Statistical Machine Learning: Regression

Daniel J. McDonald

Indiana University, Bloomington

mypage.iu.edu/∼dajmcdon

February 24-26, 2015

Review

STATISTICAL MODELS

We observe data $Z_1, Z_2, \ldots, Z_n$ generated by some probability distribution P. We want to use the data to learn about P.

A statistical model is a set of distributions $\mathbb{P}$.

Some examples:

1. $\mathbb{P} = \{P(z = 1) = p,\ P(z = 0) = 1 - p,\ 0 < p < 1\}$.
2. $\mathbb{P} = \{Y \sim N(X^\top \beta, \sigma^2),\ \beta \in \mathbb{R}^p,\ \sigma > 0,\ X \text{ fixed}\}$.
3. $\mathbb{P} = \{\text{all CDF's } F\}$.
4. $\mathbb{P} = \{\text{all smooth functions } f : \mathbb{R}^p \to \mathbb{R}\}$

EVALUATION

Once I choose some way to “learn” a statistical model, I need to decide if I’m doing a good job.

How do I decide if I’m doing anything good?

Prediction risk

Prediction risk

$$R_n(\widehat{f}) = \mathbb{E}[\ell(Y, \widehat{f}(X))]$$

where the expectation is taken over the new data point (Y, X) and $\mathcal{D}_n$ (everything that is random).

For regression applications, we will use squared-error loss:

$$R_n(\widehat{f}) = \mathbb{E}[(Y - \widehat{f}(X))^2]$$

For classification applications, we will use zero-one loss:

$$R_n(\widehat{f}) = \mathbb{E}[I(Y \neq \widehat{f}(X))]$$

BIAS-VARIANCE DECOMPOSITION

So,

$$\begin{array}{lcllcl} \text{prediction risk} & = & \text{bias}^2 & + & \text{variance} & + \quad \text{irreducible error} \\ \text{estimation risk} & = & \text{bias}^2 & + & \text{variance} & \end{array}$$

What is $R(a)$ for our estimator $\widehat{Y}_a = a\overline{Y}_n$?

$$\text{bias}(\widehat{Y}_a) = \mathbb{E}[a\overline{Y}_n] - \mu = (a-1)\mu$$

$$\text{var}(\widehat{f}(x)) = \mathbb{E}[(a\overline{Y}_n - \mathbb{E}[a\overline{Y}_n])^2] = a^2\mathbb{E}[(\overline{Y}_n - \mu)^2] = \frac{a^2}{n}$$

$$\sigma^2 = \mathbb{E}[(Y-\mu)^2] = 1$$

$$\left(\text{That is: } R_n(\widehat{Y}_a) = (a-1)^2\mu^2 + \frac{a^2}{n} + 1\right)$$

BIAS-VARIANCE DECOMPOSITION

So,

$$\begin{array}{lcllcl} \text{prediction risk} & = & \text{bias}^2 & + & \text{variance} & + \quad \text{irreducible error} \\ \text{estimation risk} & = & \text{bias}^2 & + & \text{variance} & \end{array}$$

What is $R(a)$ for our estimator $\widehat{Y}_a = a\overline{Y}_n$?

$$\text{bias}(\widehat{Y}_a) = \mathbb{E}[a\overline{Y}_n] - \mu = (a-1)\mu$$

$$\text{var}(\widehat{f}(x)) = \mathbb{E}[(a\overline{Y}_n - \mathbb{E}[a\overline{Y}_n])^2] = a^2\mathbb{E}[(\overline{Y}_n - \mu)^2] = \frac{a^2}{n}$$

$$\sigma^2 = \mathbb{E}[(Y-\mu)^2] = 1$$

$$\left(\text{That is: } R_n(\widehat{Y}_a) = (a-1)^2\mu^2 + \frac{a^2}{n} + 1\right)$$

BIAS-VARIANCE TRADEOFF: OVERVIEW

- bias: how well does $\widehat{f}$ approximate the truth g
- more complicated $\mathcal{F}$, lower bias. Flexibility $\Rightarrow$ Parsimony
- more flexibility $\Rightarrow$ larger variance
- complicated models are hard to estimate precisely for fixed n
- irreducible error

Moving forward

THE SETUP

Suppose there is a scientific problem we are interested in solving

(This could be estimating a relationship between height and weight in humans)

The perspective of this class is to define these quantities as random, say

- $X =$ height
- $Y =$ weight

We want to know about the joint distribution of X and Y

This joint distribution is unknown, and hence we

1. GATHER DATA
2. ESTIMATE IT

The Setup

Now we have data

$$D_n = \{(X_1, Y_1), (X_2, Y_2), \ldots, (X_n, Y_n)\},$$

where

- $X_i \in \mathbb{R}^p$ are the covariates, explanatory variables, or predictors
  (NOT INDEPENDENT VARIABLES!)
- $Y_i \in \mathbb{R}$ are the response or supervisor variables.
  (NOT DEPENDENT VARIABLE!)

Finding the joint distribution of X and Y is usually too ambitious

A good first start is to try and get at the mean, say

$$Y = \mu(X) + \epsilon$$

where ϵ describes the random fluctuation of Y around its mean

PARAMETERIZING THIS RELATIONSHIP

Even estimating $\mu(X)$ is often too hard

A approximation is to assume that it is linear.

This means we suppose that there is a $\beta \in \mathbb{R}^p$ such that:

$$Y = \underbrace{\mu(X) + \epsilon = X^\top \beta + \epsilon}_{\text{simplification}} \in \mathbb{R}$$

(The notation $\in$ indicates 'in' and if I say $x \in \mathbb{R}^q$, that means that x is a vector with q entries)

PARAMETERIZING THIS RELATIONSHIP

Translated into using our data, we get

$$Y = \mathbb{X}\beta + \epsilon \in \mathbb{R}^n$$

where

$$\mathbb{X} = \begin{bmatrix} X_1^\top \\ X_2^\top \\ \vdots \\ X_n^\top \end{bmatrix}$$

Commonly, $\mathbb{X}_{i1} = 1$, which encodes an intercept term in the model.

$\mathbb{X}$ is known as the design or feature matrix

PARAMETERIZING THIS RELATIONSHIP: IN DETAIL

Back to the height/weight example:

$$X_1^\top = [1, 62], X_2^\top = [1, 68], \ldots, X_{10}^\top = [1, 65]$$

Estimating

$$\widehat{\beta} = \underset{\beta}{\text{argmin}} \, ||\mathbb{X}\beta - Y||_2^2 = \underset{\beta}{\text{argmin}} \sum_{i=1}^{n} \left(Y_i - X_i^\top \beta \right)^2$$

finds the usual simple linear regression estimators: $\widehat{\beta}^\top = [\widehat{\beta}_0, \widehat{\beta}]$

What’s this argmin?

NECESSARY BACKGROUND: OPTIMIZATION

Suppose I have a function

$$f(x) = (x-2)^2.$$

I want to find the minimizer:

FOC: $f'(x) = 0 \Rightarrow 2x = 4 \Rightarrow x = 2$

Necessary for minimum: $f''(x) > 0$.
$f''(x) = 2 > 0$ ✓

WHAT'S THE *argmin*?

The minimizer is

$$2 = \operatorname*{argmin}_x (x-2)^2$$

The minimum is

$$0 = \min_x (x-2)^2 = (2-2)^2$$

A HARDER ONE:

$$\widehat{\beta} = \underset{\beta}{\operatorname{argmin}} \frac{1}{n} ||Y - \mathbb{X}\beta||_2^2$$

I want to find the minimizer:

FOC: $f'(\beta) = 0 \Rightarrow 2\mathbb{X}^\top Y = 2\beta\mathbb{X}^\top\mathbb{X} \Rightarrow \beta = (\mathbb{X}^\top\mathbb{X})^{-1}\mathbb{X}^\top Y$

Necessary for minimum: $f''(x) > 0$.

$f''(x) = \frac{2}{n}\mathbb{X}^\top\mathbb{X}$ (here we need this matrix to be positive definite)

WHAT ABOUT CONSTRAINTS?

$$\widehat{\beta} = \underset{||\beta||_2^2 < t}{\text{argmin}} \frac{1}{n} ||Y - \mathbb{X}\beta||_2^2$$

I want to find the minimizer:

FOC: $f'(\beta) = 0 \Rightarrow 2\mathbb{X}^\top Y = 2\beta\mathbb{X}^\top\mathbb{X} \Rightarrow \beta = (\mathbb{X}^\top\mathbb{X})^{-1}\mathbb{X}^\top Y$

This is only a necessary condition. We must ensure that this solution satisfies the constraint as well

LAGRANGE MULTIPLIERS

It turns out we can write the dual problem

$$\widehat{\beta} = \text{argmin}\, \frac{1}{n} ||Y - \mathbb{X}\beta||_2^2 + \lambda(||\beta||_2^2 - t)$$

Differentiate first in β, then in λ.
FOC:
$f'(\beta) = 0 \Rightarrow 2\mathbb{X}^\top Y = 2\beta(\mathbb{X}^\top \mathbb{X} + \lambda I) \Rightarrow \beta = (\mathbb{X}^\top \mathbb{X} + \lambda I)^{-1}\mathbb{X}^\top Y$

FOC: $f'(\lambda) = 0 \Rightarrow ||\beta||_2^2 = t$
First condition gives the solution, the second relates λ to t.

What Makes a Good Estimator of β?

The least squares solution minimizes the training error (in regression this is sometimes called 'mean squared error' or 'residual sums of squares').

$$\widehat{\beta}_{\text{OLS}} = \underset{\beta}{\text{argmin}}\, \widehat{R}_n(\beta) = \underset{\beta}{\text{argmin}} \sum_{i=1}^{n} (Y_i - \beta^\top X_i)^2 = (\mathbb{X}^\top \mathbb{X})^{-1} \mathbb{X}^\top \mathbb{Y},$$

where:

$$\mathbb{X} = \begin{bmatrix} 1 & X_{12} & \cdots & X_{1p} \\ 1 & X_{22} & \cdots & X_{2p} \\ \vdots & & & \\ 1 & X_{n2} & \cdots & X_{np} \end{bmatrix} \quad \text{and} \quad \mathbb{Y} = \begin{bmatrix} Y_1 \\ \vdots \\ Y_n \end{bmatrix}.$$

There's a problem: we can make the training error arbitrarily small!

What Makes a Good Estimator of β?

Here's an example: Suppose we have $n = 20$ observations of pairs of data such that $X_i \in \mathbb{R}$.

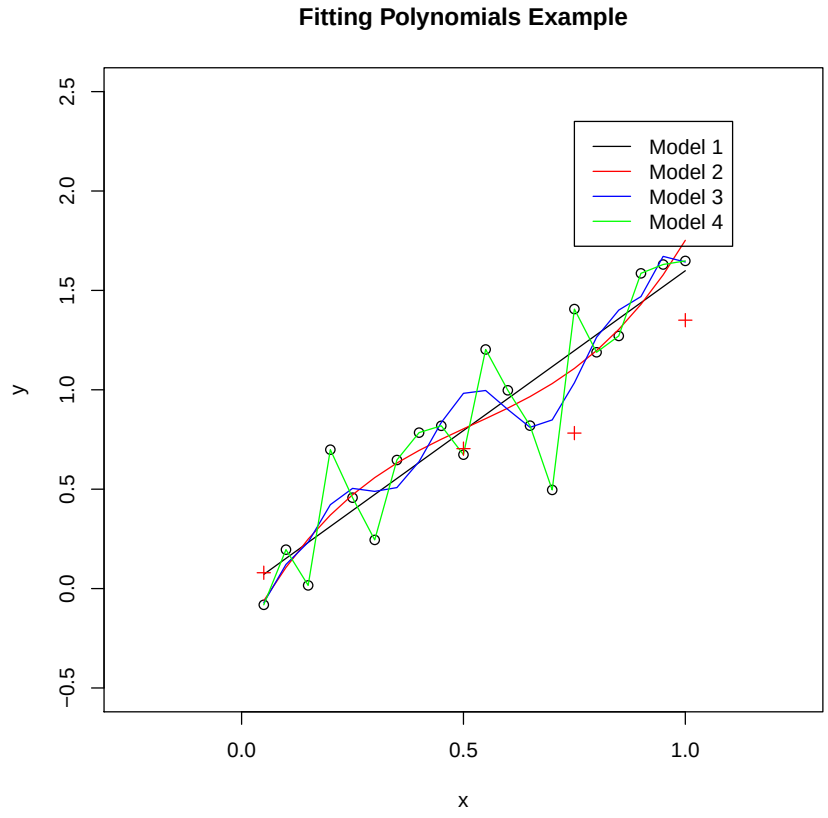
Let's fit some polynomials to this data (this is known in ML as feature creation).

We can consider the following models:

- Model 1: $m(X) = \beta_0 + \beta_1 X$
- Model 2: $m(X) = \beta_0 + \beta_1 X + \beta_2 X^2 + \beta_3 X^3$
- Model 3: $m(X) = \beta_0 + \sum_{k=1}^{10} \beta_k X^k$
- Model 4: $m(X) = \beta_0 + \sum_{k=1}^{20} \beta_k X^k$

Let's look at what happens...

What Makes a Good Estimator of β?



The MSE's are: $\widehat{R}_n(\text{Model 1}) = 0.049$, $\widehat{R}_n(\text{Model 2}) = 0.043$, $\widehat{R}_n(\text{Model 3}) = 0.034$, and $\widehat{R}_n(\text{Model 4}) = 0$.

What about predicting new observations (red crosses)?

REGULARIZATION

Minimizing the training error ($\widehat{R}_n$) as our criterion has led us to a poor quality estimator.

What can we do instead? Regularization!

Regularization is a general philosophy for trading model fit (bias) with model complexity (variance).

Ridge regression

RIDGE REGRESSION

The first method we will mention is quite old: ridge regression. Originally, it was developed as a way to deal with multicollinearity and (independently) as a way to invert ill-conditioned operators.

Ridge regression is defined as:

$$\widehat{\beta}_{ridge}(t) = \underset{||\beta||_2^2 \leq t}{\text{argmin}} \, ||\mathbb{Y} - \mathbb{X}\beta||_2^2$$

where we are using $||\beta||_2^2 = \sum_{j=1}^{p} \beta_j^2$.

Important: the solution depends on the units of the matrix $\mathbb{X}$ in a nonlinear fashion. We should scale the matrix to have zero mean and unit standard deviation columns before forming $\widehat{\beta}_{ridge}(t)$.

RIDGE REGRESSION: ALTERNATE FORMULATIONS

$$\widehat{\beta}_{ridge}(t) = \underset{||\beta||_2^2 \leq t}{\text{argmin}} \, ||\mathbb{Y} - \mathbb{X}\beta||_2^2$$

Using the theory of Lagrangian multipliers, we can rewrite this for some parameter $\lambda \geq 0$ as

$$\widehat{\beta}_{ridge}(\lambda) = \underset{\beta}{\text{argmin}} \, ||\mathbb{Y} - \mathbb{X}\beta||_2^2 + \lambda||\beta||_2^2$$

Also, suppose that $\mathbb{Y}|\mathbb{X}, \beta \sim N(\mathbb{X}\beta, I)$ and $\beta \sim N(0, \lambda^{-1/2}I)$.

