Detect model miscalibration via your nearest neighbor

Bernoulli-*ims*
Aug 14 , 2024

Ziang Niu

Collaborators

Anirban Chatterjee

Bhaswar Bikram Bhattacharya

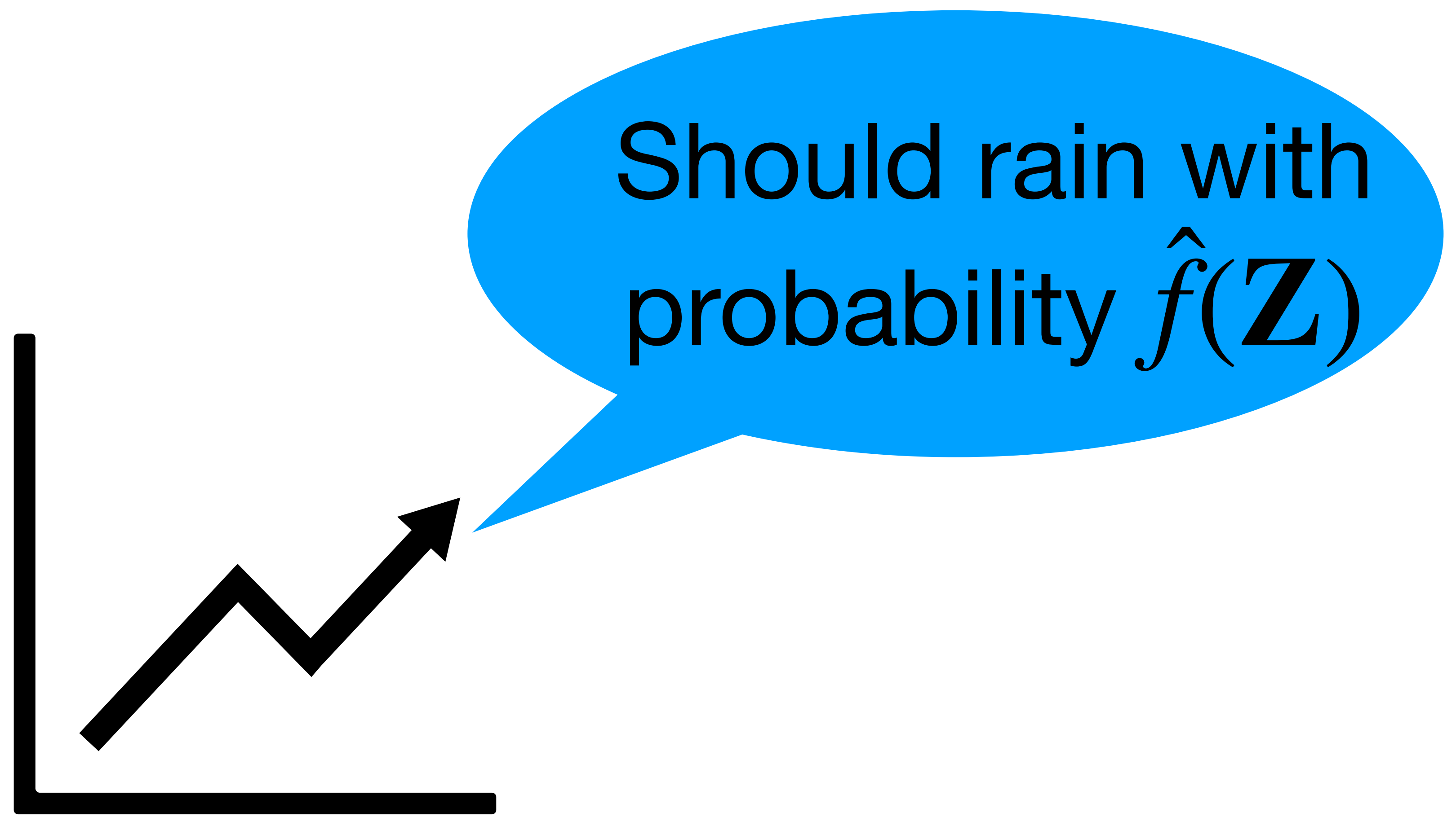
Model calibration

Model calibration

Statistical task (loose): find out whether a trained supervised model $\hat{f} : \mathbb{R}^d \mapsto \{0,1\}$ can produce “reliable” prediction.

Model calibration

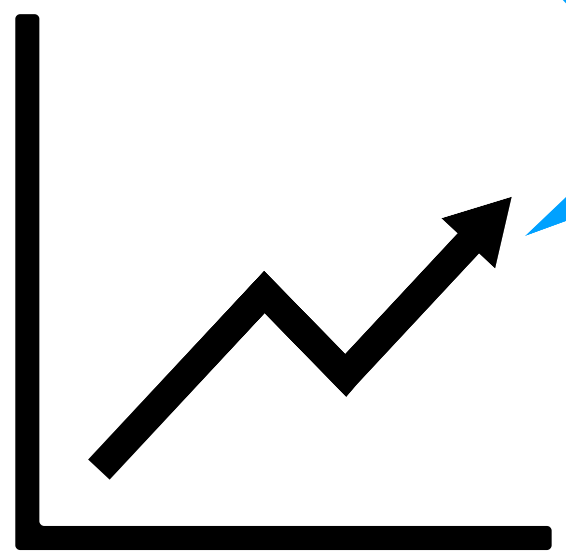
Statistical task (loose): find out whether a trained supervised model $\hat{f} : \mathbb{R}^d \mapsto \{0,1\}$ can produce "reliable" prediction.

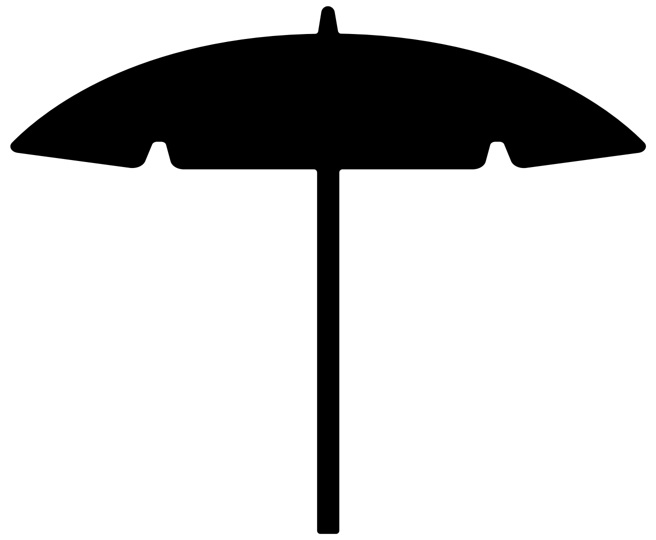


Model calibration

Statistical task (loose): find out whether a trained supervised model $\hat{f}:\mathbb{R}^d \mapsto \{0,1\}$ can produce "reliable" prediction.

Should rain with probability $\hat{f}(\mathbf{Z})$

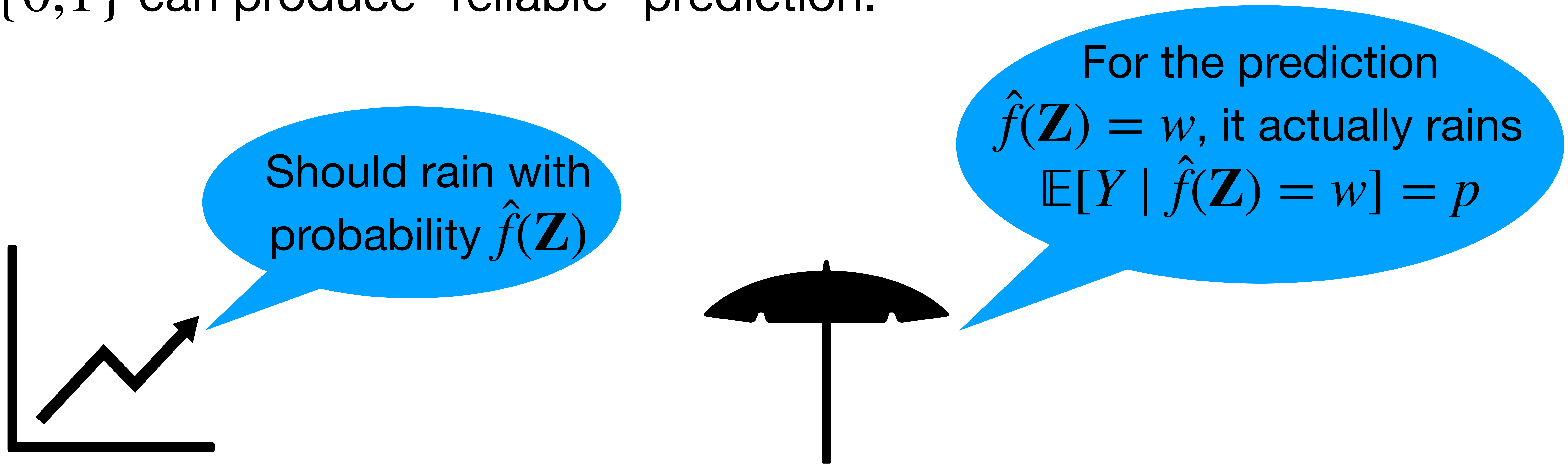




For the prediction $\hat{f}(\mathbf{Z}) = w$, it actually rains $\mathbb{E}[Y \mid \hat{f}(\mathbf{Z}) = w] = p$

Model calibration

Statistical task (loose): find out whether a trained supervised model $\hat{f} : \mathbb{R}^d \mapsto \{0,1\}$ can produce "reliable" prediction.



If $w \approx p$, it is a reliable prediction at prediction w.

