{X} \times \mathcal{Y}$ is $A_x = \{y : (x, y) \in A\} \subset \mathcal{Y}$.

There is a bijection between functions $\mathcal{X} \to 2^{\mathcal{Y}}$ and $2^{\mathcal{X} \times \mathcal{Y}}$.

$$y \in A_x \Leftrightarrow (x, y) \in A$$

Coverage rate: If $A \subset \mathcal{X} \times \mathcal{Y}$ satisfies

$$P_\theta(A) \equiv P_\theta(\{(x, y) : y \in A_x\}) = 1 - \alpha \ \forall \theta \in \Theta$$

then it is a $1 - \alpha$ constant coverage prediction region.

Prediction procedures, sections and graphs

Model:

- Random variables X and Y taking values in $\mathcal{X}$ and $\mathcal{Y}$.
- $\Pr((X, Y) \in A) = P(A)$ for some $P \in \{P_\theta : \theta \in \Theta\}$.

Objective: Predict values of Y based on observing $X = x$.

Prediction procedure: A set-valued function $\mathcal{X} \to 2^{\mathcal{Y}}$, written $x \mapsto A_x \subset \mathcal{Y}$.

- The *graph* of $\{A_x : x \in \mathcal{X}\}$ is $A = \{(x, y) : y \in A_x\} \subset \mathcal{X} \times \mathcal{Y}$.
- The *x-section* of $A \subset \mathcal{X} \times \mathcal{Y}$ is $A_x = \{y : (x, y) \in A\} \subset \mathcal{Y}$.

There is a bijection between functions $\mathcal{X} \to 2^{\mathcal{Y}}$ and $2^{\mathcal{X} \times \mathcal{Y}}$.

$$y \in A_x \Leftrightarrow (x, y) \in A$$

Coverage rate: If $A \subset \mathcal{X} \times \mathcal{Y}$ satisfies

$$P_\theta(A) \equiv P_\theta(\{(x, y) : y \in A_x\}) = 1 - \alpha \ \forall \theta \in \Theta$$

then it is a $1 - \alpha$ constant coverage prediction region.

Prediction procedures, sections and graphs

Model:

- Random variables X and Y taking values in $\mathcal{X}$ and $\mathcal{Y}$.
- $\Pr((X, Y) \in A) = P(A)$ for some $P \in \{P_\theta : \theta \in \Theta\}$.

Objective: Predict values of Y based on observing $X = x$.

Prediction procedure: A set-valued function $\mathcal{X} \to 2^{\mathcal{Y}}$, written $x \mapsto A_x \subset \mathcal{Y}$.

- The *graph* of $\{A_x : x \in \mathcal{X}\}$ is $A = \{(x, y) : y \in A_x\} \subset \mathcal{X} \times \mathcal{Y}$.
- The *x-section* of $A \subset \mathcal{X} \times \mathcal{Y}$ is $A_x = \{y : (x, y) \in A\} \subset \mathcal{Y}$.

There is a bijection between functions $\mathcal{X} \to 2^{\mathcal{Y}}$ and $2^{\mathcal{X} \times \mathcal{Y}}$.

$$y \in A_x \Leftrightarrow (x, y) \in A$$

Coverage rate: If $A \subset \mathcal{X} \times \mathcal{Y}$ satisfies

$$P_\theta(A) \equiv P_\theta(\{(x, y) : y \in A_x\}) = 1 - \alpha \ \forall \theta \in \Theta$$

then it is a $1 - \alpha$ constant coverage prediction region.

Prediction procedures, sections and graphs

Model:

- Random variables X and Y taking values in $\mathcal{X}$ and $\mathcal{Y}$.
- $\Pr((X, Y) \in A) = P(A)$ for some $P \in \{P_\theta : \theta \in \Theta\}$.

Objective: Predict values of Y based on observing $X = x$.

Prediction procedure: A set-valued function $\mathcal{X} \to 2^{\mathcal{Y}}$, written $x \mapsto A_x \subset \mathcal{Y}$.

- The *graph* of $\{A_x : x \in \mathcal{X}\}$ is $A = \{(x, y) : y \in A_x\} \subset \mathcal{X} \times \mathcal{Y}$.
- The *x-section* of $A \subset \mathcal{X} \times \mathcal{Y}$ is $A_x = \{y : (x, y) \in A\} \subset \mathcal{Y}$.

There is a bijection between functions $\mathcal{X} \to 2^{\mathcal{Y}}$ and $2^{\mathcal{X} \times \mathcal{Y}}$.

