Talk 4: Stepwise regression and friends

Dean Foster

Amazon

August 24, 2022

Preamble:Three ways to think about data

Dean Foster

Amazon

Mental models

Three ways of thinking about data:

- Probabilistic modelling
- Individual sequences
- Information theory

Information theory aside

- Key concept: Good models compress the data well.
- Key idea: Describing the model and describing the data can both be done using bits and bytes
- Describing the model:
  - Hypothesis test: takes 1 bit to describe the model (point alternative)
  - $\theta \in [-M, M]$ takes $\log_2(2M/\sqrt{n})$ bits
  - Non-parametric takes creativity to describe the model
- Describing the data:
  - Use $\log_2(P(Y_1, \ldots, Y_n|\theta))$ bits for discrete distributions
  - Use $\log_2(f(Y_1, \ldots, Y_n|\theta))$ bits for continuous densities
- Best method is shortest total for model plus data

The wins of each

- Information theory:
  - Beating LZ is hard!
  - Forces you to think about wild alternatives
- Individual sequences:
  - Think about algorithms
  - Allows you to ignore the question "Do you believe this model?"
- Probabilistic models:
  - Source of inspiration for codes and algorithms!
  - minimax lower bounds
  - Two sample t-test alone is enough to justify studing models
  - Interpretability, Explainablity, partial slopes, etc

Costs of each

- Information theory:
  - A trap for the unwary–it pretends to solve all problems
  - bit and bytes don't really matter, predictions do!
  - (story: Getting sucked down the Kolmogorov complexity well)

Costs of each

- Information theory:
  - A trap for the unwary–it pretends to solve all problems
  - bit and bytes don't really matter, predictions do!
  - (story: Getting sucked down the Kolmogorov complexity well)
- Individual sequence:
  - The space of algorithms is huge: most are impossible to analyze
  - Hard to tell what "beliefs" are implied by a algorithm
  - (story: What no interaction term?)

Costs of each

- Information theory:
  - A trap for the unwary–it pretends to solve all problems
  - bit and bytes don't really matter, predictions do!
  - (story: Getting sucked down the Kolmogorov complexity well)
- Individual sequence:
  - The space of algorithms is huge: most are impossible to analyze
  - Hard to tell what "beliefs" are implied by a algorithm
  - (story: What no interaction term?)
- Probabilistic modelling:
  - An optimal answer for a model will not be robust
  - Sometimes the world is ugly
    - No model captures it well.
    - Continuing adding bells and whistles takes time away from looking at data.
  - (story: Geographic modeling of demand)

Which is the best?

Which is the best?

Ignore everything and run a Neural Net?

Which is the best?

Ignore everything and run a Neural Net?

- Know at least a little of each one
- Translate the solution of your problem from one view to another
  - If it doesn't make sense–re-think your solution!
  - Ideally, it should make sense in all three views
- But, nothing beats simply looking at your data
  - Outliers are a problem in all three
  - Influential points cause problems everywhere
  - Looking at data cures believing something completely false!

Chalk talk: Blackwell approachability

August 24, 2022

Quick introduction to Blackwell approachability

- Original paper is unreadable
- My 1999 version is unreadable
- But the idea is simple

Talk 4: Stepwise regression and friends

Dean Foster

Amazon

Stepwise regression gets “no respect”

- Quite commonly used, but not often studied
- Most statisticians think of it as “evil” or at best useful only to “lazy” scientists

Stepwise regression gets “no respect”

- Quite commonly used, but not often studied
- Most statisticians think of it as “evil” or at best useful only to “lazy” scientists
- But I’m a fan
- This talk will review some of the theoretical results that are known about it
- I’ll give some examples of its value in applied problems

Problem statement: As a scientist

- Goal: predict Y
- Inputs: you have millions of X's that can be used to predict Y
- Most X's are garbage
- How do you find a small subset of X's that will predict Y well?