Model calibration

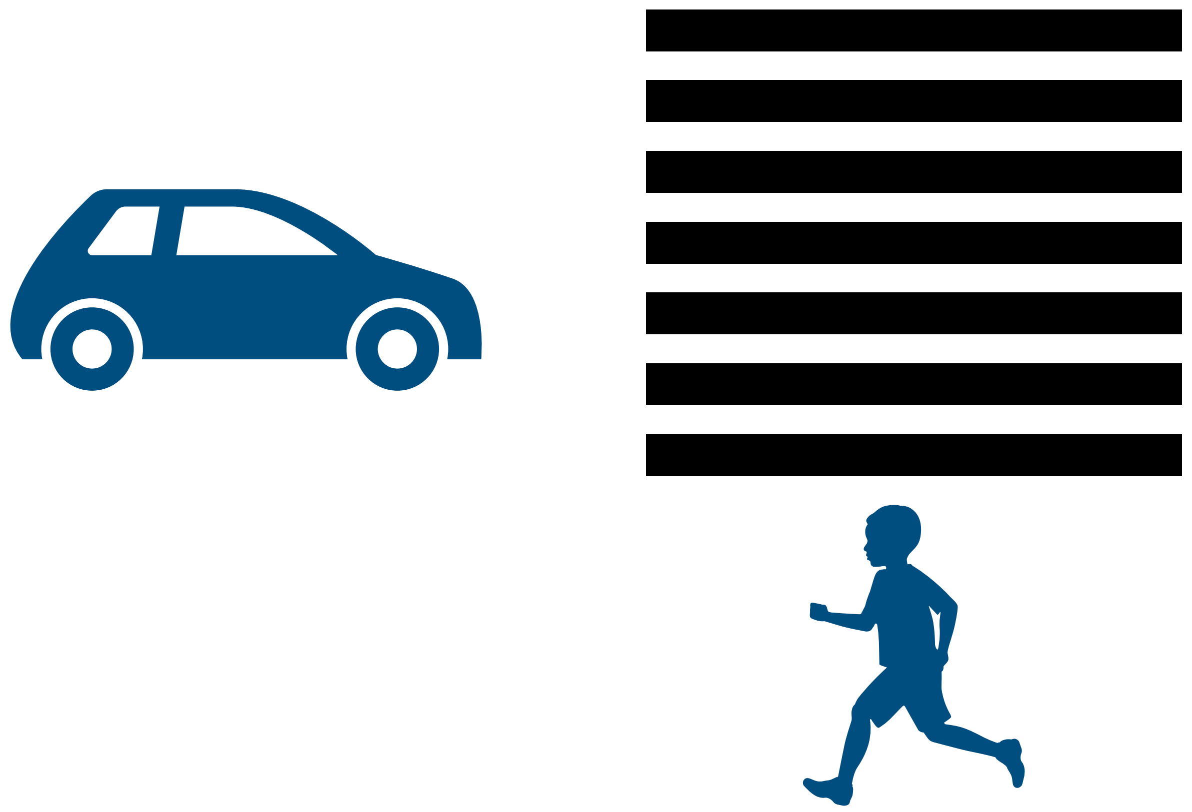
Model calibration

High-stakes application: auto-drive program

Model calibration

High-stakes application: auto-drive program

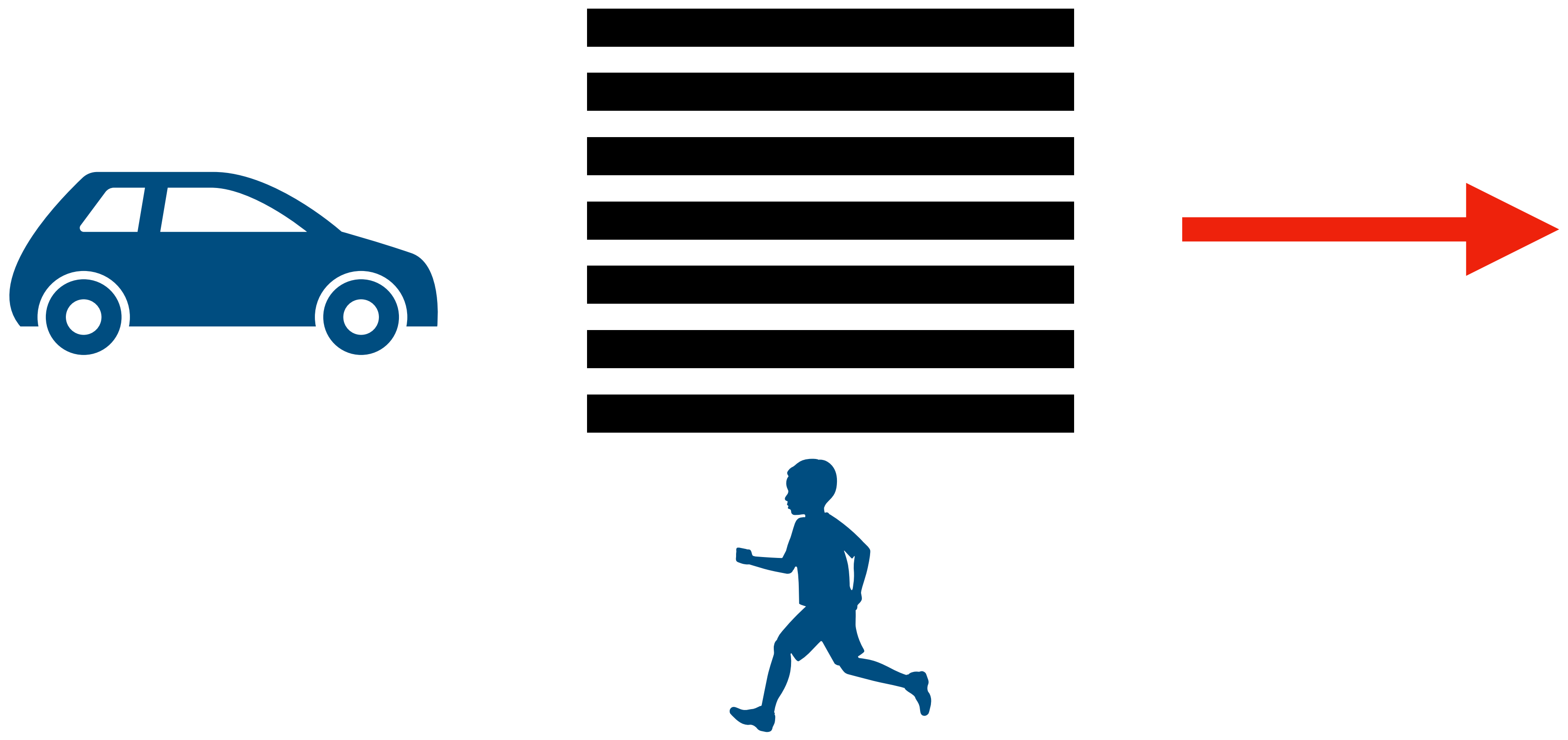
$f(\mathbf{Z})$: Car hit pedestrian



Model calibration

High-stakes application: auto-drive program

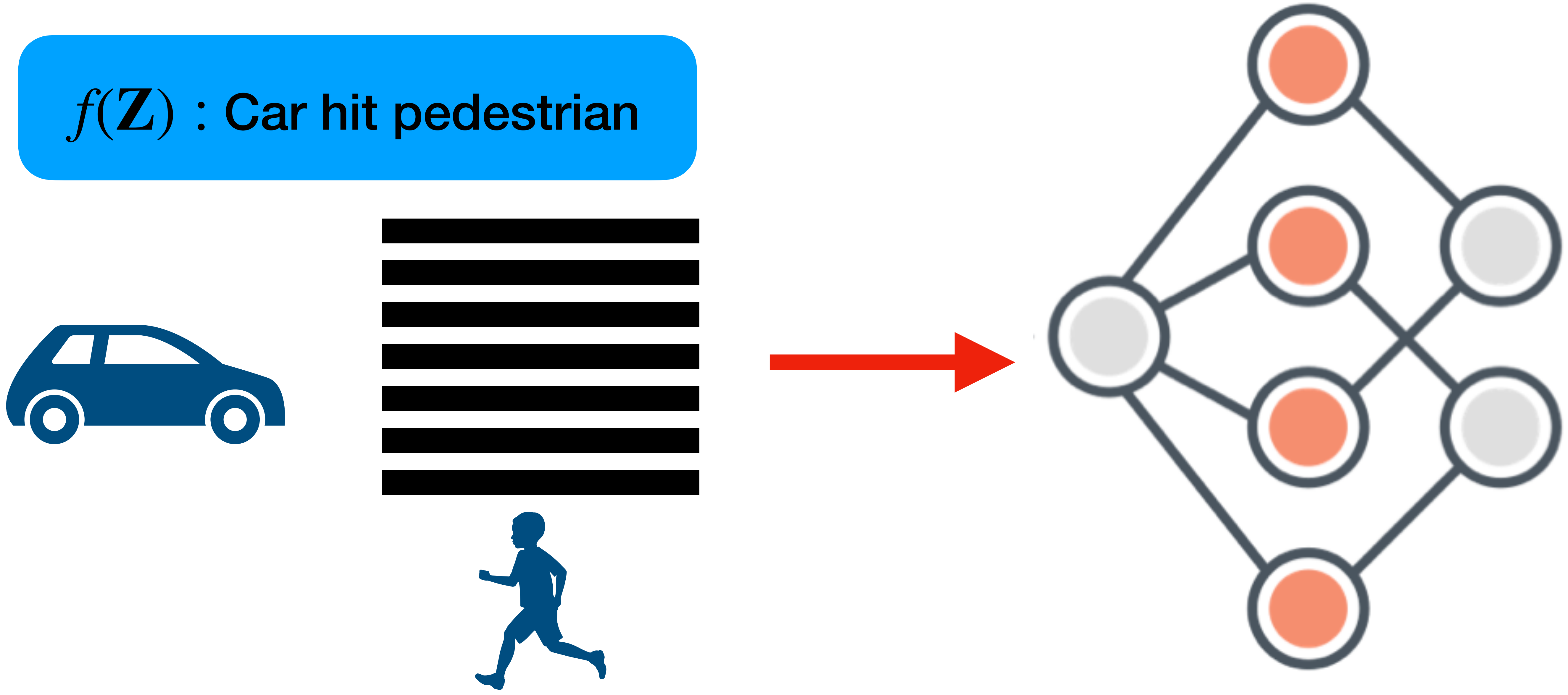
$f(\mathbf{Z})$: Car hit pedestrian



Model calibration

High-stakes application: auto-drive program

$f(\mathbf{Z})$: Car hit pedestrian

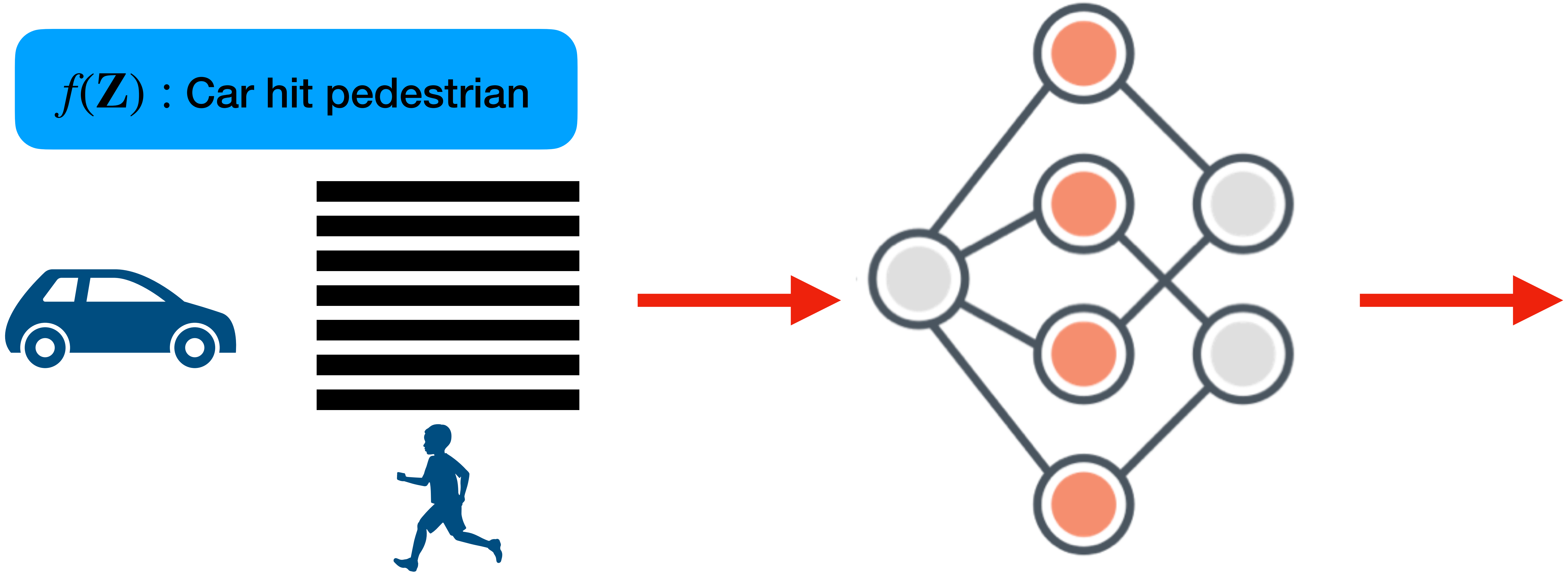


$\hat{f}(\mathbf{Z}) \approx 0.0001$

Model calibration

High-stakes application: auto-drive program

$f(\mathbf{Z})$: Car hit pedestrian

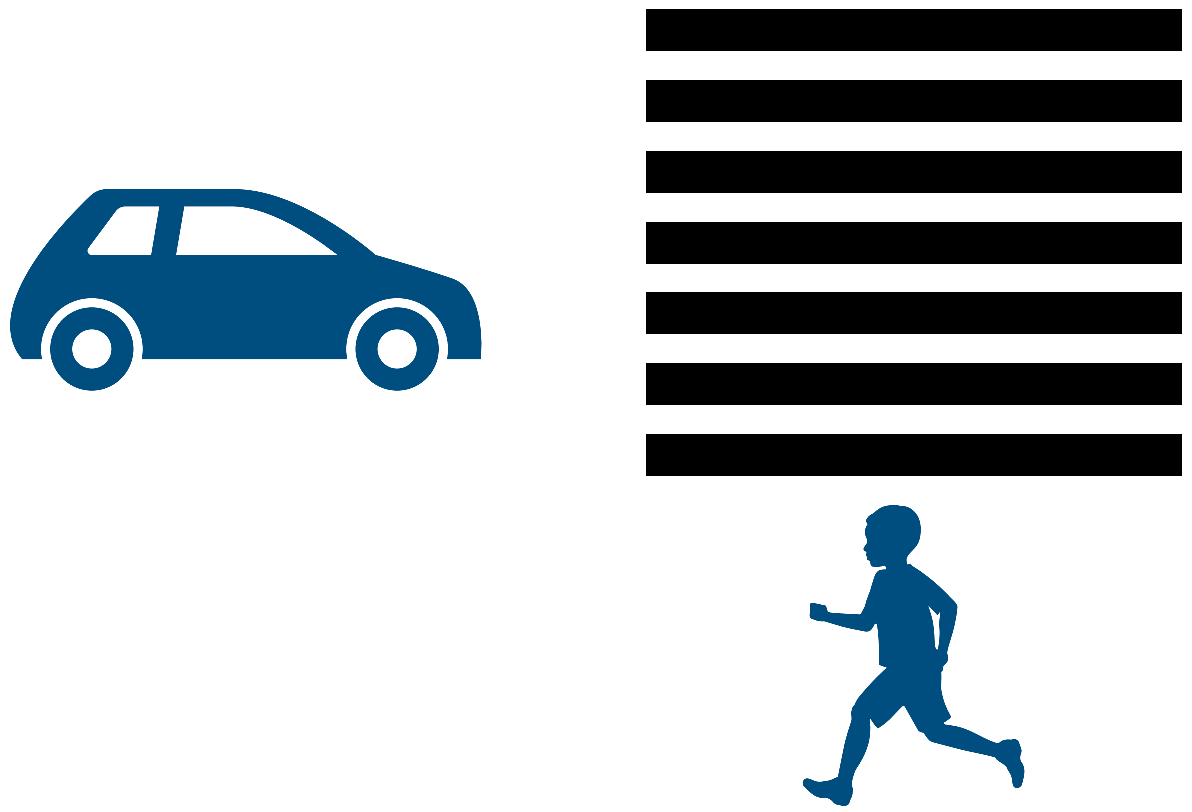


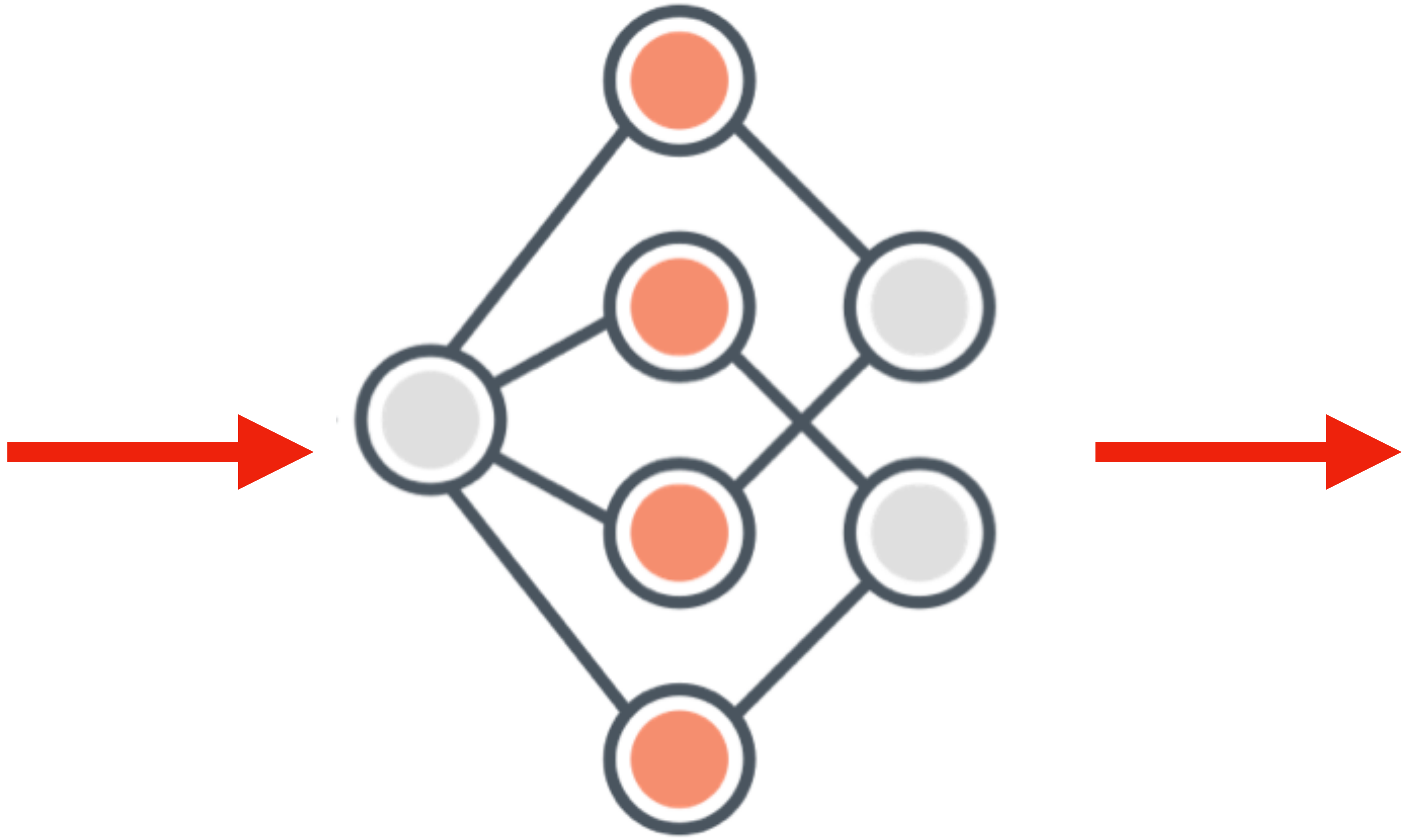
$\hat{f}(\mathbf{Z}) \approx 0.0001$

Model calibration

High-stakes application: auto-drive program

$f(\mathbf{Z})$: Car hit pedestrian





$\hat{f}(\mathbf{Z}) \approx 0.0001$

$f(\mathbf{Z})$: Car hit pedestrian

$\mathbb{E}[\mathbf{Y} \,|\, \hat{f}(\mathbf{Z}) = 0.0001] = 0.1$

Statistical formulation

Statistical formulation

Hypothesis formulation (classification): For joint distribution $\mathscr{L}_n(\mathbf{Y}, \mathrm{Bern}(\hat{f}(\mathbf{Z})))$, test the null hypothesis of calibration:

Statistical formulation

Hypothesis formulation (classification): For joint distribution $\mathscr{L}_n(\mathbf{Y}, \mathrm{Bern}(\hat{f}(\mathbf{Z})))$, test the null hypothesis of calibration:

$$H_0 : \hat{f}(\mathbf{Z}) = \mathbb{P}[\mathbf{Y} = 1 \mid \hat{f}(\mathbf{Z})] \text{ almost surely} .$$

Statistical formulation

Hypothesis formulation (classification): For joint distribution $\mathscr{L}_n(\mathbf{Y}, \mathrm{Bern}(\hat{f}(\mathbf{Z})))$, test the null hypothesis of calibration:

$$H_0 : \hat{f}(\mathbf{Z}) = \mathbb{P}[\mathbf{Y} = 1 \mid \hat{f}(\mathbf{Z})] \text{ almost surely}.$$

Equivalently,

Statistical formulation

Hypothesis formulation (classification): For joint distribution $\mathscr{L}_n(\mathbf{Y}, \mathrm{Bern}(\hat{f}(\mathbf{Z})))$, test the null hypothesis of calibration:

$$H_0 : \hat{f}(\mathbf{Z}) = \mathbb{P}[\mathbf{Y} = 1 \mid \hat{f}(\mathbf{Z})] \text{ almost surely} .$$

Equivalently,

$$H_0 : \mathbb{P}[\mathbf{X} = 1 \,|\, \hat{f}(\mathbf{Z})] = \mathbb{P}[\mathbf{Y} = 1 \mid \hat{f}(\mathbf{Z})] \text{ almost surely}, \mathbf{X} \sim \mathrm{Bern}(\hat{f}(\mathbf{Z})) .$$

Statistical formulation

Hypothesis formulation (classification): For joint distribution $\mathscr{L}_n(\mathbf{Y}, \mathrm{Bern}(\hat{f}(\mathbf{Z})))$, test the null hypothesis of calibration:

$$H_0 : \hat{f}(\mathbf{Z}) = \mathbb{P}[\mathbf{Y} = 1 \mid \hat{f}(\mathbf{Z})] \text{ almost surely}\,.$$

Equivalently,

$$H_0 : \mathbb{P}[\mathbf{X} = 1 \,|\, \hat{f}(\mathbf{Z})] = \mathbb{P}[\mathbf{Y} = 1 \mid \hat{f}(\mathbf{Z})] \text{ almost surely}, \mathbf{X} \sim \mathrm{Bern}(\hat{f}(\mathbf{Z}))\,.$$

We are interested in the hypothesis

Statistical formulation

Hypothesis formulation (classification): For joint distribution $\mathscr{L}_n(\mathbf{Y}, \mathrm{Bern}(\hat{f}(\mathbf{Z})))$, test the null hypothesis of calibration:

$$H_0 : \hat{f}(\mathbf{Z}) = \mathbb{P}[\mathbf{Y} = 1 \mid \hat{f}(\mathbf{Z})] \text{ almost surely} .$$

Equivalently,

$$H_0 : \mathbb{P}[\mathbf{X} = 1 \mid \hat{f}(\mathbf{Z})] = \mathbb{P}[\mathbf{Y} = 1 \mid \hat{f}(\mathbf{Z})] \text{ almost surely}, \mathbf{X} \sim \mathrm{Bern}(\hat{f}(\mathbf{Z})) .$$

We are interested in the hypothesis

$$H_0 : \mathbf{X} \mid \mathbf{W} \overset{d}{=} \mathbf{Y} \mid \mathbf{W}$$

Statistical formulation

Hypothesis formulation (classification): For joint distribution $\mathscr{L}_n(\mathbf{Y}, \mathrm{Bern}(\hat{f}(\mathbf{Z})))$, test the null hypothesis of calibration:

$$H_0 : \hat{f}(\mathbf{Z}) = \mathbb{P}[\mathbf{Y} = 1 \mid \hat{f}(\mathbf{Z})] \text{ almost surely} .$$

Equivalently,

$$H_0 : \mathbb{P}[\mathbf{X} = 1 \,|\, \hat{f}(\mathbf{Z})] = \mathbb{P}[\mathbf{Y} = 1 \mid \hat{f}(\mathbf{Z})] \text{ almost surely, } \mathbf{X} \sim \mathrm{Bern}(\hat{f}(\mathbf{Z})) .$$

We are interested in the hypothesis

$$H_0 : \mathbf{X} \,|\, \mathbf{W} \overset{d}{=} \mathbf{Y} \,|\, \mathbf{W}$$

$$\mathbf{W} = \hat{f}(\mathbf{Z}) \text{ in calibration test}$$

Relevant literature

Relevant literature

Regression curve comparison: Consider two regression models $X = f(\mathbf{W}) + \varepsilon, Y = g(\mathbf{W}) + \eta$. $H_0 : f = g \Leftrightarrow H_0 : \mathbf{X} \mid \mathbf{W} \stackrel{d}{=} \mathbf{Y} \mid \mathbf{W}$.

Relevant literature

Regression curve comparison: Consider two regression models $X = f(\mathbf{W}) + \varepsilon, Y = g(\mathbf{W}) + \eta$. $H_0 : f = g \Leftrightarrow H_0 : \mathbf{X} \,|\, \mathbf{W} \overset{d}{=} \mathbf{Y} \,|\, \mathbf{W}$.

(Dette et. al. 1998, AoS; Neumeyer et. al. 2003, AoS)

Relevant literature

Regression curve comparison: Consider two regression models $X = f(\mathbf{W}) + \varepsilon, Y = g(\mathbf{W}) + \eta$. $H_0 : f = g \Leftrightarrow H_0 : \mathbf{X} \,|\, \mathbf{W} \overset{d}{=} \mathbf{Y} \,|\, \mathbf{W}$.

(Dette et. al. 1998, AoS; Neumeyer et. al. 2003, AoS)

Conditional goodness-of-fit test: Given a conditional distribution $f(x \,|\, w)$, we are interested in if the observed data $(Y_i, W_i), i = 1, \ldots, n$ fit the distribution well or not. $H_0 : \mathbf{X} \,|\, \mathbf{W} \overset{d}{=} \mathbf{Y} \,|\, \mathbf{W}$

Relevant literature

Regression curve comparison: Consider two regression models $X = f(\mathbf{W}) + \varepsilon, Y = g(\mathbf{W}) + \eta$. $H_0 : f = g \Leftrightarrow H_0 : \mathbf{X} \mid \mathbf{W} \overset{d}{=} \mathbf{Y} \mid \mathbf{W}$.

(Dette et. al. 1998, AoS; Neumeyer et. al. 2003, AoS)

Conditional goodness-of-fit test: Given a conditional distribution $f(x \mid w)$, we are interested in if the observed data $(Y_i, W_i), i = 1,\ldots,n$ fit the distribution well or not. $H_0 : \mathbf{X} \mid \mathbf{W} \overset{d}{=} \mathbf{Y} \mid \mathbf{W}$

(Andrews 1997, Econometrica, Zheng 2000, Econometric Theory)

Relevant literature

Regression curve comparison: Consider two regression models $X = f(\mathbf{W}) + \varepsilon, Y = g(\mathbf{W}) + \eta$. $H_0 : f = g \Leftrightarrow H_0 : \mathbf{X} \,|\, \mathbf{W} \overset{d}{=} \mathbf{Y} \,|\, \mathbf{W}$.

