Computational Data Analysis (02582)

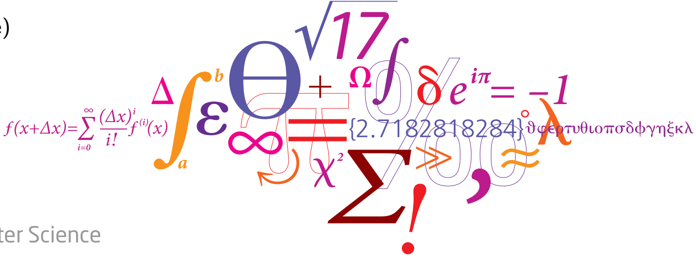
SPARSE REGRESSION

Sneha Das

Assistant Professor

Department of Applied Mathematics and Computer Science,

Technical University of Denmark (DTU Compute)



DTU Compute
Department of Applied Mathematics and Computer Science

Outline I

- Part I: Recap and outstanding questions
- Part II: Assumptions as models
  - Curse of dimensionality
  - Linear models and explicit regulations

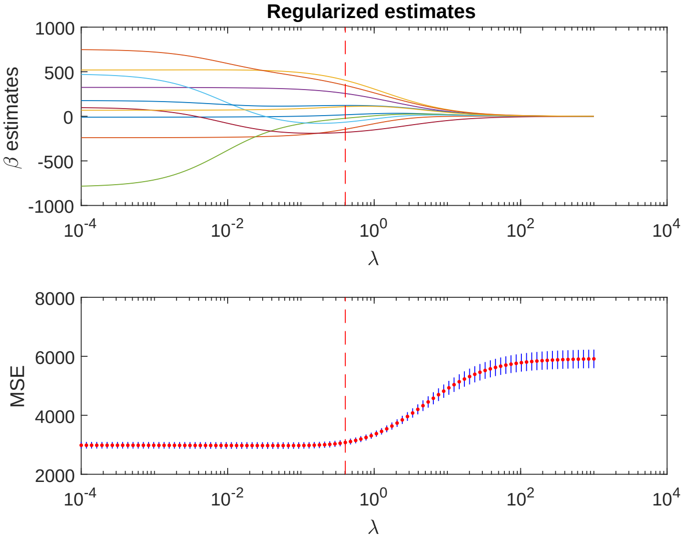
    Ridge Regression

    Bias from regularization
  - Sparsity and high-dimensional inference

    LASSO
  - Elastic Net
  - Variable selection instability

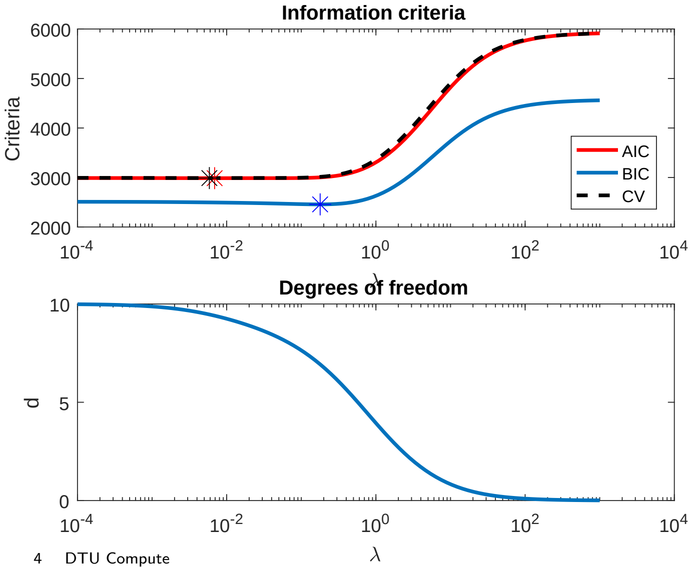
    Multiple hypothesis testing
- Exercise 3: Multiple Testing

1-SE rule



Why 1-SE rule?! Breiman, Friedman, Olshen, Stone (1984), Classification and Regression Trees (CART) monograph *'argument for the 1 SE rule is that in simulation studies it yields a stable tree size across replications whereas the 0 SE tree size can vary substantially across replications.'*

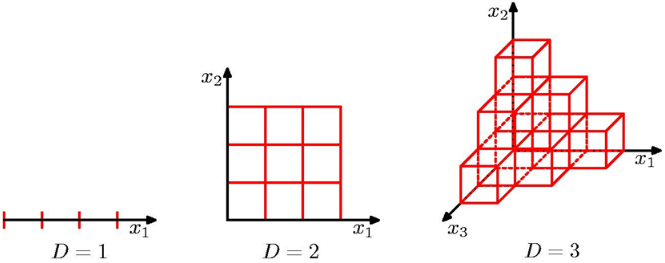
Methods for Model selection



BIC grows with number of samples (N)?! *the 'more data' → 'less penalty' idea is aligned with 'I can afford more complexity for prediction' which is AIC-ish.*
BIC is instead 'with more data, I can confidently detect whether extra parameters are truly warranted', which requires a penalty that increases (even slowly).

Curse of dimensionality

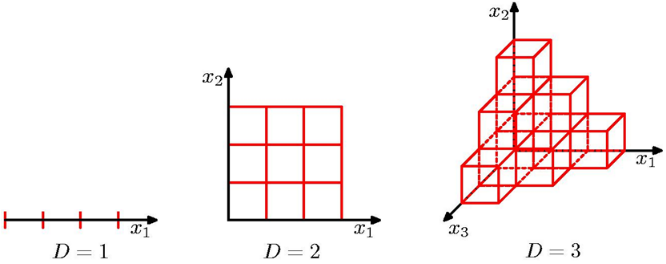
What happens when the dimension of the solution space grows, ie the number of variables grows?



Curse of dimensionality

What happens when the dimension of the solution space grows, ie the number of variables grows?

- The number of regions grows exponentially with the dimensionality D



Think-Pair-Share: The Curse of Dimensionality

Context: In high-dimensional settings ($p \gg N$), our intuition about space and distance often fails.

Task (3 min):

1. **Think:** What specific problems arise for learning algorithms when the number of features p becomes very large?
2. **Pair:** Discuss your identified problems with your neighbor.
3. **Share:** Enter your keywords into the Vevox word cloud.

Go to Vevox and enter text...

Solution: The Curse Manifests in Many Ways

- **Sparsity:** Data becomes incredibly sparse; 'local' neighborhoods become empty.
- **Distances:** Euclidean distances lose meaning; points become roughly equidistant.
- **Overfitting:** With $p > N$, models can perfectly fit noise (degrees of freedom issues).
- **Edge Effect:** Most data points reside at the boundaries (corners) of the sample space.
- **Computational Cost:** Search algorithms slow down significantly.

Part II: Assumptions as models

Blessings of dimensionality

It's not all bad...
In 2000, Donoho pinpointed **3 blessings of dimensionality.**

1. Several features will be correlated and we can average over them

Donoho, D. L., August 2000. High-dimensional data analysis: The curses and blessings of dimensionality. In: Conf. Math Challenges of the 21st Century, Los Angeles.

