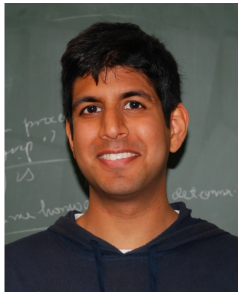
Hunt-and-test strategies for ML-powered hypothesis testing

F. Richard Guo

Department of Statistics, University of Michigan

1st May, 2026
Statistics, Oxford



Rajen D. Shah
Cambridge

Aditya Dhawan
Cambridge

☞ This talk is based on:

- *Rank-transformed subsampling: inference for multiple data splitting and exchangeable p-values*, JRSS-B 2025.
- *The debiased score test: Hunt and test for semi-parametric hypotheses*, 2026+.

Hunt and test

Hunt and test: General idea

In hypothesis testing, we seek tests that are

1. calibrated under the null H_0, and
2. powerful to detect alternatives in H_1

$\hookrightarrow$ What if H_1 is large or unspecified, and there can be all kinds of alternatives...

Hunt and test: General idea

In hypothesis testing, we seek tests that are

1. calibrated under the null H_0, and
2. powerful to detect alternatives in H_1

$\hookrightarrow$ What if H_1 is large or unspecified, and there can be all kinds of alternatives...

☞ Learn the alternative $\widehat{P} \in H_1$ from which the data appear to have arisen, and choose test statistic according to target the alternative.
$\hookrightarrow$ However, hunt and test on the same data inflates the type-I error. ☞ double-dipping

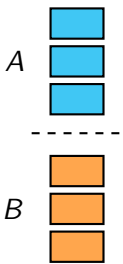
Hunt and test: General idea

In hypothesis testing, we seek tests that are

1. calibrated under the null H_0, and
2. powerful to detect alternatives in H_1

$\hookrightarrow$ What if H_1 is large or unspecified, and there can be all kinds of alternatives...

☞ Learn the alternative $\widehat{P} \in H_1$ from which the data appear to have arisen, and choose test statistic according to target the alternative.
$\hookrightarrow$ However, hunt and test on the same data inflates the type-I error. ☞ double-dipping

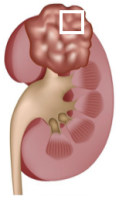


❶ **Hunt** for signal and choose the test statistic accordingly.
$\hookrightarrow$ ML powered

❷ **Test** the significance of the hunted signal.

Example: Significance of clustering

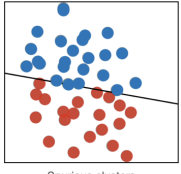
☞ Imagine a patient is diagnosed with a common subtype of kidney cancer. Before delivering the therapy, how do we rule out co-existence of other subtypes?



Kidney tumor

	Gene 1	Gene 2	Gene 3	...
Cell 1	10	10	0	
Cell 2	0	15	4	
Cell 3	600	0	20	
⋮				

Single-cell RNA read count

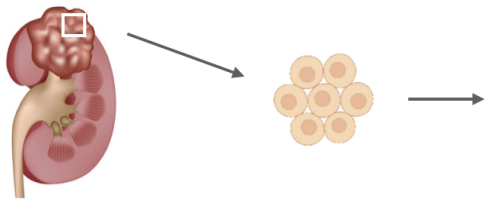


Spurious clusters

☞ Just running a clustering algorithm won't work.

Example: Significance of clustering

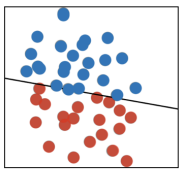
☞ Imagine a patient is diagnosed with a common subtype of kidney cancer. Before delivering the therapy, how do we rule out co-existence of other subtypes?



Kidney tumor

	Gene 1	Gene 2	Gene 3	...
Cell 1	10	10	0	
Cell 2	0	15	4	
Cell 3	600	0	20	
⋮				

Single-cell RNA read count

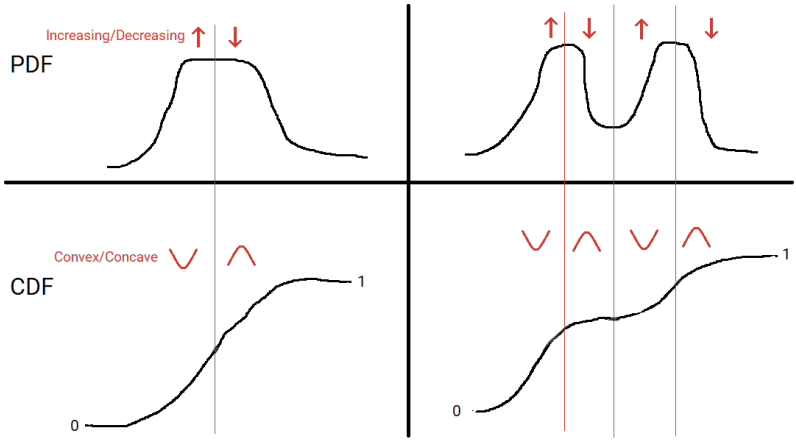


Spurious clusters

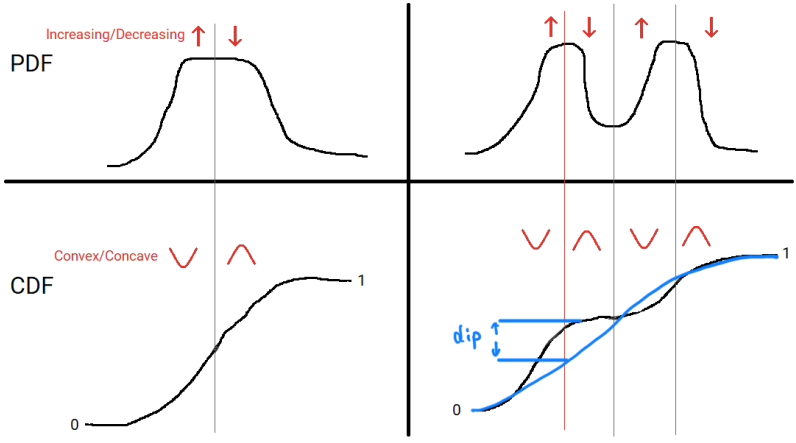
☞ Just running a clustering algorithm won't work.

✻ Given high-dimensional random vectors $X_1, \ldots, X_n \overset{\text{iid}}{\sim} P$, we want to test H_0: P is unimodal.

✳ Unimodality test on $\mathbb{R}$ is a solved problem.



✳ Unimodality test on $\mathbb{R}$ is a solved problem.



☞ Dip test for univariate unimodality (J. A. Hartigan and P. M. Hartigan, 1985; Cheng and Hall, 1998).

Example: Significance of clustering

* Can we use hunt and test to extend this to higher dimensions?

