Approximation regularization for analysis of large data sets

Daniel J. McDonald
Indiana University, Bloomington
mypage.iu.edu/ dajmcdon

21 October 2016

OBLIGATORY "DATA IS BIG" SLIDE

Modern statistical applications — genomics, neural image analysis, text analysis, weather prediction — have large numbers of covariates p

Also frequently have lots of observations n.

Need algorithms which can handle these kinds of data sets, with good statistical properties

LESSON OF THE TALK

Many statistical methods use (perhaps implicitly) a singular value decomposition (SVD).

The SVD is computationally expensive.

We want to understand the statistical properties of some approximations which speed up computation and save storage.

Spoiler: sometimes approximations actually improve the statistical properties

CORE TECHNIQUES

Suppose we have a matrix $\mathbb{X} \in \mathbb{R}^{n \times p}$ and vector $Y \in \mathbb{R}^n$

LEAST SQUARES:

$$\widehat{\beta} = \underset{\beta}{\operatorname{argmin}} \, ||\mathbb{X}\beta - Y||_2^2$$

CORE TECHNIQUES

If $\mathbb{X}$ fits into RAM, there exist excellent algorithms in LAPACK that are

- Double precision
- Very stable
- cubic complexity, $O(\min\{np^2, pn^2\})$, with small constants
- require extensive random access to matrix

There is a lot of interest in finding and analyzing techniques that extend these approaches to large(r) problems

OUT-OF-CORE TECHNIQUES

If $\mathbb{X}$ is too large to manipulate in RAM, there are other approaches

- (Stochastic) gradient descent
- Conjugate gradient
- Rank-one QR updates
- Krylov subspace methods

OUT-OF-CORE TECHNIQUES

Many techniques focus on randomized compression

This is sometimes known as sketching or preconditioning

- Rokhlin, Tygert, (2008) “A fast randomized algorithm for overdetermined linear least-squares regression”.
- Drineas, Mahoney, et al., (2011) “Faster least squares approximation”.
- Woodruff (2013) “Sketching as a Tool for Numerical Linear Algebra”.
- Ma, Mahoney, and Yu, (2015), “A statistical perspective on algorithmic leveraging”.

COMPRESSION

BASIC IDEA:

- Choose some matrix $Q \in \mathbb{R}^{q \times n}$.
- Use $Q\mathbb{X}$ and QY instead

Finding $Q\mathbb{X}$ for arbitrary Q and $\mathbb{X}$ takes $O(qnp)$ computations

This can be expensive,

To get this approach to work, we need some structure on Q

THE Q MATRIX

- Gaussian:
  Well behaved distribution and eas(ier) theory. Dense matrix
- Fast Johnson-Lindenstrauss Methods
- Randomized Hadamard (or Fourier) transformation:
  Allows for $O(np\log(p))$ computations.
- $Q = \pi\tau$ for π a permutation of I and $\tau = [I_q\ 0]$:
  $Q\mathbb{X}$ means "read q (random) rows"
- Sparse Bernoulli:

$$Q_{ij} \overset{i.i.d.}{\sim} \begin{cases} 1 & \text{with probability } 1/(2s) \\ 0 & \text{with probability } 1 - 1/s \\ -1 & \text{with probability } 1/(2s) \end{cases}$$

This means $Q\mathbb{X}$ takes $O\left(\frac{qnp}{s}\right)$ "computations" on average

TYPICAL RESULTS

The general philosophy: Find an approximation that is as close as possible to the solution of the original problem

A typical result would be to find an $\tilde{\beta}$ such that

$$\left|\left|\mathbb{X}\tilde{\beta} - Y\right|\right|_2^2 \leq (1+\epsilon)\left(\min_{\beta} ||\mathbb{X}\beta - Y||_2^2\right)$$

Here, $\tilde{\beta}$ should be 'easier' to compute than the minimization

Collaborator & Grant Support

Collaborator:

Darren Homrighausen, Southern Methodist
Department of Statistics

Grant Support:

NSF, Institute for New Economic Thinking

Atypical (better) result: Compressed regression

Risk

Form a loss function $\ell : \Theta \times \Theta \to \mathbb{R}^+$

The quality of an estimator is given by its risk

$$R(\widehat{\theta}) = \mathbb{E}\left[\ell(\widehat{\theta}, \theta)\right]$$

We could use ℓ_2 estimation risk:

$$R(\widehat{\theta}) = \mathbb{E}\left|\left|\theta - \widehat{\theta}\right|\right|_2^2$$

or excess ℓ_2 prediction risk

$$\begin{aligned} R(\widehat{\theta}) &= \mathbb{E}\left|\left|Y - \mathbb{X}\widehat{\theta}\right|\right|_2^2 = \mathbb{E}\left|\left|Y - \mathbb{X}\theta + \mathbb{X}\theta - \mathbb{X}\widehat{\theta}\right|\right|_2^2 \\ &\propto \text{constant} + \mathbb{E}\left|\left|\mathbb{X}(\theta - \widehat{\theta})\right|\right|_2^2 \\ &\propto \mathbb{E}\left|\left|\theta - \widehat{\theta}\right|\right|_2^2 \end{aligned}$$

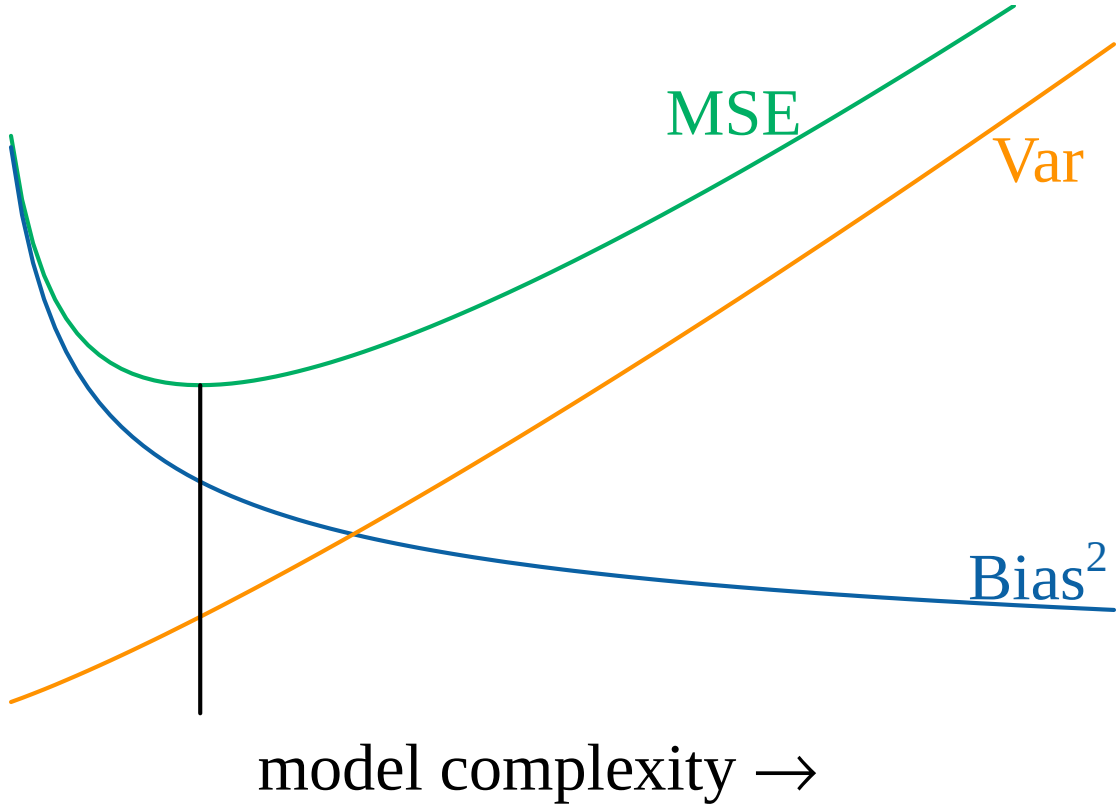
RISK DECOMPOSITION

For an approximation $\tilde{\theta}$ of $\widehat{\theta}$,

$$\begin{aligned}\mathbb{E}\left|\left|\theta-\tilde{\theta}\right|\right|_2^2 &= \mathbb{E}\left|\left|\tilde{\theta}-\widehat{\theta}+\widehat{\theta}-\mathbb{E}\widehat{\theta}+\mathbb{E}\widehat{\theta}-\theta\right|\right|_2^2 \\ &\leq \text{Approx. error}^2 + \text{Variance} + \text{Bias}^2\end{aligned}$$

- Previous analyses focus only on the approximation error.
- Specifically, they compare the approximation to the UMVUE.

