Variable Selection in Regression Analysis using Ridge, LASSO, Elastic Net, and Best Subsets

Brenda Gillespie
University of Michigan
Consulting for Statistics, Computing and Analytics Research
Michigan SAS Users' Group (MSUG) Conference
June 7, 2018

Outline

- Why we need new methods
- Ordinary least squares (OLS) regression
- Beyond linear regression
- Ridge regression and its tuning parameter, k
- LASSO (least absolute shrinkage and selection operator)
- Elastic Net
- Best subsets
- When are these methods useful (vs OLS)?

Why we need new methods

- Genetic and other data with '**small n, large p**'
  - small **n** (#observations), large **p** (# parameters)
  - So many genes – thousands of them (**p**)
  - We want to know which genes predict disease, but we only have hundreds of patients per disease (**n**)
  - i.e., not feasible with OLS
- Interest in **PREDICTION** rather than ESTIMATION
  - We want to **predict** who will get disease
  - **Estimation** of gene effects is a secondary goal
- OLS has trouble with **highly collinear variables**
  - Estimates have large standard errors
  - wrong (counter-intuitive) signs problem

Linear Regression
(i.e., OLS =ordinary least squares)

- $Y = \beta_0 + \beta_1 * X + \text{error}$
- OLS finds estimates of β_0 and β_1 that minimize the MSE

- MSE = Mean squared error = $SSE/df_{(error)}$
  $=SSE/(n-p)$

- ASE = Average squared error = SSE/n

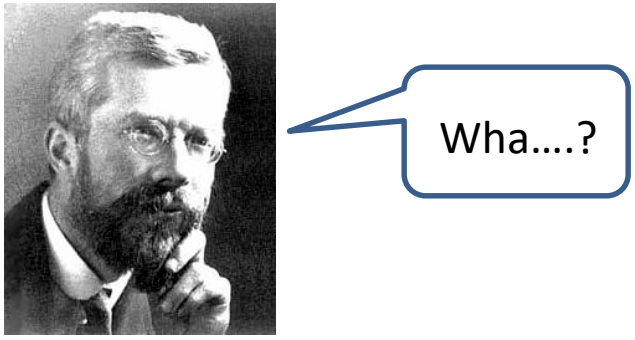
Beyond Linear Regression

SAS Software

- Proc REG
  - Ridge regression
- Proc GLMSelect
  - LASSO
  - Elastic Net
- Proc HPreg
  - High Performance for linear regression with variable selection (lots of options, including LAR, LASSO, adaptive LASSO)
  - Hybrid versions: Use LAR and LASSO to select the model, but then estimate the regression coefficients by ordinary weighted least squares. (suggested by Efron!)

What you know is no longer relevant

- **The following are OBSOLETE!**
- When testing interactions, main effects should also be included
- Rule of 10: You need 10 observations for each variable entered in the model
  - You should never have more variables than observations!
- In linear models, calculating p-values, and adding or deleting a variable is based on an F-test (F is for Fisher!)



Training, validation, and test datasets

- Training data: Data used to fit a model
- Validation data: Data used for calculating prediction error ‘on the fly’ while fitting training data
  - Used to select variables for inclusion, to minimize over-fitting: Enter a variable (Trn), check prediction error (Val), keep or exclude the variable.
- Test data: External data used after Training and Validation to confirm model results and prediction error estimates

Goals of regression …

- Prediction
- Detecting and Interpreting effects
  - Often an unstated goal, and may be sacrificed in the rush to prediction

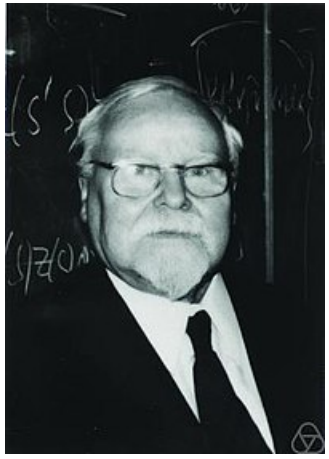
- Variable selection is an attempt to find a parsimonious model (setting some betas to zero)
  - Methods: AIC, BIC (SBC), Cp, Adjusted R^2, MSE, ASE, prediction error, cross-validation

- Regularization (shrinkage) is an attempt to shrink estimates so that those "too big" by chance are given less weight

Ridge Regression

- Developed to deal with collinearity
  - OLS: Beta estimates are unbiased, but have large standard errors
- Ridge estimates are **biased**, but have **smaller standard errors**
- A successful Ridge regression: the reduction in variance is greater than the squared bias
  - The bias/variance trade-off depends on the tuning parameter, k

Origins of Ridge Regression



Andrey Tikhonov, born in Russia (1906), developed **Tikhonov regularization**, a common way to regularize ill-posed problems.

- Known in statistics as **ridge regression**
- In machine learning as **weight decay**
- With multiple independent discoveries, it is also variously known as the **Tikhonov–Miller method**, the **Phillips–Twomey method**, the **constrained linear inversion** method, and the method of **linear regularization**.

Original way to think of Ridge Regression

- With standardized variables and little collinearity, the X'X matrix has more weight on the diagonal (squares) than off the diagonal (cross-products).
  - Beta estimation is stable.

$$X'X = \begin{pmatrix} \mathbf{1} & 0.1 & 0.03 \\ 0.1 & \mathbf{1} & 0 \\ 0.03 & 0 & \mathbf{1} \end{pmatrix}$$

- With collinearity, there is more weight off than on the diagonal ($\Rightarrow$large standard errors for beta estimates).
- By adding more weight to the diagonal, we can stabilize (shrink) the betas ($\Rightarrow$reduce variance, at a cost of bias)
- Add a small number, k (say, between 0 and 0.2) to the diagonal of X'X, and continue with OLS as usual.

$$X'X = \begin{pmatrix} 1 & .9 & 0 \\ .9 & 1 & 0 \\ 0 & 0 & 1 \end{pmatrix} \qquad X'X + kI = \begin{bmatrix} 1+k & .9 & 0 \\ .9 & 1+k & 0 \\ 0 & 0 & 1+k \end{bmatrix}$$

Newer way to think of Ridge Regression

- Estimate β by minimizing the sum of squared residuals, with a **penalty** for high sum of β_j^2
  - all X's centered and scaled
- i.e., minimize sum of squared residuals subject to the condition that $\sum_{j=1}^{p} \beta_j^2 \leq t$ (This yields a spherical region)
- Ridge regression (Lagrange form) minimizes:

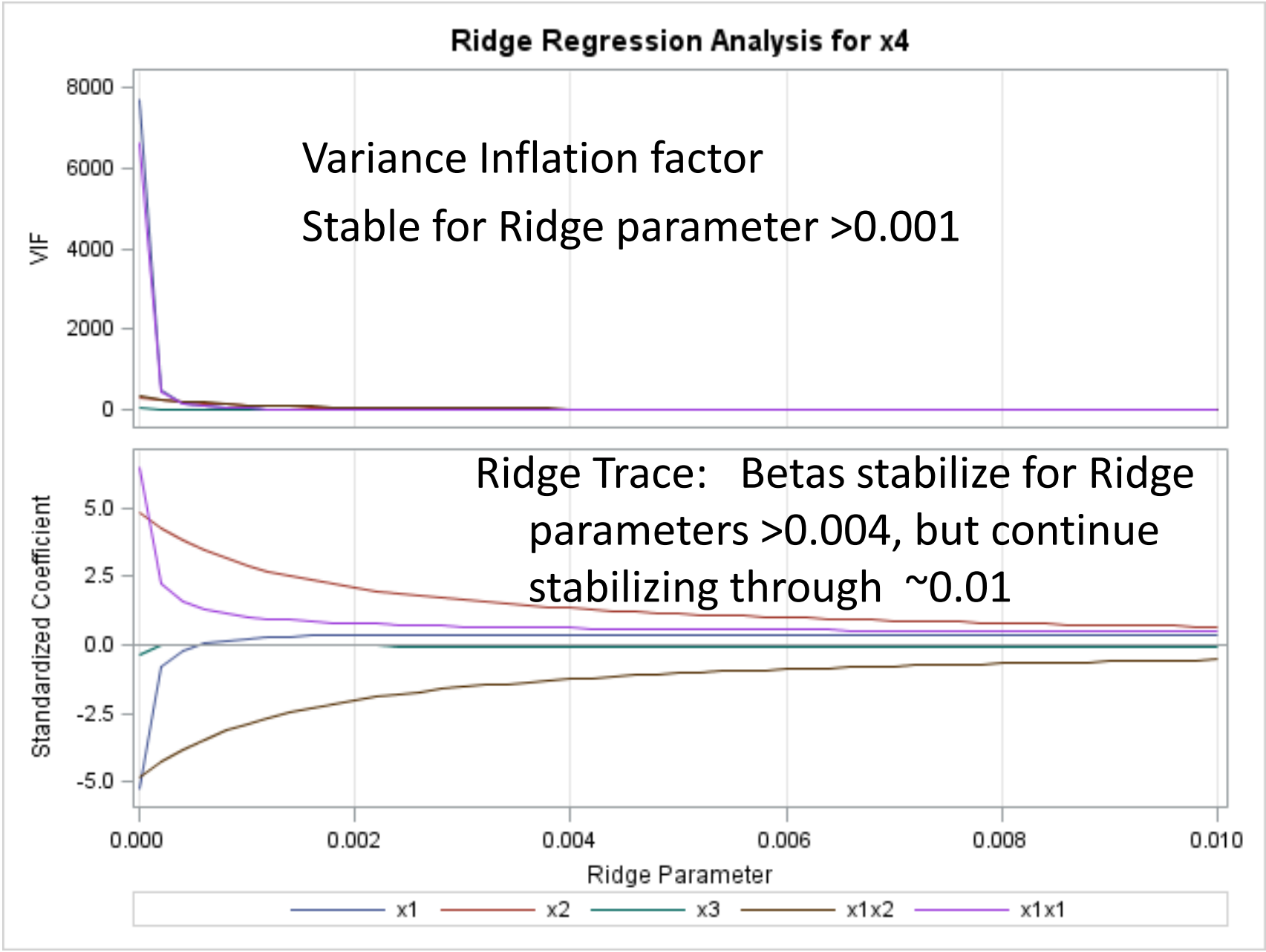
$$\arg\min_{\beta} \left\{ \underbrace{\sum_{i=1}^{n} (y_i - (X\beta)_i)^2}_{\text{sum of squared residuals}} + \underbrace{k \sum_{j=1}^{p} \beta_j^2}_{\text{penalty}} \right\}$$

(sum of squared residuals … penalty)

Note: k is the same value added to the X'X diagonal in the previous slide

Note: k and t are inversely scaled: k=0 $\Rightarrow$OLS, and t=∞ $\Rightarrow$ OLS

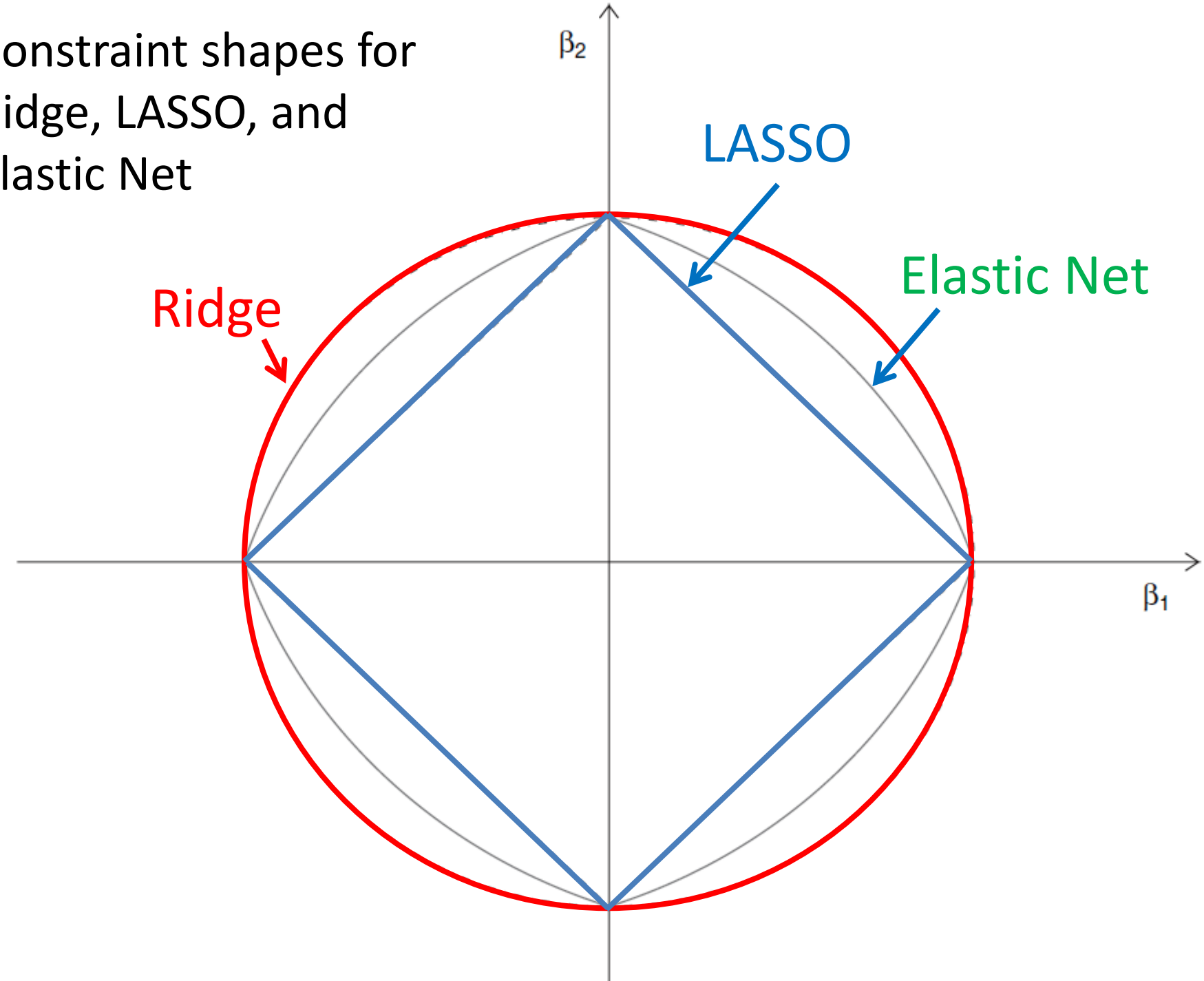
VIF, Ridge trace, to find stable point for all coefficients



