Statistical Learning with Sparsity
The Lasso and Generalizations

Trevor Hastie
Robert Tibshirani
Martin Wainwright
I like simple ideas. Good ideas are often simple, so sometimes I’m in luck! Simple methods often work as well as complex ones — we are all in luck!

I learned a few good tricks early on, then spent the rest of my career finding places to use them!

Good collaborators are essential, especially ones armed with whips!

I love developing algorithms and software. For a technique to get used and take off, good software is essential.
Confessions

- I like simple ideas.
I like simple ideas.

Good ideas are often simple, so sometimes I’m in luck!
• I like simple ideas.
  Good ideas are often simple, so sometimes I’m in luck!
  Simple methods often work as well as complex ones — we are all in luck!
I like simple ideas.
Good ideas are often simple, so sometimes I’m in luck!
Simple methods often work as well as complex ones — we are all in luck!

I learned a few good tricks early on, then spent the rest of my career finding places to use them!
Confessions

- I like simple ideas. Good ideas are often simple, so sometimes I’m in luck! Simple methods often work as well as complex ones — we are all in luck!
- I learned a few good tricks early on, then spent the rest of my career finding places to use them!
- Good collaborators are essential, especially ones armed with whips!
Confessions

• I like simple ideas.
  Good ideas are often simple, so sometimes I’m in luck!
  Simple methods often work as well as complex ones — we are all in luck!

• I learned a few good tricks early on, then spent the rest of my career finding places to use them!

• Good collaborators are essential, especially ones armed with whips!

• I love developing algorithms and software. For a technique to get used and take off, good software is essential.
Linear Models for Wide Data

As datasets grow *wide*—i.e. many more features than samples—the linear model has regained favor as the tool of choice.

**Document classification:** bag-of-words easily leads to $p = 20K$ features and $N = 5K$ document samples. Much more if bigrams, trigrams etc, or documents from Facebook, Google, Twitter!

**Genomics, microarray studies:** $p = 40K$ genes are measured for each of $N = 300$ subjects.

**Genome-wide association studies:** $p = 1–2M$ SNPs measured for $N = 2–500K$ subjects.

**Web activity:** Millions of possible webpages visited, search terms used.
Linear Models for Wide Data

As datasets grow *wide*—i.e. many more features than samples—the linear model has regained favor as the tool of choice.

**Document classification:** bag-of-words easily leads to \( p = 20K \) features and \( N = 5K \) document samples. Much more if bigrams, trigrams etc, or documents from Facebook, Google, Twitter!

**Genomics, microarray studies:** \( p = 40K \) genes are measured for each of \( N = 300 \) subjects.

**Genome-wide association studies:** \( p = 1–2M \) SNPs measured for \( N = 2–500K \) subjects.

**Web activity:** Millions of possible webpages visited, search terms used.

In examples like these we tend to use linear models — e.g. linear regression, logistic regression, Cox model. Since \( p \gg N \), we cannot fit these models using standard approaches.
Regularization

\[ \eta(X) = \beta_0 + \sum_{j=1}^{p} X_j \beta_j \]

We cannot fit linear models with \( p > N \) without some constraints. Common approaches are

**Forward stepwise** adds variables one at a time and stops when overfitting is detected. Regained popularity for \( p \gg N \), since it is the most feasible method among its subset cousins: best-subset, backward stepwise.

**Ridge regression** fits the model subject to the constraint
\[ \sum_{j=1}^{p} \beta_j^2 \leq t. \] Shrinks coefficients toward zero, and hence controls variance. Allows linear models of arbitrary size \( p \) to be fit.
Lasso regression (Tibshirani, 1996) fits the model subject to the constraint $\sum_{j=1}^{p} |\beta_j| \leq t$.
Lasso does variable selection and shrinkage, while ridge only shrinks (a lot to control variance!)

\[ \beta_2 \quad \beta_1 \]

\[ \hat{\beta} \]

\[ \beta_2 \quad \beta_1 \]

\[ \hat{\beta} \]
Lasso regression (Tibshirani, 1996) fits the model subject to the constraint $\sum_{j=1}^{p} |\beta_j| \leq t$. Lasso does variable selection and shrinkage, while ridge only shrinks (a lot to control variance!)

28867 citations (Google) by July 2019!
Ridge versus Lasso

- With ridge, *all* coefficients non-zero, and heavily shrunk for large $p$
- Lasso does variable selection and shrinkage
Brief History of $\ell_1$ Regularization

- Wavelet *Soft Thresholding* (Donoho and Johnstone 1994) in orthonormal setting.
- Same idea used in *Basis Pursuit* (Chen, Donoho and Saunders 1996).
- Extended to many linear-model settings e.g. Survival models (Tibshirani, 1997), logistic regression, and so on.
- Gives rise to a new field *Compressed Sensing* (Donoho 2004, Candes and Tao 2005)—near exact recovery of sparse signals in very high dimensions. In many cases $\ell_1$ a good surrogate for $\ell_0$. 
Cross Validation to select $\lambda$

![Graph showing cross-validation results over log(Lambda) and Test Deviance]

Divide the data into K folds. Fit the model path K times on K-1 folds and predict on the fold omitted. Average results. 

