

# Representing Additive Models as Mixed Models

Katharina Ring

LMU Seminar: Mixed and Semiparametric Models

January 14, 2020

# Truncated Power Basis

univariate:

$$y = \theta_0 + \theta_1 x + \theta_2 x^2 + \dots + \theta_d x^d + \sum_{k=1}^K \theta_{dk} (x - \kappa_k)_+^d + \epsilon$$

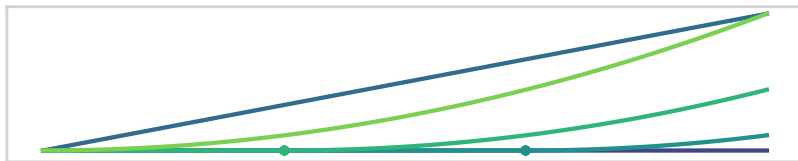
# Truncated Power Basis

univariate:

$$y = \theta_0 + \theta_1 x + \theta_2 x^2 + \dots + \theta_d x^d + \sum_{k=1}^K \theta_{dk} (x - \kappa_k)_+^d + \epsilon$$

univariate and quadratic with two knots:

$$y = \theta_0 + \theta_1 x + \theta_2 x^2 + \theta_{21} (x - \kappa_1)_+^2 + \theta_{22} (x - \kappa_2)_+^2 + \epsilon$$



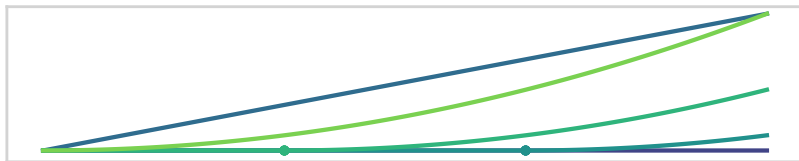
# Truncated Power Basis

univariate:

$$y = \theta_0 + \theta_1 x + \theta_2 x^2 + \dots + \theta_d x^d + \sum_{k=1}^K \theta_{dk} (x - \kappa_k)_+^d + \epsilon$$

univariate and quadratic with two knots:

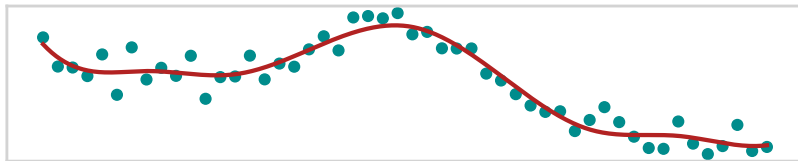
$$y = \overbrace{\theta_0 + \theta_1 x + \theta_2 x^2}^{\text{fixed effects}} + \overbrace{\theta_{21}(x - \kappa_1)_+^2 + \theta_{22}(x - \kappa_2)_+^2}^{\text{random effects: depend on } i} + \epsilon$$



# Additive Models

Semiparametric regression:

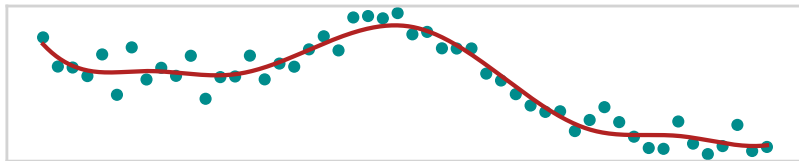
$$\hat{y}_i = f(\nu_i) + u_i^T \gamma$$



# Additive Models

Semiparametric regression:

$$\hat{y}_i = \underbrace{f(\nu_i^T \xi)}_{\nu_i^T \xi} + u_i^T \gamma$$



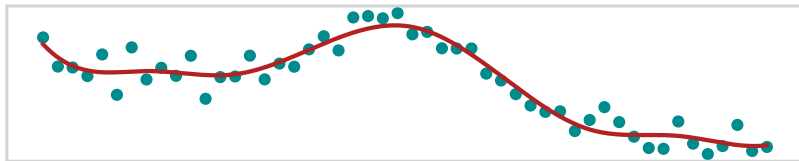
# Additive Models

Semiparametric regression:

$$\hat{y}_i = \underbrace{f(\nu_i^T \xi)}_{\nu_i^T \xi} + u_i^T \gamma$$

In matrix notation:

$$\hat{y} = V\xi + U\gamma$$

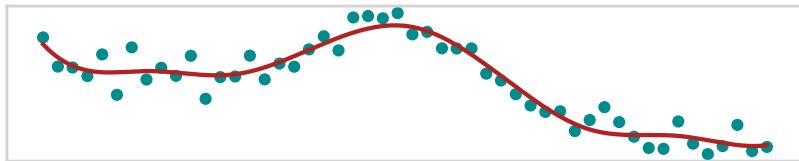


# Additive Models

Semiparametric regression:

$$\hat{y}_i = \underbrace{v_{i1}^T \xi_1}_{f_1(\nu_{i1})} + \dots + \underbrace{v_{ip}^T \xi_p}_{f_p(\nu_{ip})} + u_i^T \gamma$$

In matrix notation:





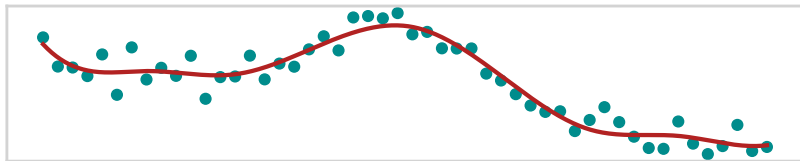
# Additive Models

Semiparametric regression:

$$\hat{y}_i = \underbrace{v_{i1}^T \xi_1}_{f_1(\nu_{i1})} + \dots + \underbrace{v_{ip}^T \xi_p}_{f_p(\nu_{ip})} + u_i^T \gamma$$

In matrix notation:

$$\hat{y} = V_1 \xi_1 + \dots + V_p \xi_p + U \gamma = \sum_{j=1}^p V_j \xi_j + U \gamma$$



## Splines

Spline functions are **piecewise polynomial segments** (called basis functions) joined together smoothly at so-called knots.

$$\hat{y} = V\xi + U\gamma$$



# Splines

Spline functions are **piecewise polynomial segments** (called basis functions) joined together smoothly at so-called knots.

$$\hat{y} = V\xi + U\gamma = \begin{pmatrix} b_1(x_1) & \dots & b_k(x_1) \\ \vdots & \ddots & \vdots \\ b_1(x_n) & \dots & b_k(x_n) \end{pmatrix} \begin{pmatrix} \xi_1 \\ \vdots \\ \xi_k \end{pmatrix} + \begin{pmatrix} 1 & x_1 \\ \vdots & \vdots \\ 1 & x_n \end{pmatrix} \begin{pmatrix} \gamma_1 \\ \gamma_2 \end{pmatrix}$$



## Splines

Spline functions are **piecewise polynomial segments** (called basis functions) joined together smoothly at so-called knots.

$$\hat{y} = V\xi + U\gamma = \begin{pmatrix} b_1(x_1) & \dots & b_k(x_1) \\ \vdots & \ddots & \vdots \\ b_1(x_n) & \dots & b_k(x_n) \end{pmatrix} \begin{pmatrix} \xi_1 \\ \vdots \\ \xi_k \end{pmatrix} + \begin{pmatrix} 1 & x_1 \\ \vdots & \vdots \\ 1 & x_n \end{pmatrix} \begin{pmatrix} \gamma_1 \\ \gamma_2 \end{pmatrix}$$

with basis functions  $b_1(\cdot), \dots, b_k(\cdot)$ , e. g. *B-spline*, truncated power basis, natural cubic spline, ...

