Model selection invalidates goodness-of-fit tests

When the best isn't good enough

Joshua Loftus Weichi Yao

New York University

2018

Work in progress

- The title is the take home message for now...
- It may seem obvious (to us), but examples are interesting
- Goodness-of-fit test vs selection algorithm: maximally contradictory!
- Progress on selective inference / conditional corrections
- What are the interesting scientific directions?
- Appreciate any feedback! Here or `loftus@nyu.edu`

Outline

1. Time series
   - Autoregressive model, AICc, and the Ljung-Box test
2. Linear regression
   - F-tests of unselected variables
   - Orthogonal case / many means
3. Conclusion
   - The quiet scandal for goodness-of-fit tests
   - Converging on standard practices?

Simple time series example: the AR(p) process

Autoregressive model set up

An AR(p) process $\{X_t\}$ with mean μ satisfies

$$(X_t - \mu) - \phi_1(X_{t-1} - \mu) - \ldots - \phi_p(X_{t-p} - \mu) = W_t$$

where $W_t \sim \mathcal{N}(0, \sigma^2)$. For simplicity, we consider the zero-mean autoregressive process with $\mu = 0$

$$X_t - \phi_1 X_{t-1} - \ldots - \phi_p X_{t-p} = W_t,$$

where $W_t \sim \mathcal{N}(0, \sigma^2)$.