$$y \in A_x \Leftrightarrow (x, y) \in A$$

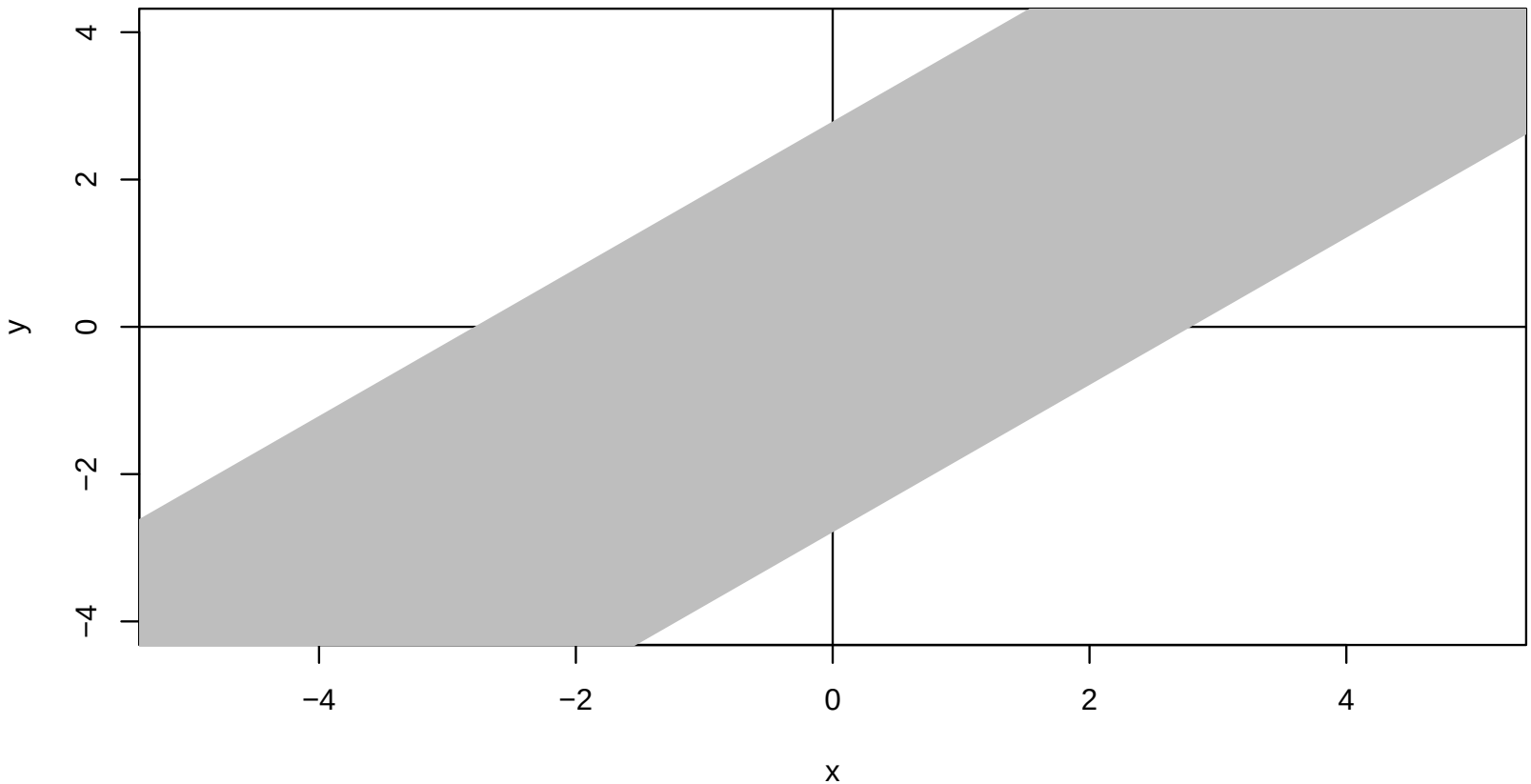
Coverage rate: If $A \subset \mathcal{X} \times \mathcal{Y}$ satisfies

$$P_\theta(A) \equiv P_\theta(\{(x, y) : y \in A_x\}) = 1 - \alpha \ \forall \theta \in \Theta$$

then it is a $1 - \alpha$ constant coverage prediction region.

Simplest example

$Y, X \sim$ i.i.d. $N(\theta, 1),\ \theta \in \mathbb{R}$.



Faulkenberry's method

Objective: Find a set $A \subset \mathcal{X} \times \mathcal{Y}$ such that $P_\theta(A) = 1 - \alpha \; \forall \theta \in \Theta$.

Faulkenberry (1973): Let $Z(x, y)$ be a sufficient statistic for $\{P_\theta : \theta \in \Theta\}$.

1. For each z, let C_z be the acceptance region of a size-α test of $H : Y \sim P_z^Y$.

$$P_z^Y(C_z) = 1 - \alpha \; \forall z. \qquad \text{(doesn't depend on } \theta\text{!)}$$

2. Let $A = \{(x, y) : y \in C_{Z(x,y)}\}$.

Coverage:

$$\begin{aligned} P_\theta(A) &= \Pr(Y \in C_{Z(X,Y)} | \theta) \\ &= \int P_z^Y(C_z) \, \nu_\theta(dz) \\ &= \int (1 - \alpha) \, \nu_\theta(dz) = 1 - \alpha. \end{aligned}$$

Summary: $y \in A_x$ if $\{Y = y\}$ does not reject $H : Y \sim P_{Z(x,y)}^Y$.

Faulkenberry's method

Objective: Find a set $A \subset \mathcal{X} \times \mathcal{Y}$ such that $P_\theta(A) = 1 - \alpha \; \forall \theta \in \Theta$.

Faulkenberry (1973): Let $Z(x, y)$ be a sufficient statistic for $\{P_\theta : \theta \in \Theta\}$.