Blessings of dimensionality

It's not all bad...
In 2000, Donoho pinpointed **3 blessings of dimensionality.**

1. Several features will be correlated and we can average over them
2. Underlying distribution will be finite, informative data will lie on a low-dimensional manifold

Donoho, D. L., August 2000. High-dimensional data analysis: The curses and blessings of dimensionality. In: Conf. Math Challenges of the 21st Century, Los Angeles.

Blessings of dimensionality

It's not all bad...
In 2000, Donoho pinpointed **3 blessings of dimensionality.**

1. Several features will be correlated and we can average over them
2. Underlying distribution will be finite, informative data will lie on a low-dimensional manifold
3. Underlying structure in data (samples from continuous processes, images etc) will give an approximate finite dimensionality.

Donoho, D. L., August 2000. High-dimensional data analysis: The curses and blessings of dimensionality. In: Conf. Math Challenges of the 21st Century, Los Angeles.

Dimension reduction

How to decrease the dimension and identify the most important variables, and get rid of the redundant or irrelevant variables.

Dimension reduction

- Regularization of parameters
  - Focus of today
- Combinatoric search, forward and backward selection
  - Previous courses - we make a recap and talk about multiple hypothesis testing
- Projection to lower dimensions - latent variables
  - Coming lectures, PCA, Unsupervised decomposition and Multi-way models
- Clustering of features
  - Lecture on Clustering
- Structuring parameter estimates
  - Related to regularization

Norms of β

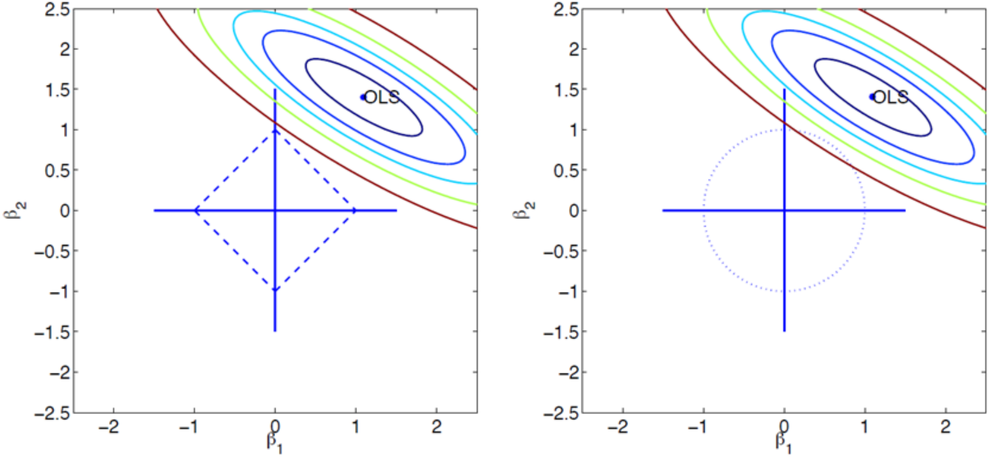
What is the definition of the L_2-norm,

$$||\beta||_2^2 =$$

What is the definition of the L_1-norm

$$||\beta||_1 =$$

Regularization



Shrinkage methods

Instead of controlling model complexity by setting a subset of coefficients to zero we can **shrink** all the coefficients some way towards zero.
Three established standard techniques

- **Ridge** regression uses quadratic shrinkage, L_2-norm
- **Lasso** regression uses absolute-value shrinkage, L_1-norm
- **Elastic net** which is a hybrid method

Ridge regression

Ridge regression solves

$$\min_{\beta}(Y - X\beta)^T(Y - X\beta) + \lambda\beta^T\beta$$

or equivalently the constrained optimization problem

$$\min_{\beta}(Y - X\beta)^T(Y - X\beta) \text{ subject to } \sum \beta_j^2 \leq s$$

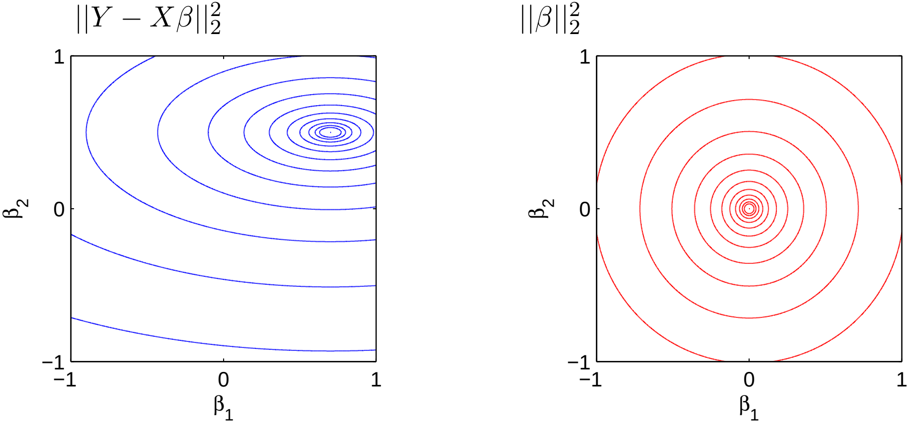
- Increased λ will make the estimated β's smaller but not exactly zero.
- We typically do not penalize the intercept β_0

Ridge regression optima

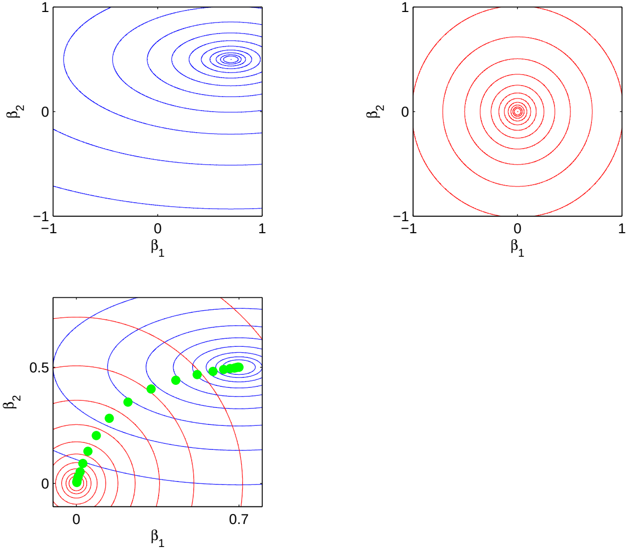
Optimization of a weighted sum

$$\beta_{Ridge} = \arg\min_{\beta} ||Y - X\beta||_2^2 + \lambda||\beta||_2^2$$

Contour plots of



Regularization path



The Lasso

The Lasso regression solves

$$\min_{\beta}(Y - X\beta)^T(Y - X\beta) + \lambda||\beta||_1$$

or equivalently the constrained optimization problem (known as basis pursuit)

$$\min_{\beta}(Y - X\beta)^T(Y - X\beta) \text{ subject to } \sum|\beta| \leq s$$

- Notice that the L_2-penalty is replaced by a L_1-penalty.

The Lasso

The Lasso regression solves

$$\min_{\beta}(Y - X\beta)^T(Y - X\beta) + \lambda||\beta||_1$$

or equivalently the constrained optimization problem (known as basis pursuit)

$$\min_{\beta}(Y - X\beta)^T(Y - X\beta) \text{ subject to } \sum|\beta| \leq s$$

- Notice that the L_2-penalty is replaced by a L_1-penalty.
- This makes the solution non-differentiable at 0, and a quadratic programming algorithm can be used to solve it.