Example AR(p) paths

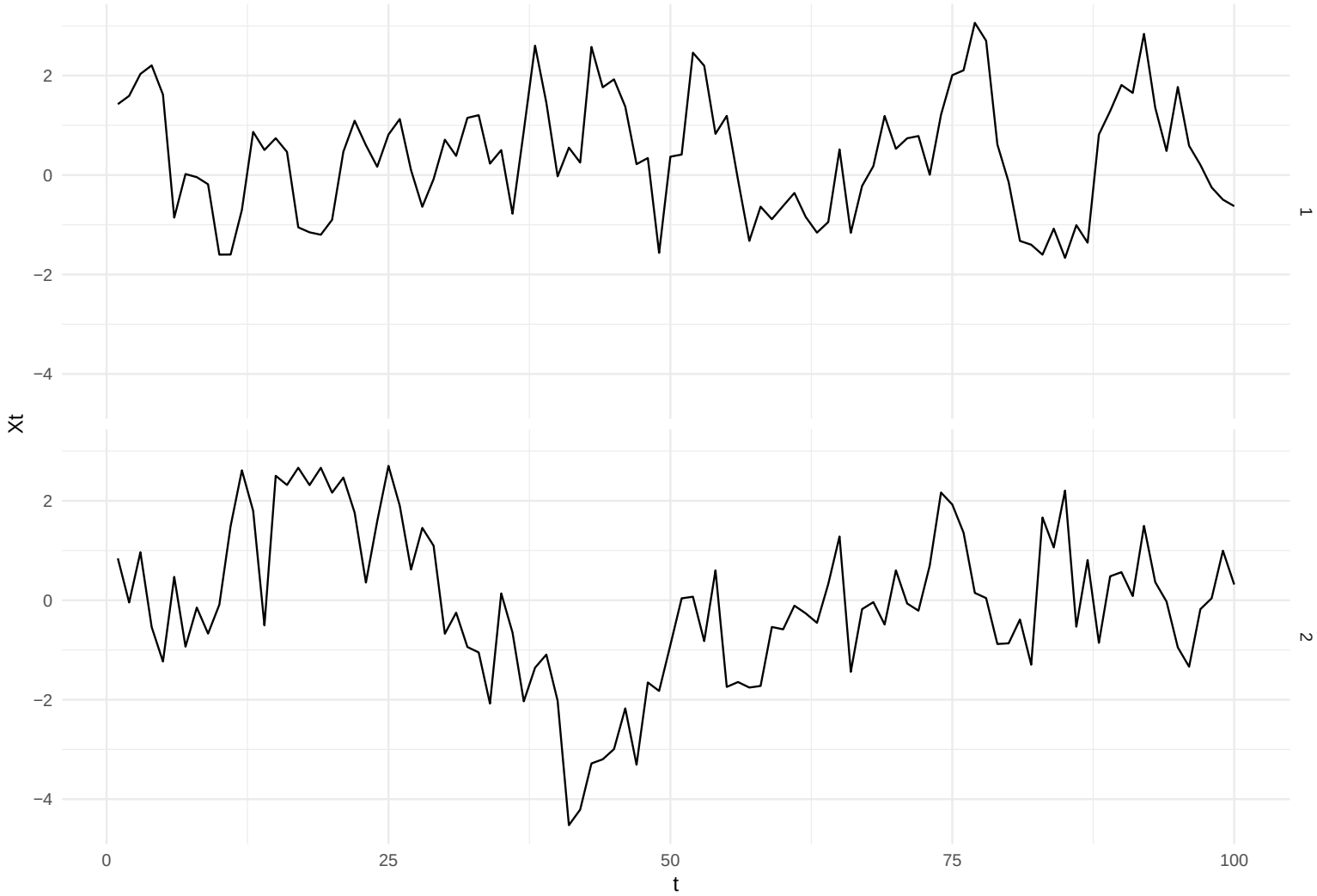


Figure: Top panel: AR(1), bottom panel: AR(2)

Goodness of fit test

Perhaps the most commonly used goodness of fit test in time series:

Ljung-Box test [?]

The Ljung-Box statistic with l lags is defined as

$$Q^{(l)}(\widehat{r}) = n(n+2)\sum_{k=1}^{l}(n-k)^{-1}\widehat{\mathrm{r}}(k)^2$$

where

$$\widehat{\mathrm{r}}(k) = \frac{\sum_{t=k+1}^{n}(Y_t - \widehat{Y}_t)(Y_{t-k} - \widehat{Y}_{t-k})}{\sum_{t=1}^{n}(Y_t - \widehat{Y}_t)^2}, \quad k = 1, ..., m$$

is the autocorrelation function.

If the data is truly AR(p) but we fit AR(q), with $q < p$, the LB test is powered to detect the residual^2 autocorrelation.

With or without selection?

In an ideal world, the order p would be chosen *a priori*, and failing to reject the LB test would be evidence that the order was indeed large enough to capture the time dependence of the errors.

But we don't live in an ideal world...

In practice, people will use the data to choose p.

Model estimation and model selection

To fit an autoregressive model to the data, the parameters ϕ_t are usually estimated through Yule-Walker/Burg/Least-Squares procedures, and the order p is chosen by minimizing the AIC/$\mathrm{AIC_C}$/BIC. For concreteness we consider the $\mathrm{AIC_C}$

$\mathrm{AIC_C}$ statistic for an autoregressive model

To choose the order p, we use the proposal of [**?**]

$$\mathrm{AIC_C} = n(\log \widehat{\sigma}_m^2 + 1) + \frac{2n(m+1)}{n-m-2}. \tag{1}$$

where $\widehat{\sigma}_m^2$ can be obtained by the Yule-Walker method, or some other asymptotically equivalent method.

An example: AR(3)

Suppose our true model is an AR(3) process:

$$Y_{t+1} = \phi_1 Y_t + \phi_2 Y_{t-1} + \phi_3 Y_{t-2} + \varepsilon_{t+1}, \quad \text{for } t \geq 3,$$

where $\varepsilon_{t+1} \sim \mathcal{N}(0, \sigma_3^2)$. In our example,

$$\phi_1 = -0.33, \phi_2 = -0.17, \phi_3 = -0.12.$$

If we choose the order 3 then estimate the model as AR(3),

$$Y_{t+1} = \widehat{\phi}_{31} Y_t + \widehat{\phi}_{32} Y_{t-1} + \widehat{\phi}_{33} Y_{t-2}$$

with estimated variance $\widehat{\sigma}_3^2$. If we choose the order 2, then

$$Y_{t+1} = \widehat{\phi}_{21} Y_t + \widehat{\phi}_{22} Y_{t-1}$$

with estimated variance $\widehat{\sigma}_2^2$, or the order 1 and as AR(1)

$$Y_{t+1} = \widehat{\phi}_{11} Y_t$$

with estimated variance $\widehat{\sigma}_1^2$.

A selection event example

By the $\mathrm{AIC_C}$ statistic defined in (1), order 2 is chosen if and only if

$$\mathrm{AIC_C}(3) > \mathrm{AIC_C}(2) < \mathrm{AIC_C}(1) \Leftrightarrow c_3\widehat{\sigma}_3^2 > c_2\widehat{\sigma}_2^2 < c_1\widehat{\sigma_1^2}$$

where c_p are constants.

We now show simulated distributions of the LB test statistic under different selection events in the following figures, first without selection and then with selection.

Remember, we reject for large values of the LB statistic.

