Penalized Splines and GAMs Summary

Monday, November 27, 2023 07:56

Introduce λ, a smoothing/tuning/hyperparameter.
↳ 1-1 mapping between λ and C

Let's say we have the basis expansion $g(x_i) = \beta_0 + \beta_1 x_i + \sum_{m=1}^{M} \beta_{1+m} b_m(x_i)$ aka $y_i = g(x_i) + \epsilon_i$ $\quad \epsilon_i \overset{iid}{\sim} N(0, \sigma^2)$
and we want to constrain the β's.

Penalties: Introduce bias to decrease variance

Least squares solution: $\boxed{\hat{\beta} = (X'X + \lambda B)^{-1} X'Y}$

↳ a diagonal matrix with 0 and 1 entries selecting which coefs are penalized

$\lambda = 0 \rightarrow$ OLS
$\uparrow \lambda \downarrow$ slope
Can use CV to pick λ.

Recall:
$\uparrow \lambda = \uparrow$ shrinkage $= \downarrow$ variance
$\downarrow \lambda = \downarrow$ shrinkage $= \uparrow$ bias

RIDGE REGRESSION / L-2 PENALTY
Minimizes residuals and $\lambda \cdot \text{slope}^2$

- better to standardize before applying ridge regression

$\beta_2^2 + \beta_3^2 + \beta_4^2 + \ldots + \beta_{M+1}^2 = \underbrace{\|\beta^*\|_2^2}_{[\beta_2, \ldots, \beta_{M+1}]'} \leq \underbrace{C}_{\text{some constant}}$

- β's only approach 0

LASSO / L-1 PENALTY
Minimizes residuals and $\lambda \cdot |\text{slope}|$

$\|\beta^*\|_1 = \sum_{j=2}^{M+1} |\beta_j|$

- Will let β's go to 0

Let S be $X(X'X + \lambda B)^{-1}X'$ aka a smoothing matrix. $\hat{Y} = SY$
EFFECTIVE DF IS

$$\boxed{df_{eff} = tr(S)}$$

↓

$$\boxed{GCV = \frac{1}{n} \frac{\sum_{i=1}^{n} (y_i - \hat{y}_i)^2}{[1 - n^{-1} tr(S)]^2}}$$

Residual Error Variance Estimate

Recall our model is

$$y_i = g(x_i) + \epsilon_i \qquad \epsilon_i \overset{iid}{\sim} N(0, \sigma^2).$$

We now have an estimate $\hat{g}(x_i)$. How about σ^2?

We have two options:

$$\hat{\sigma}^2 = \frac{\sum_{i=1}^{n} [y_i - \hat{g}(x_i)]^2}{n - df_{eff}} \qquad (6)$$

The above is a biased estimate. Some software gives you the option to use

$$\hat{\sigma}^2_{unbiased} = \frac{\sum_{i=1}^{n} [y_i - \hat{g}(x_i)]^2}{n - 2tr(\mathbf{S}) + tr(\mathbf{SS}')} \qquad (7)$$

Variance of $\hat{g}(x_i)$

Now we can calculate uncertainty associated with $\hat{g}(x_i)$ at each x_i.

With slight abuse of notation, let x_i' be the row vector of basis function values for x_i.

The variance of $\hat{g}(x_i)$ is

$$\begin{aligned} Var[\hat{g}(x_i)] &= Var[x_i'\hat{\beta}] = x_i' Var[\hat{\beta}] x_i \\ &= x_i' Var[(\mathbf{X}'\mathbf{X} + \lambda\mathbf{B})^{-1}\mathbf{X}'\mathbf{Y}] x_i \\ &= \sigma^2 x_i' (\mathbf{X}'\mathbf{X} + \lambda\mathbf{B})^{-1}(\mathbf{X}'\mathbf{X})(\mathbf{X}'\mathbf{X} + \lambda\mathbf{B})^{-1} x_i. \end{aligned}$$

Note: you should decide whether or not to include the variance due to the intercept. If $x_i[1] = 1$, then the variance estimate of $\hat{g}(x_i)$ includes this source of uncertainty.

Confidence interval and prediction interval

Obtain point-wise confidence interval derived from previous expression by plugging in $\hat{\sigma}^2$ for σ^2.

If $\lambda = 0$: the previous equation reduces to the OLS variance.