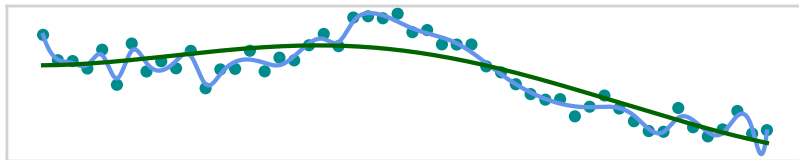


# Roughness Penalty

Penalized Regression Spline:

$$\log L(\xi, \gamma) + \lambda \int_{x_1}^{x_n} [f''(x)]^2 dx$$

Control wiggleness (bias-variance tradeoff):

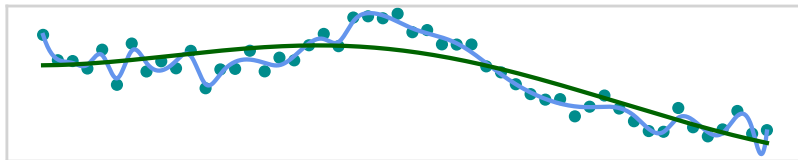


# Roughness Penalty

Penalized Regression Spline:

$$\log L(\xi, \gamma) + \lambda \xi^T K \xi$$

e. g. first order differences  $\xi^T K \xi = \sum (\xi_{k+1} - \xi_k)^2$ :

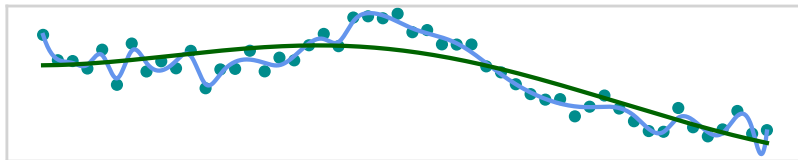


# Roughness Penalty

Penalized Regression Spline:

$$\log L(\xi, \gamma) + \lambda \xi^T K \xi$$

Problem: How to choose  $\lambda$ ?



# Mixed Models

$$y_i = \underbrace{X_i \beta}_{\text{fixed effects}} + \underbrace{Z_i b_i}_{\text{random effects}} + \epsilon_i$$



## Mixed Models

$$y_i = \underbrace{X_i\beta}_{\text{fixed effects}} + \underbrace{Z_i b_i}_{\text{random effects}} + \epsilon_i$$

### Classical View

random effects reflect that the individuals/  
clusters are a **random sample** of a larger  
population (not always appropriate)

## Mixed Models

$$y_i = \underbrace{X_i\beta}_{\text{fixed effects}} + \underbrace{Z_i b_i}_{\text{random effects}} + \epsilon_i$$

### Classical View

random effects reflect that the individuals/ clusters are a **random sample** of a larger population (not always appropriate)

### Marginal View

random effects induce a general linear model with **correlated errors**

## Mixed Models

$$y_i = \underbrace{X_i\beta}_{\text{fixed effects}} + \underbrace{Z_i b_i}_{\text{random effects}} + \epsilon_i$$

### Classical View

random effects reflect that the individuals/ clusters are a **random sample** of a larger population (not always appropriate)

### Marginal View

random effects induce a general linear model with **correlated errors**

### Bayesian View

the random effects distribution is a **prior** on the random effects

## Mixed Models

$$y_i = \underbrace{X_i\beta}_{\text{fixed effects}} + \underbrace{Z_i b_i}_{\text{random effects}} + \epsilon_i$$

### Classical View

random effects reflect that the individuals/clusters are a **random sample** of a larger population (not always appropriate)

### Marginal View

random effects induce a general linear model with **correlated errors**

### Bayesian View

the random effects distribution is a **prior** on the random effects

### Penalization View

the random effects distribution results in a **penalty** on the random effects leading to **shrinkage**

## Mixed Models

$$y_i = \underbrace{X_i \beta}_{\text{fixed effects}} + \underbrace{Z_i b_i}_{\text{random effects}} + \epsilon_i$$

### Classical View

random effects reflect that the individuals/clusters are a **random sample** of a larger population (not always appropriate)

### Marginal View

random effects induce a general linear model with **correlated errors**

### Bayesian View

the random effects distribution is a **prior** on the random effects

### Penalization View

the random effects distribution results in a **penalty** on the random effects leading to **shrinkage**

# Empirical Bayes

- **Idea:** Use Mixed Model inference:  $y = \sum_{j=1}^p \underbrace{V_p \xi_p}_{\text{random effects}} + \underbrace{U\gamma}_{\text{fixed effects}} + \epsilon$

# Empirical Bayes

- **Idea:** Use Mixed Model inference:  $y = \sum_{j=1}^p \underbrace{V_p \xi_p}_{\text{random effects}} + \underbrace{U\gamma}_{\text{fixed effects}} + \epsilon$

Prior:  $p(\gamma) \propto \text{const.}$

# Empirical Bayes

- **Idea:** Use Mixed Model inference:  $y = \sum_{j=1}^p \underbrace{V_p \xi_p}_{\text{random effects}} + \underbrace{U\gamma}_{\text{fixed effects}} + \epsilon$

Prior:  $p(\gamma) \propto \text{const.}$

Prior:  $p(\xi_j | \tau_j^2) \propto \exp \left( -\frac{1}{2\tau_j^2} \xi_j^T \Sigma_j^{-1} \xi_j \right)$



# Empirical Bayes

- **Idea:** Use Mixed Model inference:  $y = \sum_{j=1}^p \underbrace{V_p \xi_p}_{\text{random effects}} + \underbrace{U\gamma}_{\text{fixed effects}} + \epsilon$

Prior:  $p(\gamma) \propto \text{const.}$

Prior:  $p(\xi_j | \tau_j^2) \propto \exp\left(-\frac{1}{2\tau_j^2} \xi_j^T \Sigma_j^{-1} \xi_j\right)$

Posterior:  $p(\xi_1, \dots, \xi_p, \gamma | y) \propto L(y, \xi_1, \dots, \xi_p, \gamma) \prod_{j=1}^p p(\xi_j | \tau_j^2)$