*Need fast algorithms for fitting paths!*
Least Angle Regression


The figure below is on page 609 in ESL, and shows the similarity between lasso and a version of least-squares boosting. Brad let the step-size go to zero, with some wonderful consequences.

\[
t = \sum_{k} |\alpha_k|
\]
Lasso: \( \hat{\beta}(\lambda) = \text{argmin}_\beta \frac{1}{N} \sum_{i=1}^{N} (y_i - \beta_0 - x_i^T \beta)^2 + \lambda \| \beta \|_1 \)

[Graph showing Lasso coefficient path with standardized coefficients]
Least Angle Regression — LAR

A *democratic* version of forward stepwise regression.

- Start with $r = y, \hat{\beta}_1, \hat{\beta}_2, \ldots \hat{\beta}_p = 0$. Assume $x_j$ standardized.
- Find predictor $x_j$ most correlated with $r$.
- Increase $\beta_j$ in the direction of sign $\langle r, x_j \rangle$ until some other competitor $x_k$ has as much correlation with current residual as does $x_j$.
- Move $(\hat{\beta}_j, \hat{\beta}_k)$ in the joint least squares direction for $(x_j, x_k)$ until some other competitor $x_\ell$ has as much correlation with the current residual.
- Continue in this way until all predictors have been entered. Stop when $\langle r, x_j \rangle = 0 \ \forall \ j$, i.e. OLS solution.

This *almost* gives the lasso path; a simple tweak and it does.
History of Path Algorithms

Efficient path algorithms for $\hat{\beta}(\lambda)$ allow for easy and exact cross-validation and model selection.

- In 2001 the LARS algorithm (Efron et al) provides a way to compute the entire lasso coefficient path efficiently at the cost of a full least-squares fit.
History of Path Algorithms

Efficient path algorithms for $\hat{\beta}(\lambda)$ allow for easy and exact cross-validation and model selection.

• In 2001 the LARS algorithm (Efron et al) provides a way to compute the entire lasso coefficient path efficiently at the cost of a full least-squares fit.

• 2001 – 2008: path algorithms pop up for a wide variety of related problems: Group lasso (Yuan & Lin 2006), support-vector machine (H, Rosset, Tibshirani & Zhu 2004), elastic net (Zou & H 2004), quantile regression (Li & Zhu, 2007), logistic regression and glms (Park & H, 2007), Dantzig selector (James & Radchenko 2008), ...
History of Path Algorithms

Efficient path algorithms for $\hat{\beta}(\lambda)$ allow for easy and exact cross-validation and model selection.

- In 2001 the LARS algorithm (Efron et al) provides a way to compute the entire lasso coefficient path efficiently at the cost of a full least-squares fit.


- Many of these do not enjoy the piecewise-linearity of LARS, and seize up on very large problems.
GLMNET and coordinate descent

Team: Friedman, H, Tibshirani, Narasimhan, Simon, Qian, Balasubramaniam

- Solve the lasso problem by coordinate descent: optimize each parameter separately, holding all the others fixed. Updates are trivial. Cycle around till coefficients stabilize.
- Do this on a grid of $\lambda$ values, from $\lambda_{max}$ down to $\lambda_{min}$ (uniform on log scale), using warm starts.
- Can use a variety of loss functions and additive penalties.

Coordinate descent achieves dramatic speedups over all competitors, by factors of 10, 100 and more.

Example: Newsgroup data: 11K obs, 778K features (sparse), 100 values $\lambda$ across entire range, lasso logistic regression; time 29s on a 2010 Macbook Pro.
1997 Tibshirani’s student Wenjiang Fu at U. Toronto develops the *shooting algorithm* for the lasso. Tibshirani doesn’t fully appreciate it.
Brief History of Coordinate Descent for the Lasso

1997  Tibshirani’s student Wenjiang Fu at U. Toronto develops the *shooting algorithm* for the lasso. Tibshirani doesn’t fully appreciate it.

2002  Ingrid Daubechies gives a talk at Stanford, describes a one-at-a-time algorithm for $\ell_1$ image deblurring. Hastie implements it, makes an error, and Hastie + Tibshirani conclude that the method doesn’t work.
Brief History of Coordinate Descent for the Lasso

1997  Tibshirani’s student Wenjiang Fu at U. Toronto develops the *shooting algorithm* for the lasso. Tibshirani doesn’t fully appreciate it.

2002  Ingrid Daubechies gives a talk at Stanford, describes a one-at-a-time algorithm for $\ell_1$ image deblurring. Hastie implements it, makes an error, and Hastie + Tibshirani conclude that the method doesn’t work.

