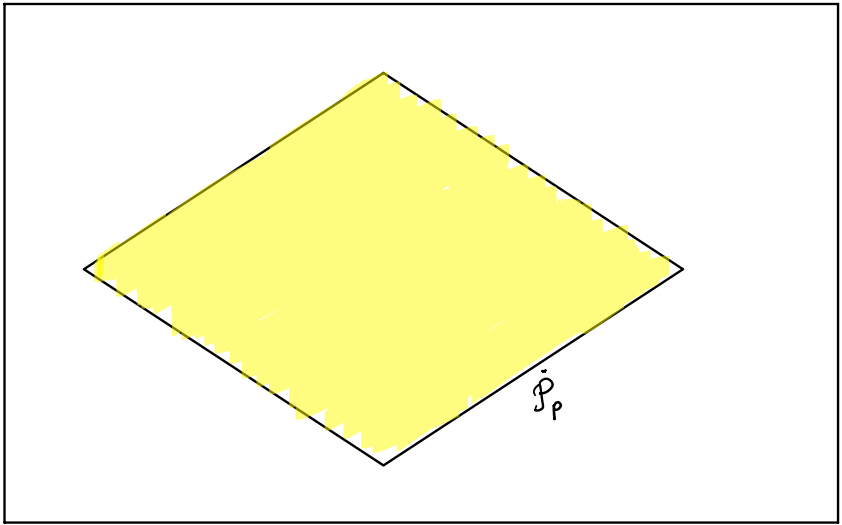
Augmenting Statistical Inference with Machine Learning III

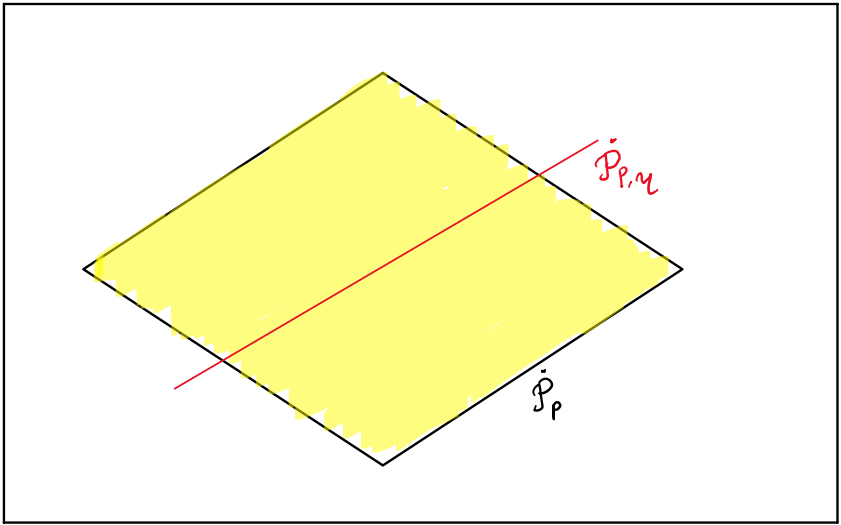
Rajen D. Shah (Statistical Laboratory, University of Cambridge)

