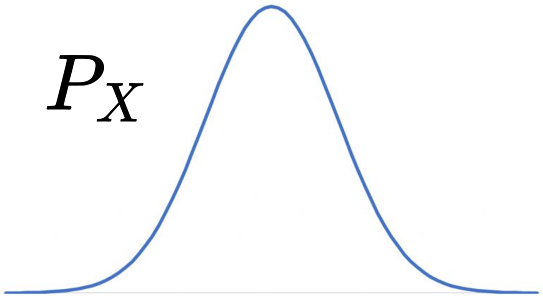
Testing for Outliers with Conformal p-values

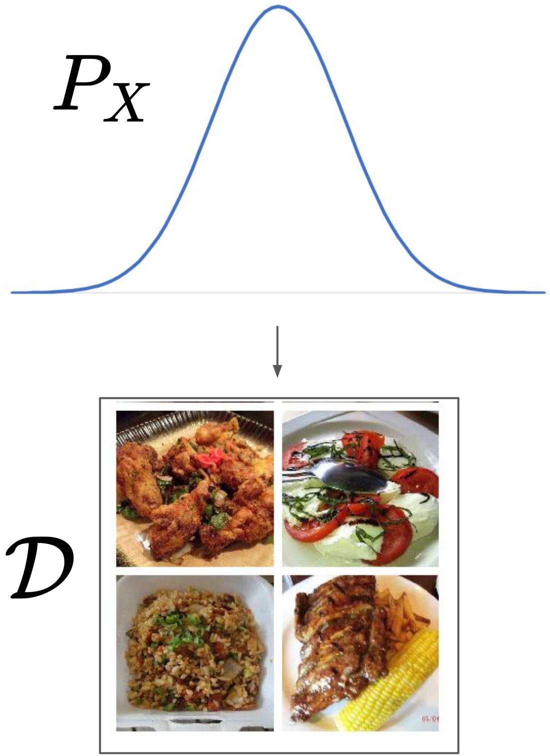
Authors: S. Bates, E. Candes, L. Lei, Y. Romano, M. Sesia