1. For each z, let C_z be the acceptance region of a size-α test of $H : Y \sim P_z^Y$.

$$P_z^Y(C_z) = 1 - \alpha \; \forall z. \qquad \text{(doesn't depend on } \theta\text{!)}$$

2. Let $A = \{(x, y) : y \in C_{Z(x,y)}\}$.

Coverage:

$$\begin{aligned} P_\theta(A) &= \Pr(Y \in C_{Z(X,Y)} | \theta) \\ &= \int P_z^Y(C_z) \, \nu_\theta(dz) \\ &= \int (1 - \alpha) \, \nu_\theta(dz) = 1 - \alpha. \end{aligned}$$

Summary: $y \in A_x$ if $\{Y = y\}$ does not reject $H : Y \sim P_{Z(x,y)}^Y$.

Faulkenberry's method

Objective: Find a set $A \subset \mathcal{X} \times \mathcal{Y}$ such that $P_\theta(A) = 1 - \alpha \; \forall \theta \in \Theta$.

Faulkenberry (1973): Let $Z(x, y)$ be a sufficient statistic for $\{P_\theta : \theta \in \Theta\}$.

1. For each z, let C_z be the acceptance region of a size-α test of $H : Y \sim P_z^Y$.

$$P_z^Y(C_z) = 1 - \alpha \; \forall z. \qquad \text{(doesn't depend on } \theta\text{!)}$$

2. Let $A = \{(x, y) : y \in C_{Z(x,y)}\}$.

Coverage:

$$\begin{aligned} P_\theta(A) &= \Pr(Y \in C_{Z(X,Y)} | \theta) \\ &= \int P_z^Y(C_z) \, \nu_\theta(dz) \\ &= \int (1 - \alpha) \, \nu_\theta(dz) = 1 - \alpha. \end{aligned}$$

Summary: $y \in A_x$ if $\{Y = y\}$ does not reject $H : Y \sim P_{Z(x,y)}^Y$.

Faulkenberry's method

Objective: Find a set $A \subset \mathcal{X} \times \mathcal{Y}$ such that $P_\theta(A) = 1 - \alpha \ \forall \theta \in \Theta$.

Faulkenberry (1973): Let $Z(x, y)$ be a sufficient statistic for $\{P_\theta : \theta \in \Theta\}$.

1. For each z, let C_z be the acceptance region of a size-α test of $H : Y \sim P_z^Y$.

$$P_z^Y(C_z) = 1 - \alpha \ \forall z. \qquad \text{(doesn't depend on } \theta\text{!)}$$

2. Let $A = \{(x, y) : y \in C_{Z(x,y)}\}$.

Coverage:

$$\begin{aligned} P_\theta(A) &= \Pr(Y \in C_{Z(X,Y)} | \theta) \\ &= \int P_z^Y(C_z)\, \nu_\theta(dz) \\ &= \int (1 - \alpha)\, \nu_\theta(dz) = 1 - \alpha. \end{aligned}$$

Summary: $y \in A_x$ if $\{Y = y\}$ does not reject $H : Y \sim P_{Z(x,y)}^Y$.

Nonparametric prediction

Model: $(Y_1, \ldots, Y_n, Y_{n+1}) \sim P \in \mathcal{P} =$ exchangeable distributions on $\mathcal{Y}^{n+1}$.

Goal: Predict Y_{n+1} from observing $(Y_1, \ldots, Y_n) = (y_1, \ldots, y_n)$.

Faulkenberry's method:

- $Z = \{Y_1, \ldots, Y_n, Y_{n+1}\}$ is a sufficient statistic.
- Distribution of Y_{n+1} given $Z = \{y_1, \ldots, y_n, y_{n+1}\}$ is $U(\{y_1, \ldots, y_n, y_{n+1}\})$.

Prediction region:

$y_{n+1} \in A_{y_1,\ldots,y_n}$ if $\{Y_{n+1} = y_{n+1}\}$ does not reject $Y_{n+1} \sim U(\{y_1, \ldots, y_n, y_{n+1}\})$.

Nonparametric prediction

Model: $(Y_1, \ldots, Y_n, Y_{n+1}) \sim P \in \mathcal{P} =$ exchangeable distributions on $\mathcal{Y}^{n+1}$.

Goal: Predict Y_{n+1} from observing $(Y_1, \ldots, Y_n) = (y_1, \ldots, y_n)$.

Faulkenberry's method:

- $Z = \{Y_1, \ldots, Y_n, Y_{n+1}\}$ is a sufficient statistic.
- Distribution of Y_{n+1} given $Z = \{y_1, \ldots, y_n, y_{n+1}\}$ is $U(\{y_1, \ldots, y_n, y_{n+1}\})$.

Prediction region:
$y_{n+1} \in A_{y_1,\ldots,y_n}$ if $\{Y_{n+1} = y_{n+1}\}$ does not reject $Y_{n+1} \sim U(\{y_1, \ldots, y_n, y_{n+1}\})$.

Nonparametric prediction

Model: $(Y_1, \ldots, Y_n, Y_{n+1}) \sim P \in \mathcal{P} =$ exchangeable distributions on $\mathcal{Y}^{n+1}$.

Goal: Predict Y_{n+1} from observing $(Y_1, \ldots, Y_n) = (y_1, \ldots, y_n)$.

Faulkenberry's method:

- $Z = \{Y_1, \ldots, Y_n, Y_{n+1}\}$ is a sufficient statistic.
- Distribution of Y_{n+1} given $Z = \{y_1, \ldots, y_n, y_{n+1}\}$ is $U(\{y_1, \ldots, y_n, y_{n+1}\})$.