(Dette et. al. 1998, AoS; Neumeyer et. al. 2003, AoS)

Conditional goodness-of-fit test: Given a conditional distribution $f(x \,|\, w)$, we are interested in if the observed data $(Y_i, W_i), i = 1, \ldots, n$ fit the distribution well or not. $H_0 : \mathbf{X} \,|\, \mathbf{W} \overset{d}{=} \mathbf{Y} \,|\, \mathbf{W}$

(Andrews 1997, Econometrica, Zheng 2000, Econometric Theory)

Calibration test:

Relevant literature

Regression curve comparison: Consider two regression models $X = f(\mathbf{W}) + \varepsilon, Y = g(\mathbf{W}) + \eta$. $H_0 : f = g \Leftrightarrow H_0 : \mathbf{X} \,|\, \mathbf{W} \overset{d}{=} \mathbf{Y} \,|\, \mathbf{W}$.

(Dette et. al. 1998, AoS; Neumeyer et. al. 2003, AoS)

Conditional goodness-of-fit test: Given a conditional distribution $f(x \,|\, w)$, we are interested in if the observed data $(Y_i, W_i), i = 1,\ldots,n$ fit the distribution well or not. $H_0 : \mathbf{X} \,|\, \mathbf{W} \overset{d}{=} \mathbf{Y} \,|\, \mathbf{W}$

(Andrews 1997, Econometrica, Zheng 2000, Econometric Theory)

Calibration test:

(Widmann et. al. 2019, NeurIPS; Widmann et. al. 2021, ICLR) SKCE method

One-sample v.s. two-sample statistics

One-sample v.s. two-sample statistics

One-sample statistics with $(Y_i, W_i)_{i=1}^n$, e.g. SKCE:

One-sample v.s. two-sample statistics

One-sample statistics with $(Y_i, W_i)_{i=1}^n$, e.g. SKCE:

$$nT_{\text{SKCE}} \xrightarrow{H_0} \sum_{m=1}^{\infty} \lambda_m (Z_k^2 - 1),\ Z_k \overset{iid}{\sim} N(0,1)$$

One-sample v.s. two-sample statistics

One-sample statistics with $(Y_i, W_i)_{i=1}^n$**, e.g. SKCE:**

$$nT_{\text{SKCE}} \xrightarrow{H_0} \sum_{m=1}^{\infty} \lambda_m (Z_k^2 - 1),\ Z_k \overset{iid}{\sim} N(0,1)$$

Two-sample statistics with $(X_i, Y_i, W_i)_{i=1}^n,\ X_i \sim \mathbb{P}_{\mathbf{Y}_i|\mathbf{W}_i}$**, e.g. ECMMD:**

One-sample v.s. two-sample statistics

One-sample statistics with $(Y_i, W_i)_{i=1}^n$, e.g. SKCE:

$$nT_{\text{SKCE}} \xrightarrow{H_0} \sum_{m=1}^{\infty} \lambda_m (Z_k^2 - 1),\ Z_k \overset{iid}{\sim} N(0,1)$$

Two-sample statistics with $(X_i, Y_i, W_i)_{i=1}^n$, $X_i \sim \mathbb{P}_{\mathbf{Y}_i|\mathbf{W}_i}$, e.g. ECMMD:

$$a_n \frac{T_{\text{ECMMD}}}{\hat{\sigma}_n} \xrightarrow{H_0} N(0,1)$$

One-sample v.s. two-sample statistics

One-sample statistics with $(Y_i, W_i)_{i=1}^n$**, e.g. SKCE:**

$$nT_{\text{SKCE}} \overset{H_0}{\to} \sum_{m=1}^{\infty} \lambda_m (Z_k^2 - 1),\ Z_k \overset{iid}{\sim} N(0,1)$$

Two-sample statistics with $(X_i, Y_i, W_i)_{i=1}^n$, $X_i \sim \mathbb{P}_{\mathbf{Y}_i|\mathbf{W}_i}$**, e.g. ECMMD:**

$$a_n \frac{T_{\text{ECMMD}}}{\hat{\sigma}_n} \overset{H_0}{\to} N(0,1)$$

Intractable distribution $\sum_{m=1}^{\infty} \lambda_m (Z_k^2 - 1)$ **versus "nice" distribution** $N(0,1)$**.**

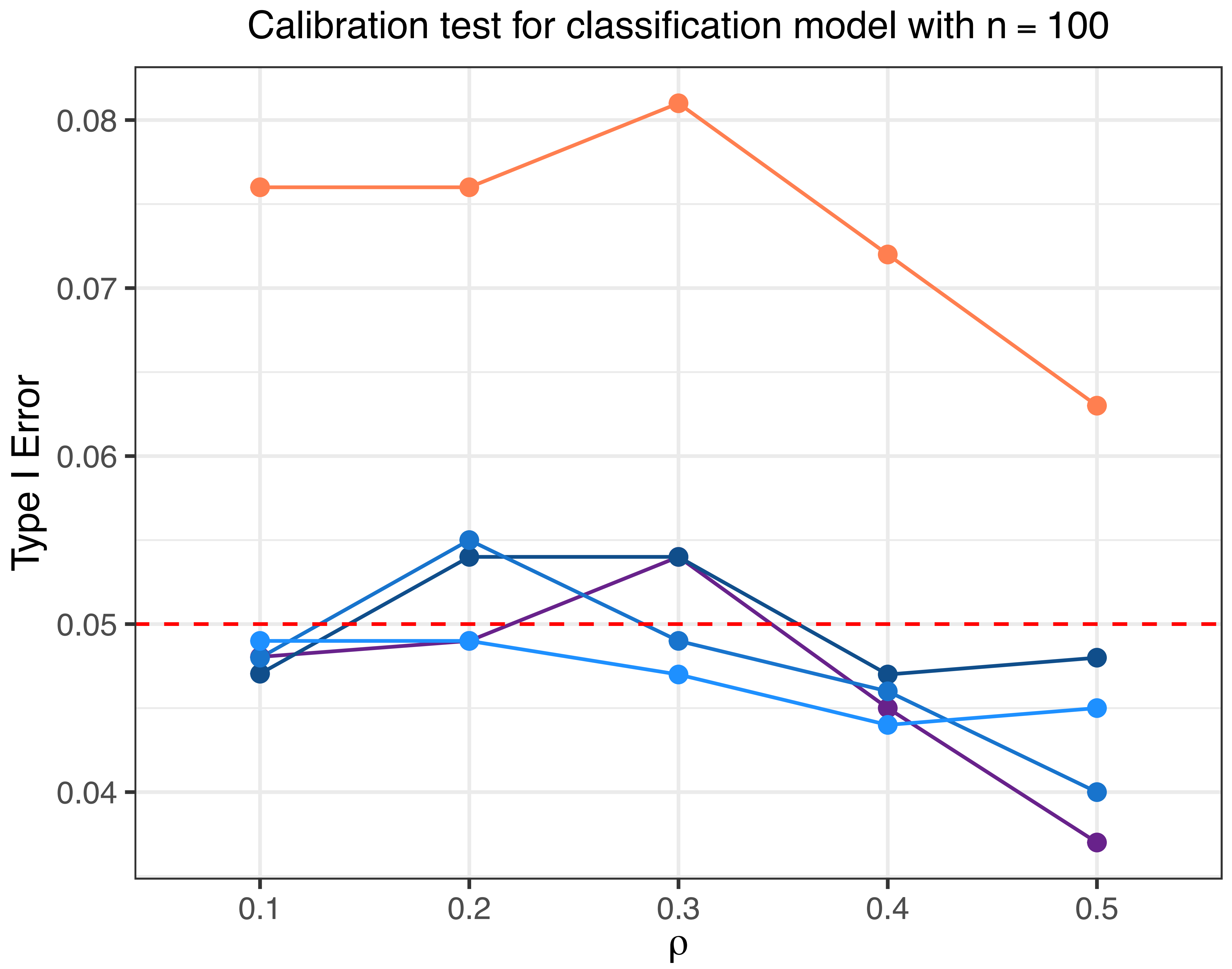
Challenge: Type-I error under null

Challenge: Type-I error under null

Type-I error inflation/deflation: degenerate U-statistic with intractable null distribution.

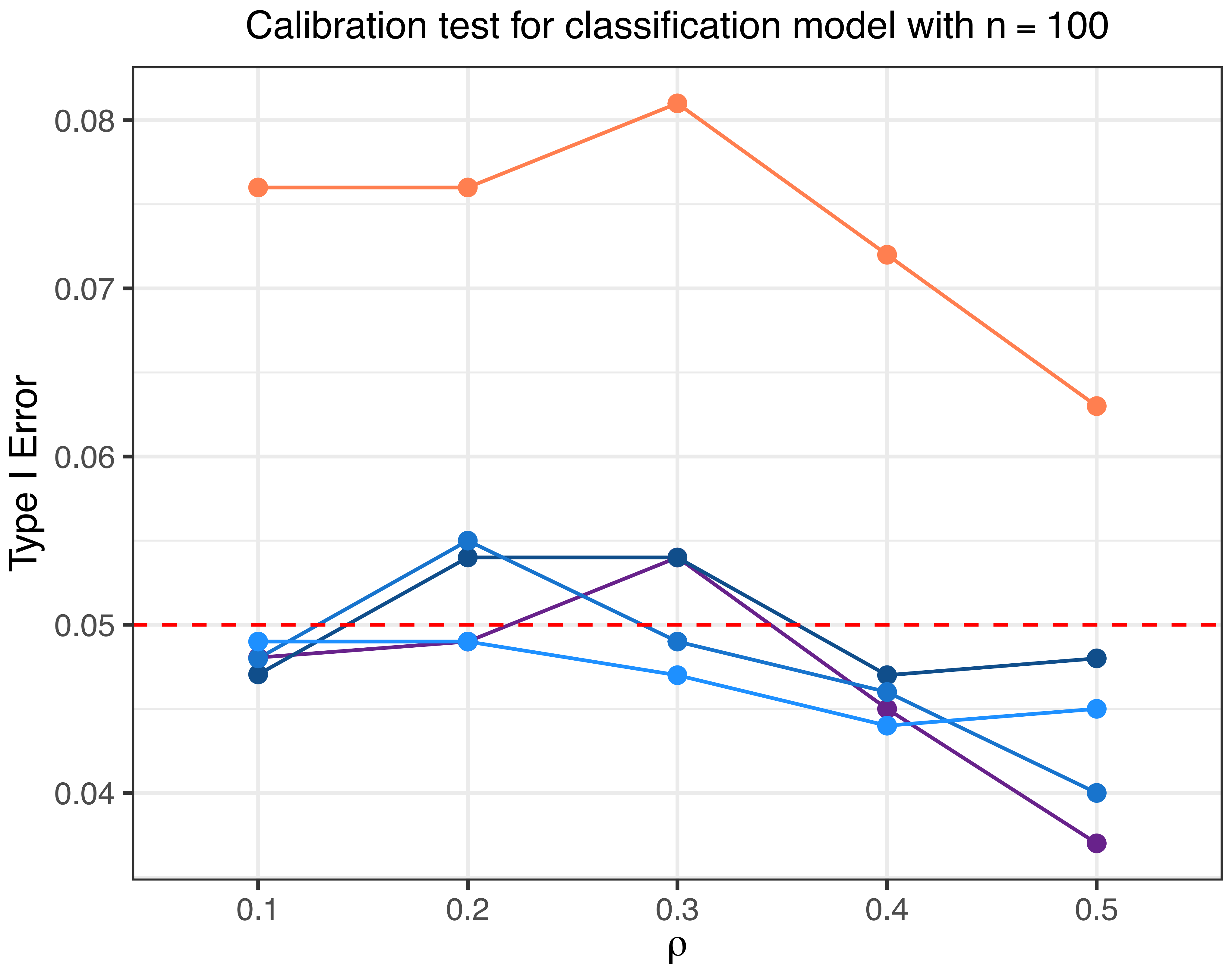
Challenge: Type-I error under null

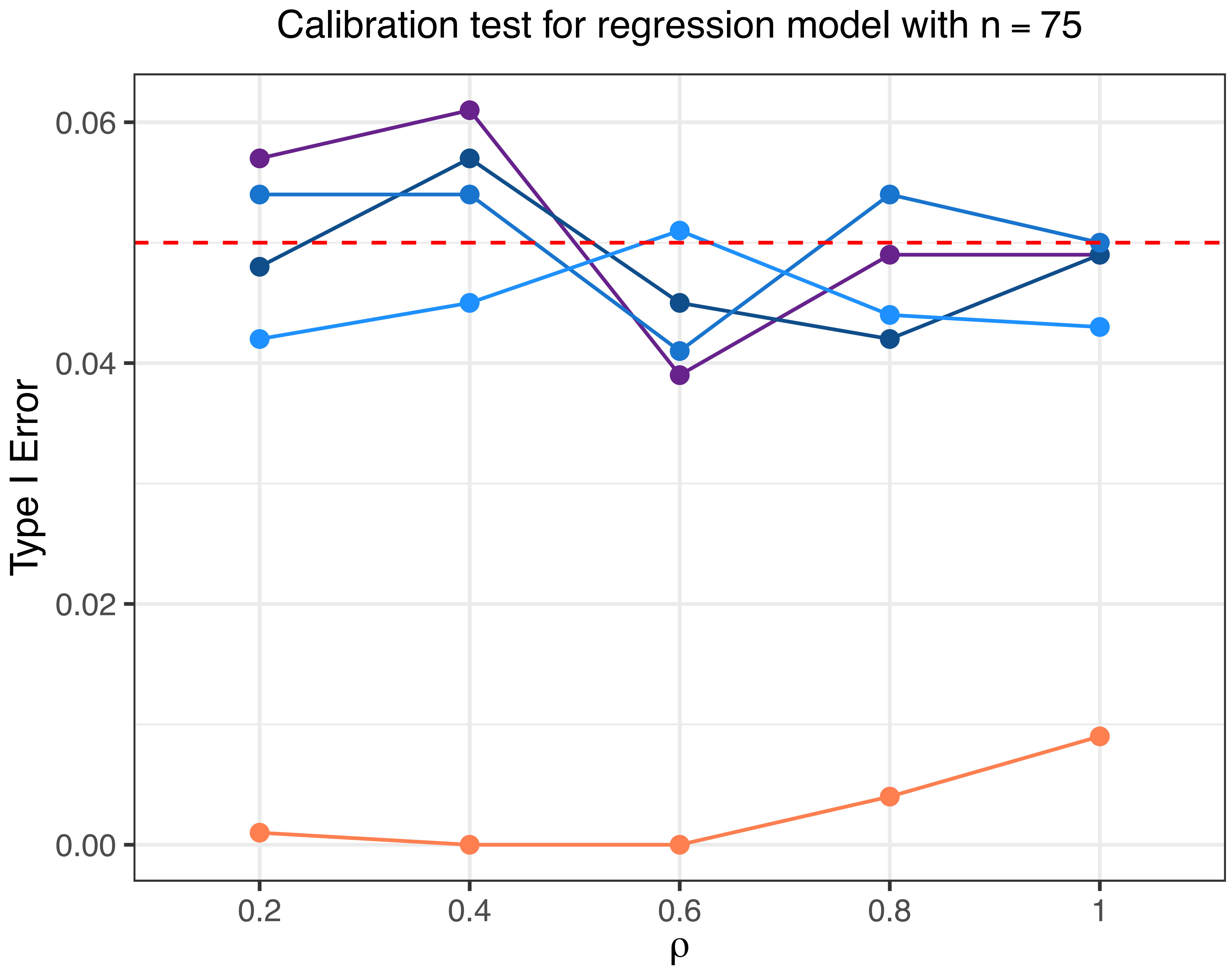
Type-I error inflation/deflation: degenerate U-statistic with intractable null distribution.



Challenge: Type-I error under null

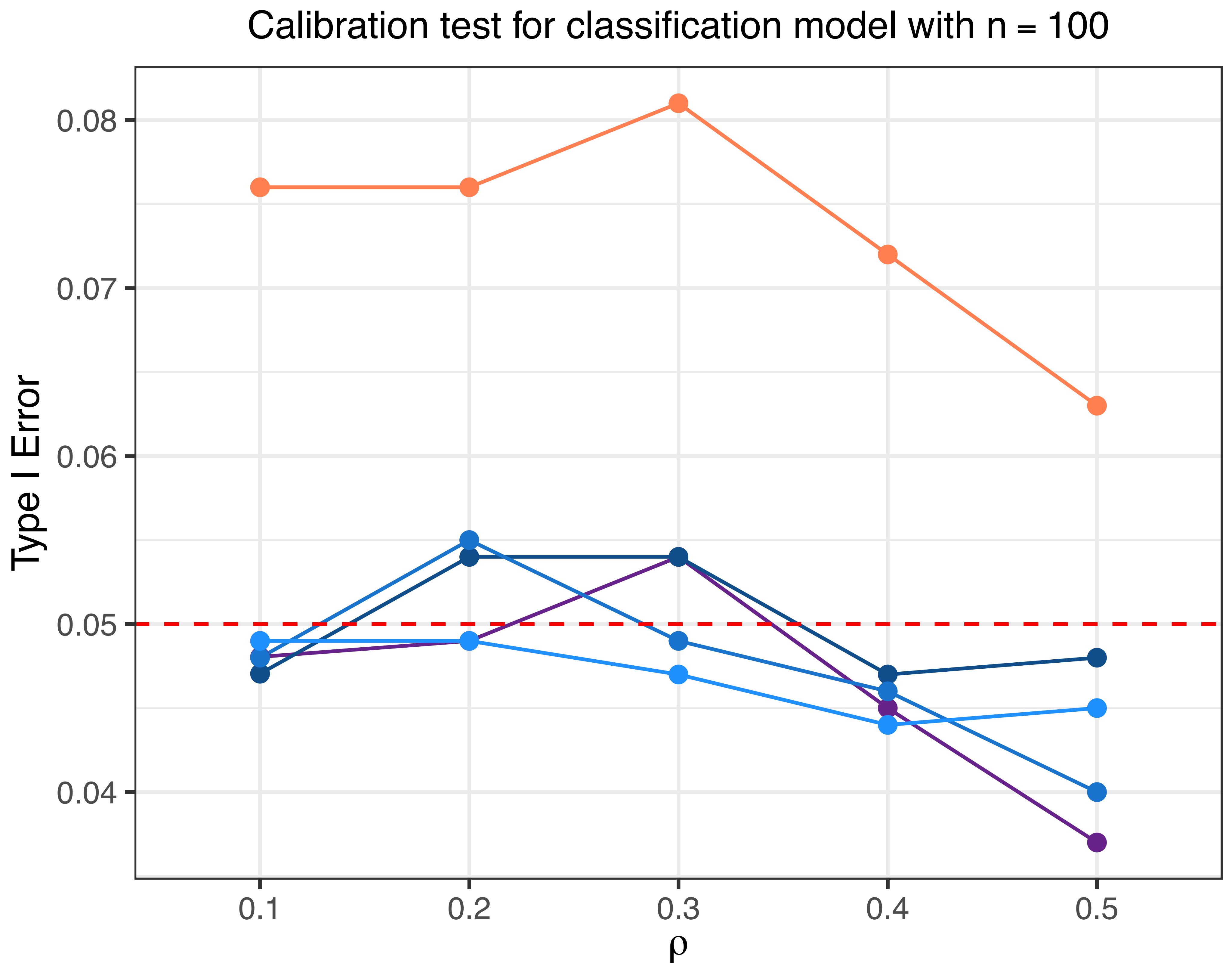
Type-I error inflation/deflation: degenerate U-statistic with intractable null distribution.