ENAR Spring Meeting 2026
18 March 2026

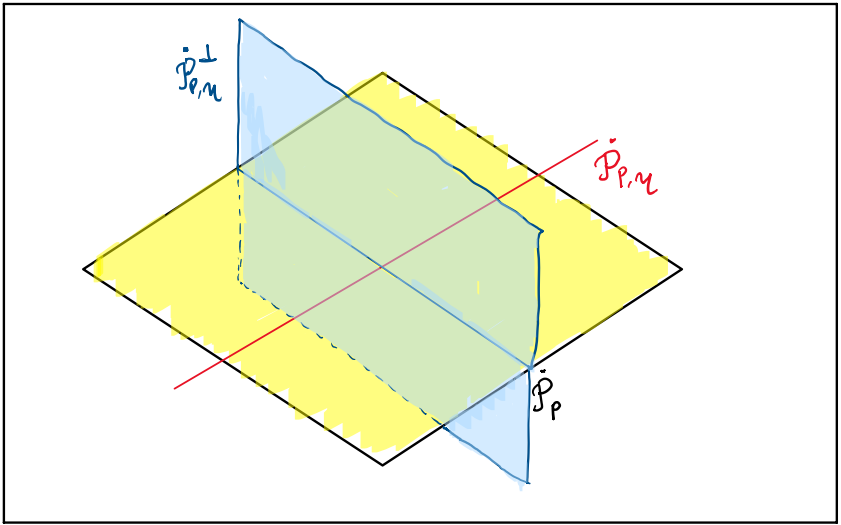
Recap: semiparametric models



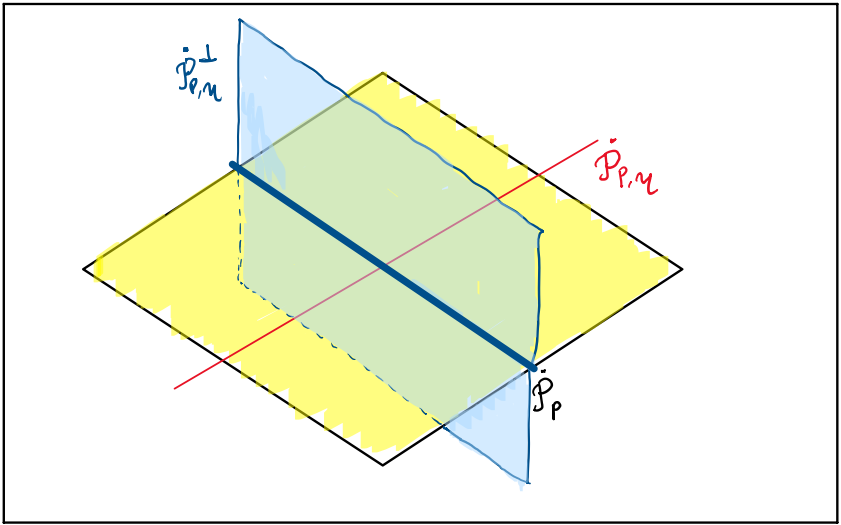
Recap: semiparametric models



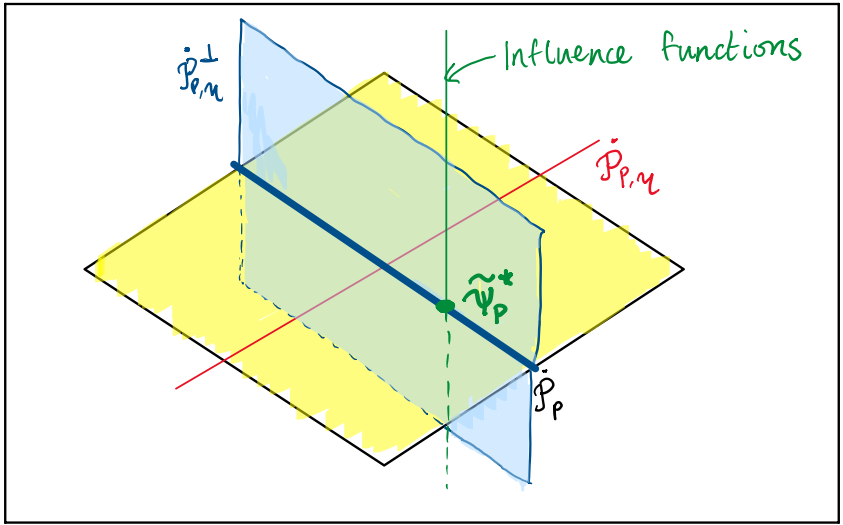
Recap: semiparametric models



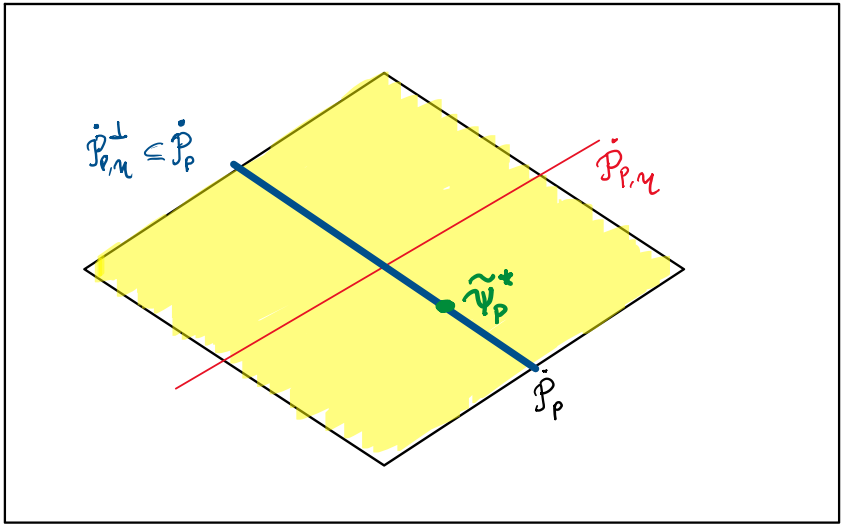
Recap: semiparametric models



Recap: semiparametric models



Recap: nonparametric models (special case)



Overview

Lecture 3

- Multiple sample splitting (Guo and Shah, 2024)
- Goodness-of-fit testing (Dhawan, Guo and Shah, *in preparation*)
- Nonparametric regression (Young, Shah and Samworth, 2026)

Sample splitting

Sample splitting (Moran, 1973; Cox, 1975)

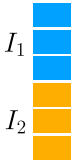
Suppose we are interested in testing a null hypothesis H_0 given iid data.

Different tests may be particularly powerful against different alternatives $P \in H_0^c$.

Sample splitting (Moran, 1973; Cox, 1975)

Suppose we are interested in testing a null hypothesis H_0 given iid data.

Different tests may be particularly powerful against different alternatives $P \in H_0^c$.

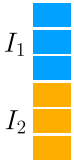


1. **'Hunt'**: Use Part A to determine which test to use to target the alternative the data appear to have come from.
2. **Test**: Apply the test to Part B.

Sample splitting (Moran, 1973; Cox, 1975)

Suppose we are interested in testing a null hypothesis H_0 given iid data.

Different tests may be particularly powerful against different alternatives $P \in H_0^c$.



1. **'Hunt'**: Use Part A to determine which test to use to target the alternative the data appear to have come from.
2. **Test**: Apply the test to Part B.

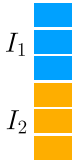
In the second step, we may treat the chosen test as fixed in advance and there is no need to account for the 'hunting' in step 1.

Hunt step can be as elaborate as needed in order to find an appropriate test.

Sample splitting (Moran, 1973; Cox, 1975)

Suppose we are interested in testing a null hypothesis H_0 given iid data.

Different tests may be particularly powerful against different alternatives $P \in H_0^c$.



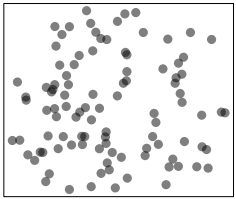
1. **'Hunt'**: Use Part A to determine which test to use to target the alternative the data appear to have come from.
2. **Test**: Apply the test to Part B.

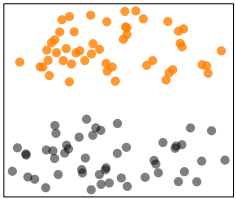
In the second step, we may treat the chosen test as fixed in advance and there is no need to account for the 'hunting' in step 1.

Hunt step can be as elaborate as needed in order to find an appropriate test.

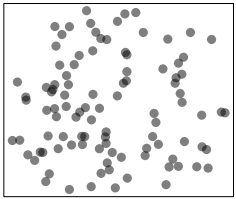
Strategy particularly useful when $H_0 = \cap_{\delta \in \mathcal{D}} H_0(\delta)$, so $H_1 = \cup_{\delta \in \mathcal{D}} H_0^c(\delta)$.

Testing for clustering structure

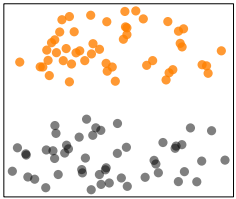




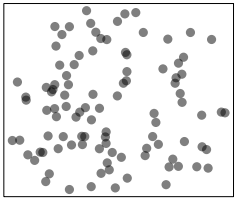
Testing for clustering structure



Clustering algorithms cannot be used directly, as they may return clusters when none are truly present.



Testing for clustering structure



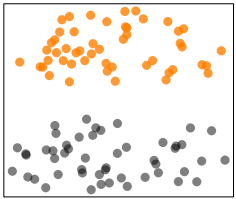
Clustering algorithms cannot be used directly, as they may return clusters when none are truly present.

We can formalise our null hypothesis as testing for unimodality.

Various notions exist in multiple dimensions including *linear unimodality*:

$$X \in \mathbb{R}^p \text{ is unimodal if } a^\top X \text{ is unimodal } \forall\, a \neq 0.$$

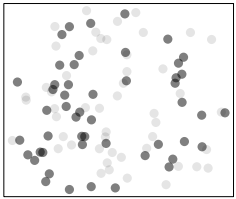
That is,

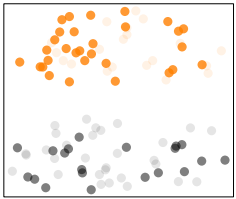


$$H_0 : \bigcap_{a \neq 0} \{a^\top X \text{ is unimodal}\},$$

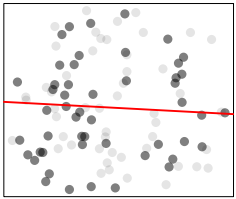
$$H_1 : \exists\, a \neq 0 \text{ such that } a^\top X \text{ is not unimodal.}$$

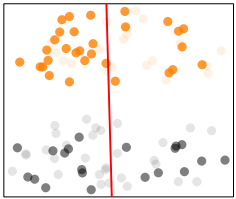
Testing for clustering structure



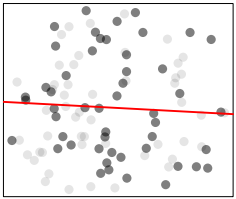


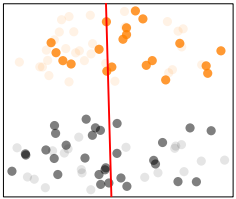
Testing for clustering structure



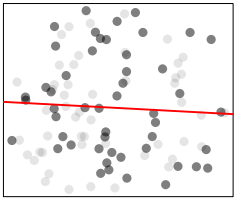


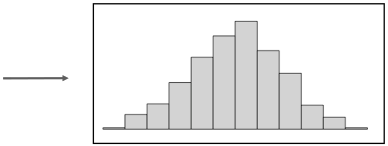
Testing for clustering structure



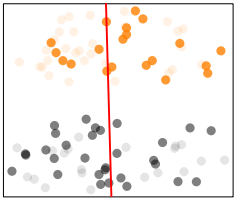


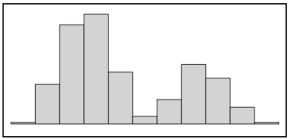
Testing for clustering structure





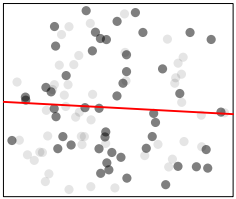
H_0

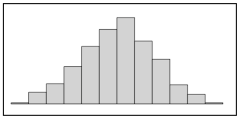




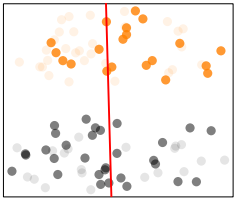
H_1

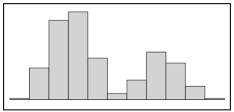
Testing for clustering structure



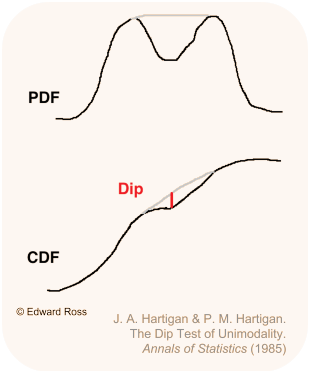


H_0





H_1



J. A. Hartigan & P. M. Hartigan.
The Dip Test of Unimodality.
Annals of Statistics (1985)

Testing for non-zero mean (Kim & Ramdas, 2024)

Let $X_1, \ldots, X_n \in \mathbb{R}^p$ be i.i.d. with mean $\mu \in \mathbb{R}^p$.