$H_0 : X \sim$ unimodal P

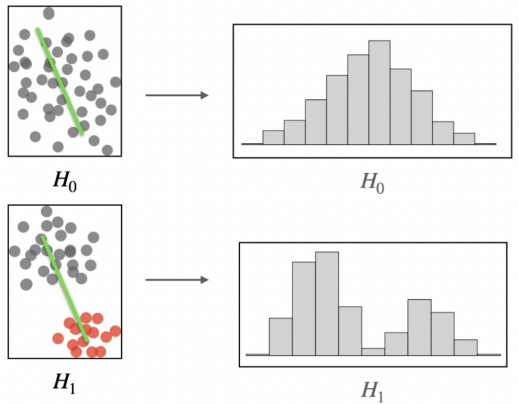
$$\iff \langle d, X \rangle \text{ is unimodal in every direction } d, \text{ i.e.,} H_0 = \bigcap_d H_0(d).$$

Example: Significance of clustering

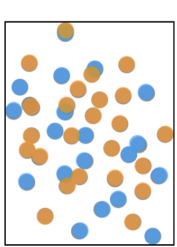
* Can we use hunt and test to extend this to higher dimensions? ☞ linear unimodality

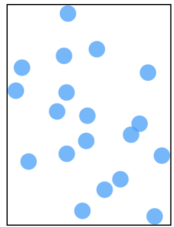
$H_0 : X \sim$ unimodal P

$\iff \langle d, X \rangle$ is unimodal in every direction d, i.e., $H_0 = \bigcap_d H_0(d)$.

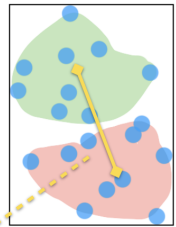


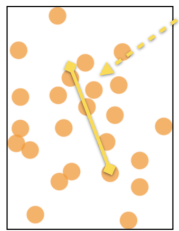
Hunt and test!

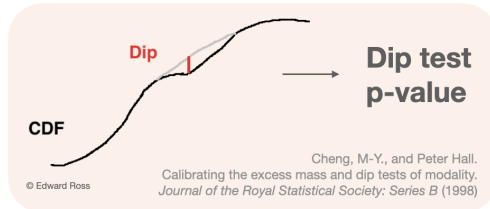




2-means



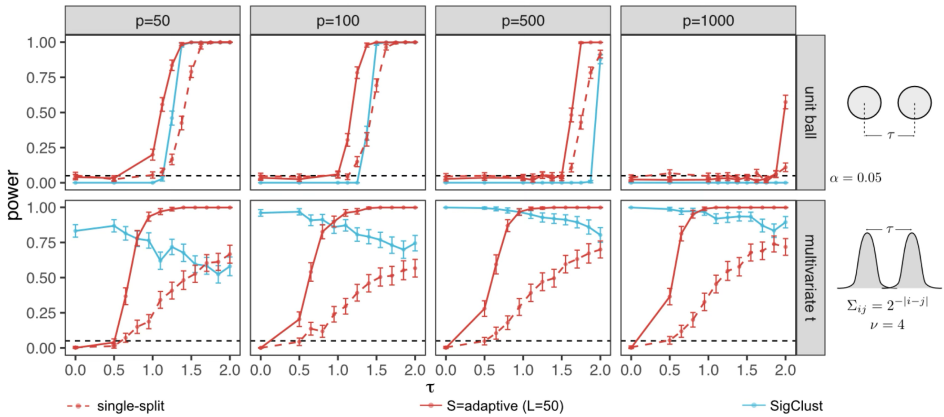




Cheng, M-Y., and Peter Hall.
Calibrating the excess mass and dip tests of modality.
Journal of the Royal Statistical Society: Series B (1998)

Example: Significance of clustering

☞ 'single-split' (- - -) is the method just described.



☞ `SigClust` (Huang, Liu, and Marron, 2022) is a state-of-the-art method based on Gaussian mixtures.

Example: Significance of clustering

☞ 'Hunt and test' is a flexible strategy one can adopt to design tests:

↪ See Guo and Shah (2024) for further examples.

- Applicable whenever one can reduce testing a difficult null hypothesis H_0 to testing easier hypotheses, e.g., $H_0 = \cap_d H_0(d)$.
- Through data splitting, data-driven ML methods can be used for hunting.
- Can repurpose existing tests for the testing step.
- Can be sensitive to the way that the data is split. ↪ More on this later

Example: Significance of clustering

☞ 'Hunt and test' is a flexible strategy one can adopt to design tests:

↪ See Guo and Shah (2024) for further examples.

- Applicable whenever one can reduce testing a difficult null hypothesis H_0 to testing easier hypotheses, e.g., $H_0 = \cap_d H_0(d)$.
- Through data splitting, data-driven ML methods can be used for hunting.
- Can repurpose existing tests for the testing step.
- Can be sensitive to the way that the data is split. ↪ More on this later

✳ Next, I will describe how to systematically and optimally hunt and test for assessing goodness-of-fit and significance of semiparametric regression models.

Hunt and test for semiparametric regression

Perspectives on goodness-of-fit testing

* I will focus on goodness-of-fit testing, although you will see that significance testing can be handled similarly.

☞ Models offer an often necessary simplification in the face of limited information from data, e.g., curse of dimensionality, latent variables, missing data.

Yet as we well know: *all models are wrong...*

$\hookrightarrow$ When is a model useful?

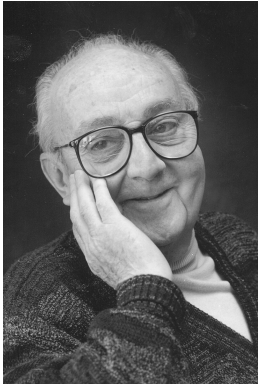
Perspectives on goodness-of-fit testing

✻ I will focus on goodness-of-fit testing, although you will see that significance testing can be handled similarly.

☞ Models offer an often necessary simplification in the face of limited information from data, e.g., curse of dimensionality, latent variables, missing data.

Yet as we well know: *all models are wrong...*

$\hookrightarrow$ When is a model useful?



☞ A necessary condition: the data does not provide strong evidence against the model we assume.

- Difficult to interpret otherwise (even for OLS!)
- Simple visual diagnostics exist for LM/GLM.
- Goodness-of-fit test is a pricipled approach.
- ML is useful for scrutinizing the potential discrepancy between data and model.

Semiparametric models

- Generalised additive models (GAM):

$$g(\mathbb{E}[Y \mid X]) = \sum_j f_j(X_j),$$

where $g(\cdot)$ is a chosen link function.

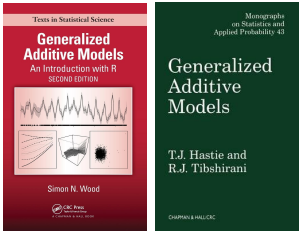
- Partially linear models:

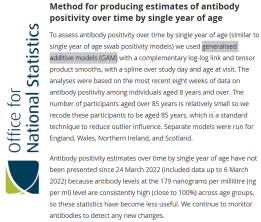
$$\mathbb{E}(Y \mid A, X) = \theta A + f(X).$$

- Marginal structural models:

$$\mathbb{E}[Y(a) - Y(0) \mid X] = f_1(a, X_S) + f_0(X),$$

where f_1 models treatment effect heterogeneity.