Problem statement: example

- 20 years ago Bob Stine and I ran a “little” regression (JASA 2004)
  - 70,000 features
  - 2 million rows
  - Y = credit card holder going bankrupt next month

Problem statement: example

- 20 years ago Bob Stine and I ran a “little” regression (JASA 2004)
  - 70,000 features
  - 2 million rows
  - Y = credit card holder going bankrupt next month
- At the time it caused jaws to drop

Problem statement: example

- 20 years ago Bob Stine and I ran a “little” regression (JASA 2004)
  - 70,000 features
  - 2 million rows
  - Y = credit card holder going bankrupt next month
- At the time it caused jaws to drop
- Tricks:
  - Linear model instead of logistic regression (Fast!)
  - Dummy variables for interactions (contain signal)
  - Interactions (non-linear structure)
  - Bennett’s bound to calculate p-values (avoiding over-fitting)
  - Stepwise regression!

Problem statement: As a mathematician

- Model:

$$Y_i \sim X_i^\top \beta + \sigma Z_i$$

- Penalized regression:

$$\widehat{\beta}_\Pi \equiv \arg\min_{\widehat{\beta}} \sum_{i=1}^{n} (Y_i - X_i^\top \widehat{\beta})^2 + \Pi \sigma^2 |\widehat{\beta}|_0$$

- $|\widehat{\beta}|_0$ is the number of non-zeros in β

Problem statement: As a mathematician

- Model:
$$Y_i \sim X_i^\top \beta + \sigma Z_i$$
- Penalized regression:
$$\widehat{\beta}_\Pi \equiv \arg\min_{\widehat{\beta}} \sum_{i=1}^{n} (Y_i - X_i^\top \widehat{\beta})^2 + \Pi \sigma^2 |\widehat{\beta}|_0$$
- $|\widehat{\beta}|_0$ is the number of non-zeros in β
- Non-convex problem
- Note: $L1$ is the convex relaxation of $L0$, which leads to Lasso.

Why we care

- Error larger by p/q if we don't do variable selection
- Huge improvement in accuracy is possible
- Precisely:

$$E(\mu_{Y|X} - \widehat{Y}_p)^2 = \frac{p}{q} \quad E(\mu_{Y|X} - \widehat{Y}_q)^2$$

  - $\widehat{Y}_p$ is best fit using all the variables
  - $\widehat{Y}_q$ is best fit using only the q correct variables
- But, can we find the right subset?

First algorithm

- Try all subsets to find best fitting subset
  - Oops: Slow, and it will say use all the variables

Second algorithm

- Try all subsets and penalize by Bonferroni
  - $|t| > \sqrt{2\log(p)}$
  - Yes, it is painfully slow. But does it at least find the right subset?

Risk Inflation

Theorem (F. and George 1994, Donoho and Johnstone 1994)

For any orthogonal X matrix, if $\Pi = 2\log(p)$, then the risk of $\widehat{\beta}_{\Pi}$ is within a $2\log(p)$ factor of the target.

Risk Inflation

Theorem (F. and George 1994, Donoho and Johnstone 1994)

For any orthogonal X matrix, if $\Pi = 2\log(p)$, then the risk of $\widehat{\beta}_\Pi$ is within a $2\log(p)$ factor of the target.

- The bound is tight.
- (The same bound works for Lasso.)

Risk Inflation

Theorem (F. and George 1994, ~~Donoho and Johnstone 1994~~)

For any ~~orthogonal~~ X matrix, if $\Pi = 2\log(p)$, then the risk of $\widehat{\beta}_\Pi$ is within a 4 $\log(p)$ factor of the target.

Risk Inflation

Theorem (F. and George 1994, ~~Donoho and Johnstone 1994~~)

For any ~~orthogonal~~ X matrix, if $\Pi = 2\log(p)$, then the risk of $\widehat{\beta}_\Pi$ is within a 4 $\log(p)$ factor of the target.

- This bound is also tight
- (Lasso is a disaster in this case.)

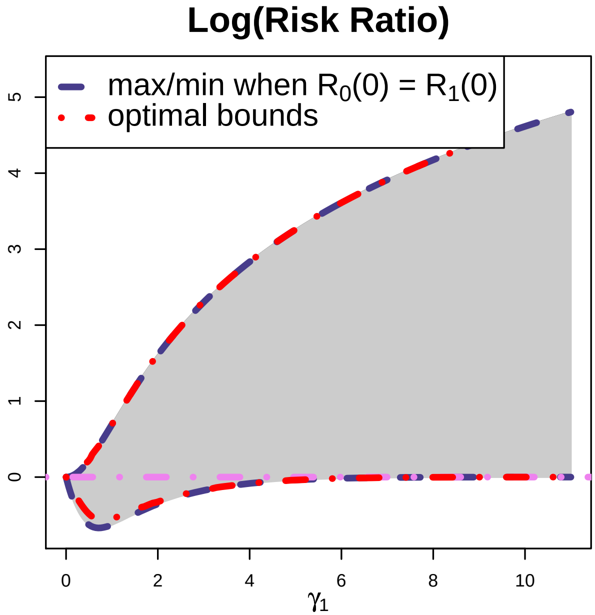
Risk Inflation

Theorem (F. and George 1994, ~~Donoho and Johnstone 1994~~)