We wish to test the null $H_0 : \mu = 0$.

Idea:

1. Form $barX_1 := \sum_{i \in I_1} X_i$,
2. Look at test statistic

$$\frac{|I_2|^{-1/2} \sum_{i \in I_2} \bar{X}_1^\top X_i}{\sqrt{\bar{X}_1^\top \hat{\Sigma}_2 \bar{X}_1}},$$

where $\hat{\Sigma}_2$ is the empirical covariance on the I_2 sample.

Replicability and Power

Replicability: Conclusions may depend delicately on the random seed used.

Power loss: Tests may not be making full use of the data.

Consider repeatedly applying the same randomised procedure to the same data.

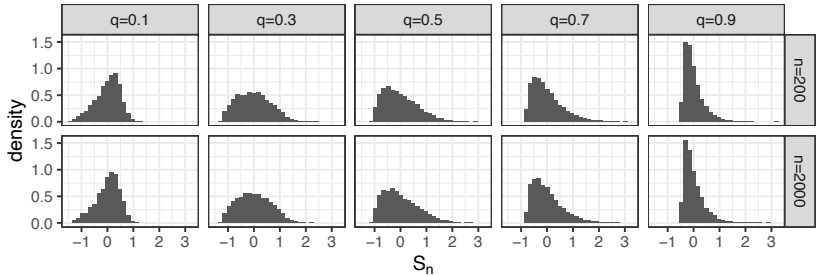
Test statistics $T^{(1)}, \ldots, T^{(L)}$ are exchangeable.

Consider aggregating them by, e.g.,

$$S := \left(T^{(1)} + \cdots + T^{(L)}\right)/L.$$

Complex dependence

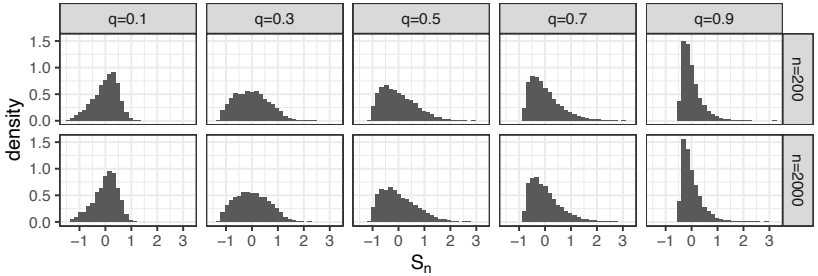
A challenge is that the dependence among $T^{(1)}, \ldots, T^{(L)}$ can be complex and there is no good description or approximation (beyond exchangeability).



Distribution of S in the hunt-and-test example case from Kim & Ramdas (2020)

Complex dependence

A challenge is that the dependence among $T^{(1)}, \ldots, T^{(L)}$ can be complex and there is no good description or approximation (beyond exchangeability).



Distribution of S in the hunt-and-test example case from Kim & Ramdas (2020)

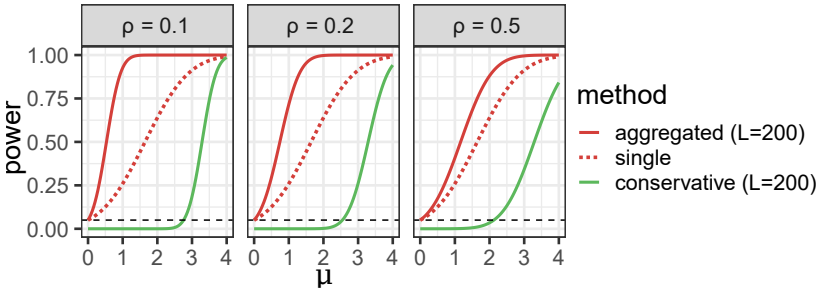
Catering for the worst case loses a lot of power.

Other conservative approaches similarly lose power. (Vovk & Wang, 2020; Vovk et al., 2021; DiCiccio et al., 2020; Meinshausen et al., 2009; ...)

Toy example: Power

Single-split test Reject H_0 when $T^{(1)}$ is large

Aggregated test Reject H_0 when $S = \left(T^{(1)} + \cdots + T^{(L)}\right)/L$ is large.



The conservative rule rejects when $S > 2z_\alpha$. [Can show that $\mathbb{P}(S \geq z_\alpha) = 2\alpha$ is possible for some dependency structures.]

Setup

Have exchangeable test statistics $T_n^{(1)}, \ldots, T_n^{(L)}$.

A1 Under $P \in H_0$, $T_n^{(1)}$ is asymptotically $U(0,1)$.
(Also works for $T_n^{(1)} \xrightarrow{d} \mathcal{N}(0,1)$).

Setup

Have exchangeable test statistics $T_n^{(1)}, \ldots, T_n^{(L)}$.

A1 Under $P \in H_0$, $T_n^{(1)}$ is asymptotically $U(0,1)$.
(Also works for $T_n^{(1)} \xrightarrow{d} \mathcal{N}(0,1)$).

Choose a deterministic Lipschitz aggregation function $S : \mathbb{R}^L \to \mathbb{R}$ to give $S_n := S(T^{(1)}, \ldots, T^{(L)})$.
E.g. $S_n = (T^{(1)} + \cdots + T^{(L)})/L$ or $S_n = \min_l T^{(l)}$.

A2 Under $P \in H_0$, S_n converges to (unknown) distribution G_P with bounded density.

We wish to construct a test / form a p-value based on S_n.

Subsampling (e.g. Politis et al. (1999))

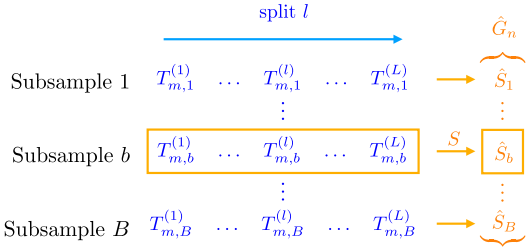
We use subsampling to estimate the asymptotic distribution G_P.

Choose $b = 1, \ldots, B$ subsamples of size $m := \lfloor n / \log n \rfloor$.

We use subsampling to estimate the asymptotic distribution G_P.

Choose $b = 1, \ldots, B$ subsamples of size $m := \lfloor n / \log n \rfloor$. $\hat{G}_n :=$ ECDF of $\{\hat{S}_1, \ldots, \hat{S}_B\}$.

By standard consistency of subsampling (e.g. Politis et al. (1999)) for $P \in H_0$, $\|\hat{G}_n - G_P\|_\infty \xrightarrow{p} 0$.



Subsampling (e.g. Politis et al. (1999))

We use subsampling to estimate the asymptotic distribution G_P.

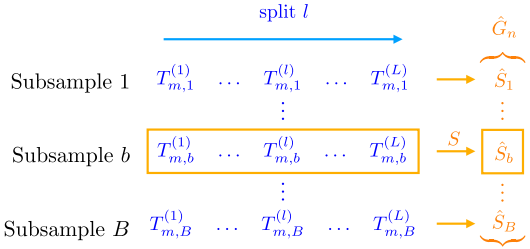
Choose $b = 1, \ldots, B$ subsamples of size $m := \lfloor n / \log n \rfloor$.

We use subsampling to estimate the asymptotic distribution G_P.