Observed LB vs null distribution, without selection

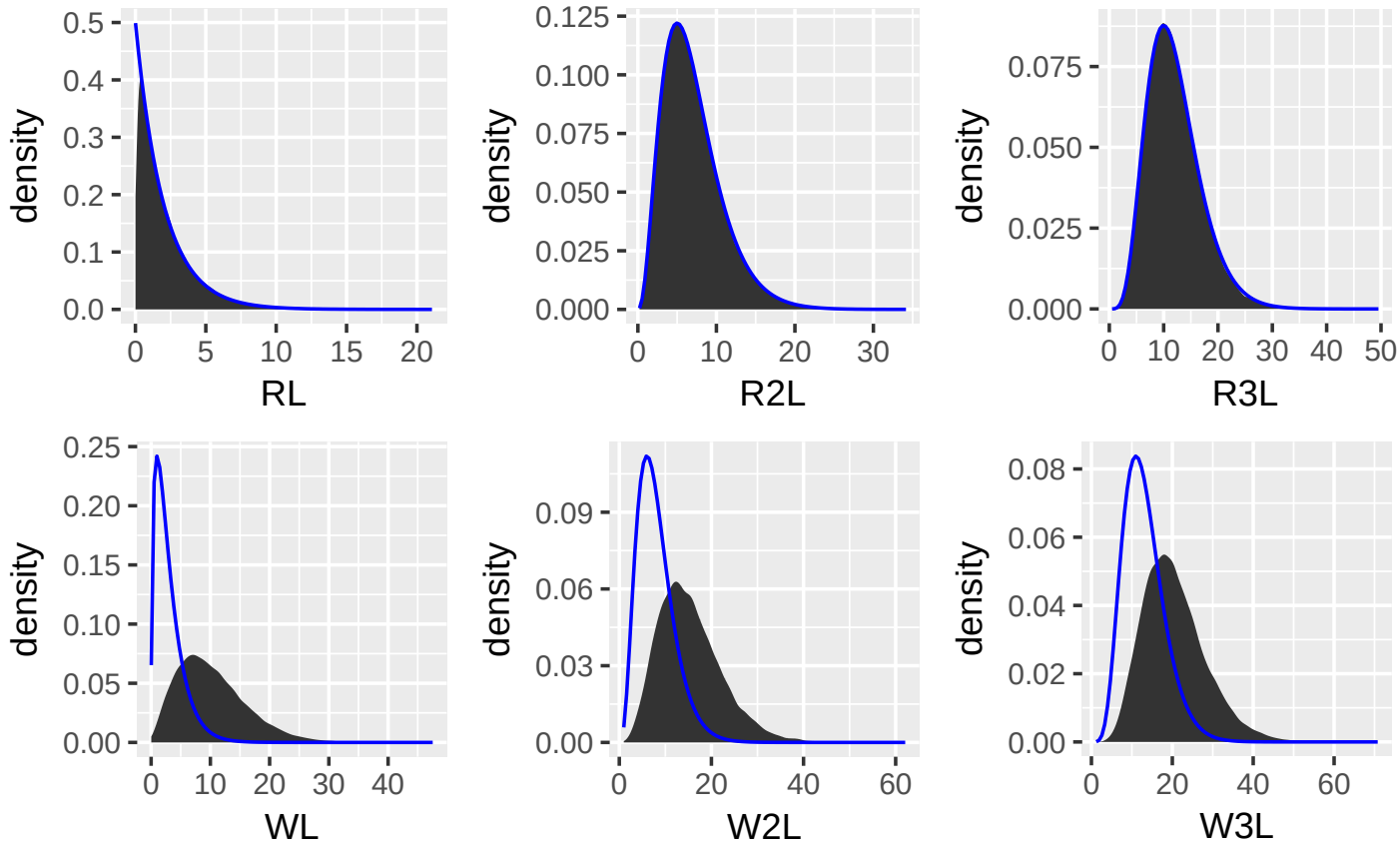


Figure: Truth is AR(3). Top row: fitted with $p = 3$. Bottom row: fitted with $p = 2$. Distributions shown for LB with lags 5, 10, 15.