Donghwan Lee

Out-of-distribution detection

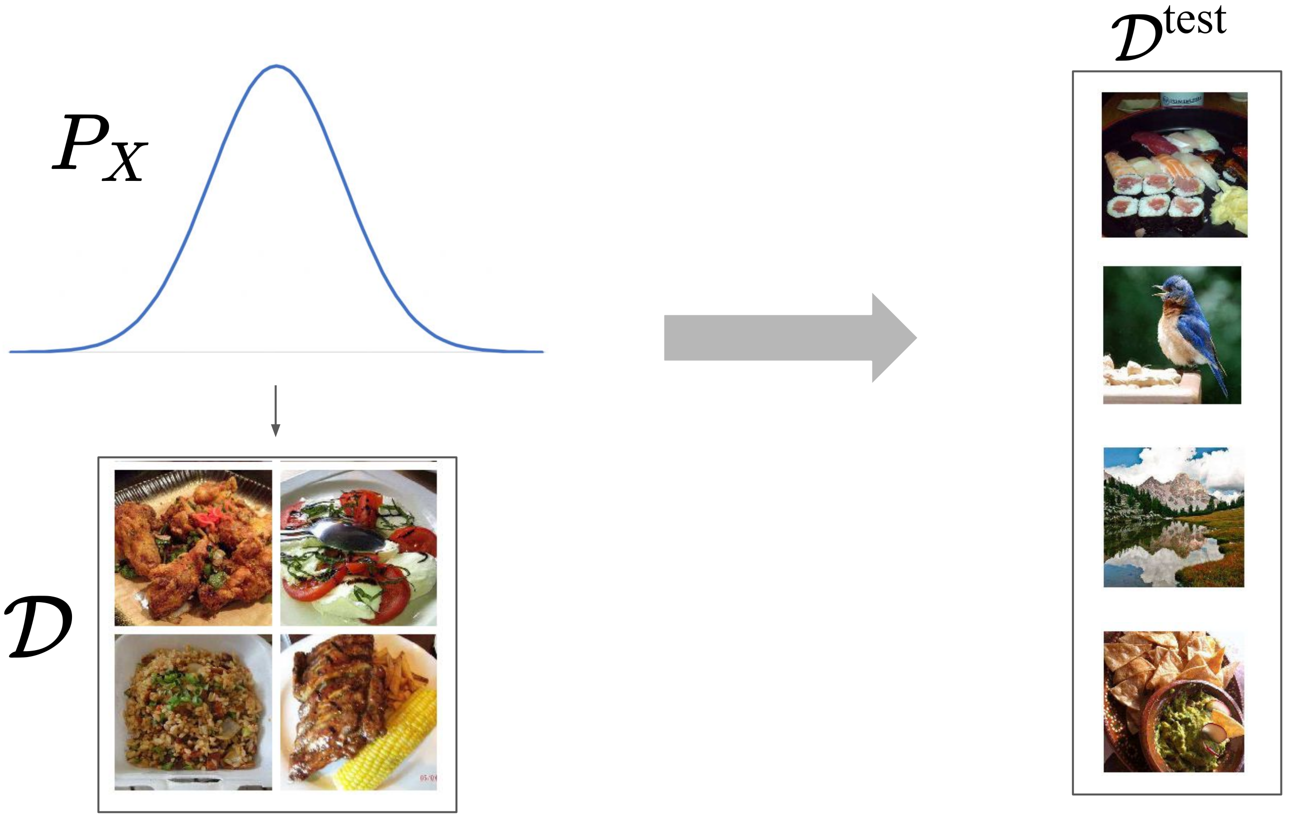


Out-of-distribution detection



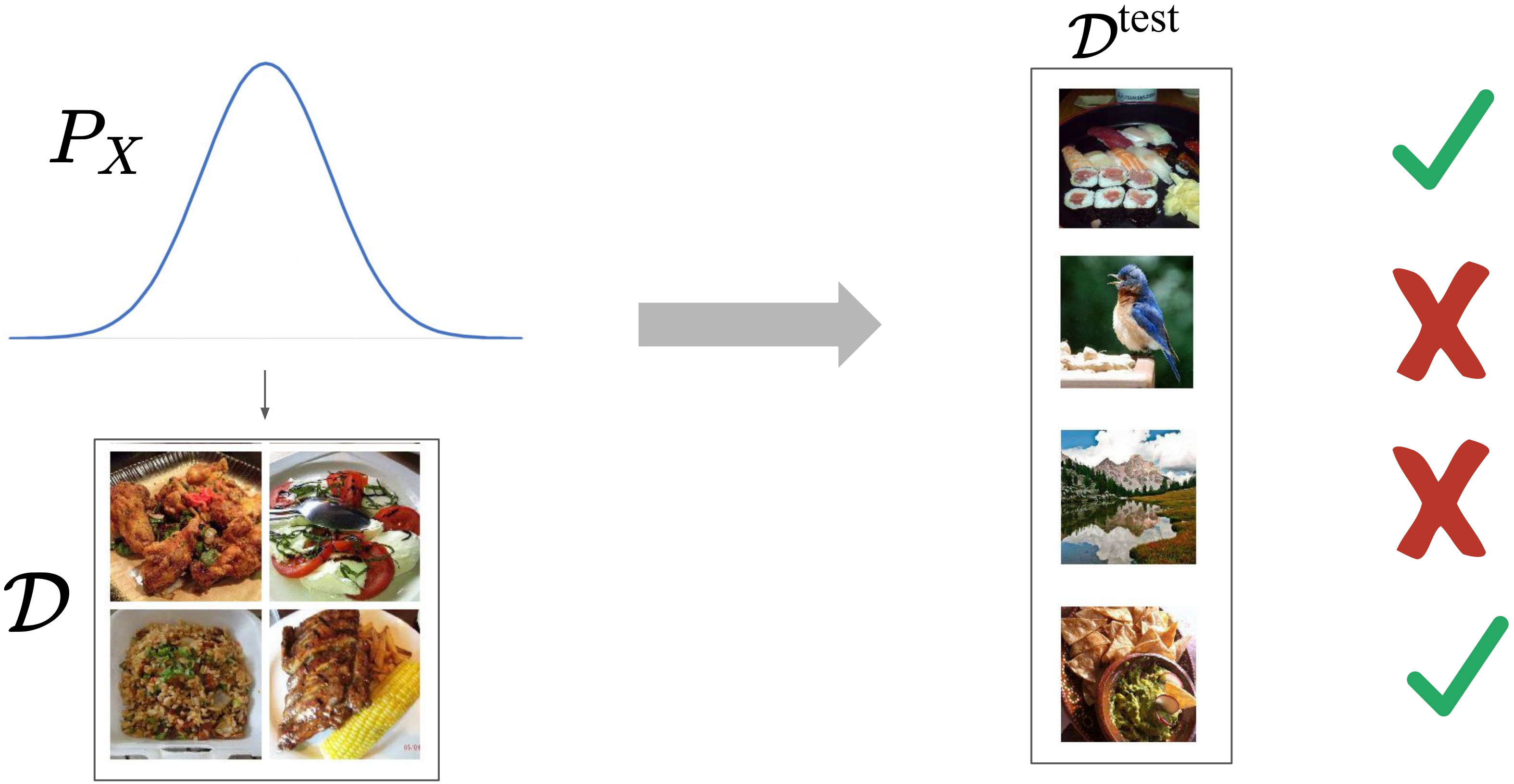
Out-of-distribution detection

Test time



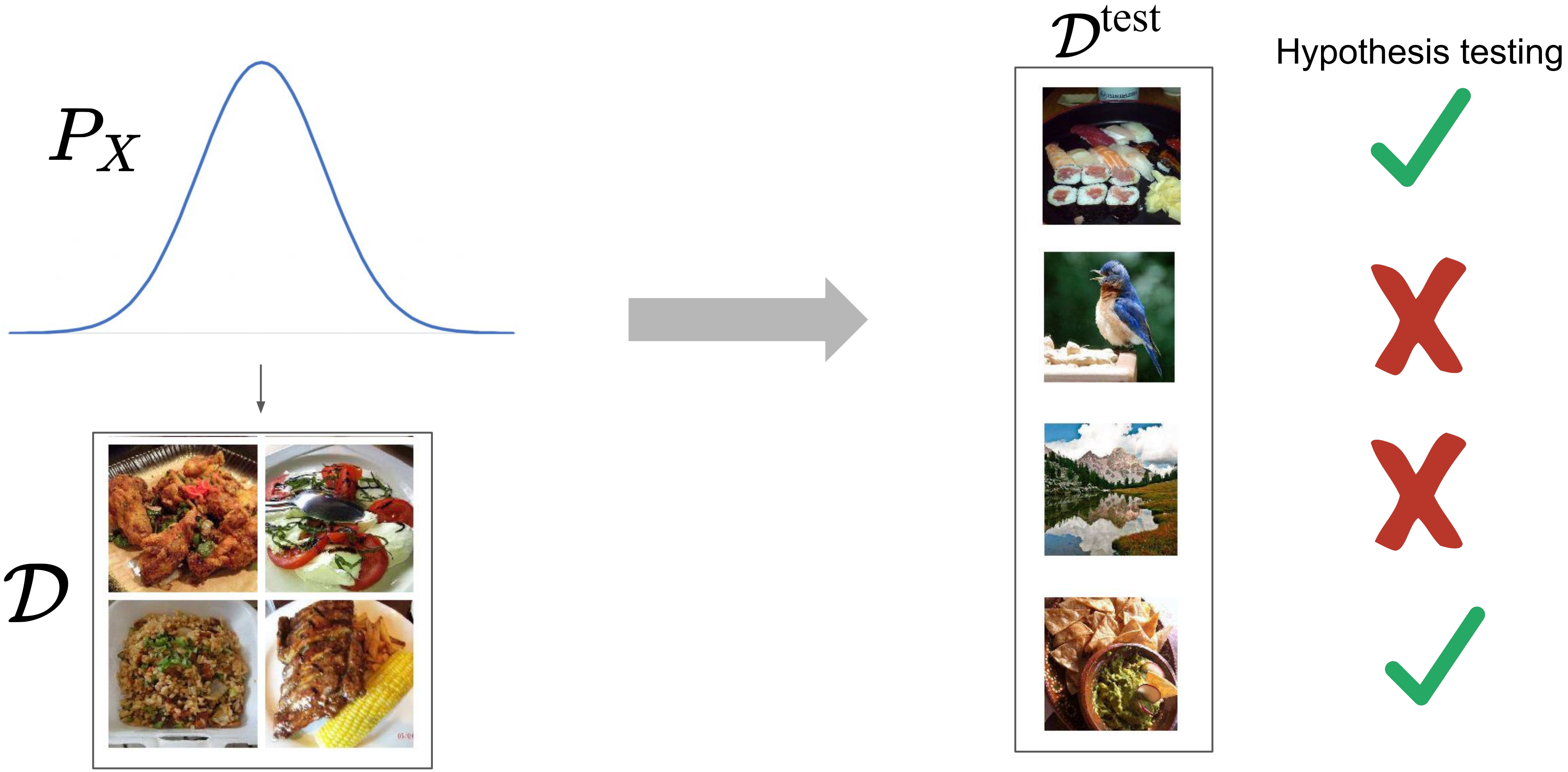
Out-of-distribution detection

Test time



Out-of-distribution detection

Test time



Previous works

Supervised

- [Lee et al., 2018] Mahalanobis distance-based score
- [Lee et al., 2018] GAN-based training

Self-supervised

- [Golan et al., 2018]
- [Bergman et al., 2020]
- [Hendricks et al., 2019]

Unsupervised

- [Liang et al., 2017] ODIN (Out-of-DIstribution detector for Neural networks)
- [Macedo et al., 2021] isomax scores
- [Mahmood et al., 2021] density estimates

Previous works

Supervised

- [Lee et al., 2018] Mahalanobis distance-based score
- [Lee et al., 2018] GAN-based training

Self-supervised

- [Golan et al., 2018]
- [Bergman et al., 2020]
- [Hendricks et al., 2019]

Unsupervised

- [Liang et al., 2017] ODIN (Out-of-DIstribution detector for Neural networks)
- [Macedo et al., 2021] isomax scores
- [Mahmood et al., 2021] density estimates

No statistical guarantees!

Goal

Given a score function $\hat{s} : \mathcal{X} \to \mathbb{R}$ and a calibration set $\mathcal{D} \sim P_X$, build statistical "wrappers" $\hat{u}$ such that for a test point $X \sim P_X$

Goal

Given a score function $\hat{s} : \mathcal{X} \to \mathbb{R}$ and a calibration set $\mathcal{D} \sim P_X$, build statistical "wrappers" $\hat{u}$ such that for a test point $X \sim P_X$

- Marginal validity:

$$\mathbb{P}[\hat{u}^{(\mathrm{marg})}(X) \leq t] \leq t \text{ for all } t \in (0,1)$$

Goal

