When bias hurts - a tale of nonparametric testing

Martin Hazelton[1]
Tilman Davies

Department of Mathematics & Statistics
University of Otago

22 November 2023

[1] Email: martin.hazelton@otago.ac.nz

Comparing Parametric Things

Two (multivariate) random samples: $\{\boldsymbol{x}_{ij} : j = 1, \ldots, n_i\}$, $i = 1, 2$

Want to compare underyling quantities θ_1 and θ_2

By STAT200 or some such:

- $H_0 : \theta_1 = \theta_2$
- Define test statistic
$$Z = \frac{\hat{\theta}_1 - \hat{\theta}_2}{\mathrm{SE}(\hat{\theta}_1 - \hat{\theta}_2)}$$
- Find p-value using approximately $N(0, 1)$ null distribution for Z

Works fine in most standard parametric settings

Comparing Parametric Things

Why bias doesn't hurt

Standard parameter estimators have:

- $\text{Bias}(\hat{\theta}) = O(n^{-1})$
- $\text{SE}(\hat{\theta}) = O(n^{-1/2})$

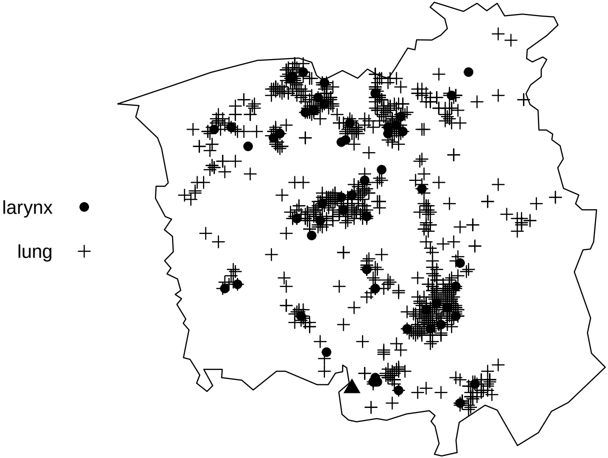
Any bias in $\hat{\theta}_1 - \hat{\theta}_2$ will lead to

$$\text{E}[Z] = \text{E}\left[\frac{\hat{\theta}_1 - \hat{\theta}_2}{\text{SE}(\hat{\theta}_1 - \hat{\theta}_2)}\right] = O(n^{-1/2})$$

- $n_1 = \gamma n$, $n_2 = (1-\gamma)n$
- Often under H_0 get perfect bias cancellation, so $E[Z] = 0$.

Comparing Nonparametric Things

Do larynx and lung cancer densities differ at incinerator location ▲?



Comparing Multivariate Densities at a Point

$H_0: f_1(\boldsymbol{x}) = f_2(\boldsymbol{x})$

- Also include equality of all derivatives at $\boldsymbol{x}$ under H_0?

Test statistic

$$z(\boldsymbol{x}) = \frac{\hat{f}_1(\boldsymbol{x}|H_1) - \hat{f}_2(\boldsymbol{x}|H_2)}{\hat{\sigma}(\boldsymbol{x})}$$

using kernel density estimates

$$\hat{f}_i(\boldsymbol{x}|H_i) = \frac{1}{n_i|H_i|^{1/2}} \sum_{j=1}^{n_i} K(H_i^{-1/2}(\boldsymbol{x} - \boldsymbol{x}_{ij}))$$

- H_1, H_2 are bandwidth matrices;
- $\hat{\sigma}(\boldsymbol{x})$ is asymptotic estimate of $\mathrm{SE}(\hat{f}_1(\boldsymbol{x}|H_1) - \hat{f}_2(\boldsymbol{x}|H_2))$.

