dm25s1

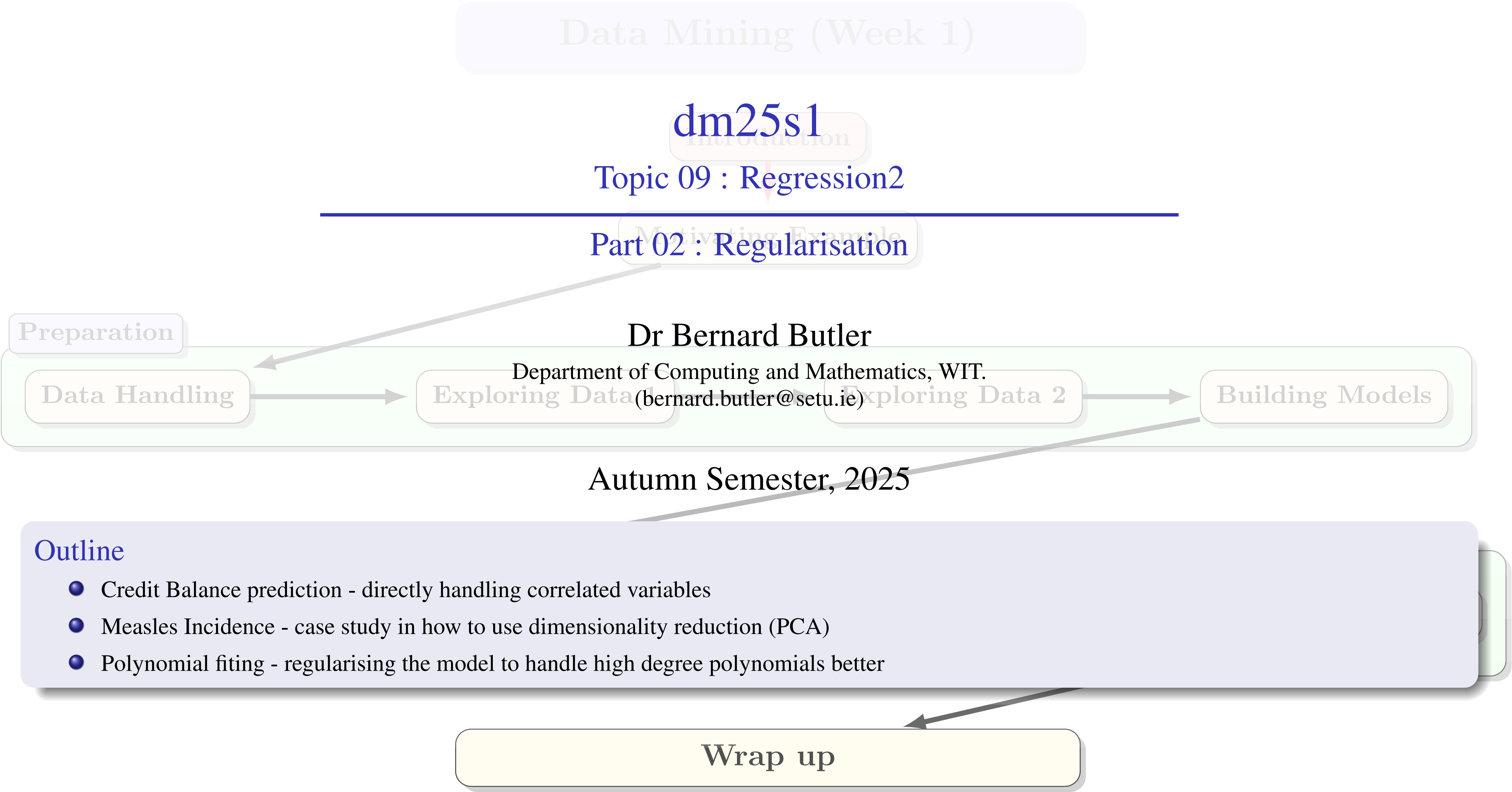
Topic 09 : Regression2

Part 02 : Regularisation

Dr Bernard Butler

Department of Computing and Mathematics, WIT.
(bernard.butler@setu.ie)

Autumn Semester, 2025



Data Mining (Week 1)

Introduction

Motivating Example

Preparation

Data Handling

Exploring Data 1

Exploring Data 2

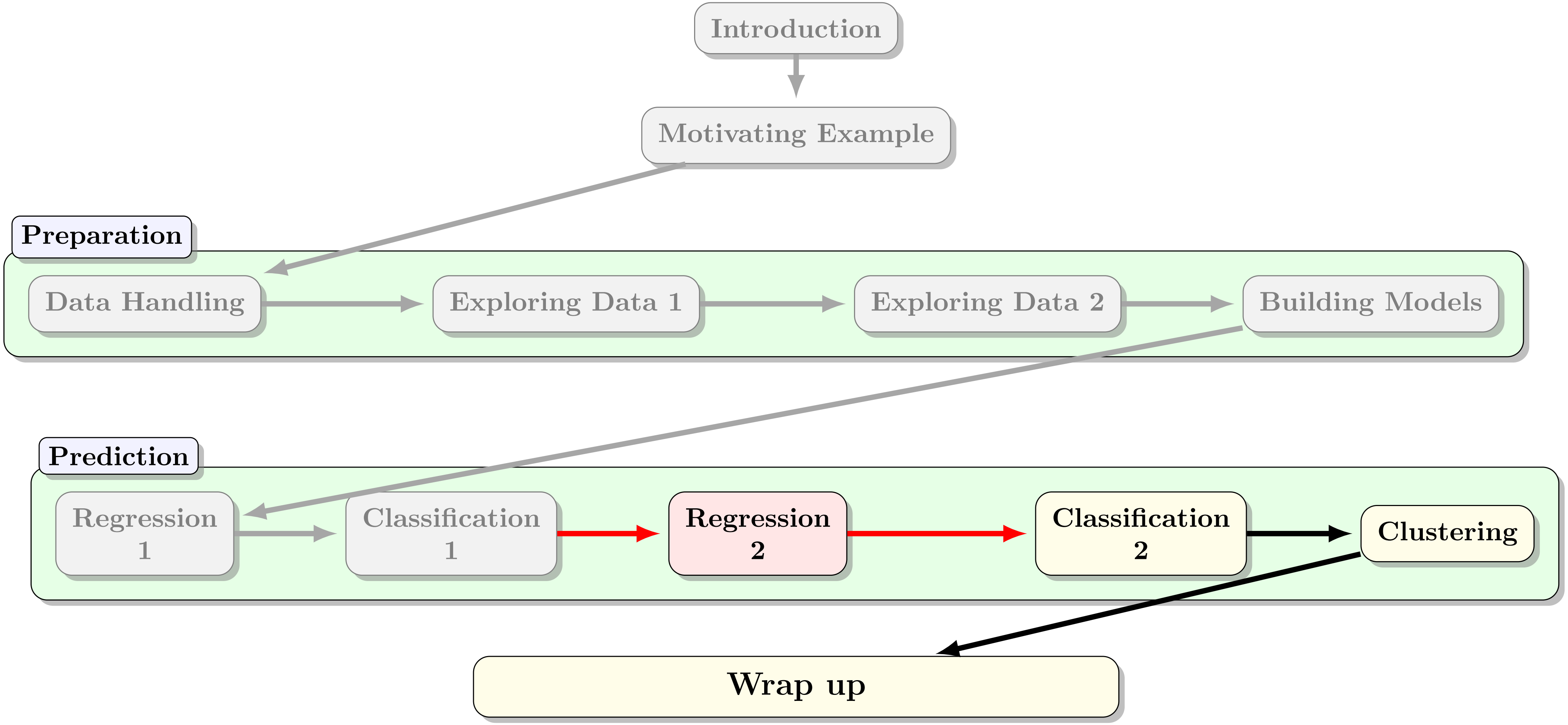
Building Models

Wrap up

Outline

- Credit Balance prediction - directly handling correlated variables
- Measles Incidence - case study in how to use dimensionality reduction (PCA)
- Polynomial fiting - regularising the model to handle high degree polynomials better