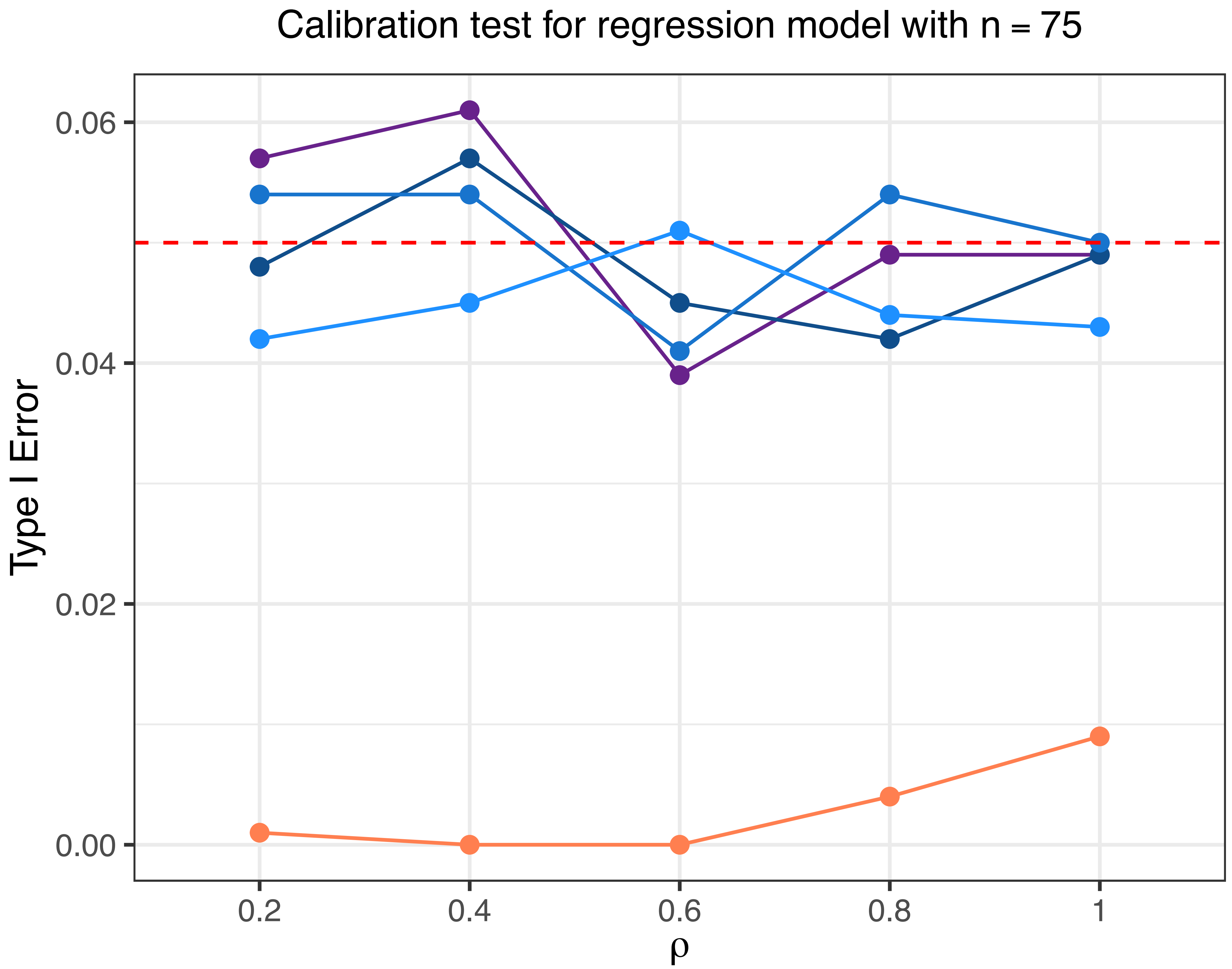




Challenge: Type-I error under null

Type-I error inflation/deflation: degenerate U-statistic with intractable null distribution.

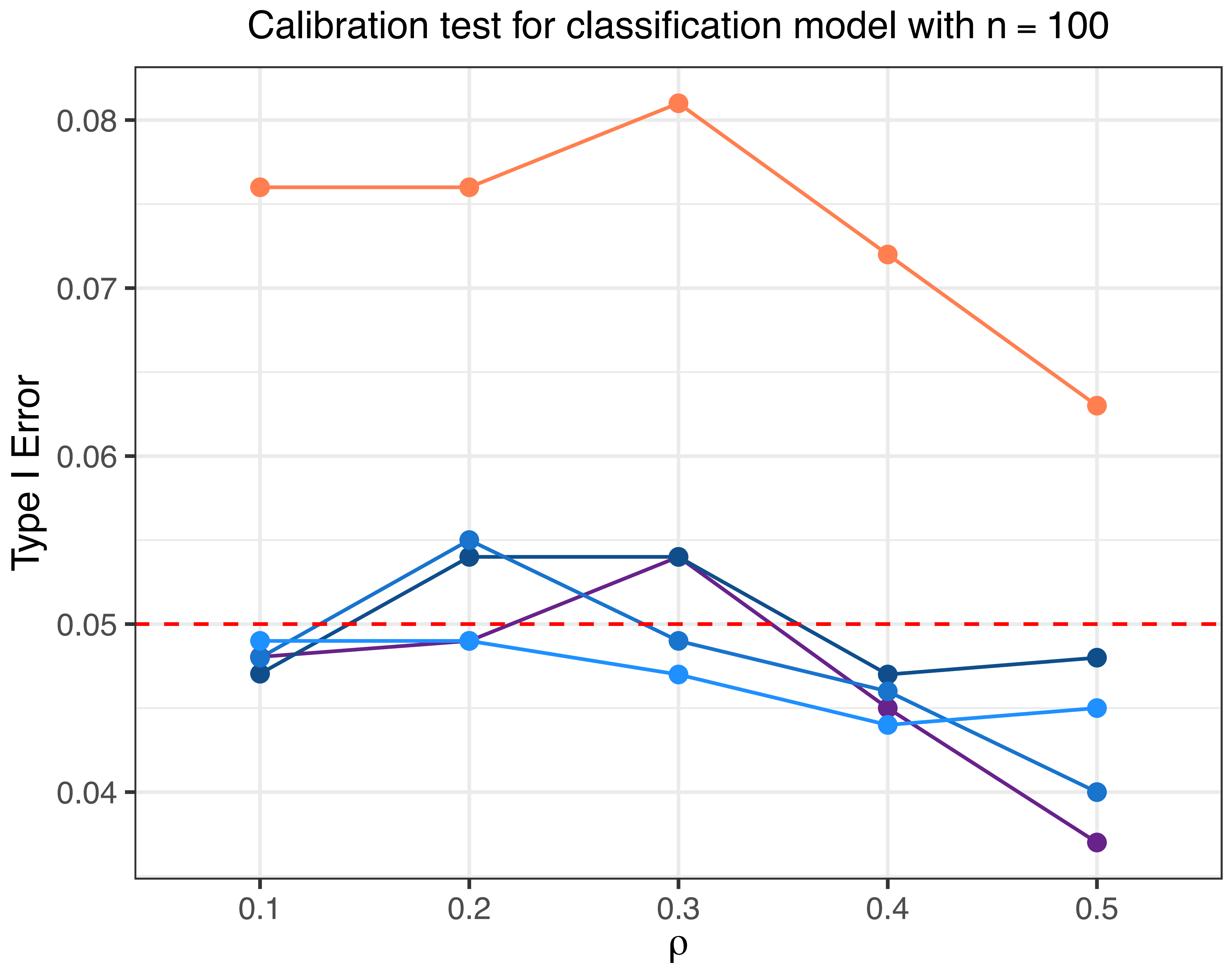


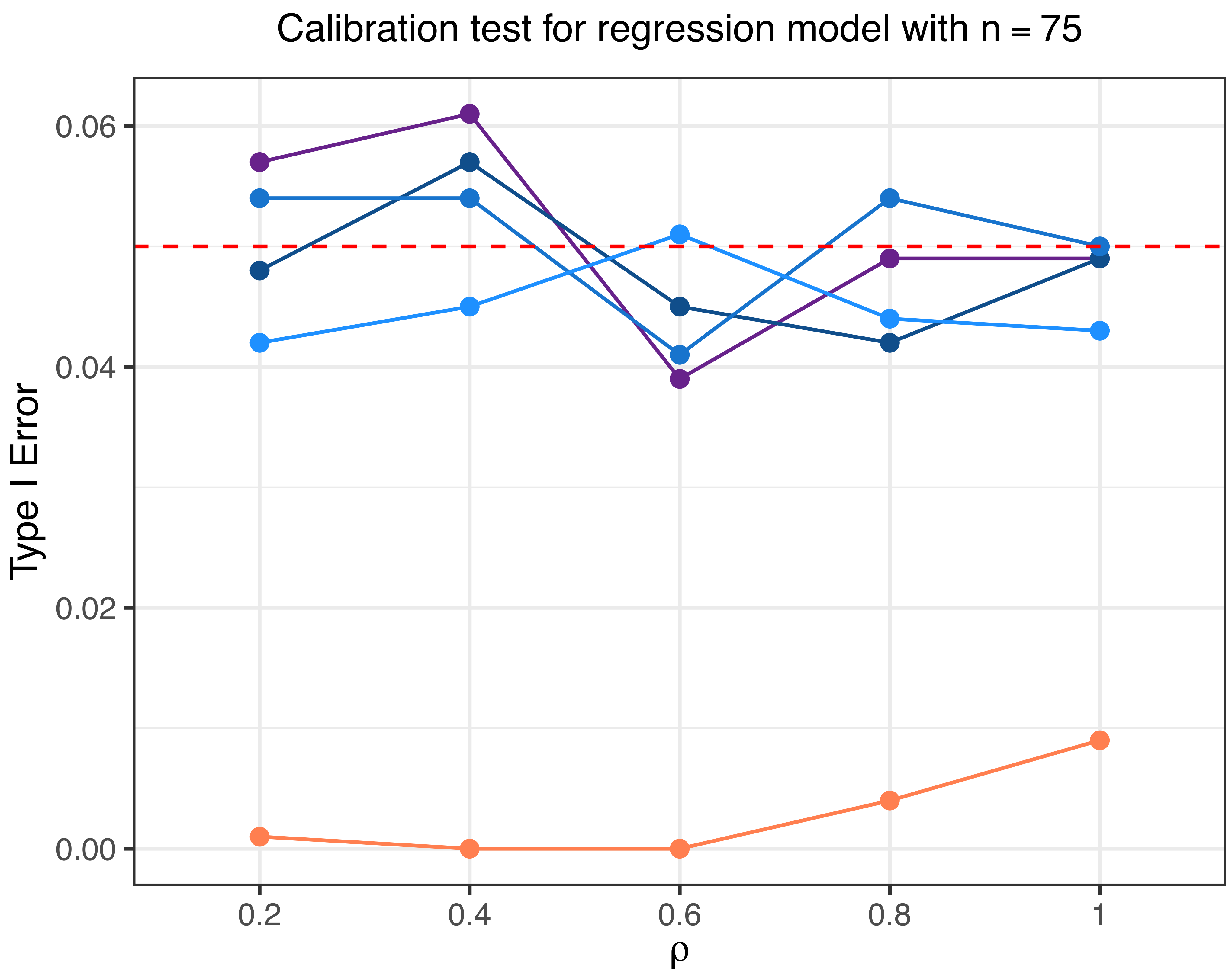


SKCE:

Challenge: Type-I error under null

Type-I error inflation/deflation: degenerate U-statistic with intractable null distribution.





SKCE:
Proposed: the other

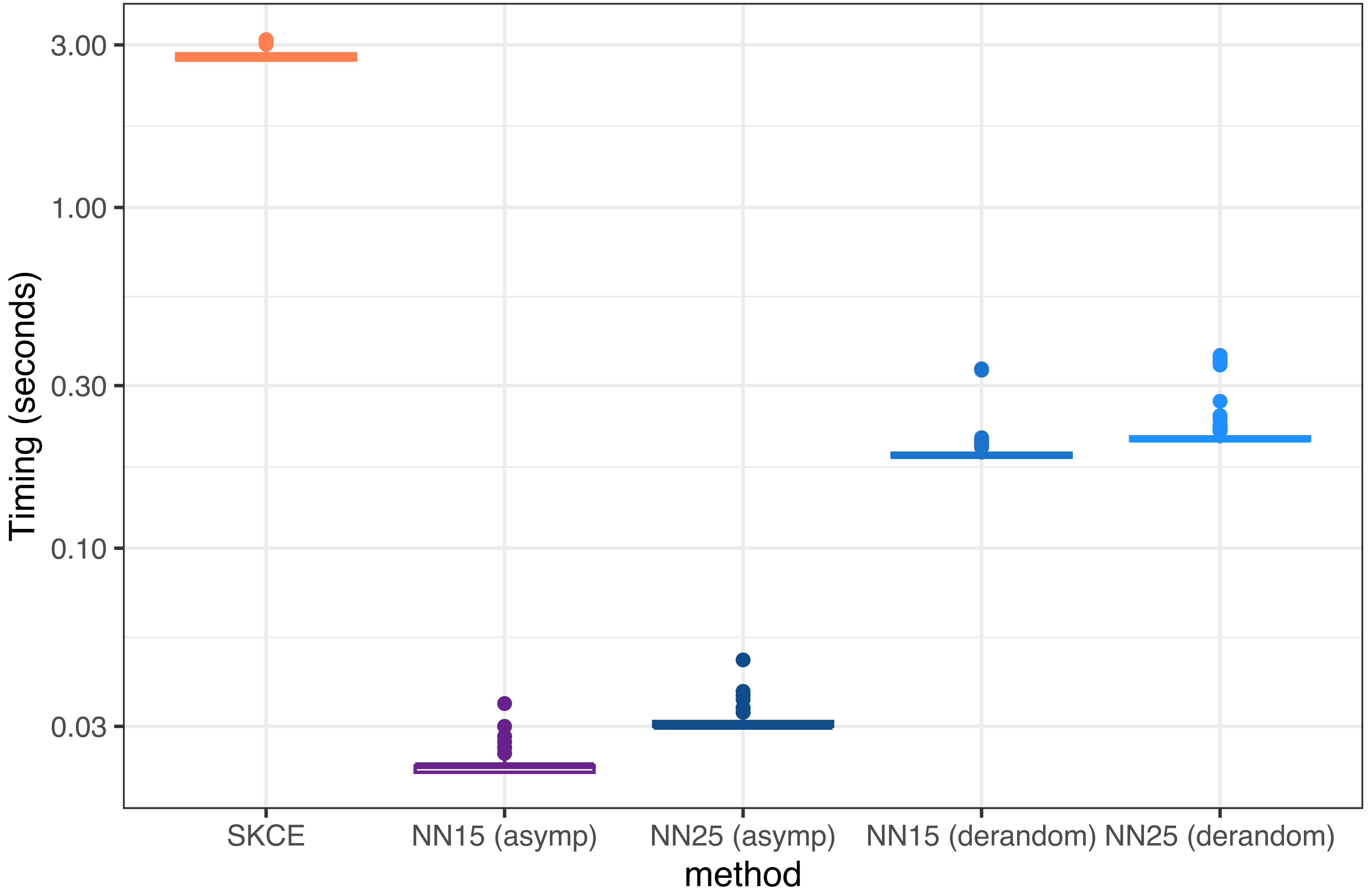
Challenge: computation

Challenge: computation

Resampling requirement: need many resamples to give reliable p-value estimate.

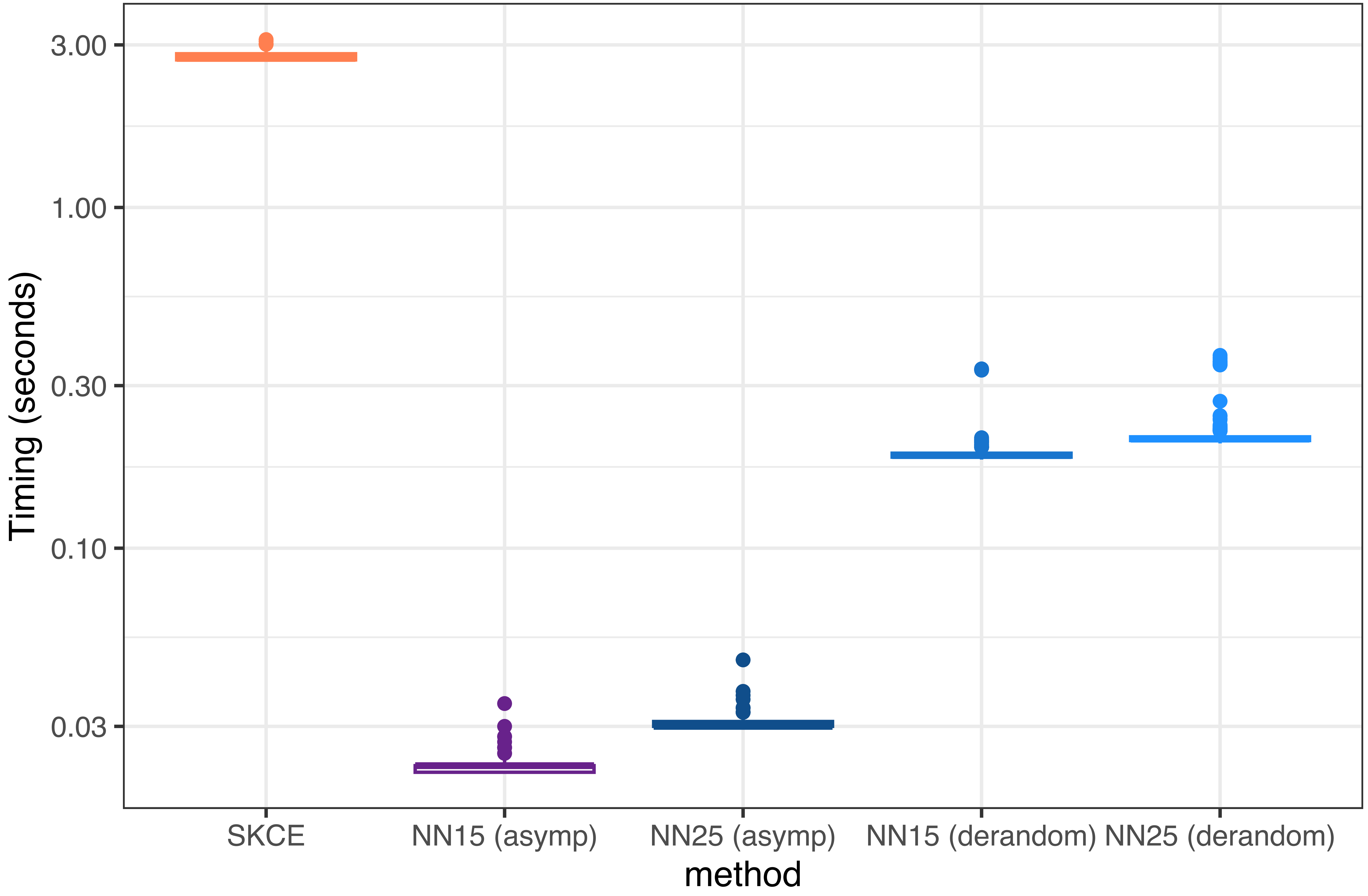
Challenge: computation

Resampling requirement: need many resamples to give reliable p-value estimate.