Some Asymptotics

Individual density estimates

Bandwidth matrices vary with n: $H_i = C_i n_i^{-2\alpha}$

$$\mathrm{E}[\hat{f}_i(\boldsymbol{x}|H_i)] = f_i(\boldsymbol{x}) + \tfrac{1}{2}\mathrm{tr}\{C_i \mathcal{H}_{f_i}(\boldsymbol{x})\} n_i^{-2\alpha} + O(n_i^{-4\alpha})$$

- $\mathcal{H}_{f_i}(\boldsymbol{x})$ is Hessian matrix for f_i

$$\mathrm{Var}[\hat{f}_i(\boldsymbol{x}|H_i)] = R(K) f_i(\boldsymbol{x}) |C_i|^{-1/2} n_i^{\alpha d - 1} + o(n_i^{\alpha d - 1}).$$

To minimize MSE($\hat{f}_i$), set $\alpha = 1/(4+d)$ for d-dimensional data

Then Bias($\hat{f}_i$) $= O(n^{-2/(4+d)})$ and $\sqrt{\mathrm{Var}(\hat{f}_i)} = O(n^{-2/(4+d)})$

Some Asymptotics

Test statistic null distribution

Theorem

Under standard regularity conditions, under H_0

$$z(\boldsymbol{x}) - \mathrm{E}[z(\boldsymbol{x})] \xrightarrow{D} \mathrm{N}(0, 1).$$

So null distribution is asymptotically standard normal iff $\mathrm{E}[z(\boldsymbol{x})] = 0$.

UNIVERSITY of OTAGO
Te Whare Wānanga o Otāgo
NEW ZEALAND

Some Asymptotics

Test statistic null mean

Using the traditionally optimal bandwidth order:

$\mathrm{E}[\hat{f}_1(\boldsymbol{x}|H_1) - \hat{f}_2(\boldsymbol{x}|H_2)] = O(n^{-2/(4+d)})$

$\hat{\sigma}(\boldsymbol{x}) = O(n^{-2/(4+d)})$

Therefore

$$\mathrm{E}[z(\boldsymbol{x})] = \mathrm{E}\left[\frac{\hat{f}_1(\boldsymbol{x}|H_1) - \hat{f}_2(\boldsymbol{x}|H_2)}{\hat{\sigma}(\boldsymbol{x})}\right] = O(1)$$

UNIVERSITY of OTAGO
Te Whare Wānanga o Otāgo
NEW ZEALAND

Some Asymptotics

But surely the biases cancel under H_0? If only ...

Recall:

$$\mathrm{E}[\hat{f}_i(\boldsymbol{x}|H_i)] = f_i(\boldsymbol{x}) + \frac{1}{2}\mathrm{tr}\{C_i \mathcal{H}_{f_i}(\boldsymbol{x})\} n_i^{-2\alpha} + O(n_i^{-4\alpha})$$

and

$$n_1 = \gamma n, n_2 = (1-\gamma)n$$

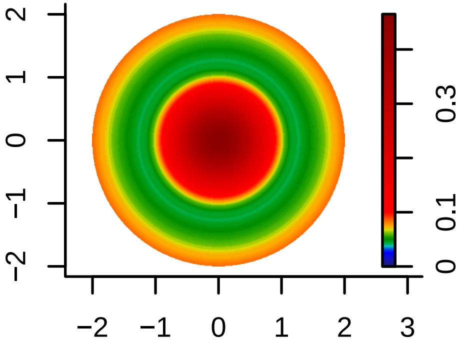
With 'optimal' $\alpha = 1/(4+d)$:

Hessians	Sample size split	$E[z(\boldsymbol{x})]$
$\mathcal{H}_{f_1}(\boldsymbol{x}) \neq \mathcal{H}_{f_2}(\boldsymbol{x})$	$0 < \gamma < 1$	$O(1)$
$\mathcal{H}_{f_1}(\boldsymbol{x}) = \mathcal{H}_{f_2}(\boldsymbol{x})$	$\gamma \neq 1/2$	$O(1)$
$\mathcal{H}_{f_1}(\boldsymbol{x}) = \mathcal{H}_{f_2}(\boldsymbol{x})$	$\gamma = 1/2$	$O(n^{-2/(d+4)})$

UNIVERSITY of OTAGO
Te Whare Wānanga o Otāgo
NEW ZEALAND

But Does it Matter in Practice?