The Lasso

The Lasso regression solves

$$\min_{\beta}(Y - X\beta)^T(Y - X\beta) + \lambda||\beta||_1$$

or equivalently the constrained optimization problem (known as basis pursuit)

$$\min_{\beta}(Y - X\beta)^T(Y - X\beta) \text{ subject to } \sum|\beta| \leq s$$

- Notice that the L_2-penalty is replaced by a L_1-penalty.
- This makes the solution non-differentiable at 0, and a quadratic programming algorithm can be used to solve it.
- For large enough λ some of the β will be set to **exactly zero**.

The Lasso

The Lasso regression solves

$$\min_{\beta}(Y - X\beta)^T(Y - X\beta) + \lambda||\beta||_1$$

or equivalently the constrained optimization problem (known as basis pursuit)

$$\min_{\beta}(Y - X\beta)^T(Y - X\beta) \text{ subject to } \sum|\beta| \leq s$$

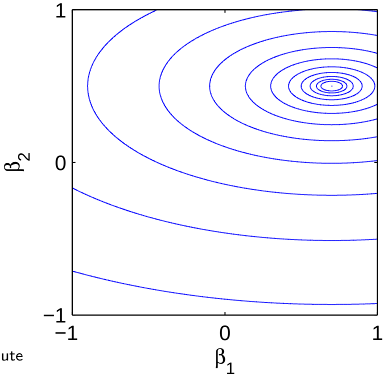
- Notice that the L_2-penalty is replaced by a L_1-penalty.
- This makes the solution non-differentiable at 0, and a quadratic programming algorithm can be used to solve it.
- For large enough λ some of the β will be set to **exactly zero**.
- The effective numbers of parameters, df, equals the number of coefficients different from zero.

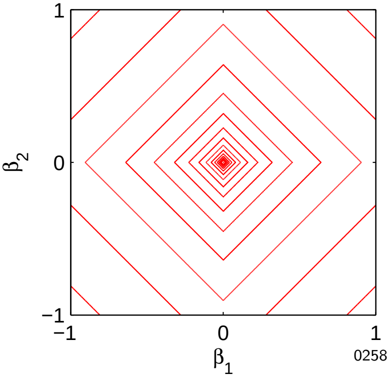
Lasso regularization

- Lasso regularization will gear parameters towards zero.

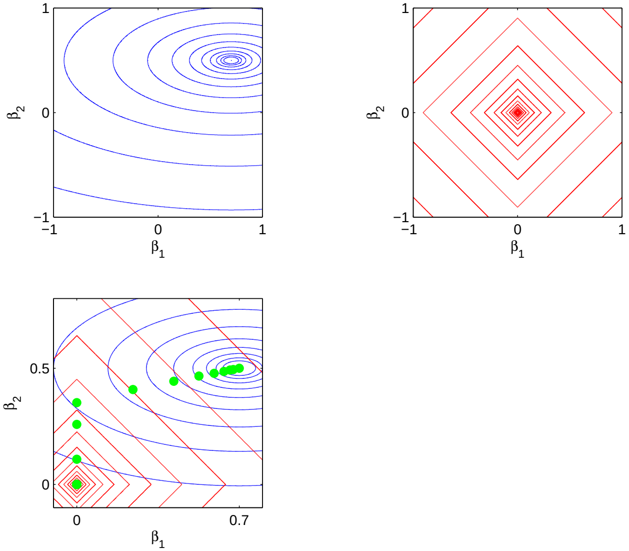
$$\begin{aligned}\hat{\beta}_{Lasso} &= \arg\min_{\beta} ||Y - X\beta||_2^2 + \lambda||\beta||_1 \\ &= \arg\min_{\beta} ||Y - X\beta||_2^2 + \lambda\sum_i |\beta_i|\end{aligned}$$

- Non-trivial optimization problem...



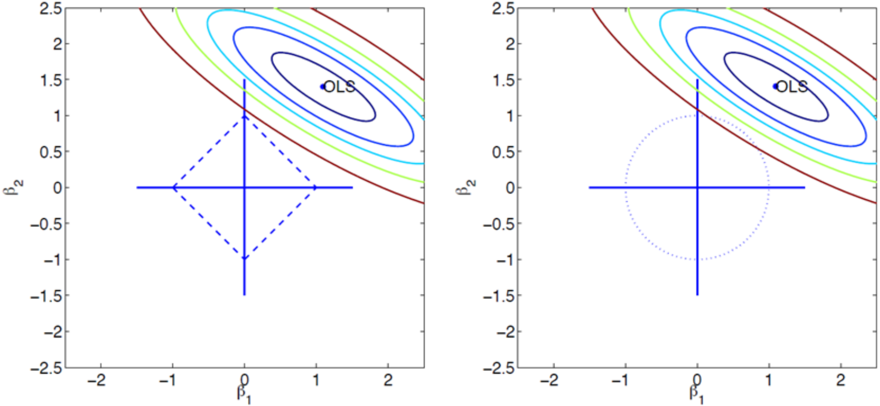


Regularization path



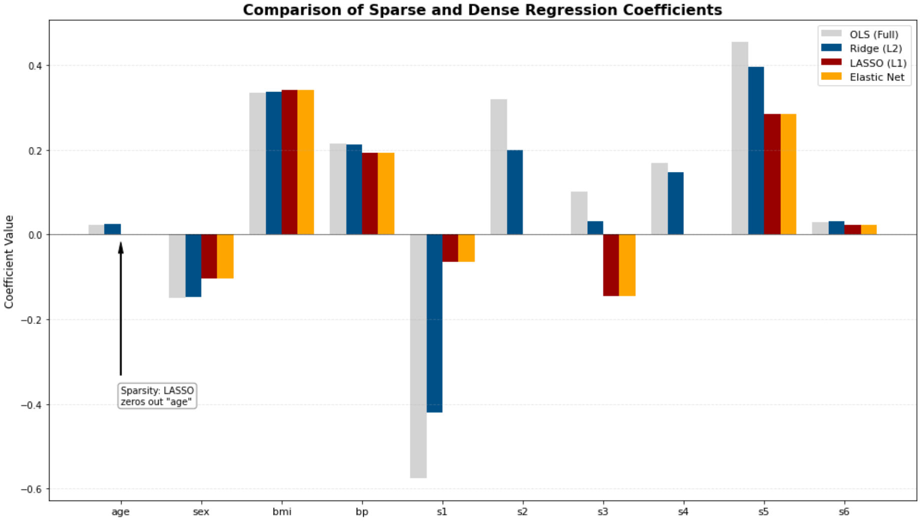
Geometry of solutions with L_1 and L_2 penalties

Visual solution to the constrained optimization problems for lasso and ridge,



The elastic net example - Diabetes dataset

Target variable is a quantitative measure of disease progression one year after baseline.



