Prep Meeting 16

# OMP With Bootstrap

Implemented function OMP\_Bootstrap, which receives a dataset X and the true data generating matrix W\_true. It will generate X’, which is generated using the same statistical model as X, but now has different random noise.

Alternatively, we can split X into two (or more parts), and get our X and X’ in that manner.

The matrix W\_OMP will be computed **with the data X**, but the scores are computed **with the data X’**. Most likely, noise in X that W\_OMP overfitted on, will be very different noise in data X’. Hence, removing the overfitted edges will actually lead to an **improvement** of the MSE. True edges will remain to be beneficial, whereas false edges will now be detrimental.

We prune the matrix W\_OMP by iteratively removing the least important edge. At each iteration (i), we compute the mean squared error of W\_OMP^(i) with respect to the data X’. Then, we pick the matrix W\_OMP^(i) that minimizes the MSE. This should hopefully resemble the true matrix W\_true.

Example:

Ten dimensional matrix with 25 edges, 100 samples. All coefficients are +/- 0.25 – 0.50:

[[ 0. 0. 0. 0. 0. 0. 0. 0. 0. 0. ].

[ 0. 0.44 0. 0. 0. 0. 0. 0. 0. 0. ]

[ 0. 0. 0. 0. 0. 0. 0. 0. 0. 0. ]

[-0.48 0. 0. 0. 0. 0. 0. 0. 0. 0. ]

[ 0. 0. 0. -0.3 0. 0. 0. 0. 0. 0. ]

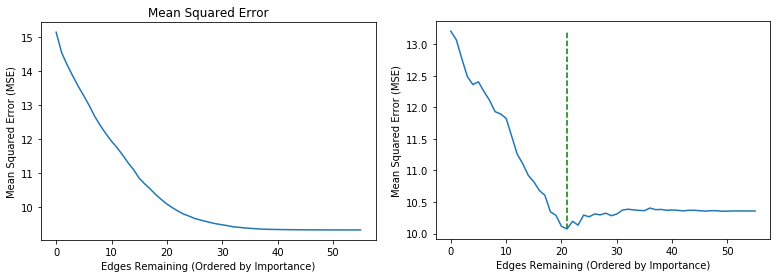
[ 0. -0.38 0.34 0.31 0. 0. 0. 0. 0. 0. ]

[-0.48 0. 0.29 0. -0.45 -0.37 0.41 0. 0. 0. ]

[-0.28 0. -0.31 0. 0.44 0.45 0. 0. 0. 0. ]

[ 0. 0.48 -0.36 0. -0.49 0.43 -0.36 0.34 0. 0. ]

[-0.44 0.43 0. 0. 0. -0.29 0. -0.32 0. 0. ]]



MSE of iteratively pruning W\_OMP on X. MSE of iteratively pruning W\_OMP on X’.

Highlighted in green, the W\_OMP^(i) that minimizes the MSE. Structural performance of this W\_OMP to the W\_true, plus MSE and R^2 on the data X:

Dense W\_OMP Results on X: Bootstrap W\_OMP Results on X:

True Positive Rate: 1.0. True Positive Rate: 1.0.

True Negative Rate: 0.6. True Negative Rate: 0.949.

Accuracy: 0.7. Accuracy: 0.96.

R-Squared: 0.385 R-Squared: 0.34

Mean Squared Error: 9.32 Mean Squared Error: 10.018

As expected, the bootstrap W\_OMP yields *higher structural performance*, as there is no overfitting on the noise anymore. However, as expected, *the mean squared error and the R-Squared are lower*, but this is for the exact same reason; we do not overfit on the noise anymore.

I also investigated doing more bootstrapping sampling, say 10 or even 100, and see if the score improved. However, doing more bootstrapping often does **not** result in better results, as the “problem” lies with the initial dataset **X**, and not with the *N* boostraps, no matter how large *N* is. How larger *N* is though, the more consistent the results are, but definitely not better.

Other things to check for with bootstrapping: Say we have our first X with T = 100. What is best, one X’ with T’ = 100, or with T’ = 200, or T’ = 50, or does it not matter? Furthermore, is it perhaps beneficial to have **two** X’, both with T’ = 100, or do multiple not really matter?

# NOTEARS Investigation

Fixed NOTEARS! NOTEARS never gave good results, so I presumed there was something wrong, as it did just as good as OLS, and did barely enforce acyclicity.