Then the posterior mean given by this Bayesian hierarchical model is $\widehat{\beta}_{ridge}(\lambda)$ as well.

- $\widehat{\beta}_{ridge}(\lambda = 0) = \widehat{\beta}_{ridge}(t = \infty) = \widehat{\beta}_{\text{OLS}}$.
- Any $\lambda > 0$ penalizes larger values of β, effectively shrinking it to zero (consider the Bayesian interpretation).

Singular Value Decomposition (SVD)

NECESSARY BACKGROUND: SVD

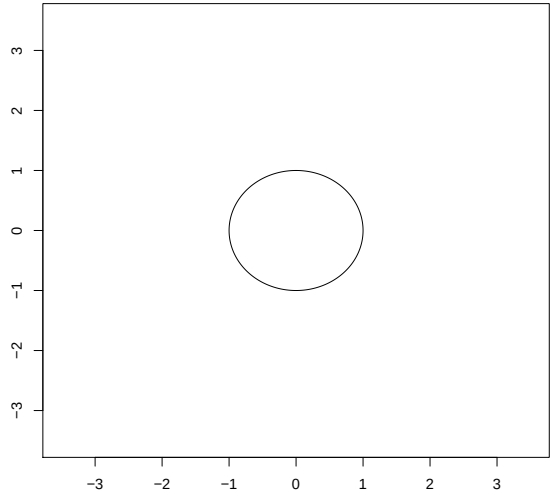
It turns out we can think of matrix multiplication in terms of circles and ellipses

(The plural is technically ellipsoids, but this term seems to freak people out)

Take a matrix $\mathbb{X}$ and let's look at the set of vectors

$$B = \{\beta : ||\beta||_2 \leq 1\}$$

This is a circle!

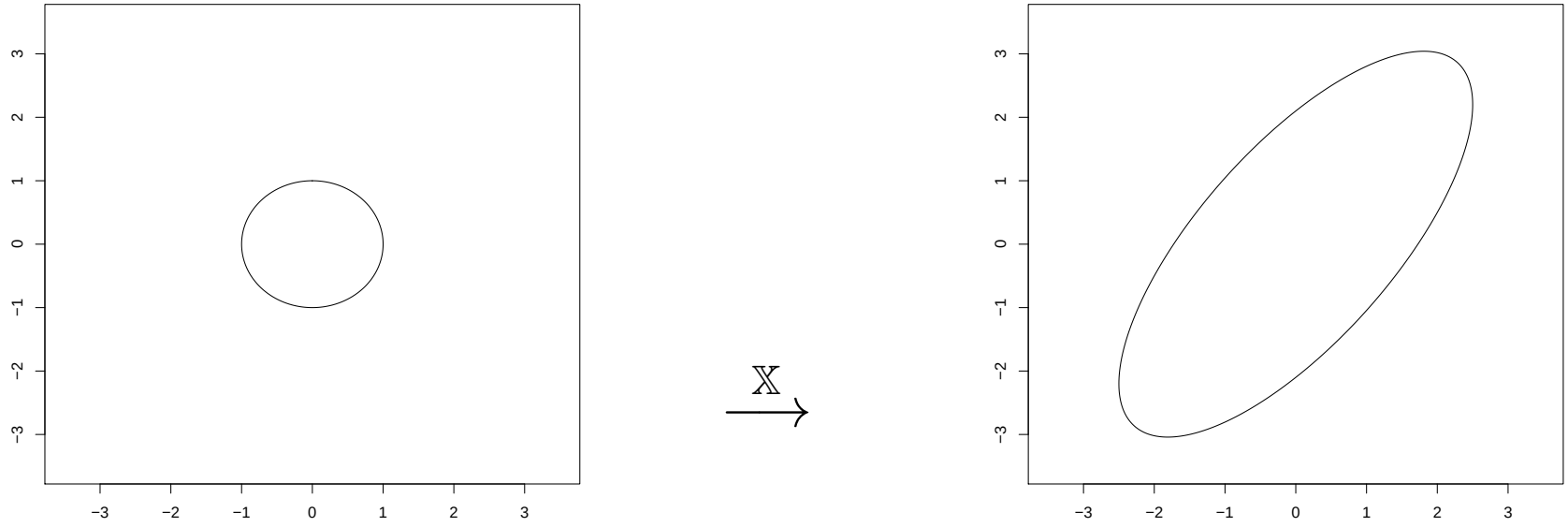


Necessary Background: SVD

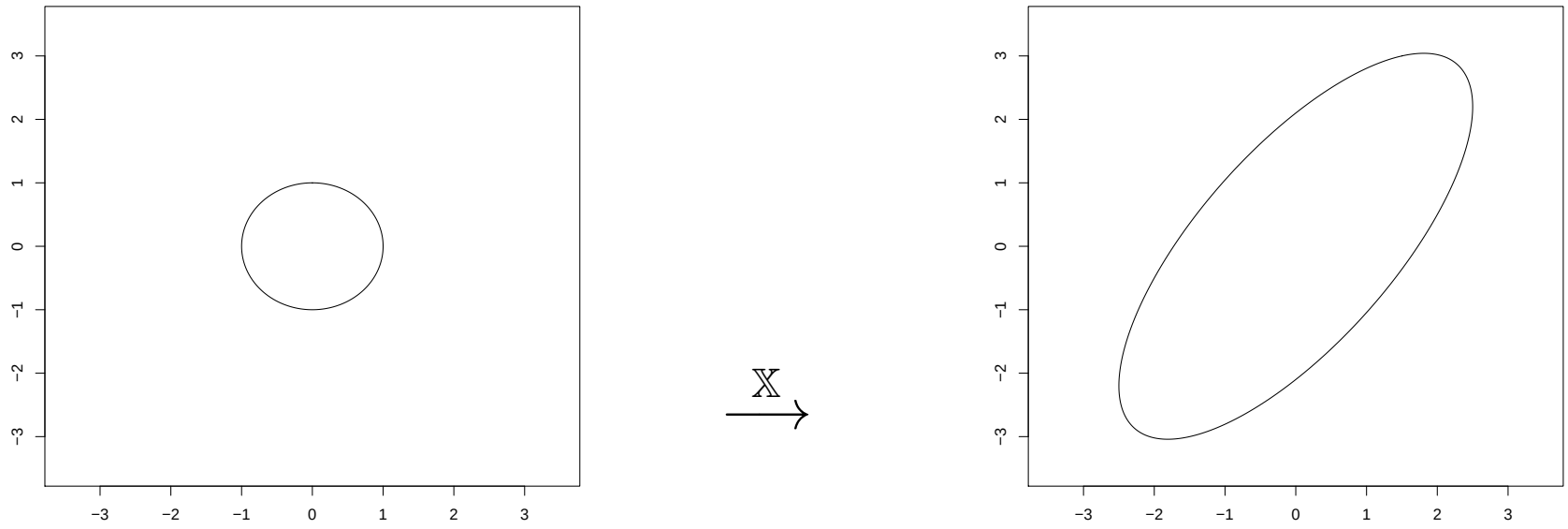
What happens when we multiply vectors in this circle by $\mathbb{X}$?

Let

$$\mathbb{X} = \begin{bmatrix} 2.0 & 0.5 \\ 1.5 & 3.0 \end{bmatrix} \text{ and } \mathbb{X}\beta = \begin{bmatrix} 2\beta_1 + 0.5\beta_2 \\ 1.5\beta_1 + 3\beta_2 \end{bmatrix}$$



NECESSARY BACKGROUND: SVD

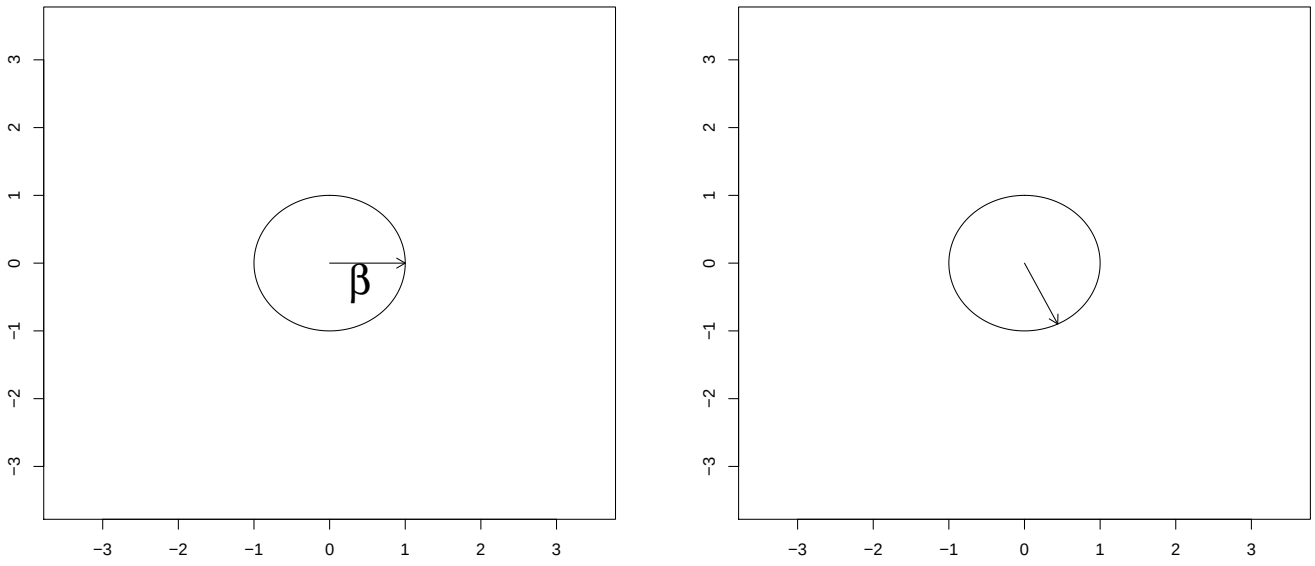


What happened?

1. The coodinate axis gets rotated
2. The new axis gets elongated (making an ellipse)
3. This ellipse gets rotated

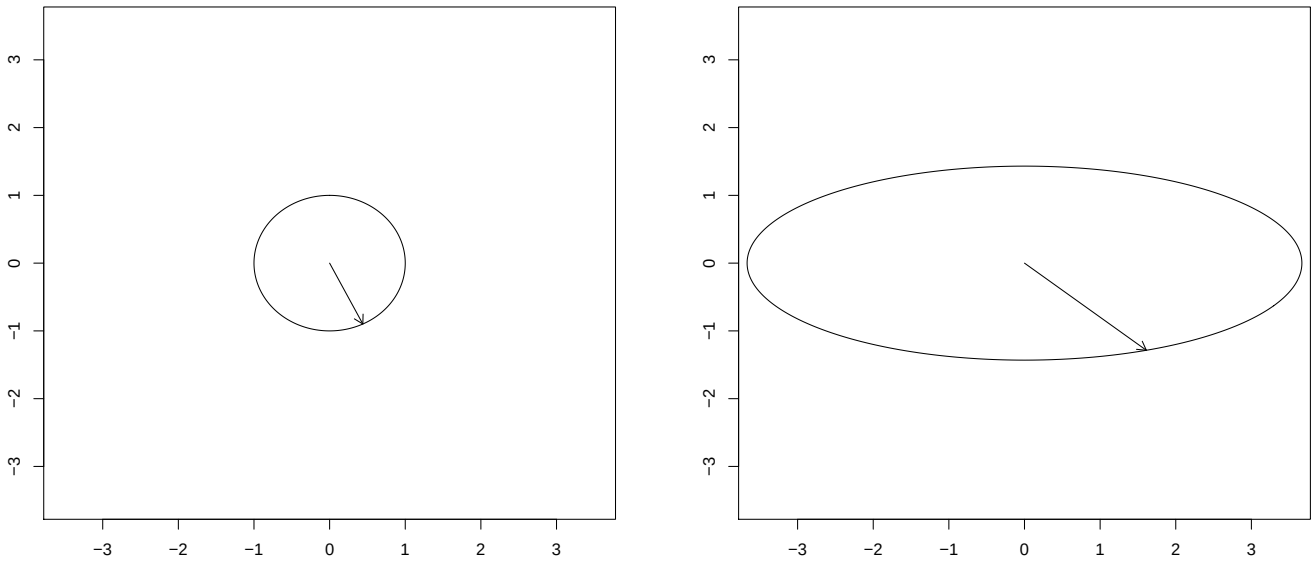
Let's break this down into parts...

NECESSARY BACKGROUND: SVD



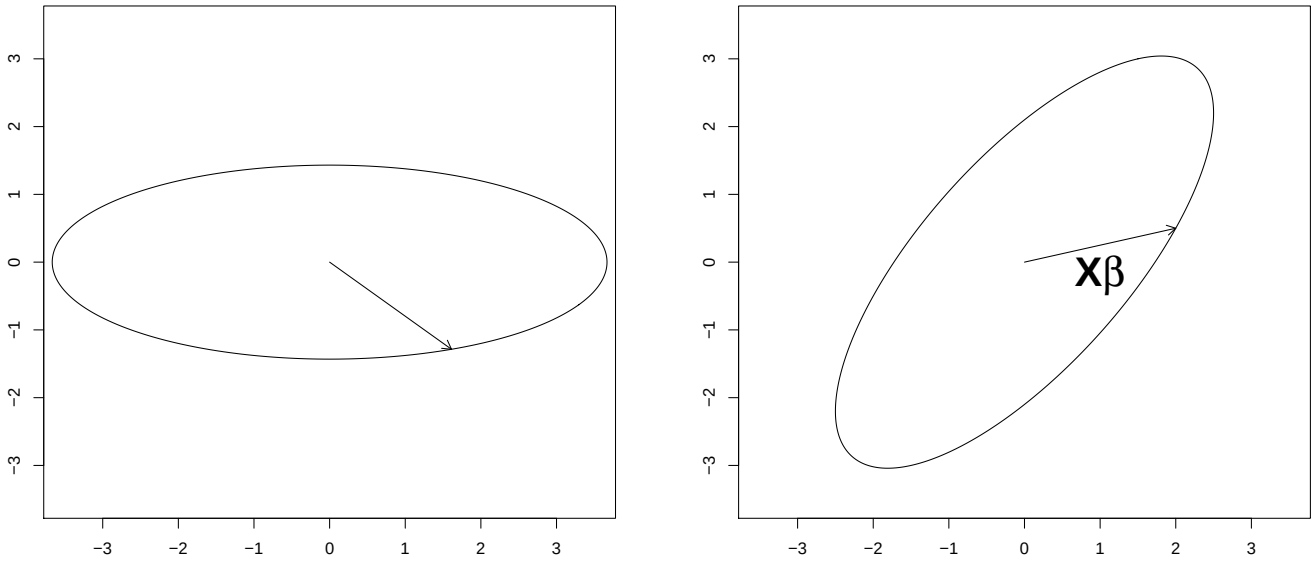
1. The coordinate axis gets rotated
2. The new axis gets elongated (making an ellipse)
3. This ellipse gets rotated

Necessary Background: SVD



1. The coordinate axis gets rotated
2. The new axis gets elongated (making an ellipse)
3. This ellipse gets rotated

NECESSARY BACKGROUND: SVD



1. The coordinate axis gets rotated
2. The new axis gets elongated (making an ellipse)
3. This ellipse gets rotated

Necessary Background: Rotation

Rotations: These can be thought of as just reparameterizing the coordinate axis. This means that they don't change the geometry.