Prediction region:
$y_{n+1} \in A_{y_1,\ldots,y_n}$ if $\{Y_{n+1} = y_{n+1}\}$ does not reject $Y_{n+1} \sim U(\{y_1, \ldots, y_n, y_{n+1}\})$.

Nonparametric prediction

Test implementation: Let $t_{n+1} = t(y_{n+1} : y_1, \ldots, y_n)$ be a test statistic.

$$
\begin{aligned}
t_1 &= t(y_1 : y_{n+1}, y_2, y_3, \ldots, y_{n-1}, y_n) \\
t_2 &= t(y_2 : y_1, y_{n+1}, y_3, \ldots, y_{n-1}, y_n) \\
&\vdots \\
t_n &= t(y_n : y_1, y_2, y_3, \ldots, y_{n-1}, y_{n+1})
\end{aligned}
$$

- Distribution of t conditional on $Z = \{y_1, \ldots, y_n, y_{n+1}\}$ is $U(\{t_1, \ldots, t_{n+1}\})$.
- Include a value y_{n+1} in the prediction region if $t_{n+1} \leq q_{1-\alpha}(t_1, \ldots, t_{n+1})$.

This is referred to as *conformal prediction* (Gammerman, Vovk, Vapnik [1998]).

Nonparametric prediction

Test implementation: Let $t_{n+1} = t(y_{n+1} : y_1, \ldots, y_n)$ be a test statistic.

$$
\begin{aligned}
t_1 &= t(y_1 : y_{n+1}, y_2, y_3, \ldots, y_{n-1}, y_n) \\
t_2 &= t(y_2 : y_1, y_{n+1}, y_3, \ldots, y_{n-1}, y_n) \\
&\vdots \\
t_n &= t(y_n : y_1, y_2, y_3, \ldots, y_{n-1}, y_{n+1})
\end{aligned}
$$

- Distribution of t conditional on $Z = \{y_1, \ldots, y_n, y_{n+1}\}$ is $U(\{t_1, \ldots, t_{n+1}\})$.
- Include a value y_{n+1} in the prediction region if $t_{n+1} \leq q_{1-\alpha}(t_1, \ldots, t_{n+1})$.

This is referred to as *conformal prediction* (Gammerman, Vovk, Vapnik [1998]).

Bayes-optimal frequentist prediction

Risk and Bayes risk

- Let μ be a volume measure on $\mathcal{Y}$.
- Let π be a prior distribution over Θ.

$$R_\theta(A) = \int \mu(A_x)\, P_\theta(dx)$$
$$= \int\int 1((x,y) \in A)\, P_\theta^X(dx)\mu(dy).$$
$$R_\pi(A) = \int\int 1((x,y) \in A)\, P_\pi^X(dx)\mu(dy).$$

Bayes-optimal region with $1-\alpha$ frequentist coverage:

minimize $R_\pi(A)$ (Bayes risk)

subject to $P_\theta(A) = 1-\alpha\ \forall\theta \in \Theta$ (frequentist coverage constraint).

The constrained minimizer A_π is frequentist and Bayesian, or FAB.

Bayes-optimal frequentist prediction

Risk and Bayes risk

- Let μ be a volume measure on $\mathcal{Y}$.
- Let π be a prior distribution over Θ.

$$R_\theta(A) = \int \mu(A_x)\, P_\theta(dx)$$
$$= \int\int 1((x,y) \in A)\, P_\theta^X(dx)\mu(dy).$$
$$R_\pi(A) = \int\int 1((x,y) \in A)\, P_\pi^X(dx)\mu(dy).$$

Bayes-optimal region with $1-\alpha$ frequentist coverage:

minimize $R_\pi(A)$ (Bayes risk)

subject to $P_\theta(A) = 1-\alpha\ \forall\theta \in \Theta$ (frequentist coverage constraint).

The constrained minimizer A_π is frequentist and Bayesian, or FAB.

Bayes-optimal frequentist prediction

Risk and Bayes risk

- Let μ be a volume measure on $\mathcal{Y}$.
- Let π be a prior distribution over Θ.

$$R_\theta(A) = \int \mu(A_x)\, P_\theta(dx)$$
$$= \int \int 1((x,y) \in A)\, P_\theta^X(dx)\mu(dy).$$
$$R_\pi(A) = \int \int 1((x,y) \in A)\, P_\pi^X(dx)\mu(dy).$$

Bayes-optimal region with $1-\alpha$ frequentist coverage:

minimize $R_\pi(A)$ (Bayes risk)

subject to $P_\theta(A) = 1-\alpha\ \forall\theta \in \Theta$ (frequentist coverage constraint).

The constrained minimizer A_π is frequentist and Bayesian, or FAB.

Disintegration

Constant conditional coverage: If Z is a *complete* sufficient statistic then

$$P_\theta(A) = \int P_z(A)\, \nu_\theta(dz) = 1 - \alpha \quad \forall \theta \quad \Leftrightarrow$$
$$P_z(A) = 1 - \alpha \quad \forall z.$$

Risk disintegration: Even though R_π may not be a probability measure,

$$R_\pi(A) = \int R_z(A)\, \nu_\pi(dz).$$

- $\{R_z : z \in \mathcal{Z}\}$ is a *disintegration* of R_π.
- Each R_z is a probability measure, although ν_π is not.