Data Mining (Week 9)



Introduction

Motivating Example

Preparation

Data Handling

Exploring Data 1

Exploring Data 2

Building Models

Prediction

Regression 1

Classification 1

Regression 2

Classification 2

Clustering

Wrap up

Outline

1. Case Study 4: Credit Balances 4

2. Multivariate Analysis 16

3. Review and Summary 25

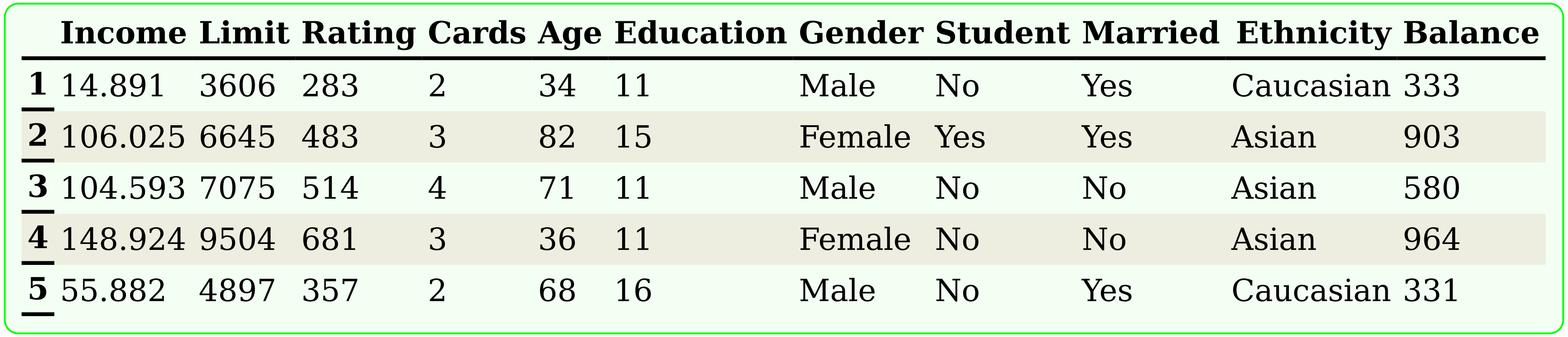
4. Resources 31

Case Study 4: Credit balances - overview

Introducing

- the `sklearn` approach to regression (we used `statsmodels` with the Diamonds and Advertising data)
- non-numeric explanatory variables like gender and ethnicity
- more advanced regression modelling, e.g., handling correlated variables

Case Study 4: Credit balances - introduction



	Income	Limit	Rating	Cards	Age	Education	Gender	Student	Married	Ethnicity	Balance
1	14.891	3606	283	2	34	11	Male	No	Yes	Caucasian	333
2	106.025	6645	483	3	82	15	Female	Yes	Yes	Asian	903
3	104.593	7075	514	4	71	11	Male	No	No	Asian	580
4	148.924	9504	681	3	36	11	Female	No	No	Asian	964
5	55.882	4897	357	2	68	16	Male	No	Yes	Caucasian	331

- Note the presence of some categorical features (*Gender*, *Student*, *Married*, *Ethnicity*).
- These can participate in linear regression models to predict a numeric response, but must be coded first.
    - For example, *Gender* can become an indicator (0,1)-valued variable of the form *IsFemale*.
    - *Ethnicity* has 3 levels and is replaced by 3-1=2 indicator variables.

A single categorical feature with *n* levels becomes *n*–1 (0,1)-coded “dummy” features.

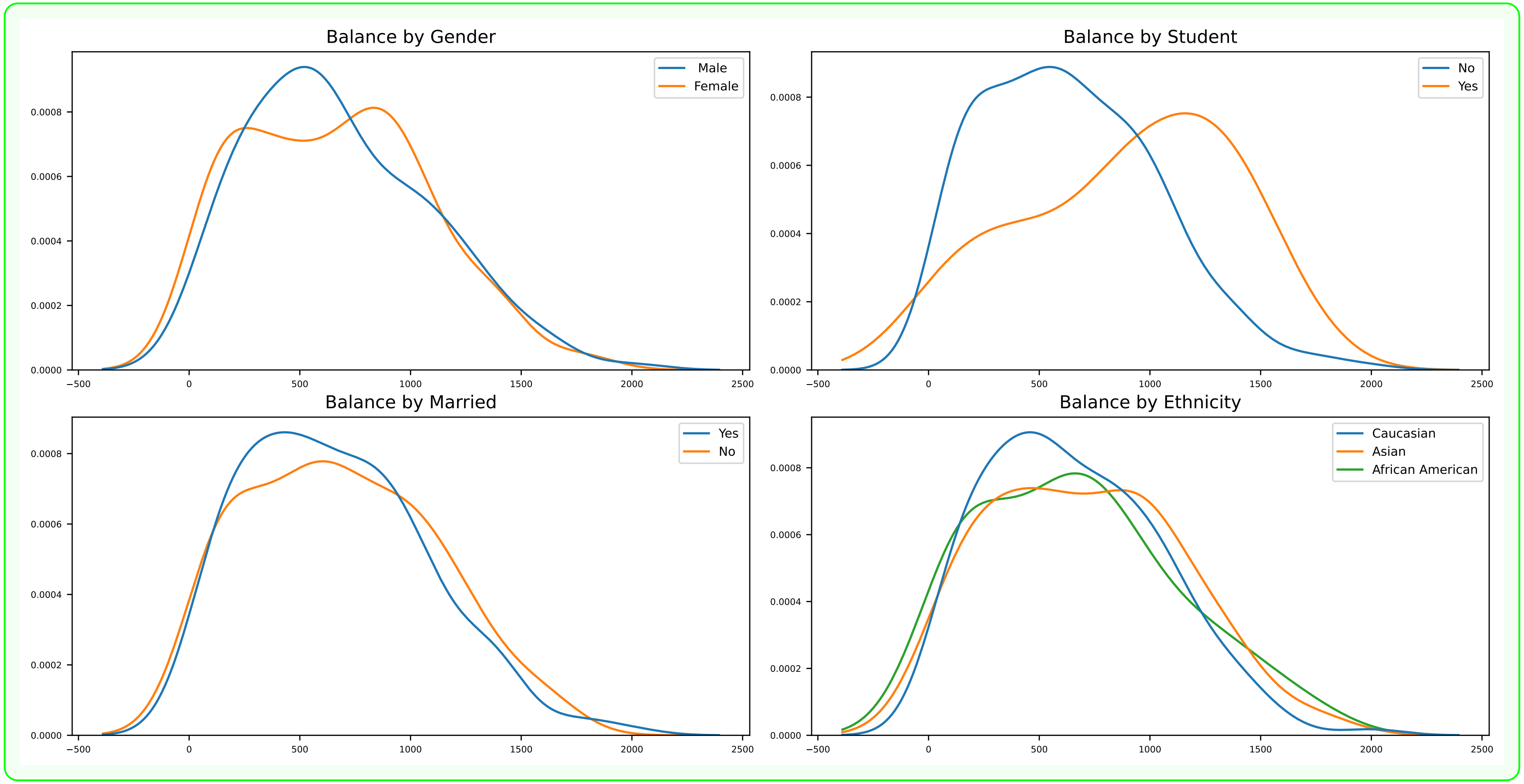
Case Study 4: Credit balances - Removing Data