As the original axis was orthogonal (that is; perpendicular), the new axis must be as well.

Necessary Background: Rotation

Let $\mathbf{v}_1, \mathbf{v}_2$ be two normalized, orthogonal vectors. This means that:

$$\mathbf{v}_1^\top \mathbf{v}_2 = 0 \quad \text{and} \quad \mathbf{v}_1^\top \mathbf{v}_1 = \mathbf{v}_2^\top \mathbf{v}_2 = 1$$

In matrix notation, if we create V as a matrix with normalized, orthogonal vectors as columns, then:

$$V^\top V = \begin{bmatrix} 1 & 0 & 0 & \dots & 0 \\ 0 & 1 & 0 & \dots & 0 \\ & & \vdots & & \\ 0 & 0 & 0 & \dots & 1 \end{bmatrix} = I$$

Here, I is the identity matrix.

NECESSARY BACKGROUND: ELONGATION

Elongation: These can be thought of as stretching vectors along the current coordinate axis. This means that they **do** change the geometry by distorting distances.

Elongations are the result of multiplication by a diagonal matrix (note: we just saw a very special case of such a matrix: the identity matrix I)

All diagonal matrices have the form:

$$D \begin{bmatrix} d_1 & 0 & 0 & \dots & 0 \\ 0 & d_2 & 0 & \dots & 0 \\ & & \vdots & & \\ 0 & 0 & 0 & \dots & d_p \end{bmatrix}$$

Necessary Background: SVD

Using this intuition, for any matrix $\mathbb{X}$ it is possible to write its SVD:

$$\mathbb{X} = UDV^\top$$

where

- U and V are orthogonal (think: rotations)
- D is diagonal (think: elongation)
- The diagonal elements of D are ordered as

$$d_1 \geq d_2 \geq \ldots \geq d_p \geq 0$$

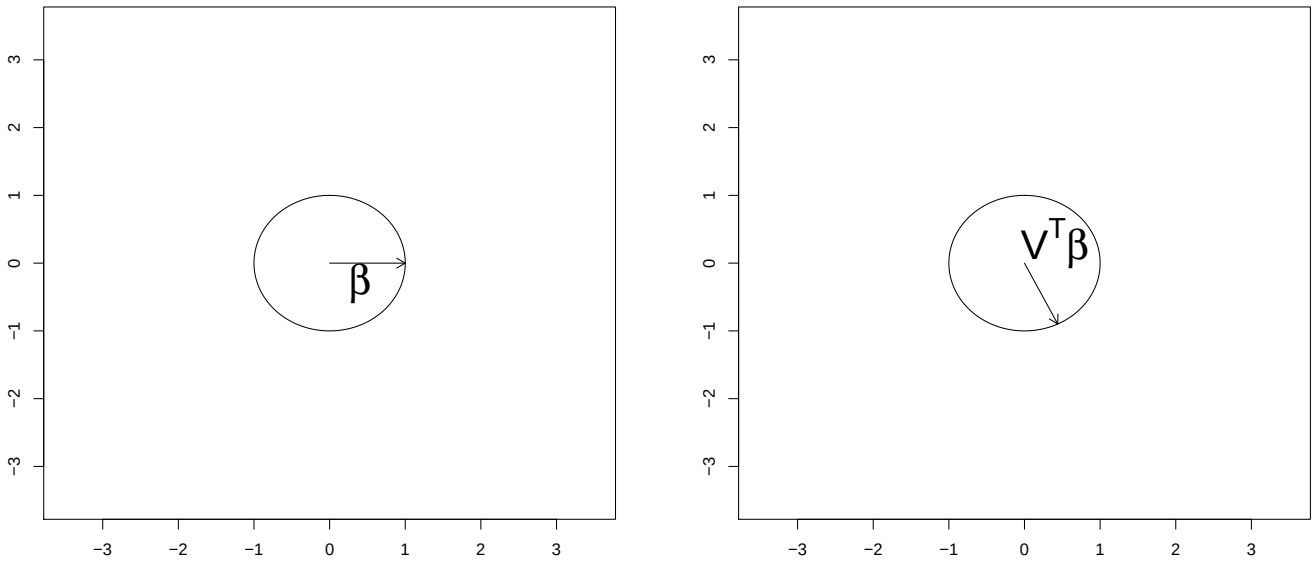
NECESSARY BACKGROUND: SVD

Many properties of matrices can be 'read off' from the SVD.

Rank: The rank of a matrix answers the question: how many dimensions does the ellipse live in? In other words, it is the number of columns of the matrix $\mathbb{X}$, not counting the columns that are 'redundant'

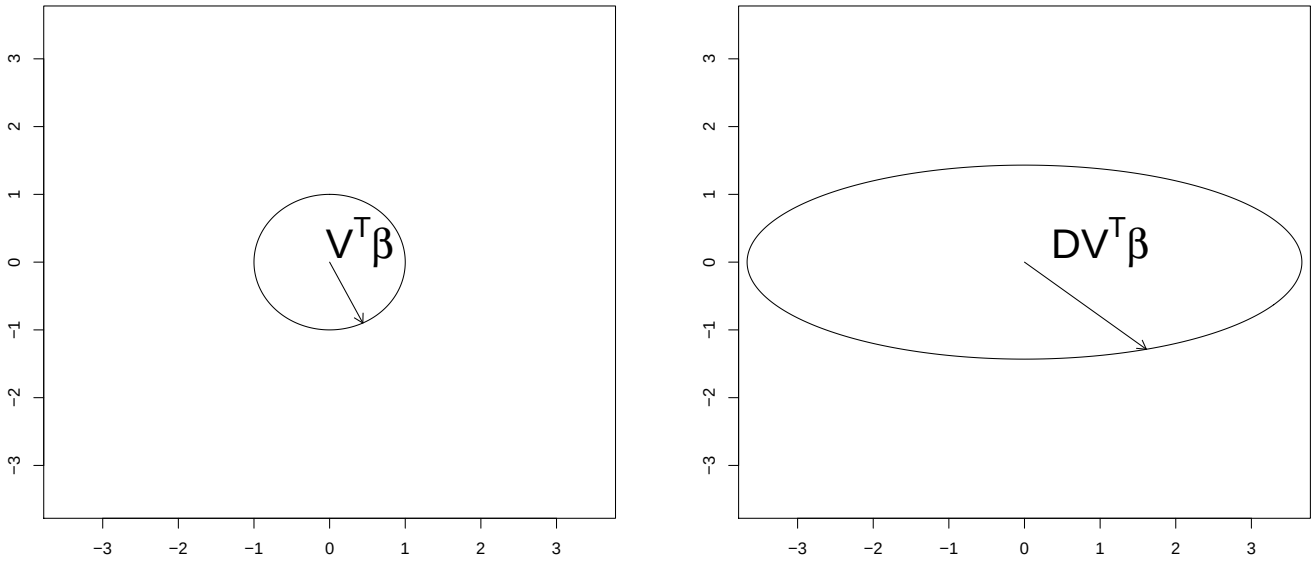
It turns out the rank is exactly the quantity q such that $d_q > 0$ and $d_{q+1} = 0$

NECESSARY BACKGROUND: SVD



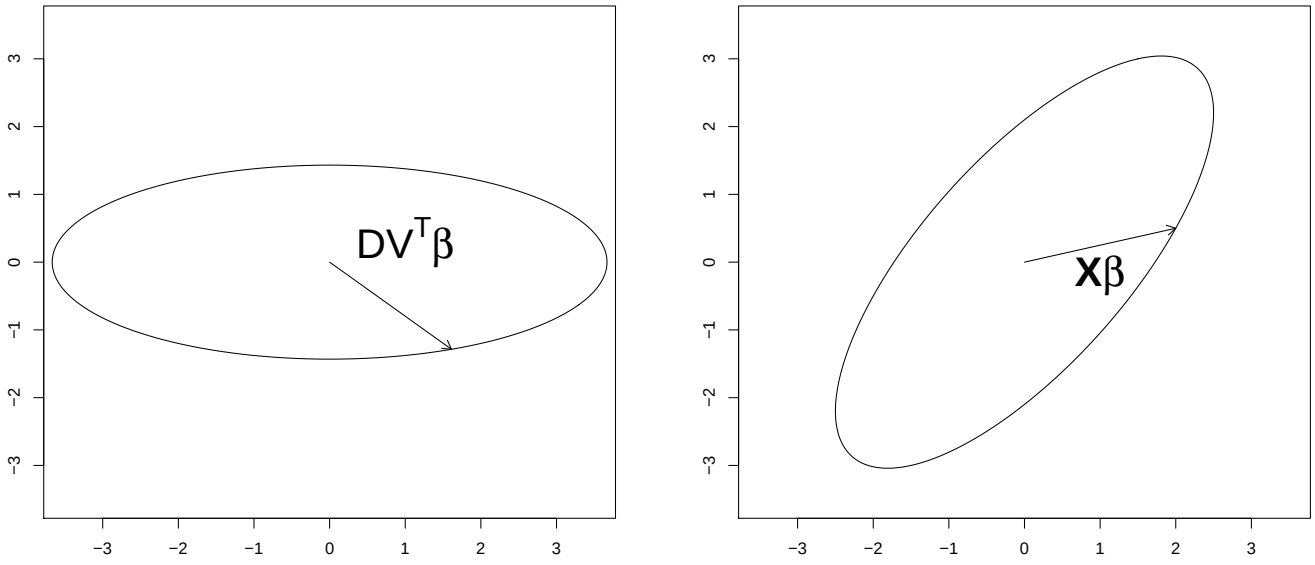
1. The coordinate axis gets rotated (Multiplication by $V^\top$)
2. The new axis gets elongated (Multiplication by D)
3. This ellipse gets rotated (Multiplication by U)

NECESSARY BACKGROUND: SVD



1. The coordinate axis gets rotated (Multiplication by $V^\top$)
2. The new axis gets elongated (Multiplication by D)
3. This ellipse gets rotated (Multiplication by U)

NECESSARY BACKGROUND: SVD



1. The coordinate axis gets rotated (Multiplication by $V^\top$)
2. The new axis gets elongated (Multiplication by D)
3. This ellipse gets rotated (Multiplication by U)

RIDGE REGRESSION

We can gain insight with the help of the SVD.

Write $\mathbb{X} = UDV^\top = \sum_{j=1}^{p} d_j \mathbf{u}_j \mathbf{v}_j^\top$ (here $\mathbf{u}_j$ is the j^{th} column of U)

Then the least squares solution can be written:

$$\widehat{\beta}_{\text{OLS}} = (\mathbb{X}^\top \mathbb{X})^{-1} \mathbb{X}^\top \mathbb{Y} = VD^{-1}U^\top \mathbb{Y} = \sum_{j=1}^{p} \mathbf{v}_j \left(\frac{1}{d_j}\right) \mathbf{u}_j^\top \mathbb{Y}.$$

We can write the ridge regression solution as:

$$\begin{aligned}\widehat{\beta}_{ridge}(\lambda) = (\mathbb{X}^\top \mathbb{X} + \lambda I)^{-1} \mathbb{X}^\top \mathbb{Y} &= V(D^2 + \lambda I)^{-1} DU^\top \mathbb{Y} \\ &= \sum_{j=1}^{p} \mathbf{v}_j \left(\frac{d_j}{d_j^2 + \lambda}\right) \mathbf{u}_j^\top \mathbb{Y}.\end{aligned}$$

Ridge shrinks the data by an additional factor of λ.

Setting the tuning parameter

How to choose λ?

If we choose λ by minimizing $\widehat{R}_n$ (the training error), we will always set $\lambda = 0$.

We want a λ such that $\widehat{\beta}_{ridge}(\lambda)$ predicts well.

That is, for a new (Y, X), we want

$$R_n = \mathbb{E}(Y - X^\top \widehat{\beta}_{ridge}(\lambda))^2$$

to be small.

However, we don't have a new (Y, X)...

How to choose λ?

... or do we?

What if instead we set aside one observation and predict that?

For example: set aside (Y_1, X_1) and fit $\widehat{\beta}^{(1)}_{ridge}(\lambda)$ on $(Y_2, X_2), \ldots, (Y_n, X_n)$.
[$\widehat{\beta}^{(1)}_{ridge}(\lambda)$ means fit without observation (Y_1, X_1)]

$$\tilde{R}_1(\widehat{\beta}^{(1)}_{ridge}(\lambda)) = \left(Y_1 - X_1^\top \widehat{\beta}^{(1)}_{ridge}(\lambda)\right)^2$$

Why stop there? We can do the same thing with the second observation as well:

$$\tilde{R}_2(\widehat{\beta}^{(2)}_{ridge}(\lambda)) = \left(Y_2 - X_2^\top \widehat{\beta}^{(2)}_{ridge}(\lambda)\right)^2$$

How to choose λ?

It can be shown that for any $i = 1, \ldots, n$

$$\mathbb{E}\left[\tilde{R}_i\left(\widehat{\beta}^{(i)}_{ridge}(\lambda)\right)\right] = R_{n-1}(\widehat{\beta}_{ridge}(\lambda))$$

Therefore,

$$\begin{aligned}\mathbb{E}\left[\frac{1}{n}\sum_{i=1}^{n}\tilde{R}_i\left(\widehat{\beta}^{(i)}_{ridge}(\lambda)\right)\right] &= R_{n-1}(\widehat{\beta}_{ridge}(\lambda)) \\ &\approx R_n(\widehat{\beta}_{ridge}(\lambda)) \\ &= \mathbb{E}(Y - X^\top \widehat{\beta}_{ridge}(\lambda))^2\end{aligned}$$

CROSS-VALIDATION

We can use this idea to form an estimate of of the prediction risk.

It is known as (leave-one-out) cross-validation (CV)

$$CV(\lambda) = \frac{1}{n}\sum_{i=1}^{n} \tilde{R}_i\left(\widehat{\beta}^{(i)}_{ridge}(\lambda)\right) = \frac{1}{n}\sum_{i=1}^{n}\left(Y_i - X_i^\top \widehat{\beta}^{(i)}_{ridge}(\lambda)\right)^2$$

Now, we pick

$$\widehat{\lambda} = \underset{\lambda \geq 0}{\operatorname{argmin}}\, CV(\lambda)$$

Problems with leave-one-out CV

There are two strong disadvantages to cross-validation

- It can be computationally demanding (we may need to fit an estimator n different times).
- It is (essentially) an unbiased estimator of the prediction risk (which means it can be very high variance).