Algorithms for Lasso

There exist several implementations to solve the Lasso problem. We will take a closer look at two

- Least angle regression selection (LARS)
- Cyclical coordinate descent

Least Angle Regression (LARS)

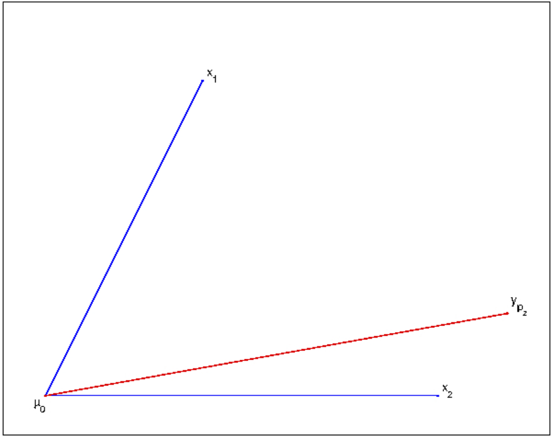
LARS is the computational "engine" for finding the LASSO/Elastic Net path.

The Algorithm:

1. Start with all $\beta = 0$. Find variable x_j most correlated with y.
2. Move β_j in the direction of its least-squares coefficient.
3. Stop when another variable x_k has as much correlation with the current residual as x_j.
4. Move in a direction equiangular to both x_j and x_k.

LASSO is a modification of LARS where a variable is dropped if its coefficient hits zero.

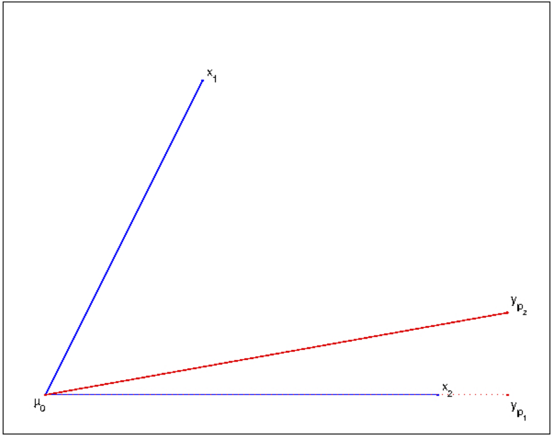
Least angle regression selection (LARS) - Algorithm



$\boldsymbol{\mu}_0 = \mathbf{0}$

$\boldsymbol{\mu}$ is the current estimate
$\mathbf{x}_i$ is the ith coordinate
$\mathbf{y}_p$ is the projection of $\mathbf{y}$ into $\{\mathbf{x}\}$

Least angle regression selection (LARS) - Algorithm

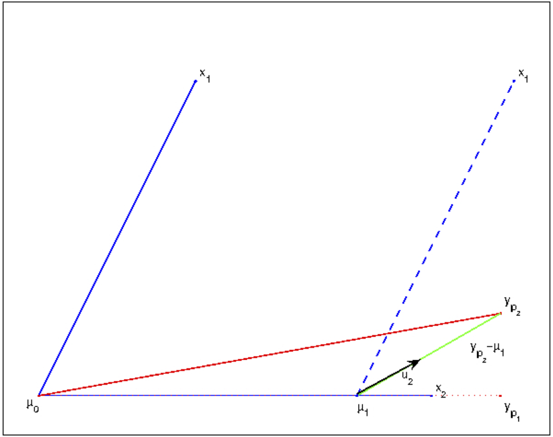


$$\boldsymbol{\mu}_0 = \mathbf{0}$$
$$\boldsymbol{\mu}_1 = \boldsymbol{\mu}_0 + \gamma_1 \mathbf{x}_2$$

$\boldsymbol{\mu}$ is the current estimate
$\mathbf{x}_i$ is the ith coordinate
$\mathbf{y}_p$ is the projection of $\mathbf{y}$ into $\{\mathbf{x}\}$

What should the size of the parameter γ_1 be?

Least angle regression selection (LARS) - Algorithm



$$\boldsymbol{\mu}_0 = \mathbf{0}$$
$$\boldsymbol{\mu}_1 = \boldsymbol{\mu}_0 + \gamma_1 \mathbf{x}_2$$
$$\boldsymbol{\mu}_2 = \boldsymbol{\mu}_1 + \gamma_2 \mathbf{x}_1$$

$\boldsymbol{\mu}$ is the current estimate
$\mathbf{x}_i$ is the ith coordinate
$\mathbf{y}_p$ is the projection of $\mathbf{y}$ into $\{\mathbf{x}\}$

γ_1 is chosen such that the residual bisects the angle between $\mathbf{x}_1$ and $\mathbf{x}_2$ (*the equiangular direction*)

The Intuition: Greedy vs. Polite

Forward Selection (Greedy)

- Finds the single best variable.
- Moves along it *completely* until it can't improve anymore.
- Only then looks for a new variable.
- **Result:** Aggressive, jerky path.

LARS (Polite)

- Finds the single best variable.
- Moves along it **only until** a second variable becomes *equally helpful*.
- Then moves in a joint direction (bisecting the angle).
- **Result:** Efficient, equiangular path.

Step 1: Initialization & First Correlation

Theory: Initialization

1. Start with coefficients $\beta = 0$.

$$\mu_0 = 0, \quad r = y - \mu_0 = y$$

2. Calculate correlations:

$$c = X^T r$$

3. Find variable x_j with max absolute correlation.

Example (2D Data)

$$y = \begin{bmatrix} 2 \\ 1 \end{bmatrix}, x_1 = \begin{bmatrix} 1 \\ 0 \end{bmatrix}, x_2 = \begin{bmatrix} 0.5 \\ 0.866 \end{bmatrix}$$

Calculations:

$c_1 = x_1^T y = 1(2) + 0(1) =$ **2.0 (Max)**

$c_2 = x_2^T y = 0.5(2) + 0.866(1) = 1.866$

Action: We start moving along x_1.

Step 2: Calculating the Step Size (γ)

Theory: How far do we go?

We move along x_j until the residual is **equally correlated** with a new variable x_k.

The formula for 2 variables:

$$\gamma = \frac{c_j - c_k}{1 - \rho_{jk}}$$

- c_j, c_k: Current correlations
- ρ_{jk}: Correlation between x_j and x_k

Example Calculation

Knowns:

$c_1 = 2.0, \quad c_2 = 1.866$
$\rho_{12} = x_1^T x_2 = 0.5$

Plug into formula:

$$\gamma = \frac{2.0 - 1.866}{1 - 0.5}$$

$$\gamma = \frac{0.134}{0.5}$$

$$\gamma = \mathbf{0.268}$$

We move 0.268 units along x_1.

Step 3: Update & Verify

Theory: The Update

1. Update the prediction vector:

$$\mu_{new} = \mu_{old} + \gamma x_j$$

2. Update residuals:

$$r_{new} = y - \mu_{new}$$

3. Verify that correlations are now equal (in absolute value).

Example Verification

Update:

$$\mu_{new} = 0 + 0.268 \begin{bmatrix} 1 \\ 0 \end{bmatrix} = \begin{bmatrix} 0.268 \\ 0 \end{bmatrix}$$

New Residual:

$$r_{new} = \begin{bmatrix} 2 \\ 1 \end{bmatrix} - \begin{bmatrix} 0.268 \\ 0 \end{bmatrix} = \begin{bmatrix} 1.732 \\ 1 \end{bmatrix}$$

Check Correlations:

$c_1 = 1(1.732) + 0(1) = \mathbf{1.732}$

$c_2 = 0.5(1.732) + 0.866(1) \approx \mathbf{1.732}$

Success! x_2 now enters the model.

