Representing Additive Models as Mixed Models

Katharina Ring

LMU Seminar: Mixed and Semiparametric Models

January 14, 2020

Truncated Power Basis

univariate:

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

Truncated Power Basis

univariate:

$$y = \theta_0 + \theta_1 x + \theta_2 x^2 + ... + \theta_d x^d + \sum_{k=1}^{K} \theta_{dk}(x - \kappa_k)_+^d + \epsilon$$
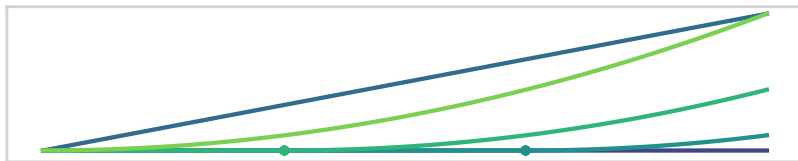
univariate and quadratic with two knots:

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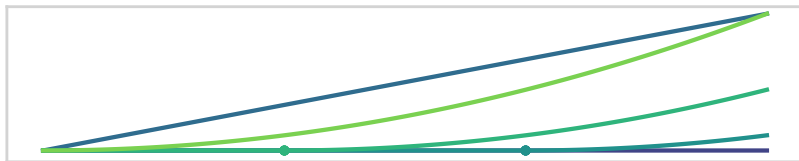
Truncated Power Basis

univariate:

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

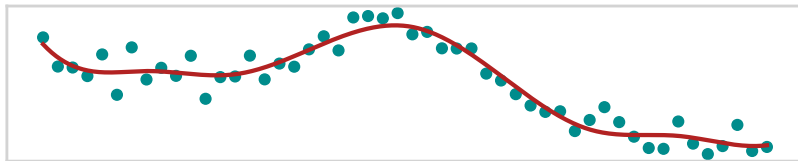univariate and quadratic with two knots:

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

Introduction
○

Recap
●○○○

Representation
○○○○○○

Inference
○○○○
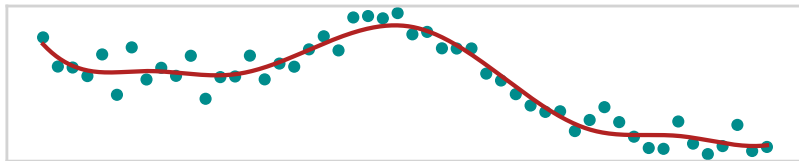
Conclusion
○○

Additive Models

Semiparametric regression:

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

Additive Models

Semiparametric regression:

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

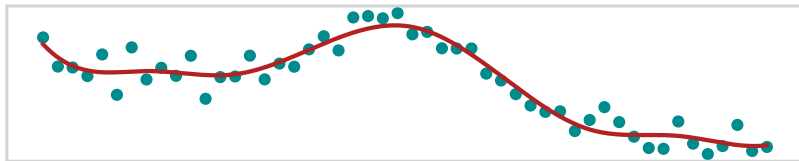# Additive Models

Semiparametric regression:

$$\hat{y}_i = \overbrace{f(\nu_i)}^{v_i^T \xi} + u_i^T \gamma$$
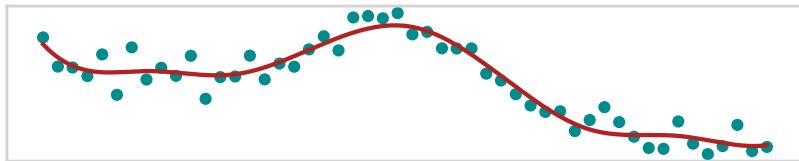
In matrix notation:

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

Additive Models

Semiparametric regression:

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

In matrix notation:

Introduction
○

Recap
●○○○

Representation
○○○○○○

Inference
○○○○

Conclusion
○○

Additive Models

Semiparametric regression:

$$\hat{y}_i = \overbrace{f_1(\nu_{i1})}^{v_{i1}^T \xi_1} + ... + \overbrace{f_p(\nu_{ip})}^{v_{ip}^T \xi_p} + u_i^T \gamma$$
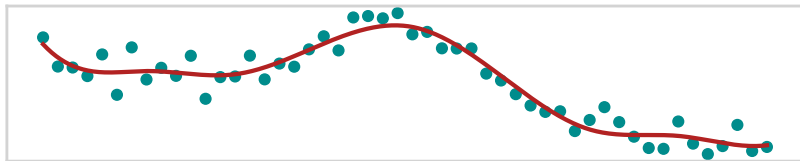
In matrix notation:

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

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.