# Empirical Bayes

- Idea:** Use Mixed Model inference:  $y = \sum_{j=1}^p \underbrace{V_p \xi_p}_{\text{random effects}} + \underbrace{U\gamma}_{\text{fixed effects}} + \epsilon$

**Prior:**  $p(\gamma) \propto \text{const.}$

**Prior:**  $p(\xi_j | \tau_j^2) \propto \exp \left( -\frac{1}{2\tau_j^2} \xi_j^T \Sigma_j^{-1} \xi_j \right)$

**Posterior:**  $p(\xi_1, \dots, \xi_p, \gamma | y) \propto L(y, \xi_1, \dots, \xi_p, \gamma) \prod_{j=1}^p p(\xi_j | \tau_j^2)$

- Maximum Likelihood for  $\tau_j^2$  (so far treated as fixed):

$$\max_{\tau_1, \dots, \tau_p} \log L(\gamma, \xi_1, \dots, \xi_p) - \sum_{j=1}^p \underbrace{\frac{1}{2\tau_j^2}}_{\lambda_j} \xi_j^T \underbrace{\Sigma_j^{-1}}_{K_j} \xi_j$$

# Empirical Bayes

- Idea:** Use Mixed Model inference:  $y = \sum_{j=1}^p \underbrace{V_p \xi_p}_{\text{random effects}} + \underbrace{U\gamma}_{\text{fixed effects}} + \epsilon$

**Prior:**  $p(\gamma) \propto \text{const.}$

**Prior:**  $p(\xi_j | \tau_j^2) \propto \exp \left( -\frac{1}{2\tau_j^2} \xi_j^T \Sigma_j^{-1} \xi_j \right)$

**Posterior:**  $p(\xi_1, \dots, \xi_p, \gamma | y) \propto L(y, \xi_1, \dots, \xi_p, \gamma) \prod_{j=1}^p p(\xi_j | \tau_j^2)$

- Maximum Likelihood for  $\tau_j^2$  (so far treated as fixed):

$$\max_{\tau_1, \dots, \tau_p} \log L(\gamma, \xi_1, \dots, \xi_p) - \sum_{j=1}^p \underbrace{\frac{1}{2\tau_j^2}}_{\lambda_j} \xi_j^T \underbrace{\Sigma_j^{-1}}_{K_j} \xi_j$$

$\Rightarrow$  Empirical Bayes is equivalent to penalized Maximum Likelihood

# Mixed Model Representation

**Idea:** Use Mixed Model inference:  $y = \sum_{j=1}^p \underbrace{V_p \xi_p}_{\text{random effects}} + \underbrace{U\gamma}_{\text{fixed effects}} + \epsilon$

## Mixed Model Representation

**Idea:** Use Mixed Model inference:  $y = \sum_{j=1}^p \underbrace{V_p \xi_p}_{\text{random effects}} + \underbrace{U\gamma}_{\text{fixed effects}} + \epsilon$

**Problem:**  $K_j$  as precision matrix is problematic as  $K_j$  is often rank deficient, e. g.  $\xi^T K \xi = \sum (\xi_{k+1} - \xi_k)^2 \rightarrow \xi_1$  not penalized:

The Gaussian prior  $p(\xi_j | \tau_j^2) \propto \exp\left(-\frac{1}{2\tau_j^2} \xi_j^T K_j \xi_j\right)$  is improper.

## Mixed Model Representation

**Idea:** Use Mixed Model inference:  $y = \underbrace{\sum_{j=1}^p V_p \xi_p}_{\text{random effects}} + \underbrace{U\gamma}_{\text{fixed effects}} + \epsilon$

**Problem:**  $K_j$  as precision matrix is problematic as  $K_j$  is often rank deficient, e. g.  $\xi^T K \xi = \sum (\xi_{k+1} - \xi_k)^2 \rightarrow \xi_1$  not penalized:

The Gaussian prior  $p(\xi_j | \tau_j^2) \propto \exp\left(-\frac{1}{2\tau_j^2} \xi_j^T K_j \xi_j\right)$  is improper.

**Solution:** Separate  $\xi_j$  into  $\xi_j = \tilde{X}_j \beta_j + \tilde{Z}_j b_j$ :

## Mixed Model Representation

**Idea:** Use Mixed Model inference:  $y = \underbrace{\sum_{j=1}^p V_p \xi_p}_{\text{random effects}} + \underbrace{U\gamma}_{\text{fixed effects}} + \epsilon$

**Problem:**  $K_j$  as precision matrix is problematic as  $K_j$  is often rank deficient, e. g.  $\xi^T K \xi = \sum (\xi_{k+1} - \xi_k)^2 \rightarrow \xi_1$  not penalized:

The Gaussian prior  $p(\xi_j | \tau_j^2) \propto \exp\left(-\frac{1}{2\tau_j^2} \xi_j^T K_j \xi_j\right)$  is improper.

**Solution:** Separate  $\xi_j$  into  $\xi_j = \tilde{X}_j \beta_j + \tilde{Z}_j b_j$ :

- $\beta$ : non-penalized parts with a flat prior  
 $\dim(\beta_j) = \dim(\xi_j) - \text{rank}(K_j)$

## Mixed Model Representation

**Idea:** Use Mixed Model inference:  $y = \underbrace{\sum_{j=1}^p V_p \xi_p}_{\text{random effects}} + \underbrace{U\gamma}_{\text{fixed effects}} + \epsilon$

**Problem:**  $K_j$  as precision matrix is problematic as  $K_j$  is often rank deficient, e. g.  $\xi^T K \xi = \sum (\xi_{k+1} - \xi_k)^2 \rightarrow \xi_1$  not penalized:

The Gaussian prior  $p(\xi_j | \tau_j^2) \propto \exp\left(-\frac{1}{2\tau_j^2} \xi_j^T K_j \xi_j\right)$  is improper.