Challenge: computation

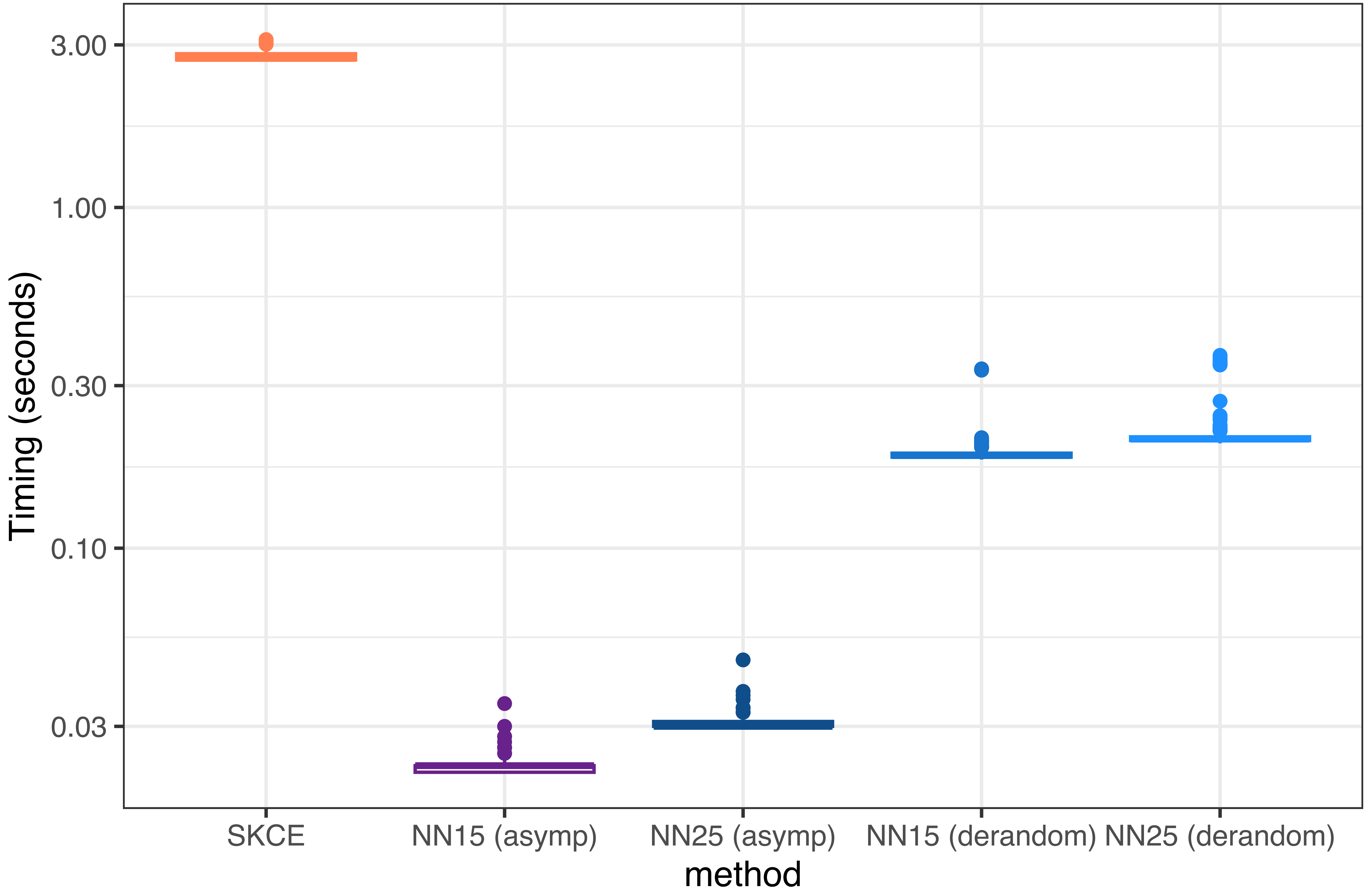
Resampling requirement: need many resamples to give reliable p-value estimate.



SKCE:

Challenge: computation

Resampling requirement: need many resamples to give reliable p-value estimate.



SKCE:
Proposed: the other

RKHS preliminary

RKHS preliminary

Kernel and RKHS: $K(x, y) : \mathbb{R} \times \mathbb{R} \mapsto \mathbb{R}, \mathscr{H}_K.$

RKHS preliminary

Kernel and RKHS: $K(x, y) : \mathbb{R} \times \mathbb{R} \mapsto \mathbb{R},\ \mathscr{H}_K.$

Any positive semidefinite kernel is associated with a unique Hilbert space $\mathscr{H}_K$ satisfying:

RKHS preliminary

Kernel and RKHS: $K(x, y) : \mathbb{R} \times \mathbb{R} \mapsto \mathbb{R}$, $\mathscr{H}_K$.

Any positive semidefinite kernel is associated with a unique Hilbert space $\mathscr{H}_K$ satisfying:[1]

Wainwright, 2019

RKHS preliminary

Kernel and RKHS: $K(x, y) : \mathbb{R} \times \mathbb{R} \mapsto \mathbb{R}, \mathscr{H}_K$.

Any positive semidefinite kernel is associated with a unique Hilbert space $\mathscr{H}_K$ satisfying:[1]

(1) $K(\cdot, x) \in \mathscr{H}_K$;

Wainwright, 2019

RKHS preliminary

Kernel and RKHS: $K(x, y) : \mathbb{R} \times \mathbb{R} \mapsto \mathbb{R}, \mathscr{H}_K$.

Any positive semidefinite kernel is associated with a unique Hilbert space $\mathscr{H}_K$ satisfying:[1]

$$(1)\ K(\cdot, x) \in \mathscr{H}_K;$$

$$(2)\ \langle f(\cdot), K(\cdot, x) \rangle_{\mathscr{H}_K} = f(x),\ \forall f \in \mathscr{H}_K.$$

Wainwright, 2019

RKHS preliminary

Kernel and RKHS: $K(x, y) : \mathbb{R} \times \mathbb{R} \mapsto \mathbb{R}$, $\mathscr{H}_K$.

Any positive semidefinite kernel is associated with a unique Hilbert space $\mathscr{H}_K$ satisfying:[1]

$$(1)\ K(\cdot, x) \in \mathscr{H}_K;$$

$$(2)\ \langle f(\cdot), K(\cdot, x) \rangle_{\mathscr{H}_K} = f(x),\ \forall f \in \mathscr{H}_K.$$

Kernel mean embedding: $\mu_{\mathbb{P}}$ satisfying $\langle \mu_{\mathbb{P}}, f \rangle_{\mathscr{H}_K} = \mathbb{E}_{X \sim \mathbb{P}}[f(X)]$, $\forall f \in \mathscr{H}_K$.

Wainwright, 2019

RKHS preliminary

Kernel and RKHS: $K(x, y) : \mathbb{R} \times \mathbb{R} \mapsto \mathbb{R}$, $\mathscr{H}_K$.

Any positive semidefinite kernel is associated with a unique Hilbert space $\mathscr{H}_K$ satisfying:[1]

(1) $K(\cdot, x) \in \mathscr{H}_K$;

(2) $\langle f(\cdot), K(\cdot, x) \rangle_{\mathscr{H}_K} = f(x), \forall f \in \mathscr{H}_K$.

Kernel mean embedding: $\mu_{\mathbb{P}}$ satisfying $\langle \mu_{\mathbb{P}}, f \rangle_{\mathscr{H}_K} = \mathbb{E}_{X \sim \mathbb{P}}[f(X)], \forall f \in \mathscr{H}_K$.

Linear kernel: $K(x, y) = x \cdot y$ and $\mu_{\mathbb{P}}(y) = \mathbb{E}_{X \sim \mathbb{P}}[K(X, y)] = \mathbb{E}_{X \sim \mathbb{P}}[X] \cdot y$.

Wainwright, 2019

A characteristic measure for linear kernel

A characteristic measure for linear kernel

Maximum mean discrepancy (MMD, Gretton et al. 2012):

A characteristic measure for linear kernel

Maximum mean discrepancy (MMD, Gretton et al. 2012):

$$\mathrm{MMD}^2(\mathbf{X}, \mathbf{Y}) \equiv \|\mu_{\mathbb{P}_\mathbf{X}} - \mu_{\mathbb{P}_\mathbf{Y}}\|^2_{\mathscr{H}_K} = (\mathbb{E}_{X\sim\mathbb{P}_\mathbf{X}}[X] - \mathbb{E}_{Y\sim\mathbb{P}_\mathbf{Y}}[Y])^2$$

A characteristic measure for linear kernel

Maximum mean discrepancy (MMD, Gretton et al. 2012):

$$\mathrm{MMD}^2(\mathbf{X}, \mathbf{Y}) \equiv \|\mu_{\mathbb{P}_\mathbf{X}} - \mu_{\mathbb{P}_\mathbf{Y}}\|^2_{\mathscr{H}_K} = (\mathbb{E}_{X\sim\mathbb{P}_\mathbf{X}}[X] - \mathbb{E}_{Y\sim\mathbb{P}_\mathbf{Y}}[Y])^2$$

Expected conditional maximum mean discrepancy (ECMMD):

A characteristic measure for linear kernel

Maximum mean discrepancy (MMD, Gretton et al. 2012):

$$\mathrm{MMD}^2(\mathbf{X}, \mathbf{Y}) \equiv \|\mu_{\mathbb{P}_\mathbf{X}} - \mu_{\mathbb{P}_\mathbf{Y}}\|^2_{\mathscr{H}_K} = (\mathbb{E}_{X\sim\mathbb{P}_\mathbf{X}}[X] - \mathbb{E}_{Y\sim\mathbb{P}_\mathbf{Y}}[Y])^2$$

Expected conditional maximum mean discrepancy (ECMMD):

$$\mathrm{ECMMD}^2(\mathbf{X}, \mathbf{Y} \,|\, \mathbf{W}) \equiv \mathbb{E}_\mathbf{W}[\mathrm{MMD}^2(\mathbf{X} \,|\, \mathbf{W}, \mathbf{Y} \,|\, \mathbf{W})] = \mathbb{E}_\mathbf{W}[(\mathbb{E}[\mathbf{X} \,|\, \mathbf{W}] - \mathbb{E}[(\mathbf{Y} \,|\, \mathbf{W}])^2]$$

A characteristic measure for linear kernel

Maximum mean discrepancy (MMD, Gretton et al. 2012):

$$\mathrm{MMD}^2(\mathbf{X}, \mathbf{Y}) \equiv \|\mu_{\mathbb{P}_\mathbf{X}} - \mu_{\mathbb{P}_\mathbf{Y}}\|^2_{\mathscr{H}_K} = (\mathbb{E}_{X\sim\mathbb{P}_\mathbf{X}}[X] - \mathbb{E}_{Y\sim\mathbb{P}_\mathbf{Y}}[Y])^2$$

Expected conditional maximum mean discrepancy (ECMMD):

$$\mathrm{ECMMD}^2(\mathbf{X}, \mathbf{Y} \,|\, \mathbf{W}) \equiv \mathbb{E}_\mathbf{W}[\mathrm{MMD}^2(\mathbf{X} \,|\, \mathbf{W}, \mathbf{Y} \,|\, \mathbf{W})] = \mathbb{E}_\mathbf{W}[(\mathbb{E}[\mathbf{X} \,|\, \mathbf{W}] - \mathbb{E}[(\mathbf{Y} \,|\, \mathbf{W}])^2]$$

Linear kernel $K(x, y) = x \cdot y$is characteristic with binary outcome $\mathbf{X}, \mathbf{Y}$:

A characteristic measure for linear kernel

Maximum mean discrepancy (MMD, Gretton et al. 2012):

$$\mathrm{MMD}^2(\mathbf{X}, \mathbf{Y}) \equiv \|\mu_{\mathbb{P}_\mathbf{X}} - \mu_{\mathbb{P}_\mathbf{Y}}\|^2_{\mathscr{H}_K} = (\mathbb{E}_{X\sim\mathbb{P}_\mathbf{X}}[X] - \mathbb{E}_{Y\sim\mathbb{P}_\mathbf{Y}}[Y])^2$$

Expected conditional maximum mean discrepancy (ECMMD):

$$\mathrm{ECMMD}^2(\mathbf{X}, \mathbf{Y} \mid \mathbf{W}) \equiv \mathbb{E}_\mathbf{W}[\mathrm{MMD}^2(\mathbf{X} \mid \mathbf{W}, \mathbf{Y} \mid \mathbf{W})] = \mathbb{E}_\mathbf{W}[(\mathbb{E}[\mathbf{X} \mid \mathbf{W}] - \mathbb{E}[(\mathbf{Y} \mid \mathbf{W}])^2]$$

Linear kernel $K(x, y) = x \cdot y$is characteristic with binary outcome $\mathbf{X}, \mathbf{Y}$:

$$\mathrm{ECMMD}^2 = 0 \text{ if and only if } \mathbf{X} \mid \mathbf{W} \overset{d}{=} \mathbf{Y} \mid \mathbf{W}\,.$$

A characteristic measure for linear kernel

Maximum mean discrepancy (MMD, Gretton et al. 2012):

$$\mathrm{MMD}^2(\mathbf{X}, \mathbf{Y}) \equiv \|\mu_{\mathbb{P}_\mathbf{X}} - \mu_{\mathbb{P}_\mathbf{Y}}\|^2_{\mathscr{H}_K} = (\mathbb{E}_{X\sim\mathbb{P}_\mathbf{X}}[X] - \mathbb{E}_{Y\sim\mathbb{P}_\mathbf{Y}}[Y])^2$$

Expected conditional maximum mean discrepancy (ECMMD):

$$\mathrm{ECMMD}^2(\mathbf{X}, \mathbf{Y} \mid \mathbf{W}) \equiv \mathbb{E}_\mathbf{W}[\mathrm{MMD}^2(\mathbf{X} \mid \mathbf{W}, \mathbf{Y} \mid \mathbf{W})] = \mathbb{E}_\mathbf{W}[(\mathbb{E}[\mathbf{X} \mid \mathbf{W}] - \mathbb{E}[(\mathbf{Y} \mid \mathbf{W}])^2]$$

Linear kernel $K(x, y) = x \cdot y$is characteristic with binary outcome $\mathbf{X}, \mathbf{Y}$:

$\mathrm{ECMMD}^2 = 0$ if and only if $\mathbf{X} \mid \mathbf{W} \overset{d}{=} \mathbf{Y} \mid \mathbf{W}$.

$\mathrm{ECMMD}^2 = 0$ if and only if H_0 is true.

Biased estimation and reformulation of ECMMD

Biased estimation and reformulation of ECMMD

Plug-in estimation is biased even under null: with any nonparametric estimator for $\mathbb{E}[\mathbf{X} \mid \mathbf{W}]$ and $\mathbb{E}[\mathbf{Y} \mid \mathbf{W}]$, e.g. KNN or kernel regression estimator, the resulting estimate is not $\sqrt{n}$ unbiased:

Biased estimation and reformulation of ECMMD

Plug-in estimation is biased even under null: with any nonparametric estimator for $\mathbb{E}[\mathbf{X} \mid \mathbf{W}]$ and $\mathbb{E}[\mathbf{Y} \mid \mathbf{W}]$, e.g. KNN or kernel regression estimator, the resulting estimate is not $\sqrt{n}$ unbiased:

$$\sqrt{n}\left(\mathbb{E}\left[\frac{1}{n}\sum_{i=1}^{n}(\hat{\mathbb{E}}[X_i \mid W_i] - \hat{\mathbb{E}}[Y_i \mid W_i])^2\right] - \mathrm{ECMMD}^2\right) \overset{H_0}{\to} \infty$$

Biased estimation and reformulation of ECMMD

Plug-in estimation is biased even under null: with any nonparametric estimator for $\mathbb{E}[\mathbf{X} \mid \mathbf{W}]$ and $\mathbb{E}[\mathbf{Y} \mid \mathbf{W}]$, e.g. KNN or kernel regression estimator, the resulting estimate is not $\sqrt{n}$ unbiased:

$$\sqrt{n}\left(\mathbb{E}\left[\frac{1}{n}\sum_{i=1}^{n}(\hat{\mathbb{E}}[X_i \mid W_i] - \hat{\mathbb{E}}[Y_i \mid W_i])^2\right] - \text{ECMMD}^2 \right) \overset{H_0}{\to} \infty$$

An alternative form of ECMMD: $\mathbf{W} \sim \mathbb{P}_{\mathbf{W}}, \ (\mathbf{X}, \mathbf{Y}), (\mathbf{X}', \mathbf{Y}') \overset{\text{i.i.d.}}{\sim} \mathbb{P}_{(\mathbf{X},\mathbf{Y})|\mathbf{W}}$

Biased estimation and reformulation of ECMMD

Plug-in estimation is biased even under null: with any nonparametric estimator for $\mathbb{E}[\mathbf{X} \mid \mathbf{W}]$ and $\mathbb{E}[\mathbf{Y} \mid \mathbf{W}]$, e.g. KNN or kernel regression estimator, the resulting estimate is not $\sqrt{n}$ unbiased:

$$\sqrt{n}\left(\mathbb{E}\left[\frac{1}{n}\sum_{i=1}^{n}(\hat{\mathbb{E}}[X_i \mid W_i] - \hat{\mathbb{E}}[Y_i \mid W_i])^2\right] - \mathrm{ECMMD}^2\right) \xrightarrow{H_0} \infty$$

An alternative form of ECMMD: $\mathbf{W} \sim \mathbb{P}_{\mathbf{W}}$, $(\mathbf{X}, \mathbf{Y}), (\mathbf{X}', \mathbf{Y}') \overset{\text{i.i.d.}}{\sim} \mathbb{P}_{(\mathbf{X},\mathbf{Y})|\mathbf{W}}$

$$H((x, y), (x', y')) \equiv xx' + yy' - xy' - x'y$$

Biased estimation and reformulation of ECMMD

Plug-in estimation is biased even under null: with any nonparametric estimator for $\mathbb{E}[\mathbf{X} \mid \mathbf{W}]$ and $\mathbb{E}[\mathbf{Y} \mid \mathbf{W}]$, e.g. KNN or kernel regression estimator, the resulting estimate is not $\sqrt{n}$ unbiased:

$$\sqrt{n}\left(\mathbb{E}\left[\frac{1}{n}\sum_{i=1}^{n}(\hat{\mathbb{E}}[X_i \mid W_i] - \hat{\mathbb{E}}[Y_i \mid W_i])^2\right] - \mathrm{ECMMD}^2\right) \xrightarrow{H_0} \infty$$

An alternative form of ECMMD: $\mathbf{W} \sim \mathbb{P}_{\mathbf{W}}, \ (\mathbf{X}, \mathbf{Y}), (\mathbf{X}', \mathbf{Y}') \overset{\text{i.i.d.}}{\sim} \mathbb{P}_{(\mathbf{X},\mathbf{Y})|\mathbf{W}}$

$$H((x, y), (x', y')) \equiv xx' + yy' - xy' - x'y$$

$$\mathrm{ECMMD}^2 = \mathbb{E}_{\mathbf{W}}[(\mathbb{E}[\mathbf{X} - \mathbf{Y} \mid \mathbf{W}])^2] = \mathbb{E}_{\mathbf{W}}[\mathbb{E}[H((\mathbf{X}, \mathbf{Y}), (\mathbf{X}', \mathbf{Y}')) \mid \mathbf{W}]]$$

An intuitive estimator

An intuitive estimator

Nearest neighbor replacement: generate k_n nearest neighbor graph with data $(W_1, \ldots, W_n)$. $\mathcal{N}(i)$ is the k_n nearest neighbors set of $\mathbf{W}_i$.

