SharpNullTest

Maxime Rischard

March 23, 2017

Contents

Testing against a sharp null hypothesis seems like an obvious thing that people will want to do with spatial regression discontinuity designs. To be specific, the sharp null hypothesis would be that the treatment effect along the boundary is zero everywhere: $\tau(\partial) = 0$. I call it the sharp null to distinguish it from a hypothesis that the average treatment effect is zero.

So far I've been experimenting with various iterations of chi square tests, the simplest of which is to compute a χ^2 statistic $\mu^\intercal \Sigma^{-1} \mu$ on the posterior treatment effect at the sentinels, and compare it to a χ^2 distribution. But numerical instabilities make it difficult to work out how many degrees of freedom this distribution should have. I've tried various alternatives and algorithms, which all make some kind of sense, but contradict each other, and suffer from numerical issues.

In this notebook, I implement two tests that rely on parametric bootstraps to derive a null distribution rather than obtaining a theoretical null distribution. The first uses the χ^2 statistic above. To be explicit, here's the full procedure:

1. fit $\mathcal{GP}$s to the treatment and control regions
2. optimize their (shared) hyperparameters
3. obtain the "cliff face" posterior means and covariance μ and Σ at the sentinels following our $2\mathcal{GP}$ procedure
4. compute the observed χ^2 statistic $\chi^2_{\mathrm{obs}} = \mu^\intercal \Sigma^{-1} \mu$
5. create a null $\mathcal{GP}$, which is a single $\mathcal{GP}$ covering both regions, with the same hyperparameters
6. simulate from this null $\mathcal{GP}$, and for each simulation
    - fit the two $\mathcal{GP}$s on the simulated data
    - obtain and save the χ^2 statistic
7. compare the distribution of the simulated χ^2 statistics to χ^2_{obs}
8. a p-value is obtained as the proportion of simulated values above the observed value

In this notebook, I apply this method to simulated data, with a fairly strong constant treatment effect, and the test unambiguously detects the treatment effect. Obtaining the χ^2 statistic on 1000 random samples takes 4.5 seconds. None of the samples are above the observed value.

The second procedure I implement is the one suggested by LukeM, that uses the log-likelihood rather than the χ^2 statistic at the boundary. This could be described as a parametric bootstrap version of a likelihood ratio test. Here's the procedure:

1. fit $\mathcal{GP}$s to the treatment and control regions
2. optimize their (shared) hyperparameters
3. obtain the log-likelihood (sum of the two log-likelihoods)
4. create a null $\mathcal{GP}$, which is a single $\mathcal{GP}$ covering both regions, with the same hyperparameters
5. obtain its log-likelihood
6. compute $\Delta \log P_{\text{obs}}$, the difference between the $2\mathcal{GP}$ log-likelihood and the null model log-likelihood
7. simulate from the null $\mathcal{GP}$, and for each simulation
    - update the mean parameter of the null and alternative $\mathcal{GP}$s to their respective empirical values
    - fit the two $\mathcal{GP}$s on the simulated data
    - fit the null model to the simulated data
    - obtain the log-likelihoods for the null and alternative model
    - store their difference, $\Delta \log P_{\text{sim}}$
8. compare the distribution of the simulated $\Delta \log P_{\text{sim}}$ to $\Delta \log P_{\text{obs}}$
9. a p-value is obtained as the proportion of simulated values above the observed value

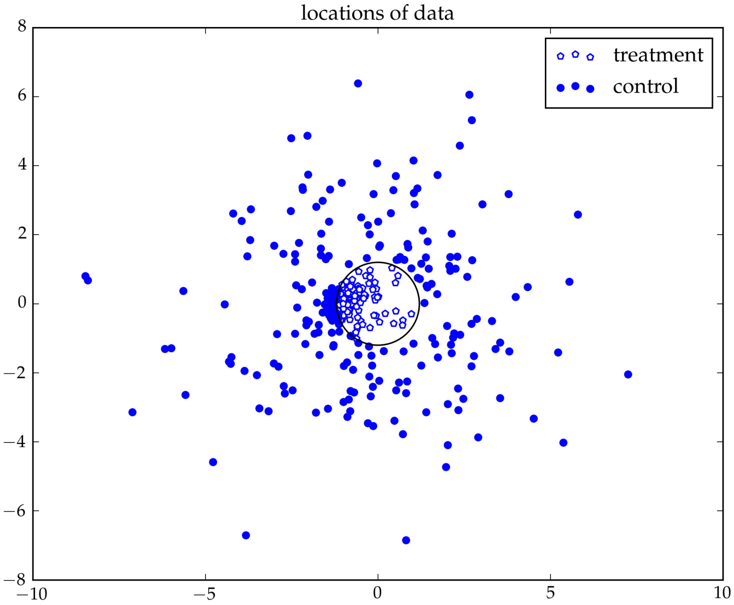
Within each simulation, I'm not optimizing the hyperparameters, other than setting the mean to its empirical value. This is mostly for computational reasons, as I get to reuse the Cholesky decomposition. This means on my laptop I can do 10,000 simulations in just over 6 seconds.