2006  Friedman is external examiner at PhD oral of Anita van der Kooij (Leiden) who uses coordinate descent for elastic net. Friedman, Hastie + Tibshirani revisit this problem.
Brief History of Coordinate Descent for the Lasso

1997  Tibshirani’s student Wenjiang Fu at U. Toronto develops the *shooting algorithm* for the lasso. Tibshirani doesn’t fully appreciate it.

2002  Ingrid Daubechies gives a talk at Stanford, describes a one-at-a-time algorithm for $\ell_1$ image deblurring. Hastie implements it, makes an error, and Hastie + Tibshirani conclude that the method doesn’t work.

2006  Friedman is external examiner at PhD oral of Anita van der Kooij (Leiden) who uses coordinate descent for elastic net. Friedman, Hastie + Tibshirani revisit this problem.

LARS and GLMNET
GLMNET package in R

Fits coefficient paths for a variety of different GLMs and the *elastic net* family of penalties.

Some features of glmnet:

- Models: linear, logistic, multinomial (grouped or not), Poisson, Cox model, and multiple-response grouped linear.
- Elastic net penalty includes *ridge* and *lasso*, and hybrids in between.
- Can handle large number of variables $p$. Along with screening rules we can fit GLMs on GWAS scale.
- Cross-validation functions for all models.
- Can allow for sparse matrix formats for $\mathbf{X}$, and hence massive problems (eg $N = 11K$, $p = 750K$ logistic regression).

*Python and Matlab versions as well on my webpage*
GLMNET package in R (continued)

- Can provide lower and upper bounds for each coefficient; eg: positive lasso
- Useful bells and whistles:
  - offsets — as in GLM, can have part of the linear predictor that is given and not fit. Often used in Poisson models (sampling frame)
  - penalty strengths — can alter relative strength of penalty on different variables. Zero penalty means a variable is *always in* the model. Useful for adjusting for demographic variables
  - observation weights
  - can fit no-intercept models
  - session-wise parameters can be set with new glmnet.options command
Coordinate descent for the lasso

\[
\min_{\beta} \frac{1}{2N} \sum_{i=1}^{N} (y_i - \sum_{j=1}^{p} x_{ij} \beta_j)^2 + \lambda \sum_{j=1}^{p} |\beta_j|
\]

Suppose the \( p \) predictors and response are standardized to have mean zero and variance 1. Initialize all the \( \beta_j = 0 \).

Cycle over \( j = 1, 2, \ldots, p, 1, 2, \ldots \) till convergence:

- Compute the partial residuals \( r_{ij} = y_i - \sum_{k \neq j} x_{ik} \beta_k \).
- Compute the simple least squares coefficient of these residuals on \( j \)th predictor: \( \beta_j^* = \frac{1}{N} \sum_{i=1}^{N} x_{ij} r_{ij} \)
- Update \( \beta_j \) by soft-thresholding:

\[
\beta_j \leftarrow S(\beta_j^*, \lambda) = \text{sign}(\beta_j^*)(|\beta_j^*| - \lambda)_+
\]
Elastic-net penalty family

Family of convex penalties proposed in Zou and H (2005) for $p \gg N$ situations, where predictors are correlated in groups.

\[
\min_{\beta} \frac{1}{2N} \sum_{i=1}^{N} (y_i - \sum_{j=1}^{p} x_{ij}\beta_j)^2 + \lambda \sum_{j=1}^{p} \left[ \frac{1}{2}(1 - \alpha)\beta_j^2 + \alpha|\beta_j| \right]
\]

\(\alpha\) creates a compromise between the lasso and ridge.

Coordinate update is now

\[
\beta_j \leftarrow \frac{S(\beta_j^*, \lambda\alpha)}{1 + \lambda(1 - \alpha)}
\]

where \(\beta_j^* = \frac{1}{N} \sum_{i=1}^{N} x_{ij}r_{ij}\) as before.
Leukemia Data, Logistic, N=72, p=3571, first 10 steps shown
Sparser than lasso — $\ell_0$ and best-subset

Mazumder, Friedman and H. (JASA, 2011).

**SPARSENET**: extends elastic-net family into concave domain. Bridges $\ell_1$ and $\ell_0$ via MC+ penalty (Zhang 2010) and coordinate descent.

See also **PICASSO** package.
Mazumder, Friedman and H. (JASA, 2011).

**SPARSENET**: extends elastic-net family into concave domain. Bridges $\ell_1$ and $\ell_0$ via MC+ penalty (Zhang 2010) and coordinate descent.

See also **PICASSO** package.

Mazumder (with Bertsimas and King, AoS 2016) crack forty year old best-subset selection bottleneck! They use *mixed-integer programming (MIO)* along with the Gurobi solver.
Best-subset, forward-stepwise or lasso?

Discussion and comparisons via simulations in H, Tibshirani and Tibshirani SS 2019
Variety of settings varying SNR, # variables, correlations, and model structure.