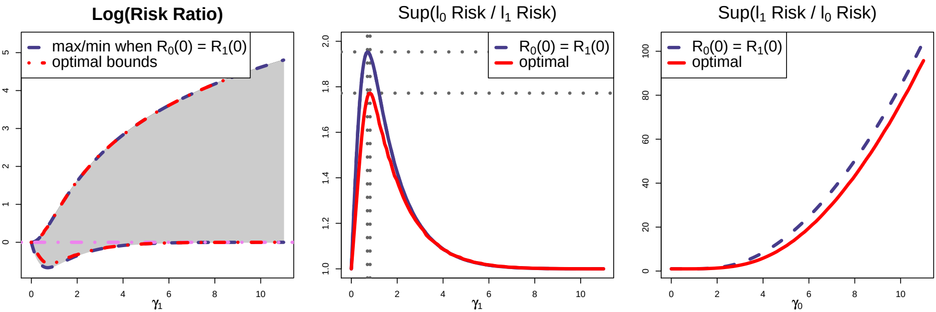
For any ~~orthogonal~~ X matrix, if $\Pi = 2\log(p)$, then the risk of $\widehat{\beta}_{\Pi}$ is within a 4 $\log(p)$ factor of the target.

- So finding the right subset of variables can generate a huge win

Why L0 instead of L1?



Why L0 instead of L1?



Greedy = Stepwise regression

- instead of exhaustive search, we can use search
- Greedy runs fast
- Called stepwise regression in statistics
- How well does it perform?

Greedy = Stepwise regression

- instead of exhaustive search, we can use search
- Greedy runs fast
- Called stepwise regression in statistics
- How well does it perform?
- For orthogonal problems, it works perfectly
- For many X's it will work well.
- But, . . .

Nasty example for stepwise

Y	D1	D2	D3	D4	...	Dn/2	X1	X2
1	1	0	0	0	...	0	$-1+\delta$	$+1+\delta$
1	1	0	0	0	...	0	$+1+\delta$	$-1+\delta$
1	0	1	0	0	...	0	$-1+\delta$	$+1+\delta$
1	0	1	0	0	...	0	$+1+\delta$	$-1+\delta$
1	0	0	1	0	...	0	$-1+\delta$	$+1+\delta$
1	0	0	1	0	...	0	$+1+\delta$	$-1+\delta$
1	0	0	0	1	...	0	$-1+\delta$	$+1+\delta$
1	0	0	0	1	...	0	$+1+\delta$	$-1+\delta$
$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$
1	0	0	0	0	...	1	$-1+\delta$	$+1+\delta$
1	0	0	0	0	...	1	$+1+\delta$	$-1+\delta$

Nasty example for stepwise

Y	D1	D2	D3	D4	...	Dn/2	X1	X2
1	1	0	0	0	...	0	$-1+\delta$	$+1+\delta$
1	1	0	0	0	...	0	$+1+\delta$	$-1+\delta$
1	0	1	0	0	...	0	$-1+\delta$	$+1+\delta$
1	0	1	0	0	...	0	$+1+\delta$	$-1+\delta$
1	0	0	1	0	...	0	$-1+\delta$	$+1+\delta$
1	0	0	1	0	...	0	$+1+\delta$	$-1+\delta$
1	0	0	0	1	...	0	$-1+\delta$	$+1+\delta$
1	0	0	0	1	...	0	$+1+\delta$	$-1+\delta$
$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$
1	0	0	0	0	...	1	$-1+\delta$	$+1+\delta$
1	0	0	0	0	...	1	$+1+\delta$	$-1+\delta$

