Gaussian Process Regression and Emulation

STAT8810, Fall 2017

M.T. Pratola

August 29, 2017

Today

More on GP Regression

Emulating Outputs from a Simulator

- Best Linear Unbiased Predictions (these actually don't require the Normality assumption assuming the statistical model's parameters are known)

Emulating Outputs from a Simulator

- Best Linear Unbiased Predictions (these actually don't require the Normality assumption assuming the statistical model's parameters are known)
- Frequentist prediction

Emulating Outputs from a Simulator

- Best Linear Unbiased Predictions (these actually don't require the Normality assumption assuming the statistical model's parameters are known)
- Frequentist prediction
- Bayesian prediction (we'll return to this later)

Frequentist Prediction

We have our data $\mathbf{y} \sim f$ where

$$E[y(\mathbf{x})] = \mu(\mathbf{x})$$

and $\mu(\mathbf{x}) = \mu$ or $\mu(\mathbf{x}) = \mathbf{f}^T(\mathbf{x})\beta$ are common choices, and

$$Cov(y(\mathbf{x}), y(\mathbf{x}')) = c(\mathbf{x} - \mathbf{x}').$$

- A popular class of predictors are linear in the data:

Frequentist Prediction

We have our data $\mathbf{y} \sim f$ where

$$E[y(\mathbf{x})] = \mu(\mathbf{x})$$

and $\mu(\mathbf{x}) = \mu$ or $\mu(\mathbf{x}) = \mathbf{f}^T(\mathbf{x})\beta$ are common choices, and

$$Cov(y(\mathbf{x}), y(\mathbf{x}')) = c(\mathbf{x} - \mathbf{x}').$$

- A popular class of predictors are linear in the data:
  - linear predictor: $\hat{y}(\mathbf{x}) = \mathbf{c}^T\mathbf{y}$

Frequentist Prediction

We have our data $\mathbf{y} \sim f$ where

$$E[y(\mathbf{x})] = \mu(\mathbf{x})$$

and $\mu(\mathbf{x}) = \mu$ or $\mu(\mathbf{x}) = \mathbf{f}^T(\mathbf{x})\beta$ are common choices, and

$$Cov(y(\mathbf{x}), y(\mathbf{x}')) = c(\mathbf{x} - \mathbf{x}').$$

- A popular class of predictors are linear in the data:
  - linear predictor: $\hat{y}(\mathbf{x}) = \mathbf{c}^T\mathbf{y}$
  - unbiased linear predictor: $\hat{y}(\mathbf{x}) = \mathbf{c}^T\mathbf{y}$ s.t. $E[\hat{y}(\mathbf{x})] = \mu(\mathbf{x})$

Frequentist Prediction

We have our data $\mathbf{y} \sim f$ where

$$E[y(\mathbf{x})] = \mu(\mathbf{x})$$

and $\mu(\mathbf{x}) = \mu$ or $\mu(\mathbf{x}) = \mathbf{f}^T(\mathbf{x})\beta$ are common choices, and

$$Cov(y(\mathbf{x}), y(\mathbf{x}')) = c(\mathbf{x} - \mathbf{x}').$$

- A popular class of predictors are linear in the data:
  - linear predictor: $\hat{y}(\mathbf{x}) = \mathbf{c}^T\mathbf{y}$
  - unbiased linear predictor: $\hat{y}(\mathbf{x}) = \mathbf{c}^T\mathbf{y}$ s.t. $E[\hat{y}(\mathbf{x})] = \mu(\mathbf{x})$
  - Best MSPE predictor: $\min_{\mathbf{c}(\mathbf{x})} \text{MSPE}(\hat{y}(\mathbf{x}) - y(\mathbf{x}))$ where $\text{MSPE} = E\left[(\hat{y}(\mathbf{x}) - y(\mathbf{x}))^2\right]$