- the purpose of the analysis is to predict credit balances.
- Basic exploratory data techniques (histograms) soon indicated that there were 2 cohorts
  1. those who do not use their cards and/or clear their balance each month
  2. those who use their cards and have nonzero balances
- Removing data relating to the first cohort meant that the remaining data looked more cohesive and also made linear regression easier
- Take-away: look for inconsistent subsets in the data, if possible: either remove them or develop a separate model for each subset

Case Study 4: Credit balances - Removing correlated feature

- Correlations between predictors are relatively high, but that between “Limit” and “Rating” is 1
- Generally, customers with a high rating are allowed to have high credit limits
- Conversely, customers will not be allowed high credit limits unless they have a high credit rating
- “Limit” was removed from the data used for analysis
- Take-away: remove all but 1 correlated features from a set of such features, because they increase the standard error (hence variance) and make the solver’s job much more difficult (larger condition number)

Case Study 4: Credit balances - Contribution of Categorical Variables



Which of these categorical features has a significant effect on Balance?

Case Study 4: Credit balances - Model building

- Using forward selection as before, the best model was found to be “Balance $\sim$ poly(Income,2) + Rating + Age + Student + Income:Rating”
- Could also use Backward Elimination to prune from a complex model
- For this data, high correlations between features can cause difficulties - we need techniques to handle this

Difficulties caused by correlated features

The Problem : Several features are highly correlated, so the solver has difficulty assigning an importance independently to each.

How it shows up : The condition score is large and several model coefficients take large values with opposite signs. Sometimes the solver gives up.

Solution options :

1. Remove selected features from the model (simple, does not always work and requires care)
2. Use *dimensionality reduction* (linear PCA) to derive an uncorrelated subset of the features with least loss in explanatory power (principal components can be opaque)
3. Use *regularisation*, to “penalise” large model coefficients (solve a related problem with a different loss function)

Regularisation introduction

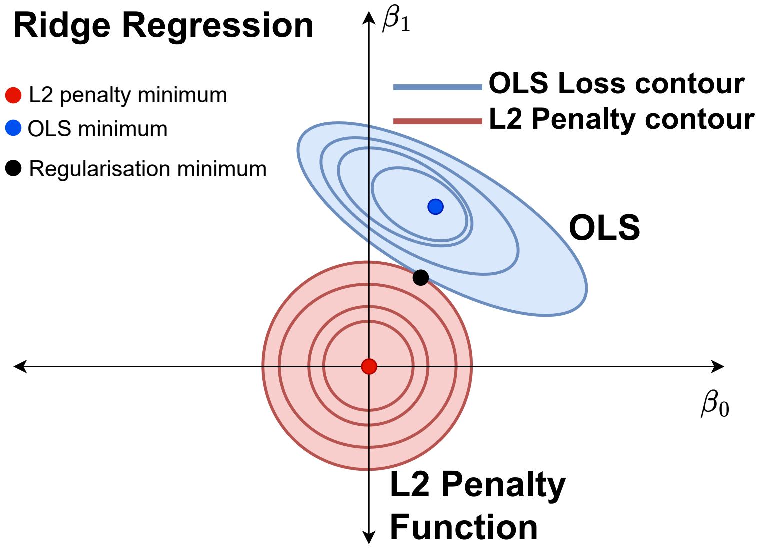
Add *regularisation* constraints to make the model work: $\min_{\boldsymbol{\beta}} ||\boldsymbol{\epsilon}||_2^2 + \lambda p(\boldsymbol{\beta})$

- Options are
  1. *Ridge Regression* where the penalty term takes the form $p(\boldsymbol{\beta}) = ||\beta||_2$
  2. *Lasso* where the penalty term takes the form $p(\boldsymbol{\beta}) = ||\beta||_1$
- Regularisation has a metaparameter λ - the challenge is to choose a suitable value
  - if too large: tries less to match the data, increases the bias
  - if too small: tries too hard to match the data so $\beta \to \infty$ and increases the variance

Ridge vs Lasso Regression

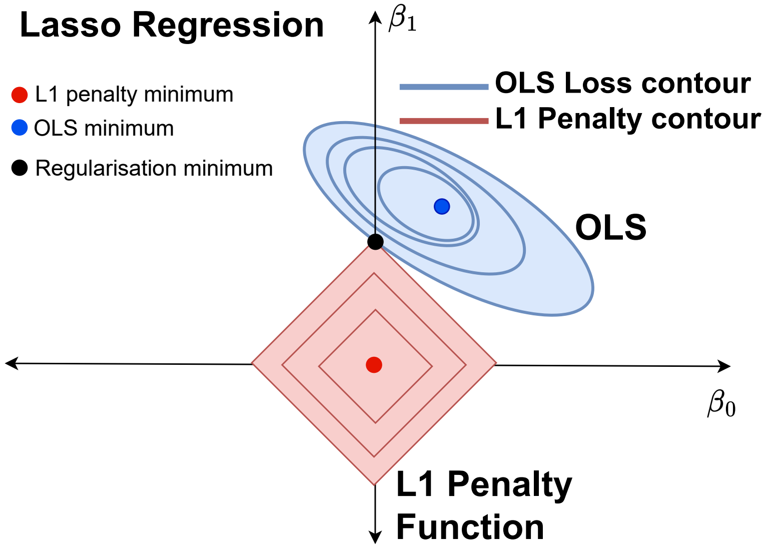
Because lasso regression favours the “corners” in parameter space, it tends to set some parameter values to 0 (essentially dropping the associated features). This has the added benefit of making the model smaller and easier to interpret.

Ridge Regression



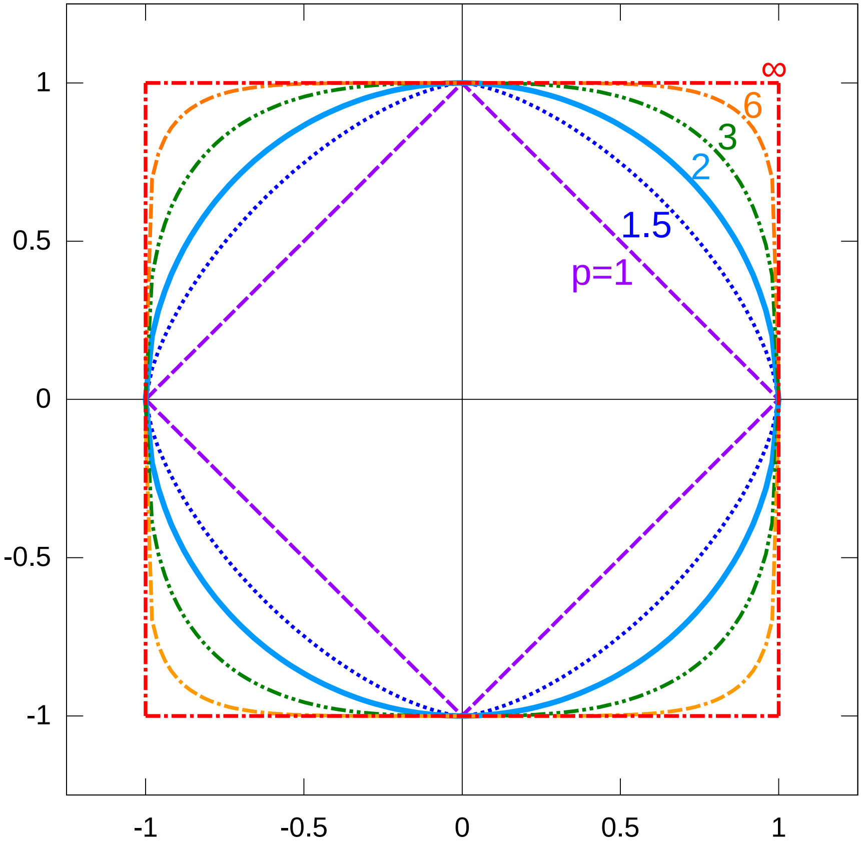
Intersection point has $\beta_0 \neq 0$ and $\beta_1 \neq 0$ so both features are needed.