Choose $b = 1, \ldots, B$ subsamples of size $m := \lfloor n / \log n \rfloor$. $\hat{G}_n :=$ ECDF of $\{\hat{S}_1, \ldots, \hat{S}_B\}$.

By standard consistency of subsampling (e.g. Politis et al. (1999)) for $P \in H_0$, $\|\hat{G}_n - G_P\|_\infty \xrightarrow{p} 0$.

But under alternatives, $\hat{G}_n$ may be shifted away from the null distribution, leading to a loss in power.



Rank transform

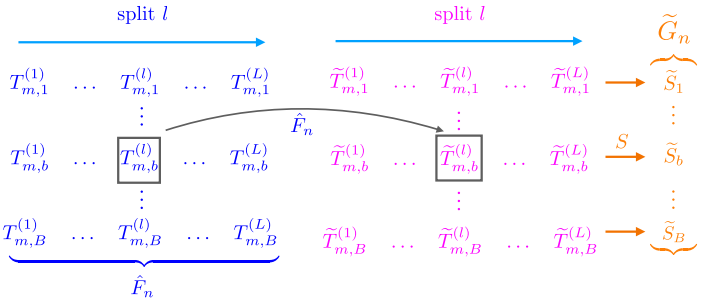
We have not yet used that we *know* the asymptotic null distribution of $T_n^{(1)}$ (A1).

Rank transform

We have not yet used that we *know* the asymptotic null distribution of $T_n^{(1)}$ (A1).

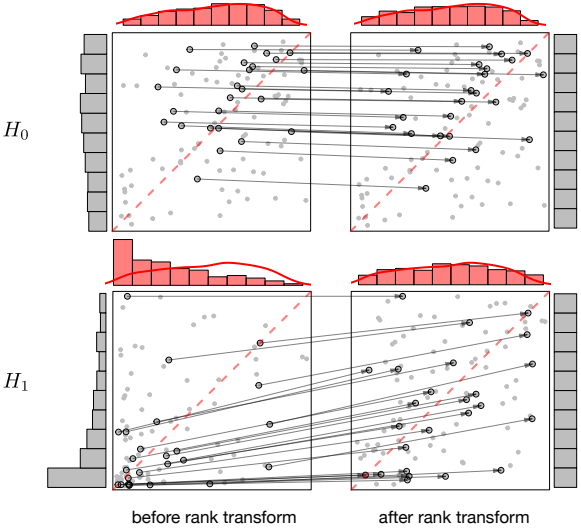
Writing $\hat{F}_n$ for the ECDF of $\{T_{m,b}^{(l)}\}$, replace each $T_{m,b}^{(l)}$ by its normalised rank within the matrix:

$$\tilde{T}_{m,b}^{(l)} := |\{T_{m,b'}^{(l')} \leq T_{m,b}^{(l)}\}|/BL = \hat{F}_n\left(T_{m,b}^{(l)}\right).$$

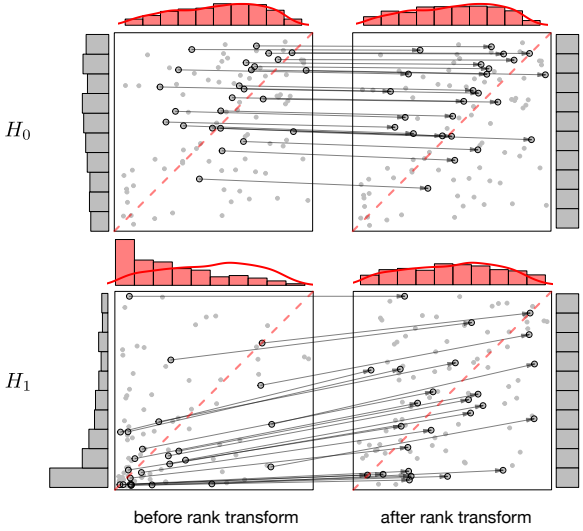


We use $\tilde{G}_n :=$ ECDF of $\{\tilde{S}_b\}$ as our reference for testing.

Rank transform



Rank transform

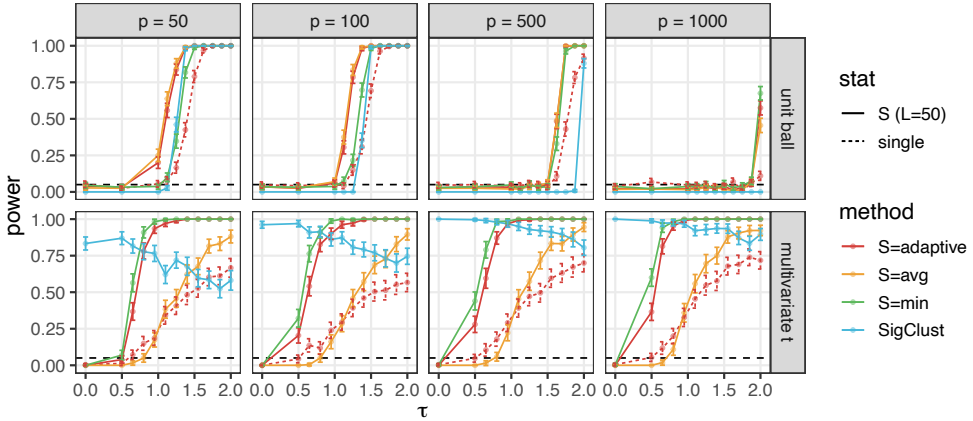


We show that:

- the approach has (uniform) asymptotic size control;
- one has power against local alternatives.

Testing unimodality

Hunt: 2-means clustering, Test: T_n = asymptotic dip test p-value. $L = 50$ splits.
Setting: Mixture of two p-dimensional (unit ball, multivariate t) distributions separated τ away.
Aggregation: Consider $S = \text{avg}$, $S = \min$.

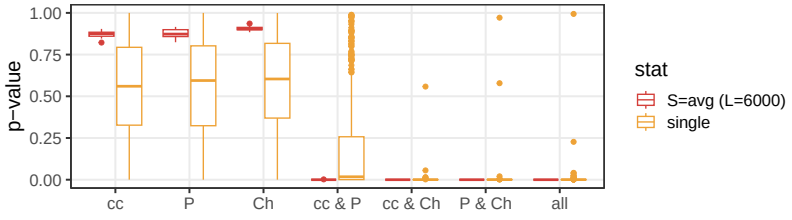


SigClust is a competing method based on multivariate normal mixture.

Gene expression of cancer subtypes

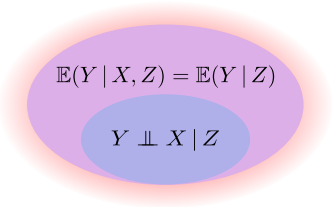
Three types of renal cell carcinoma:
clear cell (cc), papillary (P) and chromophobe (Ch).

ICGC/TCGA Pan-Cancer dataset: Expression levels of 1,000 genes. $L = 6000$ splits.



Testing semiparametric models

Variable significance testing

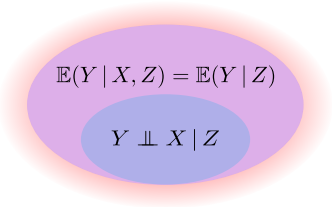


Model-free null hypothesis of conditional mean independence: $\mathbb{E}(Y \mid X, Z) = \mathbb{E}(Y \mid Z)$.

Contrast this with the null of conditional independence $Y \perp\!\!\!\perp X \mid Z$ which asks for $F_{Y|X,Z} = F_{Y|Z}$.

Even the smaller null $Y \perp\!\!\!\perp X \mid Z$ is however still fundamentally hard to test.

Variable significance testing



Model-free null hypothesis of conditional mean independence: $\mathbb{E}(Y \mid X, Z) = \mathbb{E}(Y \mid Z)$.

Contrast this with the null of conditional independence $Y \perp\!\!\!\perp X \mid Z$ which asks for $F_{Y|X,Z} = F_{Y|Z}$.

Even the smaller null $Y \perp\!\!\!\perp X \mid Z$ is however still fundamentally hard to test.