Step 4:

LARS moves along $u_{\mathcal{A}}$ - equiangular direction, until a 3rd variable becomes equally correlated (or residuals hit 0). Why equiangular?

Computing the Equiangular Vector

Theory: Matrix Formula To move equally between variables in active set $\mathcal{A}$, we calculate vector $u_{\mathcal{A}}$.

$$u_{\mathcal{A}} = X_{\mathcal{A}} w$$

$$w = A(X_{\mathcal{A}}^T X_{\mathcal{A}})^{-1} \mathbf{1}$$

- $X_{\mathcal{A}}^T X_{\mathcal{A}}$: The Gram (correlation) matrix.
- A: Normalization factor so $||u_{\mathcal{A}}|| = 1$.

Example: The Bisector For 2 vectors, u is simply the normalized sum:

$$u = \frac{x_1 + x_2}{||x_1 + x_2||}$$

Calculation:

Sum: $v = \begin{bmatrix} 1 \\ 0 \end{bmatrix} + \begin{bmatrix} 0.5 \\ 0.866 \end{bmatrix} = \begin{bmatrix} 1.5 \\ 0.866 \end{bmatrix}$

Norm: $||v|| = \sqrt{1.5^2 + 0.866^2} = \sqrt{3} \approx 1.732$

Result: $u = \frac{1}{1.732} \begin{bmatrix} 1.5 \\ 0.866 \end{bmatrix} = \begin{bmatrix} 0.866 \\ 0.5 \end{bmatrix}$

We now move along this direction.

Least angle regression selection (LARS) - Algorithm

Assumptions: Data is centered and normalized (each variable has length one). This means that: $X^T X \approx Corr(X)$.

Lasso modification: If the parameter estimate of an active variable crosses zero, set it to zero and re-compute the direction.

- Gives a piecewise linear path to obtain lasso solutions for all relevant values of lambda.

Least angle regression selection (LARS)

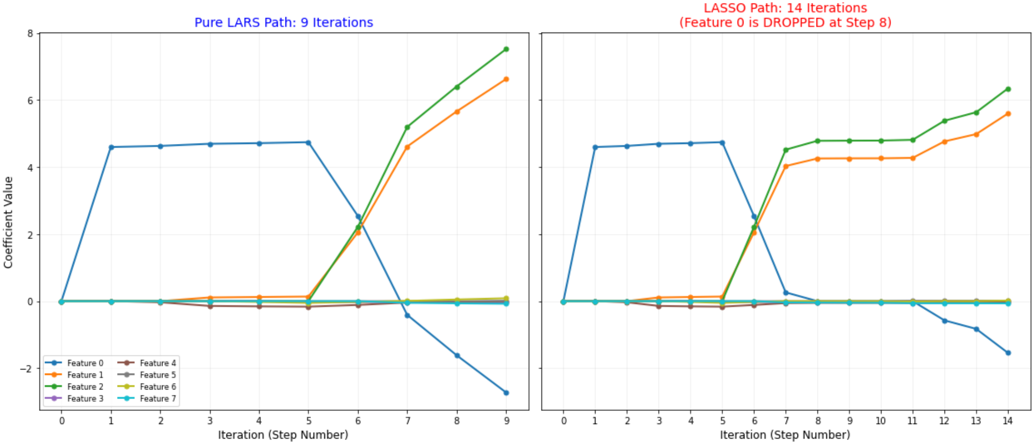
- Fast - calculates the entire path (all λ values) in the speed of one OLS fit.
- Easy to implement, intuitive.
- C_p-like statistic for choosing the number of steps.

$$C_p = \frac{1}{\hat{\sigma}^2} \sum_{i=1}^{n} (y_i - \hat{y}_i)^2 - n + 2k$$

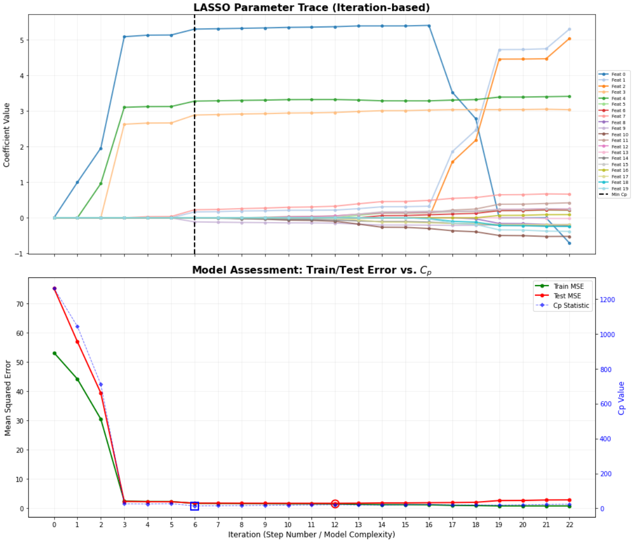
where k is the number of steps.

Hesterberg et al., 2008, Least angle and L1 penalized regression: A review, Statistics Surveys, Vol. 2, p. 61-93.

Parameter trace example



C_p in LARS for Diabetes example



Cyclical coordinate descent

Fix λ. Solve

$$\min_{\beta} \frac{1}{2n} \sum_{i=1}^{n} (y_i - x_i\beta)^2 + \lambda|\beta|$$

iteratively by cyclic updating one coordinate β_j at a time, while holding the others fixed in the current estimate $\tilde{\beta}_k$.

Cyclical coordinate descent

Fix λ. Solve

$$\min_{\beta} \frac{1}{2n} \sum_{i=1}^{n} (y_i - x_i \beta)^2 + \lambda |\beta|$$

iteratively by cyclic updating one coordinate β_j at a time, while holding the others fixed in the current estimate $\tilde{\beta}_k$.
Compute the partial residual $r_i^{(j)} = y_i - \tilde{y}_i^{(j)}$ for $\tilde{\beta}_k$ excluding parameter $\tilde{\beta}_j$,

$$r_i^{(j)} = y_i - \sum_{k \neq j} x_{ik} \tilde{\beta}_k(\lambda)$$

Cyclical coordinate descent

Fix λ. Solve

$$\min_{\beta} \frac{1}{2n} \sum_{i=1}^{n} (y_i - x_i \beta)^2 + \lambda |\beta|$$

iteratively by cyclic updating one coordinate β_j at a time, while holding the others fixed in the current estimate $\tilde{\beta}_k$.

