Scientific Computing in Python Lecture 5

# Least squares, principal components, and clustering

Lecture notes

A dataset can be used to estimate a response, represent variation in fewer coordinates, or partition observations by distance. These operations solve different optimization problems. We derive each objective and then examine how the choice of units and the train/test split enter the calculation.

## 1 Linear least squares

Slides for this section

For observations $(x_i, y_i)$, let $A$ have rows $(1, x_i)$ and let $\beta = (\beta_0, \beta_1)^\mathsf{T}$. Least squares minimizes the squared Euclidean distance between observed and fitted responses. Differentiation gives the normal equations. In computation a QR or SVD solve avoids explicitly forming an inverse.

$$J(\beta) = \|A\beta - y\|_2^2.$$

$$\nabla J = 2A^\mathsf{T}(A\beta - y) = 0 \quad \Longrightarrow \quad A^\mathsf{T}A\beta = A^\mathsf{T}y.$$

### Residual orthogonality

At a least-squares minimizer,

$$A^\mathsf{T}r = 0, \qquad r = A\widehat{\beta} - y.$$

With an intercept column, $\sum_i r_i = 0$ on the training set. There is no corresponding identity for held-out residuals.

## 2 Units of a regression coefficient

Slides for this section

The intercept has the response unit. A slope has response units divided by predictor units. Changing predictor units changes the numerical slope but should preserve predictions if coefficients are transformed consistently. The fitted relationship does not by itself establish a causal mechanism.

$$\widehat{y} = \beta_0 + \beta_1 x.$$

If $x' = ax$ with $a \neq 0$, then

$$\widehat{y} = \beta_0 + \frac{\beta_1}{a}x'.$$

### Training and test error

$$\mathrm{RMSE}_{\mathrm{test}} = \sqrt{\frac{1}{n_{\mathrm{test}}} \sum_{i \in \mathrm{test}} (y_i - \widehat{y}_i)^2}.$$

Fit the coefficients using training rows only. Test RMSE has the same physical unit as $y$.

## 3 Principal component directions

Slides for this section

Let $X_c$ be a data matrix whose columns have been centered using training means. A unit vector $v$ specifies a one-dimensional projection $z = X_c v$. Principal component analysis chooses the direction with maximal squared projected variation. A Lagrange multiplier converts the constrained optimization into an eigenvalue problem.