Frequentist Prediction

We have our data $\mathbf{y} \sim f$ where

$$E[y(\mathbf{x})] = \mu(\mathbf{x})$$

and $\mu(\mathbf{x}) = \mu$ or $\mu(\mathbf{x}) = \mathbf{f}^T(\mathbf{x})\beta$ are common choices, and

$$Cov(y(\mathbf{x}), y(\mathbf{x}')) = c(\mathbf{x} - \mathbf{x}').$$

- A popular class of predictors are linear in the data:
  - linear predictor: $\hat{y}(\mathbf{x}) = \mathbf{c}^T\mathbf{y}$
  - unbiased linear predictor: $\hat{y}(\mathbf{x}) = \mathbf{c}^T\mathbf{y}$ s.t. $E[\hat{y}(\mathbf{x})] = \mu(\mathbf{x})$
  - Best MSPE predictor: $\min_{\mathbf{c}(\mathbf{x})}$ MSPE$(\hat{y}(\mathbf{x}) - y(\mathbf{x}))$ where MSPE $= E\left[(\hat{y}(\mathbf{x}) - y(\mathbf{x}))^2\right]$
  - Best Linear Unbiased Predictor (BLUP): $\hat{y}(\mathbf{x}) = \mathbf{c}^T\mathbf{y}$ s.t. $E[\hat{y}(\mathbf{x})] = \mu(\mathbf{x})$ and $\min_{\mathbf{c}(\mathbf{x})}$ MSPE$(\hat{y}(\mathbf{x}) - y(\mathbf{x}))$.

Frequentist Prediction

- Suppose $(y(\mathbf{x}_0), \mathbf{y})) \sim f$ whose conditional mean $E[y(\mathbf{x}_0)|\mathbf{y}] := \hat{y}(\mathbf{x}_0)$ exists. Then $\hat{y}(\mathbf{x}_0)$ is the best MSPE predictor.

Proof: Let $\tilde{y}(\mathbf{x}_0)$ be another predictor of $y(\mathbf{x}_0)$.

$$\begin{aligned}
\text{MSPE}(\tilde{y}(\mathbf{x}_0)) &= E[(\tilde{y}(\mathbf{x}_0) - y(\mathbf{x}_0))^2|\mathbf{y}] \\
&= E[(\tilde{y}(\mathbf{x}_0) - \hat{y}(\mathbf{x}_0) + \hat{y}(\mathbf{x}_0) - y(\mathbf{x}_0))^2|\mathbf{y}] \\
&= E[(\tilde{y}(\mathbf{x}_0) - \hat{y}(\mathbf{x}_0))^2|\mathbf{y}] + \text{MSPE}(\hat{y}(\mathbf{x}_0)) \\
&\quad +2E\left[(\tilde{y}(\mathbf{x}_0) - \hat{y}(\mathbf{x}_0))(\hat{y}(\mathbf{x}_0) - y(\mathbf{x}_0))|\, \mathbf{y}\right]
\end{aligned}$$

But $E\left[(\tilde{y}(\mathbf{x}_0) - \hat{y}(\mathbf{x}_0))(\hat{y}(\mathbf{x}_0) - y(\mathbf{x}_0))|\, \mathbf{y}\right] = 0$ since $E\left[\hat{y}(\mathbf{x}_0) - y(\mathbf{x}_0)|\mathbf{y}\right] = E[y(\mathbf{x}_0)|\mathbf{y}] - E[y(\mathbf{x}_0)|\mathbf{y}] = 0$

Therefore, $\text{MSPE}(\tilde{y}(\mathbf{x}_0)) = E\left[(\tilde{y}(\mathbf{x}_0) - \hat{y}(\mathbf{x}_0))^2\right] + \text{MSPE}(\hat{y}(\mathbf{x}_0)) \geq \text{MSPE}(\hat{y}(\mathbf{x}_0))$.