Lasso Regression

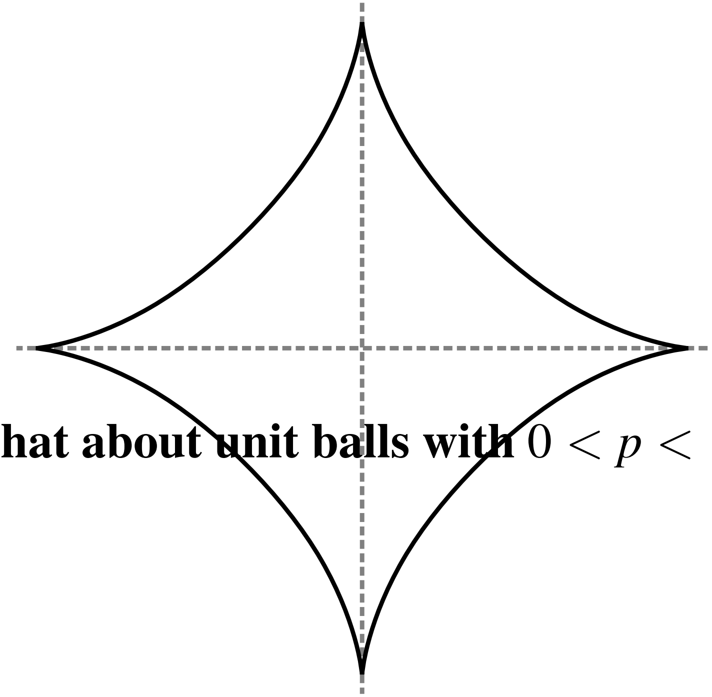


Intersection point has $\beta_0 = 0$ so its feature is no longer needed.

Sidebar - vector norms and their unit balls in 2D



By Quartl - Own work, CC BY-SA 3.0, `https://commons.wikimedia.org/w/index.php?curid=17428655`



What about unit balls with $0 < p < 1$?

Unit ball, $p = \frac{2}{3}$

Public Domain, `https://commons.wikimedia.org/w/index.php?curid=1770616`

Case Study 4: Credit balances - Regularisation - Searching for λ

1. Choose a set of candidate λ values
2. For each candidate λ, use K-fold cross-validation on data subsets to estimate the prediction error for the regularised fit with that λ
3. Choose the λ for which the expected error is least
4. Now fit all the training data again with this choice of λ

Note that lasso (but not ridge regression) can set particular β_j to 0 (effectively removing them from the model), so it operates more like the *backwards elimination* model building procedure in terms of creating a more frugal model having fewer terms.
Ridge regression downweights certain terms but does not set them to zero. However, it can be more performant, because it keeps some contribution from each feature.

Feature independence in Multivariate Data

Definition 1 (Covariance)

$\sigma_{12} = E[(X_1 - \mu_1)(X_2 - \mu_2)]$. In words, for two features X_1 and X_2, with means μ_1 and μ_2, respectively, σ_{12} is a measure of the the linear dependence between them. If they are independent, we can show that $\sigma_{12} = 0$.

Definition 2 ((Variance-)Covariance Matrix)

When there are n numeric features, there are $n \times n$ pairs of covariances $\sigma_{ij}, i = 1, \ldots, n; j = 1, \ldots, n$. The resulting covariance matrix is symmetric and diagonally dominant. This matrix captures the covariance structure of the set of n features $\{X_i\}$.

- Sometimes it is convenient to work with the correlation matrix, which is a scaled version pf the covariance matrix, with elements $\rho_{ij} = \frac{\sigma_{ij}}{\sigma_i \sigma_j}$, which is scaled so that all the diagonal elements are 1 and the off diagonal elements satisfy $-1 < \rho_{ij} < 1$.
- If two features are highly correlated, adding the second into the model does not increase the explanatory power of the model.
- Therefore, it pays to determine the covariance matrix from the data before building any models.

Multivariate data with correlated measurements

Example 3 (Measles cases, by city, per week from 1948–1985)

This data spans the period before and after the introduction of vaccination for measles (during the mid 1960s). Measles cases are recorded per week in 7 English cities. Although the cities are not adjacent, it is likely that there will be some spatial autocorrelation. Also, by the nature of disease outbreaks, there will also be some temporal autocorrelation per city.

	Date	London	Bristol	Liverpool	Manchester	Newcastle	Birmingham	Sheffield
1	1948-01-17	240	4	51	19	52	84	11
2	1948-01-24	284	3	54	23	34	65	11
3	1948-01-31	340	5	54	31	25	106	4
4	1948-02-07	511	1	89	66	27	142	7
5	1948-02-14	649	3	73	60	47	143	3
6	1948-02-21	766	13	169	87	46	191	6
7	1948-02-28	932	5	212	61	66	208	9
8	1948-03-06	1303	4	283	79	57	290	7
9	1948-03-13	1257	15	285	56	82	310	10
10	1948-03-20	1716	9	279	85	92	425	5
11	1948-03-27	1425	3	424	63	94	481	10

Removing redundant attribues, based on correlation filters

Pearson Correlation

	London	Bristol	Liverpool	Manchester	Newcastle	Birmingham	Sheffield
London	1.000000	0.474016	0.295005	0.519947	0.520185	0.707410	0.539053
Bristol	0.474016	1.000000	0.228214	0.437572	0.374370	0.546398	0.680336
Liverpool	0.295005	0.228214	1.000000	0.431414	0.482269	0.365078	0.329118
Manchester	0.519947	0.437572	0.431414	1.000000	0.554188	0.472575	0.522391
Newcastle	0.520185	0.374370	0.482269	0.554188	1.000000	0.645766	0.535574
Birmingham	0.707410	0.546398	0.365078	0.472575	0.645766	1.000000	0.690961
Sheffield	0.539053	0.680336	0.329118	0.522391	0.535574	0.690961	1.000000

London-Birmingham has correlation greater than 0.7.

Spearman Correlation

	London	Bristol	Liverpool	Manchester	Newcastle	Birmingham	Sheffield
London	1.000000	0.654859	0.399211	0.589346	0.559762	0.764533	0.581148
Bristol	0.654859	1.000000	0.356830	0.598125	0.471088	0.636617	0.613336
Liverpool	0.399211	0.356830	1.000000	0.580160	0.558448	0.383332	0.421292
Manchester	0.589346	0.598125	0.580160	1.000000	0.491076	0.507557	0.577990
Newcastle	0.559762	0.471088	0.558448	0.491076	1.000000	0.591156	0.633679
Birmingham	0.764533	0.636617	0.383332	0.507557	0.591156	1.000000	0.599110
Sheffield	0.581148	0.613336	0.421292	0.577990	0.633679	0.599110	1.000000

London-Birmingham has correlation greater than 0.7.