BIAS-VARIANCE TRADEOFF



Typical result compares to the zero-bias estimator which is assumed to have small variance.

Atypical result considers whether adding some bias might reduce variance.

COMPRESSED REGRESSION

Let $Q \in \mathbb{R}^{q \times n}$

Let's solve the fully compressed least squares problem

$$\widehat{\beta}_{FC} = \underset{\beta}{\text{argmin}} \, ||Q(\mathbb{X}\beta - Y)||_2^2 \approx \underset{\beta}{\text{argmin}} \, ||\mathbb{X}\beta - Y||_2^2$$

$\rightarrow$ A common way to solve least squares problems that are:

- Very large or
- Poorly conditioned

The numerical/theoretical properties generally depend on Q, q

Family of 4

Full compression:

$$\widehat{\beta}_{FC} = (\mathbb{X}^\top Q^\top Q \mathbb{X})^{-1} \mathbb{X}^\top Q^\top Q Y$$

Partial compression:[1]

$$\widehat{\beta}_{PC} = (\mathbb{X}^\top Q^\top Q \mathbb{X})^{-1} \mathbb{X}^\top Y$$

Linear and Convex combination compression:

$$W = [\, \widehat{Y}_{FC},\ \widehat{Y}_{PC} \,] \qquad b = [\, \widehat{\beta}_{FC},\ \widehat{\beta}_{PC} \,]$$

$$\widehat{\alpha}_{lin} = \underset{\alpha}{\operatorname{argmin}} \, ||W\alpha - Y||_2^2 \qquad \widehat{\alpha}_{con} = \underset{\substack{0<\alpha \\ \sum \alpha = 1}}{\operatorname{argmin}} \, ||W\alpha - Y||_2^2$$

$$\widehat{\beta}_{lin} = b\widehat{\alpha}_{lin} \qquad \widehat{\beta}_{con} = b\widehat{\alpha}_{con}$$

[1] see the work of Stephen Becker CU Boulder Applied Math

WHY THESE?

Note:

$$||Q(\mathbb{X}\beta - Y)||_2^2 \propto \beta^\top \mathbb{X}^\top Q^\top Q \mathbb{X}\beta - 2\beta^\top \mathbb{X}^\top Q^\top QY$$

- Turns out that FC is (approximately) unbiased, and therefore worse than OLS (has high variance)
- On the other hand, PC is biased and empirics demonstrate low variance
- Combination should give better statistical properties

COMPRESSED REGRESSION

With this Q, compression "works" in practice:

- Computational savings: $O\left(\frac{qnp}{s} + qp^2\right)$
- Approximately the same estimation risk as OLS

This is good, but we had a realization:

If ridge regression is better than OLS, why not "point" the approximation at ridge?

COMPRESSED RIDGE REGRESSION

This means introducing a tuning parameter λ and defining:

$$\widehat{\beta}_{PC}(\lambda) = (\mathbb{X}^\top Q^\top Q\mathbb{X} + \lambda I)^{-1}\mathbb{X}^\top Y$$
$$\widehat{\beta}_{FC}(\lambda) = (\mathbb{X}^\top Q^\top Q\mathbb{X} + \lambda I)^{-1}\mathbb{X}^\top Q^\top QY$$

Everything else about the procedure is the same

This has the same computational complexity, but has much lower risk

Let's look at an (a)typical result...

Evidence from simulations

Simulation setup

- Draw $\mathbb{X}_i \sim \text{MVN}(0, (1-\rho)I_p + \rho \mathbf{1}\mathbf{1}^\top)$
  - We use $\rho = \{0.2,\ 0.8\}$.
- Draw $\beta \sim \text{N}(0, \tau^2 I_p)$
- Draw $Y_i = \mathbb{X}_i^\top \beta + \epsilon_i$ with $\epsilon_i \sim \text{N}(0, \sigma^2)$.

BAYES ESTIMATOR

- For this model, the optimal estimator (in MSE) is

$$\widehat{\beta}_B = (\mathbb{X}^\top \mathbb{X} + \lambda_* I_p)^{-1} \mathbb{X}^\top Y$$

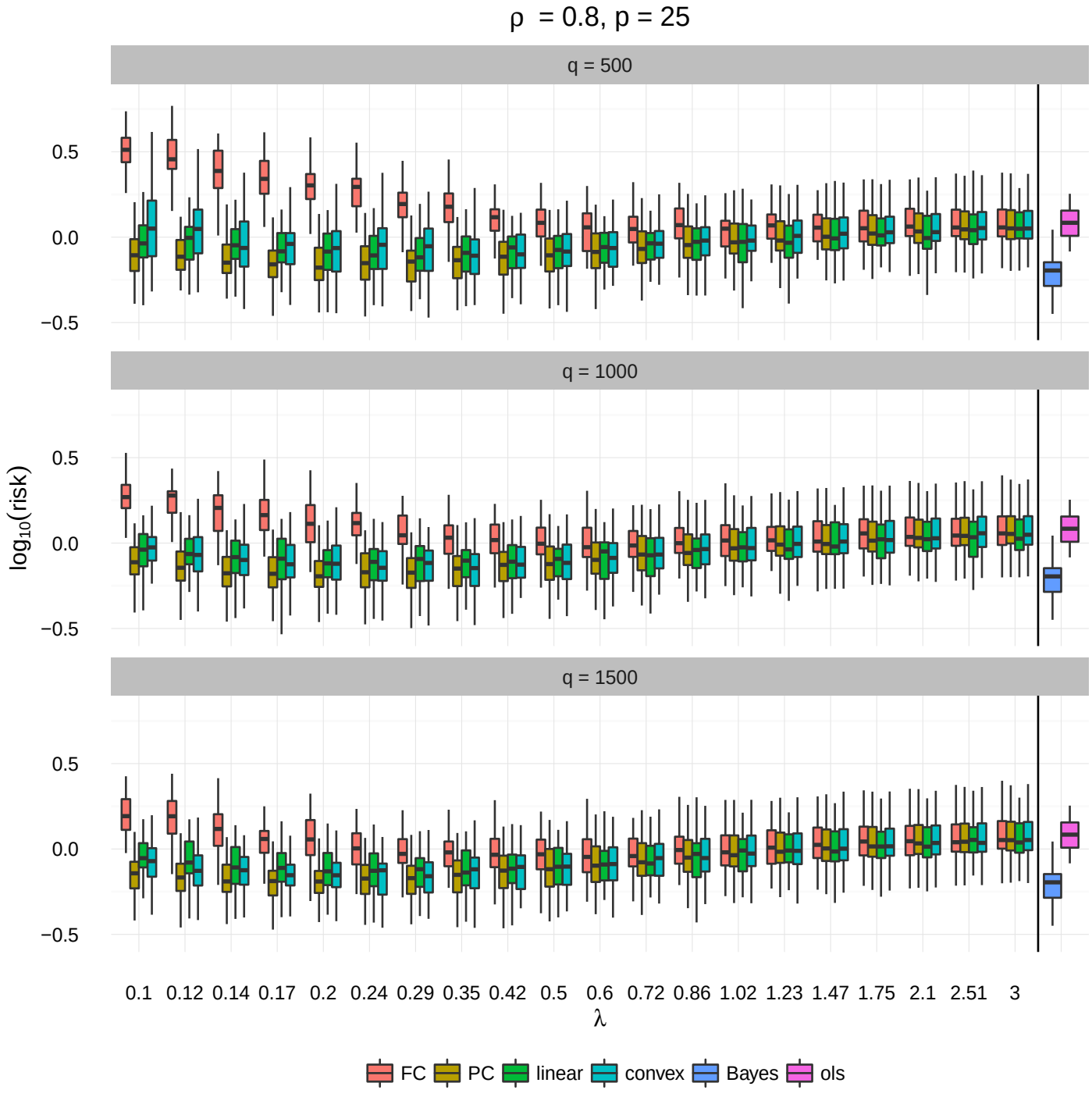
- In particular, with $\lambda_* = \frac{\sigma^2}{n\tau^2}$
- This is the posterior mode of the Bayes estimator under conjugate normal prior
- It is also the ridge regression estimator for a particular λ