There indeed went something wrong, both in mine as Alex’ code at the same time. When we evaluate h(W), we do not count the diagonals of W as penalty. Therefore, we set this to zero when evalutating h(W). However, due to some scoping issues, we did it wrong.

Everytime we evaluated h(W), we set the diagonal of W to zero, but we did not change this back! This meant that at every evaluation, the matrix W is modified in strange ways.

After fixing this, the results were superb, perhaps better than OMP! Seems to have fixed it for VAR models!

# OMP for SEM

Also made OMP for SEM, and compared this to notears, and it seems to be quite a bit worse, which makes sense as the directionality is an extra difficulty now. If a -> b, then b <- a is also a possibility in SEM, but not necessarily in VAR.

# Paper investigation for proof

Seems to me that independence is used in Lemma 8. Very difficult to say, but it is mentioned in the lemma. It seems also there is a concentration inequality.

# Paper investigation for assumptions

I checked how the assumptions fared for a “regular” VAR model in our setting, what the values for mu and rho were.

Most likely, the assumptions do **not** hold. Nevertheless, do we still get the same results from the theorem? Let us check.

**What is mu?** We struggled with this a bit, but it is a quantity of how well we can predict a variable X\_i using the “correct” variables X\_F\_bar, whereas there should not be a way to predict X\_i using X\_F\_bar.

We regress the atoms that **are not** optimal on the atoms that **are** optimal. We compute the sum of absolute values of the parameters for each non-optimal atom. We then take the maximum, and that is mu.

It can be seen as a quantity that says that all using the optimal atoms to estimate the untrue coefficients remain below a certain quantity in terms of size.

Example:

This means that when we use OLS to estimate these values, the coefficients will not be large.

What value for epsilon and smallest beta would we need then?

Values were often quite high, higher than 1.0 which complicates things.

Furthermore, using the algorithm, did sort of work as expected.

Normalizing things was difficult.

# Theorem Validation Very Easy setting

W; T = 500 Samples.

[[0.5 0. 0. ]

[0. 0.5 0. ]

[0. 0. 0.5]].

Mu: Should be < 1: 0.14072.

Rho: Should be > 0: 1.0.

* Lower bound on epsilon: 4.4152 with eta = 0.1
* Lower bound on coefficients: 0.3423.

OMP Paper approach yields:

1. Gain: 22.57.

Betas: [0.58 0. 0. 0. 0. 0. 0. 0. 0. ].

1. Gain: 20.5.

Betas: [0.58 0. 0. 0. 0. 0. 0. 0. 0.53].

1. Gain: 20.04.

Betas: [0.58 0. 0. 0. 0.52 0. 0. 0. 0.53].

1. Gain: 3.57.

* Lower than Epsilon! We break, and indeed, betas = bar{F}!

So, for the simple case, Theorem 1 holds, although we did not have

independence!

# More Difficult Setting

W; T = 500 Samples.

[[ 0.5 -0.45 -0.33]

[ 0. 0.5 0. ]

[ 0. 0. 0. ]]

Mu: Should be < 1: 0.65605.

Rho: Should be > 0: 0.66387.

* Lower bound on epsilon: 14.4348 with eta = 0.1
* Lower bound on coefficients: 1.6859.

OMP Paper approach yields:

1. Gain: 25.38.

Betas: [ 0. 0. 0. -0.66 0. 0. 0. 0. 0. ].

1. Gain: 18.73.

Betas: [ 0.48 0. 0. -0.66 0. 0. 0. 0. 0. ].

1. Gain: 16.25.

Betas: [ 0.48 0. 0. -0.5 0.47 0. 0. 0. 0. ].

1. Gain: 14.0.

* Lower than Epsilon! We break, and indeed, betas C bar{F}!

So, for the more difficult case, Theorem 2 holds!

Small conclusion: The situations seem to be in accordance with the theory, but the assumptions of epsilon and coefficient sizes seem unsatisfactory. For the more difficult setting, the epsilon was so large that we missed a coefficient, so that seems a bit crude.

Furthermore, a smallest size of > 1.0 for coefficients does not seem very satisfactory.

There is also quite a big difference in importance of diagonal and off-diagonal elements. Our setting is I think more complicated than “just” regression with independent y.