Choosing the ‘best’ Ridge Parameter

- Choice of k depends on several factors
- In general, larger k $\Rightarrow$more shrinkage of betas (bias), but smaller variance
- There are many algorithms
  - Some optimize prediction, others emphasize interpretability
  - Many choices, but lots of useful guidance
- No variable deletion; the constraint space is circular

Constraint shapes for Ridge, LASSO, and Elastic Net



MODEL 1: x4 = Intercept + β_1*x1 + β_2*x2 + β_3*x3 + error

Coefficients below are on the raw scale (not standardized)

Obs	_DEPVAR	_RIDGE_	_RMSE_	Intercept	x1	x2	x3
1	x4	.	3.767	-121.3	0.12685	0.34816	-19.0217
2	x4	0.002	3.768	-117.6	0.12405	0.34722	-25.8455
3	x4	0.004	3.770	-114.2	0.12149	0.34634	-32.0568
4	x4	0.006	3.772	-111.1	0.11914	0.34552	-37.7326
5	x4	0.008	3.776	-108.3	0.11698	0.34474	-42.9378
6	x4	0.010	3.780	-105.7	0.11498	0.34401	-47.7270
7	x4	0.020	3.803	-95.0	0.10685	0.34084	-66.8525
8	x4	0.050	3.873	-76.5	0.09269	0.33404	-98.1368

Ridge Example SAS code

```
ods graphics on;
proc REG   data=acetyl outvif
        outest=b ridge=0 to 0.02 by .002;
   model x4 = x1 x2  x3 ;
run;
```

Ridge Summary

- Ridge regression is available in SAS Proc REG
- Ridge performs **regularization**, but **not variable selection**
  - **All variable coefficients are estimated (even those not predictive)**
  - **Not useful if parsimony is desired**
- Useful in the case of **strong multicollinearity**
  - Tends to give similar coefficients for collinear variables (does not drop the less predictive one)
  - Avoids the "wrong signs problem" of collinear variables having opposite effects that are hard to interpret.
- Estimates depend on the Ridge parameter chosen, which is a somewhat subjective decision
- Bayesian ties: Ridge is a Bayesian estimate of linear model with a Gaussian prior

LASSO!

A **lasso** is made from stiff rope

The noose stays open when the **lasso** is thrown so the steer can be roped.

The noose can be re-opened from horseback to release the steer.

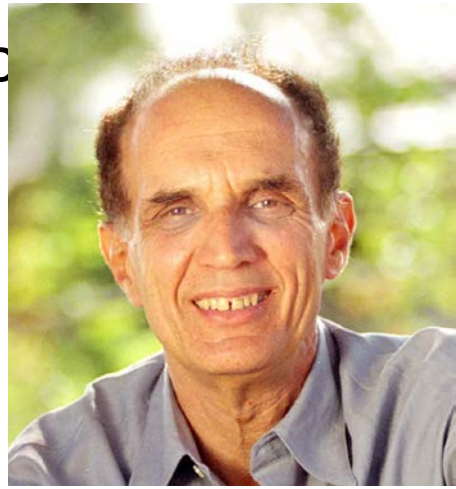
https://www.r-bloggers.com/ridge-regression-and-the-lasso/

LASSO (least absolute shrinkage and selection operator)

- Lasso variables you want, and let go (zero) those you don't
- LASSO performs both variable selection (zeroing out variables) and regularization (shrinkage) to enhance prediction accuracy
  - Variable selection $\Rightarrow$ parsimony!
- Introduced by **Robert Tibshirani** (1996)
- A different penalty than used in Ridge regression:
- Minimizes sum of squared residuals subject to

$$\sum_{j=1}^{p} |\beta_j| \leq t$$

- LAR (Least angle regression), related to LASSO, was developed by Brad Efron.
  - Easier to find best regularization parameter