1 Simulate data and fit 2GP

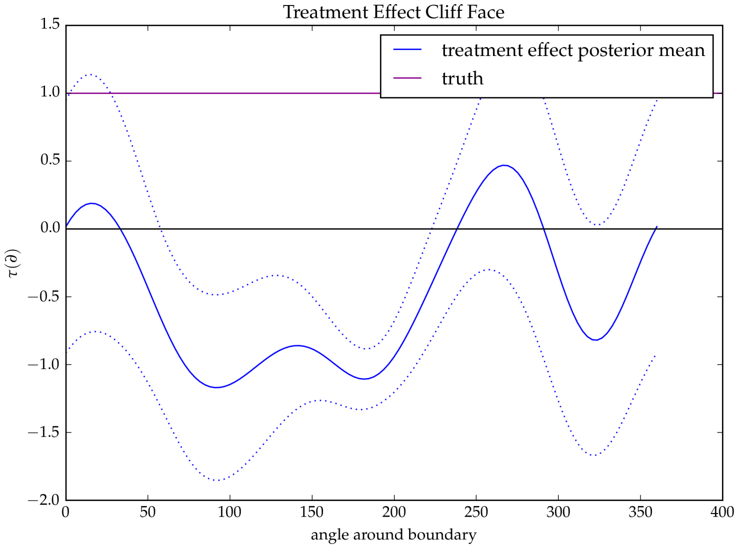
This section is just setting up the simulation. The treatment effect is just a constant shift, and I'm simulating from a GP with known hyperparameters.

```
GeoRDD
```

```
(1.0,0.002,0.011313294738755307)
```



```
(2,120)
```



2 Benavoli and Mangili 2015 χ^2 test

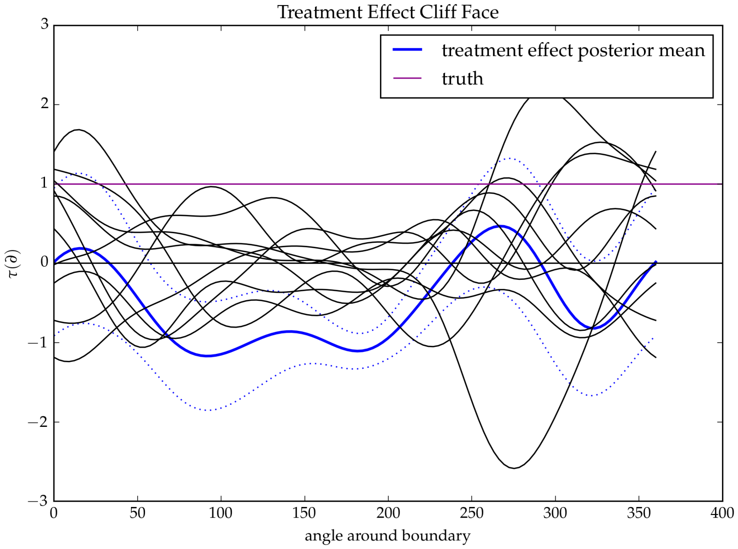
Following the 2GP procedure, we are now equipped with the posterior treatment effect along the boundary, approximated by its mean vector $\mu_{\partial|Y} = \mathbb{E}\left(\tau(\boldsymbol{\partial}) \mid Y\right)$ and covariance matrix $\Sigma_{\partial|Y} = \operatorname{cov}\left(\tau(\boldsymbol{\partial}) \mid Y\right)$ at the sentinel positions $\boldsymbol{\partial}$. Naturally, we wonder whether there is a significant treatment effect anywhere along the boundary, by setting up a hypothesis test where the null hypothesis is that $\tau(\partial) = 0$ everywhere along the boundary. Benavoli and Mangili 2015 develop a Chi-squared test for equality of two functions fitted by Gaussian processes, which can be adapted for our purposes. They compare the statistic $\mu_{\partial|Y}^{\intercal}\Sigma_{\partial|Y}^{-1}\mu_{\partial|Y}$ to a χ^2 distribution with ν degrees of freedom, where ν is the rank of the posterior covariance $\Sigma_{\partial|Y}$. However, because small eigenvalues can destabilize the χ^2 statistic, they suggest eliminating the covariance matrix's eigenvalues λ_i lower than $\epsilon\sum_{j=1}^{k}\lambda_j$, with ϵ a small number (they use 0.01). Benavoli and Mangili's simulations under the null hypothesis of function equality show the test to reject the null hypothesis less than 5% of the time. Our simulations drawing from a single Gaussian process with known covariance hyperparameters exhibit the same conservatism.