Observed LB vs null distribution, with selection

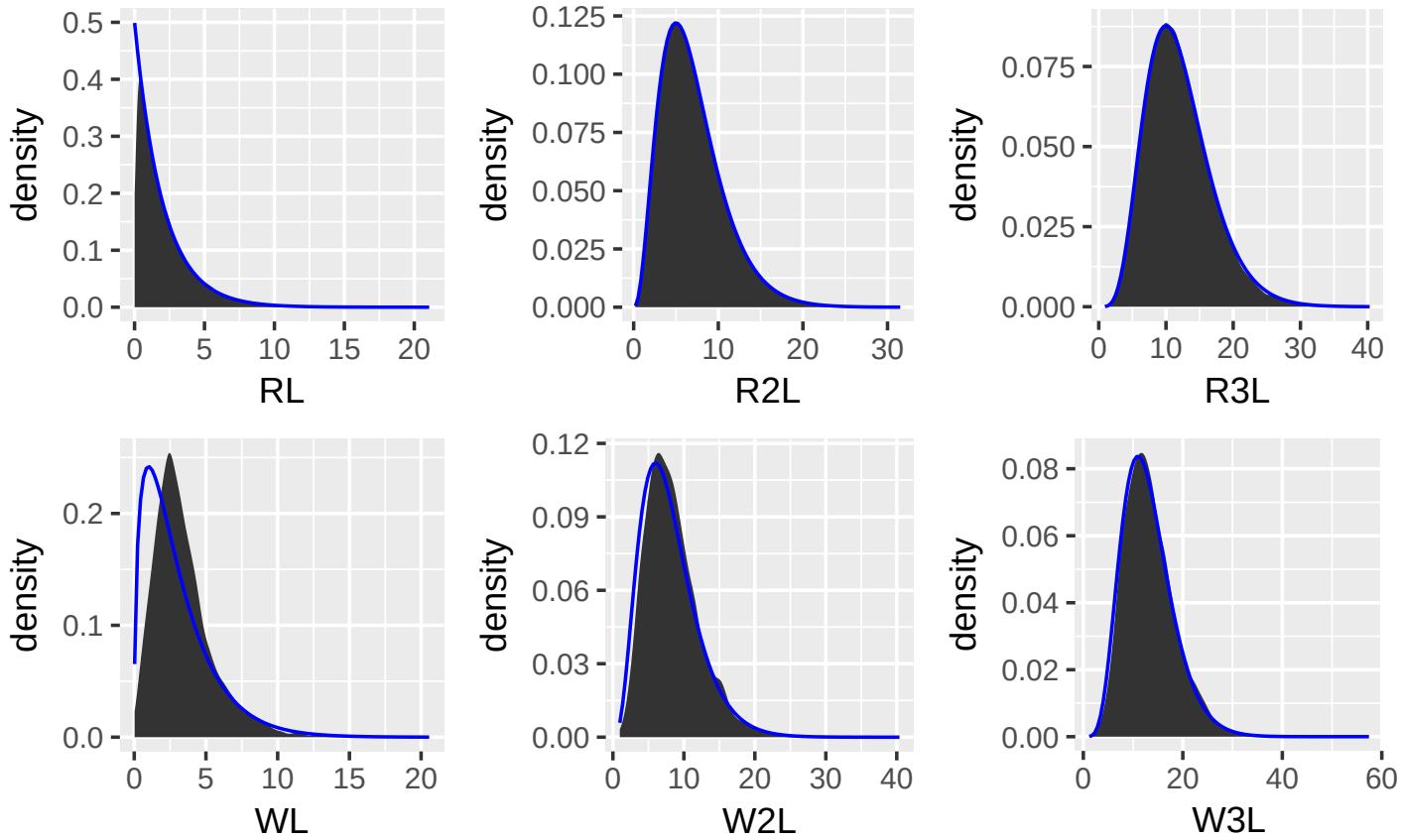


Figure: Truth is AR(3). Top row: when $\mathrm{AIC_C}$ selects $p = 3$. Bottom row: when $p = 2$ is selected. Distributions shown for LB with lags 5, 10, 15.

No power conditional on selecting wrong order!

- When we fit the right model (R, top row) the test statistic matches its null distribution
- When we fit the wrong model *deterministically* (W, 2nd row, 1st fig) the test statistic is stochastically larger, indicating power to reject
- But when we fit the wrong model *by selection* (W, 2nd row, 2nd fig) the test statistic approximately matches its null distribution. The test is (nearly) no better than an α-coin toss at telling us the chosen model is incorrect

Regression example: F-tests (of unselected variables)

- Regression models $E[Y] = X_A \beta_A$ for some subset A of columns of a matrix X.
- With nested subsets $A \subsetneq A'$, we'll conduct an F-test and consider this as a goodness-of-fit test for the model with variables A.
- In `R` we just use the `anova` function with these two linear models.