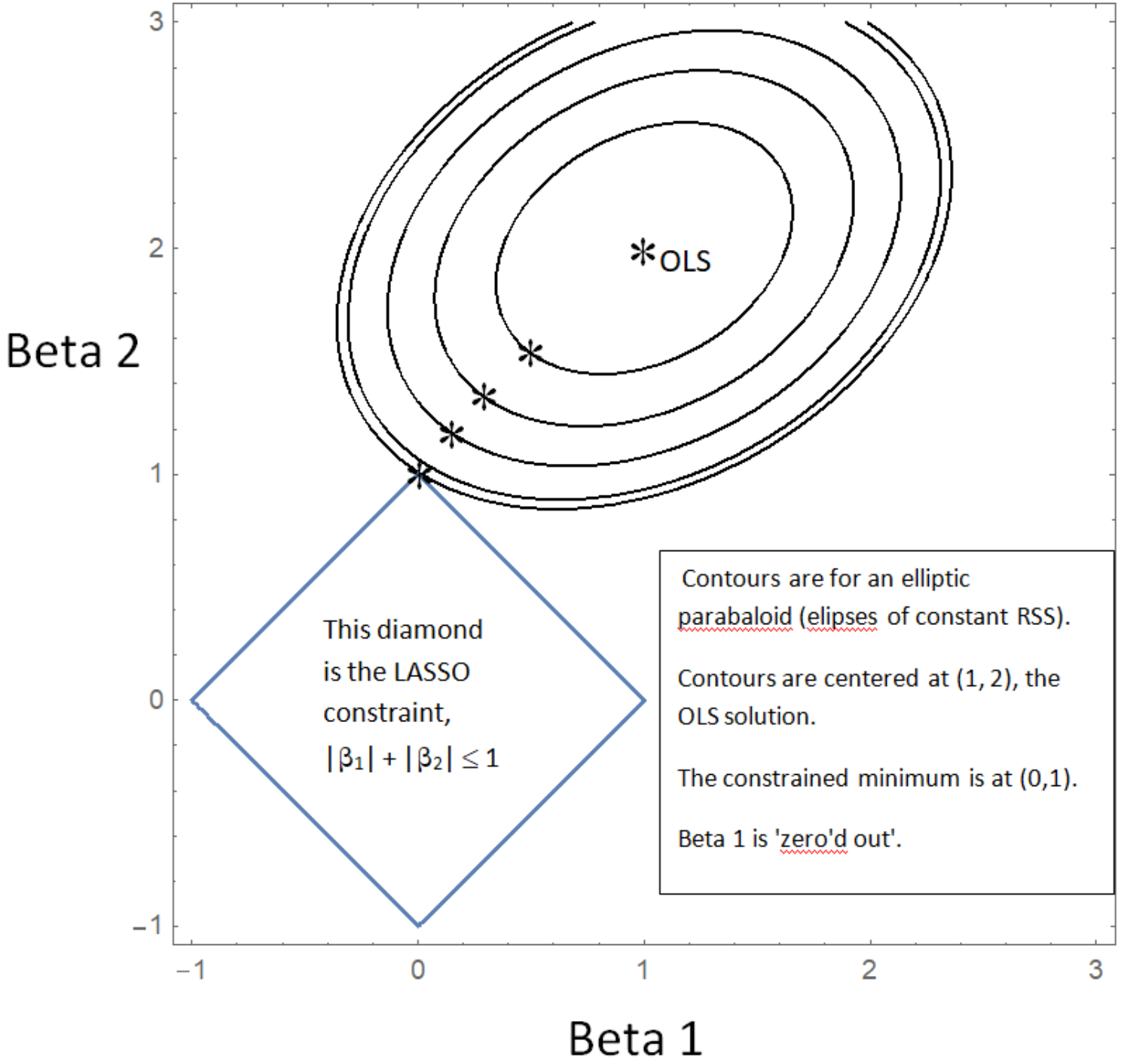
LASSO

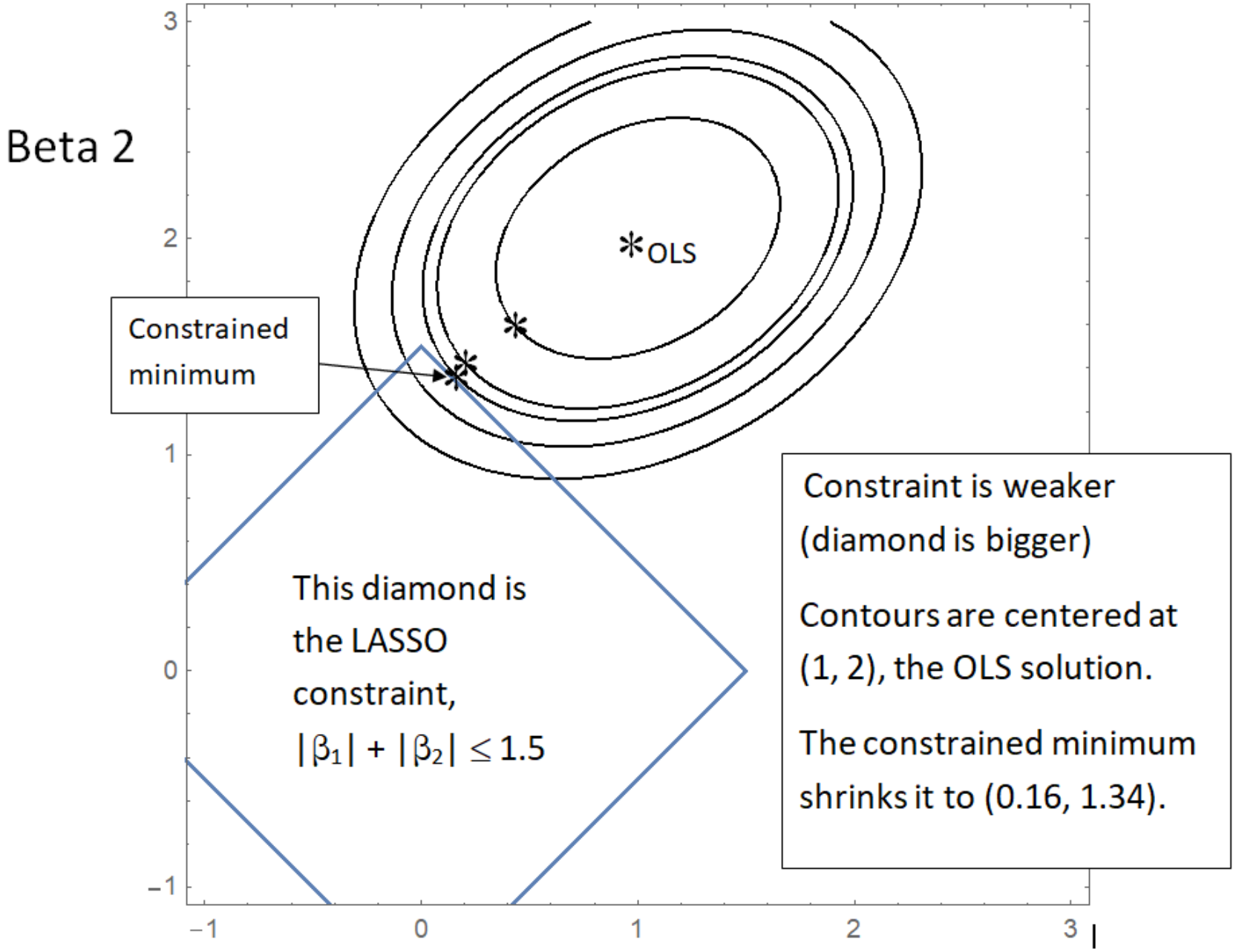
LASSO penalty parameter, k

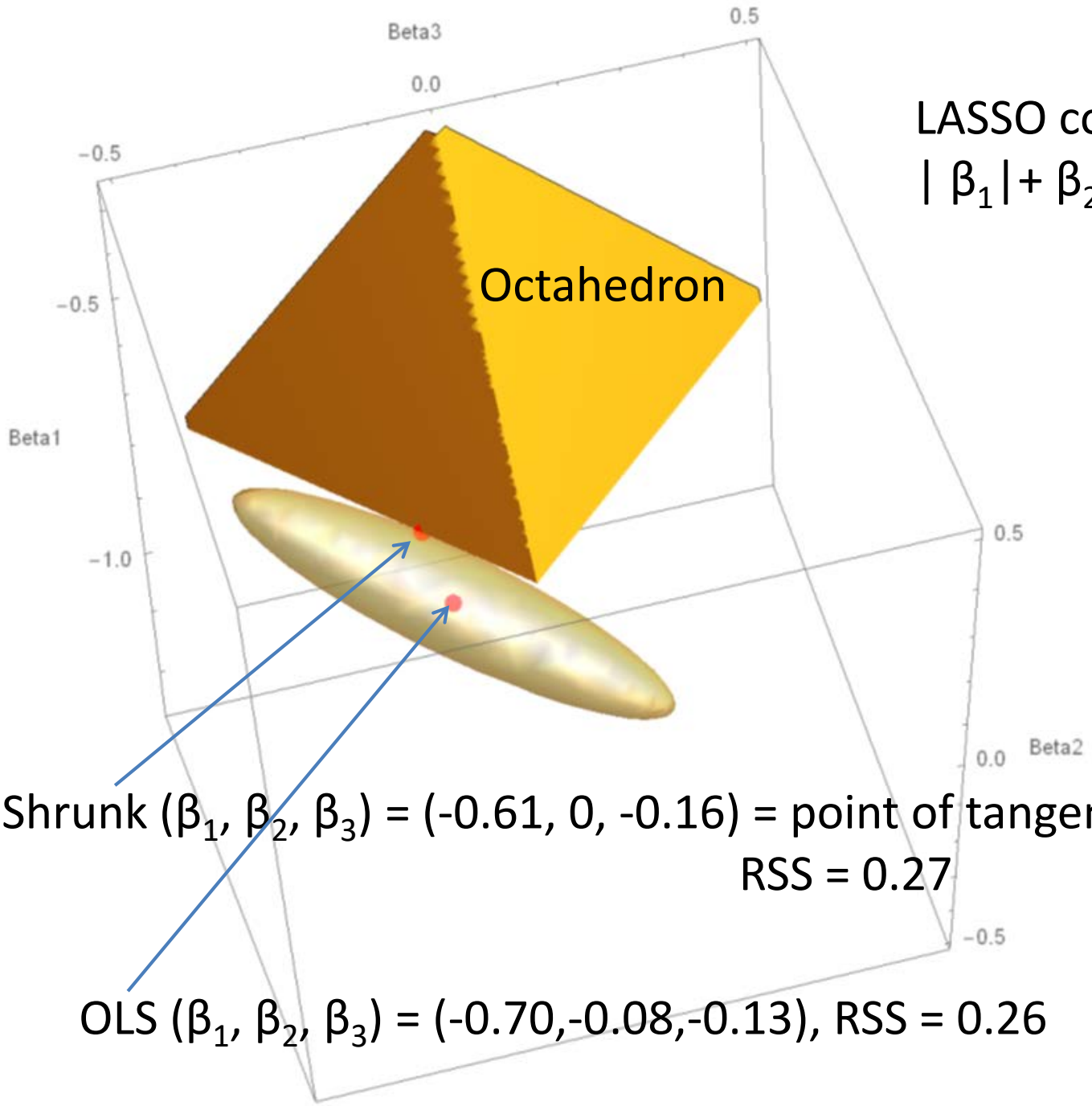
$$\arg\min_{\beta}\left\{\sum_{i=1}^{n}(y_i-(X\beta)_i)^2+k\sum_{j=1}^{p}|\beta_j|\right\}$$

- $k\rightarrow 0$ (and $t\rightarrow\infty$) $\Rightarrow$ OLS solution
- $k\rightarrow\infty$ (and $t\rightarrow 0$) $\Rightarrow$ betas=0 (complete shrinkage)

- Advantages of LASSO over Ridge
  - less biased for variables that 'really matter'
  - Allows p >> n (but will only include up to n variables)
  - Is good at getting rid of (zeroing) non-useful variables.
- Disadvantage (depending on how you look at it)
  - Given 3 collinear variables, LASSO will select one, and zero out the other two. Some people prefer to see the variables 'share' the effect.