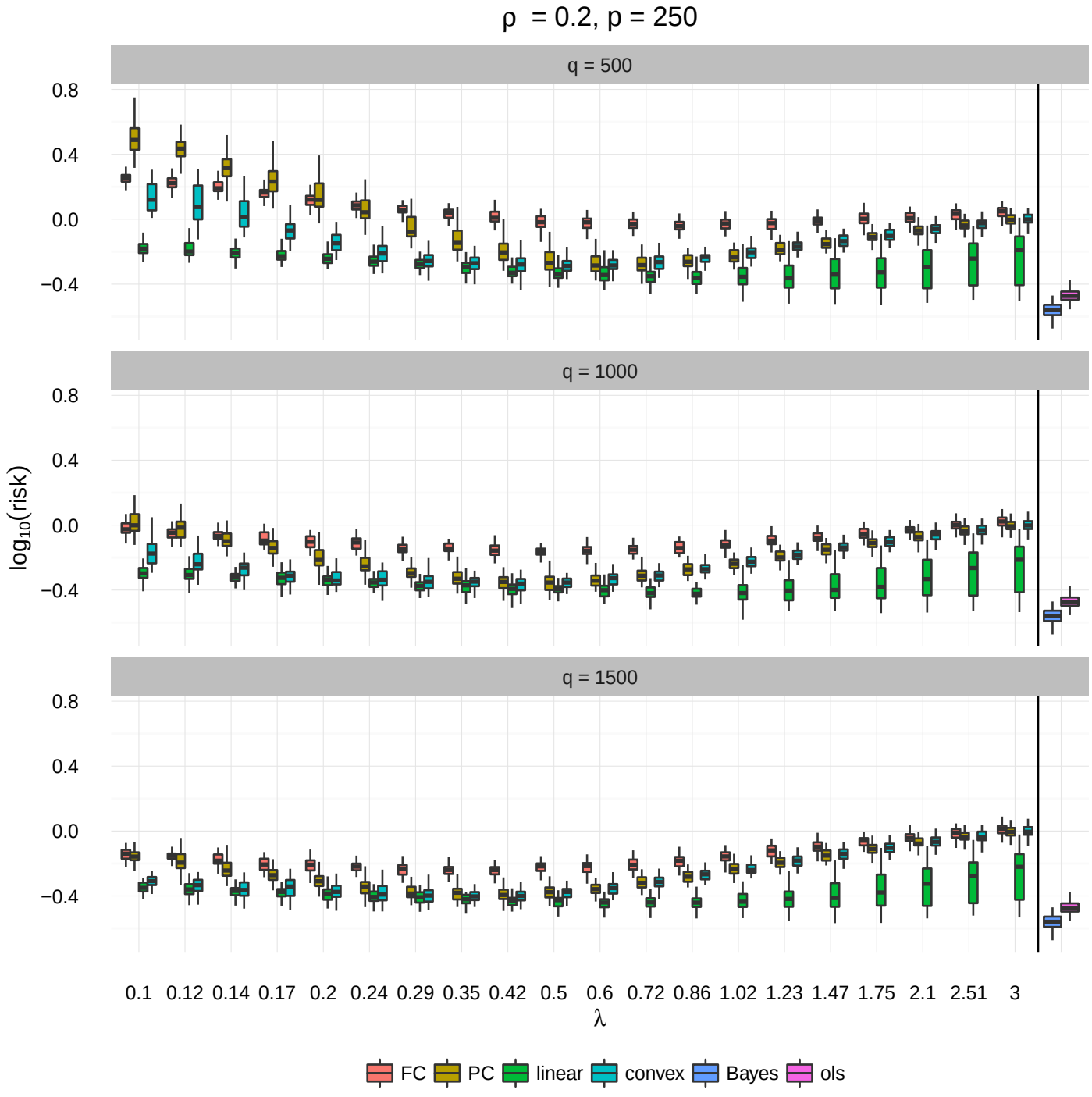
GOLDILOCKS

With $p < n$

1. If λ_* is too big, we will tend to shrink all coefficients to 0.
   - This problem is too hard.
2. If λ_* is too small, OLS will be very close to the optimal estimator.
   - This problem is too easy.
3. Need τ^2, σ^2 "just right".

- Take $\tau^2 = \pi/2$. This implies $\mathbb{E}[|\beta_i|] = 1$ (convenient)
- Take $n = 2500$. Big but computable.
- Take $\sigma^2 = 25 \Rightarrow \lambda_* \approx 0.15$ (reasonable)
- Take $p \in \{25,\ 50,\ 125,\ 250\}$