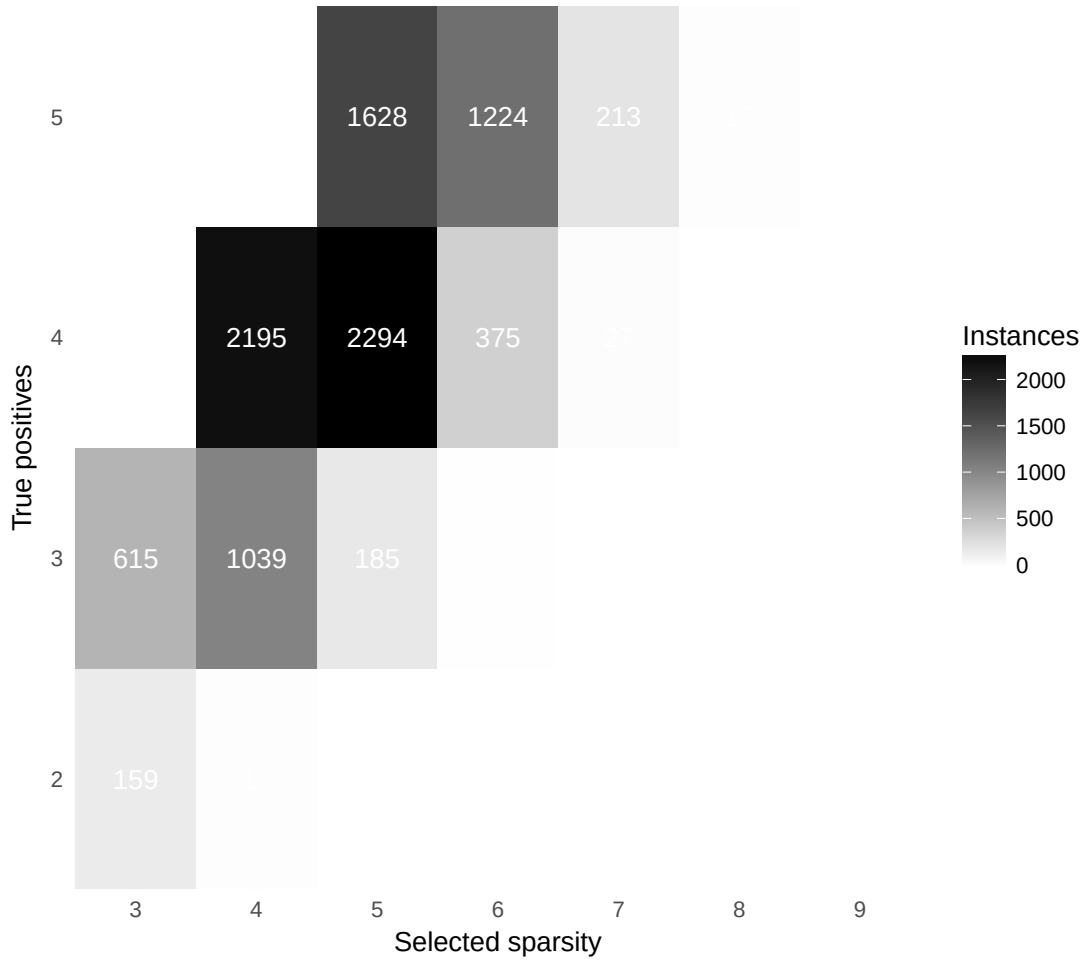
The distribution of the F-statistic is derived, of course, under the assumption that A and A' have been chosen *a priori*...

(An idea similar to this is mentioned in [**?**])

Regression variable selection

- For concreteness, consider selecting variables using forward stepwise with BIC, i.e. in `R` with `step(..., k = log(n))`.
- Simulation with $n = 100$ observations of $p = 10$ variables, the first two coefficients are larger than the next 3, and the last 5 are all 0.
- In this low-dimensional example, we'll take $A' = \{1, \ldots, 10\}$ for simplicity.
- We'll consider the F-test (1) with A chosen deterministically as the first 2 variables, and (2) with A selected by forward stepwise.