An intuitive estimator

Nearest neighbor replacement: generate k_n nearest neighbor graph with data $(W_1, \ldots, W_n)$. $\mathcal{N}(i)$ is the k_n nearest neighbors set of $\mathbf{W}_i$.

$$\mathbb{E}[H((\mathbf{X}, \mathbf{Y}), (\mathbf{X}', \mathbf{Y}')) \mid \mathbf{W}_i] \approx \frac{1}{k_n} \sum_{j \in \mathcal{N}(i)} H((\mathbf{X}_i, \mathbf{Y}_i), (\mathbf{X}_j, \mathbf{Y}_j))$$

An intuitive estimator

Nearest neighbor replacement: generate k_n nearest neighbor graph with data $(W_1, \ldots, W_n)$. $\mathcal{N}(i)$ is the k_n nearest neighbors set of $\mathbf{W}_i$.

$$\mathbb{E}_{\mathbf{W}_i}[\mathbb{E}[H((\mathbf{X}, \mathbf{Y}), (\mathbf{X}', \mathbf{Y}')) \,|\, \mathbf{W}_i]] \approx \frac{1}{n}\sum_{i=1}^{n}\frac{1}{k_n}\sum_{j\in\mathcal{N}(i)} H((X_i, Y_i), (X_j, Y_j))$$

An intuitive estimator

Nearest neighbor replacement: generate k_n nearest neighbor graph with data $(W_1, \ldots, W_n)$. $\mathcal{N}(i)$ is the k_n nearest neighbors set of $\mathbf{W}_i$.

$$\mathbb{E}_{\mathbf{W}_i}[\mathbb{E}[H((\mathbf{X}, \mathbf{Y}), (\mathbf{X}', \mathbf{Y}')) \mid \mathbf{W}_i]] \approx \frac{1}{n} \sum_{i=1}^{n} \frac{1}{k_n} \sum_{j \in \mathcal{N}(i)} H((X_i, Y_i), (X_j, Y_j))$$

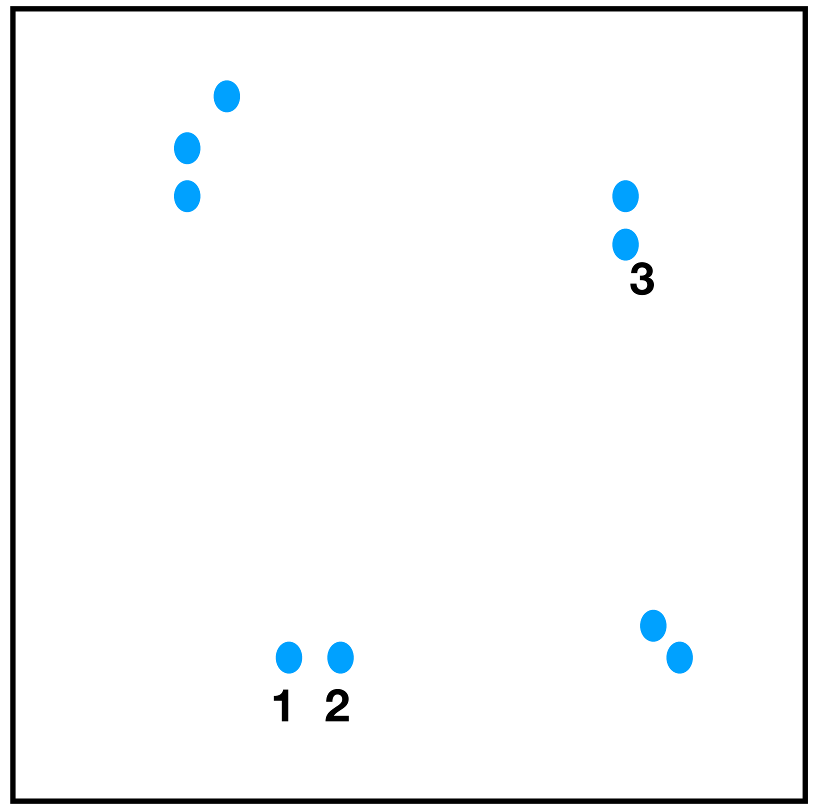
Intuition behind the estimator: $k_n = 1,\ n = 9.$

An intuitive estimator

Nearest neighbor replacement: generate k_n nearest neighbor graph with data $(W_1, \ldots, W_n)$. $\mathcal{N}(i)$ is the k_n nearest neighbors set of $\mathbf{W}_i$.

$$\mathbb{E}_{\mathbf{W}_i}[\mathbb{E}[H((\mathbf{X}, \mathbf{Y}), (\mathbf{X}', \mathbf{Y}')) \mid \mathbf{W}_i]] \approx \frac{1}{n} \sum_{i=1}^{n} \frac{1}{k_n} \sum_{j \in \mathcal{N}(i)} H((X_i, Y_i), (X_j, Y_j))$$

Intuition behind the estimator: $k_n = 1,\ n = 9$.



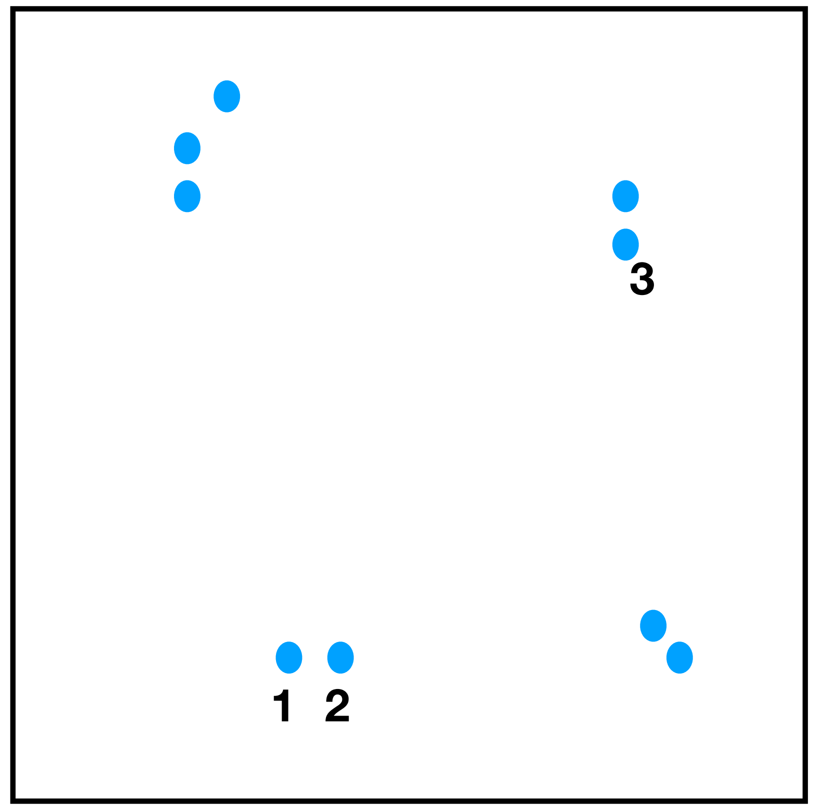
$W_1, \ldots, W_9$

An intuitive estimator

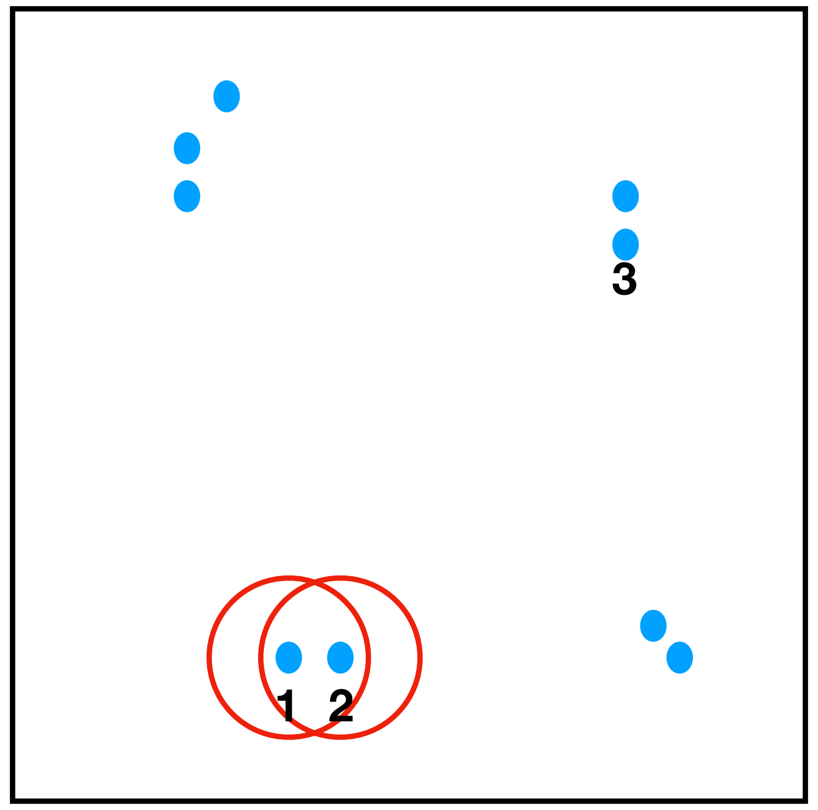
Nearest neighbor replacement: generate k_n nearest neighbor graph with data $(W_1, \ldots, W_n)$. $\mathcal{N}(i)$ is the k_n nearest neighbors set of $\mathbf{W}_i$.

$$\mathbb{E}_{\mathbf{W}_i}[\mathbb{E}[H((\mathbf{X}, \mathbf{Y}), (\mathbf{X}', \mathbf{Y}')) \mid \mathbf{W}_i]] \approx \frac{1}{n} \sum_{i=1}^{n} \frac{1}{k_n} \sum_{j \in \mathcal{N}(i)} H((X_i, Y_i), (X_j, Y_j))$$

Intuition behind the estimator: $k_n = 1,\ n = 9.$



$W_1, \ldots, W_9$



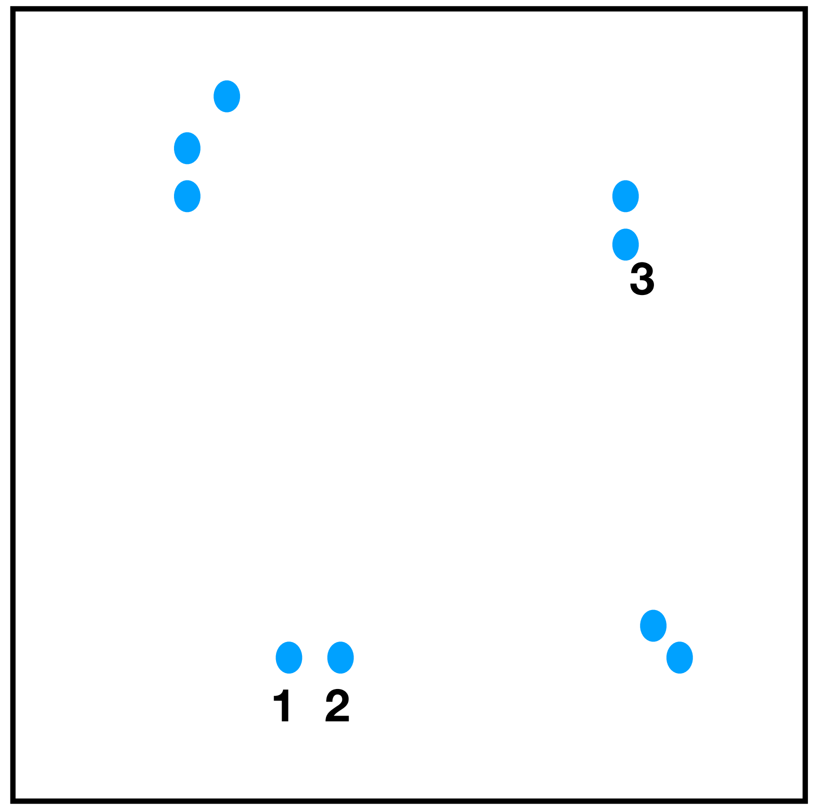
Find 1NN in W space

An intuitive estimator

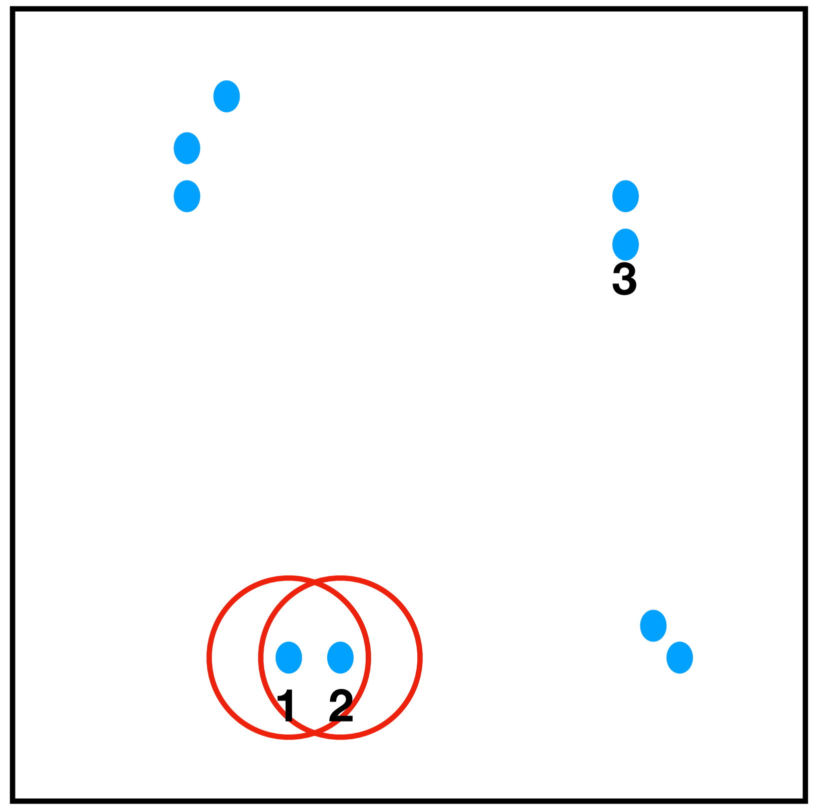
Nearest neighbor replacement: generate k_n nearest neighbor graph with data $(W_1, \ldots, W_n)$. $\mathcal{N}(i)$ is the k_n nearest neighbors set of $\mathbf{W}_i$.

$$\mathbb{E}_{\mathbf{W}_i}[\mathbb{E}[H((\mathbf{X}, \mathbf{Y}), (\mathbf{X}', \mathbf{Y}')) \mid \mathbf{W}_i]] \approx \frac{1}{n} \sum_{i=1}^{n} \frac{1}{k_n} \sum_{j \in \mathcal{N}(i)} H((X_i, Y_i), (X_j, Y_j))$$

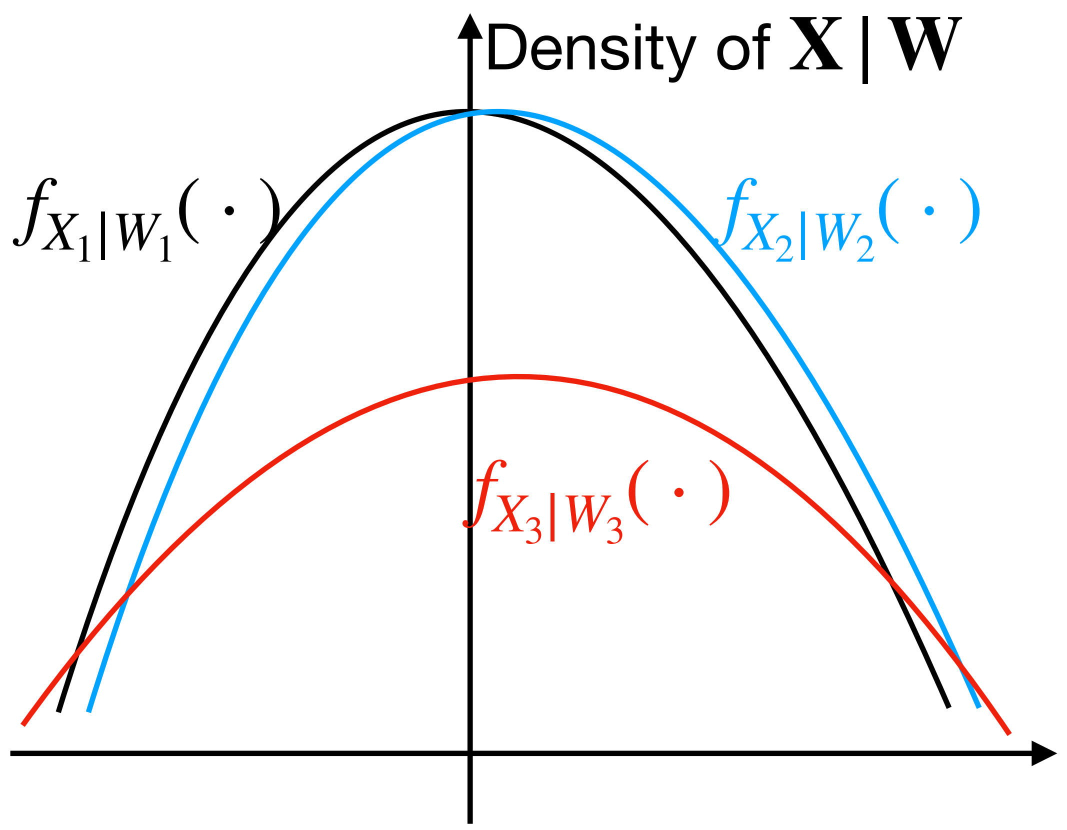
Intuition behind the estimator: $k_n = 1,\ n = 9.$