Given a score function $\hat{s} : \mathcal{X} \to \mathbb{R}$ and a calibration set $\mathcal{D} \sim P_X$, build statistical "wrappers" $\hat{u}$ such that for a test point $X \sim P_X$

- Marginal validity:

$$\mathbb{P}[\hat{u}^{(\mathrm{marg})}(X) \leq t] \leq t \text{ for all } t \in (0,1)$$

- Calibration-conditional validity:

$$\mathbb{P}[\mathbb{P}[\hat{u}^{(\mathrm{ccv})}(X) \leq t \mid \mathcal{D}] \leq t \text{ for all } t \in (0,1)] \geq 1-\delta$$

Goal

Given a score function $\hat{s} : \mathcal{X} \to \mathbb{R}$ and a calibration set $\mathcal{D} \sim P_X$, build statistical "wrappers" $\hat{u}$ such that for a test point $X \sim P_X$

- Marginal validity:

$$\mathbb{P}[\hat{u}^{(\mathrm{marg})}(X) \leq t] \leq t \text{ for all } t \in (0,1)$$

- Calibration-conditional validity:

$$\mathbb{P}[\mathbb{P}[\hat{u}^{(\mathrm{ccv})}(X) \leq t \mid \mathcal{D}] \leq t \text{ for all } t \in (0,1)] \geq 1 - \delta$$

- Multiple testing

Setup

$$\mathcal{D}^{\text{train}} = \{X_1, \ldots, X_n\} \overset{i.i.d.}{\sim} P_X$$

Setup

$$\mathcal{D}^{\text{train}} = \{X_1, \ldots, X_n\} \overset{i.i.d.}{\sim} P_X \longrightarrow \hat{\boldsymbol{s}} : \boldsymbol{\mathcal{X}} \to \mathbb{R}$$

Setup

$$\mathcal{D}^{\text{train}} = \{X_1, \ldots, X_n\} \overset{i.i.d.}{\sim} P_X \longrightarrow \quad \hat{s} : \mathcal{X} \to \mathbb{R}$$

$$\mathcal{D}^{\text{cal}} = \{X_{n+1}, \ldots, X_{2n}\} \overset{i.i.d.}{\sim} P_X$$

Setup

$$\mathcal{D}^{\text{train}} = \{X_1, \ldots, X_n\} \overset{i.i.d.}{\sim} P_X \longrightarrow \hat{s} : \mathcal{X} \to \mathbb{R}$$

$$\mathcal{D}^{\text{cal}} = \{X_{n+1}, \ldots, X_{2n}\} \overset{i.i.d.}{\sim} P_X$$

$$\mathcal{D}^{\text{test}} = \{X_{2n+1}, \ldots, X_{2n+m}\}$$

Null hypothesis: $\mathcal{H}_{0,i} : X_i \sim P_X \quad i \in \{2n+1, \ldots, 2n+m\}$