Pointwise test-sizes for comparison of bivariate t-distributions with $n_1 = 100$, $n_2 = 5000$



Solutions

Choose your bandwidth matrices carefully

Optimal

Use different ‘optimal’ bandwidth matrices for each sample

Common

Use the same bandwidth matrix for each sample, but with asymptotically optimal order

- Bias cancellation when Hessians equal

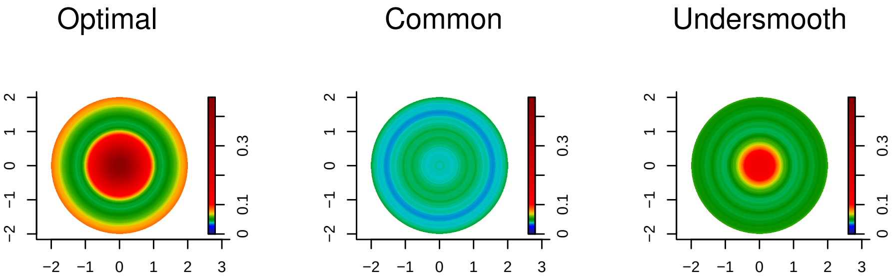
Undersmooth

Use different bandwidth matrices for each sample, but with undersmoothed asymptotic order

- $\mathrm{E}[z(\boldsymbol{x})] = o(1)$ under H_0 regardless of Hessian (in)equality

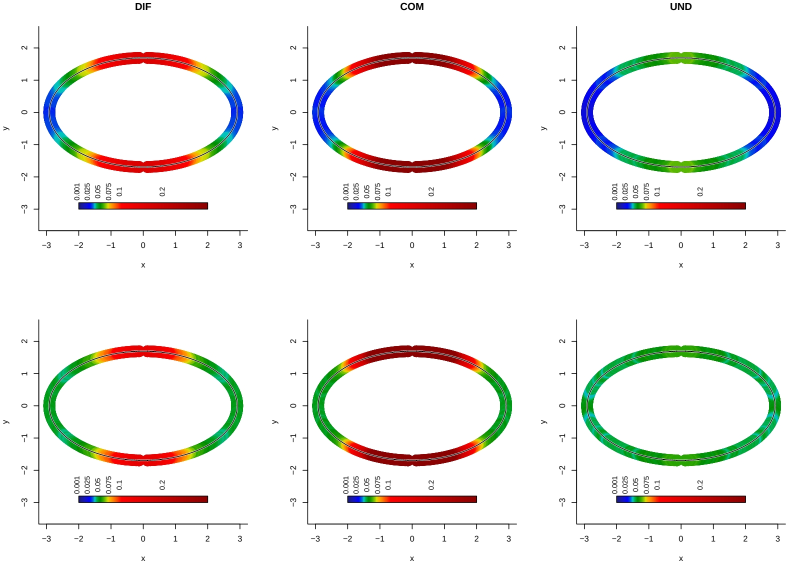
Does it Work in Practice?

Pointwise test-sizes for common distributions, $n_1 \neq n_2$



Does it Work in Practice?

Pointwise test-sizes for different distributions, $\mathcal{H}_{f_1}(\boldsymbol{x}) \neq \mathcal{H}_{f_2}(\boldsymbol{x})$



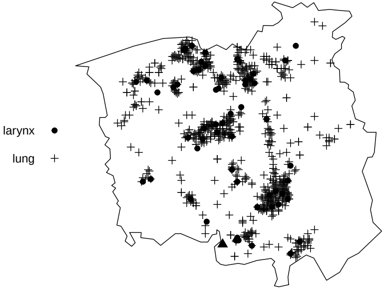
Which Method to Choose?

Its context dependent

Hessians	Sample size split	Bandwidth matrices
$\mathcal{H}_{f_1}(\boldsymbol{x}) \neq \mathcal{H}_{f_2}(\boldsymbol{x})$	$0 < \gamma < 1$	Undersmooth
$\mathcal{H}_{f_1}(\boldsymbol{x}) = \mathcal{H}_{f_2}(\boldsymbol{x})$	$\gamma \neq 1/2$	Common
$\mathcal{H}_{f_1}(\boldsymbol{x}) = \mathcal{H}_{f_2}(\boldsymbol{x})$	$\gamma = 1/2$	Optimal

Lessons from Chorley-Ribble

Test size and power from simulated data



	Test method		
	OPT	COM	UND
Test size	0.112	0.030	0.031
Test power	0.460	0.582	0.311

To Learn More ...

Hazelton, M. L., & Davies, T. M. (2022). Pointwise comparison of two multivariate density functions. *Scandinavian Journal of Statistics*, in press. `https://doi.org/10.1111/sjos.12565`