Smoothness Selection

Simon Wood
Mathematical Sciences, University of Bath, U.K.

Smoothness selection approaches

- The smoothing model $y_i = f(x_i) + \epsilon_i$, $\epsilon_i \sim N(0, \sigma^2)$, is represented via a basis expansion of f, with coefficients $\boldsymbol{\beta}$.
- The $\boldsymbol{\beta}$ estimates are $\hat{\boldsymbol{\beta}} = \arg\min_{\boldsymbol{\beta}} \|\mathbf{y} - \mathbf{X}\boldsymbol{\beta}\|^2 + \lambda \boldsymbol{\beta}^{\mathsf{T}} \mathbf{S} \boldsymbol{\beta}$ where $\mathbf{X}$ is the model matrix derived from the basis, and $\mathbf{S}$ is the wiggliness penalty matrix.
- λ controls smoothness — how should it be chosen?
- There are 3 main statistical approaches
  1. Choose λ to minimize error in predicting new data.
  2. Treat smooths as random effects, following the Bayesian smoothing model, and estimate λ as a variance parameter using a marginal likelihood approach.
  3. Go fully Bayesian by completing the Bayesian model with a prior on λ (requires simulation and not pursued here).

Prediction error: C_p/UBRE

- Suppose σ^2 is **known**, and let $\mathbf{A} = \mathbf{X}(\mathbf{X}^\mathsf{T}\mathbf{X} + \lambda\mathbf{S})^{-1}\mathbf{X}^\mathsf{T}$.
- $\hat{\boldsymbol{\mu}} = \mathbf{A}\mathbf{y}$ where $\mathbb{E}(\mathbf{y}) = \boldsymbol{\mu}$, so consider

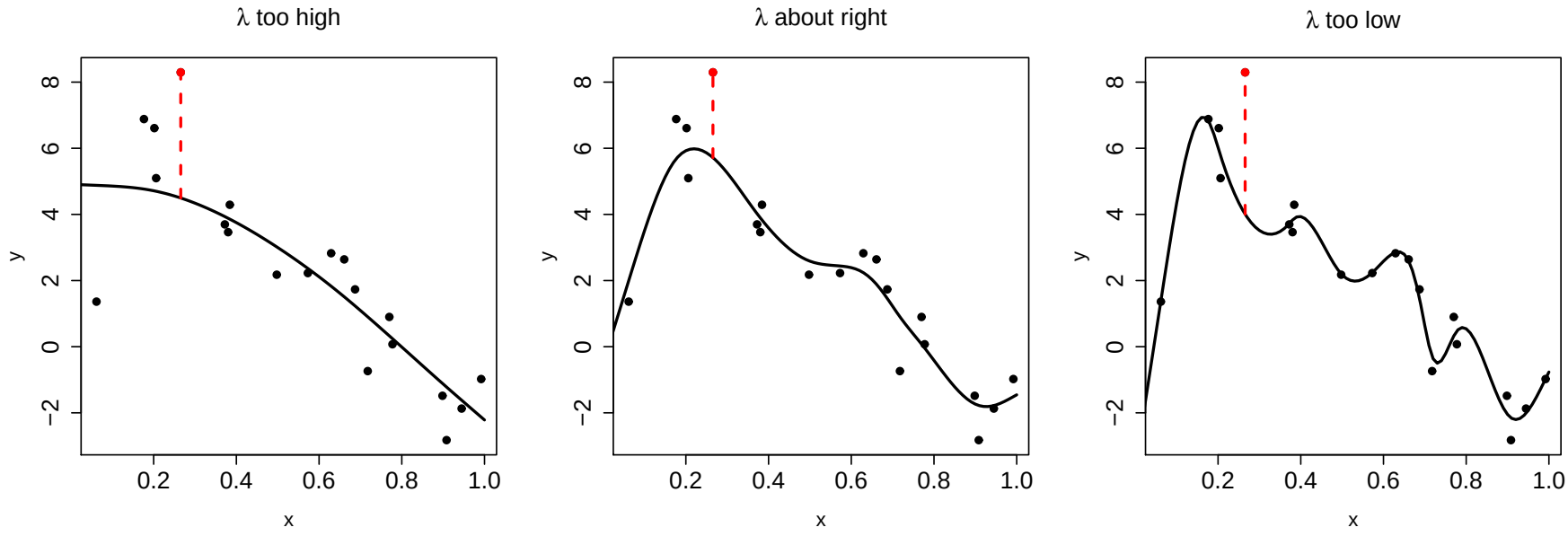
$$\begin{aligned}
\|\boldsymbol{\mu} - \hat{\boldsymbol{\mu}}\|^2 &= \|\boldsymbol{\mu} - \mathbf{A}\mathbf{y}\|^2 = \|\mathbf{y} - \mathbf{A}\mathbf{y} - \boldsymbol{\epsilon}\|^2 \\
&= \|\mathbf{y} - \mathbf{A}\mathbf{y}\|^2 + \boldsymbol{\epsilon}^\mathsf{T}\boldsymbol{\epsilon} - 2\boldsymbol{\epsilon}^\mathsf{T}(\mathbf{y} - \mathbf{A}\mathbf{y}) \\
&= \|\mathbf{y} - \mathbf{A}\mathbf{y}\|^2 + \boldsymbol{\epsilon}^\mathsf{T}\boldsymbol{\epsilon} - 2\boldsymbol{\epsilon}^\mathsf{T}(\boldsymbol{\mu} + \boldsymbol{\epsilon}) + 2\boldsymbol{\epsilon}^\mathsf{T}\mathbf{A}(\boldsymbol{\mu} + \boldsymbol{\epsilon}) \\
&= \|\mathbf{y} - \mathbf{A}\mathbf{y}\|^2 - \boldsymbol{\epsilon}^\mathsf{T}\boldsymbol{\epsilon} - 2\boldsymbol{\epsilon}^\mathsf{T}\boldsymbol{\mu} + 2\boldsymbol{\epsilon}^\mathsf{T}\mathbf{A}\boldsymbol{\mu} + 2\boldsymbol{\epsilon}^\mathsf{T}\mathbf{A}\boldsymbol{\epsilon}
\end{aligned}$$