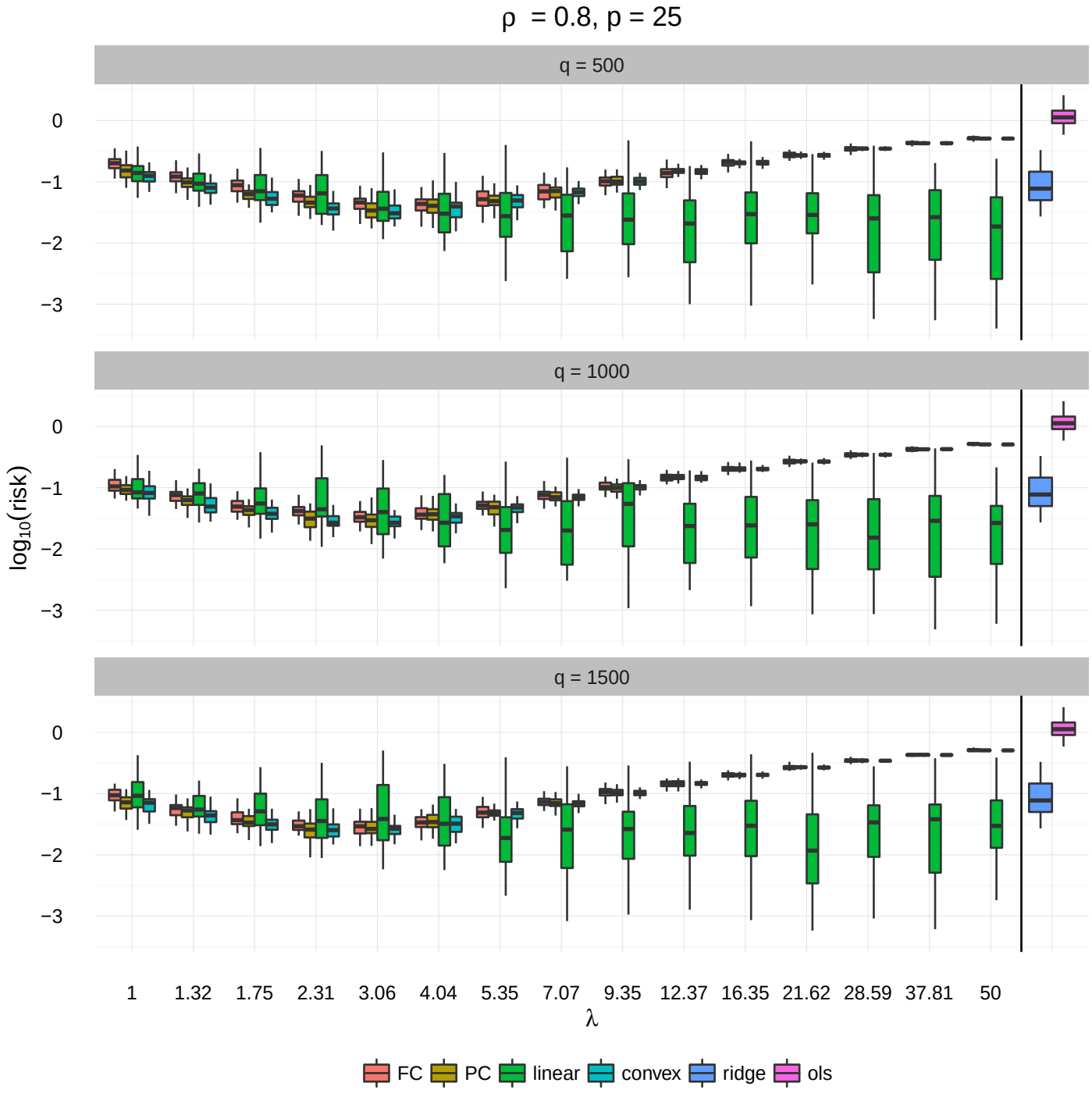

ρ = 0.8, p = 25
q = 500
q = 1000
q = 1500
log10(risk)
0.5
0.0
−0.5
0.1 0.12 0.14 0.17 0.2 0.24 0.29 0.35 0.42 0.5 0.6 0.72 0.86 1.02 1.23 1.47 1.75 2.1 2.51 3
λ
FC PC linear convex Bayes ols

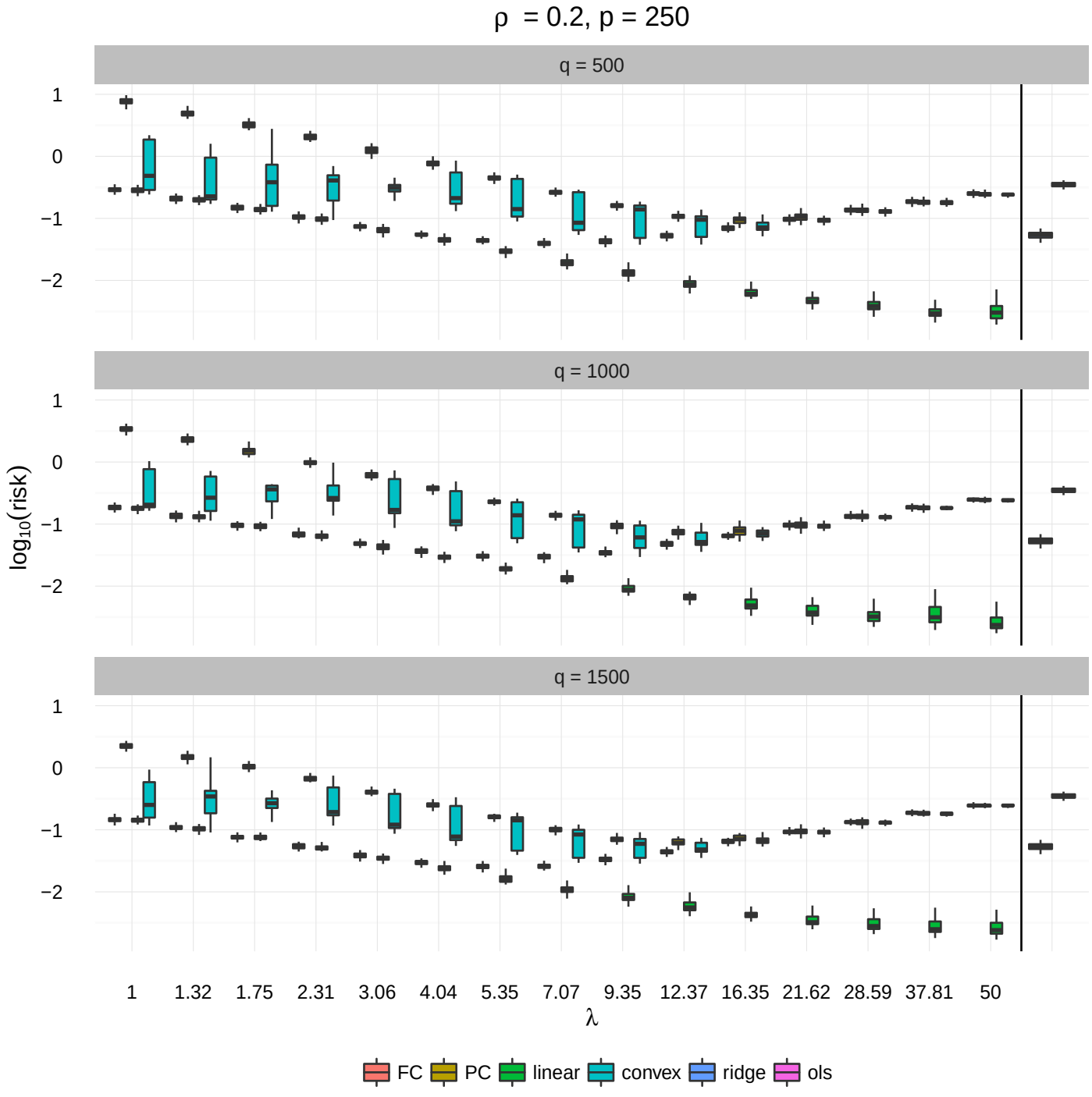

ρ = 0.2, p = 250
q = 500
q = 1000
q = 1500
0.8
0.4
0.0
−0.4
log10(risk)
0.1 0.12 0.14 0.17 0.2 0.24 0.29 0.35 0.42 0.5 0.6 0.72 0.86 1.02 1.23 1.47 1.75 2.1 2.51 3
λ
FC PC linear convex Bayes ols

SECOND VERSE. . .

In that case, ridge was optimal.

We did it again with $\beta_i \equiv 1$.


ρ = 0.8, p = 25
q = 500
q = 1000
q = 1500
log10(risk)
0
−1
−2
−3
1
1.32
1.75
2.31
3.06
4.04
5.35
7.07
9.35
12.37
16.35
21.62
28.59
37.81
50
λ
FC
PC
linear
convex
ridge
ols