Beta 2
Beta 1
3
2
1
0
−1
−1
0
1
2
3
OLS
This diamond is the LASSO constraint, $|\beta_1| + |\beta_2| \leq 1$
Contours are for an elliptic parabaloid (elipses of constant RSS).
Contours are centered at (1, 2), the OLS solution.
The constrained minimum is at (0,1).
Beta 1 is 'zero'd out'.


Beta 2
3
2
1
0
−1
−1
0
1
2
3
OLS
Constrained minimum
This diamond is the LASSO constraint, $|\beta_1| + |\beta_2| \leq 1.5$
Constraint is weaker (diamond is bigger)
Contours are centered at (1, 2), the OLS solution.
The constrained minimum shrinks it to (0.16, 1.34).
Beta 1



LASSO constraint:

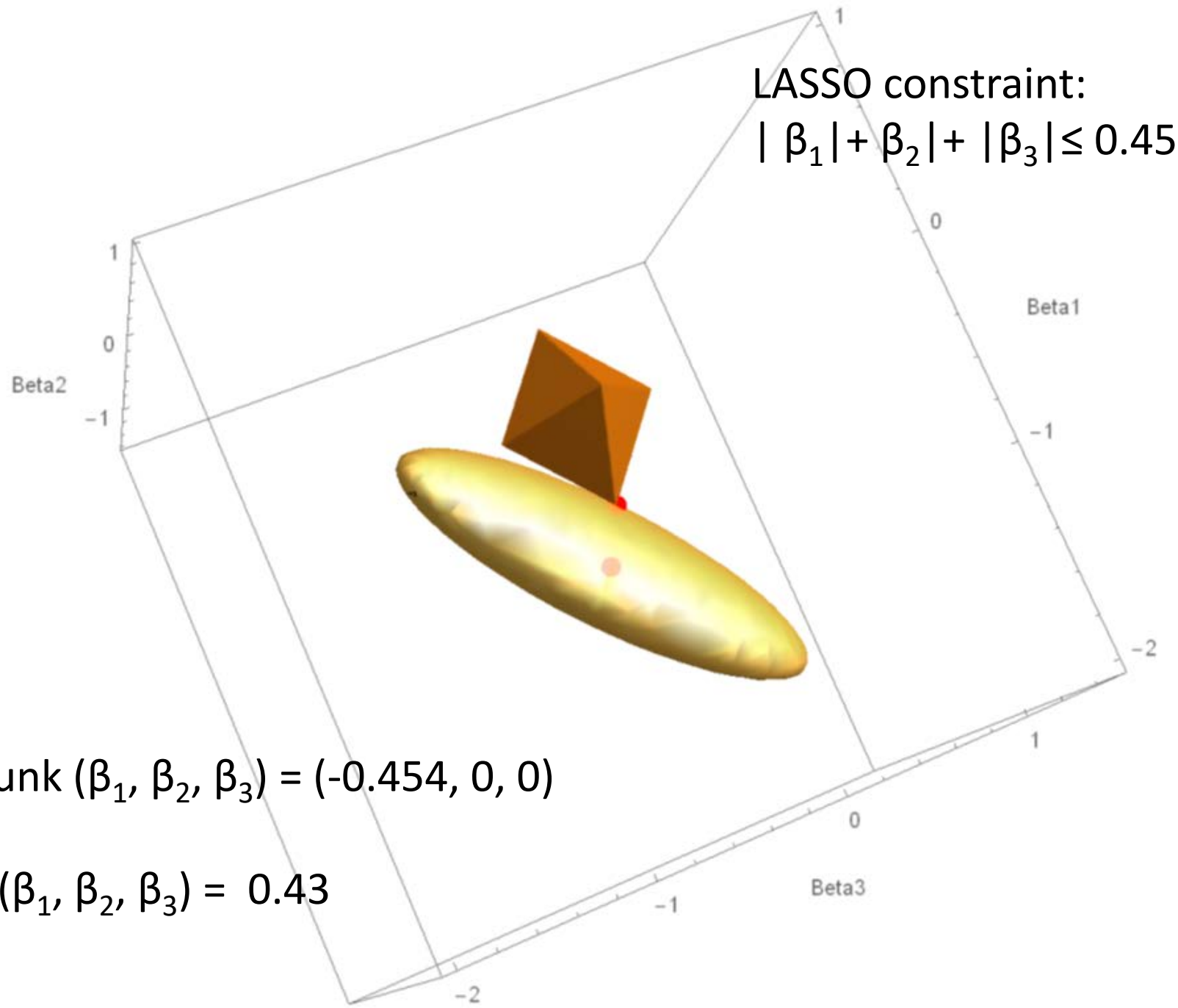
$|\beta_1| + \beta_2| + |\beta_3| \leq 0.77$