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

Splines

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

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

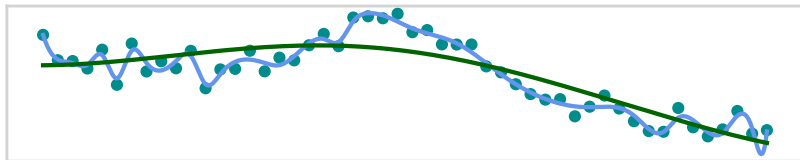with basis functions $b_1(.), ..., b_k(.)$, e. g. *B-spline*, truncated power basis, natural cubic spline, ...

Roughness Penalty

Penalized Regression Spline:

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

Control wiggliness (bias-variance tradeoff):

Introduction
○

Recap
○○○●

Representation
○○○○○○

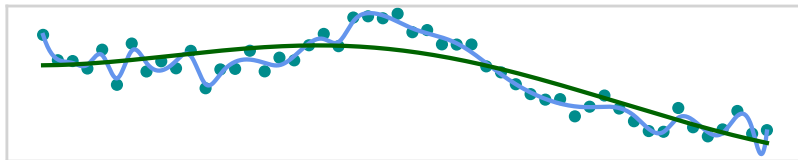Inference
○○○○

Conclusion
○○

Roughness Penalty

Penalized Regression Spline:

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

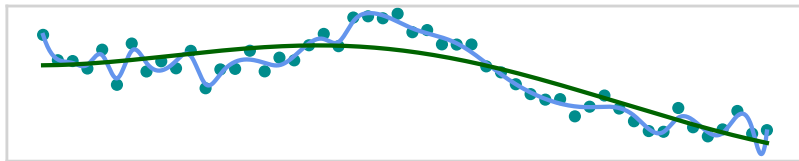e. g. first order differences $\xi^T K \xi = \sum (\xi_{k+1} - \xi_k)^2$:

Roughness Penalty

Penalized Regression Spline:

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

Problem: How to choose λ?

Mixed Models

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

Mixed Models

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

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

Mixed Models

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

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

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

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

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

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

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

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: $y = \sum_{j=1}^{p} \underbrace{V_p \xi_p}_{\text{random effects}} + \overbrace{U\gamma}^{\text{fixed effects}} + \epsilon$

Empirical Bayes

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

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

Empirical Bayes

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

Prior: $p(\gamma) \propto$ 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: $y = \sum_{j=1}^{p} \underbrace{V_p \xi_p}_{\text{random effects}} + \overbrace{U\gamma}^{\text{fixed effects}} + \epsilon$

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

Empirical Bayes

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

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

- Maximum Likelihood for τ_j^2 (so far treated as fixed):

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

Empirical Bayes

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

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

- Maximum Likelihood for τ_j^2 (so far treated as fixed):

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

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

Mixed Model Representation

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

Mixed Model Representation

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

Problem: K_j as precision matrix is problematic as K_j is often rank deficient, e. g. $\xi^T K \xi = \sum (\xi_{k+1} - \xi_k)^2 \to \xi_1$ 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: $y = \sum_{j=1}^{p} \underbrace{V_p \xi_p}_{\text{random effects}} + \overbrace{U \gamma}^{\text{fixed effects}} + \epsilon$

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 ξ_j into $\xi_j = \tilde{X}_j \beta_j + \tilde{Z}_j b_j$:

Mixed Model Representation

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

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 ξ_j into $\xi_j = \tilde{X}_j \beta_j + \tilde{Z}_j b_j$:

- β: 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: $y = \sum_{j=1}^{p} \underbrace{V_p \xi_p}_{\text{random effects}} + \overbrace{U\gamma}^{\text{fixed effects}} + \epsilon$

Problem: K_j as precision matrix is problematic as K_j is often rank deficient, e.g. $\xi^T K \xi = \sum (\xi_{k+1} - \xi_k)^2 \to \xi_1$ 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 ξ_j into $\xi_j = \tilde{X}_j \beta_j + \tilde{Z}_j b_j$:

- β: 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
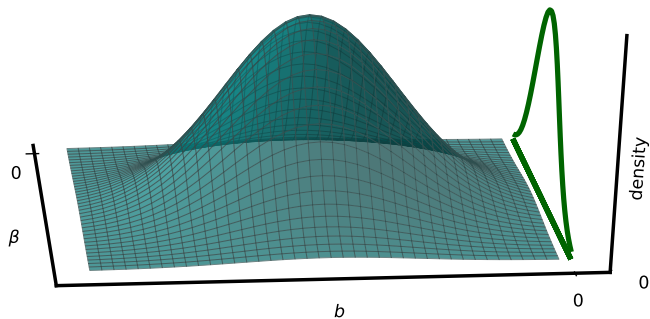 $\dim(b_j) = \text{rank}(K_j)$
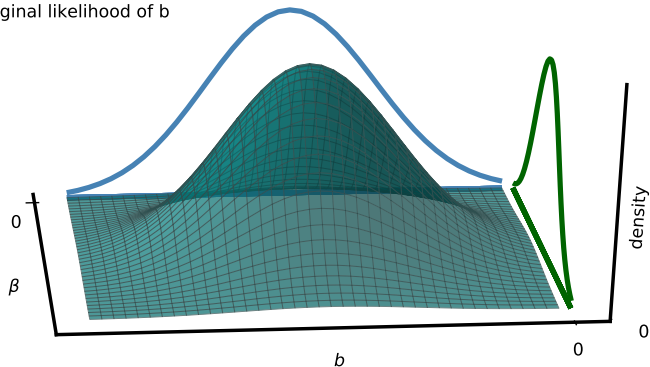
Mixed Model Representation
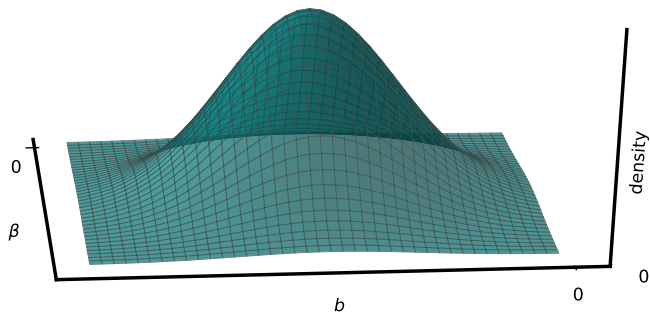
unpenalized likelihood

Mixed Model Representation

- unpenalized likelihood
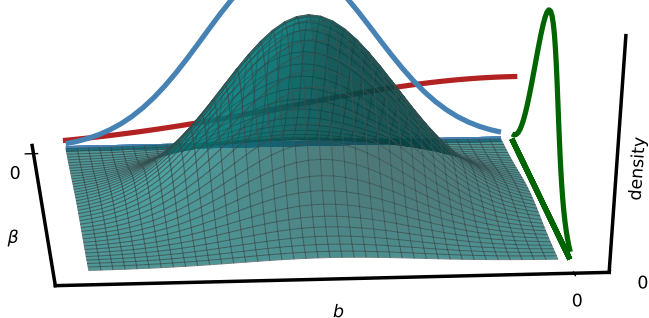- marginal likelihood of β