- Hence $\mathbb{E}\|\boldsymbol{\mu} - \hat{\boldsymbol{\mu}}\|^2 = \mathbb{E}\|\mathbf{y} - \mathbf{A}\mathbf{y}\|^2 - n\sigma^2 + 2\sigma^2\mathrm{tr}(\mathbf{A})$
- Estimating $\mathbb{E}\|\mathbf{y} - \mathbf{A}\mathbf{y}\|^2$ yields ...

$$\mathcal{C}_p = \|\mathbf{y} - \mathbf{A}\mathbf{y}\|^2 - n\sigma^2 + 2\sigma^2\mathrm{tr}(\mathbf{A})$$

- Can choose λ to minimize $\mathcal{C}_p$.

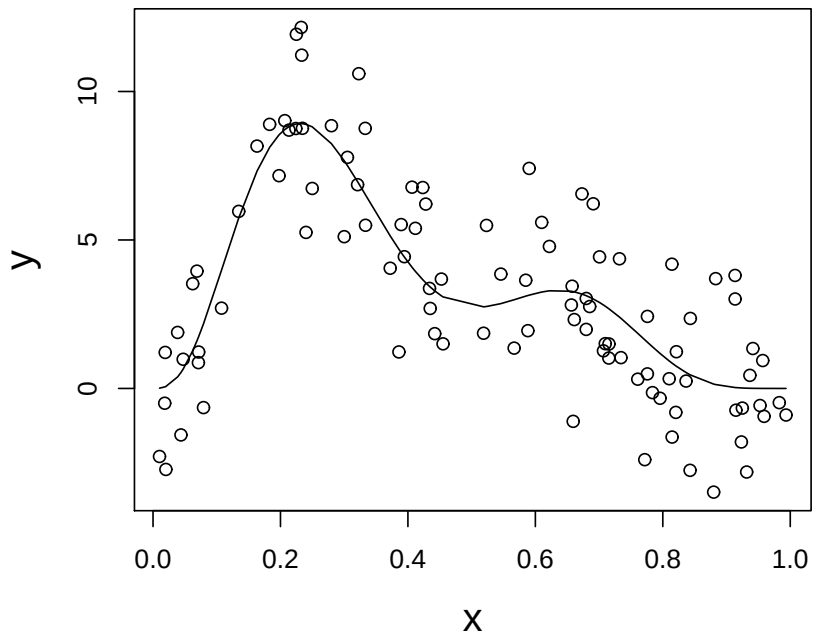
σ^2 unknown: cross validation

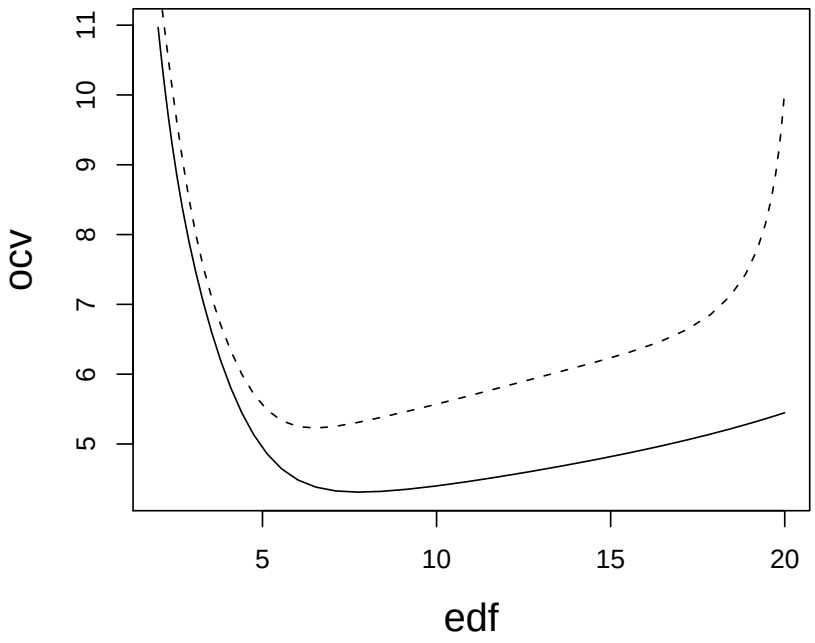


1. Choose λ to try to minimize the error predicting new data.
2. Minimize the average error in predicting single datapoints *omitted* from the fit. Each datum left out once in average.
3. It turns out that

$$\mathcal{V}_o(\lambda) = \frac{1}{n}\sum_i (y_i - \hat{\mu}_i^{[-i]})^2 = \frac{1}{n}\sum_i \frac{(y_i - \hat{\mu}_i)^2}{(1 - A_{ii})^2}$$

OCV not invariant





- OCV is not invariant in an odd way. If $\mathbf{Q}$ is orthogonal then fitting objective

$$\|\mathbf{Qy} - \mathbf{QX}\boldsymbol{\beta}\|^2 + \lambda\boldsymbol{\beta}^{\mathsf{T}}\mathbf{S}\boldsymbol{\beta}$$

yields identical inferences about $\boldsymbol{\beta}$ as the original objective, but it gives a different $\mathcal{V}_o$.

GCV: generalized cross validation

- If we find the $\mathbf{Q}$ that causes the leading diagonal elements of $\mathbf{A}$ to be constant, and then perform OCV, the result is the invariant alternative GCV:

$$\mathcal{V}_g = \frac{n\|\mathbf{y} - \hat{\boldsymbol{\mu}}\|^2}{\{n - \mathrm{tr}(\mathbf{A})\}^2}$$