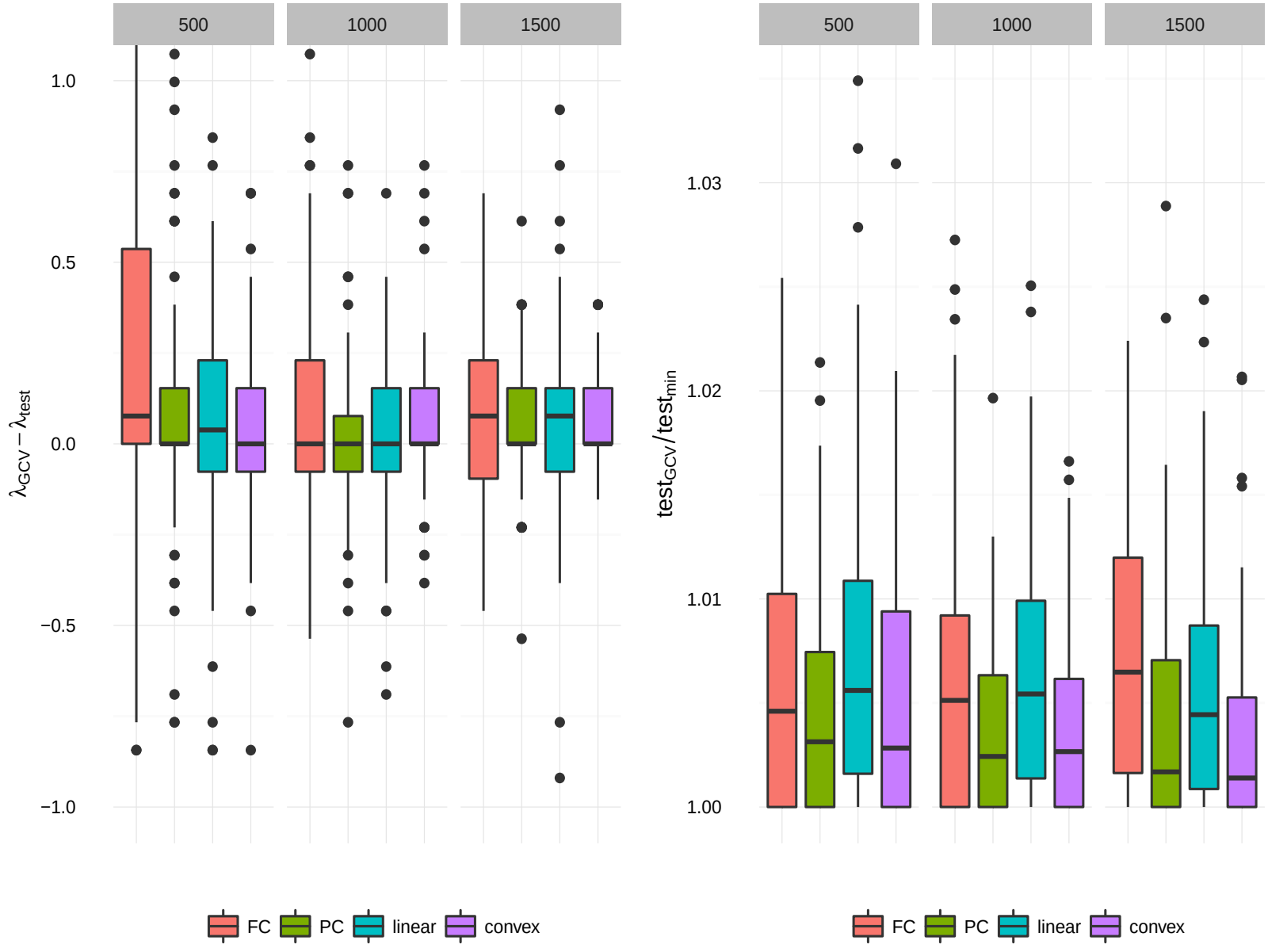

ρ = 0.2, p = 250
q = 500
q = 1000
q = 1500
log10(risk)
1
0
−1
−2
1
1.32
1.75
2.31
3.06
4.04
5.35
7.07
9.35
12.37
16.35
21.62
28.59
37.81
50
λ
FC
PC
linear
convex
ridge
ols

SELECTING TUNING PARAMETERS

- We use GCV with the degrees of freedom:

$$\text{GCV}(\lambda) = \frac{\left|\left|\mathbb{X}\widehat{\beta}(\lambda) - Y\right|\right|_2^2}{(1 - df/n)^2}$$

- df is easy for full or partial compression
- For the other cases, we do a dumb approximation: weighted average using $\widehat{\alpha}$
- Also calculate the divergence to derive Stein's Unbiased Risk Estimate
- After tedious algebra, it had odd behavior in practice (can be huge, or negative!?!)



$$n = 2500 \;\; p = 250 \;\; \rho = 0.8$$

Theoretical results (sketch)

STANDARD RIDGE RESULTS

Theorem

$$\text{bias}^2\left(\widehat{\beta}_{ridge}(\lambda) \mid \mathbb{X}\right) = \lambda^2 \beta_*^\top V (D^2 + \lambda I_p)^{-2} V^\top \beta_*.$$

$$\text{tr}\left(\mathbb{V}[\widehat{\beta}_{ridge}(\lambda) \mid \mathbb{X}]\right) = \sigma^2 \sum_{i=1}^{p} \frac{d_i^2}{(d_i^2 + \lambda)^2}.$$

WHAT'S THE TRICK?

- Similar results are hard for compressed regression.
- All the estimators depend (at least) on

$$(\mathbb{X}^\top Q^\top Q \mathbb{X} + \lambda I_p)^{-1}$$

- We derived properties of $Q^\top Q$

$$\mathbb{E}\left[\frac{s}{q} Q^\top Q\right] = I_n$$

$$\mathbb{V}\left[\text{vec}\left(\frac{s}{q} Q^\top Q\right)\right] = \frac{(s-3)_+}{q} \text{diag}(\text{vec}\,(I_n)) + \frac{1}{q} I_{n^2} + \frac{1}{q} K_{nn}$$

- So the technique is to do a Taylor expansion around $\frac{s}{q} Q^\top Q = I_n$.

Main result

Theorem

$$\text{bias}^2[\widehat{\beta}_{FC} \mid \mathbb{X}] = \lambda^2 \beta_*^\top V (D^2 + \lambda I_p)^{-2} V^\top \beta_* + \mathbb{E}[R_A \mid \mathbb{X}]$$

$$\begin{aligned}\text{tr}(\mathbb{V}[\widehat{\beta}_{FC} \mid \mathbb{X}]) = \sigma^2 \sum_{i=1}^{p} \frac{d_i^2}{(d_i^2 + \lambda)^2} + \mathbb{E}[R_A \mid \mathbb{X}] + \text{tr}(\mathbb{V}[R_A \mid \mathbb{X}]) \\ + \frac{(s-3)_+}{q} \text{tr}\left(\text{diag}(\text{vec}\,(I_n)) M^\top M \otimes (I - H) M \beta_* \beta_*^\top M^\top (I - H)\right) \\ + \frac{\beta_*^\top M^\top (I - H)^2 M \beta_*}{q} \text{tr}(M M^\top) \\ + \frac{1}{q} \text{tr}\left((I - H) M \beta_* \beta_*^\top M^\top (I - H) M^\top M\right).\end{aligned}$$

Note: $M = (\mathbb{X}^\top \mathbb{X} + \lambda I_p)^{-1} \mathbb{X}^\top$ and $H = \mathbb{X} M$ (hat matrix for ridge regression)

Applications

RNA-SEQ

- short-read RNA sequence data
- using a (Poisson) linear model to predict read counts based on the surrounding nucleotides
- 8 different tissues
- data is publicly available, preprocessing already done

THE DATASETS

- 3 Wold group mouse transcriptomes
  - brain
  - liver
  - skeletal muscle
- 3 Burge group, 15 human tissues
  - merged into 3 groups based on tissue similarities
- 2 Grimmond group, more mouse
  - embryonic stem cells
  - embryoid bodies

PROCESSING

- Use the top 100 highly-expressed genes and surrounding counts
- Need a window size use 39 neighbors (other people did this first)
- Each nucleotide represents a factor (ACTG) so this gives $p = 3 * (39 + 1) + 1 = 121$.
- Take $\log(count + 1)$ to approximate Poisson regression.

RESULTS

				$q = 10000$				$q = 20000$			
dataset	n	ols	ridge	convex	linear	full	partial	convex	linear	full	partial
B1	157614	1.21	1.21	1.21	1.21	1.22	1.21	1.21	1.21	1.22	1.21
B2	125056	1.50	1.50	1.50	1.50	1.51	1.50	1.50	1.50	1.51	1.50
B3	103394	1.47	1.47	1.47	1.47	1.48	1.47	1.47	1.47	1.47	1.47
G1	51751	2.59	2.59	2.60	2.60	2.61	2.60	2.60	2.60	2.60	2.59
G2	64966	2.02	2.02	2.03	2.03	2.04	2.03	2.02	2.02	2.03	2.03
W1	146828	0.91	0.91	0.91	0.91	0.92	0.91	0.91	0.91	0.91	0.91
W2	171776	1.60	1.60	1.60	1.60	1.61	1.60	1.60	1.60	1.61	1.60
W3	143570	1.77	1.77	1.77	1.77	1.78	1.77	1.77	1.77	1.77	1.77