K-Fold Cross-validation

A commonly used compromise is to randomly divide your data into K groups.

Let $v_1, \ldots, v_K$ be these groups.

$$CV_K(\lambda) = \frac{1}{K} \sum_{k=1}^{K} \frac{1}{|v_k|} \sum_{i \in v_k} \left(Y_i - X_i^\top \widehat{\beta}_{ridge}^{(v_k)}(\lambda) \right)^2$$

If $|v_1| = \cdots = |v_K| = c$, then

$$\begin{aligned} \mathbb{E}\left[CV_K(\lambda)\right] &= R_{n-c}(\widehat{\beta}_{ridge}(\lambda)) \\ &\approx R_n(\widehat{\beta}_{ridge}(\lambda)) \end{aligned}$$

But this approximation is less precise

$\left[R_{n-c}(\widehat{\beta}_{ridge}(\lambda)) > R_n(\widehat{\beta}_{ridge}(\lambda)) \text{ in most cases}\right]$

Coding ridge (if time)

Example: predicting sex (and I don't mean gender)

SOME STIMULATING DATA

- Randomized controlled trial
- Do happier couples have more sex or does more sex make people happier?
- Couples in "treatment" group $\rightarrow$ told to have more sex
- Observed for 3 months, daily survey completed
- Our covariates are the responses to these survey questions.
- Here: which covariates predict more sex?

Source: Data courtesy of George Loewenstein (CMU)

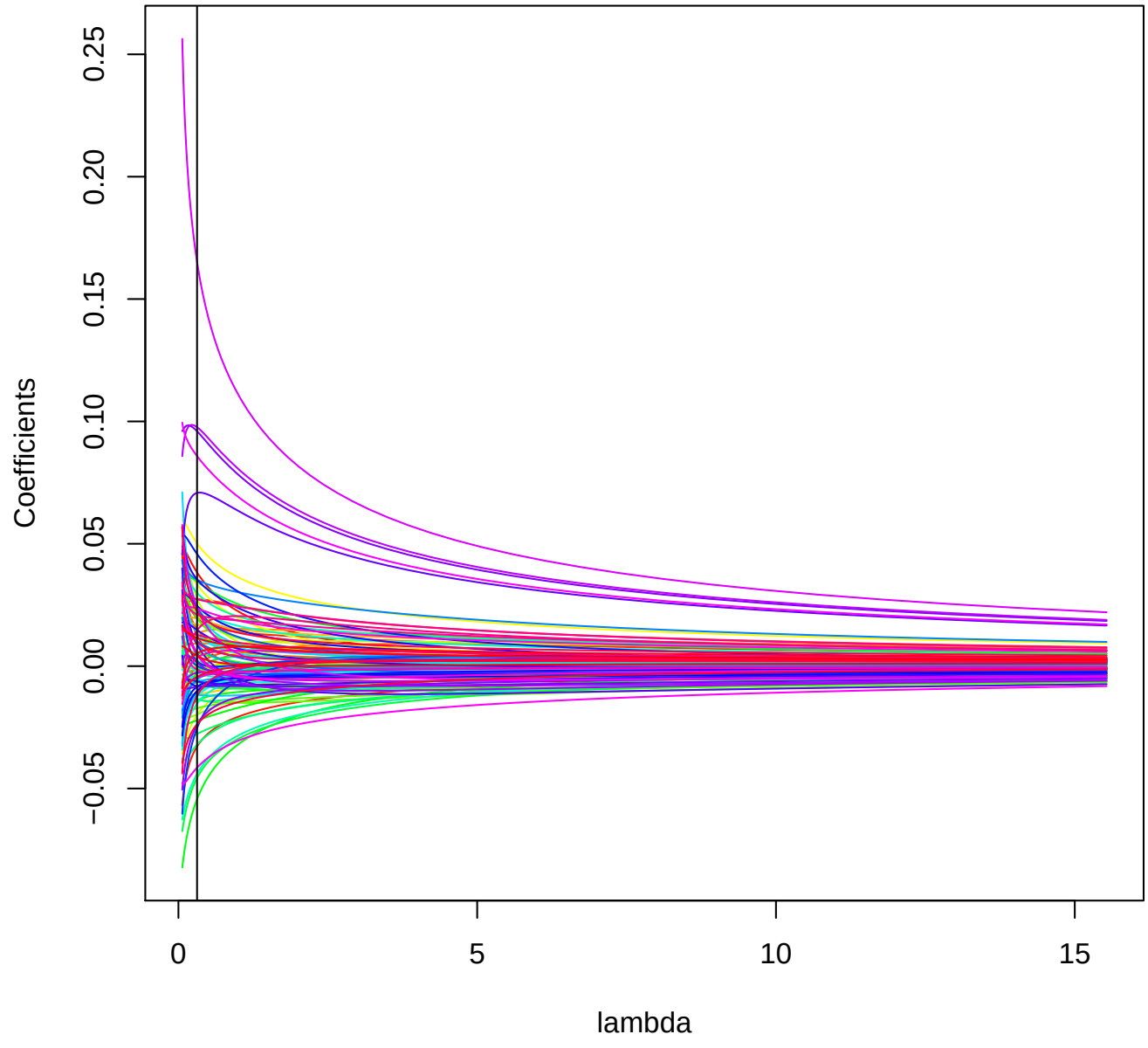
FOR YOUR EDIFICATION. . .

Excess happiness does not seem to lead to more sex.

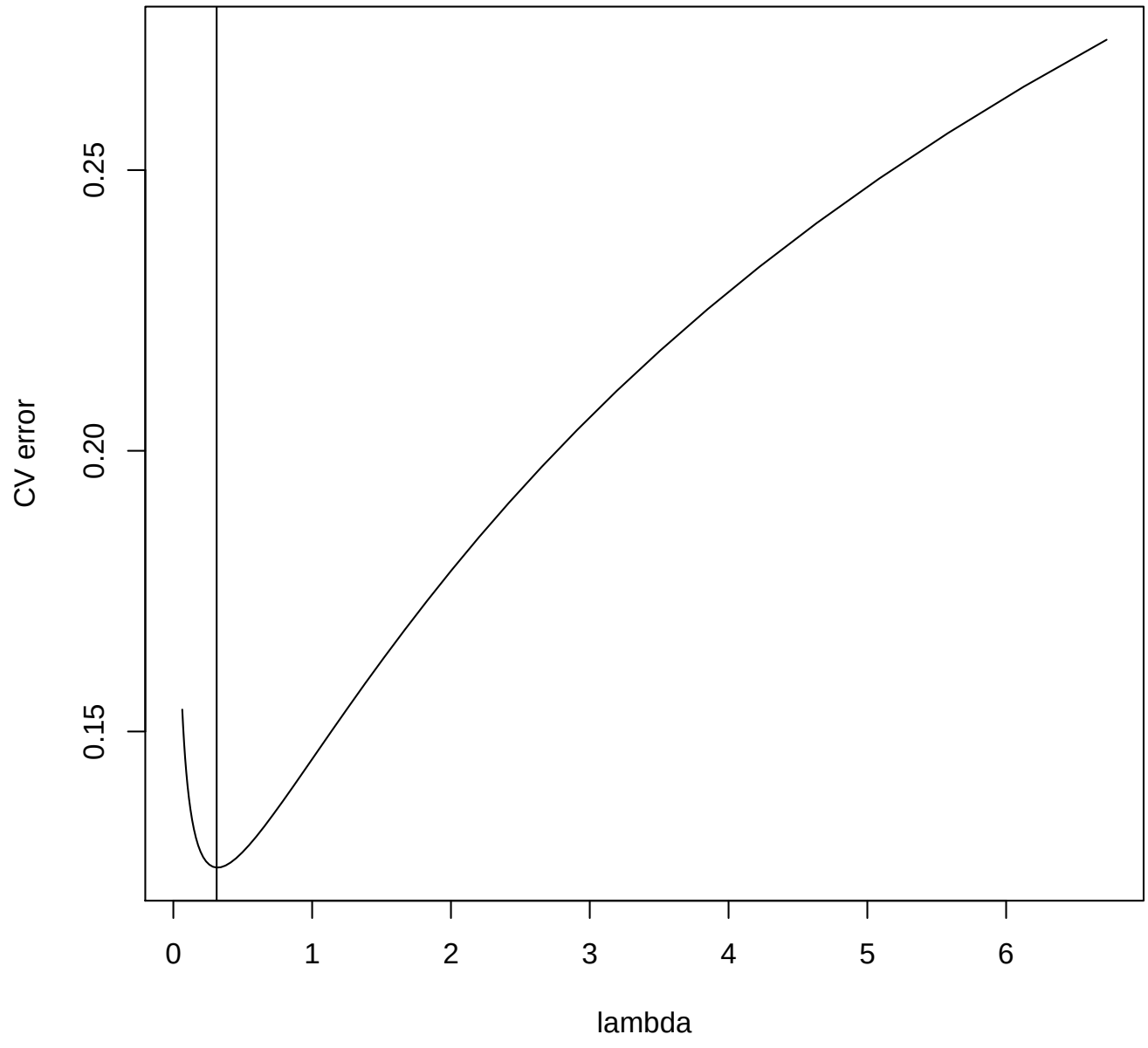
More sex does not seem to lead to happiness.

More good sex does lead to happiness.

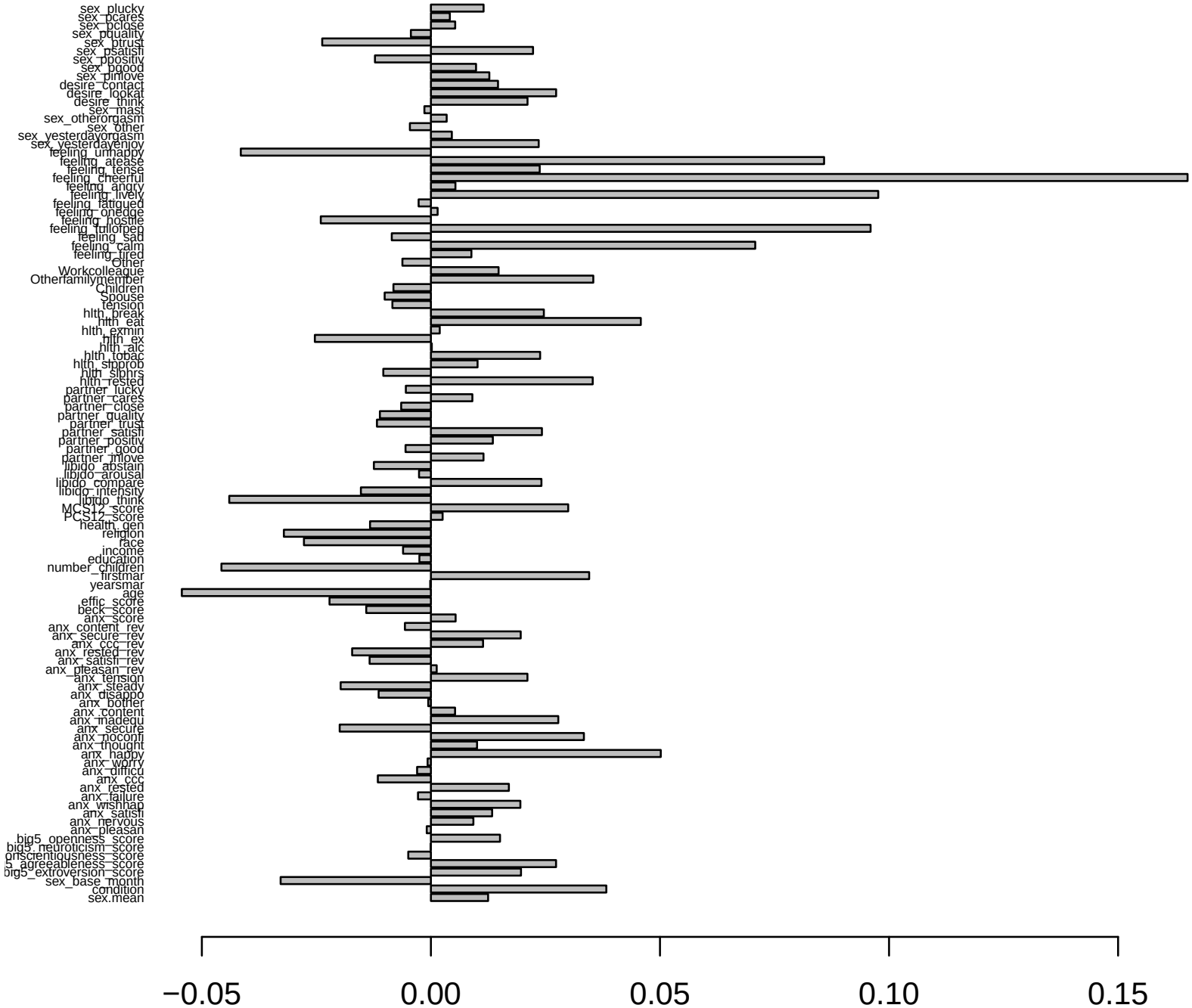
Ridge results: Regularization path



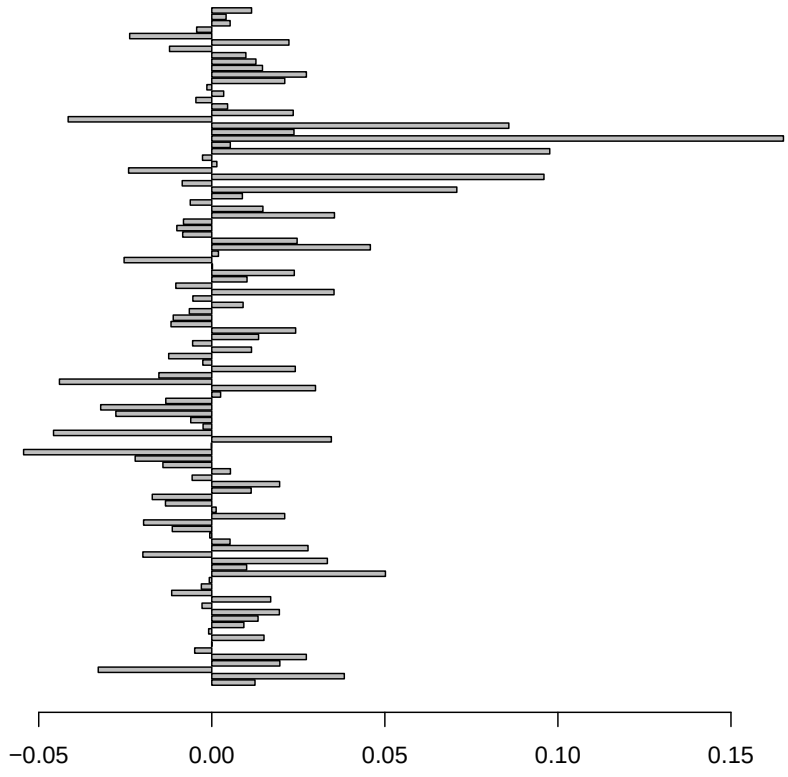
RIDGE RESULTS: CROSS-VALIDATION CURVE



RIDGE RESULTS: COEFFICIENTS OF SELECTED MODEL



RIDGE RESULTS: COEFFICIENTS OF SELECTED MODEL



There are quite a few coefficients near zero.