Conclusions:

• forward stepwise very close to best subset, but much faster.
• In wide-data settings, and low SNR, lasso can beat best subset and forward stepwise.
• relaxed lasso overall winner (except in very high SNR regime, where best-subset wins), and fastest by far.
Lasso versus Subset Regression: $\ell_1$ vs $\ell_0$

$n=70, p=30, s=5, \text{SNR}=0.71, \text{PVE}=0.42$
Effective degrees of freedom $df$

$n=70$, $p=30$, $s=5$, SNR=0.71, PVE=0.42

\[
df = \frac{1}{\sigma^2} \sum_{i=1}^{n} \text{Cov}(y_i, \hat{y}_i)
\]
Relaxed Lasso

Originally proposed by Meinshausen (2006). We present a simplified version.

- Suppose $\hat{\beta}_\lambda$ is the lasso solution at $\lambda$, and let $A_\lambda$ be the active set of indices with nonzero coefficients in $\hat{\beta}_\lambda$.
- Let $\hat{\beta}^{LS}_{A_\lambda}$ be the coefficients in the least squares fit, using only the variables in $A_\lambda$. Let $\hat{\beta}^LS_\lambda$ be the full-sized version of this coefficient vector, padded with zeros.
- $\hat{\beta}^{LS}_\lambda$ debiases the lasso, while maintaining its sparsity.
- Define the Relaxed Lasso

$$\hat{\beta}^{RELAX}_\lambda (\gamma) = \gamma \cdot \hat{\beta}_\lambda + (1 - \gamma) \cdot \hat{\beta}^{LS}_\lambda$$
Relaxed Lasso

Originally proposed by Meinshausen (2006). We present a simplified version.

• Suppose \( \hat{\beta}_\lambda \) is the lasso solution at \( \lambda \), and let \( A_\lambda \) be the active set of indices with nonzero coefficients in \( \hat{\beta}_\lambda \).

• Let \( \hat{\beta}^{LS}_{\lambda A} \) be the coefficients in the least squares fit, using only the variables in \( A_\lambda \). Let \( \hat{\beta}^{LS}_\lambda \) be the full-sized version of this coefficient vector, padded with zeros.

\( \hat{\beta}^{LS}_\lambda \) debiases the lasso, while maintaining its sparsity.

• Define the Relaxing Lasso

\[
\hat{\beta}_\lambda^{RELAX}(\gamma) = \gamma \cdot \hat{\beta}_\lambda + (1 - \gamma) \cdot \hat{\beta}^{LS}_\lambda
\]

Once \( \hat{\beta}^{LS}_\lambda \) is computed at desired values of \( \lambda \), the whole family \( \hat{\beta}_\lambda^{RELAX}(\gamma) \) comes free of charge!
Relaxed lasso now in **GLMNET** package
Relaxed lasso now in \textit{GLMNET} package (almost!)
Screening Rules

Logistic regression for small GWAS: \( N = 2000, p \sim \) million, (Wu et al, 2009)

- Compute \(|\langle x_j, y - \bar{y} \rangle|\) for each Snp \( j = 1, 2, \ldots, 10^6 \), where \( \bar{y} \) is the mean of (binary) \( y \).

Note: the largest of these is \( \lambda_{max} \): smallest value of \( \lambda \) for which all coefficients are zero.
Screening Rules

Logistic regression for small GWAS: \( N = 2000, \ p \sim \text{million}, \) (Wu et al, 2009)

- Compute \( \left| \langle x_j, y - \bar{y} \rangle \right| \) for each Snp \( j = 1, 2, \ldots, 10^6 \), where \( \bar{y} \) is the mean of (binary) \( y \).

  *Note: the largest of these is \( \lambda_{\max} \): smallest value of \( \lambda \) for which all coefficients are zero.*

- Fit lasso logistic regression path using only largest \( M \) of these (e.g. 1000; for small GWAS models with 20–30 terms common).

- Simple KKT confirmations check that omitted SnpS would not have entered the model.
Strong Rules

For lasso fit, with *active set* $A$:

$$\left| \langle \mathbf{x}_j, \mathbf{y} - \mathbf{X}\hat{\beta}(\lambda) \rangle \right| = \lambda \quad \forall j \in A$$

$$\leq \lambda \quad \forall j \notin A$$

So variables *nearly in* $A$ will have inner-products with the residuals *nearly* equal to $\lambda$. 

*Tibshirani, Bien, Friedman, H, Simon, Taylor, Tibshirani (JRSSB 2012)*
Strong Rules

For lasso fit, with *active set* $\mathcal{A}$:

$$\left| \langle \mathbf{x}_j, \mathbf{y} - \mathbf{X}\hat{\beta}(\lambda) \rangle \right| = \lambda \quad \forall j \in \mathcal{A}$$

$$\leq \lambda \quad \forall j \notin \mathcal{A}$$

So variables *nearly in* $\mathcal{A}$ will have inner-products with the residuals *nearly* equal to $\lambda$.