☞ When misspecified, however, the truth can be of any shape.

Existing work

Most existing work:

- Purpose-built for each model: GAM (Härdle, Sperlich, and Spokoiny, 2001; Gozalo and Linton, 2001), additive quantile regression (Fasiolo et al., 2021), effect homogeneity (Dukes et al., 2024), significance test (Williamson, Gilbert, Carone, et al., 2021; Williamson, Gilbert, Simon, et al., 2021);
- Kernel smoothing with a tuning parameter: Sperlich, Tjøstheim, and Yang (2002) and Fan and Jiang (2005).

$\hookrightarrow$ We would like an approach that is generic (works for a variety of models), user-friendly (no tuning or bootstrap) and powerful to detect all kinds of misspecifications.

Formulation

Suppose the true regression function f^* can be identified as a risk minimizer. We test

$$H_0: \quad f^* := \underset{f \in L^2(P)}{\arg\min}\ \mathbb{E}\,\ell(f(X), Y) \in \mathcal{F},$$

where $\hookrightarrow L^2(P) = \{P\text{-square-integrable functions of } X\}$

- $\mathcal{F}$ is a **linear subspace** of $L^2(P)$: $f_1, f_2 \in \mathcal{F},\ a_1, a_2 \in \mathbb{R} \implies a_1 f_1 + a_2 f_2 \in \mathcal{F}$
  $\hookrightarrow$ e.g., $\mathcal{F} = \{\textit{additive}\}$, $\mathcal{F} = \{\text{linear}(x_1) + \text{nonlinear}(x_2)\}$, $\mathcal{F} = \{f : f(x) = f(x_S)\}$.
- ℓ is a loss function convex in its first argument.

Formulation

Suppose the true regression function f^* can be identified as a risk minimizer. We test

$$H_0: \quad f^* := \underset{f \in L^2(P)}{\arg\min}\ \mathbb{E}\,\ell(f(X), Y) \in \mathcal{F},$$

where $\hookrightarrow L^2(P) = \{P\text{-square-integrable functions of } X\}$

- $\mathcal{F}$ is a **linear subspace** of $L^2(P)$: $f_1, f_2 \in \mathcal{F},\ a_1, a_2 \in \mathbb{R} \implies a_1 f_1 + a_2 f_2 \in \mathcal{F}$
  $\hookrightarrow$ e.g., $\mathcal{F} = \{\textit{additive}\}$, $\mathcal{F} = \{\text{linear}(x_1) + \text{nonlinear}(x_2)\}$, $\mathcal{F} = \{f : f(x) = f(x_S)\}$.
- ℓ is a loss function convex in its first argument.

$\hookrightarrow$ For simplicity, let us take the square loss $\ell = \{(\mu(f)(X) - Y\}^2$ and focus on testing the conditional mean specification

$$\mathbb{E}(Y \mid X) = \mu(f^*)(X), \quad f^* \in \mathcal{F},$$

where $\mu(\cdot)$ is a known, increasing and smooth link function.

Hunt and test with the score

☞ We start from the equivalence

$$f^* = \underset{f \in L^2(P)}{\arg\min}\ \mathbb{E}\,\ell(\mu(f)(X), Y) \iff \frac{\partial}{\partial f}\,\mathbb{E}\{Y - \mu(f)(X)\}^2|_{f^*} = 0$$

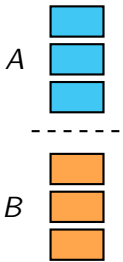
$$\iff \mathbb{E}[\mu'(f^*)(X)\,\{Y - \mu(f^*)(X)\}\,h(X)] = 0, \quad \forall h \in L^2(P).$$

Hunt and test with the score

☞ We start from the equivalence

$$\boxed{\mathbb{E}[Y \mid X] = \mu(f^*) \iff \mathbb{E}[\mu'(f^*)(X)\,\{Y - \mu(f^*)(X)\}\,h(X)] = 0, \quad \forall h \in L^2(P),}$$

where $\mu'(f^*)(Y - \mu(f)(X))$ is the score associated with the loss function.



❶ **Hunt**: With sample A, fit $\widetilde{f} \in \mathcal{F}$. Then, train an ML algorithm $\widehat{h}(X)$ to predict the residuals $Y_i - \mu(\widetilde{f})(X_i)$ from X_i.

❷ **Test**: With sample B, fit the null model $\widehat{f} \in \mathcal{F}$ again. Then, test that the residuals $Y_i - \mu(\widehat{f})(X_i)$ are uncorrelated with the hunted signal $\widehat{h}(X_i)$.
$\hookrightarrow$ CLT based on $L_i = \mu'(\widehat{f})(X_i)\{Y_i - \mu(\widehat{f})(X_i)\}\widehat{h}(X_i)$ studentized.

Testing: Bias

Consider the studentized test

$$T_n := \frac{\sum_i L_i}{\sqrt{n\,\widehat{\mathrm{var}} L}}, \quad L_i := \mu'(\widehat{f})(X_i)\{Y_i - \mu(\widehat{f})(X_i)\}\widehat{h}(X_i).$$

Testing: Bias

Consider the studentized test

$$T_n := \frac{\sum_i L_i}{\sqrt{n\,\widehat{\mathrm{var}} L}}, \quad L_i := \mu'(\widehat{f})(X_i)\{Y_i - \mu(\widehat{f})(X_i)\}\widehat{h}(X_i).$$

☞ Under H_0, we want the CLT $T_n \to_d \mathcal{N}(0,1)$ to hold, which requires

$$\begin{aligned}\sqrt{n}\,\mathbb{E}\left[\mu'(\widehat{f})(X_i)\left\{Y_i - \mu(\widehat{f})(X_i)\right\}\widehat{h}(X_i)\right] &= \sqrt{n}\,\mathbb{E}\left[\mu'(\widehat{f})(X_i)\left\{\mu(f^*)(X_i) - \mu(\widehat{f})(X_i)\right\}\widehat{h}(X_i)\right] \\ &\approx \sqrt{n}\,\mathbb{E}\left[[\mu'(\widehat{f})]^2\left\{f^*(X_i) - \widehat{f}(X_i)\right\}\widehat{h}(X_i)\right] \approx 0.\end{aligned}$$

By Cauchy–Schwarz,

$$n\left(\mathbb{E}\left[[\mu'(\widehat{f})]^2\{f(X_i) - f^*(X_i)\}\,\widehat{h}(X_i)\right]\right)^2 \lesssim \underbrace{n\,\mathbb{E}\left[\left\{f^*(X_i) - \widehat{f}(X_i)\right\}^2\right]}_{\text{bias}\to 0?} \cdot \underbrace{\mathbb{E}\,\widehat{h}^2(X_i)}_{=O(1)}.$$