$W_1, \ldots, W_9$



Find 1NN in W space

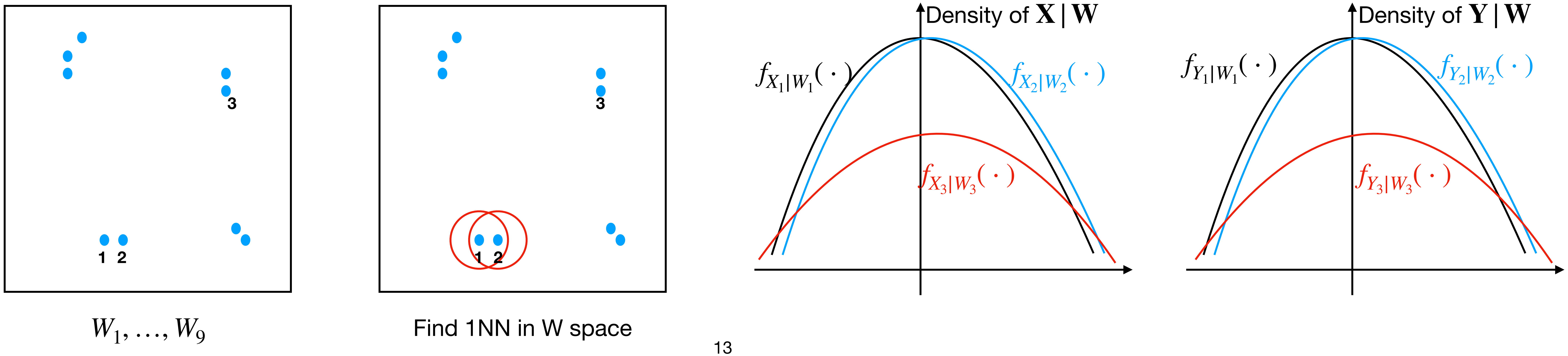


An intuitive estimator

Nearest neighbor replacement: generate k_n nearest neighbor graph with data $(W_1, \ldots, W_n)$. $\mathcal{N}(i)$ is the k_n nearest neighbors set of $\mathbf{W}_i$.

$$\mathbb{E}_{\mathbf{W}_i}[\mathbb{E}[H((\mathbf{X}, \mathbf{Y}), (\mathbf{X}', \mathbf{Y}')) \mid \mathbf{W}_i]] \approx \frac{1}{n}\sum_{i=1}^{n}\frac{1}{k_n}\sum_{j\in\mathcal{N}(i)} H((X_i, Y_i), (X_j, Y_j))$$

Intuition behind the estimator: $k_n = 1,\ n = 9.$



Unbiasedness and asymptotic consistency

Unbiasedness and asymptotic consistency

Recall

Unbiasedness and asymptotic consistency

Recall

$$T = \frac{1}{n}\sum_{i=1}^{n}\frac{1}{k_n}\sum_{j\in\mathcal{N}(i)} H((\mathbf{X}_i, \mathbf{Y}_i), (\mathbf{X}_j, \mathbf{Y}_j))$$

Unbiasedness and asymptotic consistency

Recall

$$T = \frac{1}{n}\sum_{i=1}^{n}\frac{1}{k_n}\sum_{j\in\mathcal{N}(i)} H((\mathbf{X}_i, \mathbf{Y}_i), (\mathbf{X}_j, \mathbf{Y}_j))$$

Unbiasedness and asymptotic consistency

Recall

$$T = \frac{1}{n} \sum_{i=1}^{n} \frac{1}{k_n} \sum_{j \in \mathcal{N}(i)} H((\mathbf{X}_i, \mathbf{Y}_i), (\mathbf{X}_j, \mathbf{Y}_j))$$

Theorem (informal):

Unbiasedness and asymptotic consistency

Recall

$$T = \frac{1}{n} \sum_{i=1}^{n} \frac{1}{k_n} \sum_{j \in \mathcal{N}(i)} H((\mathbf{X}_i, \mathbf{Y}_i), (\mathbf{X}_j, \mathbf{Y}_j))$$

Theorem (informal):

- Under H_0, $\mathbb{E}[T] = 0$.

Unbiasedness and asymptotic consistency

Recall

$$T = \frac{1}{n} \sum_{i=1}^{n} \frac{1}{k_n} \sum_{j \in \mathcal{N}(i)} H((\mathbf{X}_i, \mathbf{Y}_i), (\mathbf{X}_j, \mathbf{Y}_j))$$

Theorem (informal):

- Under H_0, $\mathbb{E}[T] = 0$.
- Under mild conditions, $T \xrightarrow{\mathbb{P}} \mathrm{ECMMD}^2(\mathbf{X}, \mathbf{Y} \mid \mathbf{W})$ if $k_n = o(n/\log(n))$.

Asymptotic behavior under null

Asymptotic behavior under null

Recall

Asymptotic behavior under null

Recall

$$T = \frac{1}{n}\sum_{i=1}^{n}\frac{1}{k_n}\sum_{j\in\mathcal{N}(i)} H((\mathbf{X}_i, \mathbf{Y}_i), (\mathbf{X}_j, \mathbf{Y}_j))$$

Asymptotic behavior under null

Recall

$$T = \frac{1}{n}\sum_{i=1}^{n}\frac{1}{k_n}\sum_{j\in\mathcal{N}(i)} H((\mathbf{X}_i, \mathbf{Y}_i), (\mathbf{X}_j, \mathbf{Y}_j))$$

Asymptotic behavior under null

Recall

$$T = \frac{1}{n} \sum_{i=1}^{n} \frac{1}{k_n} \sum_{j \in \mathcal{N}(i)} H((\mathbf{X}_i, \mathbf{Y}_i), (\mathbf{X}_j, \mathbf{Y}_j))$$

Theorem (informal): Under H_0, we have

Asymptotic behavior under null

Recall

$$T = \frac{1}{n} \sum_{i=1}^{n} \frac{1}{k_n} \sum_{j \in \mathcal{N}(i)} H((\mathbf{X}_i, \mathbf{Y}_i), (\mathbf{X}_j, \mathbf{Y}_j))$$

Theorem (informal): Under H_0, we have

$$\frac{\sqrt{nk_n}T}{\hat{\sigma}_n} \to N(0,1), \text{ if } k_n = o(n^\delta) \text{ for some small } \delta > 0.$$

Asymptotic behavior under null

Recall

$$T = \frac{1}{n} \sum_{i=1}^{n} \frac{1}{k_n} \sum_{j \in \mathcal{N}(i)} H((\mathbf{X}_i, \mathbf{Y}_i), (\mathbf{X}_j, \mathbf{Y}_j))$$

Theorem (informal): Under H_0, we have

$$\frac{\sqrt{nk_n}T}{\hat{\sigma}_n} \to N(0,1), \text{ if } k_n = o(n^\delta) \text{ for some small } \delta > 0.$$

where $\hat{\sigma}_n^2$ is a variance estimate.

Asymptotic behavior under null

Recall

$$T = \frac{1}{n}\sum_{i=1}^{n}\frac{1}{k_n}\sum_{j\in\mathcal{N}(i)} H((\mathbf{X}_i, \mathbf{Y}_i), (\mathbf{X}_j, \mathbf{Y}_j))$$

Theorem (informal): Under H_0, we have

$$\frac{\sqrt{nk_n}T}{\hat{\sigma}_n} \to N(0,1), \text{ if } k_n = o(n^\delta) \text{ for some small } \delta > 0.$$

where $\hat{\sigma}_n^2$ is a variance estimate.

Highly non-trivial proof:

Asymptotic behavior under null

Recall

$$T = \frac{1}{n}\sum_{i=1}^{n}\frac{1}{k_n}\sum_{j\in\mathcal{N}(i)} H((\mathbf{X}_i,\mathbf{Y}_i),(\mathbf{X}_j,\mathbf{Y}_j))$$

Theorem (informal): Under H_0, we have

$$\frac{\sqrt{nk_n}T}{\hat{\sigma}_n} \to N(0,1), \text{ if } k_n = o(n^\delta) \text{ for some small } \delta > 0.$$

where $\hat{\sigma}_n^2$ is a variance estimate.

Highly non-trivial proof:

Stein's method for dependency graph + dedicate analysis on $\hat{\sigma}_n$!

Property of the test

Property of the test

A consistent test:

Property of the test

A consistent test:

$$\phi \equiv \mathbf{1}\{\,|\sqrt{nk_n}T/\hat{\sigma}_n| \geq z_{1-\alpha/2}\},\ \limsup_{n\to\infty} \mathbb{E}_{H_0}[\phi] \leq \alpha, \liminf_{n\to\infty} \mathbb{E}_{H_1}[\phi] = 1$$

Property of the test

A consistent test:

$$\phi \equiv \mathbf{1}\{\,|\sqrt{nk_n}T/\hat{\sigma}_n| \geq z_{1-\alpha/2}\},\ \limsup_{n\to\infty} \mathbb{E}_{H_0}[\phi] \leq \alpha, \liminf_{n\to\infty} \mathbb{E}_{H_1}[\phi] = 1$$

Computation efficiency: linear comp time in n if k_n is a constant

Property of the test

A consistent test:

$$\phi \equiv \mathbf{1}\{\,|\sqrt{nk_n}T/\hat{\sigma}_n| \geq z_{1-\alpha/2}\},\;\; \limsup_{n\to\infty} \mathbb{E}_{H_0}[\phi] \leq \alpha, \liminf_{n\to\infty} \mathbb{E}_{H_1}[\phi] = 1$$

Computation efficiency: linear comp time in n if k_n is a constant

Easy calibration: no need to do resampling

Property of the test

A consistent test:

$$\phi \equiv \mathbf{1}\{\,|\sqrt{nk_n}T/\hat{\sigma}_n| \geq z_{1-\alpha/2}\},\;\; \limsup_{n\to\infty} \mathbb{E}_{H_0}[\phi] \leq \alpha, \liminf_{n\to\infty} \mathbb{E}_{H_1}[\phi] = 1$$

Computation efficiency: linear comp time in n if k_n is a constant

Easy calibration: no need to do resampling

Agnostic to hyperparameter: no rate lower bound on k_n

Property of the test

A consistent test:

$$\phi \equiv \mathbf{1}\{ |\sqrt{nk_n} T/\hat{\sigma}_n| \geq z_{1-\alpha/2}\}, \ \limsup_{n\to\infty} \mathbb{E}_{H_0}[\phi] \leq \alpha, \liminf_{n\to\infty} \mathbb{E}_{H_1}[\phi] = 1$$

Computation efficiency: linear comp time in n if k_n is a constant

Easy calibration: no need to do resampling

Agnostic to hyperparameter: no rate lower bound on k_n

This is not the end of the story!

What has been ignored?

What has been ignored?

Recall the ECMMD test statistic construction:

What has been ignored?

Recall the ECMMD test statistic construction:

- **Given** $(Y_1, W_1), \ldots, (Y_n, W_n)$ **and predictive distribution** $\mathrm{Bern}(\hat{f}(\mathbf{Z}))$**;**

What has been ignored?

Recall the ECMMD test statistic construction:

- **Given** $(Y_1, W_1), \ldots, (Y_n, W_n)$ **and predictive distribution** $\mathrm{Bern}(\hat{f}(\mathbf{Z}))$**;**
- **Sample** $(X_i, W_i) \sim \mathrm{Bern}(W_i),\ W_i = \hat{f}(Z_i)$**;**

What has been ignored?

Recall the ECMMD test statistic construction:

- **Given** $(Y_1, W_1), \ldots, (Y_n, W_n)$ **and predictive distribution** $\mathrm{Bern}(\hat{f}(\mathbf{Z}))$**;**
- **Sample** $(X_i, W_i) \sim \mathrm{Bern}(W_i),\ W_i = \hat{f}(\mathrm{Z}_i)$**;**
- **Compute the test statistic** T **with** $(X_1, Y_1, W_1), \ldots, (X_n, Y_n, W_n)$ **and standard deviation estimate** $\hat{\sigma}_n$**.**

What has been ignored?

Recall the ECMMD test statistic construction:

- **Given** $(Y_1, W_1), \ldots, (Y_n, W_n)$ **and predictive distribution** $\mathrm{Bern}(\hat{f}(\mathbf{Z}))$**;**
- **Sample** $(X_i, W_i) \sim \mathrm{Bern}(W_i),\ W_i = \hat{f}(Z_i)$**;**
- **Compute the test statistic** T **with** $(X_1, Y_1, W_1), \ldots, (X_n, Y_n, W_n)$ **and standard deviation estimate** $\hat{\sigma}_n$**.**

Sampling $X_i \sim \mathbb{P}_{\mathbf{X}_i|\mathbf{W}_i}$ **will induce a random test!**

Reduce randomness with derandomized test

Reduce randomness with derandomized test

Reduce randomness with derandomized test

1. Given $(Y_i, W_i), i = 1,\ldots,n$. Construct the nearest neighbor graph using $W_1, \ldots, W_n$;

Reduce randomness with derandomized test

1. Given $(Y_i, W_i), i = 1, \ldots, n$. Construct the nearest neighbor graph using $W_1, \ldots, W_n$;

2. Get M_n samples $(\widetilde{X}_i^{(1)}, W_i)_{i=1,\ldots,n}, \ldots, (\widetilde{X}_i^{(M_n)}, W_i)_{i=1,\ldots,n}$ from $\mathbb{P}_{X_i|W_i}$

Reduce randomness with derandomized test

1. Given $(Y_i, W_i), i = 1,\ldots,n$. Construct the nearest neighbor graph using $W_1, \ldots, W_n$;

2. Get M_n samples $(\widetilde{X}_i^{(1)}, W_i)_{i=1,\ldots,n}, \ldots, (\widetilde{X}_i^{(M_n)}, W_i)_{i=1,\ldots,n}$ from $\mathbb{P}_{X_i|W_i}$

$$T^{(m)} \equiv \frac{1}{n}\sum_{i=1}^{n}\frac{1}{k_n}\sum_{j\in\mathcal{N}(i)} H((Y_i, \widetilde{X}_i^{(m)}), (Y_j, \widetilde{X}_j^{(m)})), \; \widetilde{T} = \frac{1}{M_n}\sum_{m=1}^{M_n} T^{(m)};$$

Reduce randomness with derandomized test

1. Given $(Y_i, W_i), i = 1,\ldots,n$. Construct the nearest neighbor graph using $W_1, \ldots, W_n$;

2. Get M_n samples $(\widetilde{X}_i^{(1)}, W_i)_{i=1,\ldots,n}, \ldots, (\widetilde{X}_i^{(M_n)}, W_i)_{i=1,\ldots,n}$ from $\mathbb{P}_{X_i|W_i}$

$$T^{(m)} \equiv \frac{1}{n}\sum_{i=1}^{n}\frac{1}{k_n}\sum_{j\in\mathcal{N}(i)} H((Y_i, \widetilde{X}_i^{(m)}), (Y_j, \widetilde{X}_j^{(m)})), \widetilde{T} = \frac{1}{M_n}\sum_{m=1}^{M_n} T^{(m)};$$

3. Return the test statistic $\widetilde{T}/\widetilde{\sigma}$ with standard deviation estimate $\widetilde{\sigma}$;

Reduce randomness with derandomized test

1. Given $(Y_i, W_i), i = 1,\ldots,n$. Construct the nearest neighbor graph using $W_1, \ldots, W_n$;

2. Get M_n samples $(\widetilde{X}_i^{(1)}, W_i)_{i=1,\ldots,n}, \ldots, (\widetilde{X}_i^{(M_n)}, W_i)_{i=1,\ldots,n}$ from $\mathbb{P}_{X_i|W_i}$

$$T^{(m)} \equiv \frac{1}{n}\sum_{i=1}^{n}\frac{1}{k_n}\sum_{j\in\mathcal{N}(i)} H((Y_i, \widetilde{X}_i^{(m)}), (Y_j, \widetilde{X}_j^{(m)})), \widetilde{T} = \frac{1}{M_n}\sum_{m=1}^{M_n} T^{(m)};$$

3. Return the test statistic $\widetilde{T}/\widetilde{\sigma}$ with standard deviation estimate $\widetilde{\sigma}$;

Reduce randomness with derandomized test

1. Given $(Y_i, W_i), i = 1, \ldots, n$. Construct the nearest neighbor graph using $W_1, \ldots, W_n$;

2. Get M_n samples $(\widetilde{X}_i^{(1)}, W_i)_{i=1,\ldots,n}, \ldots, (\widetilde{X}_i^{(M_n)}, W_i)_{i=1,\ldots,n}$ from $\mathbb{P}_{X_i|W_i}$

$$T^{(m)} \equiv \frac{1}{n}\sum_{i=1}^{n}\frac{1}{k_n}\sum_{j\in\mathcal{N}(i)} H((Y_i, \widetilde{X}_i^{(m)}), (Y_j, \widetilde{X}_j^{(m)})), \widetilde{T} = \frac{1}{M_n}\sum_{m=1}^{M_n} T^{(m)};$$

3. Return the test statistic $\widetilde{T}/\widetilde{\sigma}$ with standard deviation estimate $\widetilde{\sigma}$;

Theorem (informal): Under H_0, as long as $M_n \to \infty$ at any rate we have

Reduce randomness with derandomized test

1. Given $(Y_i, W_i), i = 1,\ldots,n$. Construct the nearest neighbor graph using $W_1, \ldots, W_n$;

2. Get M_n samples $(\widetilde{X}_i^{(1)}, W_i)_{i=1,\ldots,n}, \ldots, (\widetilde{X}_i^{(M_n)}, W_i)_{i=1,\ldots,n}$ from $\mathbb{P}_{X_i|W_i}$

$$T^{(m)} \equiv \frac{1}{n}\sum_{i=1}^{n}\frac{1}{k_n}\sum_{j\in\mathcal{N}(i)} H((Y_i, \widetilde{X}_i^{(m)}), (Y_j, \widetilde{X}_j^{(m)})), \widetilde{T} = \frac{1}{M_n}\sum_{m=1}^{M_n} T^{(m)};$$

3. Return the test statistic $\widetilde{T}/\widetilde{\sigma}$ with standard deviation estimate $\widetilde{\sigma}$;

Theorem (informal): Under H_0, as long as $M_n \to \infty$ at any rate we have

$$\sqrt{nk_n}\widetilde{T}/\widetilde{\sigma}_n \to N(0,1), \text{ if } k_n = o(n^\delta) \text{ for some small } \delta > 0.$$

Numerical simulation: statistical efficiency

Numerical simulation: statistical efficiency

Classification calibration

Numerical simulation: statistical efficiency

Classification calibration

$$(W_i, 1 - W_i) \overset{iid}{\sim} \text{Dirichlet}(\rho),\ \rho \in \{0.1, 0.2, 0.3, 0.4, 0.5\}$$