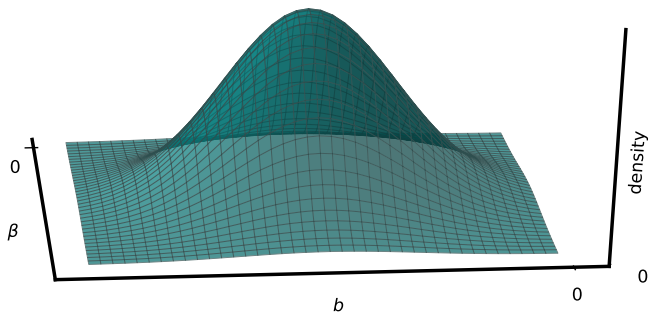
**Solution:** Separate  $\xi_j$  into  $\xi_j = \tilde{X}_j \beta_j + \tilde{Z}_j b_j$ :

- $\beta$ : non-penalized parts with a flat prior  
 $\dim(\beta_j) = \dim(\xi_j) - \text{rank}(K_j)$
- $b$ : penalized parts with a proper (Gaussian) prior  
 $\dim(b_j) = \text{rank}(K_j)$



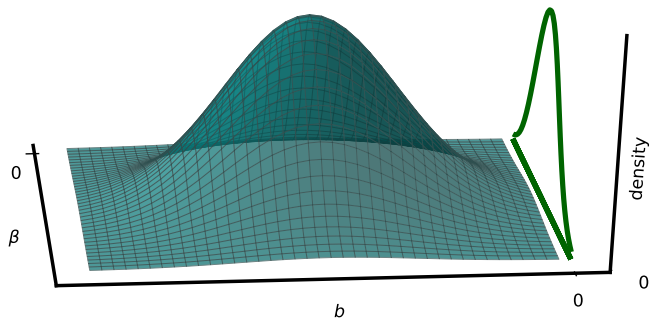
# Mixed Model Representation

unpenalized likelihood



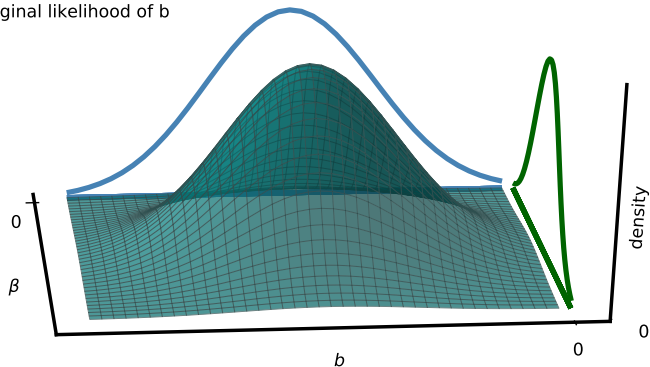
# Mixed Model Representation

- unpenalized likelihood
- marginal likelihood of  $\beta$



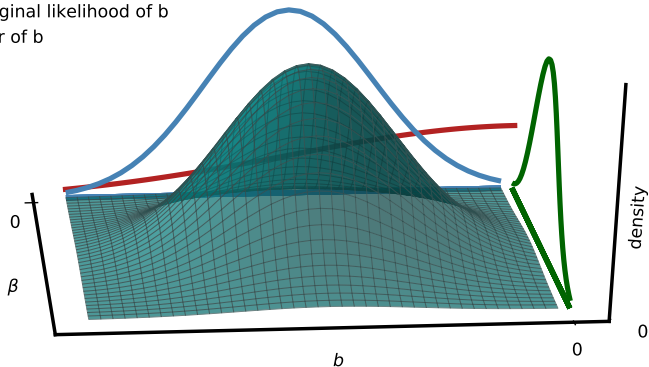
# Mixed Model Representation

- unpenalized likelihood
- marginal likelihood of  $\beta$
- marginal likelihood of  $b$

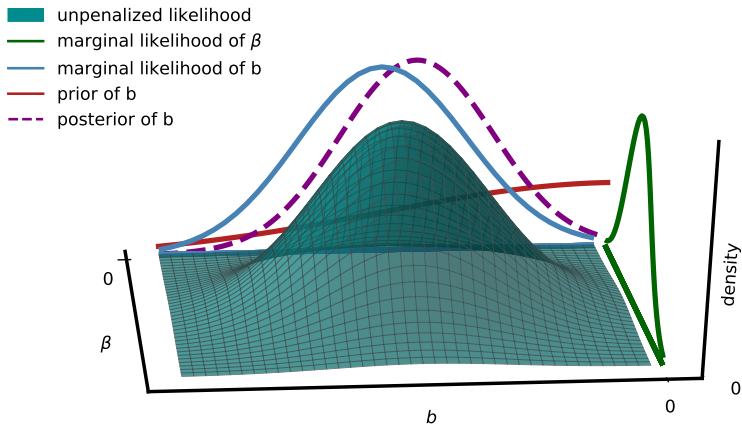


# Mixed Model Representation

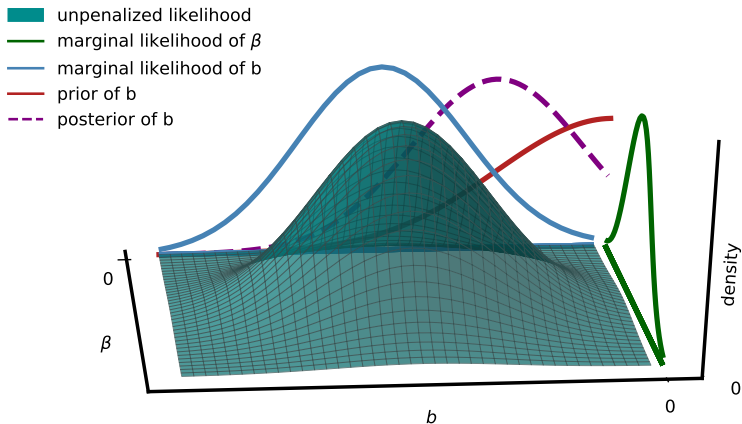
- unpenalized likelihood
- marginal likelihood of  $\beta$
- marginal likelihood of  $b$
- prior of  $b$



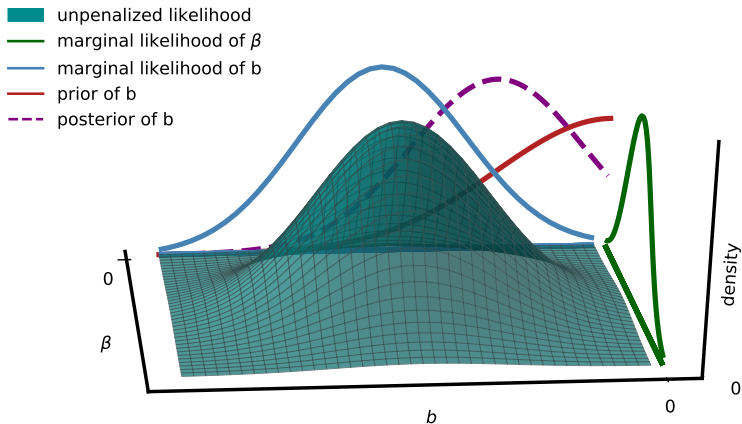
# Mixed Model Representation



# Mixed Model Representation



# Mixed Model Representation



The lower the prior variance, the higher the penalty!

