Lecture 8: Cross-Validation and the Smoothing Matrix

MATH5824 Generalised Linear and Additive Models

Richard P Mann

MATH5824 Generalised Linear and Additive Models

Reading

Course notes: Chapter 5, Sections 5.3–5.6

`www.richardpmann.com/MATH5824`

Recap: The Problem

Goal: Choose the smoothing parameter λ for the penalised criterion

$$R_\nu(f, \lambda) = \sum_{i=1}^{n} (y_i - f(t_i))^2 + \lambda\, J_\nu(f)$$

Leave-one-out CV:

$$Q_{\mathrm{OCV}}(\lambda) = \frac{1}{n} \sum_{j=1}^{n} \Big(y_j - \hat{f}_{\lambda,-j}(t_j) \Big)^2$$

Apparently requires n separate spline fits. Can we do better?

$\Rightarrow$ Yes, using the **smoothing matrix**.

The Smoothing Matrix

Key observation: The fitted values are a *linear* function of the data.

From Proposition 4.3, the smoothing spline solution gives:

$$\hat{\mathbf{f}} = \mathbf{S}_\lambda \mathbf{y}$$

where $\mathbf{S}_\lambda$ is the $n \times n$ **smoothing matrix** (or hat matrix).

Properties of $\mathbf{S}_\lambda$:

- Symmetric and positive definite (for $\lambda > 0$)
- Depends on λ and the knot locations, but *not* on $\mathbf{y}$
- Analogous to the hat matrix $\mathbf{H} = \mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'$ in linear regression

The OCV Shortcut

Result: The leave-one-out CV criterion can be computed from a *single* fit:

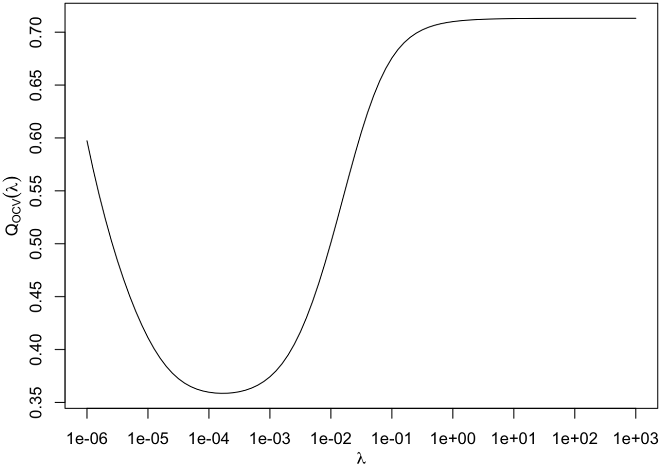
$$Q_{\mathrm{OCV}}(\lambda) = \frac{1}{n}\sum_{j=1}^{n}\left(\frac{y_j - \hat{f}_\lambda(t_j)}{1 - s_{jj}}\right)^2$$

where s_{jj} is the jth diagonal element of $\mathbf{S}_\lambda$.

Interpretation:

- The numerator is the ordinary residual $r_j = y_j - \hat{f}_\lambda(t_j)$
- Dividing by $(1 - s_{jj})$ inflates the residual to account for the influence of observation j on its own fitted value
- No need to refit the spline n times

Choosing λ by Cross-Validation



$Q_{\mathrm{OCV}}(\lambda)$ plotted against λ. The minimum gives the optimal smoothing parameter. Note: there is no theoretical guarantee of a unique minimum, but in practice the curve is

Effective Degrees of Freedom

In OLS regression: the hat matrix $\mathbf{H}$ satisfies $\mathrm{tr}(\mathbf{H}) = p$ (number of parameters).

For smoothing splines: define the **effective degrees of freedom**:

$$\mathrm{edf}_\lambda = \mathrm{tr}(\mathbf{S}_\lambda)$$

Limiting behaviour:

- $\lambda \to 0$: $\mathrm{edf}_\lambda \to n$ (interpolation, one "parameter" per data point)
- $\lambda \to \infty$: $\mathrm{edf}_\lambda \to \nu$ (polynomial regression with ν parameters)

For cubic smoothing splines ($\nu = 2$): edf ranges between 2 (straight line) and n (interpolation).

Generalised Cross-Validation (GCV)

Problem with OCV: The diagonal elements s_{jj} vary across observations.

GCV idea: Replace each s_{jj} with its average $\frac{1}{n}\mathrm{tr}(\mathbf{S}_\lambda) = \frac{\mathrm{edf}_\lambda}{n}$:

$$Q_{\mathrm{GCV}}(\lambda) = \frac{\frac{1}{n}\sum_{j=1}^{n}\left(y_j - \hat{f}_\lambda(t_j)\right)^2}{\left(1 - \frac{\mathrm{edf}_\lambda}{n}\right)^2}$$

Advantages of GCV over OCV:

- Numerically more stable
- Only requires $\mathrm{tr}(\mathbf{S}_\lambda)$, not individual diagonal elements
- Implemented in the `mgcv` package

Interpreting the GCV Formula

$$Q_{\mathrm{GCV}}(\lambda) = \frac{\text{Mean squared residual}}{(1 - \mathrm{edf}_\lambda / n)^2}$$

Numerator: Average squared residual — measures fit to data.

- Decreases as $\lambda \to 0$ (better fit)

Denominator: Penalty for model complexity.

- Approaches 0 as edf $\to n$ (overfitting penalty)
- Close to 1 when edf $\ll n$ (simple model)

Minimising GCV balances these two forces.

Connection to Linear Regression

	Linear Regression	Smoothing Spline
Fitted values	$\hat{\mathbf{y}} = \mathbf{H}\mathbf{y}$	$\hat{\mathbf{f}} = \mathbf{S}_\lambda \mathbf{y}$
Hat matrix	$\mathbf{H} = \mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'$	$\mathbf{S}_\lambda$
Degrees of freedom	$\mathrm{tr}(\mathbf{H}) = p$	$\mathrm{tr}(\mathbf{S}_\lambda) = \mathrm{edf}_\lambda$
Residual df	$n - p$	$n - \mathrm{edf}_\lambda$

Smoothing splines generalise linear regression: the smoothing matrix $\mathbf{S}_\lambda$ plays the role of the hat matrix, and edf_λ replaces the integer parameter count p.

Summary

Key points:

- Fitted values satisfy $\hat{\mathbf{f}} = \mathbf{S}_\lambda \mathbf{y}$ where $\mathbf{S}_\lambda$ is the smoothing matrix
- OCV can be computed from a single fit using diagonal elements of $\mathbf{S}_\lambda$
- Effective degrees of freedom: $\text{edf}_\lambda = \text{tr}(\mathbf{S}_\lambda)$
- edf ranges from ν (polynomial) to n (interpolation)
- GCV replaces individual s_{jj} with their average for numerical stability
- GCV is the standard method for choosing λ in practice

Next lecture: Generalised additive models — combining smoothing splines with GLMs.