That conservatism arises because if Y_T and Y_C come from a single Gaussian process with no discontinuity at the boundary, then the sampling distribution of $\mu_{\tau|Y}$ is not, as the test implicitly assumes, $\mathcal{N}\left(0, \Sigma_{\partial|Y}\right)$. Because μ is a linear transformation of the data, it is indeed normally distributed with mean zero, but with covariance

$$\Sigma_\mu = K_{\partial T}\tilde{K}_{TT}^{-1}K_{T\partial} + K_{\partial C}\tilde{K}_{CC}^{-1}K_{\partial C}^{\intercal} + K_{\partial C}\tilde{K}_{CC}^{-1}K_{CT}\tilde{K}_{TT}^{-1}K_{T\partial} + \left(K_{\partial C}\tilde{K}_{CC}^{-1}K_{CT}\tilde{K}_{TT}^{-1}K_{T\partial}\right)^{\intercal} \tag{1}$$

A test with better frequentist properties can therefore be obtained by substitituing Σ_μ for $\Sigma_{\partial|Y}$ in Benavoli and Mangili's procedure.

3 Bootstrap cliff face

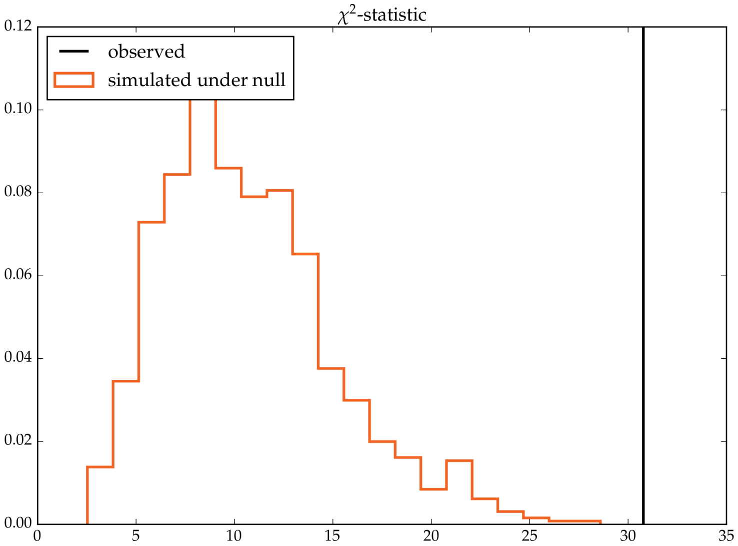


We can see pretty well from the bootstrapped cliff faces that there is a segment of the boundary (at angle 180 °) where there is a very clear signal. The standard deviation envelope is low, and all boostrapped samples move towards zero (the truth in those cases). In other regions of the boundary, we confirm that the signal is very weak.

4 Simulating χ^2 statistic under null

```
30.780223666060692
```

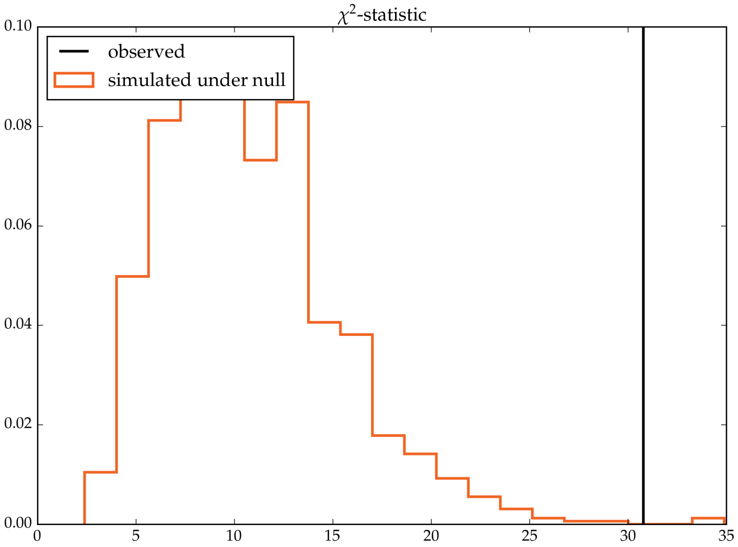
```
30.774546992181772
```



```
0.0
```