Numerical simulation: statistical efficiency

Classification calibration

$(W_i, 1 - W_i) \overset{iid}{\sim} \text{Dirichlet}(\rho),\ \rho \in \{0.1, 0.2, 0.3, 0.4, 0.5\}$

$n = 100,\ k_n \in \{15, 25\},\ M_n = 100$

Numerical simulation: statistical efficiency

Classification calibration

$(W_i, 1 - W_i) \overset{iid}{\sim} \text{Dirichlet}(\rho),\ \rho \in \{0.1, 0.2, 0.3, 0.4, 0.5\}$

$n = 100,\ k_n \in \{15, 25\},\ M_n = 100$

Null

Numerical simulation: statistical efficiency

Classification calibration

$(W_i, 1 - W_i) \overset{iid}{\sim} \text{Dirichlet}(\rho),\ \rho \in \{0.1, 0.2, 0.3, 0.4, 0.5\}$

$n = 100,\ k_n \in \{15, 25\},\ M_n = 100$

Null

$Y_i \sim \text{Bern}(W_i),\ X_i \sim \text{Bern}(W_i)$

Numerical simulation: statistical efficiency

Classification calibration

$(W_i, 1 - W_i) \overset{iid}{\sim} \text{Dirichlet}(\rho),\ \rho \in \{0.1, 0.2, 0.3, 0.4, 0.5\}$

$n = 100,\ k_n \in \{15, 25\},\ M_n = 100$

Null

$Y_i \sim \text{Bern}(W_i),\ X_i \sim \text{Bern}(W_i)$

Alternative

Numerical simulation: statistical efficiency

Classification calibration

$(W_i, 1 - W_i) \overset{iid}{\sim} \text{Dirichlet}(\rho),\ \rho \in \{0.1, 0.2, 0.3, 0.4, 0.5\}$

$n = 100,\ k_n \in \{15, 25\},\ M_n = 100$

Null

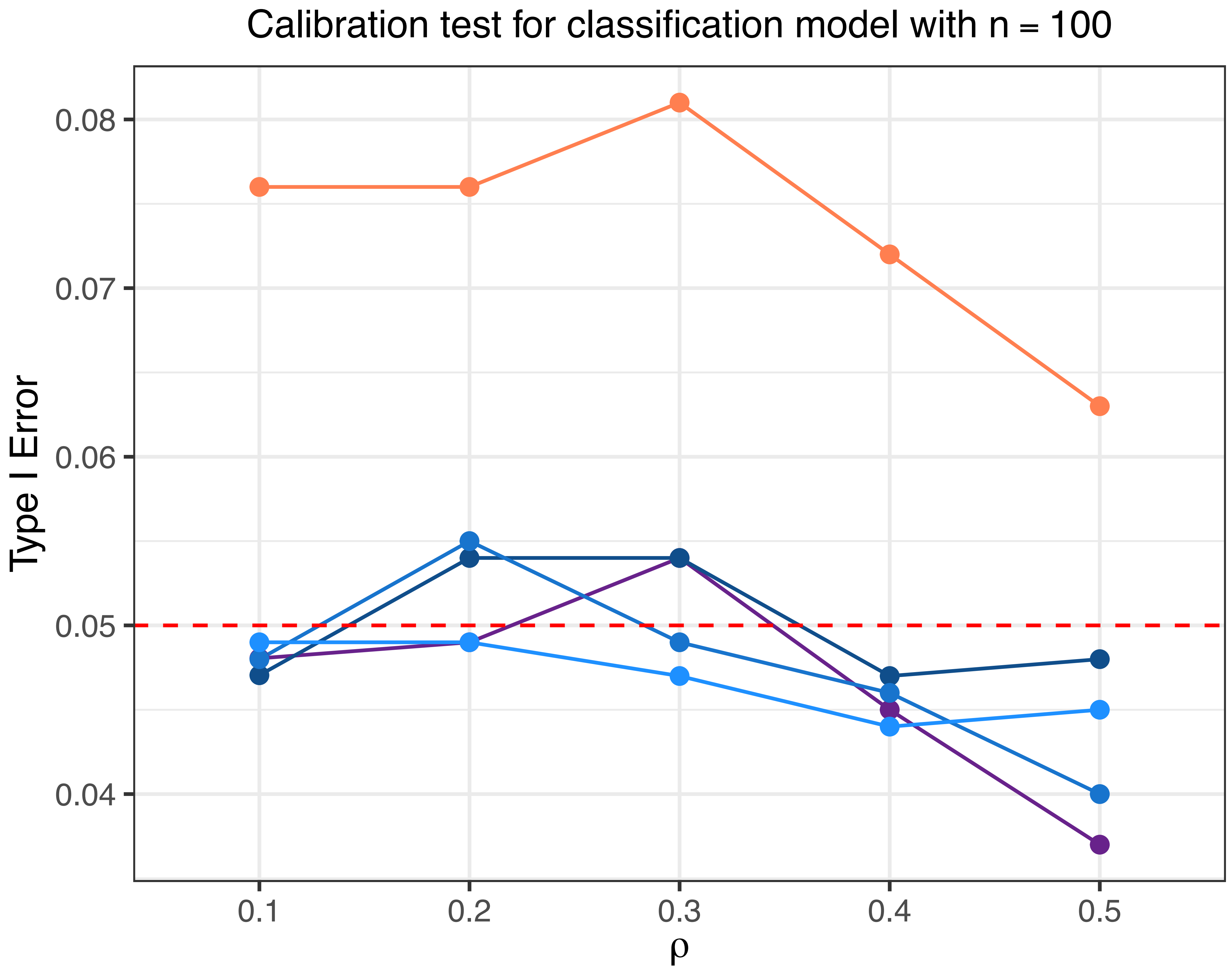
$Y_i \sim \text{Bern}(W_i),\ X_i \sim \text{Bern}(W_i)$

Alternative

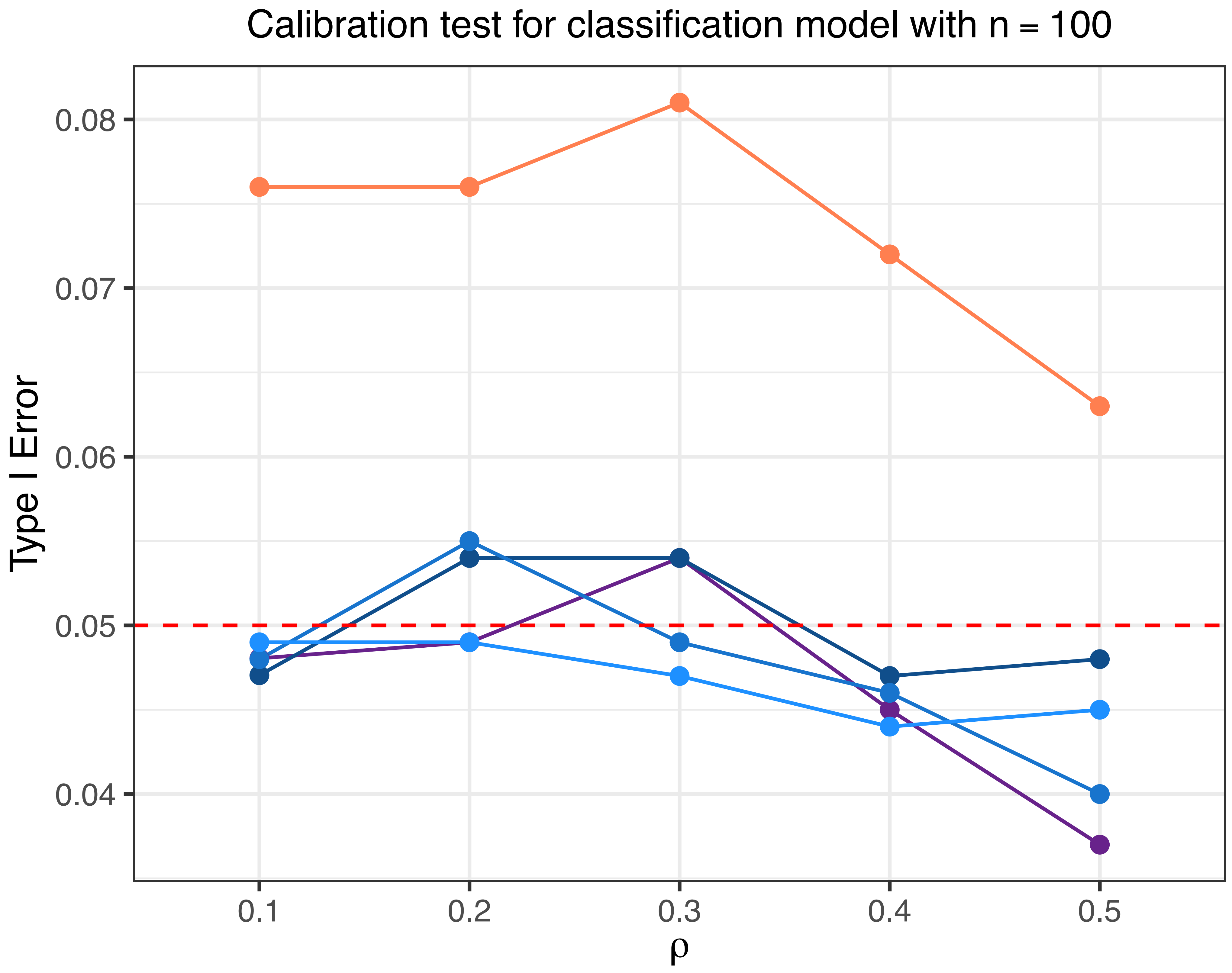
$Y_i \sim \text{Bern}(W_i - W_i^5),\ X_i \sim \text{Bern}(W_i)$

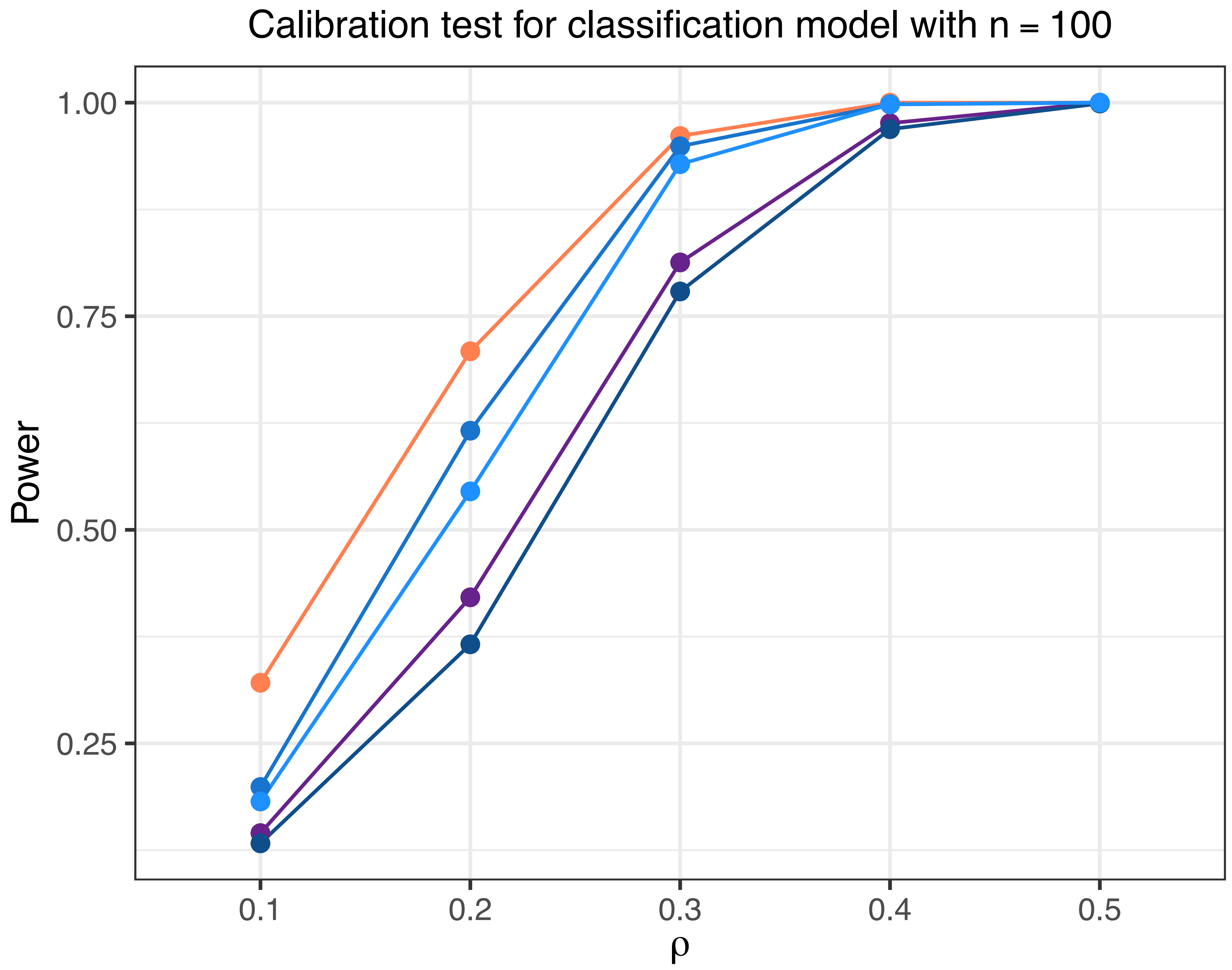
Classification calibration

Classification calibration

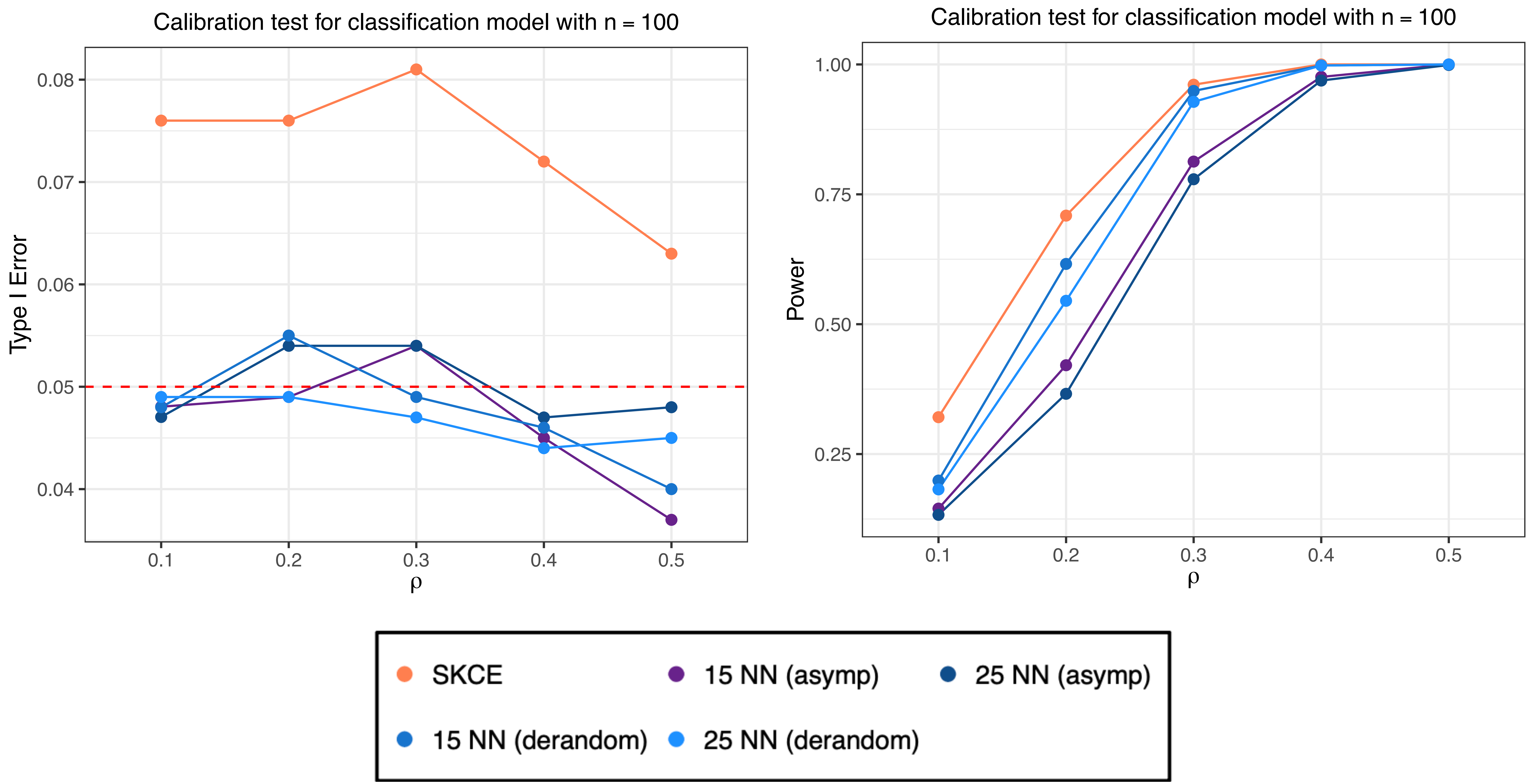


Classification calibration





Classification calibration



Discussion

Discussion

Take-home messages:

Discussion

Take-home messages:

- Formulate the model calibration test to a conditional two-sample problem and bridge the classical inference literature with calibration problem;

Discussion

Take-home messages:

- Formulate the model calibration test to a conditional two-sample problem and bridge the classical inference literature with calibration problem;
- Nearest neighbor-based test has statistical and computational advantages;

Discussion

Take-home messages:

- Formulate the model calibration test to a conditional two-sample problem and bridge the classical inference literature with calibration problem;
- Nearest neighbor-based test has statistical and computational advantages;
- Derandomization is beneficial for the power of the test;

Discussion

Take-home messages:

- Formulate the model calibration test to a conditional two-sample problem and bridge the classical inference literature with calibration problem;
- Nearest neighbor-based test has statistical and computational advantages;
- Derandomization is beneficial for the power of the test;

Open questions:

Discussion

Take-home messages:

- Formulate the model calibration test to a conditional two-sample problem and bridge the classical inference literature with calibration problem;
- Nearest neighbor-based test has statistical and computational advantages;
- Derandomization is beneficial for the power of the test;

Open questions:

- Is the proposed test powerful against local alternatives?

Discussion

Take-home messages:

- Formulate the model calibration test to a conditional two-sample problem and bridge the classical inference literature with calibration problem;
- Nearest neighbor-based test has statistical and computational advantages;
- Derandomization is beneficial for the power of the test;

Open questions:

- Is the proposed test powerful against local alternatives?
- What if there are multiple candidate models?

Discussion

Take-home messages:

- Formulate the model calibration test to a conditional two-sample problem and bridge the classical inference literature with calibration problem;
- Nearest neighbor-based test has statistical and computational advantages;
- Derandomization is beneficial for the power of the test;

Open questions:

- Is the proposed test powerful against local alternatives?
- What if there are multiple candidate models?
- High-stakes application with the proposed method?

Thank you!

Thank you!

Questions?