Compute the partial residual $r_i^{(j)} = y_i - \tilde{y_i}^{(j)}$ for $\tilde{\beta}_k$ excluding parameter $\tilde{\beta}_j$,

$$r_i^{(j)} = y_i - \sum_{k \neq j} x_{ik} \tilde{\beta}_k(\lambda)$$

Calculate the OLS solution to $r_i^{(j)}$. This is

$$\tilde{\beta}_j^{OLS} = \frac{1}{n} \sum_{i=1}^{n} x_{ij} r_i^{(j)}$$

(Assume standardization $\sum_i x_{ij} = 0$ and $\frac{1}{n} \sum_i x_{ij}^2 = 1$, $j = 1, ..., p$)

Cyclical coordinate descent, cont'd

Obtain the new lasso coordinate $\tilde{\beta}_j$ by shrinking the OLS estimate and set it to zero if it is close to zero,

$$\tilde{\beta}_j(\lambda) = sign(\tilde{\beta}_j^{OLS})(|\tilde{\beta}_j^{OLS}| - \lambda)_+$$

this is called **soft thresholding**.
Cycle through $j = 1, ..., p$ repeatedly until convergence.

Soft thresholding

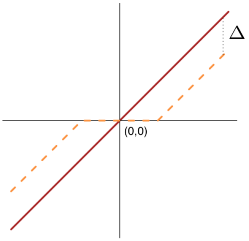


FIGURE 18.2. *Soft thresholding function* $\mathrm{sign}(x)(|x| - \Delta)_+$ *is shown in orange, along with the* 45° *line in red.*

Motivation of EN: Limitations of LASSO

While LASSO is powerful, it has three major limitations:

1. **High Dimensionality:** In the $p > n$ case, the LASSO selects at most n variables.
2. **Grouping Effect:** If there is a group of variables with high pairwise correlation, LASSO tends to arbitrarily select one variable from the group.
3. **Predictive Power:** If $n > p$ and there are high correlations between predictors, Ridge regression tends to outperform LASSO.

Source: Zou and Hastie (2005), "Regularization and variable selection via the elastic net"

The elastic net

By combining the L_1 and the L_2-norm we obtain sparsity and shrinkage

$$\min_{\beta} \frac{1}{2n}||Y - X\beta||_2^2 + \lambda \left(\frac{1}{2}(1-\alpha)||\beta||_2^2 + \alpha||\beta||_1 \right)$$

or equivalently

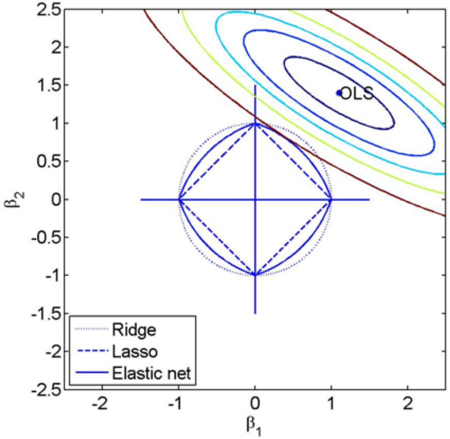
$$\min_{\beta} \frac{1}{2n}||Y - X\beta||_2^2 \quad \text{such that} \quad \frac{1}{2}(1-\alpha)||\beta||_2^2 + \alpha||\beta||_1 \ \leq t$$

for some t.

- $\alpha = 1$: Equivalent to LASSO.
- $\alpha = 0$: Equivalent to Ridge.
- $0 < \alpha < 1$: The 'Elastic' region.

Advantage: Combines the shrinkage of ridge and parameter selection of the lasso to obtain a robust sparse estimate.

Contour plot



Contour plot of OLS criteria,

$$||Y - X\beta||_2^2$$

and the elastic net constraint,

$$\frac{1}{2}(1-\alpha)||\beta||_2^2 + \alpha||\beta||_1$$

In figure $\alpha = 0.5$.

Augmented Data: Step 1a (Construction)

To solve the Elastic Net using standard LASSO solvers, we 'hide' the L_2 penalty inside the data.

Define Augmented Matrices (y^*, X^*): We stack the original data with m additional rows (the 'Ridge rows'):

$$X^*_{(n+m)\times m} = \begin{pmatrix} X \\ \sqrt{\lambda_2} I_m \end{pmatrix}, \quad y^*_{(n+m)} = \begin{pmatrix} y \\ \mathbf{0}_m \end{pmatrix}$$

- I_m is a $m \times m$ identity matrix.
- The bottom m rows of y^* are zeros.
- The factor $\sqrt{\lambda_2}$ controls the Ridge influence within the new feature matrix.

Augmented Data: Step 1b (The Absorption)

Why does this work? Let's look at the Residual Sum of Squares (RSS) for the new data:

$$\|y^* - X^*\beta\|^2 = \left\| \begin{pmatrix} y \\ \mathbf{0} \end{pmatrix} - \begin{pmatrix} X \\ \sqrt{\lambda_2} I \end{pmatrix} \beta \right\|^2$$

$$= \underbrace{\|y - X\beta\|^2}_{\text{Original RSS}} + \underbrace{\|\mathbf{0} - \sqrt{\lambda_2} I\beta\|^2}_{\text{The Ridge Penalty!}}$$

The result

By calculating OLS on y^* and X^*, we are *automatically* minimizing the Ridge-penalized loss. The L_2 term is now part of the 'error' term.

The Equivalent LASSO Problem

Once the L_2 part is absorbed into X^*, only the L_1 penalty remains standing:

$$\min_{\beta^*} \|y^* - X^* \beta^*\|^2 + \lambda_1 \|\beta^*\|_1$$

Crucial Implication:

- We can now use **LARS** or **Coordinate Descent** on (y^*, X^*).
- These algorithms are 'blind' to the fact that we have a hybrid penalty; they just see a standard LASSO problem.
- This is computationally efficient and allows us to use mature software for complex problems.

Augmented problem

We can change an elastic net problem into a Lasso problem,

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

by extending data,

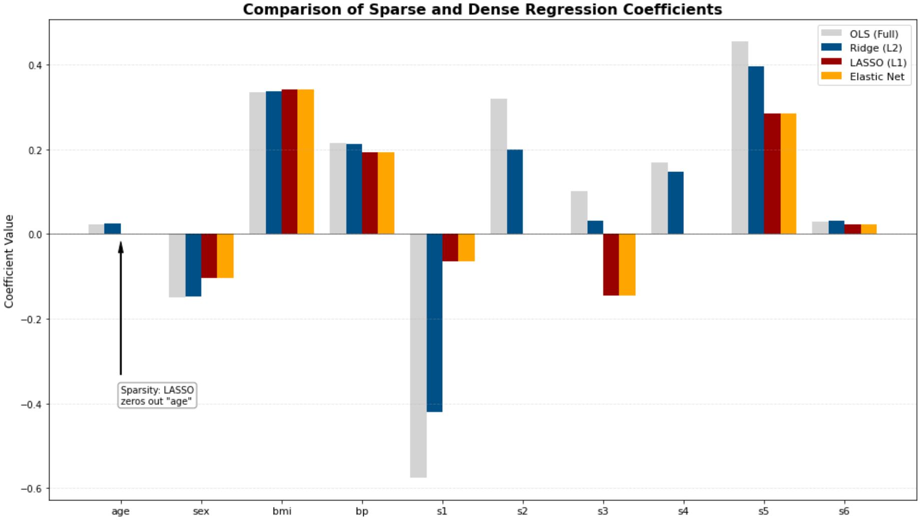
$$X^* = (1+\lambda_2)^{-1/2} \begin{bmatrix} X \\ \sqrt{\lambda_2} I_p \end{bmatrix} \text{ and } y^* = \begin{bmatrix} y \\ 0_p \end{bmatrix}$$