Suppose fit at $\lambda_\ell$ is $\mathbf{X}\hat{\beta}(\lambda_\ell)$, and we want to compute the fit at $\lambda_{\ell+1} < \lambda_\ell$.

Strong rules gamble on set

$$\mathcal{S}_{\ell+1} = \left\{ j : \left| \langle \mathbf{x}_j, \mathbf{y} - \mathbf{X}\hat{\beta}(\lambda_\ell) \rangle \right| > \lambda_{\ell+1} - (\lambda_\ell - \lambda_{\ell+1}) \right\}$$

*GLMNET* screens at every $\lambda$ step, and after convergence, checks if any violations. Mostly $\mathcal{A}_{\ell+1} \subseteq \mathcal{S}_{\ell+1}$.

* Tibshirani, Bien, Friedman, H, Simon, Taylor, Tibshirani (JRSSB 2012)
Strong rules inspired by El Ghaoui, Viallon and Rabbani (2010)
Sequential DPP due to Wang, Lin, Gong, Wonka and Ye (2013)
Example: UK Biobank

- Work with PhD students Junyang Qian, Yosuke Tanigawa, Rob Tibshirani and the Manny Rivas group at Stanford DBDS
- Large cohort of 500K British adults (Bycroft et al 2018)
- Each individual genotyped at 805K locations (AA, Aa, aa or NA)
- 100s of phenotypes measured on each subject
- We looked at white British subset of 337K, and illustrate with height, asthma, and MI phenotypes.
- Divided the data 60% training, 20% validation, 20% test.

Package SNPNET available on Github (link on my website, and to 2019 report)
R-squared or heritability shown for regularization paths
Detailed results

4.4 Results

We present results of the lasso and related methods for quantitative traits including standing height and BMI, and for qualitative traits including asthma and high cholesterol. A comparison of the univariate p-values and the lasso coefficients for all these traits is showed in the form of Manhattan plots in the Appendix A (Supplementary Figure 13, 14).

4.4.1 Quantitative Traits

Standing Height

Height is a polygenic and heritable trait that has been studied for a long time. It has been used as a model for other quantitative traits, since it is easy to measure reliably. From twin and sibling studies, the narrow sense heritability is estimated to be 70-80% (Silventoinen et al., 2003; Visscher et al., 2006, 2010). Recent estimates controlling for shared environmental factors present in twin studies calculate heritability at 0.69 (Zaitlen et al., 2013; Hemani et al., 2013). A linear based model with common SNPs explains 45% of the variance (Yang et al., 2010) and a model including imputed variants explains 56% of the variance, almost matching the estimated heritability (Yang et al., 2015). So far, GWAS studies have discovered 697 associated variants that explain one fifth of the heritability (Lango Allen et al., 2010; Wood et al., 2014). Recently, a large sample study was able to identify more variants with low frequencies that are associated with height (Marouli et al., 2017). Using lasso with the larger UK Biobank dataset allows both a better estimate of the proportion of variance that can be explained by genomic predictors and simultaneous selection of SNPs that may be associated. We obtain $R^2$ values that are close to the estimated heritability. The results are summarized in Table 3. The associated $R^2$ curves for the lasso and the relaxed lasso are shown in Figure 3. The residuals of the optimal lasso prediction are plotted in Figure 4.

<table>
<thead>
<tr>
<th>Model</th>
<th>Form</th>
<th>$R^2_{train}$</th>
<th>$R^2_{val}$</th>
<th>$R^2_{test}$</th>
<th>Size</th>
</tr>
</thead>
<tbody>
<tr>
<td>(1)</td>
<td>Age + Sex</td>
<td>0.5300</td>
<td>0.5260</td>
<td>0.5288</td>
<td>2</td>
</tr>
<tr>
<td>(2)</td>
<td>Age + Sex + 10 PCs</td>
<td>0.5344</td>
<td>0.5304</td>
<td>0.5336</td>
<td>12</td>
</tr>
<tr>
<td>(3)</td>
<td>Strong Single SNP</td>
<td>0.5364</td>
<td>0.5323</td>
<td>0.5355</td>
<td>13</td>
</tr>
<tr>
<td>(4)</td>
<td>10K Combined</td>
<td>0.5482</td>
<td>0.5408</td>
<td>0.5444</td>
<td>10,012</td>
</tr>
<tr>
<td>(5)</td>
<td>100K Combined</td>
<td>0.5833</td>
<td>0.5515</td>
<td>0.5551</td>
<td>100,012</td>
</tr>
<tr>
<td>(6)</td>
<td>Sequential LR</td>
<td>0.7416</td>
<td>0.6596</td>
<td>0.6601</td>
<td>17,012</td>
</tr>
<tr>
<td>(7)</td>
<td>Lasso</td>
<td>0.8304</td>
<td>0.6992</td>
<td><strong>0.6999</strong></td>
<td>47,673</td>
</tr>
<tr>
<td>(8)</td>
<td>Relaxed Lasso</td>
<td>0.7789</td>
<td>0.6718</td>
<td>0.6727</td>
<td>13,395</td>
</tr>
</tbody>
</table>
Algorithmic details in pictures

Figure 1: The lasso coefficient profile that shows the progression of the BASIL algorithm. The previously finished part of the path is colored grey, the newly completed and verified is in green, and the part that is newly computed but failed the verification is colored red.