Disintegration

Constant conditional coverage: If Z is a *complete* sufficient statistic then

$$P_\theta(A) = \int P_z(A)\, \nu_\theta(dz) = 1 - \alpha \quad \forall \theta \quad \Leftrightarrow$$
$$P_z(A) = 1 - \alpha \quad \forall z.$$

Risk disintegration: Even though R_π may not be a probability measure,

$$R_\pi(A) = \int R_z(A)\, \nu_\pi(dz).$$

- $\{R_z : z \in \mathcal{Z}\}$ is a *disintegration* of R_π.
- Each R_z is a probability measure, although ν_π is not.

Optimization via conditional optimization

Implication:

$$\text{minimizer of } R_\pi \text{ among } A : P_\theta(A) = 1 - \alpha \ \forall \theta$$
$$\text{obtained by minimizing } R_z \text{ among } A_z : P_z(A_z) = 1 - \alpha \ \forall z.$$

Result: Bayes-optimal prediction region with $1 - \alpha$ frequentist coverage is

$$A_\pi = \left\{ (x, y) : \frac{dP_{Z(x,y)}}{dR_{Z(x,y)}}(x, y) > k_{Z(x,y)} \right\}$$

Summary:

- Prediction region are inversions of conditional tests of $(X, Y) \sim P_z$.
- The Bayes-optimal tests use the test statistic dP_z/dR_z.

Optimization via conditional optimization

Implication:

$$\text{minimizer of } R_\pi \text{ among } A : P_\theta(A) = 1 - \alpha \ \forall \theta$$
$$\text{obtained by minimizing } R_z \text{ among } A_z : P_z(A_z) = 1 - \alpha \ \forall z.$$

Result: Bayes-optimal prediction region with $1 - \alpha$ frequentist coverage is

$$A_\pi = \left\{ (x, y) : \frac{dP_{Z(x,y)}}{dR_{Z(x,y)}}(x, y) > k_{Z(x,y)} \right\}$$

Summary:

- Prediction region are inversions of conditional tests of $(X, Y) \sim P_z$.
- The Bayes-optimal tests use the test statistic dP_z/dR_z.

Optimization via conditional optimization

Implication:

$$\text{minimizer of } R_\pi \text{ among } A : P_\theta(A) = 1 - \alpha \ \forall \theta$$
$$\text{obtained by minimizing } R_z \text{ among } A_z : P_z(A_z) = 1 - \alpha \ \forall z.$$

Result: Bayes-optimal prediction region with $1 - \alpha$ frequentist coverage is

$$A_\pi = \left\{ (x, y) : \frac{dP_{Z(x,y)}}{dR_{Z(x,y)}}(x, y) > k_{Z(x,y)} \right\}$$

Summary:

- Prediction region are inversions of conditional tests of $(X, Y) \sim P_z$.
- The Bayes-optimal tests use the test statistic dP_z/dR_z.

Example: Multivariate normal prediction

Task: Predict $Y \sim N_p(\theta, \Sigma)$ from $X \sim N_p(\theta, k\Sigma)$, with $\theta \sim N_p(\mu, \lambda\Sigma)$.

$$A_x{}^E = \{y : \|\Sigma^{-1/2}(x-y)/\sqrt{k+1}\|^2 < \chi^2_{p,0,1-\alpha}\}$$
$$A_x{}^\pi = \{y : \|\Sigma^{-1/2}(x-y)/\sqrt{k+1} + \delta_{Z(x,y)}\|^2 < \chi^2_{p,\|\delta_{Z(x,y)}\|^2,1-\alpha}\},$$

where $\delta_z = \Sigma^{-1/2}(\mu - z)v^{1/2}/(v_\lambda - v)$.

Reexpression:

$$A_x{}^B = \{y : \|\Sigma^{-1/2}(y-\hat{\theta}^\pi)/v_\lambda^{1/2}\|^2 < \chi^2_{p,0,1-\alpha}\}$$
$$A_x{}^\pi = \{y : \|\Sigma^{-1/2}(y-\hat{\theta}^\pi)/v_\lambda^{1/2}\|^2 \times (k+1)/v_\lambda < \chi^2_{p,\|\delta_{Z(x,y)}\|^2,1-\alpha}\},$$

where $\hat{\theta}^\pi = \mathrm{E}[\theta|X=x]$.

Example: Multivariate normal prediction

Task: Predict $Y \sim N_p(\theta, \Sigma)$ from $X \sim N_p(\theta, k\Sigma)$, with $\theta \sim N_p(\mu, \lambda\Sigma)$.

$$A_x{}^E = \{y : \|\Sigma^{-1/2}(x-y)/\sqrt{k+1}\|^2 < \chi^2_{p,0,1-\alpha}\}$$
$$A_x{}^\pi = \{y : \|\Sigma^{-1/2}(x-y)/\sqrt{k+1} + \delta_{Z(x,y)}\|^2 < \chi^2_{p,\|\delta_{Z(x,y)}\|^2,1-\alpha}\},$$

where $\delta_z = \Sigma^{-1/2}(\mu - z)v^{1/2}/(v_\lambda - v)$.