Testing: Bias

Consider the studentized test

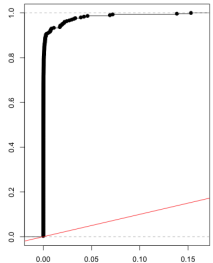
$$T_n := \frac{\sum_i L_i}{\sqrt{n\,\widehat{\mathrm{var}} L}}, \quad L_i := \mu'(\widehat{f})(X_i)\{Y_i - \mu(\widehat{f})(X_i)\}\widehat{h}(X_i).$$

☞ Under H_0, we want the CLT $T_n \to_d \mathcal{N}(0,1)$ to hold, which requires

$$\sqrt{n}\,\mathbb{E}\left[\mu'(\widehat{f})(X_i)\left\{Y_i - \mu(\widehat{f})(X_i)\right\}\widehat{h}(X_i)\right] = \sqrt{n}\,\mathbb{E}\left[\mu'(\widehat{f})(X_i)\left\{\mu(f^*)(X_i) - \mu(\widehat{f})(X_i)\right\}\widehat{h}(X_i)\right]$$
$$\approx \sqrt{n}\,\mathbb{E}\left[[\mu'(\widehat{f})]^2\left\{f^*(X_i) - \widehat{f}(X_i)\right\}\widehat{h}(X_i)\right] \approx 0.$$

For semiparametric models, this bias due to estimating f^* cannot be ignored in general.

$\hookrightarrow$ Distribution of p-values under the null ☞



Testing: Killing the bias

Under H_0, there is hardly any control we have over the estimation error

$$\sqrt{n}\,\mathbb{E}\left[[\mu'(\widehat{f})]^2 \underbrace{\left\{f^*(X_i) - \widehat{f}(X_i)\right\}}_{\text{error}} \widehat{h}(X_i)\right]$$

Testing: Killing the bias

Under H_0, there is hardly any control we have over the estimation error

$$\sqrt{n}\,\mathbb{E}\left[[\mu'(\widehat{f})]^2 \underbrace{\left\{f^*(X_i) - \widehat{f}(X_i)\right\}}_{\in \mathcal{F}} \widehat{h}(X_i)\right]$$

☜ This holds by design: $\widehat{f} \in \mathcal{F}$ under H_0, $\widehat{f}$ is fitted within $\mathcal{F}$ and $\mathcal{F}$ is linear.

Testing: Killing the bias

Under H_0, there is hardly any control we have over the estimation error

$$\sqrt{n}\,\mathbb{E}\left[[\mu'(\widehat{f})]^2 \underbrace{\left\{f^*(X_i) - \widehat{f}(X_i)\right\}}_{\in \mathcal{F}} \widehat{h}(X_i)\right]$$

☜ This holds by design: $\widehat{f} \in \mathcal{F}$ under H_0, $\widehat{f}$ is fitted within $\mathcal{F}$ and $\mathcal{F}$ is linear.

❋ While estimation cannot be controlled, we do have control over the test function $\widehat{h}$.
↪ Indeed, project it to be orthogonal to $\mathcal{F}$ under weights $w = (\mu')^2$ to kill the bias!

$$\sqrt{n}\,\mathbb{E}\left[\underbrace{[\mu'(\widehat{f})]^2}_{w} \underbrace{\left\{f^*(X_i) - \widehat{f}(X_i)\right\}}_{\in \mathcal{F}} \underbrace{\left(\widehat{h}(X_i) - m_{\widehat{h}}(X_i)\right)}_{\in \mathcal{F}_w^{\perp}}\right] = 0.$$

"The killing time; Unwillingly mine."

— the Killing Moon, Echo & the Bunnymen.

Testing: Killing the bias

Let $\mathcal{F}_w := \{[\mu'(f^*(\cdot))]^2 f(\cdot) : f \in \mathcal{F}\}$.

For orthogonal complement $\mathcal{F}_w^{\perp} := \{g : \mathbb{E}\, g(X)f(X) = 0 \,\forall f \in \mathcal{F}_w\}$, we seek $\widehat{h} \in \mathcal{F}_w^{\perp}$:

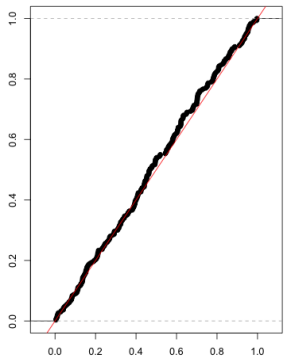
$$m_{\widehat{h}} := \arg\min_{g \in \mathcal{F}} \mathbb{E}\left[[\mu'(f^*)]^2 \left\{\widehat{h}(X) - g(X)\right\}^2 \mid \widehat{h}\right] \implies \widehat{h} - m_{\widehat{h}} \in \mathcal{F}_w^{\perp}.$$

☞ Estimate $\widehat{m}_{\widehat{h}}$ through a weighted least squares

$$\widehat{m}_{\widehat{h}} = \arg\min_{m \in \mathcal{F}} \sum_{i \in B} \left[\mu'(\widehat{f})(X_i)\right]^2 \left\{\widehat{h}(X_i) - m(X_i)\right\}^2$$

and form the debiased test statistic using

$$L_i := \mu'(\widehat{f})(X_i)\left\{Y_i - \mu(\widehat{f})(X_i)\right\}\{\widehat{h}(X_i) - \widehat{m}_{\widehat{h}}(X_i)\}.$$



Hunting

We still have to determine how to hunt.

☞ Consider an oracle version of the test using oracle weighted residuals R_i:

$$T = \frac{1}{\sqrt{n \cdot \mathrm{var}\{R_i(h - m_h)\}}} \sum_{i \in B} R_i\{h(X_i) - m_h(X_i)\}, \quad R_i := \mu'(f^*)(X_i)\,\{Y_i - \mu(f^*)(X_i)\}.$$

✻ The optimal hunt is determined by maximizing

$$\mathrm{SNR}(h) = \frac{\mathbb{E}[R\{h(X) - m_h(X)\}]}{\sqrt{\mathrm{var}(R(h(X) - m_h(X)))}}. \qquad (\star)$$

Hunting

We still have to determine how to hunt.

☞ Consider an oracle version of the test using oracle weighted residuals R_i:

$$T = \frac{1}{\sqrt{n \cdot \mathrm{var}\{R_i(h - m_h)\}}} \sum_{i \in B} R_i\{h(X_i) - m_h(X_i)\}, \quad R_i := \mu'(f^*)(X_i)\,\{Y_i - \mu(f^*)(X_i)\}.$$

☞ Since $h - m_h \in \mathcal{F}_w^{\perp}$, we can instead directly maximize over $h \in \mathcal{F}_w^{\perp}$

$$\mathrm{SNR}(h) = \frac{\mathbb{E}[Rh]}{\sqrt{\mathrm{var}(Rh)}}. \qquad (\star)$$

Hunting

We still have to determine how to hunt.