# Mixed Model Representation

Decomposition  $\xi_j = \tilde{X}_j\beta_j + \tilde{Z}_jb_j$ :

$$y = \sum_{j=1}^p V_j \xi_j + U\gamma$$



# Mixed Model Representation

Decomposition  $\xi_j = \tilde{X}_j\beta_j + \tilde{Z}_jb_j$ :

$$y = \sum_{j=1}^p V_j \overbrace{(\tilde{X}_j\beta_j + \tilde{Z}_jb_j)}^{\xi_j} + U\gamma$$

# Mixed Model Representation

Decomposition  $\xi_j = \tilde{X}_j\beta_j + \tilde{Z}_jb_j$ :

$$y = \sum_{j=1}^p V_j \overbrace{(\tilde{X}_j\beta_j + \tilde{Z}_jb_j)}^{\xi_j} + U\gamma = X\beta + Zb$$

$$\beta := (\beta_1^T, \dots, \beta_p^T, \gamma^T)$$

$$b := (b_1^T, \dots, b_p^T)$$

$$Z := V_j\tilde{Z}_j$$

$$X := (V_j\tilde{X}_j, U)$$

# Mixed Model Representation

Decomposition  $\xi_j = \tilde{X}_j\beta_j + \tilde{Z}_jb_j$ :

$$y = \sum_{j=1}^p V_j \overbrace{(\tilde{X}_j\beta_j + \tilde{Z}_jb_j)}^{\xi_j} + U\gamma = X\beta + Zb$$

$$\beta := (\beta_1^T, \dots, \beta_p^T, \gamma^T)$$

$$b := (b_1^T, \dots, b_p^T)$$

$$Z := V_j\tilde{Z}_j$$

$$X := (V_j\tilde{X}_j, U)$$

Requirements:

# Mixed Model Representation

Decomposition  $\xi_j = \tilde{X}_j\beta_j + \tilde{Z}_jb_j$ :

$$\beta := (\beta_1^T, \dots, \beta_p^T, \gamma^T)$$

$$b := (b_1^T, \dots, b_p^T)$$

$$Z := V_j\tilde{Z}_j$$

$$X := (V_j\tilde{X}_j, U)$$

$$y = \sum_{j=1}^p V_j \overbrace{(\tilde{X}_j\beta_j + \tilde{Z}_jb_j)}^{\xi_j} + U\gamma = X\beta + Zb$$

Requirements:

1. 1-to-1 transformation: matrix  $(\tilde{X}_j \ \tilde{Z}_j)$  has full rank

## Mixed Model Representation

Decomposition  $\xi_j = \tilde{X}_j\beta_j + \tilde{Z}_jb_j$ :

$$y = \sum_{j=1}^p V_j \overbrace{(\tilde{X}_j\beta_j + \tilde{Z}_jb_j)}^{\xi_j} + U\gamma = X\beta + Zb$$

$$\beta := (\beta_1^T, \dots, \beta_p^T, \gamma^T)$$

$$b := (b_1^T, \dots, b_p^T)$$

$$Z := V_j\tilde{Z}_j$$

$$X := (V_j\tilde{X}_j, U)$$

Requirements:

1. 1-to-1 transformation: matrix  $(\tilde{X}_j \ \tilde{Z}_j)$  has full rank
2.  $\tilde{X}_j$  and  $\tilde{Z}_j$  are orthogonal:  $\tilde{X}_j^T \tilde{Z}_j = 0$

# Mixed Model Representation

Decomposition  $\xi_j = \tilde{X}_j\beta_j + \tilde{Z}_jb_j$ :

$$y = \sum_{j=1}^p V_j \overbrace{(\tilde{X}_j\beta_j + \tilde{Z}_jb_j)}^{\xi_j} + U\gamma = X\beta + Zb$$

$$\beta := (\beta_1^T, \dots, \beta_p^T, \gamma^T)$$

$$b := (b_1^T, \dots, b_p^T)$$

$$Z := V_j\tilde{Z}_j$$

$$X := (V_j\tilde{X}_j, U)$$

Requirements:

1. 1-to-1 transformation: matrix  $(\tilde{X}_j \ \tilde{Z}_j)$  has full rank
2.  $\tilde{X}_j$  and  $\tilde{Z}_j$  are orthogonal:  $\tilde{X}_j^T \tilde{Z}_j = 0$
3.  $\beta_j$  not penalized by  $K_j$ :  $\tilde{X}_j^T K_j \tilde{X}_j = 0$

# Mixed Model Representation

Decomposition  $\xi_j = \tilde{X}_j\beta_j + \tilde{Z}_jb_j$ :

$$y = \sum_{j=1}^p V_j \overbrace{(\tilde{X}_j\beta_j + \tilde{Z}_jb_j)}^{\xi_j} + U\gamma = X\beta + Zb$$

$$\beta := (\beta_1^T, \dots, \beta_p^T, \gamma^T)$$

$$b := (b_1^T, \dots, b_p^T)$$

$$Z := V_j\tilde{Z}_j$$

$$X := (V_j\tilde{X}_j, U)$$

Requirements:

1. 1-to-1 transformation: matrix  $(\tilde{X}_j \ \tilde{Z}_j)$  has full rank
2.  $\tilde{X}_j$  and  $\tilde{Z}_j$  are orthogonal:  $\tilde{X}_j^T \tilde{Z}_j = 0$
3.  $\beta_j$  not penalized by  $K_j$ :  $\tilde{X}_j^T K_j \tilde{X}_j = 0$
4. Gaussian prior for  $b_j$ :  $\tilde{Z}_j^T K_j \tilde{Z}_j = I_{k_j}$

# Choosing $\tilde{X}_j$ and $\tilde{Z}_j$ for Mixed Model Representation

## Recap: Requirements

1. 1-on-1 transformation: matrix  $(\tilde{X}_j \ \tilde{Z}_j)$  has full rank
2.  $\tilde{X}_j$  and  $\tilde{Z}_j$  are orthogonal:  $\tilde{X}_j^T \tilde{Z}_j = 0$
3.  $\beta_j$  not penalized by  $K_j$ :  $\tilde{X}_j^T K_j \tilde{X}_j = 0$
4. Gaussian prior for  $b_j$ :  $\tilde{Z}_j^T K_j \tilde{Z}_j = I_{k_j}$

## Setup

- $\tilde{X}_j$  is a basis of the null space of  $K_j$  (condition 3)
- $\tilde{Z}_j = L_j(L_j^T L_j)^{-1}$  with  $K_j = L_j L_j^T$  (conditions 1 and 4)
- Choose  $L_j$  s. t.  $L_j^T \tilde{X}_j = 0$  and  $\tilde{X}_j L_j^T = 0$  (condition 2)  
e. g. spectral decomposition:  $K_j = \Gamma_j \Lambda_j \Gamma_j^T$ , so  $L_j = \Gamma \Lambda_j^{1/2}$