Yields the OLS solution

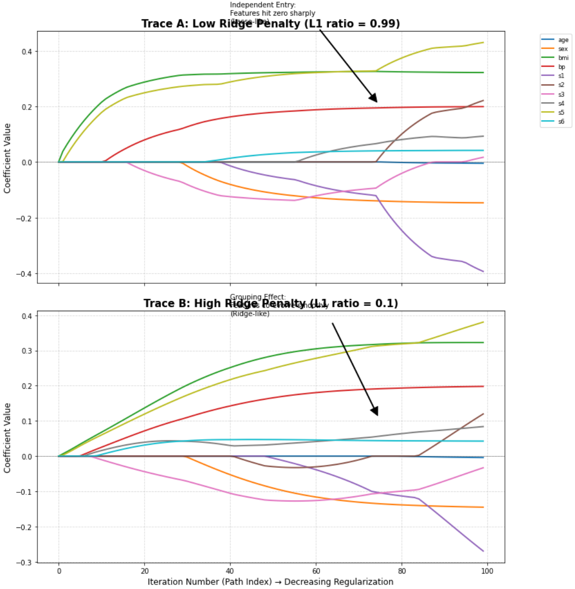
$$\frac{1}{\sqrt{1+\lambda_2}}(X^t X + \lambda_2 I_p^T I_p)\beta^* = X^T y$$

We see that $1/\sqrt{1+\lambda_2}\beta^*$ is a scaled ridge solution.
Why? Because now we can use the LARS algorithm to obtain the whole parameter trace.

The elastic net example - Diabetes



Parameter traces for Diabetes example



Combinatoric search, forward and backward selection I

Combinatoric search: Try all possible combinations of features and select the optimal one.

Pro: You will find the best combination.

Con: Number of combinations to test may be extremely large.

Forward selection: Add variables with highest information criterion one at a time.

Pro:
- Reasonable number of models to test.
- Can be used when $p > n$

Con: Might not give the best combination of features.

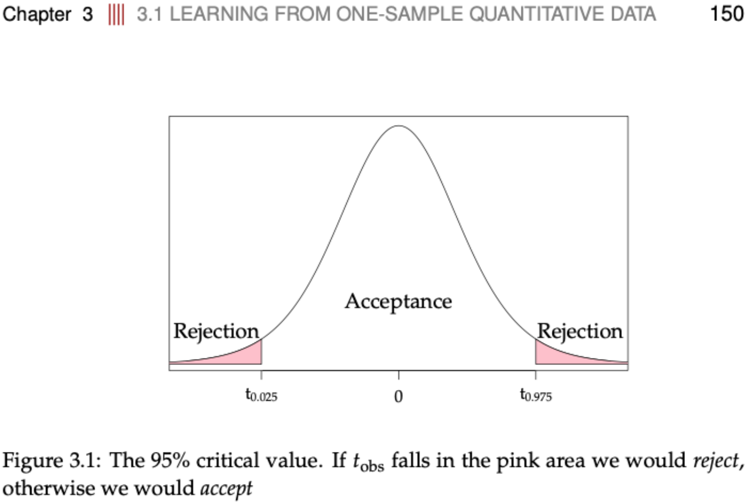
Backward elimination: Remove irrelevant features one at a time.

Pro: Reasonable number of models to test.

Con:
- Numerical issues when computing differences between models with many features.
- Might not give the best combination of features
  - Usually better than forward selection

Part II: Assumptions as models

Multiple testing



Chapter 3 |||| 3.1 LEARNING FROM ONE-SAMPLE QUANTITATIVE DATA 150

Figure 3.1: The 95% critical value. If t_{obs} falls in the pink area we would *reject*, otherwise we would *accept*

Figure from Introduction to Statistics, Brockhoff et al.

Feature assessment

Assessing the significance of each of the p features.

- Traditional t-test of difference between groups.
  - Testing for differences in mean.
- Traditional F-test of parameter significance.
  - Testing if the estimated parameters are zero.

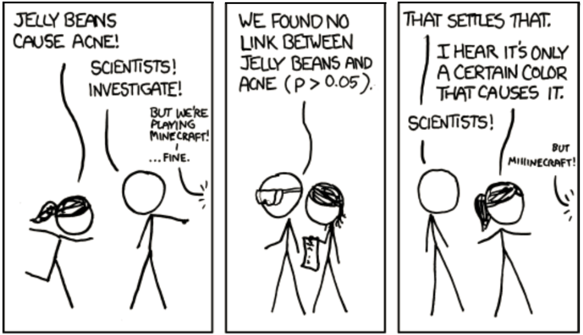
Feature assessment - the issue

If we test one hypothesis at an α-level of significance there is a chance α of falsely rejecting the hypothesis.
This is no longer the case if we do many tests!
The family-wise error rate (FWER) is the probability of at least one false rejection.
If the features are independent and each tested at an α-level, then $FWER >> \alpha$ for large p.
For M independent test at significance level α,

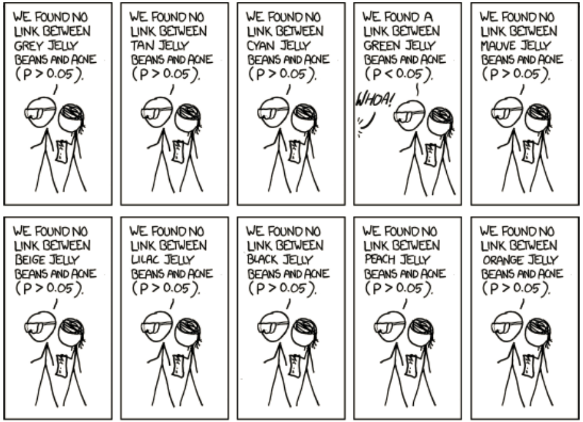
$$FWER = 1 - (1 - \alpha)^M$$

Part II: Assumptions as models

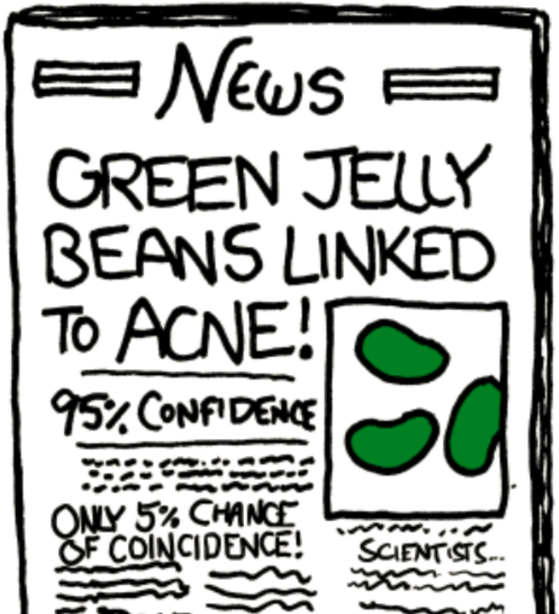
Example from www.xkcd.com


JELLY BEANS CAUSE ACNE!
SCIENTISTS! INVESTIGATE!
BUT WE'RE PLAYING MINECRAFT!
... FINE.
WE FOUND NO LINK BETWEEN JELLY BEANS AND ACNE (P > 0.05).
THAT SETTLES THAT.
I HEAR IT'S ONLY A CERTAIN COLOR THAT CAUSES IT.
SCIENTISTS!
BUT MIIIINECRAFT!