- It is easy to show that $\mathrm{tr}(\mathbf{A}) = \mathrm{tr}(\mathbf{F})$, where $\mathbf{F}$ is the degrees of freedom matrix.
- In addition to invariance, GCV is much easier to optimize efficiently in the multiple smoothing parameter case.

REML/ML λ estimation

- The Bayesian smooth model is

$$\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\epsilon}, \quad \boldsymbol{\beta} \sim N(\mathbf{0}, \mathbf{S}^{-}\sigma^2/\lambda), \quad \boldsymbol{\epsilon} \sim N(\mathbf{0}, \mathbf{I}\sigma^2)$$

- This can be viewed as a mixed model for computational purposes, but the impropiety of $f(\boldsymbol{\beta})$ is awkward.
- To fix this, find the eigen-decomposition $\mathbf{S} = \mathbf{U}\boldsymbol{\Lambda}\mathbf{U}^{\mathsf{T}}$
- Reparameterize $\boldsymbol{\beta}' = \mathbf{U}^{\mathsf{T}}\boldsymbol{\beta}$ and let $\boldsymbol{\Lambda}_+$ denote the diagonal matrix of +ve eigenvalues.
- Now $\boldsymbol{\beta}^{\mathsf{T}}\mathbf{S}\boldsymbol{\beta} = \boldsymbol{\beta}'^{\mathsf{T}}\boldsymbol{\Lambda}\boldsymbol{\beta}' = \mathbf{b}^{\mathsf{T}}\boldsymbol{\Lambda}_+\mathbf{b}$ where $\boldsymbol{\beta}' = (\mathbf{b}^{\mathsf{T}}, \boldsymbol{\gamma}^{\mathsf{T}})^{\mathsf{T}}$.
- Now partition $\mathbf{X}' = \mathbf{X}\mathbf{U} = (\mathbf{Z} : \tilde{\mathbf{X}})$, so that the model becomes

$$\mathbf{y} = \tilde{\mathbf{X}}\boldsymbol{\gamma} + \mathbf{Z}\mathbf{b} + \boldsymbol{\epsilon}, \quad \mathbf{b} \sim N(\mathbf{0}, \boldsymbol{\Lambda}_+^{-1}\sigma^2/\lambda), \quad \boldsymbol{\epsilon} \sim N(\mathbf{0}, \mathbf{I}\sigma^2)$$