Shrunk (β_1, β_2, β_3) = (-0.61, 0, -0.16) = point of tangency to octahedron, RSS = 0.27

OLS (β_1, β_2, β_3) = (-0.70,-0.08,-0.13), RSS = 0.26

LASSO constraint:

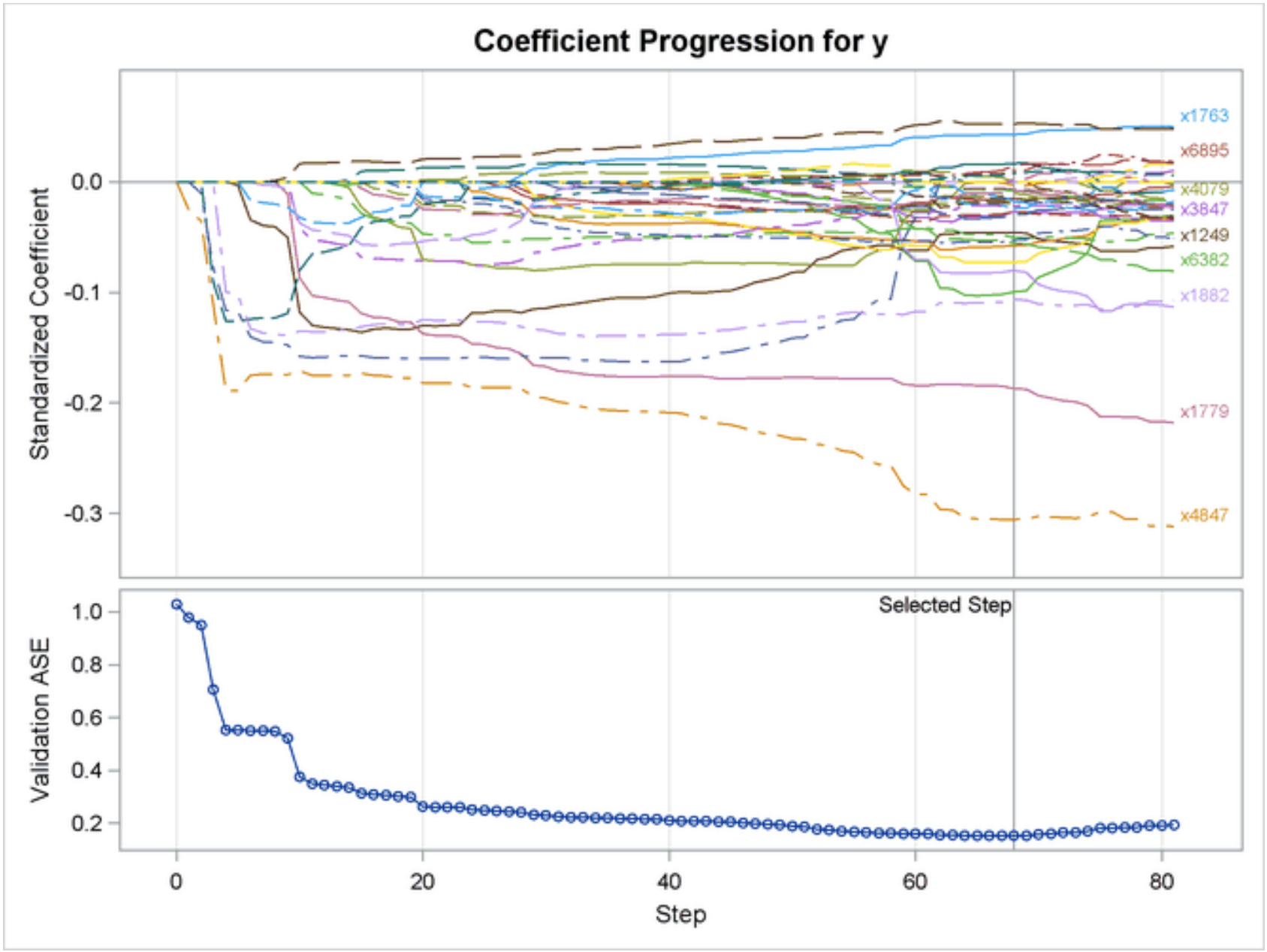
$|\beta_1| + \beta_2| + |\beta_3| \leq 0.45$



Shrunk $(\beta_1, \beta_2, \beta_3) = (-0.454, 0, 0)$

$RSS(\beta_1, \beta_2, \beta_3) = 0.43$

LASSO Coefficient Progression Plot



Step	Effect Entered	Effect Removed	Number Effects In	ASE	Validation ASE	
0	Intercept		1	0.8227	1.0287	
1	x4847		2	0.7884	0.9787	
2	x3320		3	0.7610	0.9507	
3	x2020		4	0.4935	0.7061	
4	x5039		5	0.2838	0.5527	
5	x1249		6	0.2790	0.5518	
6	x2242		7	0.2255	0.5513	
	LASSO Selection Summary		.	.	.	
			.	.	.	
			.	.	.	
62	x2534		35	0.0016	0.1554	
63		x3320	34	0.0016	0.1557	
64	x5631		35	0.0013	0.1532	
65		x6021	34	0.0012	0.1534	
66		x1745	33	0.0012	0.1535	
67	x6376		34	0.0012	0.1535	
68	x4831		35	0.0010	0.1531*	**Optimum**
69		x6184	34	0.0009	0.1539	
70	x3820		35	0.0006	0.1588	
71	x3171		36	0.0005	0.1615	