Kendall Correlation

	London	Bristol	Liverpool	Manchester	Newcastle	Birmingham	Sheffield
London	1.000000	0.471882	0.268666	0.417987	0.402433	0.570474	0.416055
Bristol	0.471882	1.000000	0.243417	0.428664	0.331594	0.460080	0.449481
Liverpool	0.268666	0.243417	1.000000	0.411598	0.400088	0.260798	0.291779
Manchester	0.417987	0.428664	0.411598	1.000000	0.346396	0.354931	0.411831
Newcastle	0.402433	0.331594	0.400088	0.346396	1.000000	0.428067	0.463323
Birmingham	0.570474	0.460080	0.260798	0.354931	0.428067	1.000000	0.432066
Sheffield	0.416055	0.449481	0.291779	0.411831	0.463323	0.432066	1.000000

Observations

- Critical level of correlation $\rho^{(\text{crit})} = 0.7$, so one of London or Birmingham can be dropped.
- The Spearman correlations are particularly high, so more correlation might be present.
- The Kendall correlations are inconclusive.

Working with high-dimensional data

Definition 4 (Curse of Dimensionality)

High dimensions do not just require more computing resources and make interpretation more difficult. They also make it more difficult to capture data that samples very high dimensional spaces efficiently. It can be shown that, as the dimension *d* increases, most of the volume of a hypercube is near the corners, not near the centre, where data might be easiest to collect. This makes estimating parameters much more difficult.

The proof is based on the fact that, as the dimension *d* tends to infinity, the *ratio* of the volume of the maximum hypersphere inscribed inside the hypercube of the same dimension, tends to 0. Thus there are good reasons to prefer low dimension approximations to high dimensional space.

In 2D: imagine the largest circle fitting inside a square; ratio is $\frac{\pi r^2}{4r^2} = \frac{\pi}{4} \approx 0.79$.

In 3D: imagine the largest sphere fitting inside a cube; ratio is $\frac{(4/3)\pi r^3}{8r^3} = \frac{\pi}{6} \approx 0.52$.
More generally

$$\frac{V_{\text{hypersphere}}}{V_{\text{hypercube}}} = \frac{\pi^{d/2}}{d2^{d-1}\Gamma(d/2)} \to 0 \text{ when } d \to \infty$$

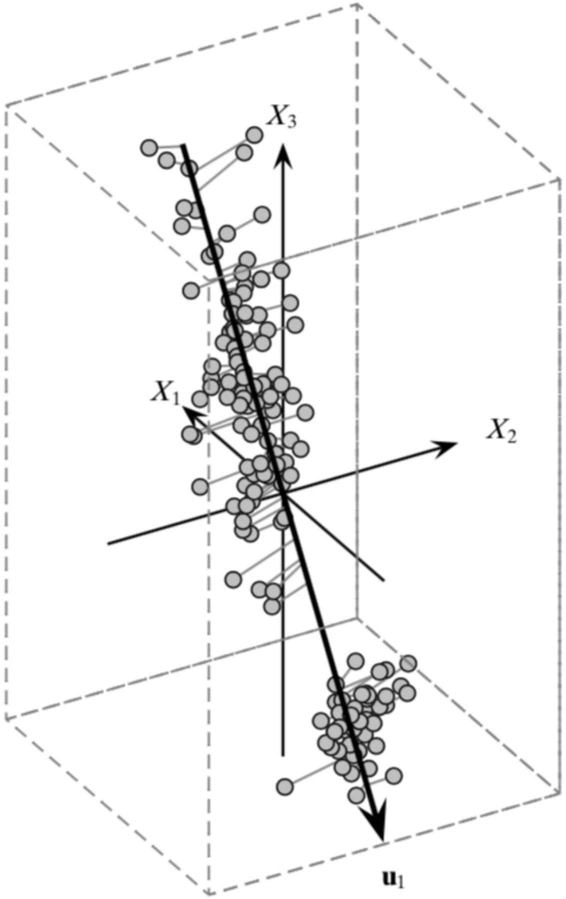
Impact: Harder to collect data that samples high-dimensional space, so harder to estimate such models.

Feature reduction

- Sometimes it is possible to use intuition to reduce the dimension, by omitting selected features.
- Another possibility is to look for groups of correlated features (c.f., *mediation*), such as the London and Birmingham measles cases above, and just choose 1 of these.
- More generally, there are techniques that search for a subspace with specified dimension d' of the features that captures most of the variance of the full set of features having dimension d, where $d' < d$ (often $d' \ll d$).
- The best known of these techniques is *Principal Components Analysis* (PCA).