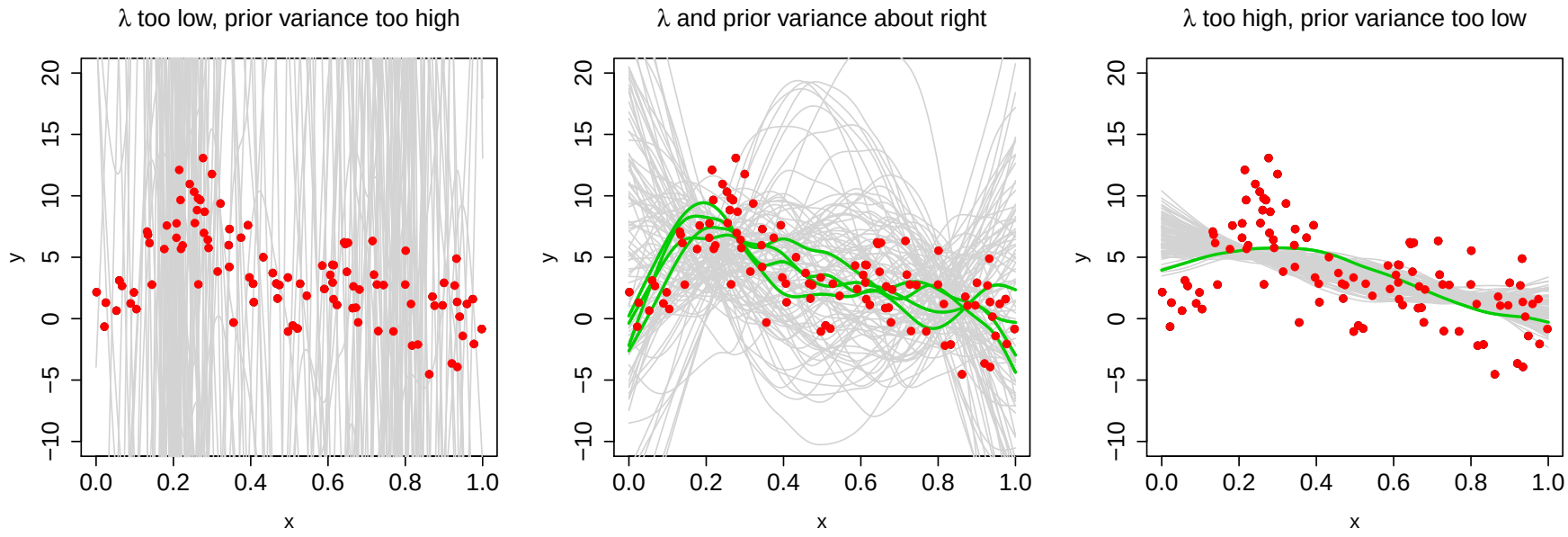
REML/ML λ estimation

- Now that the model is in standard mixed model form, mixed model methods can estimate λ as a variance parameter.
- MLE or REML can be used.
- From a Bayesian perspective we are being empirical Bayesians and using marginal likelihood.
- Notice that the restricted/marginal likelihood has the form

$$\int f(\mathbf{y}|\boldsymbol{\beta})f(\boldsymbol{\beta})d\boldsymbol{\beta}$$

