Discussion

Dean Foster

Amazon @ NYC

Differential privacy means in statistics language:

Fit the world not the data.

Differential privacy means in statistics language:

Fit the world not the data.

- You shouldn't be able to tell which data set the experiment came from.
- (I expect Gelman will say how impossible this is later.)

Differential privacy means in statistics language:

Fit the world not the data.

- You shouldn’t be able to tell which data set the experiment came from.
- (I expect Gelman will say how impossible this is later.)
- More extreme, you should not be able to tell anything about the dataset even when given all but one person.

For most of the history of statistics this wouldn't matter.

- Regression for example:
  - $EY_i = x_i^\top \beta$ with $\beta \in \Re^p$
  - $p \ll n$
- Once we have $\hat\beta$ we can estimate any thing (The estimate of: $E(g(Y))$ is simply $E(g(x^\top \hat\beta + \sigma Z))$
- For linear combination, we even have confidence intervals (Scheffe)
- There wasn't all that much more in the data then in the model
- In fact, $\hat\beta$ was "sufficient" to answer any question we could dream of asking

Stepwise regression changed all that

- Model:

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

- Penalized regression:

$$\hat{\beta} \equiv \arg\min_{\hat{\beta}} \sum_{i=1}^{n} (Y_i - X_i^\top \hat{\beta})^2 + 2q_{\hat{\beta}}\sigma^2 \log(p)$$

- $\beta \in \Re^p$
- $q_{\hat{\beta}}$ is the number of non-zeros in $\hat{\beta}$
- let q, the number of non-zeros in β
- Need $q \ll n$, but p could be large

Sample of theory

- Competitive ratios:

Risk Inflation

- Prediction risk:

$$R(\hat{\beta}, \beta) = E_{\beta}|\mathbf{X}\beta - \mathbf{X}\hat{\beta}|_2^2$$

- Target risk:

$$R(\hat{\beta}) = q\sigma^2$$

- The L-0 penalized regression is within a log factor of this target.

Theorem (Foster and George, 1994)

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

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.

- Complexity:

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.

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

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

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

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. $|\hat{\beta}|_0 \ll p$)*

bibliography: 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."

Stewise regression and beyond

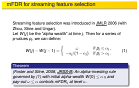
- The gready search for a best model is called stepwise regression

Stewise regression and beyond

- The gready search for a best model is called stepwise regression
- Bob Stine and I came up alpha investing:
  - It is an opportunistic search which doesn't worry about finding the best at each step
  - Try a variables sequentially and keep if if you like it

Properties of alpha investing

- "provides" mFDR protection (2008)



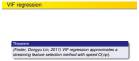
mFDR for streaming feature selection

Streaming feature selection was introduced in JMLR 2006 (with Zhou, Stine and Ungar).
Let $W(j)$ be the "alpha wealth" at time j. Then for a series of p-values p_j, we can define:

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

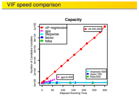
Theorem
(Foster and Stine, 2008, JRSS-B) An alpha-investing rule governed by (1) with initial alpha-wealth $W(0) \le \alpha\,\eta$ and pay-out $\omega \le \alpha$ controls $mFDR_\eta$ at level α.

- Can be done really fast (2011)



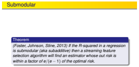
VIF regression

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



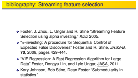
VIF speed comparison

- Works well under sub-modularity (2013)



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.



bibliography: 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."

- But it encourages dynamic variable selection

Properties of alpha investing

- “provides” mFDR protection (2008)

mFDR for streaming feature selection

Streaming feature selection was introduced in JMLR 2006 (with Zhou, Stine and Ungar).
Let $W(j)$ be the “alpha wealth” at time j. Then for a series of p-values p_j, we can define:

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

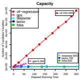
Theorem
(Foster and Stine, 2008, JRSS-B) An alpha-investing rule governed by (1) with initial alpha-wealth $W(0) \le \alpha\,\eta$ and pay-out $\omega \le \alpha$ controls mFDR$_\eta$ at level α.