# Mixed Model Representation

log-Prior:

$$\log p(\xi_j | \tau_j^2) \propto -\frac{1}{2\tau_j^2} \xi_j^T K_j \xi_j$$

# Mixed Model Representation

log-Prior:

$$\log p(\xi_j | \tau_j^2) \propto -\frac{1}{2\tau_j^2} \xi_j^T K_j \xi_j = -\frac{1}{2\tau_j^2} b_j^T b_j$$

# Mixed Model Representation

log-Prior:

$$\log p(\xi_j | \tau_j^2) \propto -\frac{1}{2\tau_j^2} \xi_j^T K_j \xi_j = -\frac{1}{2\tau_j^2} b_j^T b_j$$

$$\Rightarrow p(\beta) \propto \text{const.}$$

# Mixed Model Representation

log-Prior:

$$\log p(\xi_j | \tau_j^2) \propto -\frac{1}{2\tau_j^2} \xi_j^T K_j \xi_j = -\frac{1}{2\tau_j^2} b_j^T b_j$$

$$\Rightarrow p(\beta) \propto \text{const.}$$

$$\Rightarrow p(b_j) \sim N(0, \tau_j^2 I_{k_j})$$

# Mixed Model Representation

log-Prior:

$$\log p(\xi_j | \tau_j^2) \propto -\frac{1}{2\tau_j^2} \xi_j^T K_j \xi_j = -\frac{1}{2\tau_j^2} b_j^T b_j$$

$$\Rightarrow p(\beta) \propto \text{const.}$$

$$\Rightarrow p(b_j) \sim N(0, \tau_j^2 I_{k_j})$$

log-Posterior:

$$l_p(\beta, b|y) = l(y, \beta, b) - \sum_{j=1}^p \overbrace{\frac{1}{2\tau_j^2}}^{=\lambda} b_j^T b_j$$

# Estimates $\hat{\beta}$ and $\hat{b}$

In order to maximize the (log-)Posterior (equivalent to ML), derive estimates for  $\beta$  and  $b$  simultaneously based on known  $\sigma^2$  and  $\tau^2$ .

# Estimates $\hat{\beta}$ and $\hat{b}$

In order to maximize the (log-)Posterior (equivalent to ML), derive estimates for  $\beta$  and  $b$  simultaneously based on known  $\sigma^2$  and  $\tau^2$ .

Mixed Model equations:

$$\overbrace{\begin{pmatrix} X^T W X & X^T W Z \\ Z^T W X & Z^T W Z + Q^{-1} \end{pmatrix}}^{\text{Fisher information}} \begin{pmatrix} \hat{\beta} \\ \hat{b} \end{pmatrix} = \begin{pmatrix} X^T W y \\ Z^T W y \end{pmatrix}$$

with  $W = \text{diag}(\sigma^2)$  and  $Q = \text{blockdiag}(\tau_1^2 I_{k_1}, \dots, \tau_p^2 I_{k_p})$

# Variance Estimates

## Maximum Likelihood (integrate $b$ out)

- uses (partially) marginal distribution  $y \sim N(X\beta, \Sigma)$ 
  1. Derive  $\hat{\beta}$  analytically
  2. Plug in to get profile likelihood for  $\tau^2$  and  $\sigma^2$
  3. Maximize numerically
- estimates variance components of posterior mode



# Variance Estimates

## Maximum Likelihood (integrate $b$ out)

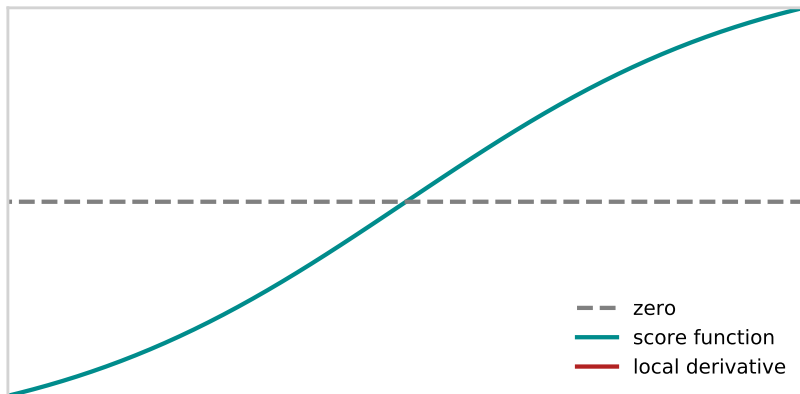
- uses (partially) marginal distribution  $y \sim N(X\beta, \Sigma)$ 
  1. Derive  $\hat{\beta}$  analytically
  2. Plug in to get profile likelihood for  $\tau^2$  and  $\sigma^2$
  3. Maximize numerically
- estimates variance components of posterior mode

## Restricted ML (integrate $b$ and $\beta$ out)

- directly uses marginal distribution of  $y$
- Advantages over ML:
  - + considers loss of degrees of freedom due to estimation of  $\beta$
  - + estimates mode of the marginal posterior for the variances

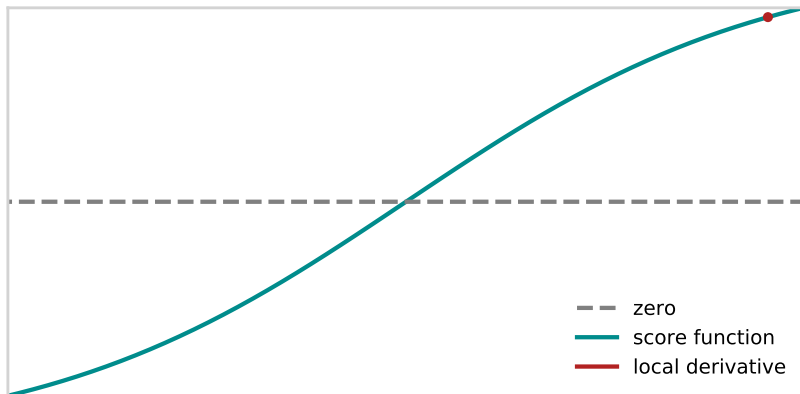
## Variance Estimates $\hat{\sigma}^2$ and $\hat{\tau}^2$

Maximize the restricted likelihood (REML) using Newton-Raphson:



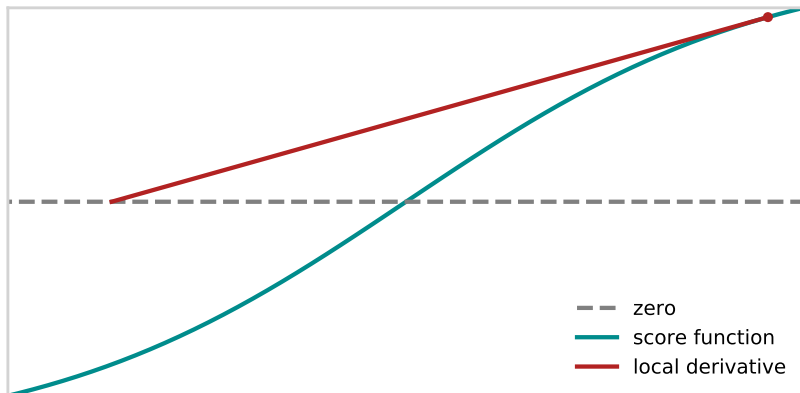
## Variance Estimates $\hat{\sigma}^2$ and $\hat{\tau}^2$