Similarly the variance for an unobserved point y_i^* with covariate x_i^* has variance

$$Var[y_i^*] = \sigma^2 + \sigma^2 x_i^{*\prime} (\mathbf{X}'\mathbf{X} + \lambda\mathbf{B})^{-1}(\mathbf{X}'\mathbf{X})(\mathbf{X}'\mathbf{X} + \lambda\mathbf{B})^{-1} x_i^*.$$

Checking gam

The default is $k = 10$, such that highest possible EDF is 9 (because of identifiability constraint).

```
> gam.check(fit)
```

→ function we use to get diagnostic plots

```
Method: GCV   Optimizer: magic
Smoothing parameter selection converged after 5 iterations.
The RMS GCV score gradient at convergence was 7.242e-08 .
The Hessian was positive definite.
Model rank =  10 / 10

Basis dimension (k) checking results. Low p-value (k-index<1) may
indicate that k is too low, especially if edf is close to k'.

          k'  edf k-index p-value
s(Temp) 9.00 6.65    1.02    0.88
```

→ H_0: Basis dimension is adequate.
– this test not very useful

In general, "k" should be "notably" larger than edf
e.g. edf ≈ 5 ... here we will check fitting model w/ a larger basis dimension

For the deviance residuals plot, the data fits well and does not deviate too much from the straight line. This suggests that our normality assumption is not violated.
For the residuals vs. linear predictor plot, the residuals appear to be randomly scattered around zero, which means that the model may be adequate for capturing the underlying relationship of our data.
For the response vs. fitted values plot, it should look linear if k is sufficient, in which it does.
Finally, the histogram of residuals appears to be normal. Visually, it appears that there may be no issues.

A different approach to smoothing rather than GCV

Penalized splines as BLUPs

- GCV may undersmooth.
- An alternative is to treat the coefficients of the truncated polynomials as random effects, and then use BLUPs.
- For concreteness, consider a linear spline:

(or whatever basis you are using)

- Applies to non-clustered data (though you could use clustered?)
- smoothing trick

random slope

Data is not longitudinal here, so this is a little trick

$$y_i = \beta_0 + \beta_1 x_i + \sum_{m=1}^{M} \theta_m (x_i - \kappa_m)_+ + \epsilon_i,$$

$$\theta_m \overset{iid}{\sim} N(0, \tau^2), \quad \epsilon_i \overset{iid}{\sim} N(0, \sigma^2)$$

↳ τ^2 becomes larger with more wiggly data
↳ σ^2 becomes larger with more noisy data → if $\sigma^2 >> \tau^2$ then we will get ...?

↑ wiggly ↑ τ^2
↑ noisy ↑ σ^2

$$\beta = \begin{bmatrix} \beta_0 \\ \beta_1 \end{bmatrix} \quad \theta = \begin{bmatrix} \theta_1 \\ \vdots \\ \theta_M \end{bmatrix} \quad \mathbf{X} = \begin{bmatrix} 1 & x_1 \\ \vdots & \vdots \\ 1 & x_n \end{bmatrix} \quad \mathbf{Z} = \begin{bmatrix} (x_1 - \kappa_1)_+ & \cdots & (x_1 - \kappa_M)_+ \\ \vdots & & \vdots \\ (x_n - \kappa_1)_+ & \cdots & (x_n - \kappa_M)_+ \end{bmatrix}$$

Mixed model for estimating a penalized spline

$\begin{bmatrix} \vdots & x_1 \\ \vdots & \vdots \\ \vdots & x_n \end{bmatrix}$ $\quad \begin{bmatrix} (x_1 - \kappa_1)_+ & \cdots & (x_1 - \kappa_M)_+ \\ \vdots & & \vdots \\ (x_n - \kappa_1)_+ & & (x_n - \kappa_M)_+ \end{bmatrix}$

Given τ^2 and σ^2, we seek to minimize

$$\frac{1}{\sigma^2}\|\mathbf{Y} - \mathbf{X}\beta - \mathbf{Z}\theta\|^2 + \frac{1}{\tau^2}\|\theta\|_2^2,$$

$\rightarrow \|Y - X\beta - Z\theta\|^2 + \frac{\sigma^2}{\tau^2}\|\theta\|_2^2$

if σ^2 is very large → large penalty
→ shrink coefs towards 0

which we can think of ridge regression with penalty $\lambda = \frac{\sigma^2}{\tau^2}$.

We estimate all parameters from the data using the mixed modeling tools we previously learned, and thus obtain a model-based estimate of λ.