p-value

Motivation: $F(\hat{s}(X)) \sim \mathrm{Unif}([0,1])$

p-value

Motivation: $F(\hat{s}(X)) \sim \mathrm{Unif}([0,1])$

$$\hat{u}(X) = (g \circ \hat{F} \circ \hat{s})(X)$$

adjustment function

Empirical CDF of $\hat{s}(X)$

Marginal conformal p-value

Choose $g^{(\mathrm{marg})}(x) = \dfrac{nx+1}{n+1}$

Marginal conformal p-value

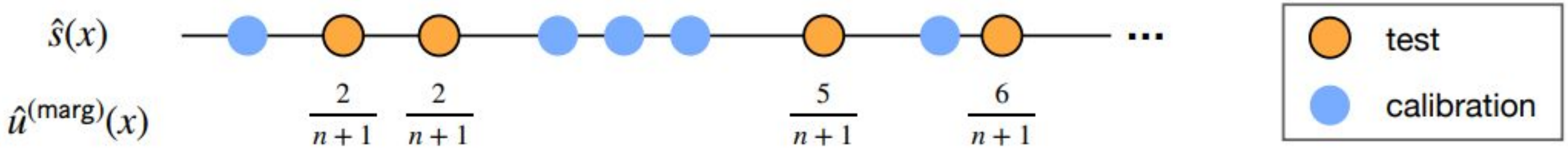
Choose $g^{(\mathrm{marg})}(x) = \dfrac{nx+1}{n+1}$

$$\hat{u}^{(\mathrm{marg})}(x) = \frac{1 + |\{i \in \mathcal{D}^{\mathrm{cal}} : \hat{s}(X_i) \leq \hat{s}(x)\}|}{n+1}$$

Marginal conformal p-value

Choose $g^{(\mathrm{marg})}(x) = \dfrac{nx+1}{n+1}$

$$\hat{u}^{(\mathrm{marg})}(x) = \frac{1 + |\{i \in \mathcal{D}^{\mathrm{cal}} : \hat{s}(X_i) \le \hat{s}(x)\}|}{n+1}$$



Marginal conformal p-value

(Assuming $\hat{s}(X)$ has a continuous distribution)

If $X \sim P_X$ is independent of $\mathcal{D}^{\text{cal}}$, then $u^{(\text{marg})}(X) \sim \text{Unif}(\{\frac{1}{n+1}, \frac{2}{n+1}, \ldots, 1\})$

Therefore,

$$\mathbb{P}[\hat{u}^{(\text{marg})}(X) \leq t] \leq t \text{ for all } t \in (0,1)$$

Multiple testing

Global null: $H_0 : X_{2n+1}, \ldots, X_{2n+m} \stackrel{i.i.d.}{\sim} P_X$

Multiple testing

Global null: $H_0 : X_{2n+1}, \ldots, X_{2n+m} \overset{i.i.d.}{\sim} P_X$

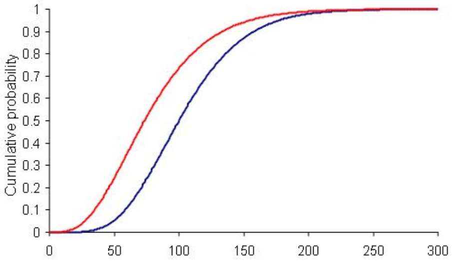
p-values: $p_i = \hat{u}^{(\text{marg})}(X_{2n+i}), i = 1, \ldots, m$

Multiple testing

Global null: $H_0 : X_{2n+1}, \ldots, X_{2n+m} \overset{i.i.d.}{\sim} P_X$

p-values: $p_i = \hat{u}^{(\mathrm{marg})}(X_{2n+i}), i = 1, \ldots, m$

Stochastic dominance: Under the global null,



$$\mathbb{P}[p_i \leq t] \leq t \text{ for all } t \in (0,1), i = 1, \ldots, m$$

Fisher's combination test

Fact: $p_i \overset{i.i.d.}{\sim} \mathrm{Unif}\left([0,1]\right), i = 1, \ldots, m$

$$\Rightarrow -2\sum_{i=1}^{m} \log p_i \sim \chi^2(2m)$$

$$\Rightarrow \mathbb{P}\left(-2\sum_{i=1}^{m} \log p_i \geq \chi^2(2m; 1-\alpha)\right) = \alpha$$

Fisher's combination test

cf. [Vovk et al., 2020], [Shafer et al., 2019]

Fact: $p_i \overset{i.i.d.}{\sim} \mathrm{Unif}([0,1]), i = 1, \ldots, m$

$$\Rightarrow -2 \sum_{i=1}^{m} \log p_i \sim \chi^2(2m)$$

$$\Rightarrow \mathbb{P}\left(-2 \sum_{i=1}^{m} \log p_i \geq \chi^2(2m; 1-\alpha)\right) = \alpha$$

Generalization: If p_i stochastically dominate $\mathrm{Unif}([0,1])$ and are independent of each other,

$$\mathbb{P}\left(-2 \sum_{i=1}^{m} \log p_i \geq \chi^2(2m; 1-\alpha)\right) \leq \alpha$$

Failure of Fisher’s combination test

Theorem 1 (Type-I error of Fisher’s combination test). *Assume that $\hat{s}(X)$ is continuous. Then, under the global null, if $m = \lfloor \gamma n \rfloor$ for some $\gamma \in (0, \infty)$, as n tends to infinity,*

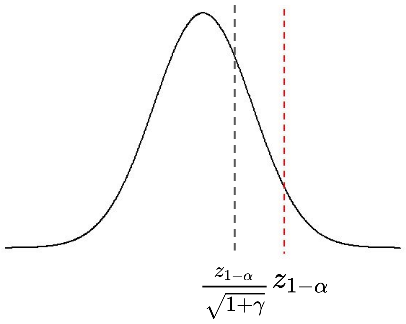
$$\mathbb{P}\left[-2\sum_{i=1}^{m} \log\left[\hat{u}^{(\mathrm{marg})}(X_{2n+i})\right] \geq \chi^2(2m; 1-\alpha)\right] \to \bar{\Phi}\left(\frac{z_{1-\alpha}}{\sqrt{1+\gamma}}\right),$$

where $z_{1-\alpha}$ and $\bar{\Phi}$ denote the $(1-\alpha)$-th quantile and tail function of the standard normal distribution, respectively. Furthermore, under the same asymptotic regime, for $W \sim N(0,1)$,

$$\mathbb{P}\left[-2\sum_{i=1}^{m} \log\left[\hat{u}^{(\mathrm{marg})}(X_{2n+i})\right] \geq \chi^2(2m; 1-\alpha) \mid \mathcal{D}\right] \xrightarrow{d} \bar{\Phi}(z_{1-\alpha} + \sqrt{\gamma} W). \tag{5}$$