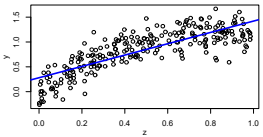


Figure: One way of restricting the null is via models...

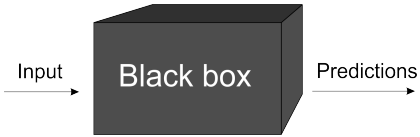


Figure: ...or we can attempt to leverage the predictive power of flexible regression methods.

Estimating the mean squared difference in regression functions

Williamson et al. (2021a) propose to estimate

$$\tau := \mathbb{E}[\{\mathbb{E}(Y \mid X, Z) - \mathbb{E}(Y \mid Z)\}^2] = \mathbb{E}[\{Y - \mathbb{E}(Y \mid Z)\}^2] - \mathbb{E}[\{Y - \mathbb{E}(Y, \mid X, Z)\}^2]$$

via

$$\hat{\tau} := \frac{1}{n}\sum_{i=1}^{n}\{Y_i - \hat{m}_{Y|Z}(Z_i)\}^2 - \frac{1}{n}\sum_{i=1}^{n}\{Y_i - \hat{m}_{Y|X,Z}(X_i, Z_i)\}^2.$$

$\hat{\tau}$ is asymptotically Gaussian centred on τ and achieves the semiparametric efficient variance bound provided $\tau > 0$.

However the functional τ has influence function $\tilde{\psi}^*_P = 0$ when $\tau = 0$. Consequently corresponding $\hat{\tau}$ becomes degenerate under the null.

Williamson et al. (2021b) and Dai et al. (2021) propose variants involving sample-splitting and adding noise to the statistic. However these sacrifice power to obtain a tractable limiting distribution when $\tau = 0$.

Perspectives on goodness-of-fit testing

Models offer an often necessary simplification in the face of the curse of dimensionality.

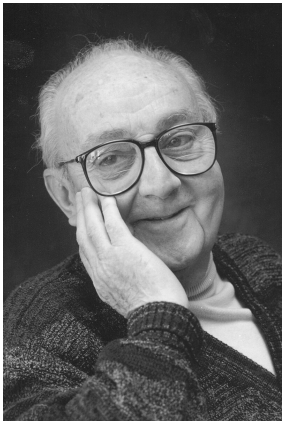
Yet as we well know, *All models are wrong...*

When is a model *useful*?

A necessary condition is that the data does not provide strong evidence against its assumptions.

Otherwise we risk producing confidence intervals around parameters that have no meaningful interpretation.

Goodness-of-fit tests provide a convenient and formal way of assessing this evidence.



Perspectives on goodness-of-fit testing

Models offer an often necessary simplification in the face of the curse of dimensionality.

Yet as we well know, *All models are wrong...*

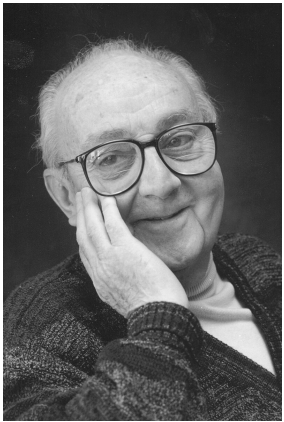
When is a model *useful*?

A necessary condition is that the data does not provide strong evidence against its assumptions.

Otherwise we risk producing confidence intervals around parameters that have no meaningful interpretation.

Goodness-of-fit tests provide a convenient and formal way of assessing this evidence.

Goodness-of-fit encompasses variable significance: consider the (semiparametric) model $\mathbb{E}(Y \mid X, Z) = \mathbb{E}(Y \mid Z)$.



Some existing work

Most existing work involves:

- Choosing bandwidth parameters in smoothing: Fan and Li (1996); Sperlich et al. (2002); Fan and Jiang (2005);...
- Bootstrap for critical values: Li and Wang (1998); Stute et al. (1998); Gu et al. (2007); Meintanis and Einbeck (2012);...
- 'Classical' nonparametric regression of Y on X: Dette et al. (2001); Gozalo and Linton (2001); Härdle et al. (2001);...

We would ideally like a method that is:

- relatively easy to use,
- can scale to moderate-dimensional settings,
- avoids bootstrapping,
- avoids choices of bandwidths that may be hard to make.

A general framework

Consider a class of null hypotheses characterised by

$$\arg\min_{f} \mathbb{E}\ell(f(X), Y) \in \mathcal{F}, \qquad \text{where}$$

- ℓ is a loss function convex in its first argument
- $\mathcal{F}$ is a class of functions closed under pointwise addition and scaling.

A general framework

Consider a class of null hypotheses characterised by

$$\arg\min_f \mathbb{E}\ell(f(X), Y) \in \mathcal{F}, \qquad \text{where}$$

- ℓ is a loss function convex in its first argument
- $\mathcal{F}$ is a class of functions closed under pointwise addition and scaling.

For this talk, we consider the special case of testing the conditional mean specification

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

and μ is a known differentiable strictly increasing inverse link function.

This may be cast in the general framework by setting $\ell(\eta, y) = -\eta y + K(\eta)$ where K is an antiderivative of μ.

A general framework

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

- $\mathcal{F} = \left\{ x \mapsto \sum_j f_j(x_j) \right\}$
- $X = (T, Z)$ and $\mathcal{F} = \{(t, z) \mapsto f(z)\}$
- As above but $\mathcal{F} = \{(t, z) \mapsto \theta t + f(z) : \theta \in \mathbb{R}\}$
- As above, but for some subset S of variables $\mathcal{F} = \{(t, z) \mapsto f(t, z_S) + g(z)\}$

Encompasses generalised additive models, (generalised) partially linear models, varying coefficient models,...

Exposing lack of fit

Starting point: The null is equivalent to $f^* := \arg\min_{f \in \mathcal{F}} \mathbb{E}\ell(f(X), Y)$ satisfying

$$\mathbb{E}\left[\{Y - \mu(f^*(X))\}\, h(X)\right] = 0 \qquad \text{for all functions } h$$

Exposing lack of fit

Starting point: The null is equivalent to $f^* := \arg\min_{f \in \mathcal{F}} \mathbb{E}\ell(f(X), Y)$ satisfying

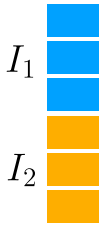
$$\mathbb{E}\left[\{Y - \mu(f^*(X))\}\, h(X)\right] = 0 \qquad \text{for all functions } h \qquad \mathbb{E}[\ell'(f^*(X), Y)h(X)] = 0.$$

Exposing lack of fit

Starting point: The null is equivalent to $f^* := \arg\min_{f \in \mathcal{F}} \mathbb{E}\ell(f(X), Y)$ satisfying

$$\mathbb{E}\left[\{Y - \mu(f^*(X))\}\, h(X)\right] = 0 \qquad \text{for all functions } h \qquad \mathbb{E}[\ell'(f^*(X), Y)h(X)] = 0.$$

Idea: Given i.i.d. data $(X_i, Y_i)_{i=1}^{2n}$, split this into two equal parts and perform the following:



1. **'Hunt'**: With observations I_1, fit the null model to yield estimate $\tilde{f}$, and search for $\hat{h}$ that has positive correlation with the residuals $Y_i - \mu(\tilde{f}(X_i))$.
2. **'Test'**: With observations I_2, fit the null model once more to yield estimate $\hat{f} \in \mathcal{F}$ and consider

$$L_i := \left\{ Y_i - \mu(\hat{f}(X_i)) \right\} \hat{h}(X_i), \qquad T_n := \frac{\frac{1}{\sqrt{n}} \sum_{i \in I_2} L_i}{\left\{ \frac{1}{n} \sum_{i \in I_2} L_i^2 - \left(\frac{1}{n} \sum_{i \in I_2} L_i \right)^2 \right\}^{1/2}}$$

Related work: Escanciano (2024) and Sancetta (2022) look at a test statistic based on averaging $(\mathbb{E}[\{Y - \mu(f(X))\}h(X)])^2$ over different h lying in an RKHS.