Profile of model selection events



Distributions of p-values for full-model F-tests

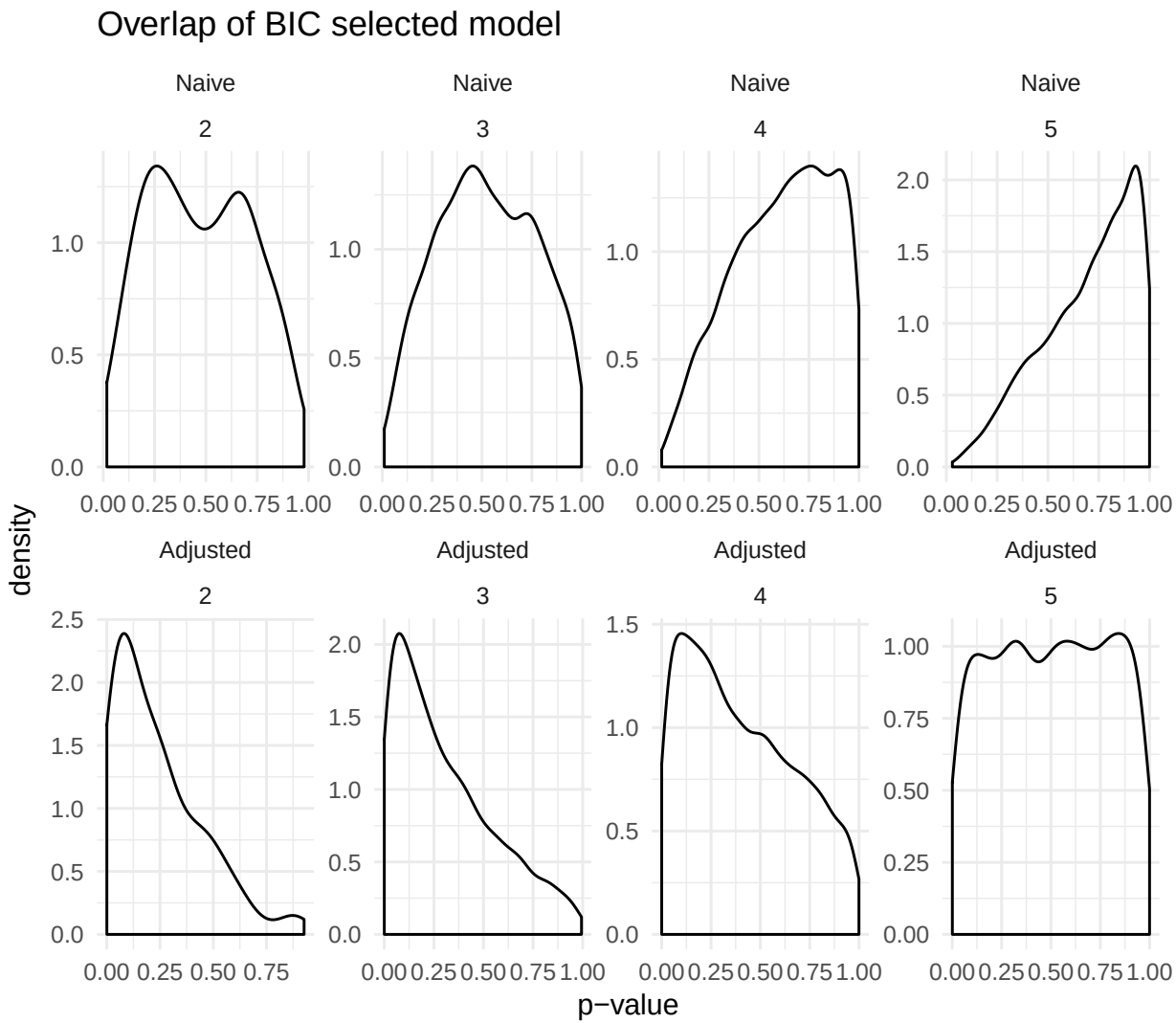


Figure: Top: unadjusted p-values. Bottom: adjusted for selection.

Probability of rejection

Table: Probability of rejection at level 0.1, conditional on size of overlap

pvalue	overlap	Pr(reject)
Naive	2	0.056
Naive	3	0.032
Naive	4	0.017
Naive	5	0.005
Adjusted	2	0.328
Adjusted	3	0.251
Adjusted	4	0.156
Adjusted	5	0.101

Power conditional on not selecting all true positives

Table: Probability of rejection at level 0.1, conditional on overlap less than 5

pvalue	Pr(reject)
Naive	0.022
Adjusted	0.186

Many means example (with new math?)

- Independent $Z_i \sim \mathcal{N}(\mu_i, 1)$, $i = 1, \ldots, p$.
- Apply hard-thresholding: discard any Z_i with $|Z_i| < C$.
- Goodness-of-fit test: χ^2 with discarded effects

Test statistic

Suppose $Z_1, \ldots, Z_m$ are discarded, i.e. they are less than C in absolute value. We will use

$$T = \sum_{i=1}^{m} Z_i^2$$

as a test statistic.

Not distributed as χ^2 – each term in the sum is truncated.