Maximize the restricted likelihood (REML) using Newton-Raphson:



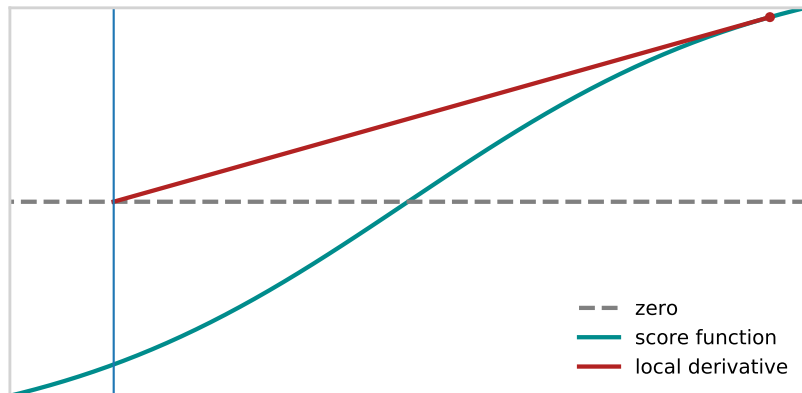
## Variance Estimates $\hat{\sigma}^2$ and $\hat{\tau}^2$

Maximize the restricted likelihood (REML) using Newton-Raphson:



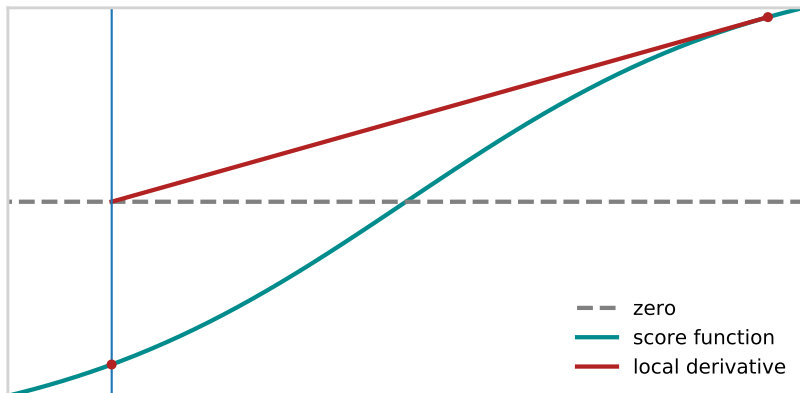
## Variance Estimates $\hat{\sigma}^2$ and $\hat{\tau}^2$

Maximize the restricted likelihood (REML) using Newton-Raphson:



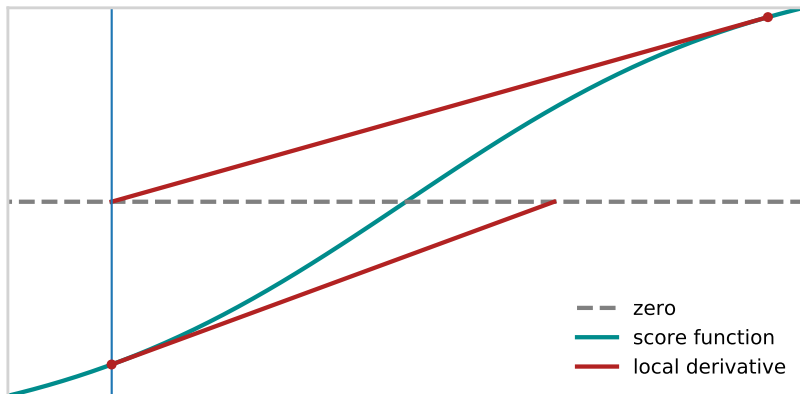
## Variance Estimates $\hat{\sigma}^2$ and $\hat{\tau}^2$

Maximize the restricted likelihood (REML) using Newton-Raphson:



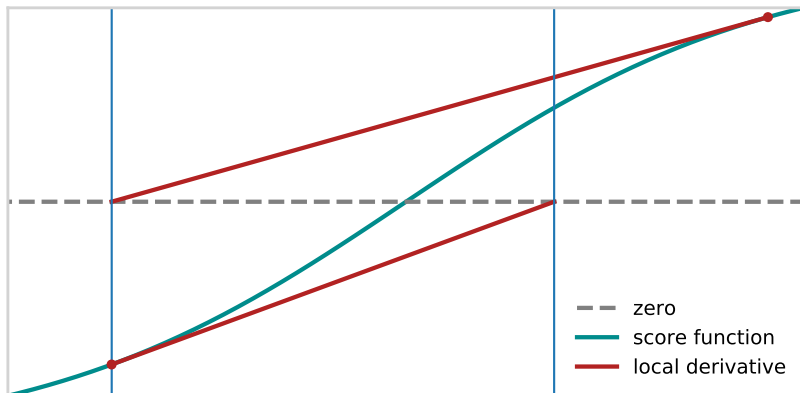
## Variance Estimates $\hat{\sigma}^2$ and $\hat{\tau}^2$

Maximize the restricted likelihood (REML) using Newton-Raphson:



## Variance Estimates $\hat{\sigma}^2$ and $\hat{\tau}^2$

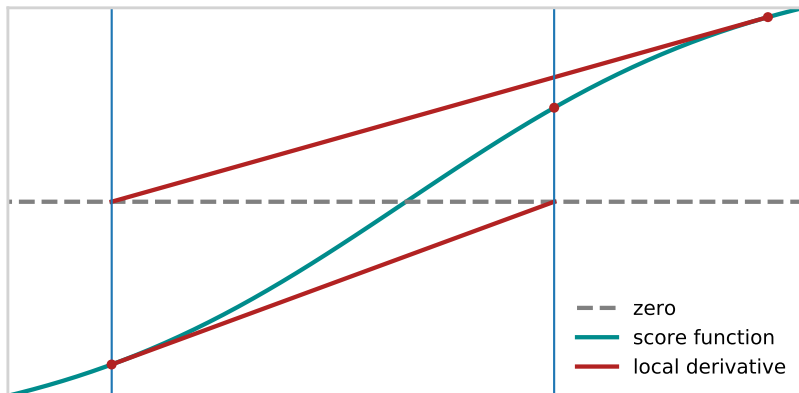
Maximize the restricted likelihood (REML) using Newton-Raphson:





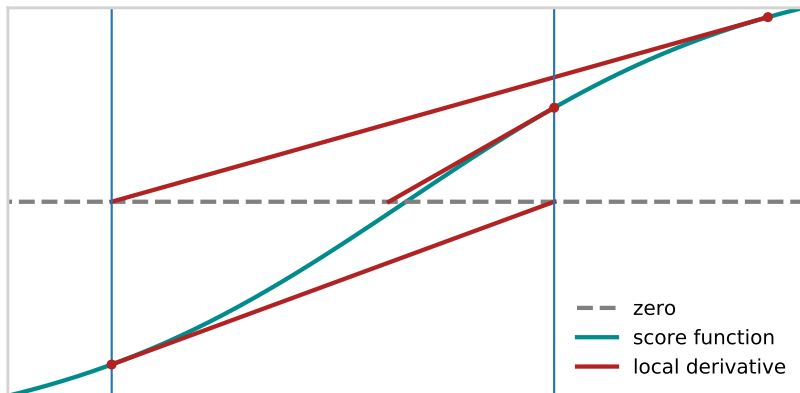
## Variance Estimates $\hat{\sigma}^2$ and $\hat{\tau}^2$

Maximize the restricted likelihood (REML) using Newton-Raphson:



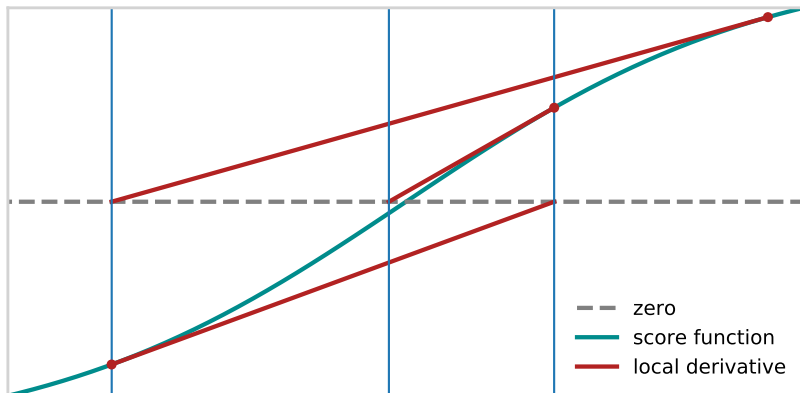
## Variance Estimates $\hat{\sigma}^2$ and $\hat{\tau}^2$

Maximize the restricted likelihood (REML) using Newton-Raphson:



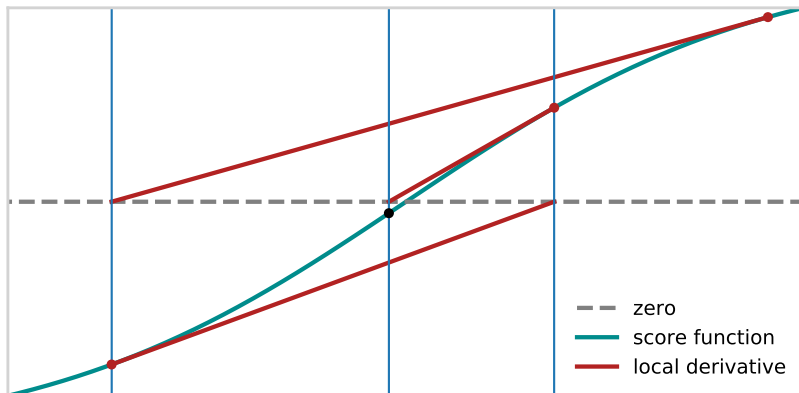
## Variance Estimates $\hat{\sigma}^2$ and $\hat{\tau}^2$

Maximize the restricted likelihood (REML) using Newton-Raphson:



## Variance Estimates $\hat{\sigma}^2$ and $\hat{\tau}^2$

Maximize the restricted likelihood (REML) using Newton-Raphson:



## (RE)ML estimation

### Single iterations (old)

- **Update  $\hat{\beta}$  and  $\hat{b}$**  given the current  $\hat{\lambda}$
- **Update  $\hat{\lambda}$**  using Fisher-Scoring (or Newton-Raphson)  
→  $\mathcal{V}_{\hat{\beta}, \hat{b}}(\lambda)$  depends on  $\hat{\beta}$  and  $\hat{b}$

⇒ Convergence is not guaranteed

# (RE)ML estimation

## Single iterations (old)

- **Update  $\hat{\beta}$  and  $\hat{b}$**  given the current  $\hat{\lambda}$
- **Update  $\hat{\lambda}$**  using Fisher-Scoring (or Newton-Raphson)  
→  $\mathcal{V}_{\hat{\beta}, \hat{b}}(\lambda)$  depends on  $\hat{\beta}$  and  $\hat{b}$

⇒ Convergence is not guaranteed

## Nested iterations (new)

- **Update  $\hat{\lambda}$**  using Newton-Raphson
  - **for each step: estimate  $\hat{\beta}_{\lambda}$  and  $\hat{b}_{\lambda}$**   
→  $\mathcal{V}(\lambda)$  depends on  $\beta$  and  $b$  only via  $\hat{\beta}_{\lambda}$  and  $\hat{b}_{\lambda}$

- **Update  $\hat{\beta}$  and  $\hat{b}$**  given the current  $\hat{\lambda}$

⇒ Convergence is guaranteed (under mild regulatory conditions)

# Comparison of Mixed Model Approach

## Fully Bayesian approach (MCMC)

- + no reparameterization needed
- Markov chain convergence is difficult to determine
- how to choose hyperpriors?

## Prediction error methods (AIC, GCV)





- + better prediction error performance
- worse resistance to overfit
- higher smoothing parameter variability
- increased tendency to multiple minima
- *more on that next week*

## Summary

- Semiparametric models can be **written as mixed models**, which enables an efficient way to estimate the penalty term  $\lambda$ .
- In order to get a proper random effects distribution, the flexible parameters have to be **separated** into sets of parameters with **flat priors** and sets with **proper priors**.
- The penalty term is proportional to the inverse of the prior variance:  $\lambda \propto \frac{1}{\tau^2}$
- For good results in mixed model inference, the **penalty term** has to be estimated in a **nested iteration** setup with the other parameters.



# References

-  Fahrmeir, L., Kneib, T., Lang, S., & Marx, B. (2013). *Regression – Models, Methods and Applications*. Springer-Verlag Berlin Heidelberg.
-  Kneib, T. (2006). *Doctoral Thesis*, LMU Munich.  
Mixed model based inference in structured additive regression.
-  Wood, S. N. (2017). *Generalized Additive Models: An Introduction with R*. Chapman and Hall/CRC.
-  Wood, S. N. (2011). *J. R. Statist. Soc. B*, 73: 3–36.  
Fast stable REML and ML estimation of semiparametric GLMs.