- Can be done really fast (2011)

VIF regression

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

VIF speed comparison



- Works well under sub-modularity (2013)

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.

bibliography: 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.”

- But it encourages dynamic variable selection
- Enter the dragon!

Sequential data collection

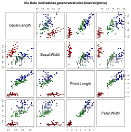
Talking points:

- We can to grow the data set as we do more queries
  - Still cheaper to collectively generate data rather than doing it fresh
  - In other words, the sample complexity of doing k queries is $O(k)$ if each is done on a seperate dataset but only $O(\sqrt{k})$ if each is done on one large dataset. (Thanks Jonathan!)

Picture = 1000 words

Talking points:

- A picture is worth a 1000 queries.
  - The adage of "always graph your data" counts as doing many queries against the distribution
  - People can pick out several different possible patterns in one glance at a graph
  - Probably not worth 1000, more like 50



Biased questions: Entropy vs number of queries

Talking points:

- In variable selection, we mostly have very wide confidence intervals when we fail to reject the null.
  - Can this be used to allow more queries?
  - Can the bound be phrase in terms of entropy of the number of yes/no questions?

Significant digits

Talking points:

- Never quote: "$\hat{\beta} = 3.2123245386703$"
  - All I have had in the past to justify not giving all these extra digits was saying something like, "do you really believe it is ...703 and not ...704?"
  - Now it is a theorem! You are leaking too much information and saying things about the data and not about the population (Thanks Cynthia!)
  - I've argued using about a 1-SD scale for approximation (based on information theory). I think differential privacy asks for even cruder scales. Can this difference be closed?

Thanks!

Sequential data collection

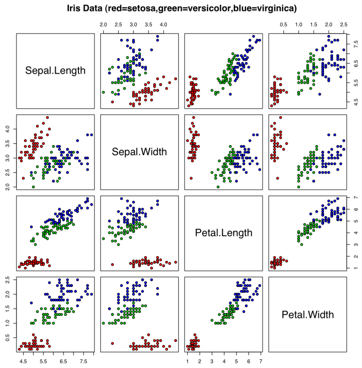
Talking points:

- We can to grow the data set as we do more queries
  - Still cheaper to collectively generate data rather than doing it fresh
  - In other words, the sample complexity of doing k queries is $O(k)$ if each is done on a seperate dataset but only $O(\sqrt{k})$ if each is done on one large dataset. (Thanks Jonathan!)

Picture = 1000 words

Talking points:

- A picture is worth a 1000 queries.
  - The adage of “always graph your data” counts as doing many queries against the distribution
  - People can pick out several different possible patterns in one glance at a graph
  - Probably not worth 1000, more like 50



Biased questions: Entropy vs description length

Talking points:

- In variable selection, we mostly have very wide confidence intervals when we fail to reject the null.
  - Can this be used to allow more queries?
  - Can the bound be phrase in terms of entropy of the number of yes/no questions?

Significant digits

Talking points:

- Never quote: "$\hat{\beta} = 3.2123245386703$"
  - All I have had in the past to justify not giving all these extra digits was saying something like, "do you really believe it is ...703 and not ...704?"
  - Now it is a theorem! You are leaking too much information and saying things about the data and not about the population (Thanks Cynthia!)
  - I've argued using about a 1-SD scale for approximation (based on information theory). I think differential privacy asks for even cruder scales. Can this difference be closed?

mFDR for streaming feature selection

Streaming feature selection was introduced in JMLR 2006 (with Zhou, Stine and Ungar).
Let $W(j)$ be the "alpha wealth" at time j. Then for a series of p-values p_j, we can define:

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

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 $O(np)$.

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.