- Stepwise regression finds:

$$Y = D_1 + D_2 + \cdots + D_{n/2}$$

Nasty example for stepwise

Y	D1	D2	D3	D4	...	Dn/2	X1	X2
1	1	0	0	0	...	0	$-1+\delta$	$+1+\delta$
1	1	0	0	0	...	0	$+1+\delta$	$-1+\delta$
1	0	1	0	0	...	0	$-1+\delta$	$+1+\delta$
1	0	1	0	0	...	0	$+1+\delta$	$-1+\delta$
1	0	0	1	0	...	0	$-1+\delta$	$+1+\delta$
1	0	0	1	0	...	0	$+1+\delta$	$-1+\delta$
1	0	0	0	1	...	0	$-1+\delta$	$+1+\delta$
1	0	0	0	1	...	0	$+1+\delta$	$-1+\delta$
$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$
1	0	0	0	0	...	1	$-1+\delta$	$+1+\delta$
1	0	0	0	0	...	1	$+1+\delta$	$-1+\delta$

- Actually:

$$Y = (X1 + X2)/\delta$$

Nasty example for stepwise

Y	D1	D2	D3	D4	...	Dn/2	X1	X2
1	1	0	0	0	...	0	$-1+\delta$	$+1+\delta$
1	1	0	0	0	...	0	$+1+\delta$	$-1+\delta$
1	0	1	0	0	...	0	$-1+\delta$	$+1+\delta$
1	0	1	0	0	...	0	$+1+\delta$	$-1+\delta$
1	0	0	1	0	...	0	$-1+\delta$	$+1+\delta$
1	0	0	1	0	...	0	$+1+\delta$	$-1+\delta$
1	0	0	0	1	...	0	$-1+\delta$	$+1+\delta$
1	0	0	0	1	...	0	$+1+\delta$	$-1+\delta$
$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$
1	0	0	0	0	...	1	$-1+\delta$	$+1+\delta$
1	0	0	0	0	...	1	$+1+\delta$	$-1+\delta$

- Stepwise regression finds the wrong model
- The model it finds is n/4 times bigger than it needs

Nasty example for stepwise

Y	D1	D2	D3	D4	...	Dn/2	X1	X2
1	1	0	0	0	...	0	$-1+\delta$	$+1+\delta$
1	1	0	0	0	...	0	$+1+\delta$	$-1+\delta$
1	0	1	0	0	...	0	$-1+\delta$	$+1+\delta$
1	0	1	0	0	...	0	$+1+\delta$	$-1+\delta$
1	0	0	1	0	...	0	$-1+\delta$	$+1+\delta$
1	0	0	1	0	...	0	$+1+\delta$	$-1+\delta$
1	0	0	0	1	...	0	$-1+\delta$	$+1+\delta$
1	0	0	0	1	...	0	$+1+\delta$	$-1+\delta$
$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$
1	0	0	0	0	...	1	$-1+\delta$	$+1+\delta$
1	0	0	0	0	...	1	$+1+\delta$	$-1+\delta$

- Lasso will also find the wrong model

How general is this example?

One example on one algorithm isn't real mathematics!

A success for stepwise regression

Theorem (Natarajan 1995)

Stepwise regression will have a prediction accuracy of at most twice optimal using at most $\approx 18|X^+|_2^2 q$ *variables.*

- This result was only recently noticed to be about stepwise regression. He didn't use that term.
- The risk inflation is a disaster.
- The $|X^+|_2$ is a measure of co-linearity.
- This bound can be arbitrarily large.
- Brings up two points: we are willing to "cheat" on both accuracy and number of variables. But hopefully not by very much.

L0 regression is hard

Theorem (Zhang, Wainwright, Jordan 2014)

There exists an design matrix X such that no polynomial time algorithm which outputs q variables achieves a risk better than

$$R(\widehat{\theta}) \gtrsim \frac{1}{\gamma^2(X)} \sigma^2 q \log(p).$$

Where γ is the RE, a measure of co-linearity.