Frequentist Prediction

- Consider $\begin{pmatrix} y_0 \\ \mathbf{y} \end{pmatrix} \sim \left[\begin{pmatrix} \mathbf{f}_0^T \\ \mathbf{F} \end{pmatrix} \beta, \sigma^2 \begin{pmatrix} 1 & \mathbf{r}_0^T \\ \mathbf{r}_0 & \mathbf{R} \end{pmatrix} \right], \mathbf{x}_i \in \mathbb{R}^d, i = 0, 1, \ldots, n.$

where $\mathbf{f}_0 = (f_1(\mathbf{x}_0), \ldots, f_p(\mathbf{x}_0))^T$,

$\mathbf{F} = [\mathbf{f}_j(\mathbf{x}_0)], \quad 1 \le i \le n, 1 \le j \le p,$

$\beta = (\beta_1, \ldots, \beta_p)^T,$

$\mathbf{r}_0 = (R(\mathbf{x}_0 - \mathbf{x}_1), \ldots, R(\mathbf{x}_0 - \mathbf{x}_n))^T$,

$\mathbf{R} = [R(\mathbf{x}_i - \mathbf{x}_j)], \quad 1 \le i, j \le n.$

$$\hat{y}(\mathbf{x}_0) = \mathbf{f}_0^T \widehat{\beta} + \mathbf{r}_0^T \mathbf{R}^{-1} \left(\mathbf{y} - \mathbf{F} \widehat{\beta} \right)$$

where $\widehat{\beta} = \left(\mathbf{F}^T \mathbf{R}^{-1} \mathbf{F} \right)^{-1} \mathbf{F}^T \mathbf{R}^{-1} \mathbf{y}.$

Frequentist Prediction

- Consider $\begin{pmatrix} y_0 \\ \mathbf{y} \end{pmatrix} \sim \left[\begin{pmatrix} \mathbf{f}_0^T \\ \mathbf{F} \end{pmatrix} \boldsymbol{\beta}, \sigma^2 \begin{pmatrix} 1 & \mathbf{r}_0^T \\ \mathbf{r}_0 & \mathbf{R} \end{pmatrix} \right], \mathbf{x}_i \in \mathbb{R}^d, i = 0, 1, \ldots, n.$

where $\mathbf{f}_0 = (f_1(\mathbf{x}_0), \ldots, f_p(\mathbf{x}_0))^T$,

$\mathbf{F} = [\mathbf{f}_j(\mathbf{x}_0)], \quad 1 \le i \le n, 1 \le j \le p,$

$\boldsymbol{\beta} = (\beta_1, \ldots, \beta_p)^T,$

$\mathbf{r}_0 = (R(\mathbf{x}_0 - \mathbf{x}_1), \ldots, R(\mathbf{x}_0 - \mathbf{x}_n))^T$,

$\mathbf{R} = [R(\mathbf{x}_i - \mathbf{x}_j)], \quad 1 \le i, j \le n.$

- The BLUP of $y_0 = y(\mathbf{x}_0)$ is

$$\hat{y}(\mathbf{x}_0) = \mathbf{f}_0^T \widehat{\boldsymbol{\beta}} + \mathbf{r}_0^T \mathbf{R}^{-1} \left(\mathbf{y} - \mathbf{F} \widehat{\boldsymbol{\beta}} \right)$$

where $\widehat{\boldsymbol{\beta}} = \left(\mathbf{F}^T \mathbf{R}^{-1} \mathbf{F} \right)^{-1} \mathbf{F}^T \mathbf{R}^{-1} \mathbf{y}.$

BLUP Properties

- $\hat{y}(\mathbf{x}_0)$ is linear in $y_i, i = 1, \quad \ldots, n$, so given the weights it is easy to calculate.

BLUP Properties

- $\hat{y}(\mathbf{x}_0)$ is linear in $y_i, i = 1, \quad \ldots, n$, so given the weights it is easy to calculate.
- MSPE of the BLUP is

$$\sigma^2 \left(1 - \mathbf{r}_0^T \mathbf{R}^{-1} \mathbf{r}_0 + (\mathbf{f}_0 - \mathbf{F}^T \mathbf{R}^{-1} \mathbf{r}_0)^T \mathbf{F}^T \mathbf{R}^{-1} \mathbf{F} (\mathbf{f}_0 - \mathbf{F}^T \mathbf{R}^{-1} \mathbf{r}_0)\right)$$

$$= \sigma^2 (1 - \text{variance improvement term}$$

$$+ \text{ penalty since we don't know } \beta)$$

BLUP Properties

- Write $\mathbf{d} = \mathbf{R}^{-1}(\mathbf{y} - \mathbf{F}\widehat{\beta})$, then

$$\hat{y}(\mathbf{x}_0) = \sum_j f_j(\mathbf{x}_0)\hat{\beta}_j + \sum_{i=1}^{n} d_i R(\mathbf{x}_0 - \mathbf{x}_i).$$