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
```

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.

bibliography: 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."

Risk Inflation

- Prediction risk:

$$R(\hat{\beta}, \beta) = E_\beta |\mathbf{X}\beta - \mathbf{X}\hat{\beta}|_2^2$$

- Target risk:

$$R(\hat{\beta}) = q\sigma^2$$

- The L-0 penalized regression is within a log factor of this target.

Theorem (Foster and George, 1994)

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

Risk Inflation

- Prediction risk:

$$R(\hat{\beta}, \beta) = E_\beta |\mathbf{X}\beta - \mathbf{X}\hat{\beta}|_2^2$$

- Target risk:

$$R(\hat{\beta}) = q\sigma^2$$

- The L-0 penalized regression is within a log factor of this target.

Theorem (Foster and George, 1994)

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

- Also proven by Donoho and Johnstone in the same year.

Risk Inflation

- Prediction risk:

$$R(\hat{\beta}, \beta) = E_\beta |\mathbf{X}\beta - \mathbf{X}\hat{\beta}|_2^2$$

- Target risk:

$$R(\hat{\beta}) = q\sigma^2$$

- The L-0 penalized regression is within a log factor of this target.

Theorem (Foster and George, 1994)

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

Risk Inflation

- Prediction risk:

$$R(\hat{\beta}, \beta) = E_\beta |\mathbf{X}\beta - \mathbf{X}\hat{\beta}|_2^2$$

- Target risk:

$$R(\hat{\beta}) = q\sigma^2$$

- The L-0 penalized regression is within a log factor of this target.

Theorem (Foster and George, 1994)

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

- This bound is also tight: I.e. there are design matrices for

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.*

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.

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	$+1+\delta$	$-1+\delta$
⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮
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	$+1+\delta$	$-1+\delta$

- “Model:”

$$Y \sim D1 + D2 + \cdots + Dn/2 + X1 + X2$$

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:

$$Y = \frac{1}{\delta} X1 + \frac{1}{\delta} X2$$

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 will add all the distractors before adding either X1 or X2.
- (If $\delta < 1/\sqrt{n}$)

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 add all the other features before adding the two “correct” features
- (True for the standardized version with $\delta < 1/\sqrt{n}$.)

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	$+1+\delta$	$-1+\delta$

- This example breaks stepwise regression and lasso. But clearly better algorithms exist.
- Or do they?

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

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

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

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

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

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

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

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

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

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

- It was previously known that that

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

L0 regression is hard

Theorem (Zhang, Wainwright, Jordan 2014)

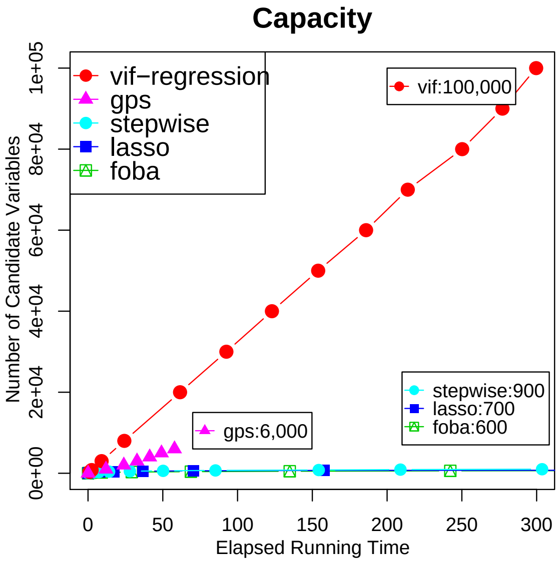
There exists an design matrix X such that no polynomial time algorithm which outputs q variables achieves a risk better than

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

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

- Note: No cheating on the dimension.
- What if we let it use $2q$ variables? Could we get good risk?

VIF speed comparison



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.* $|\hat{\beta}|_0 \ll p$*)*

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. $|\hat{\beta}|_0 \ll p$)*

- 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).

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. $|\hat{\beta}|_0 \ll p$)*

- The sparsity results depend on the assumptions used. We can get $|\hat{\beta}|_0 < cq$ easily, and $|\hat{\beta}|_0 < p^{.99}$ with difficulty.
- Difficult to improve to $|\hat{\beta}|_0 \leq p$ since then all the heavy lifting is being done by the accuracy claims.

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. $|\hat{\beta}|_0 \ll p$)*

bibliography: 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.”