Testing

Let $v_L := \mathrm{Var}(L_i)$. $T_n \xrightarrow{d} \mathcal{N}(0,1)$ under the null relies on

$$\sqrt{\frac{n}{v_L}}\mathbb{E}\left[\left\{Y_i - \mu(\hat{f}(X_i))\right\}\hat{h}(X_i)\right] = \sqrt{\frac{n}{v_L}}\mathbb{E}\left[\left\{\mu(f^*(X_i)) - \mu(\hat{f}(X_i))\right\}\hat{h}(X_i)\right] \approx 0.$$

Testing

Let $v_L := \mathrm{Var}(L_i)$. $T_n \xrightarrow{d} \mathcal{N}(0,1)$ under the null relies on

$$\sqrt{\frac{n}{v_L}}\mathbb{E}\left[\left\{Y_i - \mu(\hat{f}(X_i))\right\}\hat{h}(X_i)\right] = \sqrt{\frac{n}{v_L}}\mathbb{E}\left[\left\{\mu(f^*(X_i)) - \mu(\hat{f}(X_i))\right\}\hat{h}(X_i)\right] \approx 0.$$

Now by Cauchy–Schwarz,

$$\frac{n}{v_L}\left(\mathbb{E}\left[\left\{\mu(f^*(X_i)) - \mu(\hat{f}(X_i))\right\}\hat{h}(X_i)\right]\right)^2 \leq \underbrace{n\mathbb{E}\left[\left\{\mu(f^*(X_i)) - \mu(\hat{f}(X_i))\right\}^2\right]}_{\to 0?} \cdot \underbrace{\frac{\mathbb{E}\hat{h}^2(X_i)}{v_L}}_{=O(1)}$$

Testing

Let $v_L := \mathrm{Var}(L_i)$. $T_n \xrightarrow{d} \mathcal{N}(0,1)$ under the null relies on

$$\sqrt{\frac{n}{v_L}}\mathbb{E}\left[\left\{Y_i - \mu(\hat{f}(X_i))\right\}\hat{h}(X_i)\right] = \sqrt{\frac{n}{v_L}}\mathbb{E}\left[\left\{\mu(f^*(X_i)) - \mu(\hat{f}(X_i))\right\}\hat{h}(X_i)\right] \approx 0.$$

Now by Cauchy–Schwarz,

$$\frac{n}{v_L}\left(\mathbb{E}\left[\left\{\mu(f^*(X_i)) - \mu(\hat{f}(X_i))\right\}\hat{h}(X_i)\right]\right)^2 \leq \underbrace{n\mathbb{E}\left[\left\{\mu(f^*(X_i)) - \mu(\hat{f}(X_i))\right\}^2\right]}_{\to 0?} \cdot \underbrace{\frac{\mathbb{E}\hat{h}^2(X_i)}{v_L}}_{=O(1)}$$

Idea: Try to make $\hat{h}$ orthogonal to the bias term

$$\mu(f^*(X_i)) - \mu(\hat{f}(X_i)) \approx \mu'(f^*(X_i)) \cdot \underbrace{\{f^*(X_i) - \hat{f}(X_i)\}}_{=\Delta(X_i) \text{ for some } \Delta \in \mathcal{F}}$$

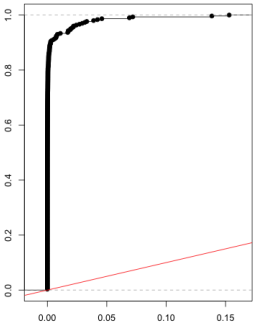


Figure: ECDF of p-values under null!

Testing

Let $\mathcal{F}_w := \{\mu'(f^*(\cdot))f(\cdot) : f \in \mathcal{F}\}$,

$\mathcal{F}_w^\perp := \{g : \mathbb{E}g(X)f(X) = 0 \text{ for all } f \in \mathcal{F}_w\}$. Seek $\hat{h} \in \mathcal{F}_w^\perp$.

Testing

Let $\mathcal{F}_w := \{\mu'(f^*(\cdot))f(\cdot) : f \in \mathcal{F}\}$,
$\mathcal{F}_w^\perp := \{g : \mathbb{E}g(X)f(X) = 0 \text{ for all } f \in \mathcal{F}_w\}$. Seek $\hat{h} \in \mathcal{F}_w^\perp$.

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

We can form an estimate $\hat{m}_{\hat{h}}$ of $m_{\hat{h}}$ through a weighted least squares regression minimising

$$\sum_{i \in I_2} \mu'(\hat{f}(X_i))\left\{\hat{h}(X_i) - g(X_i)\right\}^2 \qquad \text{over } g \in \mathcal{F}.$$

Testing

We should then have

$$\frac{n}{v_L} \cdot \left(\mathbb{E}[\{\mu(f^*(X)) - \mu(\hat{f}(X))\}\{\hat{h}(X) - \hat{m}_{\hat{h}}(X)\}] \right)^2 \approx 0.$$

Indeed, this is approximately

$$\frac{n}{v_L} \cdot \left(\mathbb{E}[\{f^*(X) - \hat{f}(X)\}\mu'(f^*(X))\{\hat{h}(X) - \hat{m}_{\hat{h}}(X)\}] \right)^2$$
$$= \frac{n}{v_L} \cdot \left(\mathbb{E}[\{f^*(X) - \hat{f}(X)\}\mu'(f^*(X))\{\hat{h}(X) - \hat{m}_{\hat{h}}(X) - \hat{h}(X) + m_{\hat{h}}(X)\}] \right)^2$$

Testing

We should then have

$$\frac{n}{v_L} \cdot \left(\mathbb{E}[\{\mu(f^*(X)) - \mu(\hat{f}(X))\}\{\hat{h}(X) - \hat{m}_{\hat{h}}(X)\}] \right)^2 \approx 0.$$

Indeed, this is approximately

$$\begin{aligned} &\frac{n}{v_L} \cdot \left(\mathbb{E}[\{f^*(X) - \hat{f}(X)\}\mu'(f^*(X))\{\hat{h}(X) - \hat{m}_{\hat{h}}(X)\}] \right)^2 \\ &= \frac{n}{v_L} \cdot \left(\mathbb{E}[\{f^*(X) - \hat{f}(X)\}\mu'(f^*(X))\{m_{\hat{h}}(X) - \hat{m}_{\hat{h}}(X)\}] \right)^2. \end{aligned}$$

Testing

We should then have

$$\frac{n}{v_L} \cdot \left(\mathbb{E}[\{\mu(f^*(X)) - \mu(\hat{f}(X))\}\{\hat{h}(X) - \hat{m}_{\hat{h}}(X)\}] \right)^2 \approx 0.$$

Indeed, this is approximately

$$\frac{n}{v_L} \cdot \left(\mathbb{E}[\{f^*(X) - \hat{f}(X)\}\mu'(f^*(X))\{\hat{h}(X) - \hat{m}_{\hat{h}}(X)\}] \right)^2$$
$$= \frac{n}{v_L} \cdot \left(\mathbb{E}[\{f^*(X) - \hat{f}(X)\}\mu'(f^*(X))\{m_{\hat{h}}(X) - \hat{m}_{\hat{h}}(X)\}] \right)^2.$$

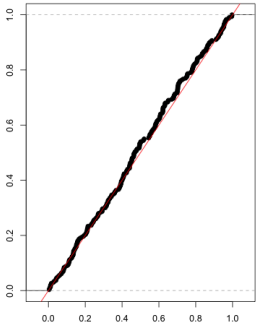
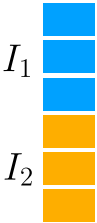


Figure: ECDF of p-values under null

Summary of approach



1. **'Hunt'**: With observations I_1, hunt for signal in residuals via weighted least squares of
   - response $\{Y_i - \mu(\tilde{f}(X_i))\}^{-1}$ onto X_i
   - with weights $\{Y_i - \mu(\tilde{f}(X_i))\}^2$.

   using your favourite machine learning method.
   Compute 'refinement' to get $\hat{h}$.