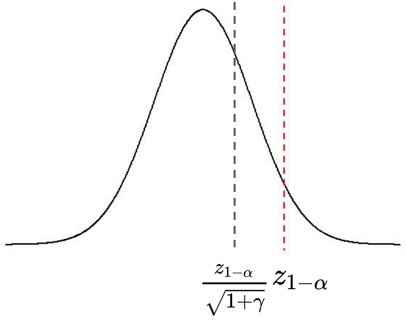
Failure of Fisher's combination test

$$\mathbb{P}\left[-2\sum_{i=1}^{m}\log\left[\hat{u}^{(\mathrm{marg})}(X_{2n+i})\right] \geq \chi^2(2m; 1-\alpha)\right] \to \bar{\Phi}\left(\frac{z_{1-\alpha}}{\sqrt{1+\gamma}}\right)$$

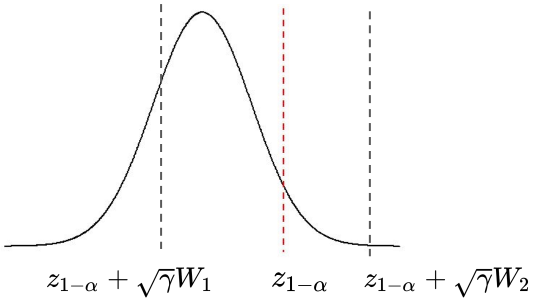


Failure of Fisher's combination test

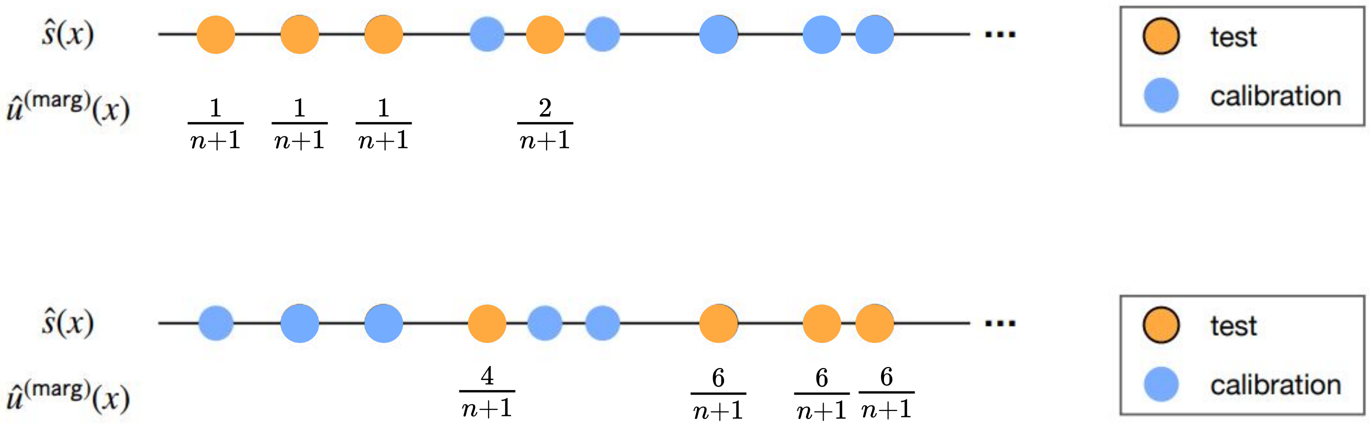
$$\mathbb{P}\left[-2\sum_{i=1}^{m}\log\left[\hat{u}^{(\mathrm{marg})}(X_{2n+i})\right] \geq \chi^2(2m;1-\alpha)\right] \to \bar{\Phi}\left(\frac{z_{1-\alpha}}{\sqrt{1+\gamma}}\right)$$



$$\mathbb{P}\left[-2\sum_{i=1}^{m}\log\left[\hat{u}^{(\mathrm{marg})}(X_{2n+i})\right] \geq \chi^2(2m;1-\alpha) \mid \mathcal{D}\right] \xrightarrow{d} \bar{\Phi}(z_{1-\alpha}+\sqrt{\gamma}W)$$



Conformal p-values are positively correlated



Conformal p-values are positively correlated

Lemma 1. *Assume that $\hat{s}(X)$ is continuous. Then, for any function $G : [0,1] \mapsto \mathbb{R}$, and for any pair of nulls (i, j),*

$$\mathrm{Cor}\left[G(\hat{u}^{(\mathrm{marg})}(X_{2n+i})), G(\hat{u}^{(\mathrm{marg})}(X_{2n+j}))\right] = \frac{1}{n+2}.$$

Conformal p-values are positively correlated

Lemma 1. *Assume that $\hat{s}(X)$ is continuous. Then, for any function $G : [0,1] \mapsto \mathbb{R}$, and for any pair of nulls (i, j),*

$$\mathrm{Cor}\left[G(\hat{u}^{(\mathrm{marg})}(X_{2n+i})), G(\hat{u}^{(\mathrm{marg})}(X_{2n+j}))\right] = \frac{1}{n+2}.$$

$$\begin{aligned}
\mathrm{Var}\left[\sum_{i=1}^{m} G(p_i)\right] &= m\mathrm{Var}\left[G(p_1)\right] + m(m-1)\mathrm{Cov}\left[G(p_1), G(p_2)\right] \\
&= \left(m + \frac{m(m-1)}{n+2}\right)\mathrm{Var}\left[G(p_1)\right] \\
&\approx (1+\gamma)m\mathrm{Var}\left[G(p_1)\right].
\end{aligned}$$

Correction of Fisher's test

Reject the global null if

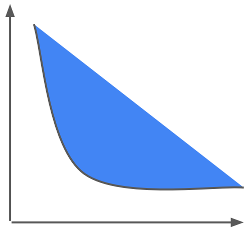
$$\frac{-2\sum_{i=1}^{m}\log\left[\hat{u}^{(\mathrm{marg})}(X_{2n+i})\right]+2(\sqrt{1+\gamma}-1)m}{\sqrt{1+\gamma}} \geq \chi^2(2m;1-\alpha)$$

Asymptotically equivalent to [Brown, 1975], [Kost et al., 2002]

Positive Regression Dependent on a Subset

Definition 1 (PRDS). *A random vector* $X = (X_1, \ldots, X_m)$ *is PRDS if for any* $i \in \{1, \ldots, m\}$ *and any increasing set* D*, the probability* $\mathbb{P}[X \in D \mid X_i = x]$ *is increasing in* x*.*

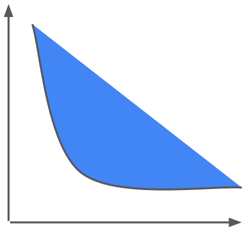
a set $D \subset \mathbb{R}^m$ is *increasing* if $a \in D$ and $b \succeq a$ implies $b \in D$.