1.4 Outline of the paper

The rest of the paper is organized as follows.

- Section 2 describes the proposed batch screening iterative lasso (BASIL) algorithm for the Gaussian family in detail and its extension to other problems such as logistic regression.
- Section 3 discusses related methods and packages for solving large-scale lasso problems.
UK Biobank — Asthma

Odds ratio plot - Polygenic Risk Score percentiles

Asthma PRS Percentile

Number of individuals

Asthma odds ratio

Cases

Controls
UK Biobank — Asthma

Note: only 6347 variants in active path for Lasso are included in the coefficient plot.
## UK Biobank — Asthma

<table>
<thead>
<tr>
<th>Variable</th>
<th>Level</th>
<th>N</th>
<th>Events</th>
<th>Hazard Ratio</th>
<th>P value</th>
</tr>
</thead>
<tbody>
<tr>
<td>Polygenic risk score</td>
<td>40-60%</td>
<td>62139</td>
<td>5017</td>
<td></td>
<td>Reference</td>
</tr>
<tr>
<td></td>
<td>Bottom 10%</td>
<td>31036</td>
<td>896</td>
<td>0.36 (0.34, 0.39)</td>
<td>&lt;0.001</td>
</tr>
<tr>
<td></td>
<td>Top 1%</td>
<td>3085</td>
<td>1494</td>
<td>7.31 (6.90, 7.74)</td>
<td>&lt;0.001</td>
</tr>
<tr>
<td></td>
<td>Top 10%</td>
<td>15480</td>
<td>3636</td>
<td>3.06 (2.93, 3.20)</td>
<td>&lt;0.001</td>
</tr>
<tr>
<td></td>
<td>Top 5%</td>
<td>12403</td>
<td>3894</td>
<td>4.27 (4.09, 4.45)</td>
<td>&lt;0.001</td>
</tr>
<tr>
<td>Eosinophil threshold</td>
<td>below .18</td>
<td>73655</td>
<td>6392</td>
<td>1.61 (1.56, 1.67)</td>
<td>&lt;0.001</td>
</tr>
<tr>
<td></td>
<td>above .18</td>
<td>50488</td>
<td>8545</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Sex</td>
<td>Female</td>
<td>66292</td>
<td>8330</td>
<td></td>
<td>Reference</td>
</tr>
<tr>
<td></td>
<td>Male</td>
<td>57851</td>
<td>6607</td>
<td>0.87 (0.85, 0.90)</td>
<td>&lt;0.001</td>
</tr>
<tr>
<td>Body Mass Index</td>
<td>Below 25</td>
<td>40969</td>
<td>4414</td>
<td></td>
<td>Reference</td>
</tr>
<tr>
<td></td>
<td>25-30</td>
<td>53273</td>
<td>6220</td>
<td>1.04 (1.00, 1.08)</td>
<td>0.06</td>
</tr>
<tr>
<td></td>
<td>30-35</td>
<td>21600</td>
<td>2880</td>
<td>1.12 (1.07, 1.17)</td>
<td>&lt;0.001</td>
</tr>
<tr>
<td></td>
<td>35-40</td>
<td>6010</td>
<td>954</td>
<td>1.26 (1.17, 1.35)</td>
<td>&lt;0.001</td>
</tr>
<tr>
<td></td>
<td>Above 40</td>
<td>2291</td>
<td>469</td>
<td>1.59 (1.44, 1.75)</td>
<td>&lt;0.001</td>
</tr>
</tbody>
</table>
Note: only 1197 variants in active path for Lasso are included in the coefficient plot.
UK Biobank — Myocardial Infarct
### UK Biobank — Myocardial Infarct