BLUP Properties

- Write $\mathbf{d} = \mathbf{R}^{-1}(\mathbf{y} - \mathbf{F}\widehat{\beta})$, then

$$\hat{y}(\mathbf{x}_0) = \sum_j f_j(\mathbf{x}_0)\hat{\beta}_j + \sum_{i=1}^{n} d_i R(\mathbf{x}_0 - \mathbf{x}_i).$$

- When $\mathbf{x}_0 - \mathbf{x}_i$ is large, $R(\cdot)$ is small so less weight is given in forming the prediction

BLUP Properties

- Write $\mathbf{d} = \mathbf{R}^{-1}(\mathbf{y} - \mathbf{F}\widehat{\beta})$, then

$$\hat{y}(\mathbf{x}_0) = \sum_j f_j(\mathbf{x}_0)\hat{\beta}_j + \sum_{i=1}^{n} d_i R(\mathbf{x}_0 - \mathbf{x}_i).$$

- When $\mathbf{x}_0 - \mathbf{x}_i$ is large, $R(\cdot)$ is small so less weight is given in forming the prediction
- When $\mathbf{x}_0 - \mathbf{x}_i$ is small, $R(\cdot)$ is large so more weight is given in forming the prediction

BLUP Properties

- Write $\mathbf{d} = \mathbf{R}^{-1}(\mathbf{y} - \mathbf{F}\widehat{\beta})$, then

$$\hat{y}(\mathbf{x}_0) = \sum_j f_j(\mathbf{x}_0)\hat{\beta}_j + \sum_{i=1}^{n} d_i R(\mathbf{x}_0 - \mathbf{x}_i).$$

- When $\mathbf{x}_0 - \mathbf{x}_i$ is large, $R(\cdot)$ is small so less weight is given in forming the prediction
- When $\mathbf{x}_0 - \mathbf{x}_i$ is small, $R(\cdot)$ is large so more weight is given in forming the prediction
- If $R(h) = exp(-\theta \sum_{l=1}^{d} h_l^2)$ (isotropic model) then

$$\hat{y}(\mathbf{x}_0) = \sum_j f_j(\mathbf{x}_0)\hat{\beta}_j + \sum_{i=1}^{n} d_i exp(-\theta \sum_{l=1}^{d}(x_{0l} - x_{il})^2)$$

is the so-called "Radial Basis Function" model, popular in machine learning and elsewhere for awhile.

BLUP Properties

- $\hat{y}(\cdot)$ interpolates the observations $y_1, \ldots, y_n$

BLUP Properties

- $\hat{y}(\cdot)$ interpolates the observations $y_1, \ldots, y_n$
- MSPE at the observed data $y_1, \ldots, y_n$ is zero

Empirical BLUP

- The BLUP uses the GLS estimator for $\boldsymbol{\beta}$. But in actuality we also have parameters σ^2 and $\boldsymbol{\theta}$ (correlation function parameters) that so far have been assumed known.

Empirical BLUP

- The BLUP uses the GLS estimator for $\boldsymbol{\beta}$. But in actuality we also have parameters σ^2 and $\boldsymbol{\theta}$ (correlation function parameters) that so far have been assumed known.
- Empirical BLUP (EBLUP): use plug-in estimators for $\boldsymbol{\theta}$ and $\sigma^2(\boldsymbol{\theta})$.

Empirical BLUP

- The BLUP uses the GLS estimator for β. But in actuality we also have parameters σ^2 and $\boldsymbol{\theta}$ (correlation function parameters) that so far have been assumed known.
- Empirical BLUP (EBLUP): use plug-in estimators for $\boldsymbol{\theta}$ and $\sigma^2(\boldsymbol{\theta})$.
- The most common approach uses the MLE.