Positive Regression Dependent on a Subset

Definition 1 (PRDS). *A random vector $X = (X_1, \dots, X_m)$ is PRDS if for any $i \in \{1, \dots, m\}$ and any increasing set D, the probability $\mathbb{P}[X \in D \mid X_i = x]$ is increasing in x.*

a set $D \subset \mathbb{R}^m$ is *increasing* if $a \in D$ and $b \succeq a$ implies $b \in D$.



Theorem 2 (Conformal p-values are PRDS). *Assume that $\hat{s}(X)$ is continuous. Consider m test points $X_{2n+1}, \dots, X_{2n+m}$ such that the first $m' \leq m$ of them are inliers, jointly independent of each other and of the data in $\mathcal{D}$. Then, the marginal conformal p-values $(\hat{u}^{(\mathrm{marg})}(X_{2n+1}), \dots, \hat{u}^{(\mathrm{marg})}(X_{2n+m'}))$ are PRDS.*

Benjamini-Hochberg procedure

[Benjamini et al., 2001]

Benjamini and Hochberg (1995) showed that when the test statistics are independent the following procedure controls the FDR at level $q \cdot m_0/m \leq q$.

THE BENJAMINI HOCHBERG PROCEDURE. Let $p_{(1)} \leq p_{(2)} \leq \cdots \leq p_{(m)}$ be the ordered observed p-values. Define

$$k = \max\left\{i\colon p_{(i)} \leq \frac{i}{m}q\right\}, \tag{1}$$

and reject $H^0_{(1)} \cdots H^0_{(k)}$. If no such i exists, reject no hypothesis.

False Discovery Rate control

Corollary 1 (Benjamini and Yekutieli [14]). *In the setting of Theorem 2, the Benjamini-Hochberg procedure applied at level* $\alpha \in (0,1)$ *to* $(\hat{u}^{(\mathrm{marg})}(X_{2n+1}), \ldots, \hat{u}^{(\mathrm{marg})}(X_{2n+m}))$ *controls the FDR at level* $\pi_0\alpha$*, where* π_0 *is the proportion of true nulls. That is,*

$$\mathbb{E}\left[\frac{|\mathcal{R} \cap \mathcal{H}_0|}{\max\{1, |\mathcal{R}|\}}\right] \leq \pi_0\alpha \leq \alpha, \tag{7}$$

where $\mathcal{H}_0 = \{i : H_{0,i} \text{ holds}\} \subseteq \{2n+1, \ldots, 2n+m\}$ *is the subset of true inliers in the test set, and* $\mathcal{R} \subseteq \{2n+1, \ldots, 2n+m\}$ *is the subset of test points reported as likely outliers.*

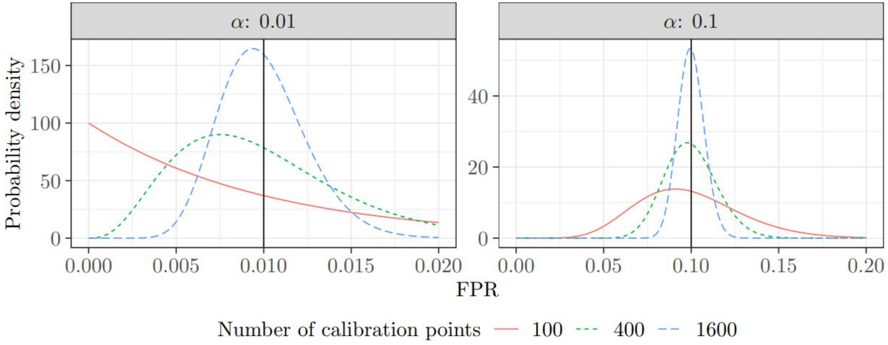
Calibration-conditional validity

$$\mathrm{FPR}(\alpha; \mathcal{D}) := \mathbb{P}\left[\hat{u}^{(\mathrm{marg})}(X_{2n+1}) \leq \alpha \mid \mathcal{D}\right]$$

Calibration-conditional validity

$$\mathrm{FPR}(\alpha; \mathcal{D}) := \mathbb{P}\left[\hat{u}^{(\mathrm{marg})}(X_{2n+1}) \leq \alpha \mid \mathcal{D}\right]$$

Proposition 1 (Pointwise FPR of marginal conformal p-values, adapted from [38])**.** *Let* $\ell = \lfloor (n+1)\alpha \rfloor$. *If* $\hat{s}(X)$ *is continuous,* $\mathrm{FPR}(\alpha; \mathcal{D})$ *follows a* $\mathrm{B{\small ETA}}(\ell, n+1-\ell)$ *distribution.*



Calibration-conditional validity

Idea: p-value adjustment $\hat{u}(X) = (g \circ \hat{F} \circ \hat{s})(X)$

Calibration-conditional validity

Idea: p-value adjustment $\hat{u}(X) = (g \circ \hat{F} \circ \hat{s})(X)$

Theorem 4 (Conditional p-value adjustment). *Let* $U_1, \ldots, U_n \overset{\text{i.i.d.}}{\sim} \text{Unif}([0,1])$, *with order statistics* $U_{(1)} \leq U_{(2)} \leq \ldots \leq U_{(n)}$, *and fix any* $\delta \in (0,1)$. *Suppose* $0 \leq b_1 \leq b_2 \leq \ldots \leq b_n \leq 1$ *are* n *reals such that*

$$\mathbb{P}\left[U_{(1)} \leq b_1, \ldots, U_{(n)} \leq b_n\right] \geq 1 - \delta. \tag{10}$$