It is very hard to interpret the results.

We should do model selection to remove unnecessary terms.

Choosing models

MODEL SELECTION REVIEW

Model selection can be accomplished by formulating two ingredients:

(1) An estimator of the risk.

(2) A way of sifting through all the possible models.

We can do (1) by an information criterion (AIC, BIC, AICc, etc.).

For (2)

There are $p = 108$ explanatory variables

This means there are $2^p - 1$ possible models

For this problem, $2^{108} - 1 > 1.8 \times 10^{32}$

(for context, there are about 1.2×10^{23} stars in the universe)

CAN WE GET THE BEST OF BOTH WORLDS?

RIDGE REGRESSION $\quad \min ||\mathbb{Y} - \mathbb{X}\beta||_2^2$ subject to $||\beta||_2^2 \leq t$

BEST LINEAR REGRESSION MODEL $\quad \min ||\mathbb{Y} - \mathbb{X}\beta||_2^2$ subject to $||\beta||_0 \leq t$

$(||\beta||_0 =$ the number of nonzero elements in $\beta)$

The Ridge optimization problem is convex (but doesn't do model selection).

The best linear regression model optimization problem is nonconvex and is an NP-hard problem.

An intuitive idea

Ridge regression $\min ||\mathbb{Y} - \mathbb{X}\beta||_2^2$ subject to $||\beta||_2^2 \leq t$

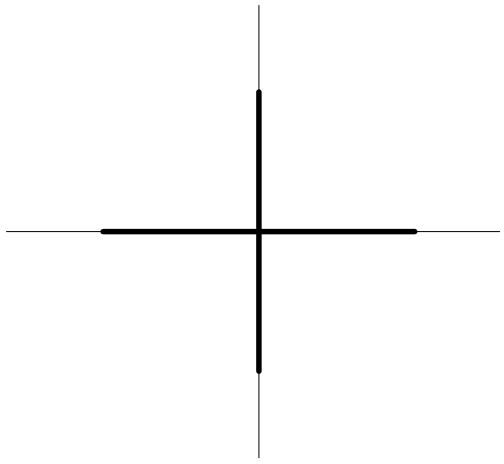
Best linear regression model $\min ||\mathbb{Y} - \mathbb{X}\beta||_2^2$ subject to $||\beta||_0 \leq t$

($||\beta||_0 =$ the number of nonzero elements in β)

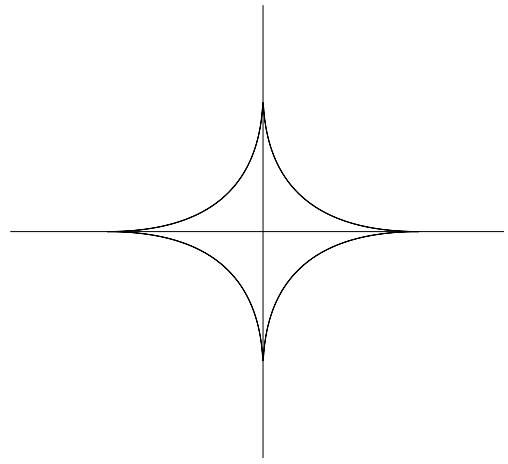
	Best linear regression model	Ridge regression
Computationally Feasible?	No	Yes
Does Model Selection?	Yes	No

Can we 'interpolate' $||\beta||_2$ and $||\beta||_0$ to find a method that does both?

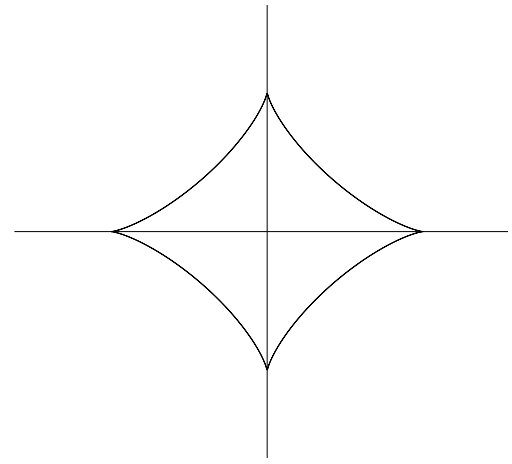
GEOMETRY OF REGULARIZATION IN $\mathbb{R}^2$



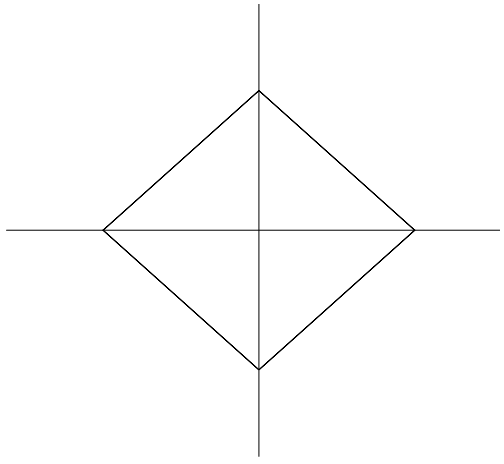
$||\beta||_0 \leq t$



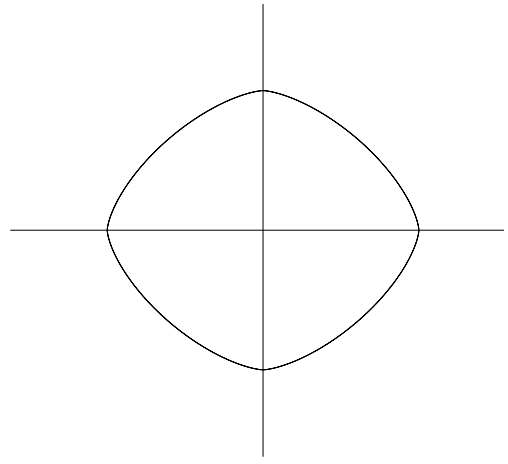
$||\beta||_{\frac{1}{2}} \leq t$



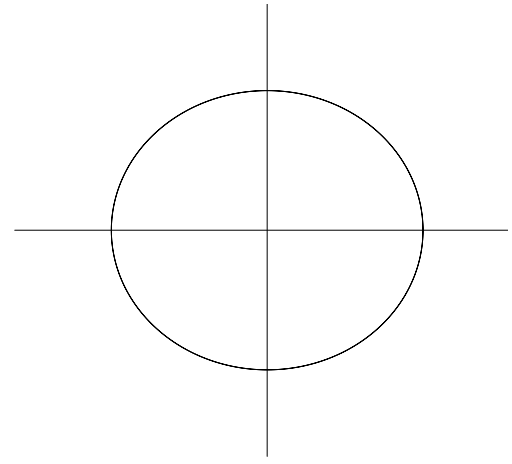
$||\beta||_{\frac{3}{4}} \leq t$



$||\beta||_1 \leq t$

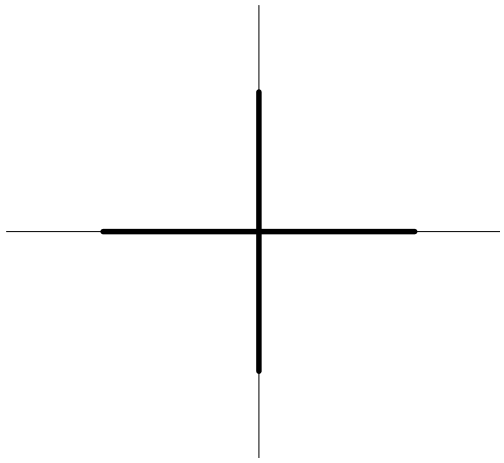


$||\beta||_{\frac{3}{2}} \leq t$

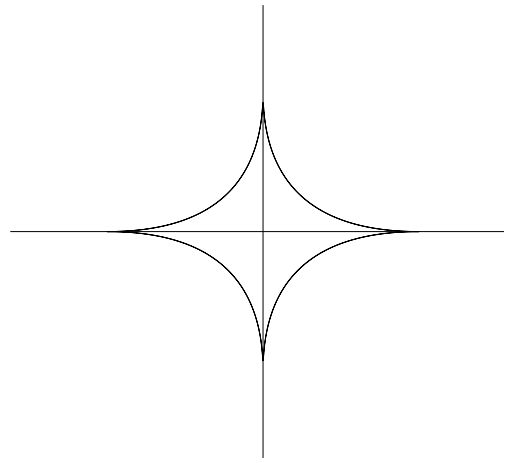


$||\beta||_2 \leq t$

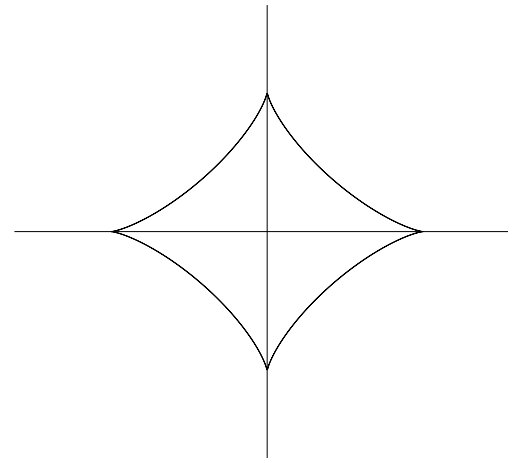
Geometry of regularization in $\mathbb{R}^2$



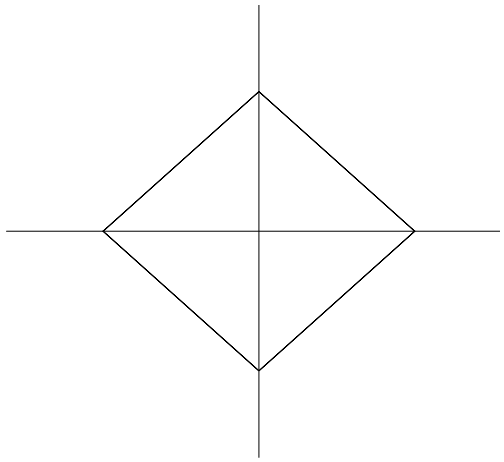
$||\beta||_0 \leq t$



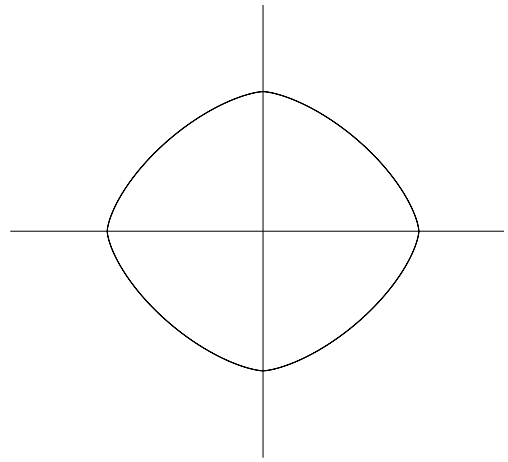
$||\beta||_{\frac{1}{2}} \leq t$



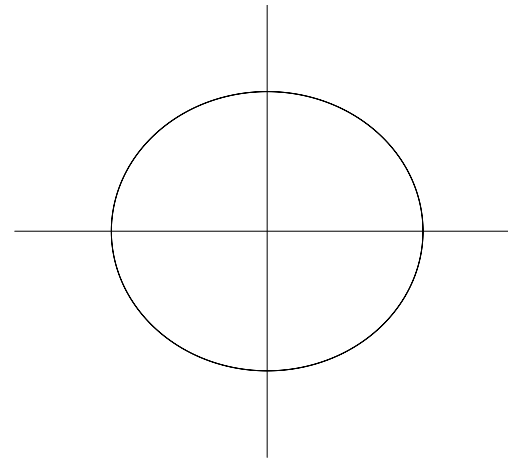
$||\beta||_{\frac{3}{4}} \leq t$



$||\beta||_1 \leq t$

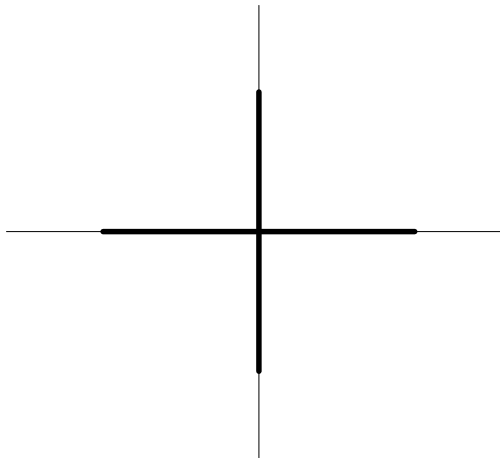


$||\beta||_{\frac{3}{2}} \leq t$

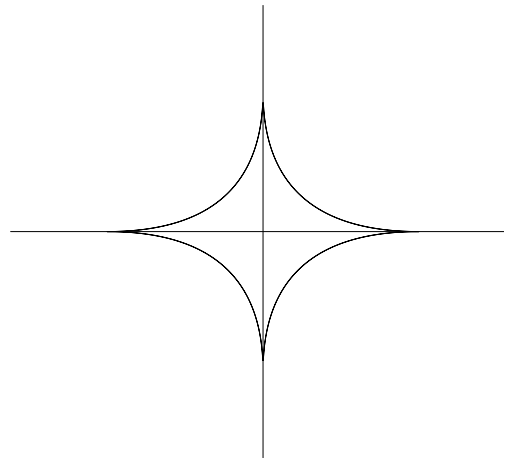


$||\beta||_2 \leq t$

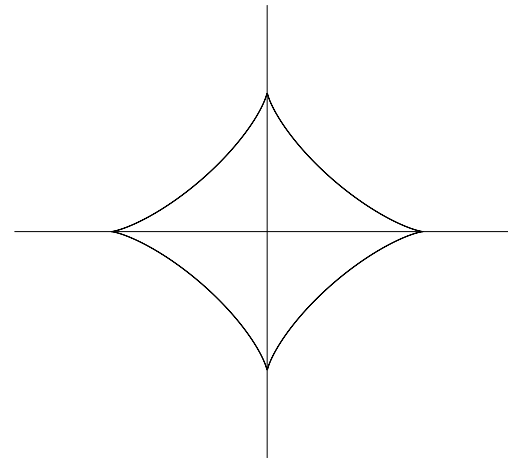
GEOMETRY OF REGULARIZATION IN $\mathbb{R}^2$