☞ Consider an oracle version of the test using oracle weighted residuals R_i:

$$T = \frac{1}{\sqrt{n \cdot \mathrm{var}\{R_i(h - m_h)\}}} \sum_{i \in B} R_i \{h(X_i) - m_h(X_i)\}, \quad R_i := \mu'(f^*)(X_i)\,\{Y_i - \mu(f^*)(X_i)\}.$$

☞ Since $h - m_h \in \mathcal{F}_w^{\perp}$, we can instead directly maximize over $h \in \mathcal{F}_w^{\perp}$

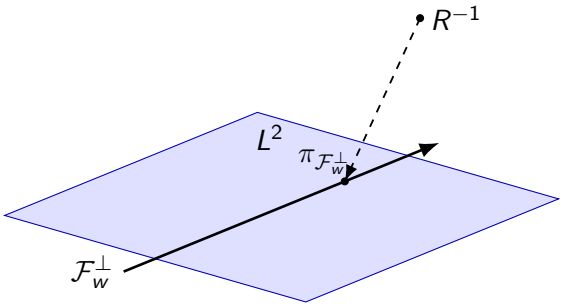
$$\mathrm{SNR}(h) = \frac{\mathbb{E}[Rh]}{\sqrt{\mathrm{var}(Rh)}}. \qquad (\star)$$

Lemma Given a function class $\mathcal{H}$, we have the weighted-least-squares representation

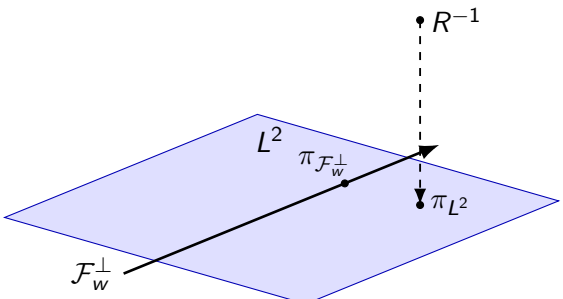
$$\arg\max_{h \in \mathcal{H}} \mathrm{SNR}(h) = \pi_{\mathcal{H}} := \arg\min_{h \in \mathcal{H}} \mathbb{E}\left[R^2 \left(R^{-1} - h(X)\right)^2\right].$$

$\hookrightarrow$ The optimal hunt is $\pi_{\mathcal{F}_w^{\perp}}$, which we cannot directly fit.

Optimal hunting

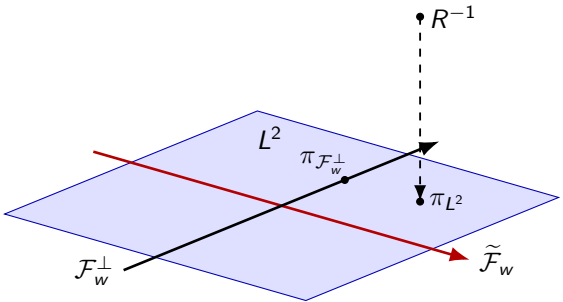


Optimal hunting



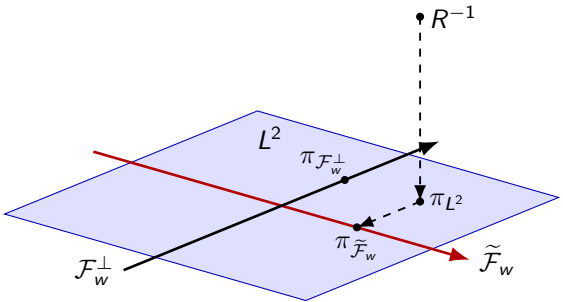
Optimal hunting

$\widetilde{F}_w := \{g : \mathbb{E}[R^2 g(X) f(X)] = 0 \; \forall \; f \in \mathcal{F}_w^{\perp}\}.$



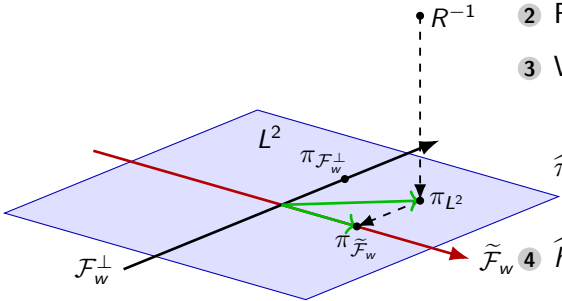
Optimal hunting

$\widetilde{F}_w := \{g : \mathbb{E}[R^2 g(X) f(X)] = 0 \; \forall \; f \in \mathcal{F}_w^{\perp}\}.$



Optimal hunting

$\widetilde{F}_w := \{g : \mathbb{E}[R^2 g(X) f(X)] = 0 \; \forall \, f \in \mathcal{F}_w^{\perp}\}$.



Write $\widehat{\mathbb{E}}$ for sample average over I_1.

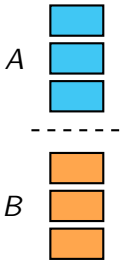
1. $\widehat{\pi}_{L^2} := \arg\min_g \widehat{\mathbb{E}}\{\widehat{R}^2(\widehat{R}^{-1} - g(X))^2\}$
2. Form estimate $\widehat{v}$ of $v(X) := \mathbb{E}(R^2 \,|\, X)$.
3. Writing $\widetilde{g} := \frac{\widehat{v}}{\mu' \circ \widetilde{f}} \widehat{\pi}_{L^2}$ and $\widetilde{w} := \frac{(\mu' \circ \widetilde{f})^2}{\widehat{v}}$, set

$$\widehat{\pi}_{\widetilde{\mathcal{F}}_w} := \frac{\mu' \circ \widetilde{f}}{\widehat{v}} \cdot \arg\min_g \widehat{\mathbb{E}}\{\widetilde{w}(X)(\widetilde{g}(X) - g(X))^2\}.$$

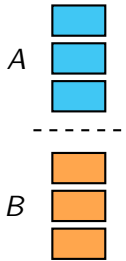
4. $\widehat{h} := \widehat{\pi}_{L^2} - \widehat{\pi}_{\widetilde{\mathcal{F}}_w}$.

$\hookrightarrow$ Think of $\widehat{\pi}_{L^2}$ as doing the 'bulk of the hunting' and $\widehat{\pi}_{\widetilde{\mathcal{F}}_w}$ as a 'refinement'.

Summary of approach



Summary of approach

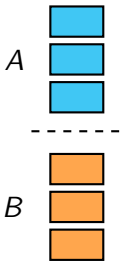


1. **Hunt**: With sample A, fit model $\widetilde{f} \in \mathcal{F}$ under the null. Next, use an ML algorithm to hunt for signal in H_1 by fitting weighted least squares

$$1/R_i \sim X_i, \quad \text{with weights } R_i^2,$$

where $R_i = \mu'(\widetilde{f})(X_i)\{Y_i - \mu(\widetilde{f})(X_i)\}$. This gives $\widehat{h}$.