Let also $b_0 = 0, b_{n+1} = 1$, *and* $h : [0,1] \mapsto [0,1]$ *be a piece-wise constant function such that*

$$h(t) = b_{\lceil (n+1)t \rceil}, \; t \in [0,1].$$

Then, $\hat{u}^{(\text{ccv})} = h \circ \hat{u}^{(\text{marg})}$ *satisfies* (4), *i.e.,* $\hat{u}^{(\text{ccv})}(X_{2n+1})$ *is a calibration-conditional valid p-value.*

$$\mathbb{P}\left[\mathbb{P}\left[\hat{u}^{(\text{ccv})}(X_{2n+1}) \leq t \mid \mathcal{D}\right] \leq t \text{ for all } t \in (0,1)\right] \geq 1 - \delta, \tag{4}$$

Calibration-conditional validity

E.g. $\hat{u}^{(\mathrm{marg})}(X) = \dfrac{25}{n+1} \approx 0.05 \to \hat{u}^{(\mathrm{ccv})}(X) = b_{25} \approx 0.075$

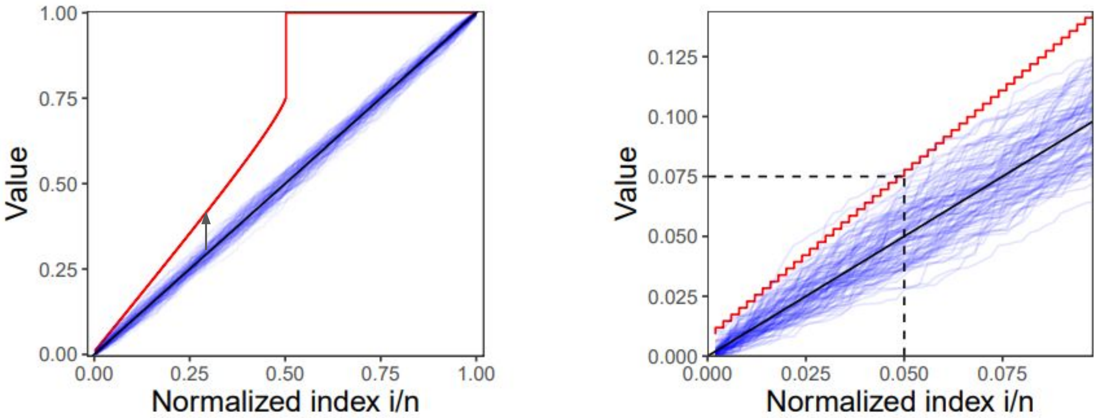


Figure 3: Illustration of Theorem 4. The red curve gives the sequence derived by generalized Simes inequality (Proposition 2) with $k = n/2 = 250$. The right panel zooms in on small indices.

Simes Inequality

Proposition 2 (Generalized Simes Inequality, from Equation (3.5) in [73]). *For any positive integer $k \leq n$, the uniform bound (10) in Theorem 4 holds with*

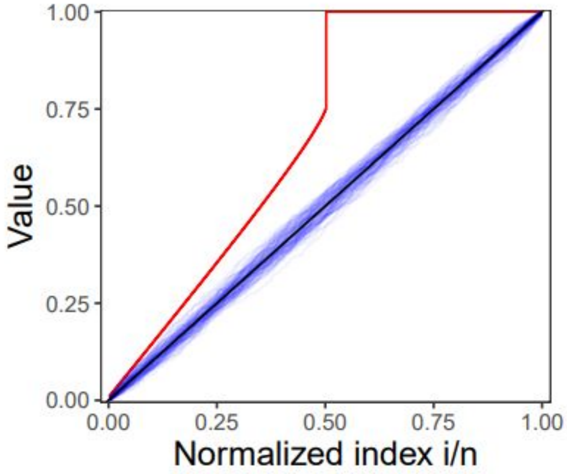
$$b_{n+1-i} = 1 - \delta^{1/k} \left(\frac{i \cdots (i-k+1)}{n \cdots (n-k+1)} \right)^{1/k}, \quad i = 1, \ldots, n.$$

Simes Inequality

Proposition 2 (Generalized Simes Inequality, from Equation (3.5) in [73]). *For any positive integer $k \leq n$, the uniform bound (10) in Theorem 4 holds with*

$$b_{n+1-i} = 1 - \delta^{1/k} \left(\frac{i \cdots (i-k+1)}{n \cdots (n-k+1)} \right)^{1/k}, \quad i = 1, \ldots, n.$$

$$k = \frac{n}{2}$$



Some extensions

- Simultaneous confidence bounds for FPR

Proposition 3. *Let F denote the true CDF of some distribution from which n i.i.d. samples, $Z_1, \ldots, Z_n$, are drawn, and denote by $\hat{F}_n$ the corresponding empirical CDF. With the same notation as in Theorem 4,*

$$\mathbb{P}\left[F(z) \leq h(\hat{F}_n(z)),\ \forall z \in \mathbb{R}\right] \geq 1 - \delta. \tag{11}$$

Some extensions

- Simultaneous confidence bounds for FPR

Proposition 3. *Let F denote the true CDF of some distribution from which n i.i.d. samples, $Z_1, \dots, Z_n$, are drawn, and denote by $\hat{F}_n$ the corresponding empirical CDF. With the same notation as in Theorem 4,*

$$\mathbb{P}\left[F(z) \leq h(\hat{F}_n(z)),\ \forall z \in \mathbb{R}\right] \geq 1 - \delta. \tag{11}$$

- Simultaneously-valid prediction sets

$$\hat{\mathcal{C}}^\alpha := \{x : \hat{u}^{(\mathrm{ccv})}(x) > \alpha\}.$$

$$\mathbb{P}\Big[\mathbb{P}\big[X_{2n+1} \in \hat{\mathcal{C}}^\alpha \mid \mathcal{D}\big] \geq 1 - \alpha \text{ for all } \alpha \in (0,1)\Big] \geq 1 - \delta.$$

Experiments (synthetic)

$X_i = \sqrt{a} V_i + W_i \in \mathbb{R}^{50}$, $a = 1$: inliers, $a > 1$:outliers