Empirical BLUP (MLE-based)

- For the GP model, the log-likelihood of the data $\mathbf{y}$ is

$$l = -\frac{n}{2} log(\sigma^2) - \frac{1}{2} log(|\mathbf{R}|) - \frac{1}{2\sigma^2}(\mathbf{y} - \mathbf{F}\beta)^T \mathbf{R}^{-1}(\mathbf{y} - \mathbf{F}\beta).$$

Empirical BLUP (MLE-based)

- For the GP model, the log-likelihood of the data $\mathbf{y}$ is

$$l = -\frac{n}{2} log(\sigma^2) - \frac{1}{2} log(|\mathbf{R}|) - \frac{1}{2\sigma^2}(\mathbf{y} - \mathbf{F}\beta)^T \mathbf{R}^{-1}(\mathbf{y} - \mathbf{F}\beta).$$

- Taking $\frac{\partial l}{\partial \beta}$ and setting equal to 0 we get what we had before:

$$\widehat{\boldsymbol{\beta}}(\boldsymbol{\theta}) = \left(\mathbf{F}^T \mathbf{R}^{-1} \mathbf{F}\right)^{-1} \mathbf{F}^T \mathbf{R}^{-1} \mathbf{y}$$

since R depends on $\boldsymbol{\theta}$.

Empirical BLUP (MLE-based)

- $\Rightarrow l(\widehat{\beta}(\boldsymbol{\theta}), \sigma^2, \boldsymbol{\theta}) = -\frac{n}{2} log(\sigma^2) - \frac{1}{2} log(|\mathbf{R}|) - \frac{1}{2\sigma^2}(\mathbf{y} - \mathbf{F}\widehat{\beta})^T \mathbf{R}^{-1}(\mathbf{y} - \mathbf{F}\widehat{\beta}).$

Empirical BLUP (MLE-based)

- $\Rightarrow l(\widehat{\beta}(\boldsymbol{\theta}), \sigma^2, \boldsymbol{\theta}) =$
  $-\frac{n}{2} log(\sigma^2) - \frac{1}{2} log(|\mathbf{R}|) - \frac{1}{2\sigma^2}(\mathbf{y} - \mathbf{F}\widehat{\beta})^T \mathbf{R}^{-1}(\mathbf{y} - \mathbf{F}\widehat{\beta}).$
- Taking partial wrt σ^2 and setting equal to zero, we get

$$\hat{\sigma}^2(\boldsymbol{\theta}) = \frac{1}{n}(\mathbf{y} - \mathbf{F}\widehat{\beta})^T \mathbf{R}^{-1}(\mathbf{y} - \mathbf{F}\widehat{\beta})$$

  since $\mathbf{R}$ and $\widehat{\beta}$ depend on $\boldsymbol{\theta}$.

Empirical BLUP (MLE-based)

- $\Rightarrow l(\widehat{\beta}(\boldsymbol{\theta}), \hat{\sigma}^2(\boldsymbol{\theta}), \boldsymbol{\theta}) = -\frac{n}{2} log(\hat{\sigma}^2) - \frac{1}{2} log(|\mathbf{R}|) - \frac{n}{2}.$

Empirical BLUP (MLE-based)

- $\Rightarrow l(\widehat{\beta}(\boldsymbol{\theta}), \hat{\sigma}^2(\boldsymbol{\theta}), \boldsymbol{\theta}) = -\frac{n}{2} log(\hat{\sigma}^2) - \frac{1}{2} log(|\mathbf{R}|) - \frac{n}{2}$.
- No closed form solution for this, so find $\widehat{\boldsymbol{\theta}}$ by taking the arg max:

$$\widehat{\boldsymbol{\theta}} = \arg\max_{\boldsymbol{\theta}} -\frac{n}{2} log(\hat{\sigma}^2(\boldsymbol{\theta})) - \frac{1}{2} log(|\mathbf{R}(\boldsymbol{\theta})|)$$

Empirical BLUP (MLE-based)

- Say $\widehat{\mathbf{R}} = \mathbf{R}(\cdot|\widehat{\boldsymbol{\theta}})$, $\widehat{\mathbf{r}} = \mathbf{r}(\cdot|\widehat{\boldsymbol{\theta}})$ and $\widehat{\beta} = \widehat{\beta}(\widehat{\boldsymbol{\theta}})$.