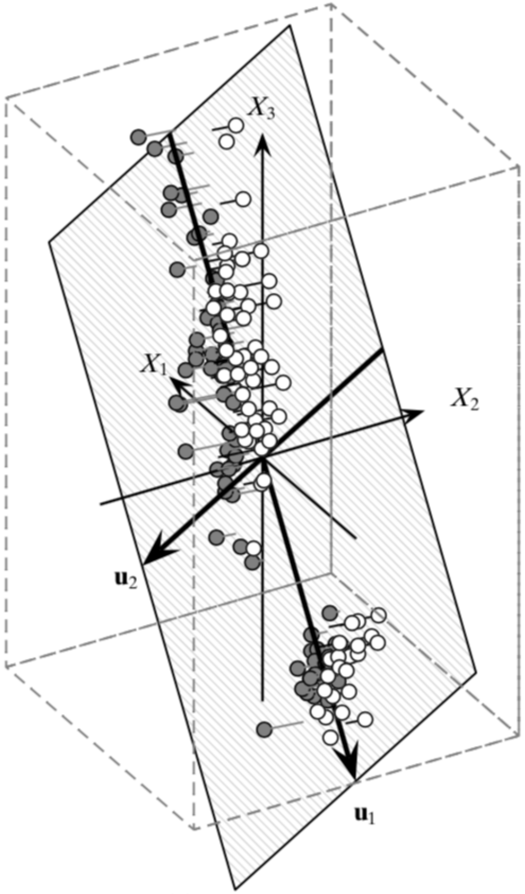
PCA visualisation

Mapping to 1 Principal Component



Mapping correlated X_1, X_2, X_3 to uncorrelated u_1

Mapping to 2 Principal Components

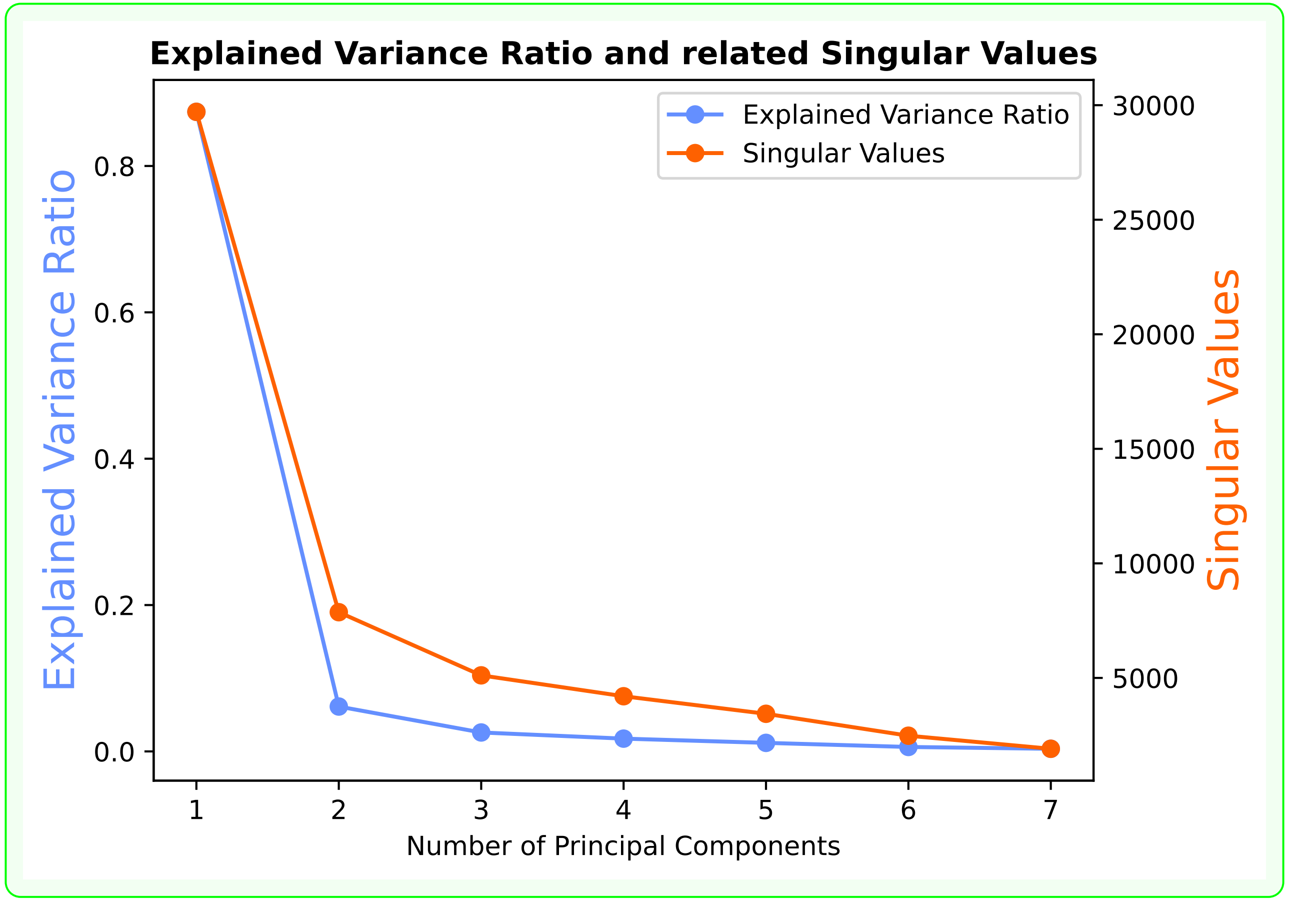


Mapping correlated X_1, X_2, X_3 to uncorrelated u_1, u_2

PCA interpretation

- Although the data has dimension $d = 3$, it is possible to find the line (on the left; $d = 1$) and plane (on the right, $d = 2$) which retain most of the variance of the data after it has been projected onto this lower dimensional subspace.
- First compute the transformations needed to align the training data with the selected subspace.
- Train the model using the transformed training data and the transformed features (principal components of the original features)
- Can then project other data, e.g., test data, onto the subspace that was derived with the original, training data, and use the model to perform predictions in the transformed space.
- Apply the inverse projection to the data, restoring it to its original orientation. However, because of the use of projections, it is not the same as the original data - the round-trip is “lossy”.
- However, it is helpful to interpret the results in terms of the original features.

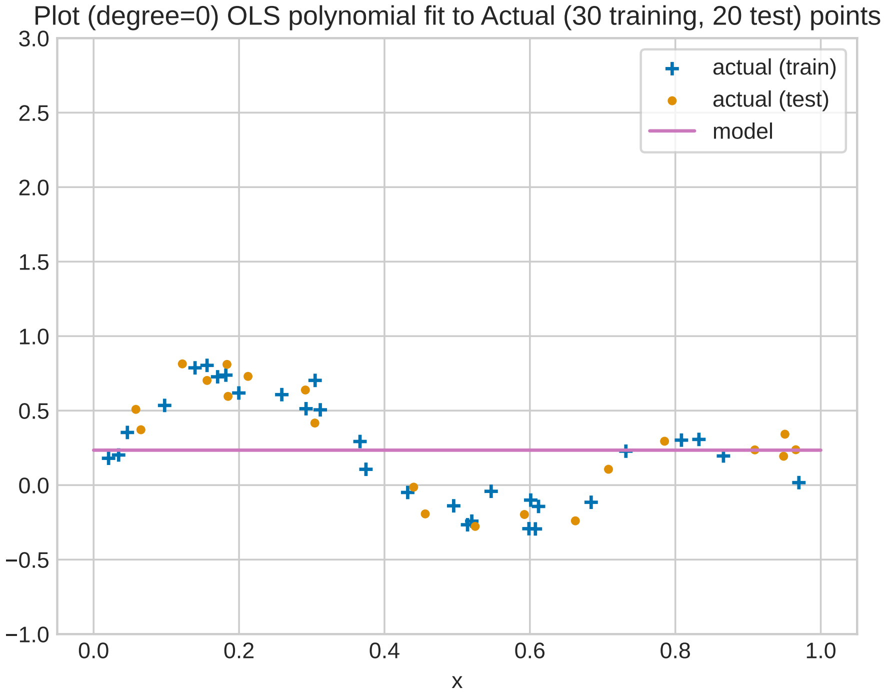
PCA example



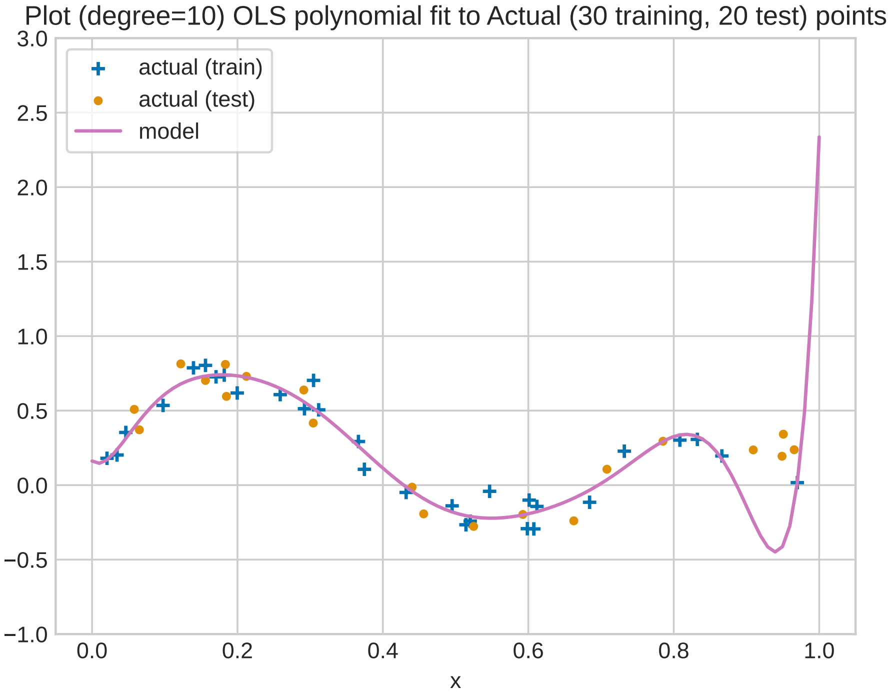
The plot shows that the first 3 singular values (associated with principal components u_1, u_2, u_3) capture the bulk of the variance in the training set. Therefore, three features, which are transformations of the other 7, are sufficient. You could interpret those features as representing the measles outbreaks in three archetypal English cities…

This youtube video describes PCA concepts well.