Mixed Model Representation

Mixed Model Representation

Introduction
o

Recap
oooo

Representation
ooo●ooo

Inference
oooo

Conclusion
oo
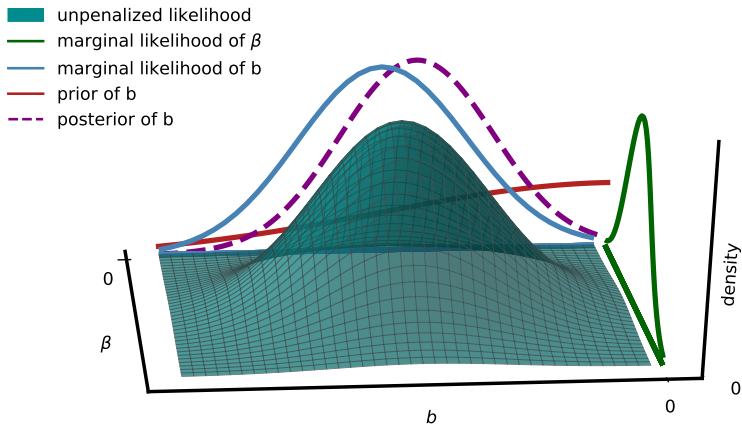
Mixed Model Representation

Mixed Model Representation

Mixed Model Representation

- ▬ unpenalized likelihood
- —— marginal likelihood of β
- —— marginal likelihood of b
- —— prior of b
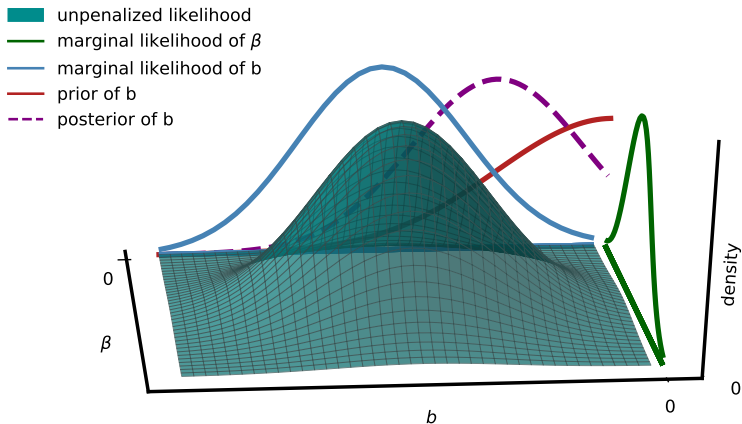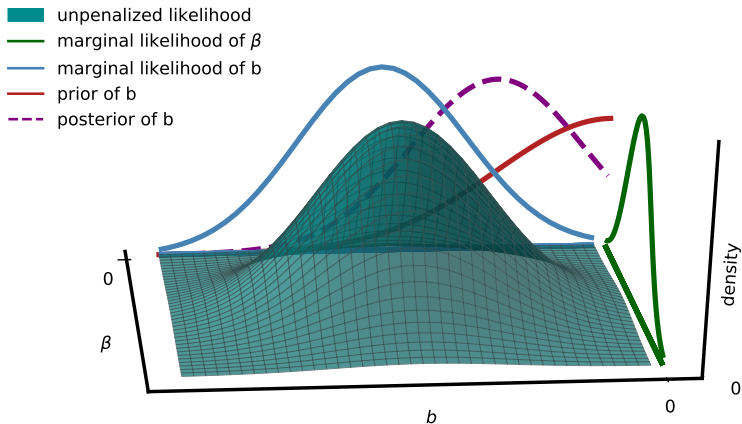- – – posterior of b

The lower the prior variance, the higher the penalty!