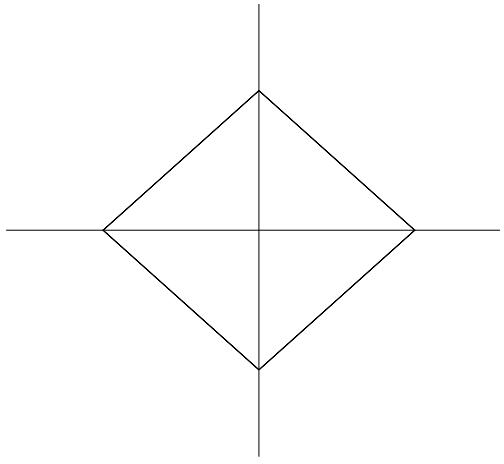
$||\beta||_0 \leq t$



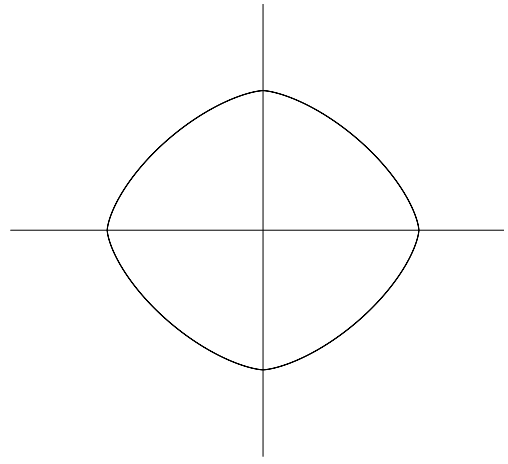
$||\beta||_{\frac{1}{2}} \leq t$



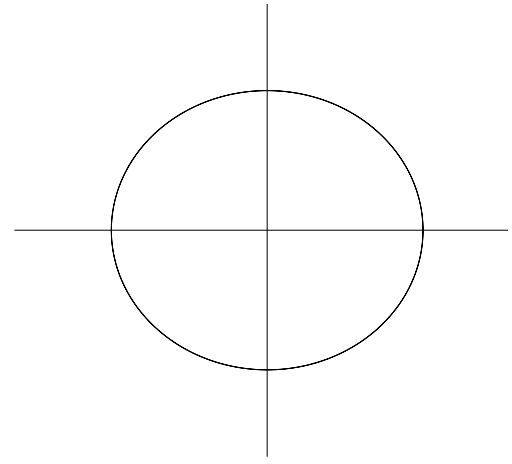
$||\beta||_{\frac{3}{4}} \leq t$



$||\beta||_1 \leq t$

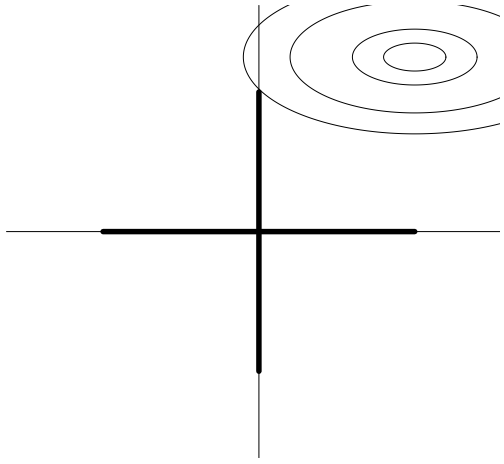


$||\beta||_{\frac{3}{2}} \leq t$

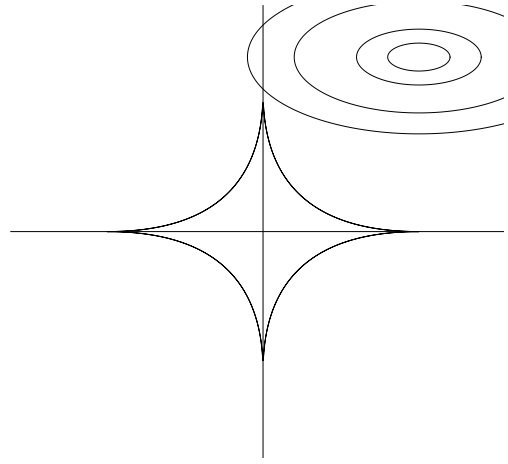


$||\beta||_2 \leq t$

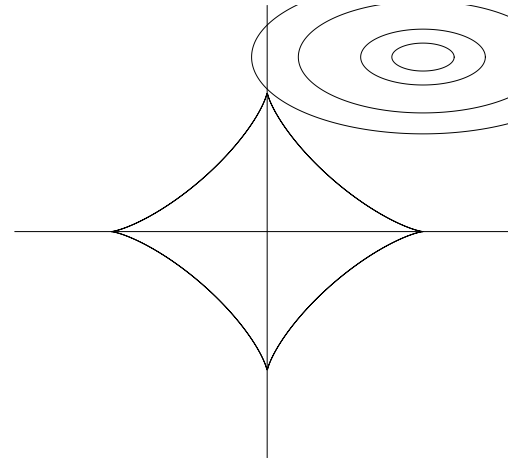
GEOMETRY OF REGULARIZATION IN $\mathbb{R}^2$



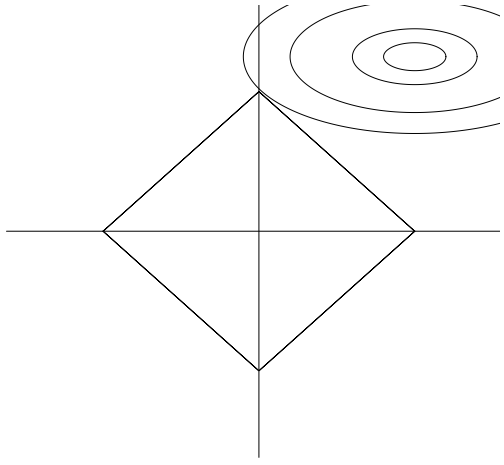
$||\beta||_0 \leq t$



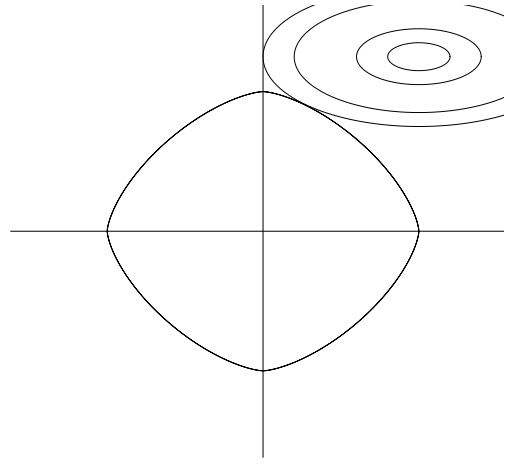
$||\beta||_{\frac{1}{2}} \leq t$



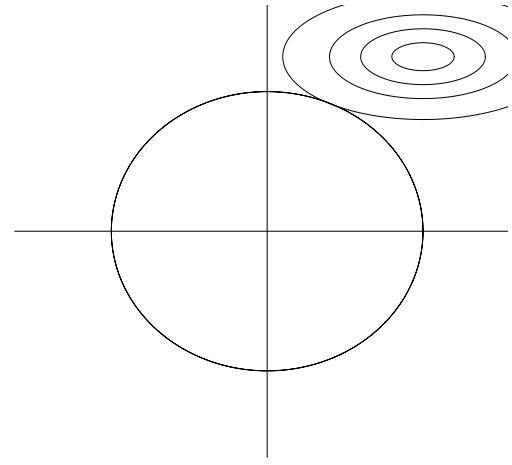
$||\beta||_{\frac{3}{4}} \leq t$



$||\beta||_1 \leq t$

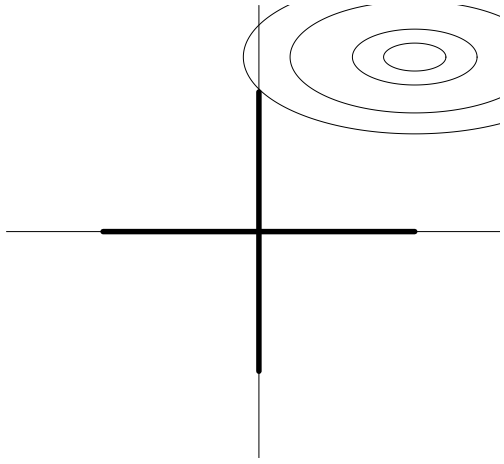


$||\beta||_{\frac{3}{2}} \leq t$

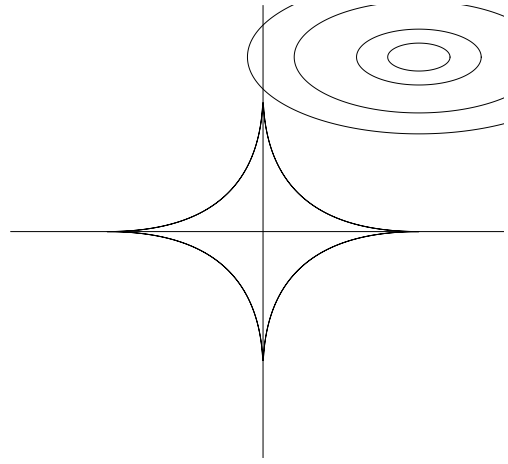


$||\beta||_2 \leq t$

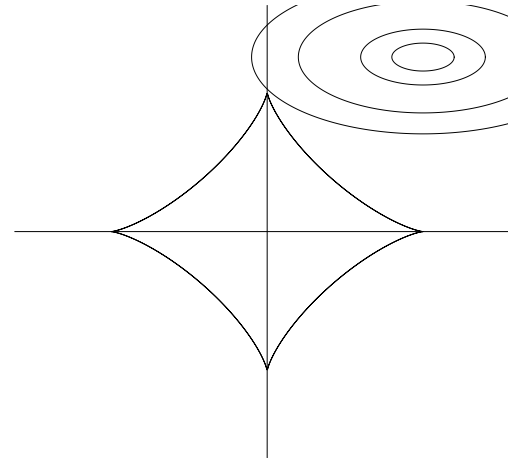
GEOMETRY OF REGULARIZATION IN $\mathbb{R}^2$



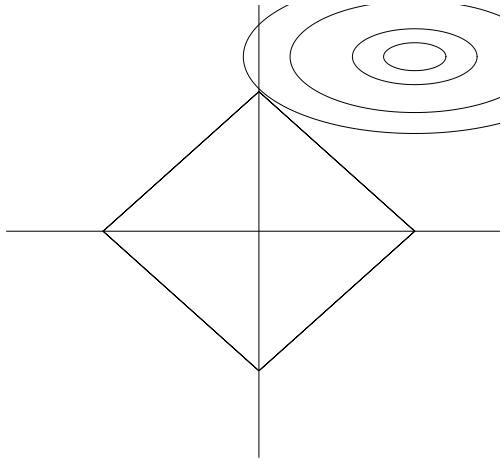
$||\beta||_0 \leq t$



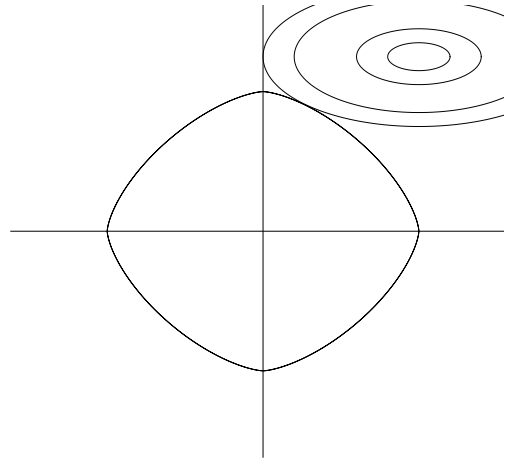
$||\beta||_{\frac{1}{2}} \leq t$



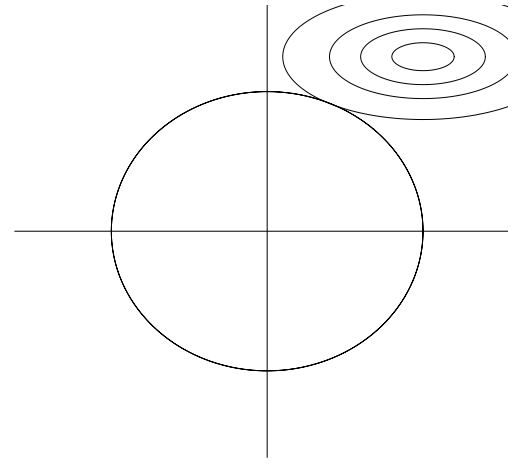
$||\beta||_{\frac{3}{4}} \leq t$



$||\beta||_1 \leq t$

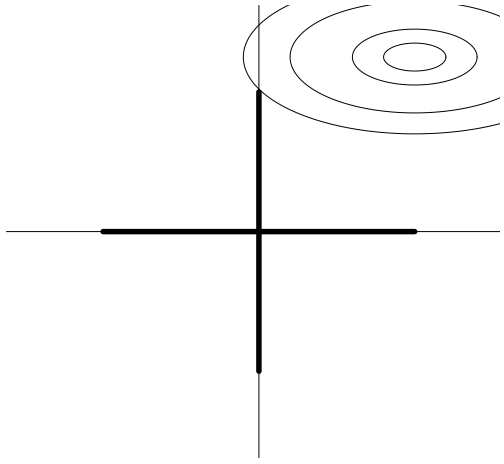


$||\beta||_{\frac{3}{2}} \leq t$

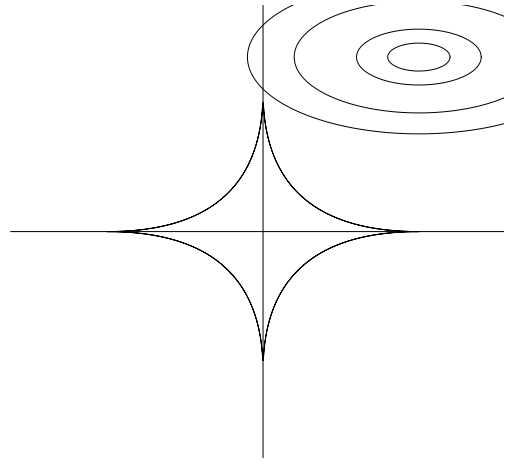


$||\beta||_2 \leq t$

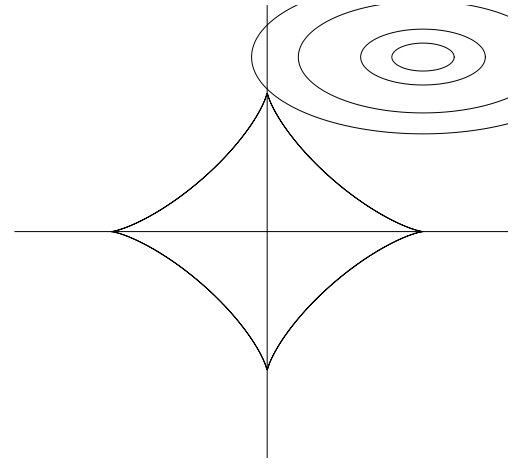
GEOMETRY OF REGULARIZATION IN $\mathbb{R}^2$