Returning to the problematic example

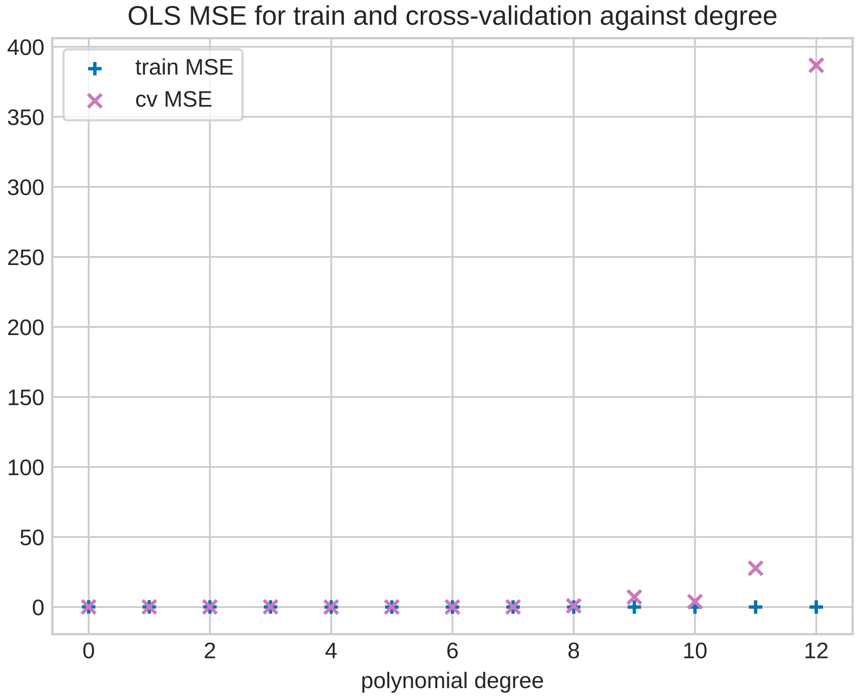


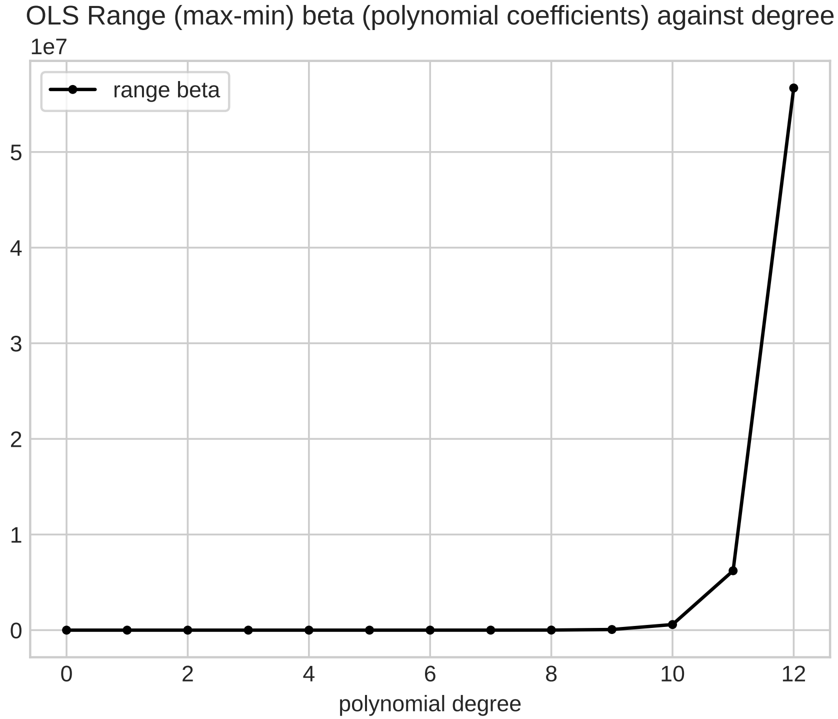
Degree 0 (constant) fit: high bias, low variance



Degree 10 (up to x^{10}) fit: low bias, high variance

Diagnosis - OLS





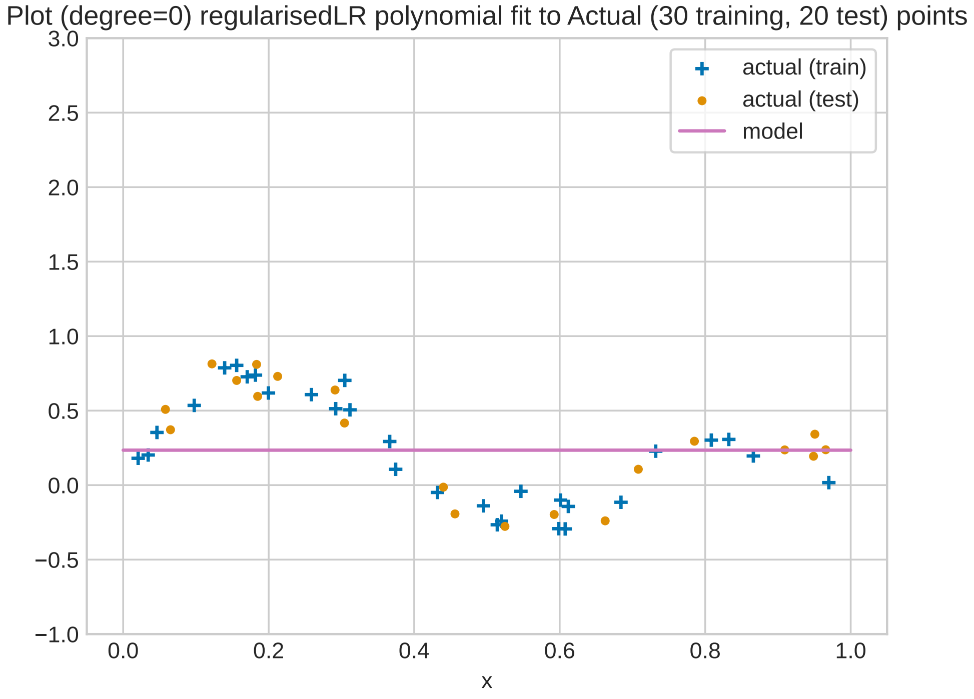
Train MSE decreases with degree, Test MSE decreases, then increases