4.1 χ^2 with mean update

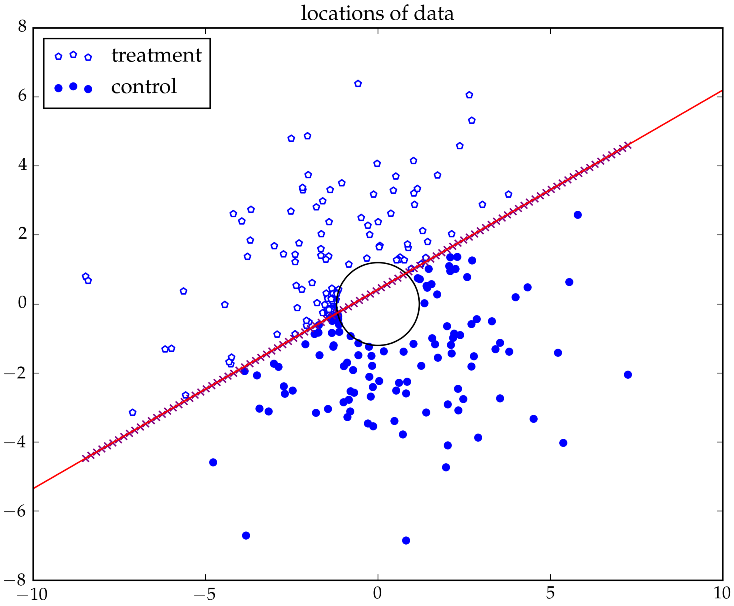
```
33.1667798196404
```

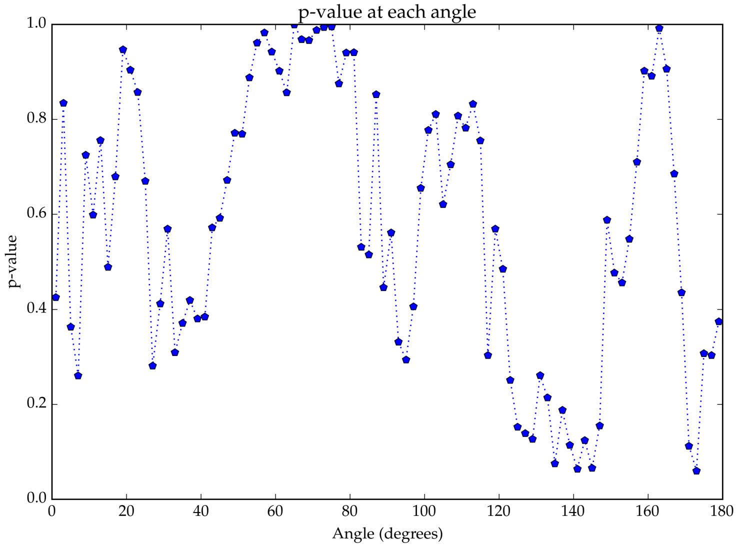


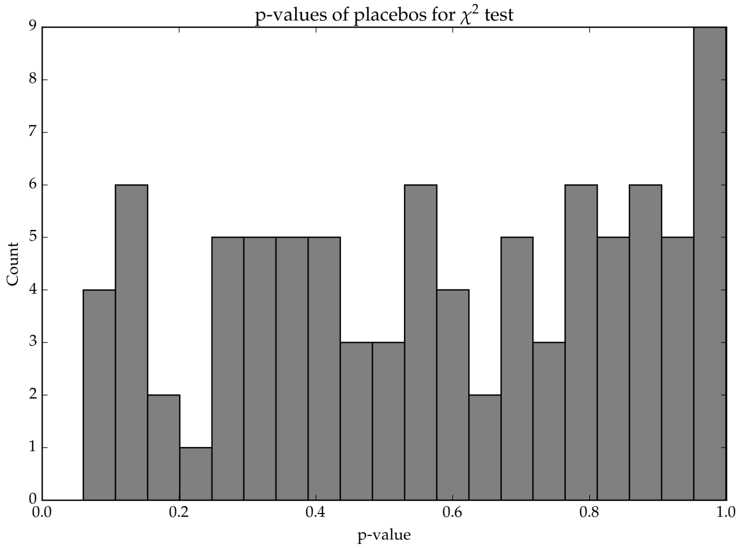
When we update the mean, the upper tail is lengthened, reducing the p-value a little bit (although actually here it's still 0 with 1000 sims), which seems right.

```
0.002
```