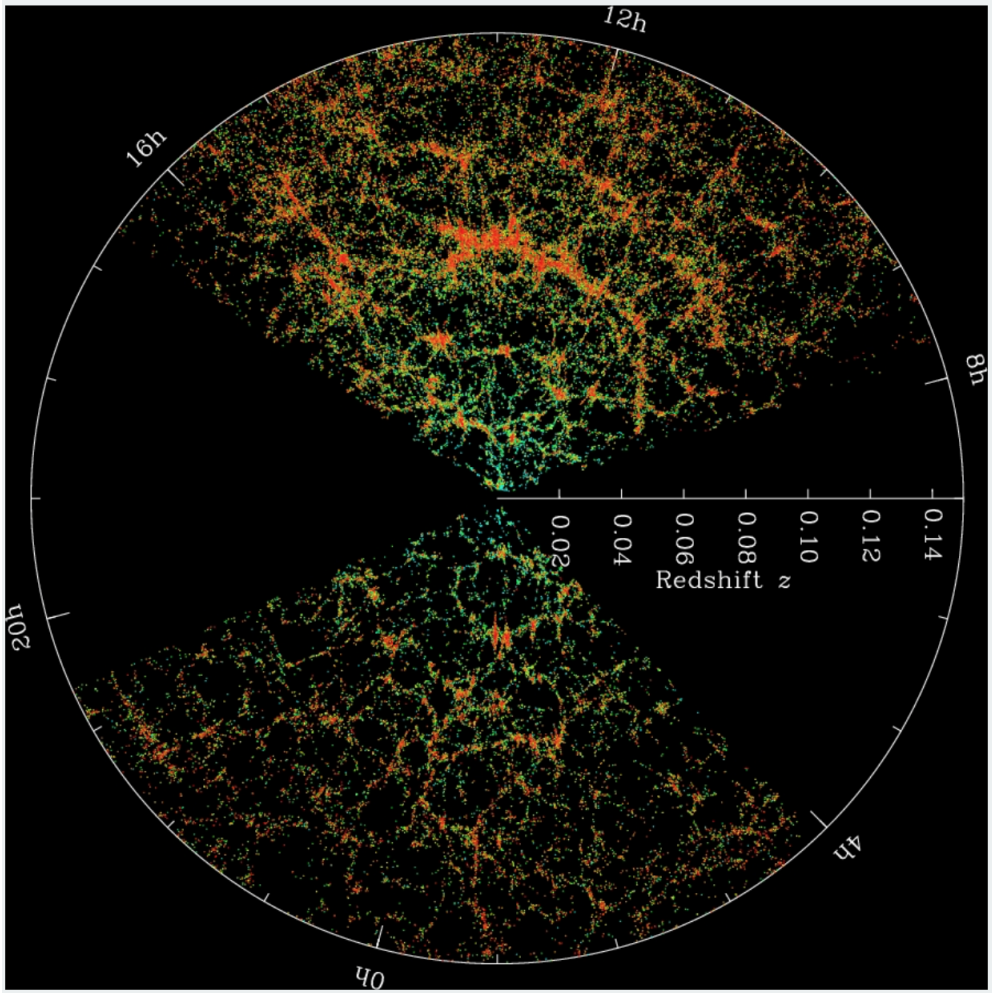
- Test error averaged over 10 replications.
- On each replication, tuning parameters were chosen with GCV.
- At the best tuning parameter, the test error was computed and then these were averaged.
- For each dataset, we randomly chose 75% of the data as training on each replication.

TAKE-AWAY MESSAGE

- $q = 10000$ results in data reductions between 74% and 93%
- $q = 20000$ gives reductions between 48% and 84%
- ridge and ordinary least squares give equivalent test set performance (differing by less than .001%)
- This is an "easy" problem.
- Despite rounding appearance, none of the compressed methods "beat" OLS
- Full compression is the worst.
- Worst performance is FC on the smallest dataset G1, but it's less than 1% off.

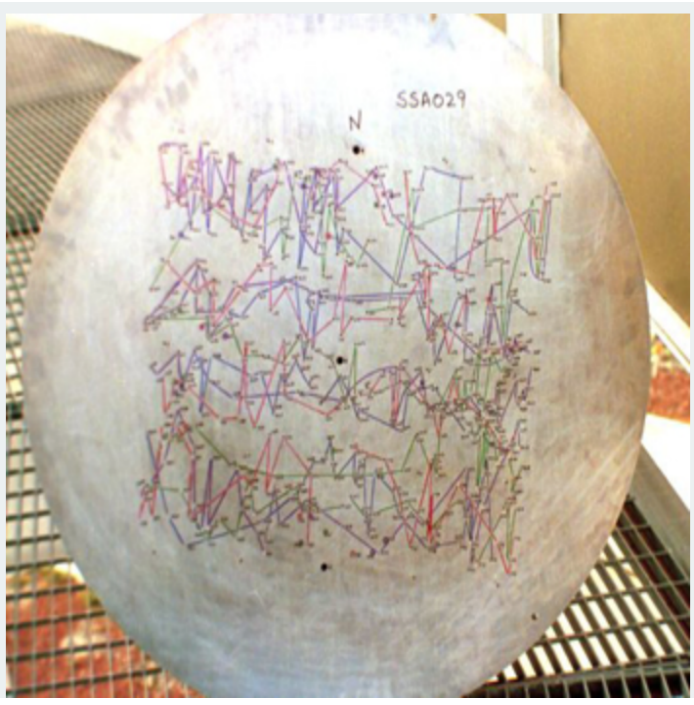
Skip galaxies

CURRENT APPLICATION



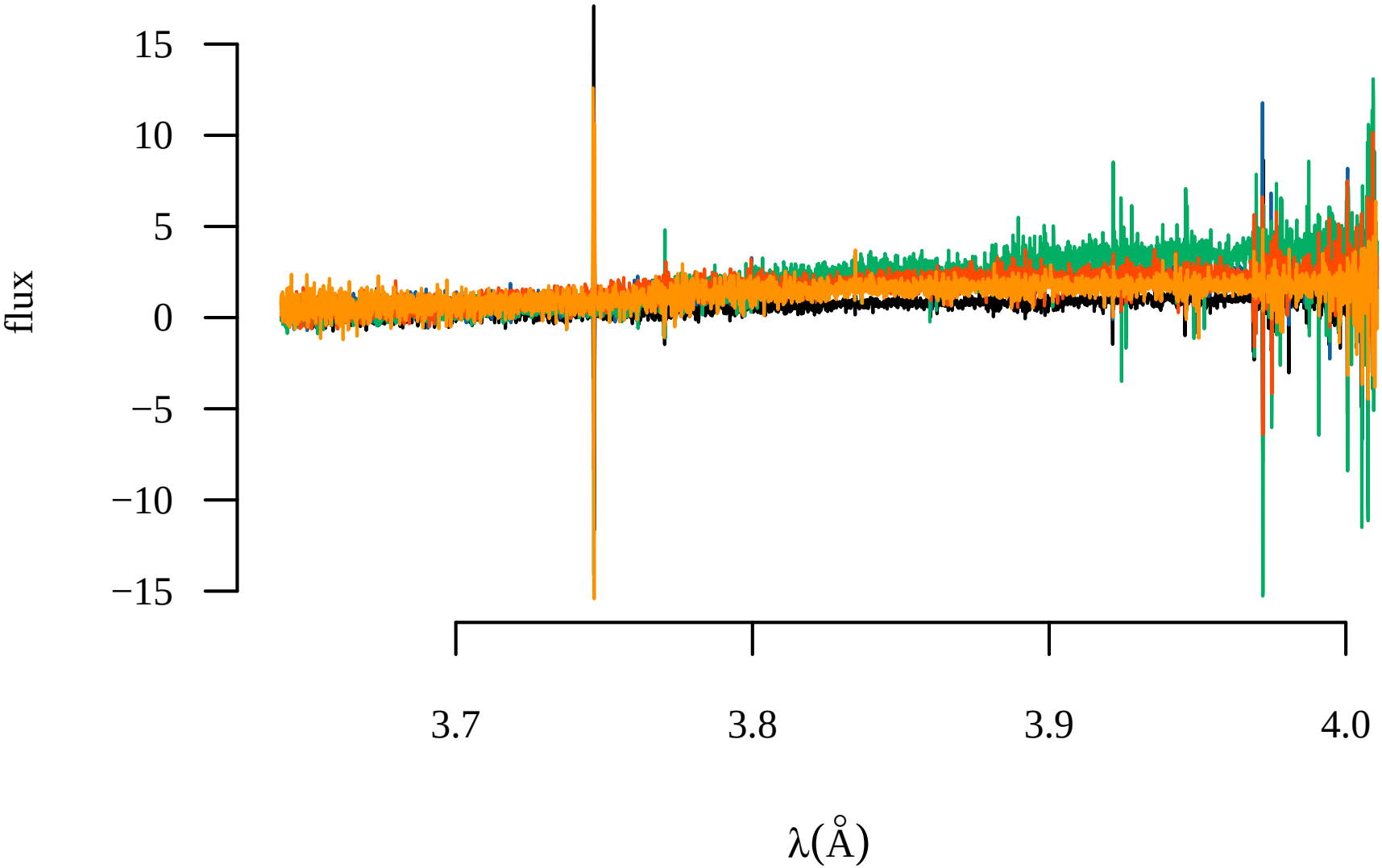
This image and the next are from the Sloan Digital Sky Survey