Mixed Model Representation

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

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

Introduction

Recap
oooo

Representation
ooo●oo

Inference
oooo

Conclusion
oo

Mixed Model Representation

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

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

Mixed Model Representation

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

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

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

$b := (b_1^T, ..., b_p^T)$

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

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

Mixed Model Representation

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

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

$$\beta := (\beta_1^T, ..., \beta_p^T, \gamma^T)$$
$$b := (b_1^T, ..., b_p^T)$$
$$Z := V_j \tilde{Z}_j$$
$$X := (V_j \tilde{X}_j, U)$$

Requirements:

Mixed Model Representation

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

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

$$\beta := (\beta_1^T, ..., \beta_p^T, \gamma^T)$$
$$b := (b_1^T, ..., b_p^T)$$
$$Z := V_j \tilde{Z}_j$$
$$X := (V_j \tilde{X}_j, U)$$

Requirements:

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

Mixed Model Representation

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

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

$$\beta := (\beta_1^T, ..., \beta_p^T, \gamma^T)$$
$$b := (b_1^T, ..., b_p^T)$$
$$Z := V_j \tilde{Z}_j$$
$$X := (V_j \tilde{X}_j, U)$$

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: $\tilde{X}_j^T \tilde{Z}_j = 0$

Mixed Model Representation

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

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

$$\beta := (\beta_1^T, ..., \beta_p^T, \gamma^T)$$
$$b := (b_1^T, ..., b_p^T)$$
$$Z := V_j \tilde{Z}_j$$
$$X := (V_j \tilde{X}_j, U)$$

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. β_j not penalized by K_j: $\tilde{X}_j^T K_j \tilde{X}_j = 0$

Mixed Model Representation

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

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

$$\beta := (\beta_1^T, ..., \beta_p^T, \gamma^T)$$
$$b := (b_1^T, ..., b_p^T)$$
$$Z := V_j \tilde{Z}_j$$
$$X := (V_j \tilde{X}_j, U)$$

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. β_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. β_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: $K_j = \Gamma_j \Lambda_j \Gamma^T$, 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:

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

$\Rightarrow p(\beta) \propto$ const.

Mixed Model Representation

log-Prior:

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

$\Rightarrow p(\beta) \propto$ const.
$\Rightarrow p(b_j) \sim N(0, \tau_j^2 I_{kj})$

Mixed Model Representation

log-Prior:

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

$\Rightarrow p(\beta) \propto$ const.
$\Rightarrow p(b_j) \sim N(0, \tau_j^2 I_{kj})$

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 β and b simultaneously based on known σ^2 and τ^2.

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

In order to maximize the (log-)Posterior (equivalent to ML), derive estimates for β and b simultaneously based on known σ^2 and τ^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 $Q = \text{blockdiag}(\tau_1^2 I_{k1}, ..., \tau_p^2 I_{kp})$

Variance Estimates

Maximum Likelihood (integrate b out)

- uses (partially) marginal distribution $y \sim \mathrm{N}(X\beta, \Sigma)$

    1. Derive $\hat{\beta}$ analytically
    2. Plug in to get profile likelihood for τ^2 and σ^2
    3. Maximize numerically

- estimates variance components of posterior mode

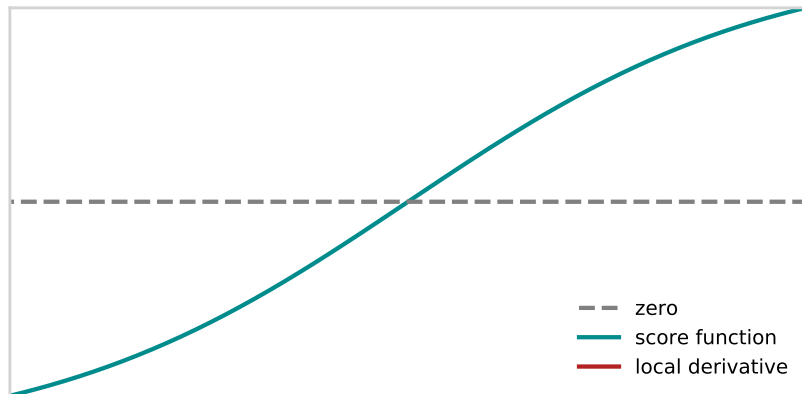# Variance Estimates

Maximum Likelihood (integrate b out)