LASSO Summary

Advantages:

- Shrinkage
- Parsimony
- Allows p>n, although only allows up to n variables in a model

Elastic Net

- Combines the penalties of both Ridge and LASSO, but with the option of unequal weights

$$\arg\min_{\beta}\left\{\sum_{i=1}^{n}(y_i-(X\beta)_i)^2+\alpha\sum_{j=1}^{p}|\beta_j|+(1-\alpha)\sum_{j=1}^{p}\beta_j^2\right\}$$

- Elastic net is better than LASSO in the setting of <u>p>n</u>
  - Although LASSO can start with p>n variables, it will delete variables until p≤n for the final model.
  - Elastic net can have a final model with p>n

LASSO vs. Elastic Net

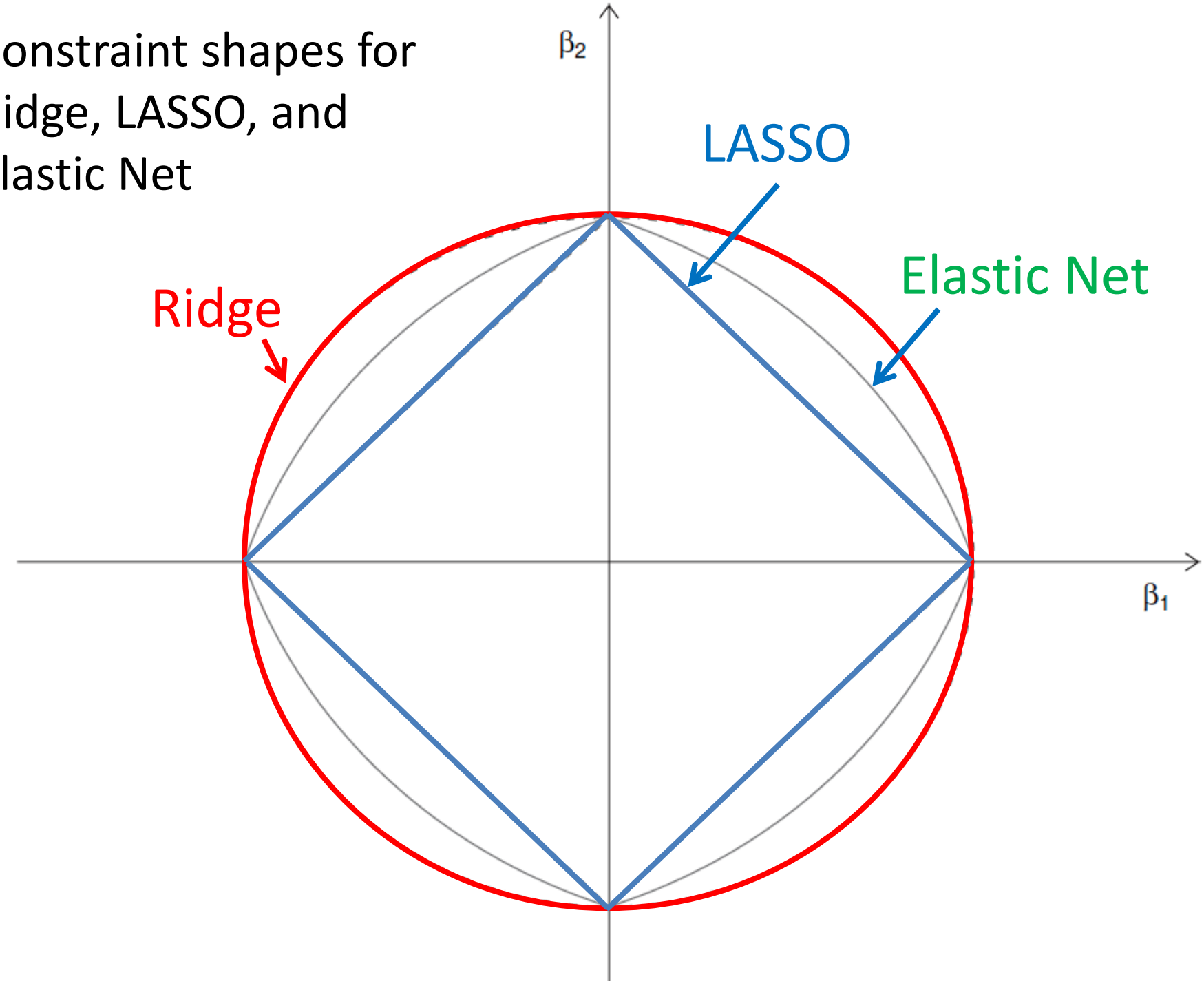
As stated by Zou and Hastie (2005), **the elastic net method can overcome the limitations of LASSO in the following three scenarios**:

- When **p > n, the LASSO method selects at most n variables**
  - The **elastic net method can select more than n variables** because of the ridge regression regularization.

- If there is a **group of variables** that have high pairwise correlations, then
  - LASSO tends to **select only one variable** from that group
  - Elastic net method **can select more than one variable**.

- In the **n > p** case, if there are **high correlations between predictors**, the prediction performance of LASSO is dominated by ridge regression. In this case, the **elastic net method can achieve better prediction performance by using ridge regression regularization**.
  - i.e., Carefully select the coefficient for $\sum_{j=1}^{p} \beta_j^2$