Reexpression:

$$A_x{}^B = \{y : \|\Sigma^{-1/2}(y-\hat{\theta}^\pi)/v_\lambda^{1/2}\|^2 < \chi^2_{p,0,1-\alpha}\}$$
$$A_x{}^\pi = \{y : \|\Sigma^{-1/2}(y-\hat{\theta}^\pi)/v_\lambda^{1/2}\|^2 \times (k+1)/v_\lambda < \chi^2_{p,\|\delta_{Z(x,y)}\|^2,1-\alpha}\},$$

where $\hat{\theta}^\pi = \mathrm{E}[\theta|X = x]$.

Example: Normal linear regression

Task: Predict $Y \sim N(v^\top \beta, \sigma^2)$ from $X \sim N_n(U\beta, \sigma^2 I)$, $\beta \sim N_p(0, \sigma^2 \Psi^{-1})$.

$$A_x^E = \{y : |(y - \hat{\beta}^\top v)/(\sigma\sqrt{w_0})| \leq \Phi^{-1}(1 - \alpha/2)\}$$
$$A_x^\pi = \{y : |(y - \hat{\beta}^\top v)/(\sigma\sqrt{w_0}) + \delta_{Z(x,y)}| \leq q_{Z(x,y)}\},$$

where δ depends on v, U, Ψ, σ.

Reexpression

$$A_x{}^B = \{y : \hat{\beta}_\psi^\top v - \Phi^{-1}(1 - \alpha/2)\sigma\sqrt{w_\psi} < y < \hat{\beta}_\psi^\top v + \Phi^{-1}(1 - \alpha/2)\sigma\sqrt{w_\psi}\}$$
$$A_x^\pi = \{y : \hat{\beta}_\psi^\top v - q_{Z(x,y)}\sigma w_\psi/\sqrt{w_0} < y < \hat{\beta}_\psi^\top v + q_{Z(x,y)}\sigma w_\psi/\sqrt{w_0}\},$$

where $\hat{\beta}_\psi = \mathrm{E}[\beta | X = x]$.

Example: Normal linear regression

Task: Predict $Y \sim N(v^\top \beta, \sigma^2)$ from $X \sim N_n(U\beta, \sigma^2 I)$, $\beta \sim N_p(0, \sigma^2 \Psi^{-1})$.

$$A_x^E = \{y : |(y - \hat{\beta}^\top v)/(\sigma\sqrt{w_0})| \leq \Phi^{-1}(1 - \alpha/2)\}$$
$$A_x^\pi = \{y : |(y - \hat{\beta}^\top v)/(\sigma\sqrt{w_0}) + \delta_{Z(x,y)}| \leq q_{Z(x,y)}\},$$

where δ depends on v, U, Ψ, σ.

Reexpression

$$A_x{}^B = \{y : \hat{\beta}_\psi^\top v - \Phi^{-1}(1 - \alpha/2)\sigma\sqrt{w_\psi} < y < \hat{\beta}_\psi^\top v + \Phi^{-1}(1 - \alpha/2)\sigma\sqrt{w_\psi}\}$$
$$A_x^\pi = \{y : \hat{\beta}_\psi^\top v - q_{Z(x,y)}\sigma w_\psi/\sqrt{w_0} < y < \hat{\beta}_\psi^\top v + q_{Z(x,y)}\sigma w_\psi/\sqrt{w_0}\},$$

where $\hat{\beta}_\psi = \mathrm{E}[\beta | X = x]$.

Example: FAB conformal prediction

Risk-optimal conditional test statistic:

$$\frac{dP_z}{dR_z}(y_1, \ldots, y_n, y_{n+1}) > k_z \Leftrightarrow p_\pi(y_{n+1}|y_1, \ldots, y_n) > c_z$$

where

$$p_\pi(y_{n+1}|y_1, \ldots, y_n) = \int p(y_{n+1}|\theta)\, \pi(d\theta|y_1, \ldots, y_n)$$

is the *posterior predictive density* of Y_{n+1} given $(Y_1, \ldots, Y_n) = (y_1, \ldots, y_n)$.

Implication: A Bayes risk-optimal conformity score is $p_\pi(y_{n+1}|y_1, \ldots, y_n)$.

Example: FAB conformal prediction

Risk-optimal conditional test statistic:

$$\frac{dP_z}{dR_z}(y_1, \ldots, y_n, y_{n+1}) > k_z \Leftrightarrow p_\pi(y_{n+1}|y_1, \ldots, y_n) > c_z$$

where

$$p_\pi(y_{n+1}|y_1, \ldots, y_n) = \int p(y_{n+1}|\theta)\, \pi(d\theta|y_1, \ldots, y_n)$$

is the *posterior predictive density* of Y_{n+1} given $(Y_1, \ldots, Y_n) = (y_1, \ldots, y_n)$.

Implication: A Bayes risk-optimal conformity score is $p_\pi(y_{n+1}|y_1, \ldots, y_n)$.