The $m = 2$ case

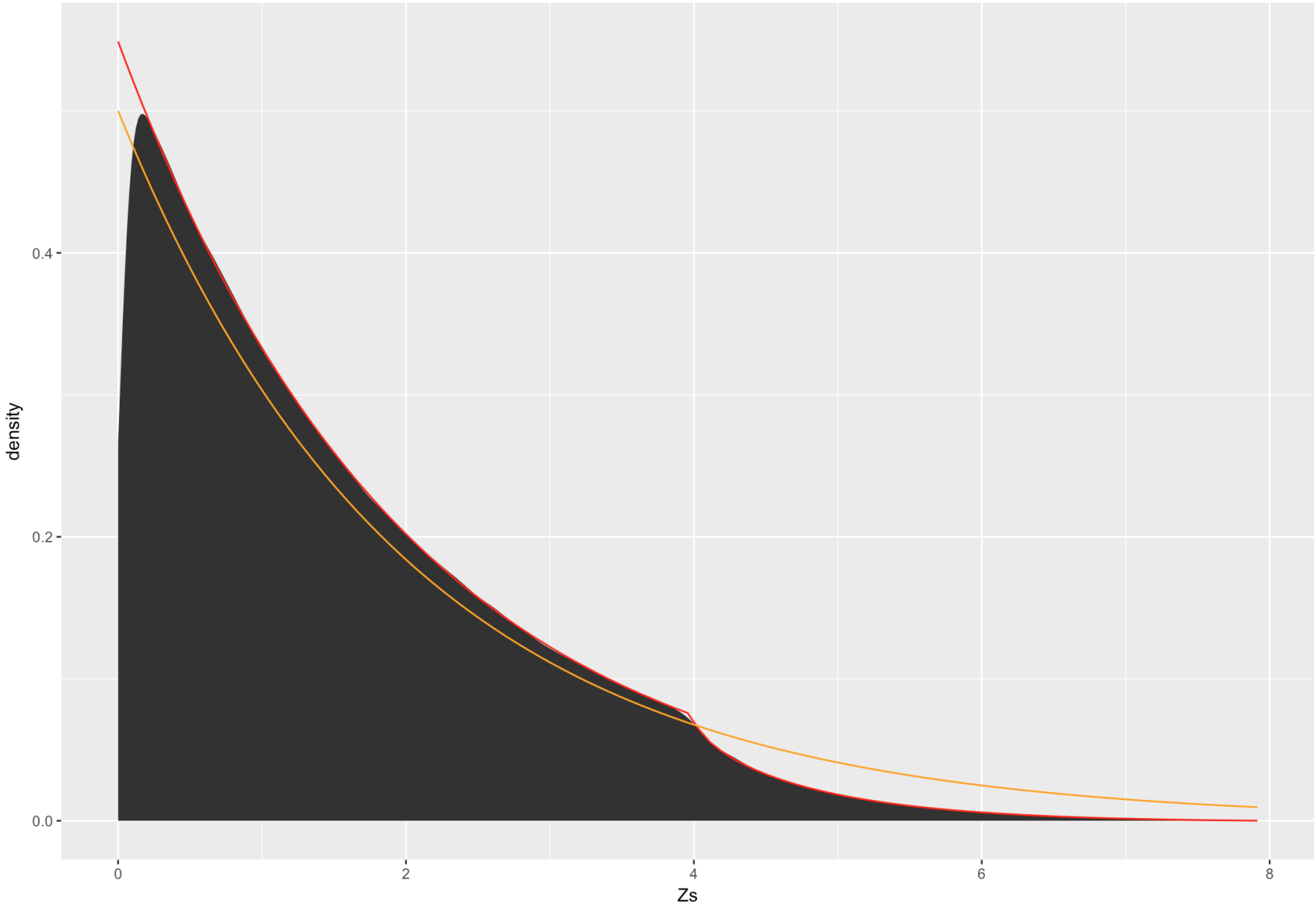
Distribution of a sum of truncated χ^2_1 random variables

Suppose $Z_i \sim \mathcal{N}(0,1)\,|_{[-C,C]}$, then the density function $f(z)$ of $Z_1^2 + Z_2^2$ is