4.2 placebo test for χ^2 test





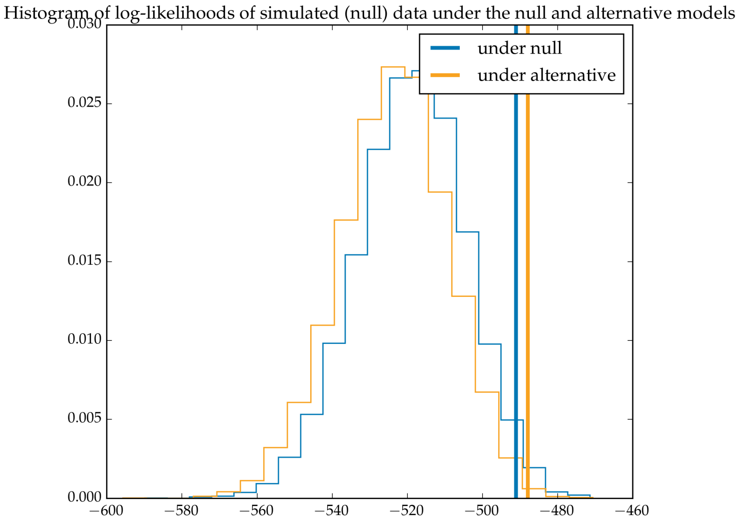


The placebo test doesn't indicate any issues with the test.

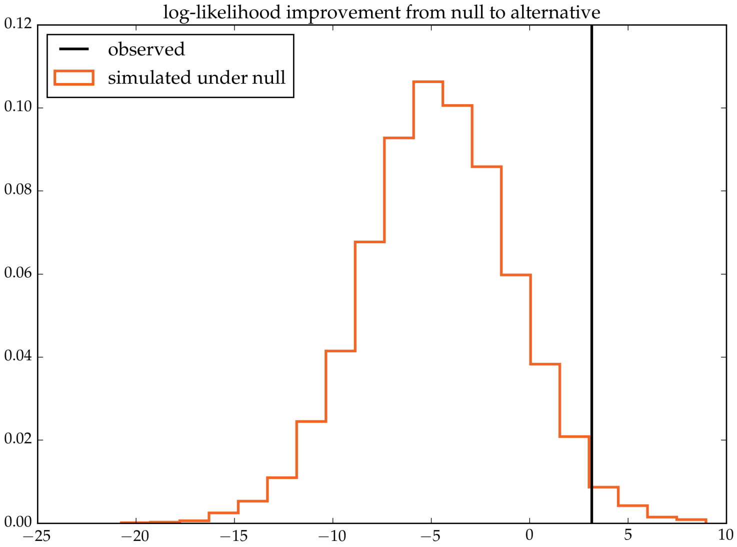
5 Comparing log-likelihoods

Now we come to the likelihood-based test. This one's interesting because it doesn't directly use the information available at the boundary. In fact, the sentinel positions are not used in this test at all, just the treatment-control indicators.

I now plot histograms of the simulated log-likelihoods. Because we're simulating from the null, it makes sense that the null-likelihood is on average higher (the blue histogram is slightly to the right of the orange histogram), and because the observed data includes a constant treatment effect, it also makes sense that the likelihood of the null model for the observed data is slightly *lower* than the null log-likelihood. That's the fundemental idea with this test: by comparing the differences in log-likelihoods between the null and alternative hypotheses, we should be able to construct a valid test.



To complete the test, we focus on the difference in log-likelihoods. Again I plot a histogram of the simulated and observed values.