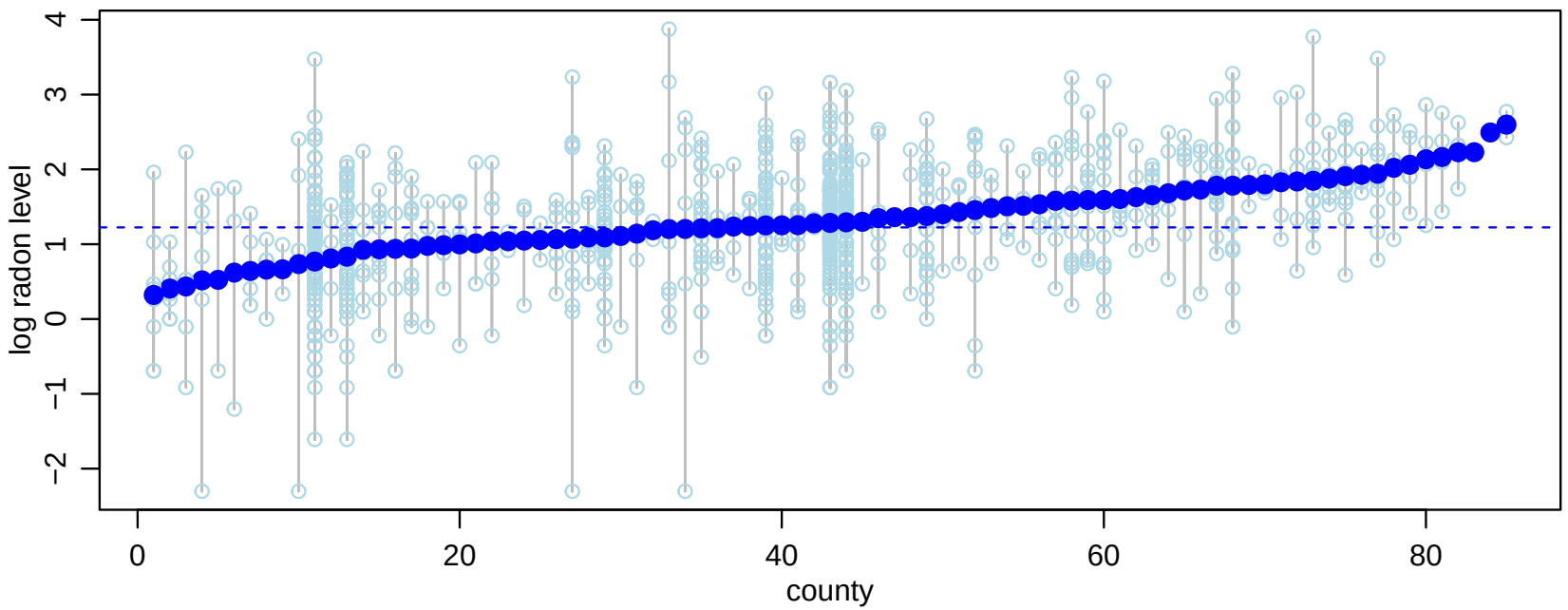
FAB conformal prediction

Test implementation: Let $p_{n+1} = p(y_{n+1}|y_1, \ldots, y_n)$.

$$
\begin{aligned}
p_1 &= p(y_1|y_{n+1}, y_2, y_3, \ldots, y_{n-1}, y_n) \\
p_2 &= p(y_2|y_1, y_{n+1}, y_3, \ldots, y_{n-1}, y_n) \\
&\vdots \\
p_n &= p(y_n|y_1, y_2, y_3, \ldots, y_{n-1}, y_{n+1})
\end{aligned}
$$

- Distribution of $p(Y_{n+1}|Y_1, \ldots, Y_n)$ conditional on $Z = \{y_1, \ldots, y_n, y_{n+1}\}$ is $U(\{p_1, \ldots, p_{n+1}\})$.
- Include a value y_{n+1} in the prediction region if $p_{n+1} \geq q_\alpha(p_1, \ldots, p_{n+1})$.

Small area FAB prediction (Bersson and Hoff 2023)



Goal: Predict Y_j^* from $\mathbf{Y}_j$, $\mathbf{Y}_{-j}$.

- direct t-intervals (parametric, exact coverage, no information sharing)
- direct conformal (nonparametric, exact coverage, no information sharing)
- Bayes HLM (parametric, inexact coverage, yes information sharing)
- FAB conformal (nonparametric, exact coverage, yes information sharing).

Conformal FAB via a working model

Spatial working model:

$$Y_{1,j}, \ldots, Y_{n_j,j} \sim \text{i.i.d. } N(\theta_j, \sigma_j^2), \; j = 1, \ldots, p$$
$$\boldsymbol{\theta} \sim N_p(\mathbf{X}\boldsymbol{\beta}, \eta^2 \mathbf{G})$$