- Actual statement is much more complex and involves a version of the assumption that $P \neq NP$.
- It was previously known that that

$$R(\widehat{\theta}_{lasso}) \lesssim \frac{1}{\gamma^2(X)} \sigma^2 q \log(p).$$

L0 regression is VERY hard

Theorem (Foster, Karloff, Thaler 2014)

No algorithm exists which achieves all three of the following goals:

- *Runs efficiently (i.e. in polynomial time)*
- *Runs accurately (i.e. risk inflation < p)*
- *Returns sparse answer (i.e. $|\widehat{\beta}|_0 \ll p$)*

- Strongest version requires an assumption about complexity (which I can't understand).
- The proof relies on "interactive proof theory." (which I also can't understand).
- The sparsity results depend on the assumptions used. We can get $|\widehat{\beta}|_0 < cq$ easily, and $|\widehat{\beta}|_0 < p^{.99}$ with difficulty.
- Difficult to improve to $|\widehat{\beta}|_0 \leq p$ since then all the heavy lifting is being done by the accuracy claims.

Theoretical analysis of several algorithms

- Several algorithms have been proposed to solve these
- In some cases they run well, in some cases they are a disaster
- Fun mathematics–but not really informative as to what to do in practice

What to do?

- Nothing will ever work perfectly
- So we have to hope the world is nice to us
- Let’s trust in this hope.

New algorithm: Alpha investing

Algorithm summary:

- Sort the variables putting the ones you like best first
  - For example, linear terms before interactions
  - put variables used last year before new ones to try
- Try each variable one at a time
- Add it to the regression if it is significant
  - Simplest rule: keep any with $|t| > \sqrt{2\log(p)}$
  - Fancy rule: Use alpha spending. But, give yourself an α bonus ever time you reject.

Alpha investing algorithm

```
Wealth = .05;
while (Wealth > 0) do
    bid = amount to bid;
    Wealth = Wealth - bid;
    let X be the next variable to try;
    if (p-value of X is less than bid) then
        Wealth = Wealth + .05;
        Add X to the model
    end
end
```

New algorithm: Alpha investing

- This is even more Greedy than stepwise regression
- provides mFDR protection
- Instead of orthogonalizing each new X, only approximately orthogonalize it.
  - Can be done via sampling
  - Can be done use fast matrix methods
- For sub-modular problems, it works well

mFDR for streaming feature selection

Let $W(j)$ be the "alpha wealth" at time j. Then for a series of p-values p_j, we can define:

$$W(j) - W(j-1) = \begin{cases} \omega & \text{if } p_j \leq \alpha_j \ , \\ -\alpha_j/(1-\alpha_j) & \text{if } p_j > \alpha_j \ . \end{cases} \qquad (1)$$

Theorem

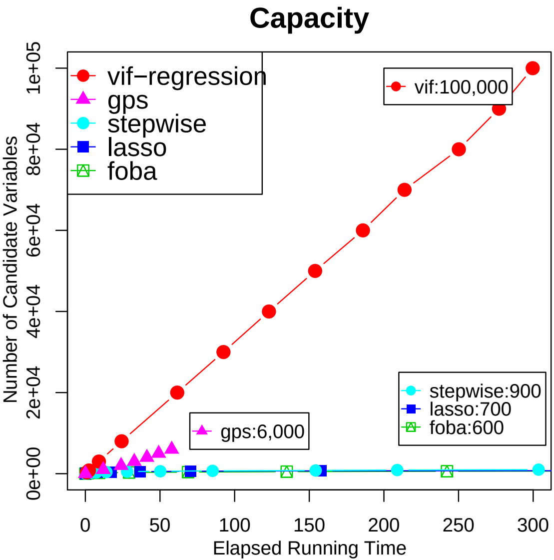
(Foster and Stine, 2008, JRSS-B) An alpha-investing rule governed by (1) with initial alpha-wealth $W(0) \leq \alpha\,\eta$ and pay-out $\omega \leq \alpha$ controls mFDR$_\eta$ at level α.

VIF regression

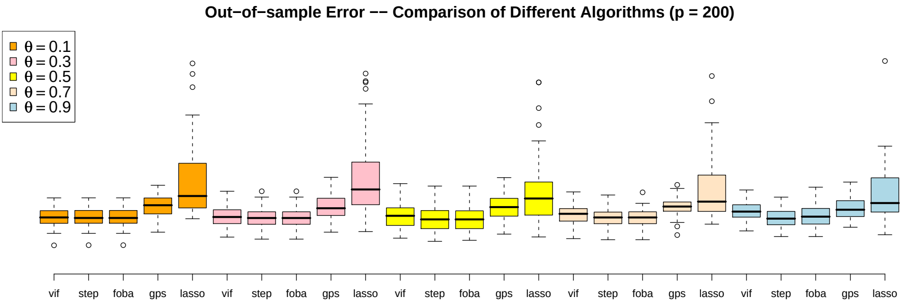
Theorem

(Foster, Dongyu Lin, 2011) VIF regression approximates a streaming feature selection method with speed $O(np)$.