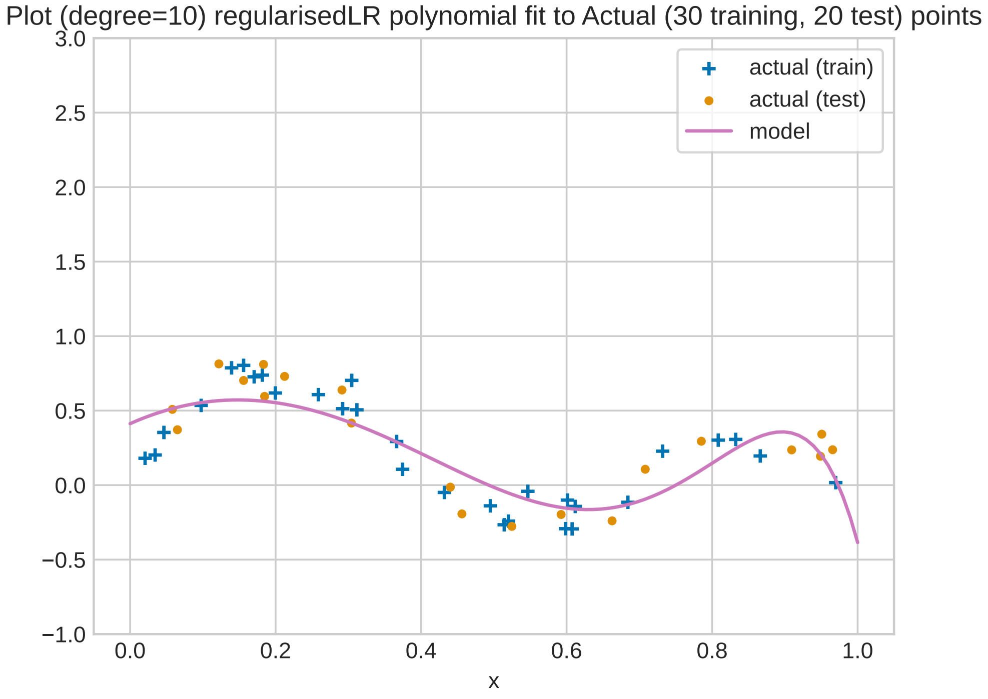
Polynomial coefficient range (max-min) increases dramatically with degree due to overfitting.

Is there any way we can use high-degree polynomials?

Yes, if we add regularisation. . .

Same data, same features, with regularisation this time

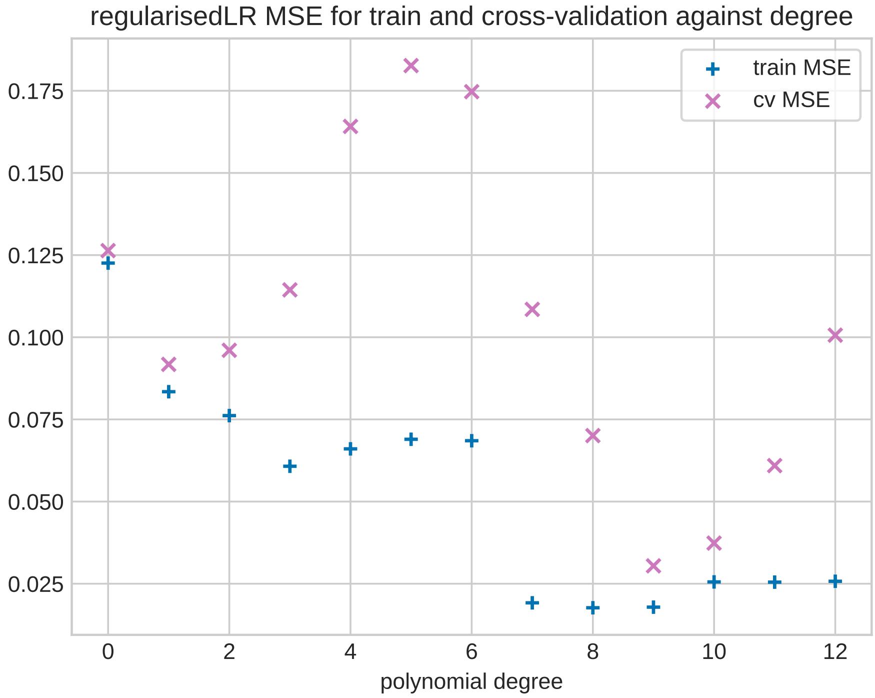


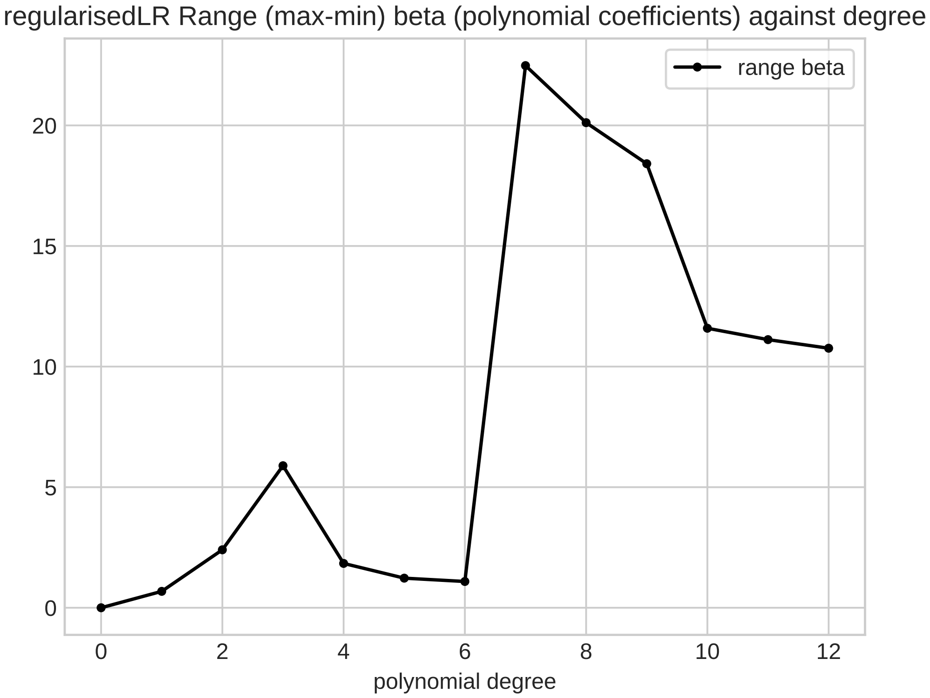


Degree 0 (constant) fit, $\lambda \approx 0$: no change

Degree 10 (up to x^{10}) fit: stabilised polynomial

Diagnosis - Regularised Linear Regression





MSE behaviour is affected by choice of λ, but degree 8 or 9 looks good

Polynomial coefficient range (max-min) is controlled - no evidence of overfitting.

So regularisation can control overfitting and/or high correlation between features

Review and summary

- Linear regression is one of the classic machine learning techniques.
- Compared to other techniques, statistics have more to offer, but ML objective (minimise prediction error on test set) is still as important!
- It has two phases, of which the first (learning from the training set) is generally the most challenging.
- It has many variants, so is quite flexible, but flexibility can be abused!
- Careful validation and model building is essential for success - it is an extension of the exploratory work done earlier in the process.
- In machine learning, prediction error is the main focus, but you need to be aware of other considerations such as
    1. model parsimony (keep model as small/simple as possible!): faster at both training and evaluation time
    2. the bias-variance dilemma: avoid overfitting and underfitting - remember, your model needs to generalise well from the training to the test set
    3. model interpretability: some models are easier to understand because the terms in the model represent concepts from the domain the data is from

Some Additional Resources

- Book: Introduction to Statistical Learning with R (2013) by James, Gareth and Witten, Daniela and Hastie, Trevor and Tibshirani, Robert.

  I strongly recommend that you read Chapter 3 of the book, as it is very well written and available online for free.
- Kaggle notebooks relating to the datasets addressed this week. There are many, but searching Kaggle should provide nice examples of data mining in action.
- I uploaded a background report on linear regression that is available for download from here.