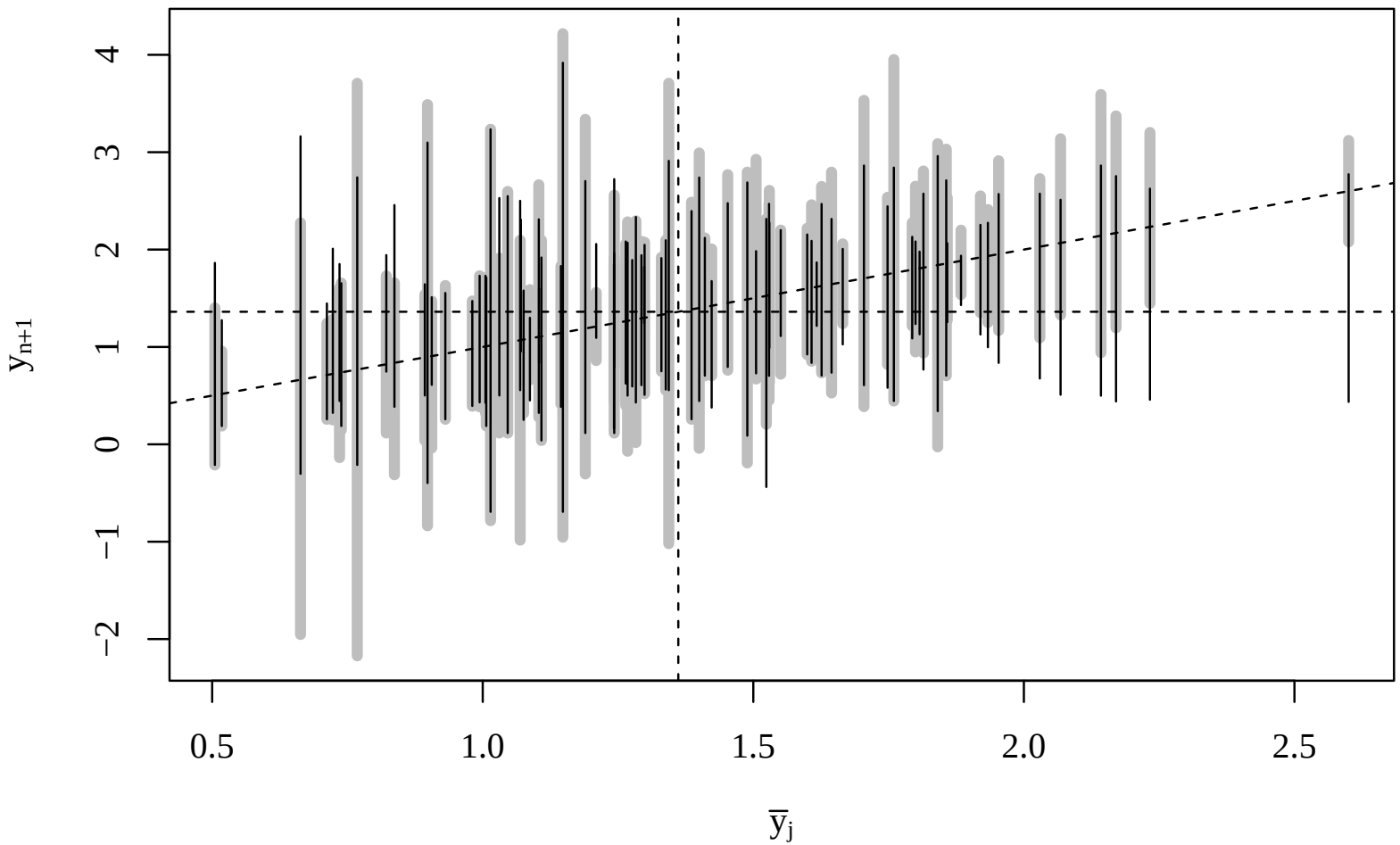
Conformal FAB: For each group $j = 1, \ldots, p$,

1. obtain estimates $(\hat{\boldsymbol{\theta}}_{-j}, \hat{\boldsymbol{\beta}}, \hat{\eta})$ using $\mathbf{Y}_{-j}$;
2. obtain an "empirical Bayes" predictive distribution for y_j^*:

$$p(y_j^* | \mathbf{Y}_j) = \int p(y_j^* | \theta_j) p(\theta_j | \mathbf{Y}_j, \hat{\boldsymbol{\theta}}_{-j}, \hat{\boldsymbol{\beta}}, \hat{\eta}) \, d\theta_j$$

3. Use $p(y_j^* | \mathbf{Y}_j)$ as a conformity score for the prediction interval.

Direct and indirect nonparametric prediction intervals



Discussion

Bayes-optimal prediction with frequentist coverage control

- permits inclusion of prior or indirect information;
- maintains frequentist coverage guarantees.

Coverage in multipopulation settings

- Direct: "Equal width" and constant coverage across groups.
- Bayes: Narrowest on average wrt prior, but coverage varies by group.
- FAB: Narrowest on average wrt prior among constant coverage procedures.

Nonparametric prediction with parametric working models

- conformal procedure guarantees coverage with any conformity statistic;
- conformity statistic can be parametric or nonparametric predictive density;
- FAB is Bayes optimal among procedures with constant coverage.

Discussion

Bayes-optimal prediction with frequentist coverage control

- permits inclusion of prior or indirect information;
- maintains frequentist coverage guarantees.

Coverage in multipopulation settings

- Direct: "Equal width" and constant coverage across groups.
- Bayes: Narrowest on average wrt prior, but coverage varies by group.
- FAB: Narrowest on average wrt prior among constant coverage procedures.

Nonparametric prediction with parametric working models

- conformal procedure guarantees coverage with any conformity statistic;
- conformity statistic can be parametric or nonparametric predictive density;
- FAB is Bayes optimal among procedures with constant coverage.

Discussion

Bayes-optimal prediction with frequentist coverage control

- permits inclusion of prior or indirect information;
- maintains frequentist coverage guarantees.

Coverage in multipopulation settings

- Direct: "Equal width" and constant coverage across groups.
- Bayes: Narrowest on average wrt prior, but coverage varies by group.
- FAB: Narrowest on average wrt prior among constant coverage procedures.

Nonparametric prediction with parametric working models

- conformal procedure guarantees coverage with any conformity statistic;
- conformity statistic can be parametric or nonparametric predictive density;
- FAB is Bayes optimal among procedures with constant coverage.