VIF speed comparison



VIF out-of-sample

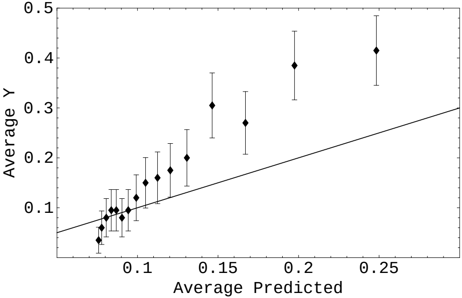


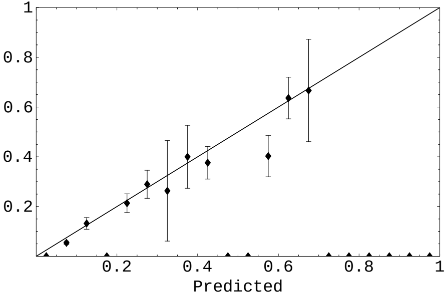
Submodular

Theorem

(Foster, Johnson, Stine, 2013) If the R-squared in a regression is submodular (aka subadditive) then a streaming feature selection algorithm will find an estimator whose out risk is within a factor of $e/(e-1)$ *of the optimal risk.*

About that calibration plot



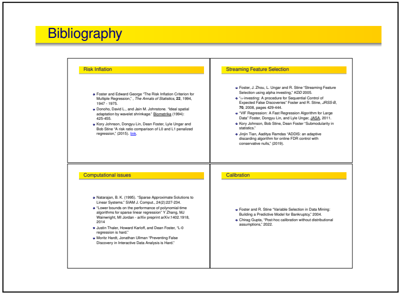


- We used PAV and crossed our fingers.
- Chirag Gupta has shown how to do this correctly.

Conclusions

- Stepwise regression when used correctly has good performance
  - include variables with $|t| > \sqrt{2\log(p)}$
  - Use interactions
  - Use dummy's for missing values
  - Use robust p-values
- Other fast alternatives
  - alpha investing (this talk)
  - Fast matrix methods (this afternoons talk)
  - gradient methods (Yichao Lu or try VW)

Thanks!



Bibliography

Risk Inflation

- Foster and Edward George "The Risk Inflation Criterion for Multiple Regression," , *The Annals of Statistics*, **22**, 1994, 1947 - 1975.
- Donoho, David L., and Jain M. Johnstone. "Ideal spatial adaptation by wavelet shrinkage." Biometrika (1994): 425-455.
- Kory Johnson, Dongyu Lin, Dean Foster, Lyle Ungar and Bob Stine "A risk ratio comparison of L0 and L1 penalized regression," (2015). link.

Streaming Feature Selection

- Foster, J. Zhou, L. Ungar and R. Stine "Streaming Feature Selection using alpha investing," *KDD* 2005.
- "α-investing: A procedure for Sequential Control of Expected False Discoveries" Foster and R. Stine, *JRSS-B*, **70**, 2008, pages 429-444.
- "VIF Regression: A Fast Regression Algorithm for Large Data" Foster, Dongyu Lin, and Lyle Ungar, JASA, 2011.
- Kory Johnson, Bob Stine, Dean Foster "Submodularity in statistics."
- Jinjin Tian, Aaditya Ramdas "ADDIS: an adaptive discarding algorithm for online FDR control with conservative nulls," (2019).

Computational issues

- Natarajan, B. K. (1995). "Sparse Approximate Solutions to Linear Systems." SIAM J. Comput., 24(2):227-234.
- "Lower bounds on the performance of polynomial-time algorithms for sparse linear regression" Y Zhang, MJ Wainwright, MI Jordan - arXiv preprint arXiv:1402.1918, 2014
- Justin Thaler, Howard Karloff, and Dean Foster, "L-0 regression is hard."
- Moritz Hardt, Jonathan Ullman "Preventing False Discovery in Interactive Data Analysis is Hard."