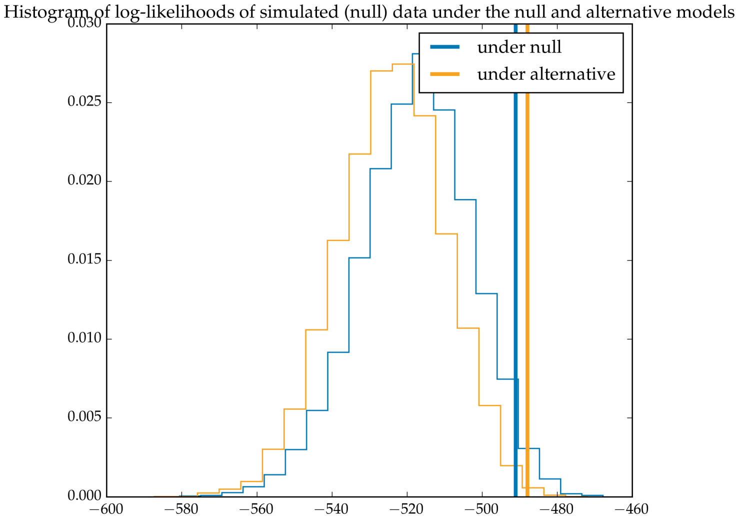


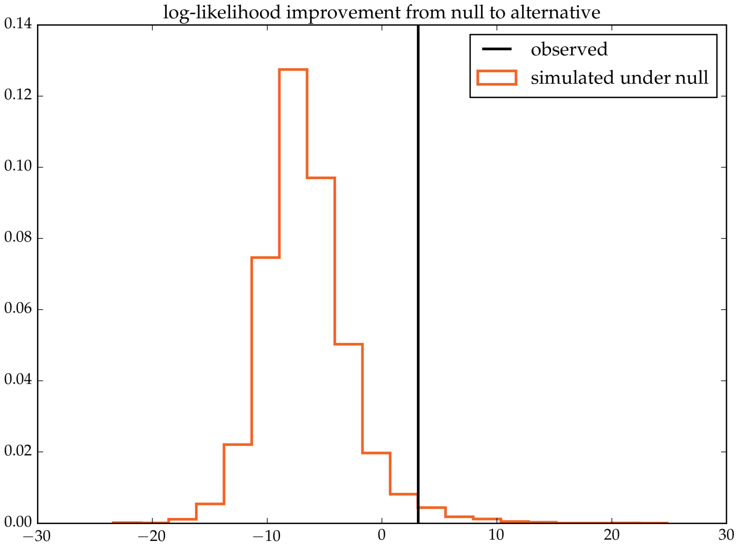
It's pretty obvious from the histogram that the p-value is zero. We can also obtain the p-value as the frequency of simulations that have a log-likelihood difference $\Delta \log P_{\mathrm{sim}}$ above the observed $\Delta \log P_{\mathrm{obs}}$.

```
0.0207
```

5.1 With mean update

This section is just a slight modification of the procedure, allowing for mean updates within each simulation. Everything else is the same.





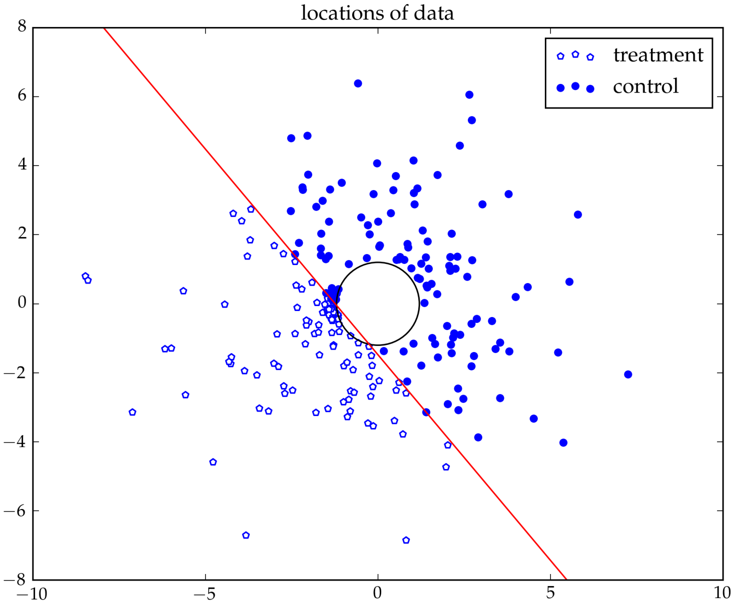
```
0.0201
```

This gives the null model a slight edge in the simulations (because it has two mean parameters), and so the p-value goes up from {{pval_nomean}} to {{p_val_meanup}}. Still a very clear rejection of the null hypothesis.

5.2 placebo test for likelihood test

Now, the validity of the test isn't immediately obvious, so I decided to implement a kind of placebo test. The idea is to take just the control data (or just the treatment data), and draw an arbitrary boundary through it, and then run the likelihood test to obtain a p-value. If we repeat this procedure over many different arbitrary boundaries, we'd hope to get some kind of uniformly distributed p-value. The way I implemented it is that I iterated over angles from 1° to 180°, drew a line through the data, and shifted it until there was

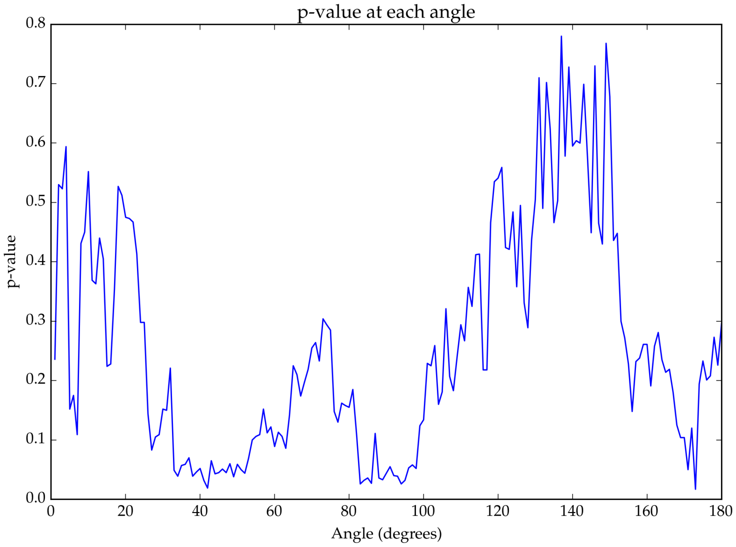
50% of the data on either side of the boundary. The picture below is an example of this procedure, with the black circle showing the real discontinuity, and the red line showing the placebo boundary, at an angle of 130° from horizontal.