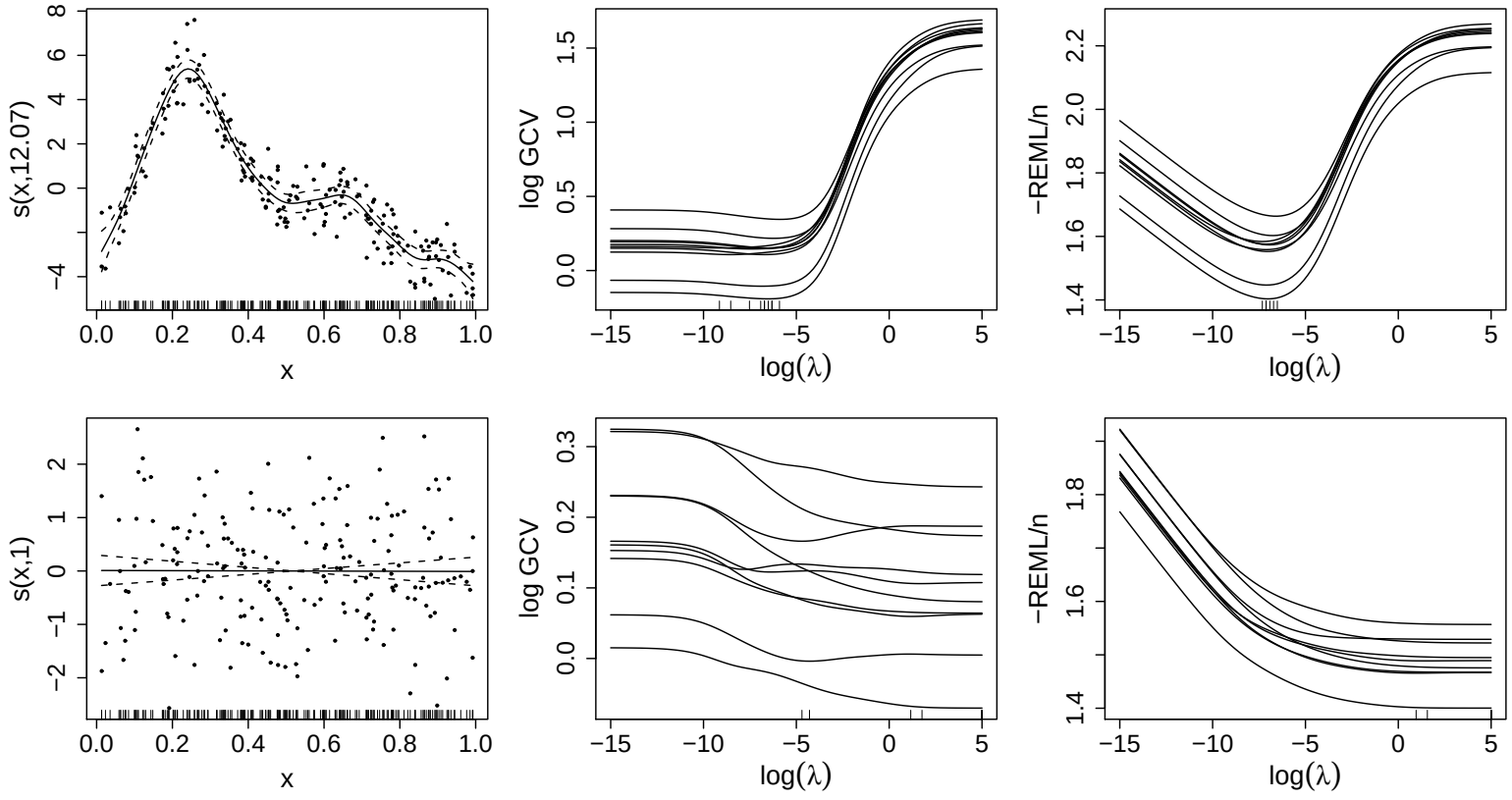
- That is, we are taking the expectation of the likelihood over the prior on $\boldsymbol{\beta}$.
- From this perspective it is possible to plot why the approach is intuitively sensible.

Basic principle of ML smoothness selection



1. Choose λ to maximize the average likelihood of random draws from the prior implied by λ.
2. If λ too low, then almost all draws are too variable to have high likelihood. If λ too high, then draws all underfit and have low likelihood. The right λ maximizes the proportion of draws close enough to data to give high likelihood.
3. Formally, maximize e.g. $\mathcal{V}_r(\lambda) = \log \int f(\mathbf{y}|\boldsymbol{\beta}) f_\lambda(\boldsymbol{\beta}) d\boldsymbol{\beta}$.

Prediction error vs. likelihood $\boldsymbol{\lambda}$ estimation



1. Pictures show GCV and REML scores for different replicates from same truth.
2. Compared to REML, GCV penalizes overfit only weakly, and so tends to undersmooth.

Are smoothers *really* random effects?

- Most times that smooth functions are used in models, the modeller believes that the function is a fixed state of nature.
- i.e. the assumption is that the true function is something that would stay fixed on replication of the dataset.
- So we are *really* being Bayesian in treating the function as random.
- If the function was a true frequentist random effect then we would expect to get a different random draw from its prior at each dataset replication. This almost never makes sense.
- Does this mean that using mixed modelling methods is wrong?
- No. It just happens that the mixed model methods can conveniently compute the Bayesian answers for us.