Constraint shapes for Ridge, LASSO, and Elastic Net

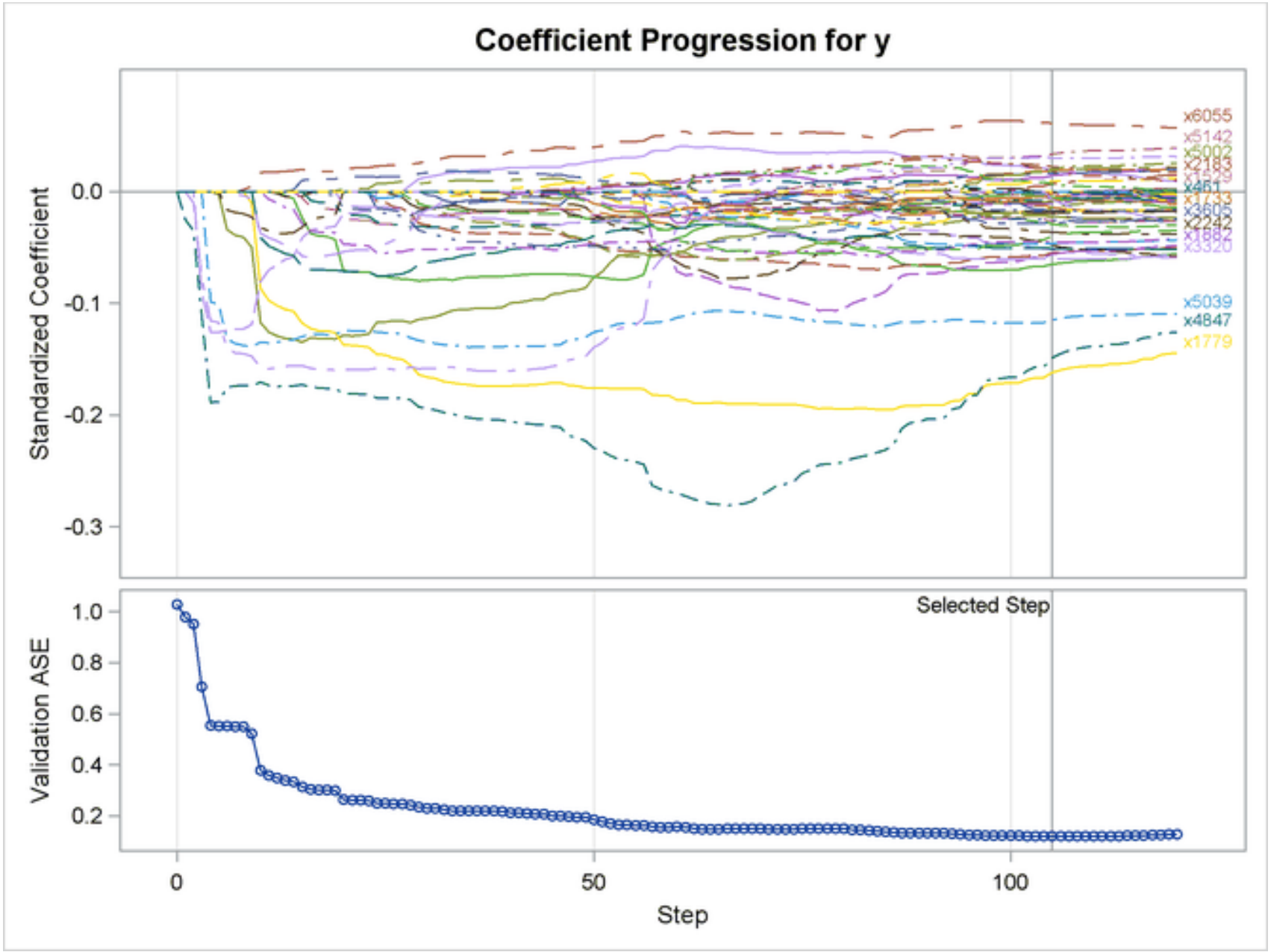


Number of variables chosen

- Because Elastic Net has a 'cushion' around the diamond in the previous slide, it is harder than LASSO to 'zero' variables
  - More variables are left in the model with Elastic Net, more are taken out by LASSO
  - SAS data example

	LASSO	Elastic Net
# Observations (Training sample)	38	38
# Variables	7129	7129
# included in model	35	53

Elastic Net Coefficient Progression Plot



LASSO and Elastic Net in SAS

These methods are useful for other regression types, too!

e.g., logistic regression, Cox regression

Proc GLMSelect

(Overview, from SAS)

- The GLMSELECT procedure performs effect selection in the framework of general linear models. A variety of model selection methods are available, including the LASSO method of Tibshirani (1996) and the related LAR method of Efron et al. (2004).

- The procedure offers **extensive capabilities for customizing the selection** with a wide variety of selection and **stopping criteria**, from traditional and computationally efficient **significance-level-based criteria** to more computationally intensive **validation-based criteria**. The procedure also provides **graphical summaries** of the selection search.

- The GLMSELECT procedure compares most closely to REG and GLM.
  - The REG procedure supports a variety of model-selection methods but **does not support a CLASS** statement.
  - The GLM procedure supports a CLASS statement but **does not include effect selection methods**.
  - The **GLMSELECT procedure fills this gap**.

- GLMSELECT offers great flexibility for and insight into the model selection algorithm.

- GLMSELECT provides results (displayed tables, output data sets, and macro variables) that make it easy to take the selected model and explore it in more detail in a subsequent procedure such as REG or GLM.

Can Best Subsets be Useful?

- The method of Best Subsets as a variable selection tool is ‘greedy’
  - i.e., it tends to over-fit and has high prediction error.

- Only feasible with modest numbers of variables
- Useful to identify and compare different models with similar fits.

Best Subset, Forward Stepwise, or Lasso?

Analysis and Recommendations Based on Extensive Comparisons

Trevor Hastie, Robert Tibshirani, Ryan J. Tibshirani

In exciting new work, Bertsimas et al. (2016) showed that the classical best subset selection problem in regression modeling can be formulated as a mixed integer optimization (MIO) problem.

Using recent advances in MIO algorithms, they demonstrated that best subset selection can now be solved at much larger problem sizes than what was thought possible in the statistics community. They presented empirical comparisons of best subset selection with other popular variable selection procedures, in particular, the lasso and forward stepwise selection. Surprisingly (to us), their simulations suggested that best subset selection consistently outperformed both methods in terms of prediction accuracy. Here we present an expanded set of simulations to shed more light on these comparisons. The summary is roughly as follows:

Neither best subset selection nor the lasso uniformly dominate the other, with best subset selection generally performing better in high signal-to-noise (SNR) ratio regimes, and the lasso better in low SNR regimes;

Best subset selection and forward stepwise perform quite similarly throughout;

The relaxed lasso (actually, a simplified version of the original relaxed estimator defined in Meinshausen, 2007) is the overall winner, performing just about as well as the lasso in low SNR scenarios, and as well as best subset selection in high SNR scenarios.

Prediction vs. Interpretation

- If prediction is the goal, then
  - LASSO and Elastic Net are extremely useful
  - They generally find important predictors without over-fitting

- If estimation is the goal, then consider
  - Use LAR, LASSO or Elastic Net to select the model
  - then estimate and interpret the regression coefficients by ordinary weighted least squares (Efron)

Wrap-up

When are these methods useful (vs OLS)?

- When you have lots of predictors, esp if more than observations (p>n)
- When PREDICTION is your goal
- When you have strong collinearity

Books of interest:

- Computing age statistical inference (Efron, Hastie)
- Elements of statistical learning (Hastie, Tibshirani, Friedman; Springer, but also online)

Future work:

- No solution yet for using these methods with:
  - repeated measures (GEE-LASSO has been suggested, but not yet available in software.
  - Spatial data
  - Dependent data