SDSS

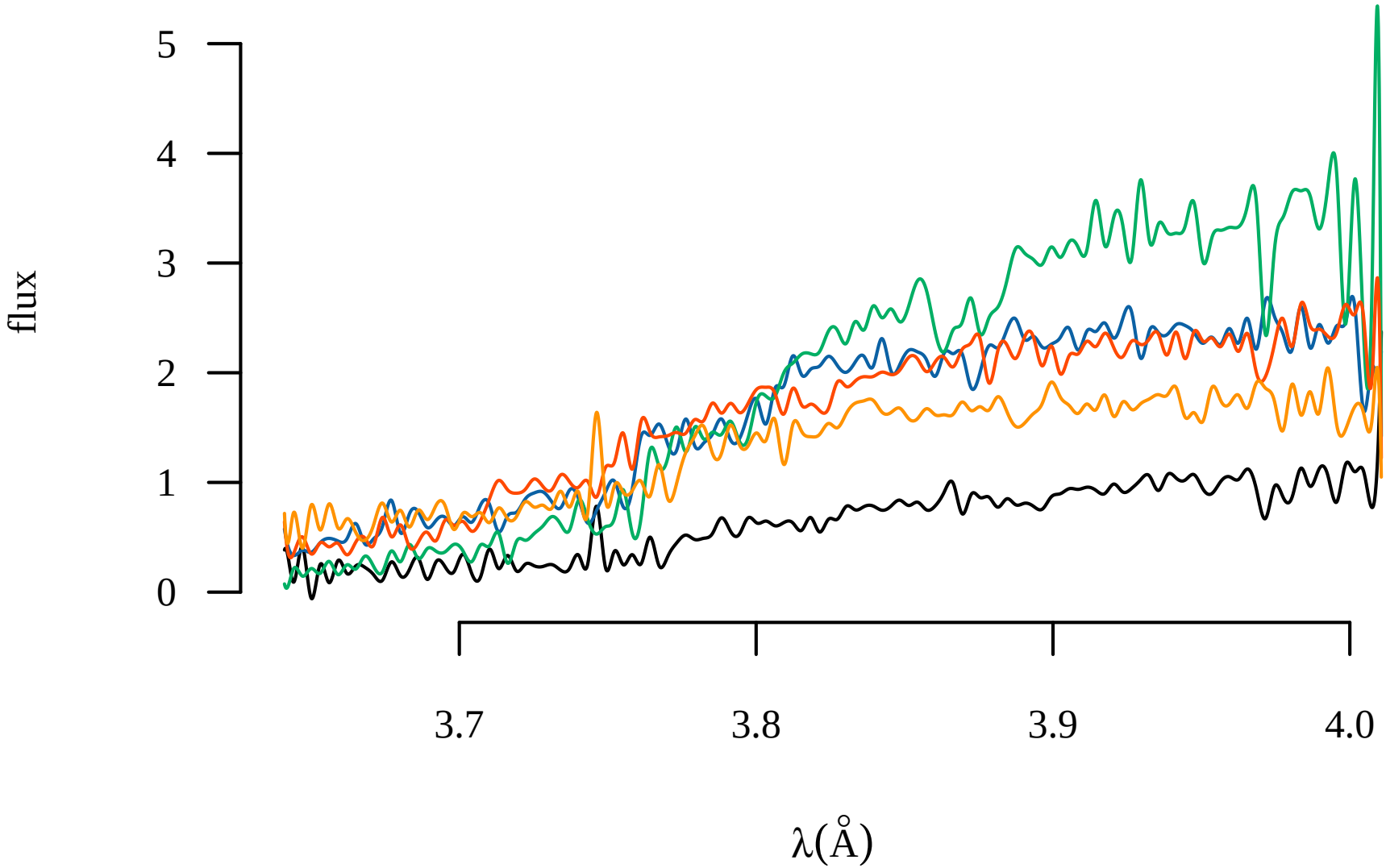


- Repeatedly “photograph” regions of the sky
- Collect information on piles of objects
- Around 200 million galaxy spectra at the moment
- Each spectra contains around 5000 wavelengths
- Want to predict the redshift from the spectra