Summary of approach



1. **Hunt**: With sample A, fit model $\widetilde{f} \in \mathcal{F}$ under the null. Next, use an ML algorithm to hunt for signal in H_1 by fitting weighted least squares

$$1/R_i \sim X_i, \quad \text{with weights } R_i^2,$$

where $R_i = \mu'(\widetilde{f})(X_i)\{Y_i - \mu(\widetilde{f})(X_i)\}$. This gives $\widehat{h}$.

2. **Test**: With sample B,
   - fit the null model $\widehat{f} \in \mathcal{F}$ and get the residuals;
   - compute the 'refinement' to the hunted signal and form $\widehat{h} - \widehat{m}_{\widehat{h}} \in \mathcal{F}_w^{\perp}$.

With $L_i := \mu'(\widehat{f})(X_i)\left\{Y_i - \mu(\widehat{f})(X_i)\right\}\left\{\widehat{h}(X_i) - \widehat{m}_{\widehat{h}}(X_i)\right\}$,

$$\text{Reject } H_0 \text{ if } \quad \frac{1}{\sqrt{n_B \widehat{\text{var}} L}} \sum_{i \in B} L_i =: T_n > \Phi^{-1}(1-\alpha).$$

Type-I error control

For theory, it is convenient to work a 3-split version, with $\widehat{h}$, $\{\widehat{m}_{\widehat{h}}, \widehat{f}\}$ and T_n each formed on different splits. For practice, a standard 2-split suffices.

Theorem Under regularity conditions, we have

$$T_n \xrightarrow{d} \mathcal{N}(0,1) \quad \text{under } H_0,$$

provided that we have $\mathcal{E}_1\mathcal{E}_2 = o_P(n^{-1})$ and $\mathcal{E}_1\mathcal{E}_3 = o_P(n^{-1})$.

$$f^* := \underset{f\in\mathcal{F}}{\arg\min}\ \mathbb{E}\,\ell(f(X), Y), \quad \xi := \widehat{h}(X) - m_{\widehat{h}}(X), \quad \sigma^2 := \mathbb{E}(\xi^2 \mid \widehat{h}),$$

$$\mathcal{E}_1 := \mathbb{E}[\{\widehat{f}(X) - f^*(X)\}^2 \mid \widehat{f}], \quad \mathcal{E}_2 := \frac{1}{\sigma^2}\,\mathbb{E}[\{\widehat{m}_{\widehat{h}}(X) - m_{\widehat{h}}(X)\}^2 \mid \widehat{h}], \quad \mathcal{E}_3 := \frac{1}{\sigma^2}\,\mathbb{E}[\xi^2\{\widehat{f}(X) - f^*(X)\}^2].$$

Power

Power depends on the size of misspecification in

$$s(X) := \mathbb{E}(Y \mid X) - \mu(f^*)(X).$$

☞ Let P_n be a sequence of alternatives $\geq$ parametric rate, namely

$$n\, \mathbb{E}_{P_n} s^2(X) \to \infty.$$

Theorem As before, suppose $\mathcal{E}_1\mathcal{E}_2 = o_P(n^{-1})$ and $\mathcal{E}_1\mathcal{E}_3 = o_P(n^{-1})$.
Suppose there exists $\rho > 0$ such that ☞ quality of hunting

$$\mathbb{P}_{P_n}\left\{\mathrm{cor}\left(s(X), \widehat{h}(X) - m_{\widehat{h}}(X) \mid \widehat{h}\right) > \rho\right\} \to 1.$$

Then, for any $\alpha \in (0,1)$,

$$\mathbb{P}_{P_n}(T_n > z_{1-\alpha}) \to 1.$$

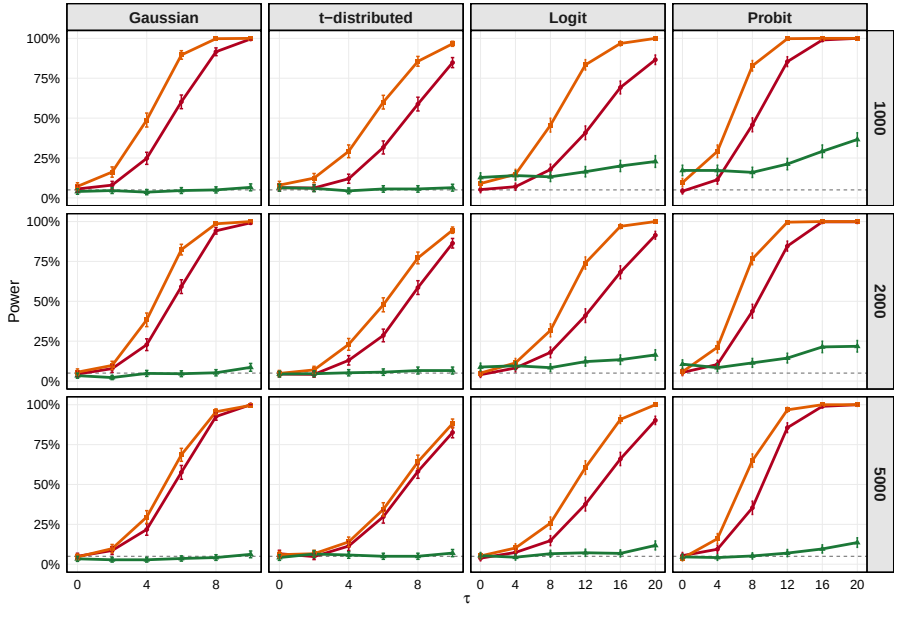
Generalised Additive Models

Continuous case: $$Y = \sin(2X_1) + \frac{\tau}{\sqrt{n}} X_1 X_3 + \frac{1}{2}\sqrt{1 + X_2^2}\,\varepsilon$$

Binary case: $$\Pr(Y = 1 \mid X) = \mu\Big(\sin(2X_1) + \frac{\tau}{\sqrt{n}} X_1 X_3\Big)$$

- $X \in \mathbb{R}^p$, $p = 10$.
- $\tau \in \{0, 2, 4, 6, 8, 10\}$ controls deviation from additivity.
- $\varepsilon \in \{\mathcal{N}(0,1), t_1\}$.
- $\mu \in$ inverse $\{$probit, logit$\}$ links .

We use `grf` (Tibshirani et al., 2024) for hunting.
$\hookrightarrow$ For comparison, we also report Williamson, Gilbert, Carone, et al. (2021), which compares the predictive performance of two regression models (GAM and `grf`).