2. **'Test'**: With observations I_2:
   - fit the null model to yield estimate $\hat{f} \in \mathcal{F}$;
   - estimate $m_{\hat{h}}$ via (regularised) weighted least squares of $\hat{h}(X_i)$ with weights $\mu'(\hat{f}(X_i))$ over function class $\mathcal{F}$ to give debiased hunted function $\hat{h} - \hat{m}_{\hat{h}}$.

   Form

$$L_i := \left\{ Y_i - \mu(\hat{f}(X_i)) \right\} \left\{ \hat{h}(X_i) - \hat{m}_{\hat{h}}(X_i) \right\}, \quad T_n := \frac{\frac{1}{\sqrt{n}} \sum_{i \in I_2} L_i}{\left\{ \frac{1}{n} \sum_{i \in I_2} L_i^2 - \left(\frac{1}{n} \sum_{i \in I_2} L_i \right)^2 \right\}^{1/2}}$$

Generalised Additive Models

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

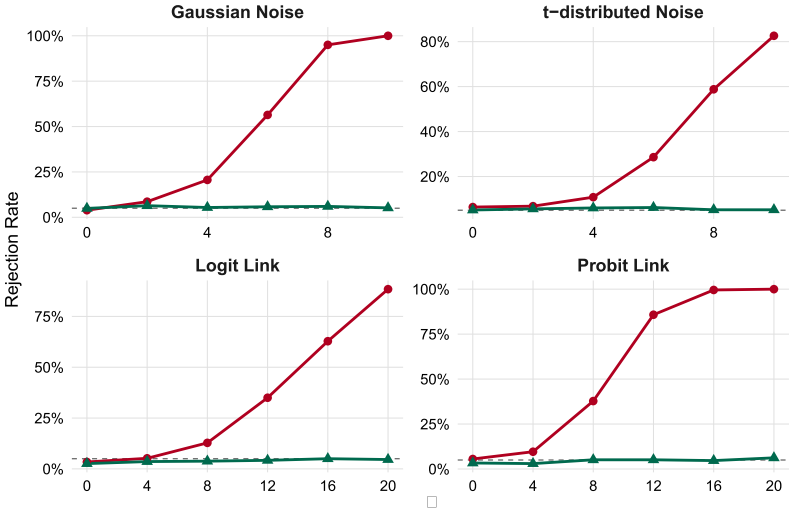
Binary case: $\Pr(Y = 1 \mid X) = \mu\Big(\sin(2X_1) + \frac{\tau}{\sqrt{n}} X_1 X_3\Big)$

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

Use `grf` (Tibshirani J, Athey S, Sverdrup E, Wager S (2024)) for hunting.

Compare to Williamson et al. (2021) which compares the predictive performance of two regression models (GAM and `grf`).

GAM, $n = 5000$


Gaussian Noise
t−distributed Noise
Logit Link
Probit Link
Rejection Rate
100%
75%
50%
25%
0%
80%
60%
40%
20%
0
4
8
12
16
20
test
Debiased Score Test
Williamson et al.

Nonparametric regression

Revisiting regression

Consider a linear model

$$Y_i = X_i^\top \beta + \varepsilon_i.$$

- In a normal linear model, ordinary least squares corresponds to maximum likelihood estimation: $\Rightarrow$ OLS is efficient.

Revisiting regression

Consider a linear model

$$Y_i = X_i^\top \beta + \varepsilon_i.$$

- In a normal linear model, ordinary least squares corresponds to maximum likelihood estimation: $\Rightarrow$ OLS is efficient.
- Other error distributions yield different efficient estimators e.g. Laplace errors $\Rightarrow$ least absolute deviation.

Revisiting regression

Consider a linear model

$$Y_i = X_i^\top \beta + \varepsilon_i.$$

- In a normal linear model, ordinary least squares corresponds to maximum likelihood estimation: $\Rightarrow$ OLS is efficient.
- Other error distributions yield different efficient estimators e.g. Laplace errors $\Rightarrow$ least absolute deviation.
- Can one try to *learn* the error distribution and perform maximum likelihood estimation on an estimated likelihood?

Revisiting regression

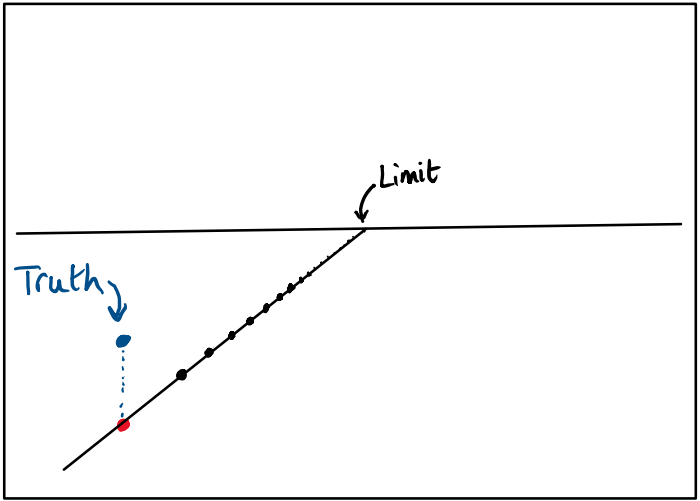
Consider a linear model

$$Y_i = X_i^\top \beta + \varepsilon_i.$$

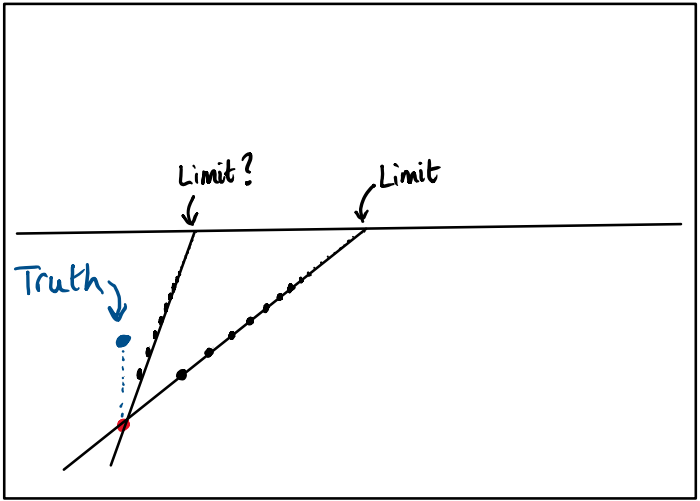
- In a normal linear model, ordinary least squares corresponds to maximum likelihood estimation: $\Rightarrow$ OLS is efficient.
- Other error distributions yield different efficient estimators e.g. Laplace errors $\Rightarrow$ least absolute deviation.
- Can one try to *learn* the error distribution and perform maximum likelihood estimation on an estimated likelihood?
- No, without making further assumptions. Moreover, OLS is an efficient estimator of the best linear predictor in a nonparametric model, i.e. the efficient estimator of

$$\arg\min_{\beta \in \mathbb{R}^p} \mathbb{E}\{(Y - X^\top \beta)^2\}.$$

Asymptotics again



Asymptotics again



Nadaraya–Watson-type estimators

Consider the local constant estimator, approximating $f \equiv \mu$ in a local region.