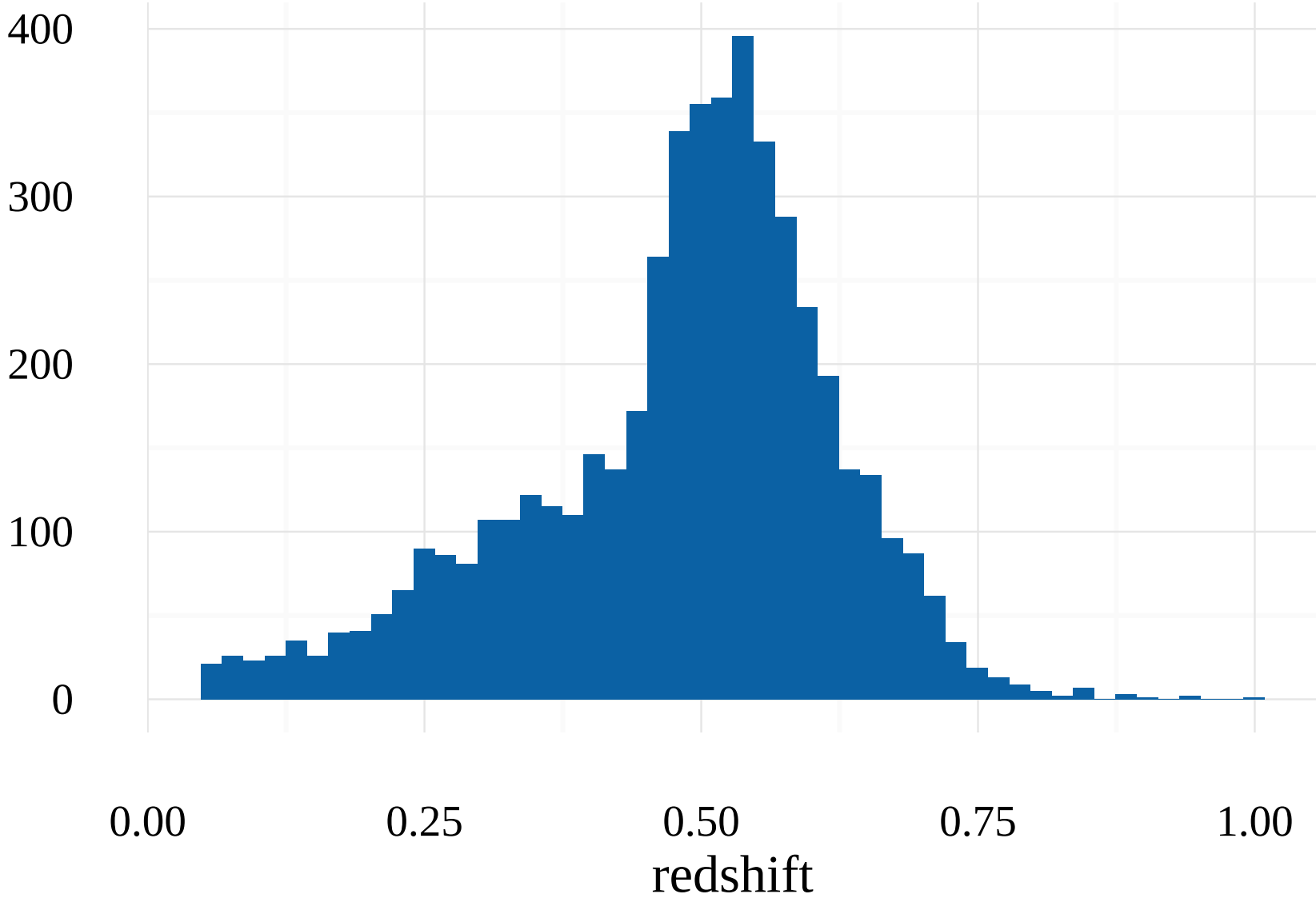
A FEW GALAXIES



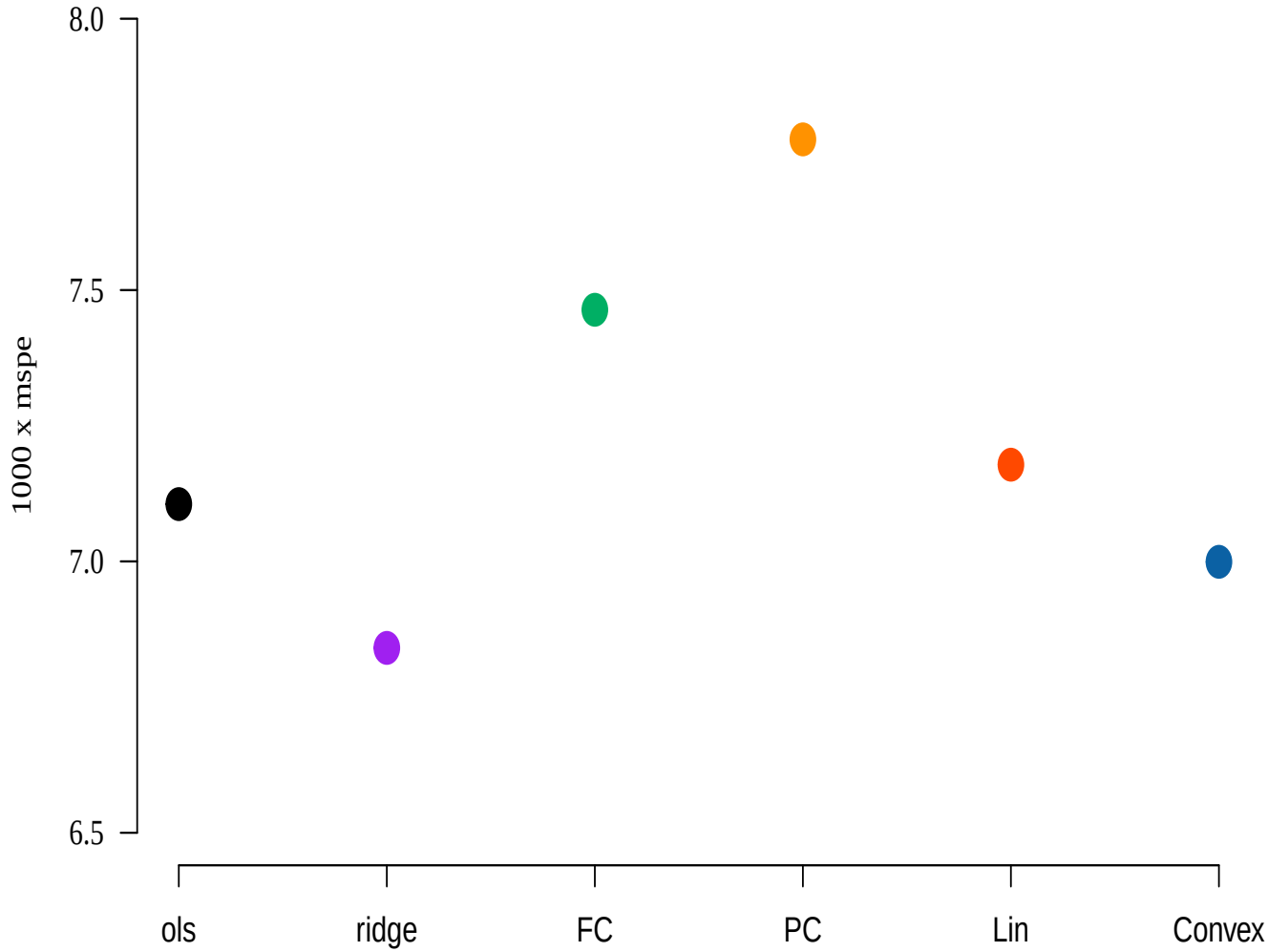
SAME GALAXIES AFTER SMOOTHING



5000 REDSHIFTS



PRELIMINARY RESULTS



Conclusions

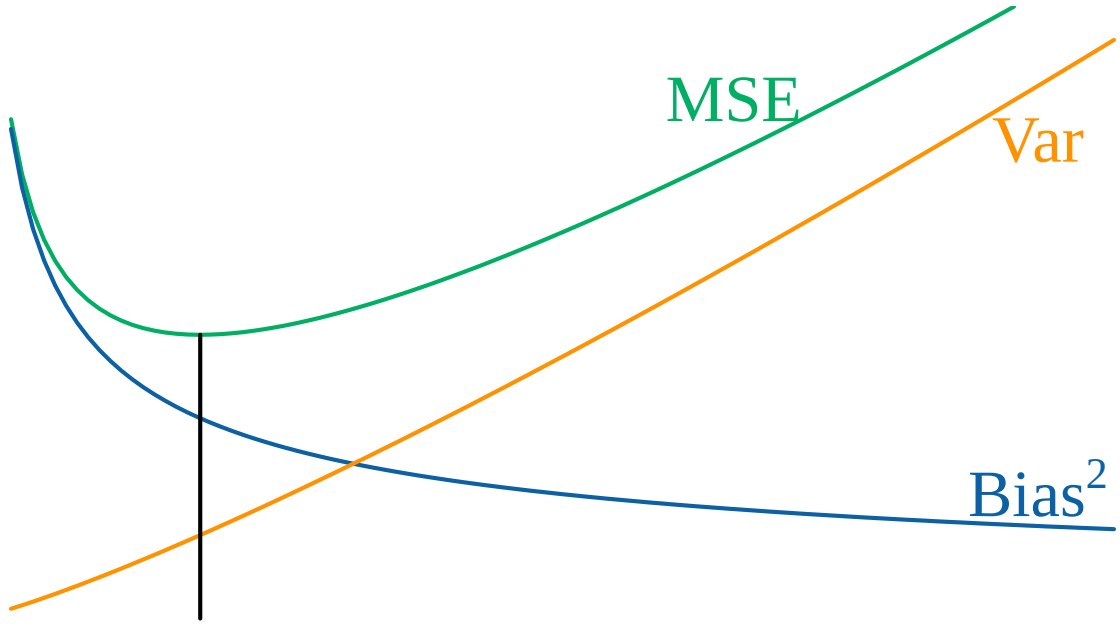
CONCLUSIONS

- Lots of people are looking at approximations to standard statistical methods (OLS, PCA, etc.)
- They characterize the approximation error
- We have been looking at whether we can actually benefit from the approximation
- Today looked at compressed regression

ONGOING AND FUTURE WORK

- We want to generalize our results here to other penalties (lasso)
- Also generalized linear models
- We're also looking at how compression interacts with PCA
- What if $p \gg n$? Can we compress on the other side?
- We have some results for PCA, developing results for PCA regression

OPEN PROBLEM



Can we develop a way to add "computation time" to this picture?