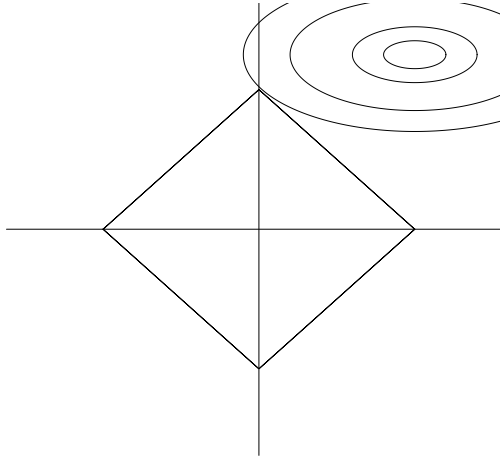
$||\beta||_0 \leq t$



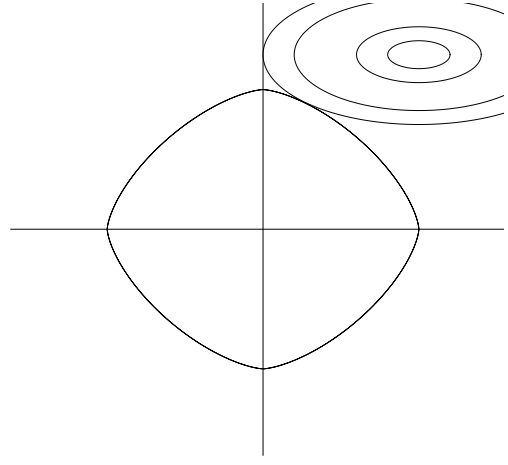
$||\beta||_{\frac{1}{2}} \leq t$



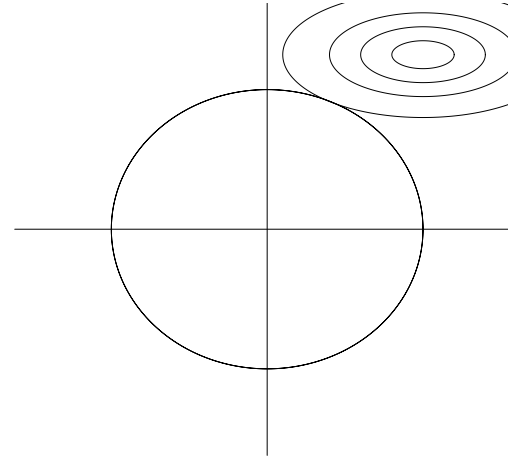
$||\beta||_{\frac{3}{4}} \leq t$



$||\beta||_1 \leq t$

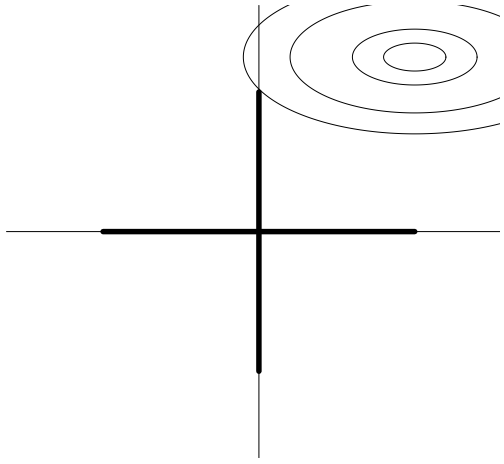


$||\beta||_{\frac{3}{2}} \leq t$

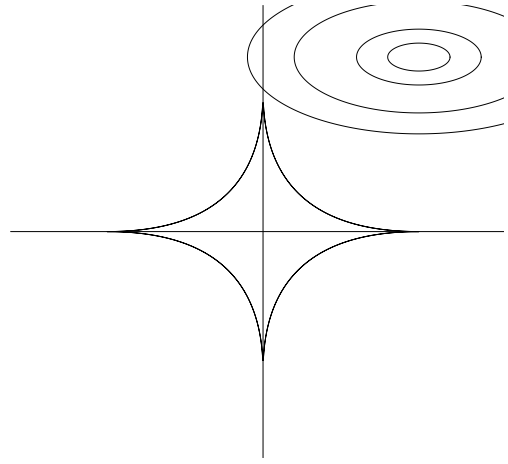


$||\beta||_2 \leq t$

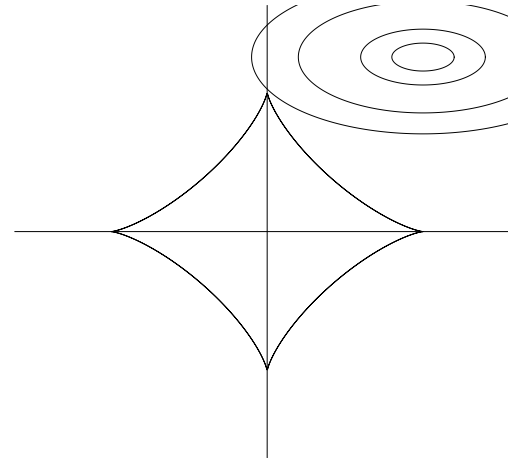
GEOMETRY OF REGULARIZATION IN $\mathbb{R}^2$



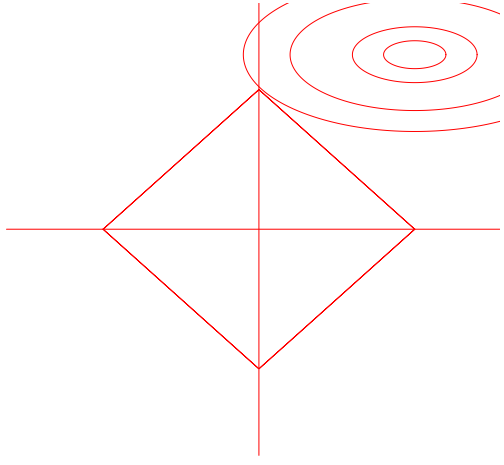
$||\beta||_0 \leq t$



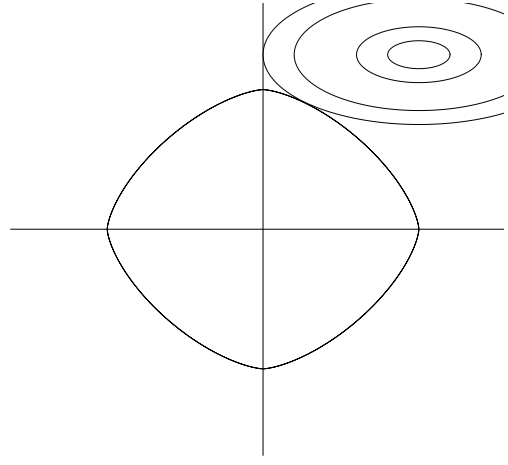
$||\beta||_{\frac{1}{2}} \leq t$



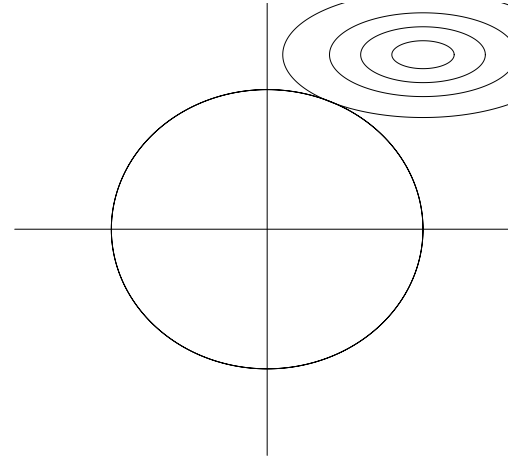
$||\beta||_{\frac{3}{4}} \leq t$



$||\beta||_1 \leq t$



$||\beta||_{\frac{3}{2}} \leq t$

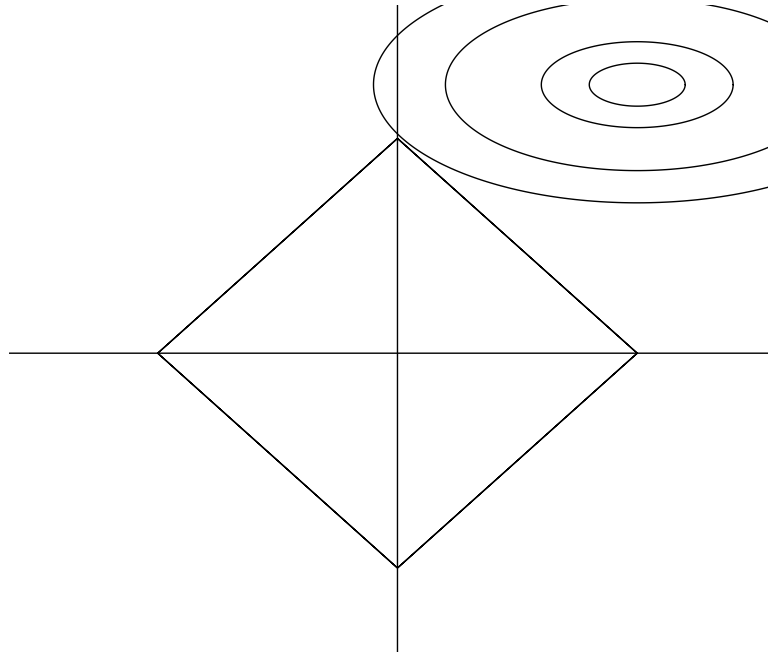


$||\beta||_2 \leq t$

SUMMARY

	CONVEX?	CORNERS?	
$\|\|\beta\|\|_0$	No	Yes	
$\|\|\beta\|\|_{\frac{1}{2}}$	No	Yes	
$\|\|\beta\|\|_{\frac{3}{4}}$	No	Yes	
$\|\|\beta\|\|_1$	Yes	Yes	✓
$\|\|\beta\|\|_{\frac{3}{2}}$	Yes	No	
$\|\|\beta\|\|_2$	Yes	No	

THE BEST OF BOTH WORLDS: $||\beta||_1$



This regularization set...

... is convex (computationally efficient)

... has corners (performs model selection)

The lasso

ℓ_1-REGULARIZED REGRESSION

Known as

- 'lasso'
- 'basis pursuit'

The estimator satisfies

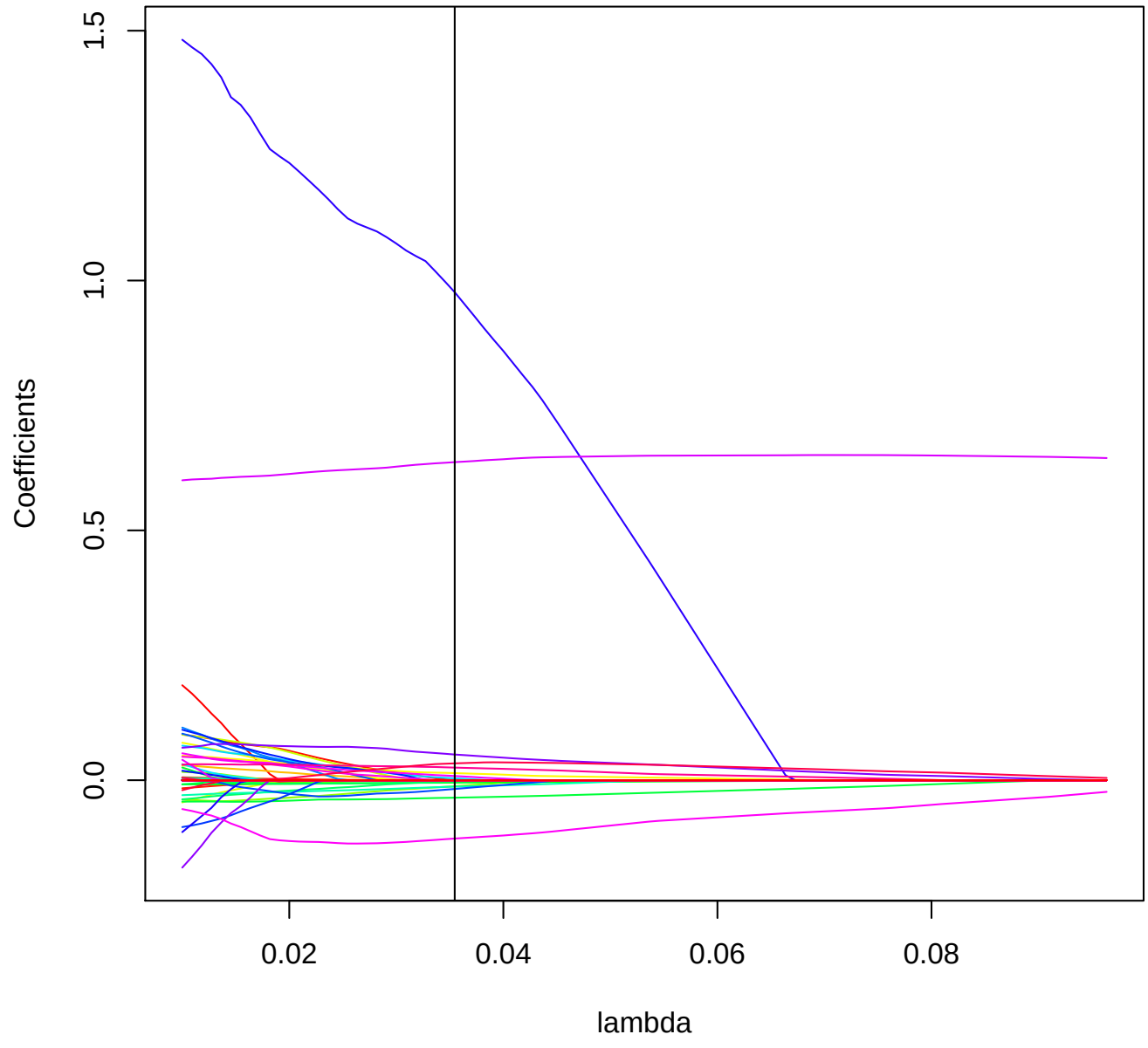
$$\widehat{\beta}_{lasso}(t) = \underset{\beta}{\text{argmin}} \, ||\mathbb{Y} - \mathbb{X}\beta||_2^2 \text{ subject to } ||\beta||_1 \leq t$$

In its corresponding Lagrangian dual form:

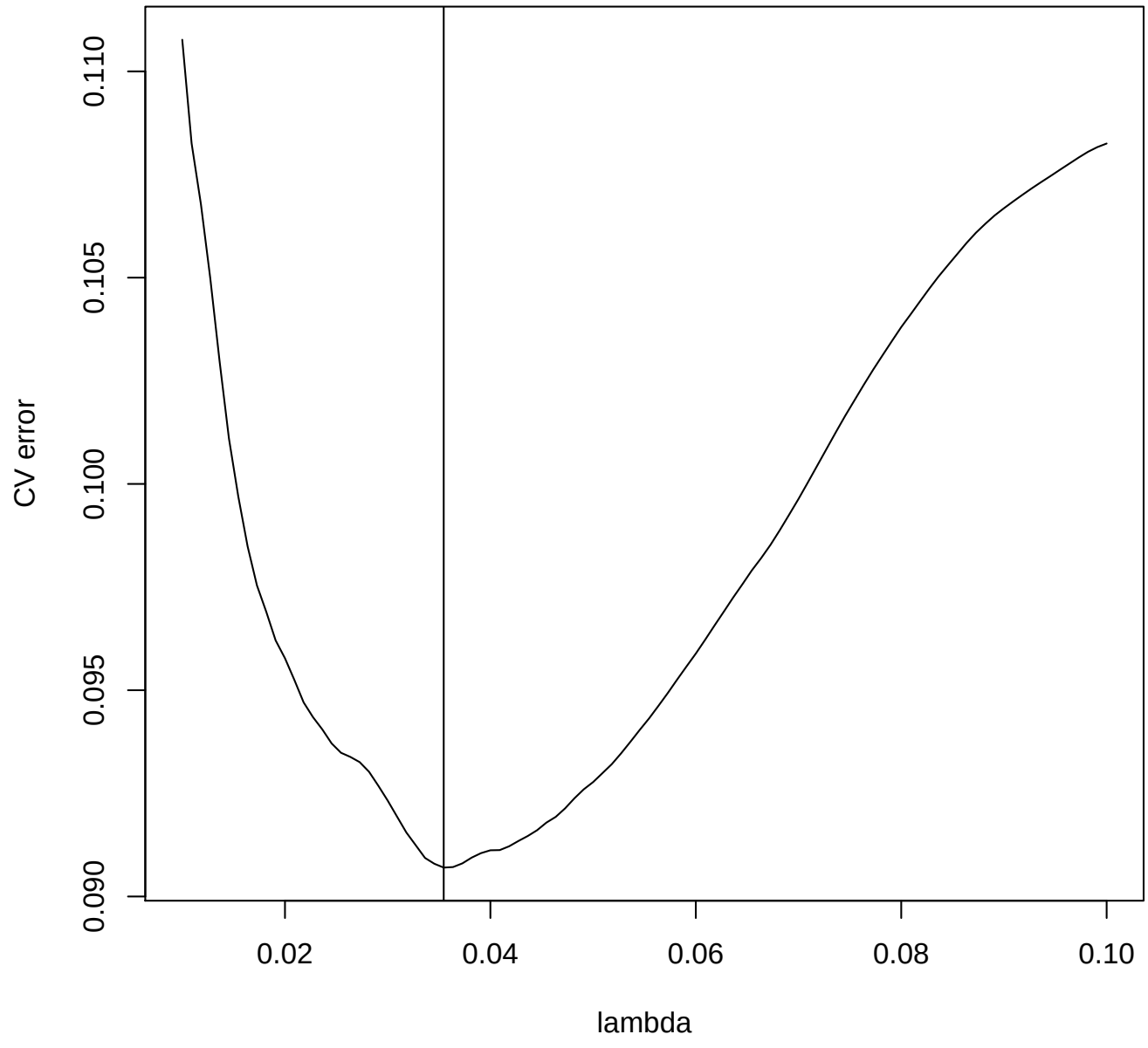
$$\widehat{\beta}_{lasso}(\lambda) = \underset{\beta}{\text{argmin}} \, ||\mathbb{Y} - \mathbb{X}\beta||_2^2 + \lambda||\beta||_1$$

Of course, just like in Ridge, we need a way of choosing this smoothing parameter. We'll just use cross-validation again (although this is still an area of active research).

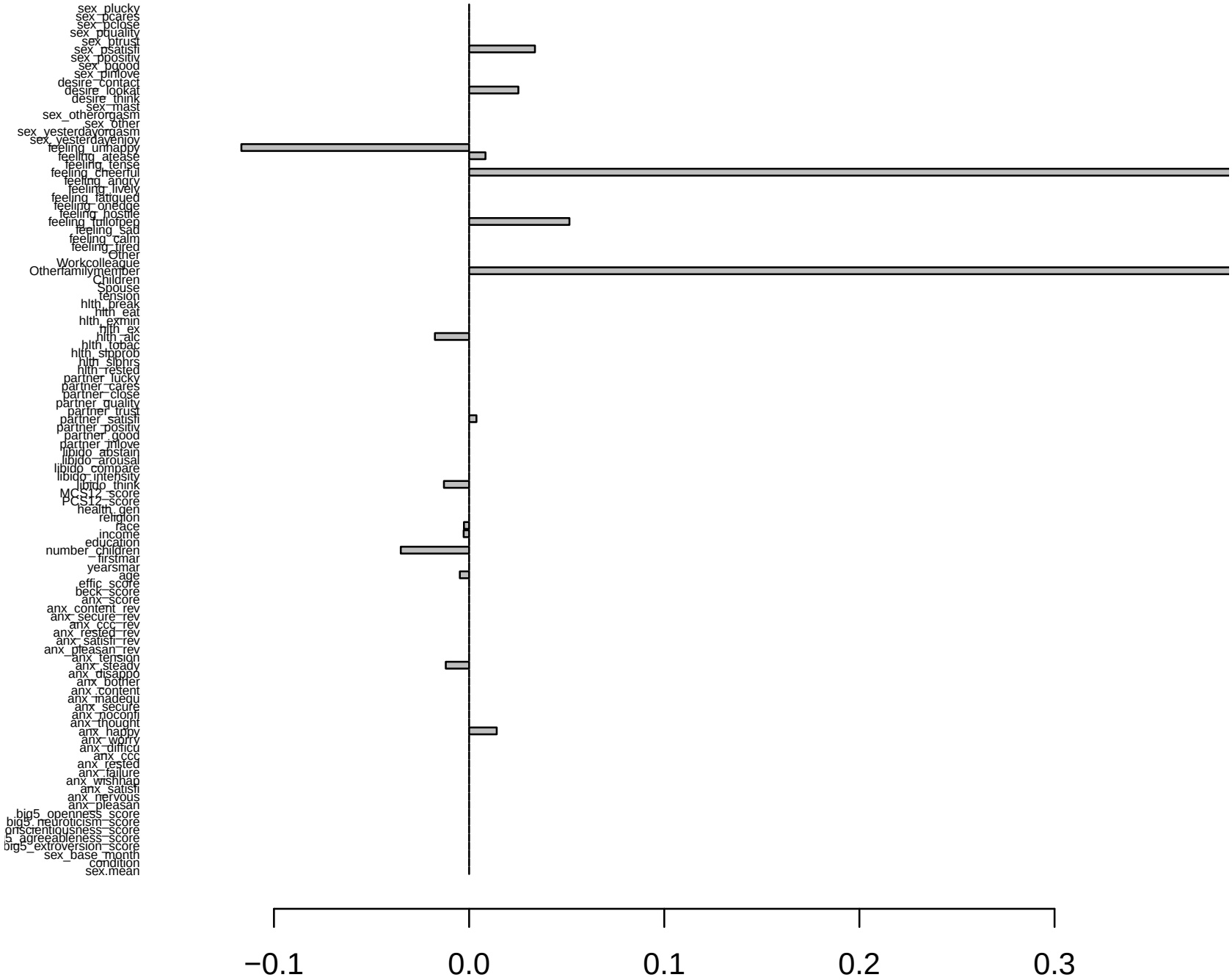
LASSO RESULTS: REGULARIZATION PATH



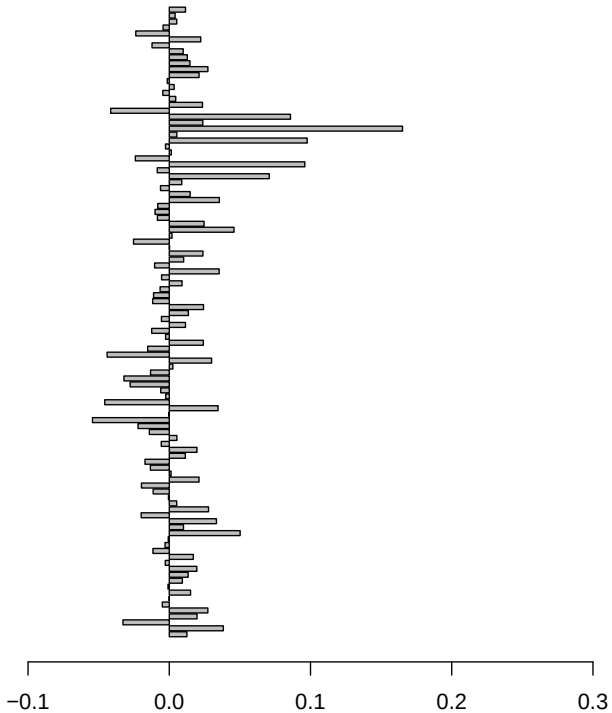
LASSO RESULTS: CROSS-VALIDATION CURVE



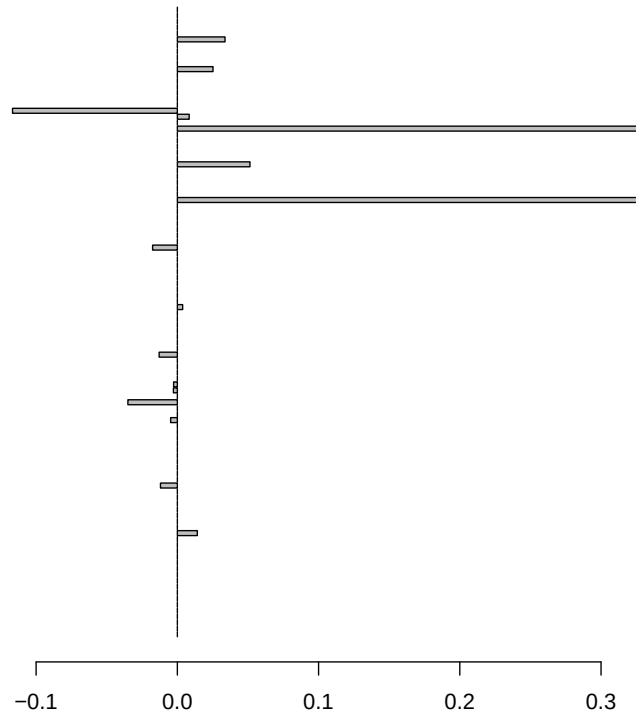
LASSO RESULTS: COEFFICIENTS OF SELECTED MODEL



COMPARISON OF RIDGE AND LASSO SOLUTIONS



Ridge



Lasso

USING LASSO FOR INFERENCE

1. Find the cross-validation minimizing $\widehat{\lambda}$
2. Form $\widehat{\beta}_{lasso}(\widehat{\lambda})$
3. Identify non-zero entries
4. Re-fit a linear model with only these non-zero entries as explanatory variables

Warning: The resulting standard errors, t-statistics, and P-values are no longer valid.

USING LASSO FOR INFERENCE

1. Find the cross-validation minimizing $\widehat{\lambda}$
2. Form $\widehat{\beta}_{lasso}(\widehat{\lambda})$
3. Identify non-zero entries
4. Re-fit a linear model with only these non-zero entries as explanatory variables

Warning: The resulting standard errors, t-statistics, and P-values are no longer valid.

LASSO INFERENCE RESULTS

	Estimate
anxiety1	0.036
anxiety2	-0.065
age	-0.009
number of children	-0.050
income	-0.011
race	-0.024
mentally aroused	-0.032
partner satisfied?	0.041
unhealthy (alcohol)	-0.081
other family member	1.902
feeling full of pep	0.097
feeling cheerful	0.595
feeling at ease	0.048
feeling unhappy	-0.163
desire when looking at partner	0.044
partner sexually satisfied?	0.011

LASSO INFERENCE RESULTS

	Estimate
anxiety1	0.036
anxiety2	-0.065
age	-0.009
number of children	-0.050
income	-0.011
race	-0.024
mentally aroused	-0.032
partner satisfied?	0.041
unhealthy (alcohol)	-0.081
other family member	1.902
feeling full of pep	0.097
feeling cheerful	0.595
feeling at ease	0.048
feeling unhappy	-0.163
desire when looking at partner	0.044
partner sexually satisfied?	0.011

LASSO INFERENCE RESULTS

	Estimate
anxiety1	0.036
anxiety2	-0.065
age	-0.009
number of children	-0.050
income	-0.011
race	-0.024
mentally aroused	-0.032
partner satisfied?	0.041
unhealthy (alcohol)	-0.081
other family member	1.902
feeling full of pep	0.097
feeling cheerful	0.595
feeling at ease	0.048
feeling unhappy	-0.163
desire when looking at partner	0.044
partner sexually satisfied?	0.011

LASSO INFERENCE RESULTS

	Estimate
anxiety1	0.036
anxiety2	-0.065
age	-0.009
number of children	-0.050
income	-0.011
race	-0.024
mentally aroused	-0.032
partner satisfied?	0.041
unhealthy (alcohol)	-0.081
other family member	1.902
feeling full of pep	0.097
feeling cheerful	0.595
feeling at ease	0.048
feeling unhappy	-0.163
desire when looking at partner	0.044
partner sexually satisfied?	0.011

LASSO INFERENCE RESULTS

	Estimate
anxiety1	0.036
anxiety2	-0.065
age	-0.009
number of children	-0.050
income	-0.011
race	-0.024
mentally aroused	-0.032
partner satisfied?	0.041
unhealthy (alcohol)	-0.081
other family member	1.902
feeling full of pep	0.097
feeling cheerful	0.595
feeling at ease	0.048
feeling unhappy	-0.163
desire when looking at partner	0.044
partner sexually satisfied?	0.011

SUMMARY

- For high dimensional linear regression, we face a bias-variance tradeoff: omitting too many variables causes bias while including too many variables causes high variance.
- The key is to select a good subset of variables.
- The lasso (ℓ_1-regularized least squares) is a fast way to select variables.
- If there are good, sparse, linear predictors, the lasso will work well.
- More generally, whati if $\mathbb{X}$ is low-rank? (more on this in the 'Dimension reduction' slides).

Encoding additional prior information

There are many generalizations to classical lasso:

Elastic Net

$$\min ||\mathbb{Y} - \mathbb{X}\beta||_2^2 + \lambda_1 ||\beta||_1 + \lambda_2 ||\beta||_2^2$$

Generalized Lasso

$$\min ||\mathbb{Y} - \mathbb{X}\beta||_2^2 + \lambda ||D\beta||_1$$

Group Lasso

$$\beta = (\underbrace{\beta_1, \ldots, \beta_k}_{g_1}, \ldots, \underbrace{\beta_l, \ldots, \beta_p}_{g_G})$$

$$\min ||\mathbb{Y} - \mathbb{X}\beta||_2^2 + \lambda \sum_{j=1}^{G} \sqrt{|g_j|}\, ||g_j||_2$$

Up next:
Predicting the macroeconomy