$$\max_{v^{\mathsf{T}} v = 1} \|X_c v\|_2^2.$$

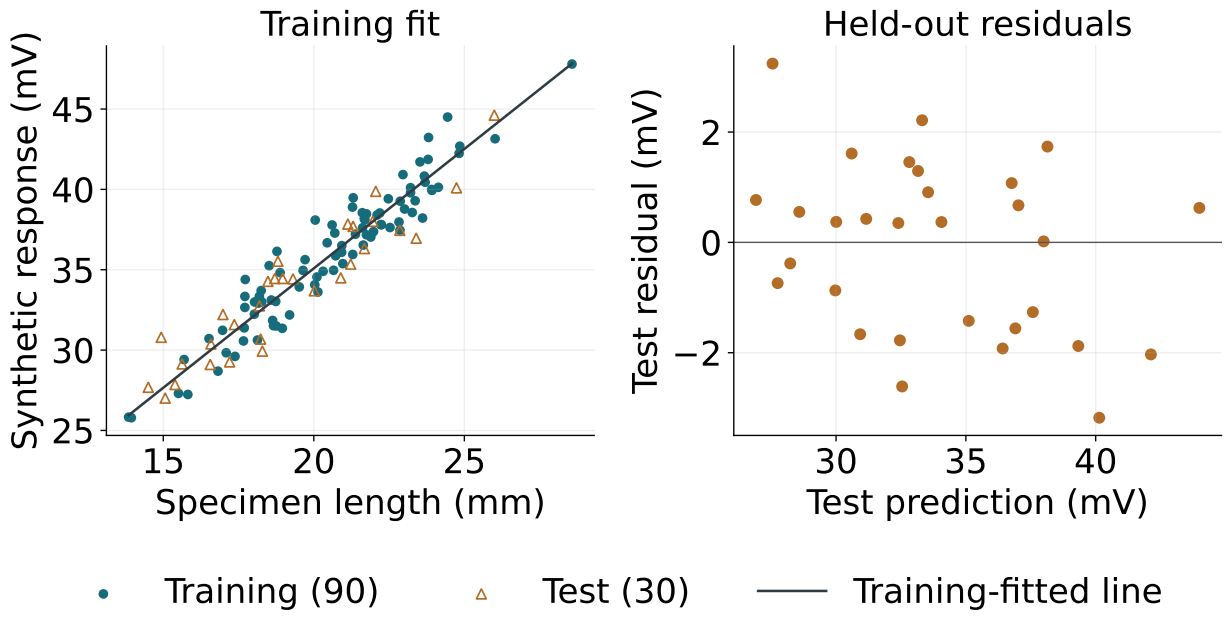
$$X_c^{\mathsf{T}} X_c v = \lambda v.$$

The leading eigenvector supplies the first principal direction.

### Projection and reconstruction

$$z = (x - \mu) v_1, \qquad \widehat{x} = \mu + z v_1^{\mathsf{T}}.$$

The reconstruction error is orthogonal to $v_1$. A large retained variance need not preserve information useful for predicting a separate response.




Ninety training rows determine the line. Thirty held-out rows determine the displayed test residuals; response and residual units are millivolts.

## 4 Scaling and distance

Slides for this section

Euclidean distance adds squared coordinate differences. Changing a mass column from grams to milligrams multiplies its squared contribution by a million. Training standardization replaces each coordinate by its deviation from the training mean divided by the training standard deviation. The same means and scales must then be used on test data.

$$d^2(x, x') = \sum_j (x_j - x'_j)^2.$$

$$z_j = \frac{x_j - \mu_{j,\text{train}}}{s_{j,\text{train}}}, \qquad s_{j,\text{train}} > 0.$$

### A zero-variance coordinate

If $s_{j,\text{train}} = 0$, division is undefined.

A fixed rule must remove or separately handle that coordinate. Test data must not determine the training scale.

# 5 The k-means objective

Slides for this section

For a fixed number of centers, k-means minimizes within-group squared distances. Given centers, each observation is assigned to its nearest center. Given assignments, each center is the arithmetic mean of its assigned observations. Alternating these steps cannot increase the objective, but may converge to different local minima from different initializations.

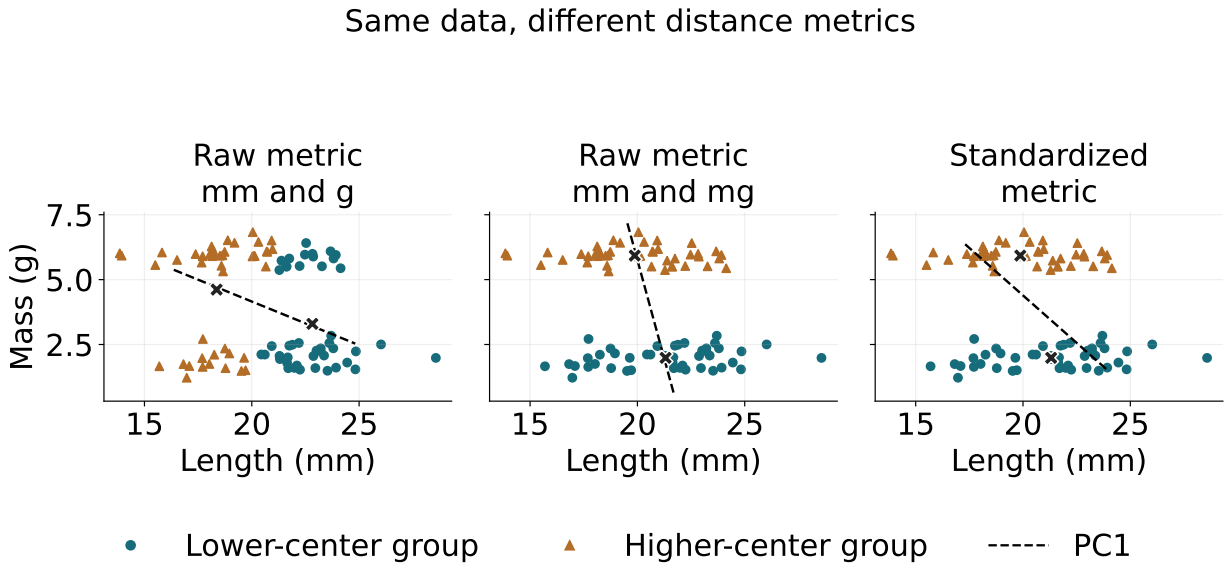
$$J(c, \mu) = \sum_i \|x_i - \mu_{c_i}\|_2^2.$$

$$c_i = \arg\min_k \|x_i - \mu_k\|_2^2, \qquad \mu_k = \frac{1}{|C_k|} \sum_{i \in C_k} x_i.$$

## Group labels

Permuting the center labels leaves the objective unchanged.