Calibration

- Foster and R. Stine "Variable Selection in Data Mining: Building a Predictive Model for Bankruptcy," 2004.
- Chirag Gupta, "Post-hoc calibration without distributional assumptions," 2022.

Bibliography

Risk Inflation

- Foster and Edward George "The Risk Inflation Criterion for Multiple Regression," , *The Annals of Statistics*, **22**, 1994, 1947 - 1975.
- Donoho, David L., and Jain M. Johnstone. "Ideal spatial adaptation by wavelet shrinkage." Biometrika (1994): 425-455.
- Kory Johnson, Dongyu Lin, Dean Foster, Lyle Ungar and Bob Stine "A risk ratio comparison of L0 and L1 penalized regression," (2015). link.

Streaming Feature Selection

- Foster, J. Zhou, L. Ungar and R. Stine "Streaming Feature Selection using alpha investing," *KDD* 2005.
- "α-investing: A procedure for Sequential Control of Expected False Discoveries" Foster and R. Stine, *JRSS-B*, **70**, 2008, pages 429-444.
- "VIF Regression: A Fast Regression Algorithm for Large Data" Foster, Dongyu Lin, and Lyle Ungar, JASA, 2011.
- Kory Johnson, Bob Stine, Dean Foster "Submodularity in statistics."
- Jinjin Tian, Aaditya Ramdas "ADDIS: an adaptive discarding algorithm for online FDR control with conservative nulls," (2019).

Computational issues

- Natarajan, B. K. (1995). "Sparse Approximate Solutions to Linear Systems." SIAM J. Comput., 24(2):227-234.
- "Lower bounds on the performance of polynomial-time algorithms for sparse linear regression" Y Zhang, MJ Wainwright, MI Jordan - arXiv preprint arXiv:1402.1918, 2014
- Justin Thaler, Howard Karloff, and Dean Foster, "L-0 regression is hard."
- Moritz Hardt, Jonathan Ullman "Preventing False Discovery in Interactive Data Analysis is Hard."

Calibration

- Foster and R. Stine "Variable Selection in Data Mining: Building a Predictive Model for Bankruptcy," 2004.
- Chirag Gupta, "Post-hoc calibration without distributional assumptions," 2022.

- Foster and Edward George "The Risk Inflation Criterion for Multiple Regression," , *The Annals of Statistics*, **22**, 1994, 1947 - 1975.
- Donoho, David L., and Jain M. Johnstone. "Ideal spatial adaptation by wavelet shrinkage." Biometrika (1994): 425-455.
- Kory Johnson, Dongyu Lin, Dean Foster, Lyle Ungar and Bob Stine "A risk ratio comparison of L0 and L1 penalized regression," (2015). link.

Streaming Feature Selection

- Foster, J. Zhou, L. Ungar and R. Stine "Streaming Feature Selection using alpha investing," *KDD* 2005.
- "α-investing: A procedure for Sequential Control of Expected False Discoveries" Foster and R. Stine, *JRSS-B*, **70**, 2008, pages 429-444.
- "VIF Regression: A Fast Regression Algorithm for Large Data" Foster, Dongyu Lin, and Lyle Ungar, JASA, 2011.
- Kory Johnson, Bob Stine, Dean Foster "Submodularity in statistics."
- Jinjin Tian, Aaditya Ramdas "ADDIS: an adaptive discarding algorithm for online FDR control with conservative nulls," (2019).

Computational issues

- Natarajan, B. K. (1995). “Sparse Approximate Solutions to Linear Systems.” SIAM J. Comput., 24(2):227-234.
- “Lower bounds on the performance of polynomial-time algorithms for sparse linear regression” Y Zhang, MJ Wainwright, MI Jordan - arXiv preprint arXiv:1402.1918, 2014
- Justin Thaler, Howard Karloff, and Dean Foster, “L-0 regression is hard.”
- Moritz Hardt, Jonathan Ullman “Preventing False Discovery in Interactive Data Analysis is Hard.”

Calibration

- Foster and R. Stine "Variable Selection in Data Mining: Building a Predictive Model for Bankruptcy," 2004.
- Chirag Gupta, "Post-hoc calibration without distributional assumptions," 2022.