- uses (partially) marginal distribution $y \sim \mathrm{N}(X\beta, \Sigma)$

  1. Derive $\hat{\beta}$ analytically
  2. Plug in to get profile likelihood for τ^2 and σ^2
  3. Maximize numerically

- estimates variance components of posterior mode

Restricted ML (integrate b and β out)

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

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

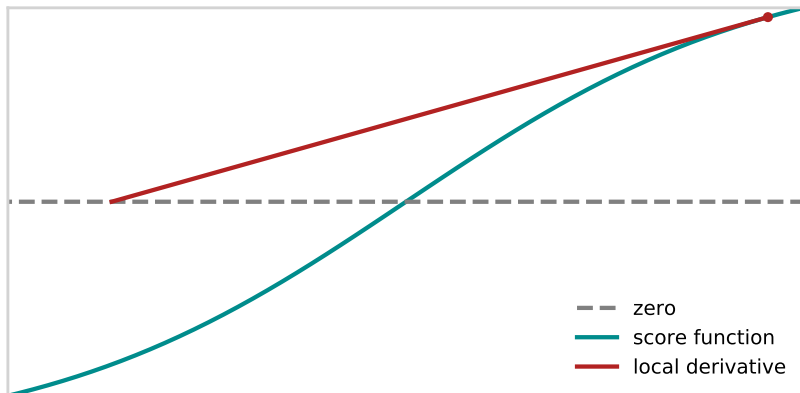Maximize the restricted likelihood (REML) using Newton-Raphson:

- - - zero
— score function
— local derivative

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

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

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

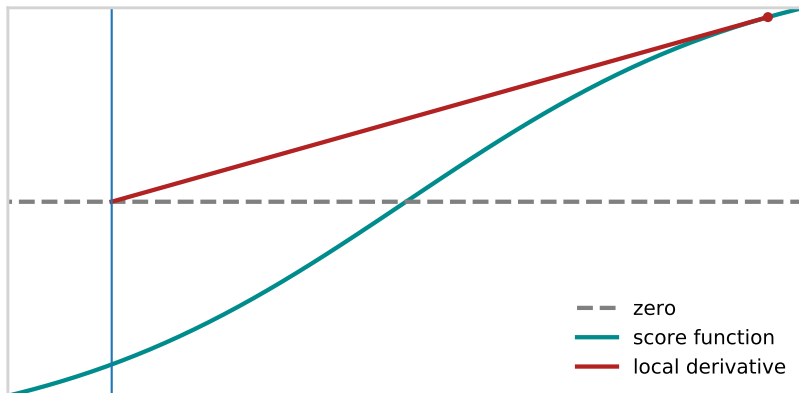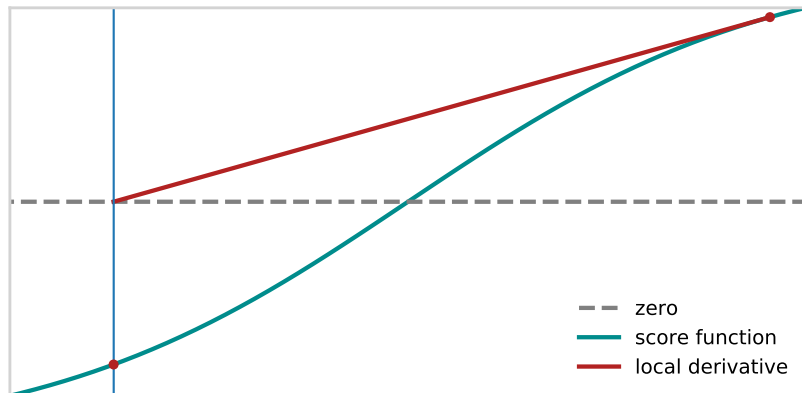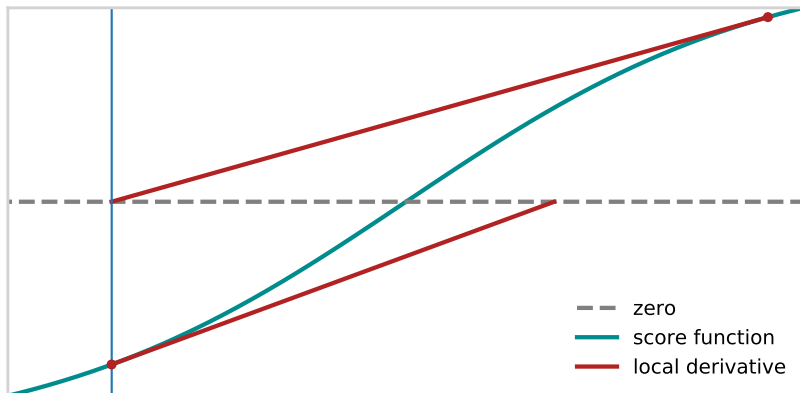Maximize the restricted likelihood (REML) using Newton-Raphson:

- - - zero
—— score function
—— local derivative

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

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

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

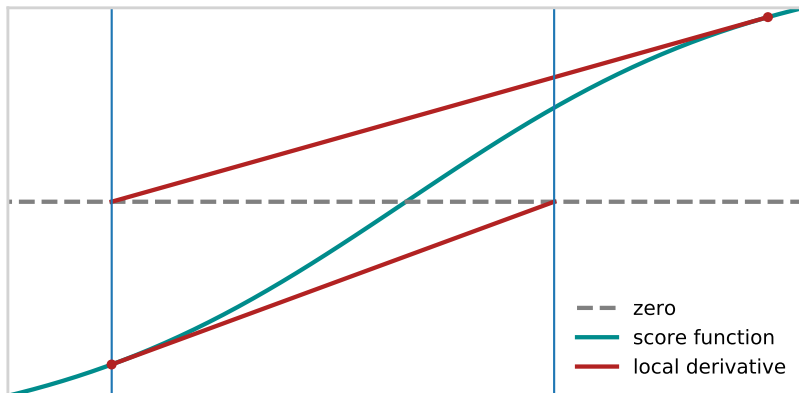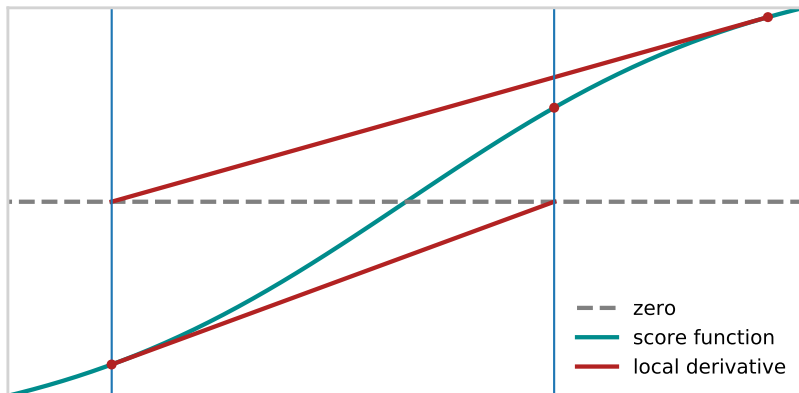Maximize the restricted likelihood (REML) using Newton-Raphson:
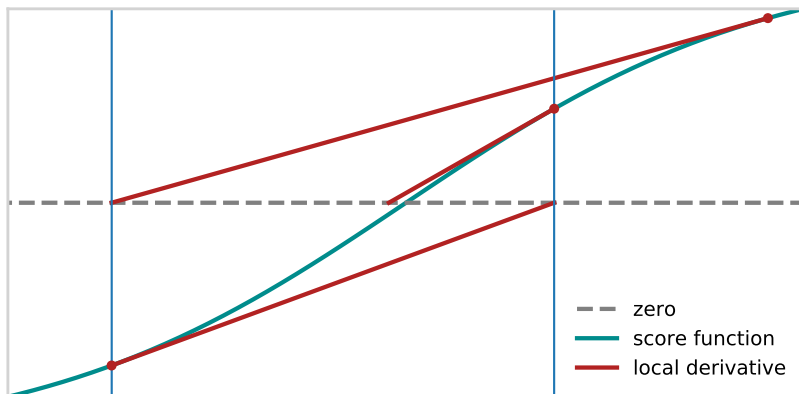