Group 0 and group 1 have no intrinsic biological meaning. Interpretation requires information beyond the distance calculation. The centroid formula assumes $|C_k| > 0$. An empty cluster has no empirical mean and requires a specified reinitialization or reassignment rule.




All panels display length in millimeters and mass in grams. Computation uses raw mm/g, raw mm/mg, or training-standardized coordinates; dashed lines are PC1 and colors are fitted groups. Circle and triangle markers redundantly distinguish the groups ordered by center mass.

# 6 Fitted transformations and evaluation

Slides for this section

A preprocessing step is part of the fitted model when it estimates means, scales, directions, or centers from data. It must be learned on the training subset. Fitting preprocessing on the full dataset allows test observations to influence the representation and changes the meaning of a held-out evaluation.

$$\text{training data} \longmapsto (\mu, s, v_1, \widehat{\beta}, \widehat{\mu}_k).$$

The fitted quantities are then held fixed when transforming and evaluating test observations.

### A numerical implementation

```
beta = np.linalg.lstsq(A_train, y_train, rcond=None)[0]
y_pred = A_test @ beta
rmse = np.sqrt(np.mean((y_test-y_pred)**2))
```

No test response enters the coefficient solve.

## Exercises

### H5-1. A line and a held-out residual

Training lengths are [0, 1, 2] mm and responses [2, 4, 6] mV. A test specimen has length 3 mm and observed response 9 mV. New practice; 10 feedback points.
**a** Find a line that exactly fits training data and state coefficient units.

**b** Compute the test prediction, observed-minus-predicted residual and single-test-point RMSE.

**c** Compare the training-mean baseline and explain why zero training error does not settle test performance.

### H5-2. Project and reconstruct

Centered feature rows are [-2, 0], [0, 0] and [2, 0]. Treat both coordinates as dimensionless. New practice; 10 feedback points.
**a** Give a first principal direction and the three one-component scores. Explain why this direction captures all variation and describe its sign ambiguity.

**b** Reconstruct the rows from those scores and state the reconstruction error.

**c** Would perfect feature reconstruction prove that the scores predict an unrelated response?

### H5-3. Units and preprocessing

A centered two-feature table has numerical covariance diag(4, 1), with zero cross-covariance. New practice; 10 feedback points.
**a** Identify the leading raw PCA coordinate, then multiply the second feature by three and identify the new leading coordinate.

**b** A training feature has mean 10 and standard deviation 2. Standardize a test value of 14. Explain why its test-set mean should not replace 10.

**c** Does standardization prove the best scientific weighting or a unique PCA direction?

### H5-4. Trace a two-center update

Four dimensionless one-feature rows are 0, 1, 7 and 8. Initial centers are 0 and 8. New practice; 10 feedback points.

**a** Assign each row to its nearest center and compute updated centers.

**b** Compute the within-group squared-distance sum before and after the update. Explain why replacing each center by its fixed-assignment group mean lowers the objective here.

**c** Explain why swapping group labels does not change this result and why the two groups are not automatically known classes.

## References

Original Python Lecture 5. PDF pages 5–18.
https://zanahmad.com/files/scientific-python/Lecture_5__Intro_to_Python.pdf

NumPy least-squares reference. Parameters, returned rank and residuals.
https://numpy.org/doc/stable/reference/generated/numpy.linalg.lstsq.html

NumPy singular-value decomposition reference. Reduced two-dimensional factorization and returned factors.
https://numpy.org/doc/stable/reference/generated/numpy.linalg.svd.html

scikit-learn PCA guide. Section 2.5.1.1: centering, scaling and principal directions.
https://scikit-learn.org/stable/modules/decomposition.html#pca

scikit-learn clustering guide. Section 2.3.2: centroid objective, alternating steps and local minima.
https://scikit-learn.org/stable/modules/clustering.html#k-means

scikit-learn common pitfalls. Sections 12.1–12.2.1: consistent preprocessing and training-only fitting.
https://scikit-learn.org/stable/common_pitfalls.html#data-leakage

Python source materials: Phillip Kerger (2023). Course adaptation and additional exposition: Zan Ahmad.