Gaussian
t−distributed
Logit
Probit
1000
2000
5000
Power
100%
75%
50%
25%
0%
0
4
8
12
16
20
τ
DST
DST.MultiSplit
Williamson

Heterogeneous treatment effects

Under no unmeasured confounding and positivity, CATE is identified as

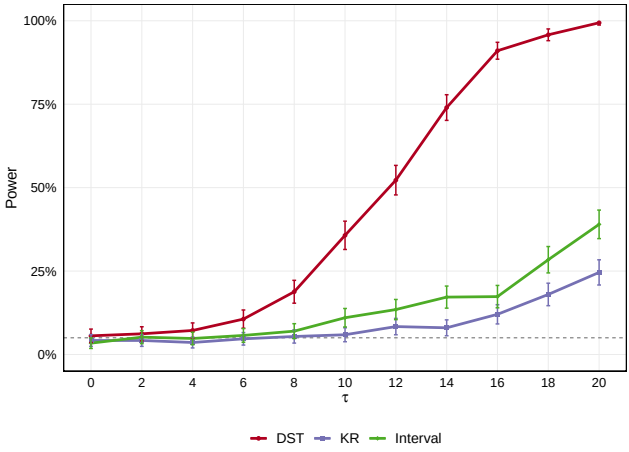
$$\tau(a, z) := \mathbb{E}[Y(a) \mid Z = z] = \mathbb{E}[Y \mid A = a, Z = z],$$

so $\ell =$ square loss.

We consider a nonparametric SEM setting from Dukes et al. (2024):

- $Z \sim \text{unif}[-1, 1]^5$; $n = 2000$.
- $\mathbb{P}(A = 1 \mid Z) = \text{expit}(\frac{1}{8} Z_1 + \frac{1}{4} \sin(\pi Z_2))$.
- $Y = \frac{3}{4} T + \text{expit}((Z_2 + Z_3)/2) + Z_1 + \frac{\tau}{\sqrt{n}} 2 T \sin(4\pi Z_3) + \frac{1}{2} \sqrt{1 + Z_2^2} \varepsilon$

Again, we use `grf` to hunt.



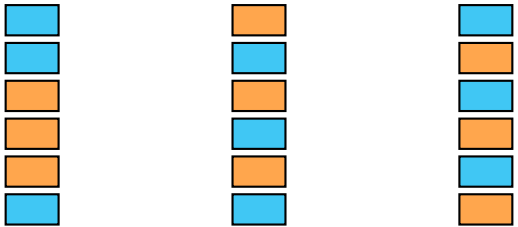
☜ Two methods are from Dukes et al. (2024):

- `KR`: doubly-robust estimation of CATE with kernel-ridge regression (Kennedy, 2023);
- `Interval`: Binning Z_3 into small intervals and use IPW to estimate the ATE in each bin.

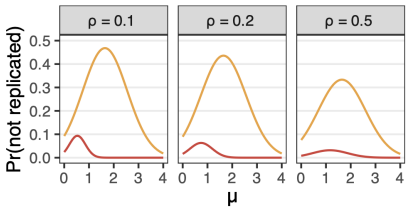
Multiple splits

Sensitivity to split

✻ As with any other method using data splitting, the result of hunt and test depends on how the dataset is split, which is typically random. ☞ randomized procedure



↪ This is an issue especially when the effect size is moderate, where the results from multiple splits may disagree.



Challenge

Suppose we randomly split data L times run the procedure for each split, getting test statistics

$$T_n^{(1)}, T_n^{(2)}, \ldots, T_n^{(L)}.$$

$\hookrightarrow$ These statistics are exchangeable.

- Each of them $\sim \mathcal{N}(0, 1)$ under H_0,
- but they are correlated in an unknown, usually complicated way.

$\hookrightarrow$ Their average is no longer a valid p-value.

Challenge

Suppose we randomly split data L times run the procedure for each split, getting test statistics

$$T_n^{(1)}, T_n^{(2)}, \ldots, T_n^{(L)}.$$

$\hookrightarrow$ These statistics are exchangeable.

- Each of them $\sim \mathcal{N}(0,1)$ under H_0,
- but they are correlated in an unknown, usually complicated way.

$\hookrightarrow$ Their average is no longer a valid p-value.

☞ How do we calibrate the aggregated test statistic, e.g.,

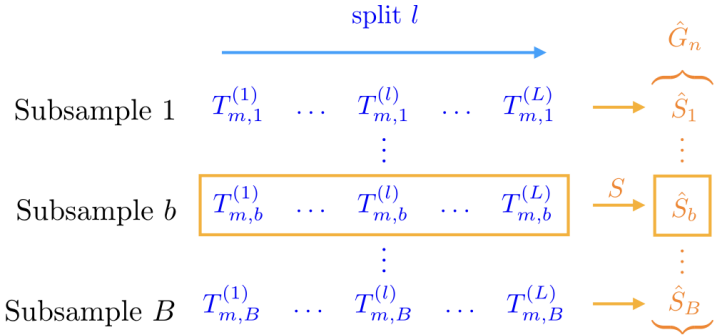
$$S_n := S(T_n^{(1)}, \ldots, T_n^{(L)}), \quad S(\cdot) = \text{average},$$

so that we have a unif$(0,1)$ p-value under H_0.

$\hookrightarrow$ Try to estimate the null joint distribution of $(T_n^{(1)}, \ldots, T_n^{(L)})$ using subsampling.

Subsampling

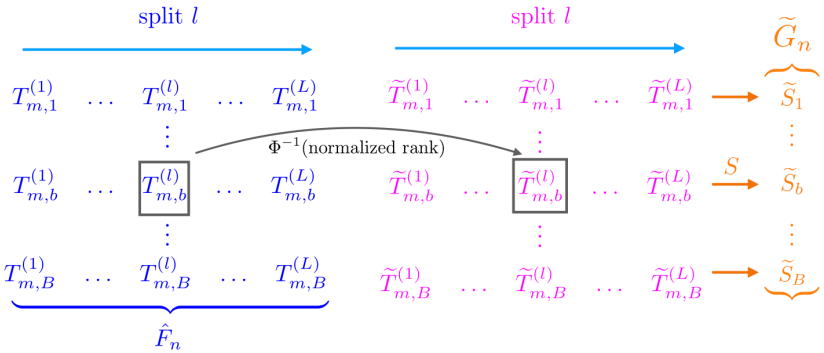
Choose $b = 1, \ldots, B$ subsamples of data each of size $m \approx n/\log n$.



☞ This tends to estimate the sampling distribution of S_n, regardless of whether we are under H_0 or H_1.

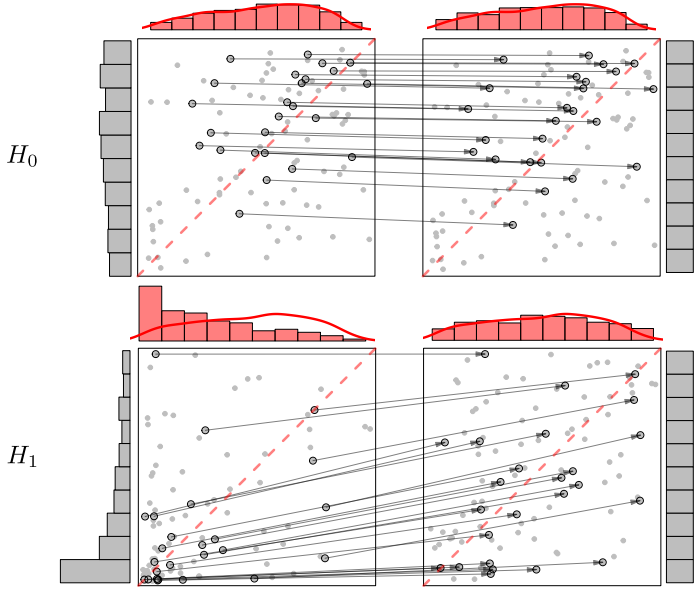
↪ Yet, for testing, we only want the null distribution.