Let $h > 0$ and take kernel $K_h : [-h, h] \to \mathbb{R}$. Consider the estimators solving:

$$\text{Least squares: (Nadaraya, 1964; Watson, 1964)} \quad \sum_{i=1}^{n} K_h(X_i - x)(Y_i - \mu) = 0,$$

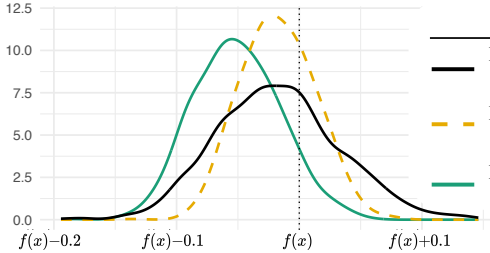
$$\text{Local likelihood (MLE): (Tibshirani and Hastie, 1987)} \quad \sum_{i=1}^{n} K_h(X_i - x)\rho(Y_i - \mu \,|\, X_i) = 0,$$

$$\text{Plug-in local likelihood:} \quad \sum_{i=1}^{n} K_h(X_i - x)\hat{\rho}(Y_i - \mu \,|\, X_i) = 0,$$

where

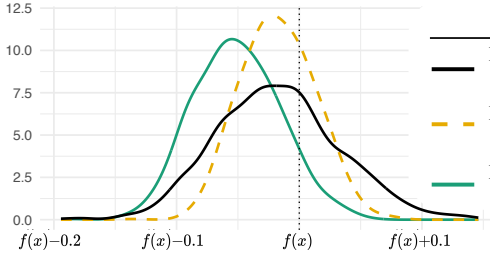
- ρ is the conditional score function $\rho(\varepsilon \,|\, x) := \frac{\partial_\varepsilon p_{\varepsilon|X}(\varepsilon \,|\, x)}{p_{\varepsilon|X}(\varepsilon \,|\, x)}$.
- $\hat{\rho}$ is an estimator for the conditional score (using auxiliary data).

Score plug-in bias



Estimator	MSE $\hat{f}(x_0)$
Least squares/ Local polynomial	3.0×10^{-3}
Local likelihood/ 'Oracle'	1.3×10^{-3}
Plug-in-score local likelihood	3.6×10^{-3}

Score plug-in bias



Estimator	MSE $\hat{f}(x_0)$
Least squares/ Local polynomial	3.0×10^{-3}
Local likelihood/ 'Oracle'	1.3×10^{-3}
Plug-in-score local likelihood	3.6×10^{-3}

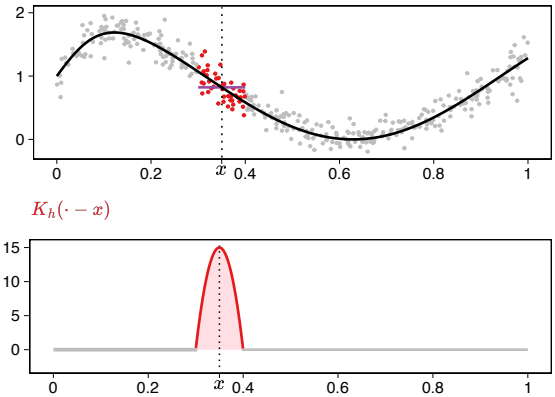
'Plugging in' an estimated score $\Rightarrow$ **worse MSE** than least squares

Issue: Plug-in incurs an additional bias term, behaving like

$$\mathbb{E}[K_h(X_i - x)\{\hat{\rho} - \rho\}(\varepsilon_i \,|\, X_i)] \approx \mathbb{E}[K_h(X_i - x)] \cdot \mathbb{E}[\hat{\rho}(\varepsilon_i \,|\, X_i) \,|\, X_i = x]$$

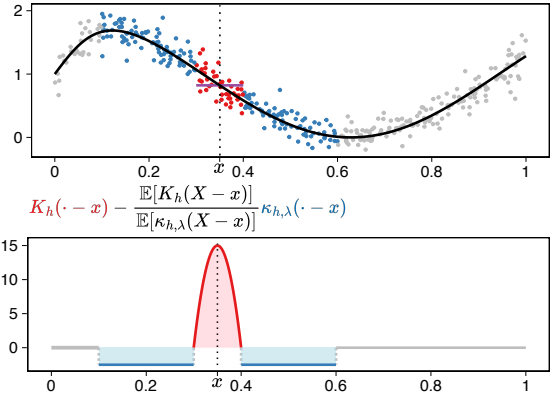
Idea: 'Centre' $K_h(X_i - x)$.

Local centring



We use the data within $[x - h, x + h]$ for our local polynomial approximation,

Local centring



We use the data within $[x - h, x + h]$ for our local polynomial approximation, and data in an expanded $[x - \lambda h, x + \lambda h] \setminus [x - h, x + h]$ for centring.

‘Outrigger’ estimation

Given an auxiliary estimator $\hat{\rho}$, we solve the estimating equation

$$\sum_{i=1}^{n} \big(K_h(X_i - x) - \hat{\gamma}(x)\kappa_{h,\lambda}(X_i - x)\big)\hat{\rho}\big(\tilde{\varepsilon}_i(\mu) \,\big|\, X_i\big) = 0,$$

$$\tilde{\varepsilon}_i(\mu) := \begin{cases} Y_i - \mu & \text{if } X_i \in x + [-h, h] \\ Y_i - \tilde{f}_{\text{pilot}}(X_i) & \text{if } X_i \in x + [-\lambda h, \lambda h] \setminus [-h, h] \end{cases}$$

with $\hat{\gamma}(x)$ an empirical estimator of $\gamma(x) := \mathbb{E}(\kappa_{h,\lambda}(X - x))^{-1}\mathbb{E}(K_h(X - x))$,
and $\tilde{f}_{\text{pilot}}$ a (debiased) pilot estimator for f.

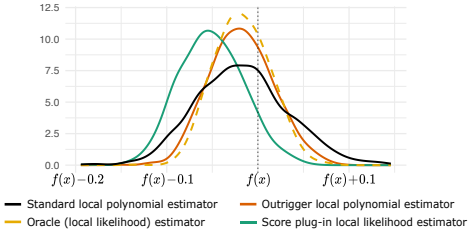
‘Outrigger’ estimation

Given an auxiliary estimator $\hat{\rho}$, we solve the estimating equation

$$\sum_{i=1}^{n} \big(K_h(X_i - x) - \hat{\gamma}(x)\kappa_{h,\lambda}(X_i - x)\big)\hat{\rho}\big(\tilde{\varepsilon}_i(\mu) \,\big|\, X_i\big) = 0,$$

$$\tilde{\varepsilon}_i(\mu) := \begin{cases} Y_i - \mu & \text{if } X_i \in x + [-h, h] \\ Y_i - \tilde{f}_{\text{pilot}}(X_i) & \text{if } X_i \in x + [-\lambda h, \lambda h] \setminus [-h, h] \end{cases}$$

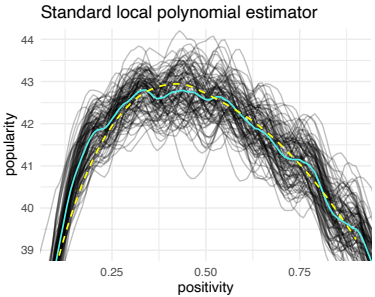
with $\hat{\gamma}(x)$ an empirical estimator of $\gamma(x) := \mathbb{E}(\kappa_{h,\lambda}(X - x))^{-1}\mathbb{E}(K_h(X - x))$, and $\tilde{f}_{\text{pilot}}$ a (debiased) pilot estimator for f.

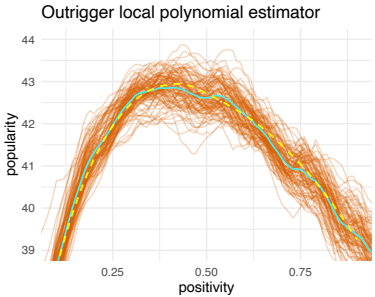


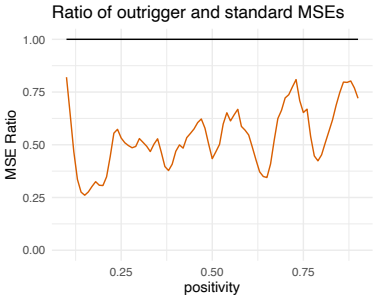
Spotify data

Dataset of $\sim 100\,000$ songs on Spotify.

Local polynomial and outrigger (both local constant) fit on subsamples.







Cyan: Average of subsampled estimators.

Yellow dashed: Full sample local quadratic estimator.

Thank you for listening.