<table>
<thead>
<tr>
<th>Variable</th>
<th>Level</th>
<th>N</th>
<th>Events</th>
<th>Hazard Ratio</th>
<th>P value</th>
</tr>
</thead>
<tbody>
<tr>
<td>Polygenic risk score</td>
<td>40-60%</td>
<td>63742</td>
<td>1228</td>
<td>Reference</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Bottom 10%</td>
<td>31918</td>
<td>280</td>
<td>0.48 (0.42, 0.54)</td>
<td>&lt;0.001</td>
</tr>
<tr>
<td></td>
<td>Top 1%</td>
<td>3187</td>
<td>349</td>
<td>4.57 (4.06, 5.15)</td>
<td>&lt;0.001</td>
</tr>
<tr>
<td></td>
<td>Top 10%</td>
<td>15948</td>
<td>771</td>
<td>2.24 (2.05, 2.45)</td>
<td>&lt;0.001</td>
</tr>
<tr>
<td></td>
<td>Top 5%</td>
<td>12738</td>
<td>871</td>
<td>3.05 (2.79, 3.33)</td>
<td>&lt;0.001</td>
</tr>
<tr>
<td>Smoke</td>
<td>Not current smoker</td>
<td>114637</td>
<td>3010</td>
<td>Reference</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Occasionally</td>
<td>3275</td>
<td>106</td>
<td>1.19 (0.98, 1.44)</td>
<td>0.081</td>
</tr>
<tr>
<td></td>
<td>Yes, most days</td>
<td>9621</td>
<td>383</td>
<td>1.35 (1.21, 1.50)</td>
<td>&lt;0.001</td>
</tr>
<tr>
<td>Sex</td>
<td>female</td>
<td>67841</td>
<td>644</td>
<td>Reference</td>
<td></td>
</tr>
<tr>
<td></td>
<td>male</td>
<td>59692</td>
<td>2855</td>
<td>5.12 (4.69, 5.58)</td>
<td>&lt;0.001</td>
</tr>
<tr>
<td>Body Mass Index</td>
<td>below 25</td>
<td>42311</td>
<td>604</td>
<td>Reference</td>
<td></td>
</tr>
<tr>
<td></td>
<td>25-30</td>
<td>54416</td>
<td>1560</td>
<td>1.13 (1.02, 1.24)</td>
<td>0.015</td>
</tr>
<tr>
<td></td>
<td>30-35</td>
<td>22228</td>
<td>945</td>
<td>1.20 (1.08, 1.34)</td>
<td>0.001</td>
</tr>
<tr>
<td></td>
<td>35-40</td>
<td>6195</td>
<td>275</td>
<td>1.22 (1.05, 1.42)</td>
<td>0.008</td>
</tr>
<tr>
<td></td>
<td>above 40</td>
<td>2383</td>
<td>115</td>
<td>1.11 (0.91, 1.37)</td>
<td>0.306</td>
</tr>
<tr>
<td>non-LDL cholesterol</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>127533</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Triglycerides</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>127533</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>LDL-c</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>127533</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Diabetes</td>
<td>non-diabetic</td>
<td>121750</td>
<td>2855</td>
<td>Reference</td>
<td></td>
</tr>
<tr>
<td></td>
<td>diabetic</td>
<td>5783</td>
<td>644</td>
<td>1.19 (1.08, 1.30)</td>
<td>&lt;0.001</td>
</tr>
</tbody>
</table>
Coordinate Descent in General

Many problems have the form

\[
\min_{\{\beta_j\}_1^p} \left[ R(y, \beta) + \lambda \sum_{j=1}^p P_j(\beta_j) \right].
\]

- If \( R \) and \( P_j \) are convex, and \( R \) is differentiable, then coordinate descent converges to the solution (Tseng, 1988).
- Often each coordinate step is trivial. E.g. for lasso, it amounts to soft-thresholding, with many steps leaving \( \hat{\beta}_j = 0 \).
- Each term \( P_j(\beta_j) \) can refer to \textit{sets} of parameters.
- Decreasing \( \lambda \) slowly means not much cycling is needed.
- Coordinate moves can exploit sparsity.
Group Lasso


$$R(y, \sum_{j=1}^{J} X_j \beta_j) + \lambda \sum_{j=1}^{J} \gamma_j \|\beta_j\|_2.$$ 

*Example: each block $X_j$ represents the levels for a categorical predictor.*

- entire groups are zero, or all elements are nonzero.
- $\gamma_j$ is penalty modifier for group $j$; $\gamma_j = \|X_j\|_F$ is good choice.
- Leads to a block-updating form of coordinate descent.
- Strong rules apply here: $\|X_j^T r\|_2 > \gamma_j [\lambda_{\ell+1} - (\lambda_{\ell} - \lambda_{\ell+1})]$
Sparse interaction models

Glinternet

With past PhD student Michael Lim (JCGS 2014 and in Chap 4 of SLS book)
Hierarchical linear + first-order interaction models selected using group lasso
Sparse interaction models

**GLINTERNET**

With past PhD student Michael Lim (JCGS 2014 and in Chap 4 of SLS book)
Hierarchical linear + first-order interaction models selected using group lasso

**Mixed Graphical Models**

Allows for both continuous and categorical variables. With past PhD student Jason Lee (JCGS 2015) (*in Wald III*)
General Markov random field representation, with edge and node potentials selected by group lasso
Sparse Generalized Additive Models

\[ \eta(x) = \alpha_0 + \sum_{j=1}^{p} f_j(x_j) \]

GAMSEL with Alexandra Chouldechova (Arxiv 2015) using overlap group lasso
Automatic, sticky selection between zero, linear, or nonlinear terms in GAMs.
Sparse Generalized Additive Models

\[ \eta(x) = \alpha_0 + \sum_{j=1}^{p} f_j(x_j) \]

**GAMSEL** with Alexandra Chouldechova (Arxiv 2015) using *overlap group lasso*.
Automatic, *sticky* selection between zero, linear, or nonlinear terms in GAMs.