Rank-transformed subsampling: Enforcing the null



☞ Replace each $T_{m,b}^{(l)}$ with

$$\widetilde{T}_{m,b}^{(l)} := \Phi^{-1}\left(\frac{\text{rank} - 1/2}{BL}\right).$$

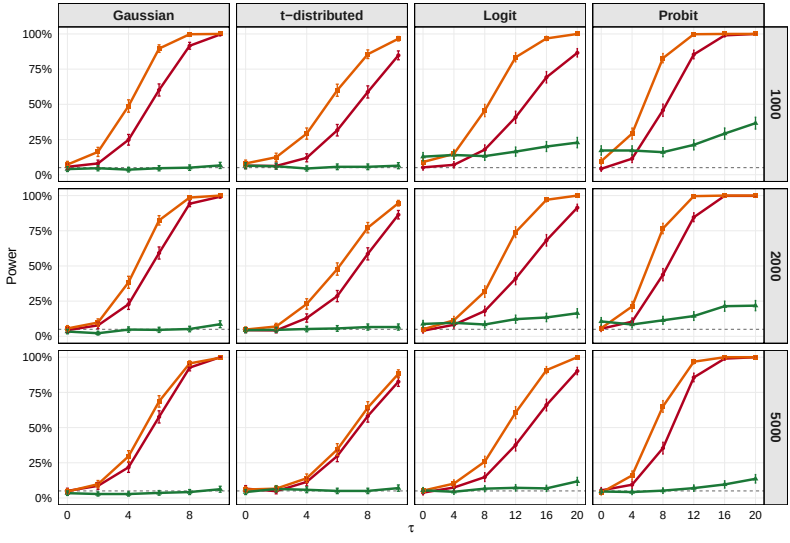

H_0
H_1
before rank transform
after rank transform

Rank-transformed subsampling

Theorem

1. Assume a mild condition on the copula of statistics from different splits. Under H_0, $\widetilde{G}_n$ approximates the null distribution function of S_n.
2. With further regularity conditions, under H_1, $\widetilde{G}_n$ still approximates the null distribution of S_n without first-order bias. ☞ fast rate/1st-order accuracy

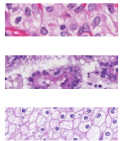
↪ For GAM goodness-of-fit test



Revisiting example: Significance of clustering

ICGC/TCGA Pan-Cancer dataset

111 Kidney Cancer (Renal Cell Carcinoma) Cases



37 Clear Cell (cc)

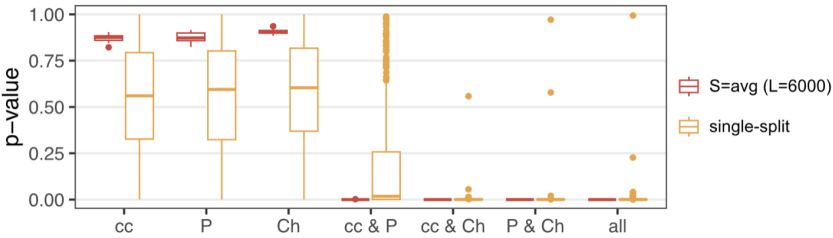
31 Papillary (P)

43 Chromophobe (Ch)

Selected 1000 genes by comparing to control

	Gene 1	Gene 2	Gene 3	...
Cell 1	-1.2	0.5	6.2	
Cell 2	0.1	12	1.1	
Cell 3	-2.2	0	-2	
⋮				

Normalized mRNA expression



THANKS

References I

Cheng, M-Y and Peter Hall (1998). "Calibrating the excess mass and dip tests of modality." In: *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* 60.3, pp. 579–589.

Dukes, Oliver et al. (2024). *Nonparametric tests of treatment effect homogeneity for policy-makers*. arXiv: 2410.00985 [stat.ME]. URL: https://arxiv.org/abs/2410.00985.

Fan, Jianqing and Jiancheng Jiang (2005). "Nonparametric inference for additive models." In: *Journal of the American Statistical Association* 100.471, pp. 890–907.

Fasiolo, Matteo et al. (2021). "Fast calibrated additive quantile regression." In: *Journal of the American Statistical Association* 116.535, pp. 1402–1412.

Gozalo, Pedro L. and Oliver B. Linton (2001). "Testing additivity in generalized nonparametric regression models with estimated parameters." In: *Journal of Econometrics* 104.1, pp. 1–48.

Guo, F Richard and Rajen D Shah (Sept. 2024). "Rank-transformed subsampling: inference for multiple data splitting and exchangeable p-values." In: *Journal of the Royal Statistical Society Series B: Statistical Methodology* 87.1, pp. 256–286. ISSN: 1369-7412. DOI: 10.1093/jrsssb/qkae091. eprint: https://academic.oup.com/jrsssb/article-pdf/87/1/256/59160848/qkae091.pdf. URL: https://doi.org/10.1093/jrsssb/qkae091.

References II

Härdle, Wolfgang, Stefan Sperlich, and Vladimir Spokoiny (2001). "Structural tests in additive regression." In: *Journal of the American Statistical Association* 96.456, pp. 1333–1347.

Hartigan, John A and Pamela M Hartigan (1985). "The dip test of unimodality." In: *The Annals of Statistics*, pp. 70–84.

Huang, Hanwen, Yufeng Liu, and J. S Marron (2022). *SigClust: Statistical Significance of Clustering*. R package version 1.1.0.1. URL: https://CRAN.R-project.org/package=sigclust.

Kennedy, Edward H (2023). "Towards optimal doubly robust estimation of heterogeneous causal effects." In: *Electronic Journal of Statistics* 17.2, pp. 3008–3049.

Sperlich, Stefan, Dag Tjøstheim, and Lijian Yang (2002). "Nonparametric estimation and testing of interaction in additive models." In: *Econometric Theory* 18.2, pp. 197–251.

Tibshirani, Julie et al. (2024). *grf: Generalized Random Forests*. R package version 2.4.0. DOI: 10.32614/CRAN.package.grf. URL: https://CRAN.R-project.org/package=grf.

Williamson, Brian D, Peter B Gilbert, Marco Carone, et al. (2021). "Nonparametric variable importance assessment using machine learning techniques." In: *Biometrics* 77.1, pp. 9–22.

References III


Williamson, Brian D, Peter B Gilbert, Noah R Simon, et al. (2021). "A general framework for inference on algorithm-agnostic variable importance." In: *Journal of the American Statistical Association*, pp. 1–14.