- - - zero
— score function
— local derivative

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

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

- - - zero
— score function
— local derivative

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

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

- - - zero
— score function
— local derivative

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

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

Legend:
- - - zero
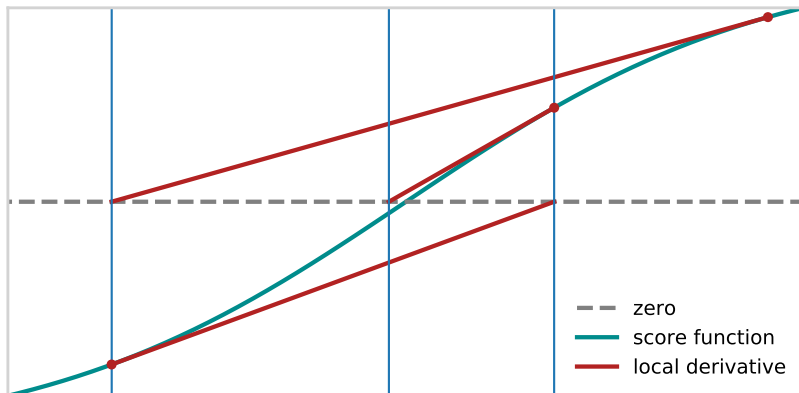- score function
- local derivative

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

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

- - - zero
— score function
— local derivative

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

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

- - - zero
— score function
— local derivative

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

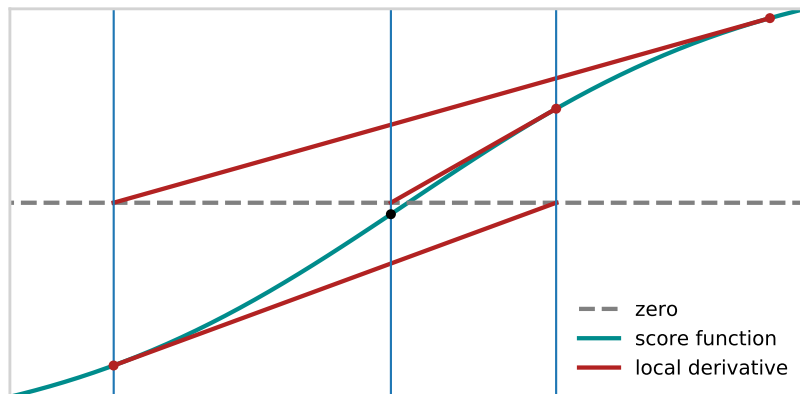Maximize the restricted likelihood (REML) using Newton-Raphson:

legend:
- – – zero
- —— score function
- —— local derivative

(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)

  $\rightarrow \mathcal{V}_{\hat{\beta},\hat{b}}(\lambda)$ depends on $\hat{\beta}$ and $\hat{b}$

$\Rightarrow$ Convergence is not guaranteed

(RE)ML estimation

Single iterations (old)

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

$\Rightarrow$ Convergence is not guaranteed

Nested iterations (new)

- **Update $\hat{\boldsymbol{\lambda}}$** using Newton-Raphson

  - **Estimate $\hat{\boldsymbol{\beta}}_{\boldsymbol{\lambda}}$ in an inner loop**

  $\rightarrow \mathcal{V}(\lambda)$ depends on β only via $\hat{\beta}_\lambda$

- **Update $\hat{\boldsymbol{\beta}}$ and $\hat{\boldsymbol{b}}$** given the current $\hat{\lambda}$
  Solve with Newton-Raphson

$\Rightarrow$ Convergence is guaranteed (under mild regulatory conditions)

Comparison of Mixed Model Approach

Fully Bayesian approach
(MCMC)

+ no reparameterization
  needed

− how to choose hyperpriors?

− Markov chain convergence is
  difficult to determine

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**.
- 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.