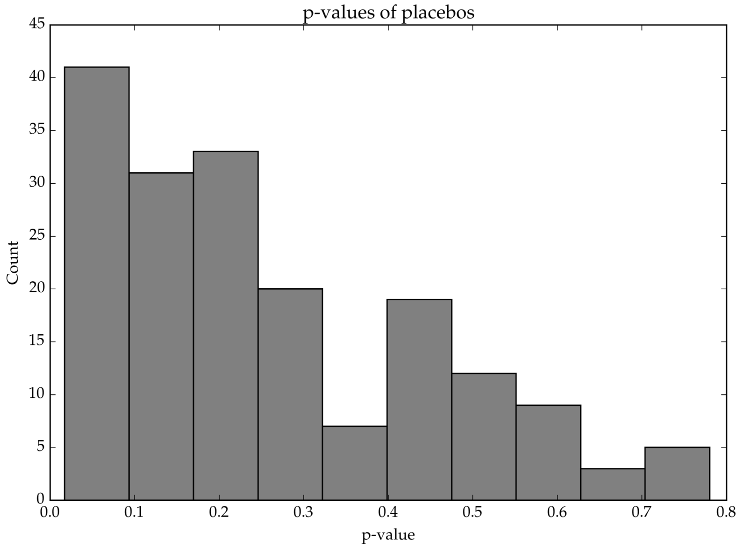


I implemented a function, `placebo`, that: * takes an angle, and the original control data * draws a linear boundary through the control data at that angle that splits the data in half * then applies our likelihood test to the data under this new boundary, * spits out the p-value.

Here's an example at 0°, with 1000 draws from the null performed for the test.

A quick look at the resulting p-values shows some issues. The p-values are strongly correlated across angles, so we definitely don't have 180 independent replications of the test. There's some sharp jumps that are particularly interesting. By eye it looks like we have maybe a dozen effective samples. The histogram of p-values is not quite uniform, but this could easily be attributed to the low number of effective samples.