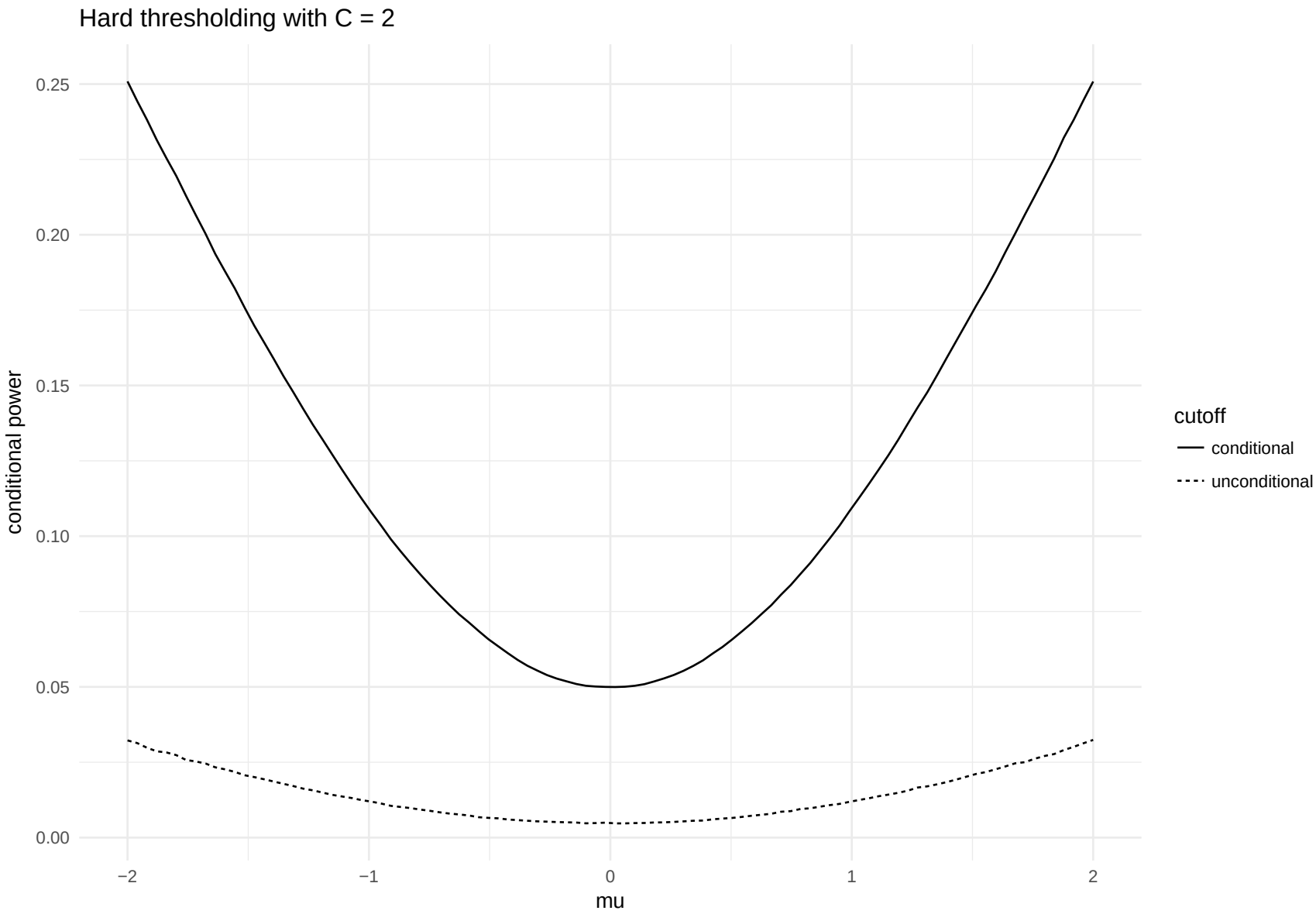
$$[1 - 2\Phi(-C)]^2 f(z) = \begin{cases} \frac{1}{2} e^{-y/2} & 0 \le z \le C^2 \\ \frac{1}{\pi} e^{-y/2} \left[2 \arcsin\left(\frac{c}{\sqrt{y}} \right) - \frac{\pi}{2} \right] & C^2 < z \le 2C^2 \\ 0 & \text{otherwise} \end{cases}$$

When $m > 2$ calculations get nastier (but for large m could use asymptotic results)

When $m = 2$



Conditional power in the $m = 1$ case



Quiet scandal

Goodness-of-fit tests: HUH!

What are they good for?

Absolutely nothing!?

(until we fix them)

Converging on standard practices?

- Large variety of goodness-of-fit tests in all kinds of different modeling settings.
- Non-tests: various measures and diagnostic plots – any kind of model assessment.
- Important message that these are *all* affected by model selection bias, not just variable significance tests or prediction/classification accuracy.
- Can we converge on a new set of standard practices?
- Proposal: we must make these tools as easy to use as `summary()` and diagnostic `plot()`

Reference