Interpretation

We interpret smoothers $\hat{g}(x_i)$ by looking at plots.

We can add some details regarding the slopes at particular x_i.

Deaths are highest at cold temperatures (< 10 degrees F) and slightly decreasing until approximately 20 degrees. Then deaths decrease at a similar rate from approximately 25 to 75 degrees. The number of deaths decreases by approximately 0.5 people / degree in a neighborhood of 40 degrees. Then the number of deaths starts to increase around 75 degrees. At 85 degrees, the number of deaths increases by approximately 0.8 for every 1 degree increase in temperature.

$\hat{g}'(40)$

$\hat{g}'(85)$

In linear regression $\hat{y}_i = \hat{\beta}_0 + \hat{\beta}_1 x_i$.

In GAMs, we have $\hat{y}_i = \hat{\beta}_0 + \hat{g}(x_i)$, and slope changes with x_i.

GAM notation

Using mathematical notation, with $g_j(x_{ij})$ for smooth terms, write out a statistical model with normal errors for the cross-sectional data with y_i denoting hippocampus volume in cm^3 and the covariates `DX` , `PTGENDER` , `APOE4` (treating APOE4 as a factor with reference level 0), a smooth for `AGE_VISIT` , and a smooth for `PTEDUCAT` . Include distributional assumptions. (You do not need to write the basis functions used for the smooth terms $g_j(x_{ij})$, nor the penalized objective function.)

$$y_i \overset{iid}{\sim} N(\mu_i, \sigma^2)$$

where y_i is hippocampus volume in cm^3 for individual i.

$$y_i = \beta_0 + \beta_1 DX_i + \beta_2 PTGENDER_i + \beta_3 APOE4_{1i} + \beta_4 APOE4_{2i} + g_1(AGEVISIT_i) + g_2(PTEDUCAT_i) + \epsilon$$

$$\epsilon \overset{iid}{\sim} N(0, \sigma^2)$$

where ϵ is our error and σ^2 is our variance.

β_0 is the baseline hippocampal volume without considering other covariates.
β_1 is the estimated effect for diagnosis, with 0 as the reference level. β_2 is the estimated effect for gender, with female as the reference level. β_3 is the estimated effect for APOE4 with 1 copy of the allele, with the reference group being 0. β_4 is the estimated effect for APOE4 with 2 copies of the allele, with the reference group being 0. $g_1(AGEVISIT_i)$ is the smoothing term for age_visit. $g_2(PTEDUCAT_i)$ is the smoothing term for pteducat.

State the null hypothesis for each p-value in the spline basis dimension diagnostics in `gam.check()` . Then state whether it is accepted or rejected at $\alpha = 0.05$.

H0: The basis dimension for AGE_VISIT is adequate
alpha = 0.05
p-value: 0.675
Decision: Fail to reject null hypothesis
Conclusion: The basis dimension for AGE_VISIT is adequate

H0: The basis dimension for PTEDUCAT is adequate
alpha = 0.05
p-value: 0.005
Decision: Reject null hypothesis
Conclusion: The basis dimension for PTEDUCAT may not be adequate

Interpretation of model

- Daily counts of non-accidental mortality were positively associated with previous-day $PM_{2.5}$ level among those 65 or above.
- Approximately linear on log scale.

```
> newd <- health[1, ] # grab any row
> newd$pm25.lag1 <- 15 - 1e-05 # subtract some small number
> y1 <- predict(fit, newd)
> newd$pm25.lag1 <- 15 + 1e-05 # add some small number
> y2 <- predict(fit, newd)
> (y2 - y1)/2e-05
          1
0.001054248
```

- We estimated a 0.00105 increase in log death count per unit ($\mu g/m^3$) increase in $PM_{2.5}$ level in the previous day.
- 10 unit increase in $PM_{2.5}$ level from the previous day increased daily death counts by $100 * (e^{(0.00105 \times 10)} - 1)\% = 1.06\%$.

Extracting Effects from a Smooth Function

We want to estimate the effect of a 10 $\mu g/m^3$ increase in $PM_{2.5}$ levels from 40 to 50 $\mu g/m^3$ when temperature = 70F.

So our estimate is 0.022 (95%CI 0.001, 0.044).

The same effect of a 10 unit increase in $PM_{2.5}$ levels from 20 to 30 $\mu g/m^3$ is 0.020 (95%CI 0.004, 0.036).

These two estimates are very similar because the $PM_{2.5}$ effect appears to be quite linear.