Example from www.xkcd.com


WE FOUND NO LINK BETWEEN GREY JELLY BEANS AND ACNE (P > 0.05).
WE FOUND NO LINK BETWEEN TAN JELLY BEANS AND ACNE (P > 0.05).
WE FOUND NO LINK BETWEEN CYAN JELLY BEANS AND ACNE (P > 0.05).
WE FOUND A LINK BETWEEN GREEN JELLY BEANS AND ACNE (P < 0.05).
WHOA!
WE FOUND NO LINK BETWEEN MAUVE JELLY BEANS AND ACNE (P > 0.05).
WE FOUND NO LINK BETWEEN BEIGE JELLY BEANS AND ACNE (P > 0.05).
WE FOUND NO LINK BETWEEN LILAC JELLY BEANS AND ACNE (P > 0.05).
WE FOUND NO LINK BETWEEN BLACK JELLY BEANS AND ACNE (P > 0.05).
WE FOUND NO LINK BETWEEN PEACH JELLY BEANS AND ACNE (P > 0.05).
WE FOUND NO LINK BETWEEN ORANGE JELLY BEANS AND ACNE (P > 0.05).

Example from www.xkcd.com



NEWS
GREEN JELLY BEANS LINKED TO ACNE!
95% CONFIDENCE
ONLY 5% CHANCE OF COINCIDENCE!
SCIENTISTS...

FWER for the jelly bean example

- 20 experiments conducted at a 5 % significance level
- Assume that the effect of different colors are independent, then $FWER = ???$.

FWER for the jelly bean example

- 20 experiments conducted at a 5 % significance level
- Assume that the effect of different colors are independent, then $FWER = 1 - (1 - 0.05)^{20} \approx 0.64$.
- There is 64 % probability of at least one false rejection.

FWER for the jelly bean example

- 20 experiments conducted at a 5 % significance level
- Assume that the effect of different colors are independent, then $FWER = ???$.

FWER for the jelly bean example

- 20 experiments conducted at a 5 % significance level
- Assume that the effect of different colors are independent, then $FWER = 1 - (1 - 0.05)^{20} \approx 0.64$.
- There is 64 % probability of at least one false rejection.

Bonferroni correction

Using the Bonferroni correction we rescale the α with the number of tests.
Reject a hypothesis if its p-value is below α/M.

1. Now we have an α-probability of making a false rejection.
    - Assuming independence
2. The resulting threshold will often result in low power.
    - We miss out on important effects

False Discovery Rate (FDR)

We can have more significant findings if we allow for a few mistakes.
The false discovery rate is a technique to control the number of falsely detected significant features.
The false discovery rate is

$$FDR = E\left(\frac{FP}{FP+TP}\right)$$

where

$$FP = \text{False positives (false discoveries)}$$
$$TP = \text{True positives (true discoveries)}$$

If we accept hypotheses where $FDR < q$ then we will expect that among our findings there will be q mistakes.

FDR

Gain: We control false positives - added power.

Cost: Increased number of false negatives.

We prefer to get a few false discoveries (percentage-wise) but gain more information, than ensuring no false discoveries and loosing some information.

First proposed by Yoav Benjamini and Yosef Hochberg in 1995.

Benjamini-Hochbergs algorithm for FDR

We have m tests with the null hypotheses $H_1, ..., H_m$ and corresponding p-values $p_1, ..., p_m$.

Denote the sorted p-values as $p_{(1)} \leq p_{(2)} \leq ... \leq p_{(m)}$.

For a given q find

$$k = \max\left\{i : p_{(i)} \leq \frac{i}{m} q\right\}$$

and reject $H_{(1)}, ..., H_{(k)}$.

Example BH algorithm for FDR

We have m sorted tests with the null hypotheses $H_{(1)}, ..., H_{(5)}$ and corresponding p-values $0.01, 0.05, 0.1, 0.4, 0.6$.

For $q = 0.1$ find

$$\left\{i = 1 : 0.01 \leq \frac{1}{5}0.1 = 0.02\right\} \tag{1}$$

$$\left\{i = 2 : 0.05 \not\leq \frac{2}{5}0.1 = 0.04\right\} \tag{2}$$

then $k = 1$ and we reject $H_{(1)}$.

Benjamini-Hochbergs algorithm intuition

1. Take your already calculated p-values and sort them from smallest to largest.
2. Walk down the sorted list and reject the hypotheses as long as $\frac{i}{m}q$ is smaller than the p-values.

q is **your choice** of acceptable fraction of mistakes. A single hypothesis is often tested at $\alpha = 0.05$ but we often accept higher values for q, say 0.1 or even 0.2.

Think-share-pair: Manual FDR Calculation

Context: You perform $m = 5$ hypothesis tests. You set your False Discovery Rate (FDR) level to $q = 0.20$.

Observed P-values (sorted):

$$0.01, \quad 0.03, \quad 0.15, \quad 0.40, \quad 0.50$$

Task (2 min): Calculate the Benjamini-Hochberg thresholds ($\frac{i}{m} \cdot q$) for $i = 1, 2, 3$... and decide which hypotheses to reject.

Vevox Poll 3: Rejection Decisions

Question: According to the Benjamini-Hochberg procedure, which hypotheses are rejected?

Ⓐ Only the first ($p = 0.01$)

Ⓑ First and second ($p = 0.01, 0.03$)

Ⓒ First, second, and third ($p = 0.01, 0.03, 0.15$)

Ⓓ None

Vevox Poll 3: Rejection Decisions

Question: According to the Benjamini-Hochberg procedure, which hypotheses are rejected?

Ⓐ Only the first ($p = 0.01$)

Ⓑ First and second ($p = 0.01, 0.03$)

Ⓒ First, second, and third ($p = 0.01, 0.03, 0.15$)

Ⓓ None

- $i = 1$: Threshold $= 1/5 \times 0.20 = 0.04$
- $i = 2$: Threshold $= 2/5 \times 0.20 = 0.08$
- $i = 3$: Threshold $= 3/5 \times 0.20 = 0.12$

Vevox Poll 3: Rejection Decisions

Question: According to the Benjamini-Hochberg procedure, which hypotheses are rejected?

Ⓐ Only the first ($p = 0.01$)

Ⓑ First and second ($p = 0.01, 0.03$)

Ⓒ First, second, and third ($p = 0.01, 0.03, 0.15$)

Ⓓ None

- $i = 1$: Threshold $= 1/5 \times 0.20 = 0.04$
- $i = 2$: Threshold $= 2/5 \times 0.20 = 0.08$
- $i = 3$: Threshold $= 3/5 \times 0.20 = 0.12$

Answer: B (Reject top 2)

Reasoning:

Check $i = 2$: $0.03 \leq 0.08$ (Pass).
Check $i = 3$: $0.15 \leq 0.12$ (Fail).
The largest k satisfying the condition is $k = 2$. Therefore, we reject all null hypotheses up to rank 2.