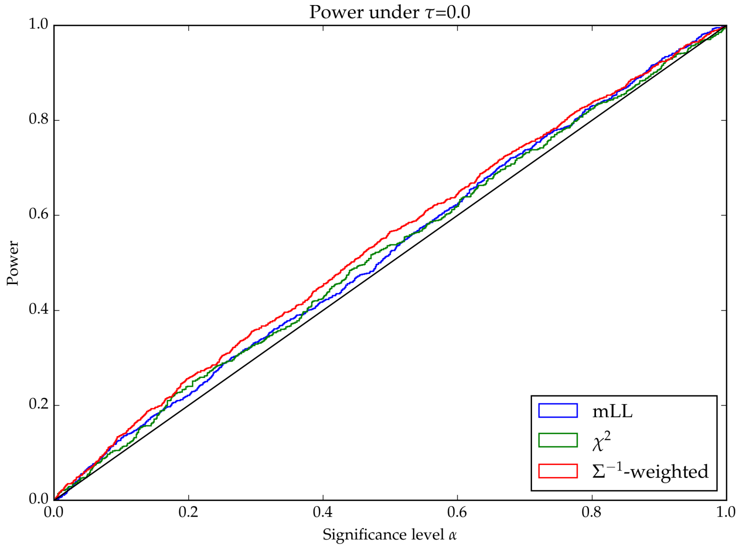


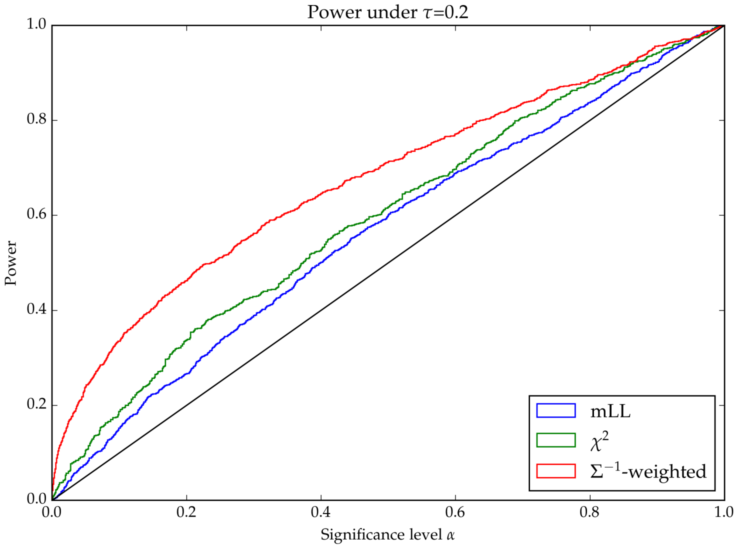


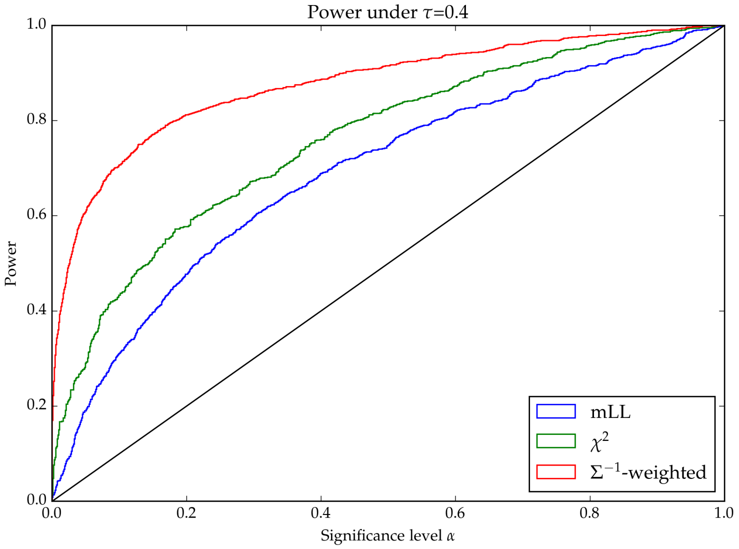
5.3 Checking the validity of the placebo test

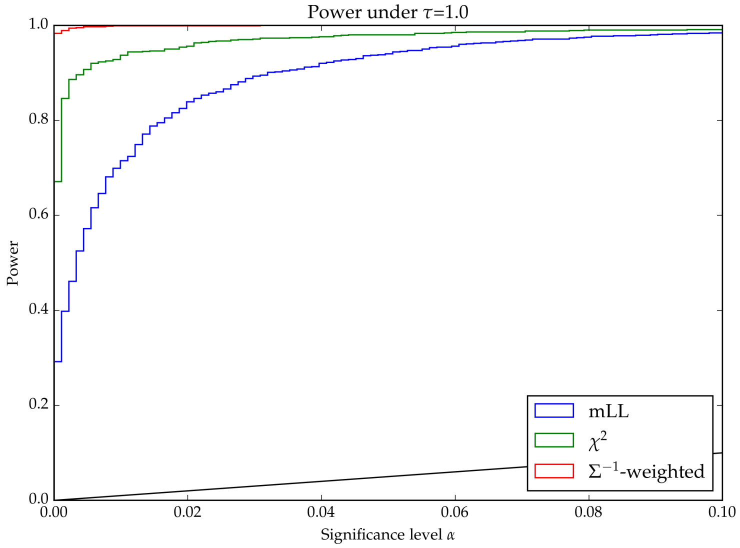
Let’s just observe the behavior of the placebo test over lots of replications from the null. We do quite commonly observe sharp jumps in the p-values as a function of angle, which we also observe in the NYC data. The histograms of p-values are overall fairly flat, though there was one with a lot of low p-values, and we do see occasional concentration of p-values at the low or high end of the scale, confirming the intuition that the effective sample size of this placebo test is fairly low.

6 Power of tests




Power under τ=0.2
Power
Significance level α
0.0
0.2
0.4
0.6
0.8
1.0
0.0
0.2
0.4
0.6
0.8
1.0
mLL
χ^2
Σ^{-1}-weighted


Power under τ=0.4
Power
Significance level α
0.0
0.2
0.4
0.6
0.8
1.0
0.0
0.2
0.4
0.6
0.8
1.0
mLL
χ^2
Σ^{-1}-weighted



This is very interesting. The sharp null hypothesis for the χ^2 is more constraining than the zero-mean null hypothesis of the inverse-variance mean test. By this I mean that the sharp null implies zero mean, but not vice versa. Intuitively, a more specific null hypothesis should be easier to reject. But here we're seeing the opposite, the inverse-variance mean test has much higher power to detect a constant treatment effect. All three tests, to a degree of approximation, are valid under the null. It's also interesting to note that the χ^2 test is in turn more powerful than the likelihood test.