**Saturated Splines** with Nick Boyd and S. Boyd, Recht and Jordan (JMLR 2018).
Adaptive GAMs using saturated piecewise-linear splines, with automatic feature and knot selection.
Matrix Completion

**SOFTIMPUTE** with past PhD students Rahul Mazumder, Jason Lee, and with Rob Tibshirani and Reza Zadeh (JMLR 2010, 2015)

\[
\text{minimize}_{Z} \sum_{\text{Observed}(i,j)} (X_{ij} - Z_{ij})^2 + \lambda \| Z \|_*
\]

where *nuclear norm* \( \| Z \|_* \) is the sum of singular values of \( Z \).

Wald II
Undirected Graphical Models — learning dependence structure via the lasso. Model the inverse covariance $\Theta$ in the Gaussian family with $L_1$ penalties applied to elements.

$$\max_{\Theta} \log \det \Theta - \text{Tr}(S\Theta) - \lambda \|\Theta\|_1$$

**GLASSO**: modified block-wise lasso algorithm, which we solve by coordinate descent (Friedman, H, Tibshirani 2007). Algorithm is very fast, and solves moderately sparse graphs with 1000 nodes in under a minute.

Topic of *Wald III*, as well as a recent application of graphical models and factor models in anomaly detection.
Theory for lasso: see chapter 11, written by our coauthor Martin Wainwright.
Theory

Theory for lasso: see chapter 11, written by our coauthor Martin Wainwright.

Also new 2019 book by Martin Wainwright (Cambridge) *High-Dimensional Statistics — a non-asymptotic viewpoint*
Statistical Inference

- Can become Bayesian! Lasso penalty corresponds to Laplacian prior. However, need priors for everything, including $\lambda$ (variance ratio). *Also posterior not sparse!*

- Bootstrap. Easier to do, with similar results.

- Post-selection inference since circa 2013...

- A Significance Test for the Lasso — Richard Lockhart, Jonathan Taylor, Ryan Tibshirani and Rob Tibshirani (AoS, 2014)

- Exact Post-Selection Inference, with application to the Lasso — Jason Lee, Dennis Sun, Yuekai Sun, Jonathan Taylor (AoS, 2016)

- Optimal Inference after Model Selection — Will Fithian, Dennis Sun, Jonathan Taylor (2015, arXiv)
Statistical Inference

- Can become Bayesian! Lasso penalty corresponds to Laplacian prior. However, need priors for everything, including $\lambda$ (variance ratio). *Also posterior not sparse!*
- Bootstrap. Easier to do, with similar results.
Statistical Inference

- Can become Bayesian! Lasso penalty corresponds to Laplacian prior. However, need priors for everything, including $\lambda$ (variance ratio). *Also posterior not sparse!*
- Bootstrap. Easier to do, with similar results.
- Post-selection inference since circa 2013 . . .
  - *Exact Post-Selection Inference, with application to the Lasso* — Jason Lee, Dennis Sun, Yuekai Sun, Jonathan Taylor (AoS, 2016)
  - *Optimal Inference after Model Selection* — Will Fithian, Dennis Sun, Jonathan Taylor (2015, arXiv)
Statistical Inference continued

• *Selective inference with a randomized response* — Xioying Tian and Jonathan Taylor (AoS 2018)
• and many more papers with students of Taylor and Tibshirani, including Keli Liu, Jelena Markovic, Jeremy Taylor, Snighda Panigrahi, Nan Bi.
• **Selective inference with a randomized response** — Xioying Tian and Jonathan Taylor (AoS 2018)
• and many more papers with students of Taylor and Tibshirani, including Keli Liu, Jelena Markovic, Jeremy Taylor, Snighda Panigrahi, Nan Bi.

• **Knockoffs** — false discovery control for sequences of models such as those produced by lasso. Since circa 2015...
  Emmanuel Candes with collaborators Rina Foygel Barber, Chiara Sabatti, and students including Stephen Bates, Lucas Jansen, Eugene Katsevich, Matteo Sessio, Asaf Weinstein.
Summary

- Sparsity via convex relaxation both effective and computable.
- Although lasso 25 years old, still super cool!
- Good public software is essential.
- Coordinate descent simple and effective, especially when models are sparse.
- Screening rules allow for massive computational savings.
- Many extensions to lasso ideas — I present some in *Wald II* and *Wald III*
Summary

- Sparsity via convex relaxation both effective and computable.
- Although lasso 25 years old, still super cool!
- Good public software is essential.
- Coordinate descent simple and effective, especially when models are sparse.
- Screening rules allow for massive computational savings.
- Many extensions to lasso ideas — I present some in *Wald II* and *Wald III*

*One down, two to go — Thank you for your attention!*