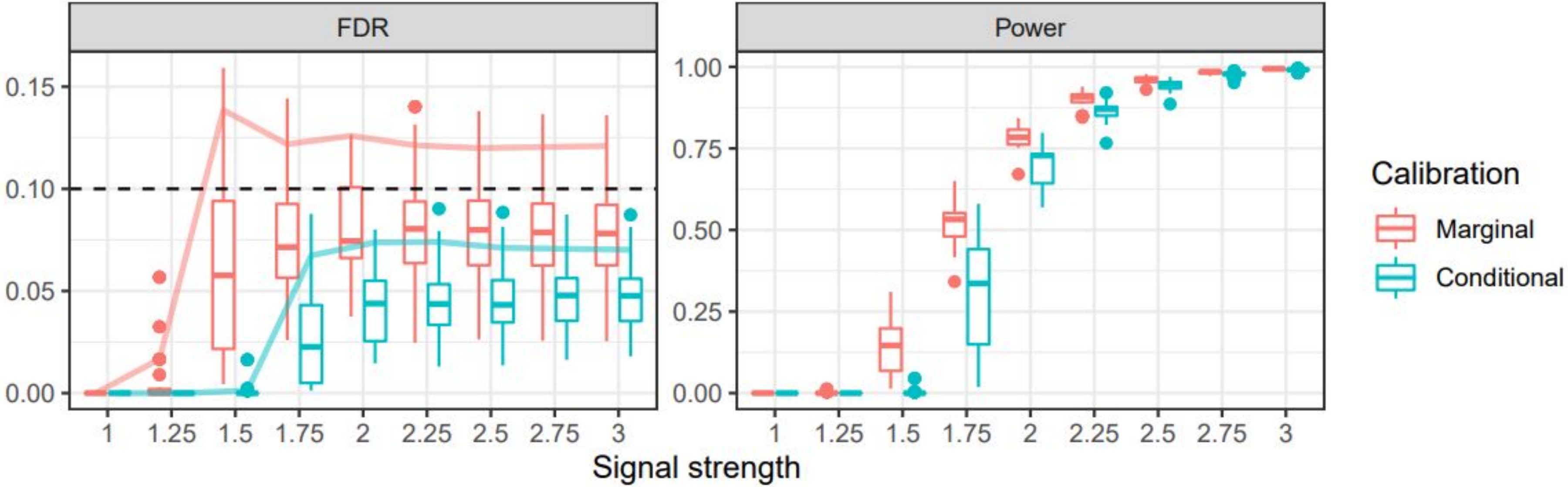
$\hat{s}$: one-class SVM

train/calibration sets: $|\mathcal{D}_j| = 2000, j = 1, \ldots, 100$

test sets: 10% outliers

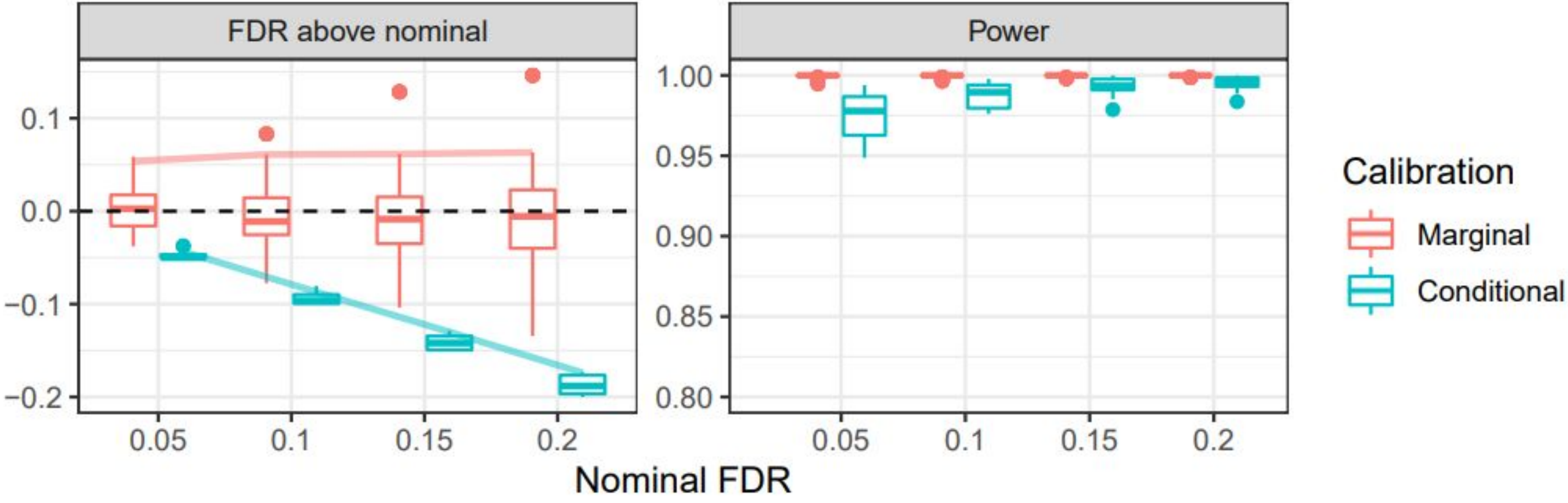
$$\widehat{\text{cFDR}}(\mathcal{D}_j) := \frac{1}{L} \sum_{l=1}^{L} \text{FDP}(\mathcal{D}_{j,l}^{\text{test}}; \mathcal{D}_j), \qquad \widehat{\text{cPower}}(\mathcal{D}_j) := \frac{1}{L} \sum_{l=1}^{L} \text{Power}(\mathcal{D}_{j,l}^{\text{test}}; \mathcal{D}_j),$$

Experiments (synthetic)



Experiments (synthetic)

Batch outlier detection



Experiments (real data)

	ALOI [78, 79]	Cover [80]	Credit card [81]	KDDCup99 [78, 82]	Mammography [83]	Digits [84]	Shuttle [85]
Features d	27	10	30	40	6	16	9
Inliers n_{inliers}	283301	286048	284315	47913	10923	6714	45586
Outliers n_{outliers}	1508	2747	492	200	260	156	3511

Experiments (real data)