Empirical BLUP (MLE-based)

- Say $\widehat{\mathbf{R}} = \mathbf{R}(\cdot|\widehat{\boldsymbol{\theta}})$, $\widehat{\mathbf{r}} = \mathbf{r}(\cdot|\widehat{\boldsymbol{\theta}})$ and $\widehat{\beta} = \widehat{\beta}(\widehat{\boldsymbol{\theta}})$.
- Then our EBLUP is

$$\hat{y}(\mathbf{x}_0) = \mathbf{f}_0^T \widehat{\beta} + \widehat{\mathbf{r}}_0^T \widehat{\mathbf{R}}^{-1}(\mathbf{y} - \mathbf{F}\widehat{\beta})$$

and our Empirical MSPE (EMSPE) of the EBLUP is

$$\begin{aligned}\hat{s}^2(\mathbf{x}_0) &= \hat{\sigma}^2(1 - \widehat{\mathbf{r}}_0^T \widehat{\mathbf{R}}^{-1} \widehat{\mathbf{r}}_0 \\ &\quad + (\mathbf{f}_0 - \mathbf{F}^T \widehat{\mathbf{R}}^{-1} \widehat{\mathbf{r}}_0)^T \mathbf{F}^T \widehat{\mathbf{R}}^{-1} \mathbf{F} (\mathbf{f}_0 - \mathbf{F}^T \widehat{\mathbf{R}}^{-1} \widehat{\mathbf{r}}_0)).\end{aligned}$$

EBLUP Properties

- Still interpolates the data

† Li and Sudjianto: Analysis of computer experiments using penalized likelihood in Gaussian Kriging models (2005).

EBLUP Properties

- Still interpolates the data
- $l(\widehat{\beta}(\boldsymbol{\theta}), \hat{\sigma}^2(\boldsymbol{\theta}), \boldsymbol{\theta})$ can be very challenging to numerically maximize.

† Li and Sudjianto: Analysis of computer experiments using penalized likelihood in Gaussian Kriging models (2005).

EBLUP Properties

- Still interpolates the data
- $l(\widehat{\beta}(\boldsymbol{\theta}), \hat{\sigma}^2(\boldsymbol{\theta}), \boldsymbol{\theta})$ can be very challenging to numerically maximize.
  - straight Newton-Raphson typically unsuccessful

† Li and Sudjianto: Analysis of computer experiments using penalized likelihood in Gaussian Kriging models (2005).

EBLUP Properties

- Still interpolates the data
- $l(\widehat{\beta}(\boldsymbol{\theta}), \hat{\sigma}^2(\boldsymbol{\theta}), \boldsymbol{\theta})$ can be very challenging to numerically maximize.
  - straight Newton-Raphson typically unsuccessful
  - A combined strategy such as Nelder-Mead with Newton-Raphson works reasonably well

† Li and Sudjianto: Analysis of computer experiments using penalized likelihood in Gaussian Kriging models (2005).

EBLUP Properties

- Still interpolates the data
- $l(\widehat{\beta}(\boldsymbol{\theta}), \hat{\sigma}^2(\boldsymbol{\theta}), \boldsymbol{\theta})$ can be very challenging to numerically maximize.
  - straight Newton-Raphson typically unsuccessful
  - A combined strategy such as Nelder-Mead with Newton-Raphson works reasonably well
- Other estimators of $\hat{\sigma}^2$, $\widehat{\boldsymbol{\theta}}$ are possible:

† Li and Sudjianto: Analysis of computer experiments using penalized likelihood in Gaussian Kriging models (2005).

EBLUP Properties

- Still interpolates the data
- $l(\widehat{\beta}(\boldsymbol{\theta}), \hat{\sigma}^2(\boldsymbol{\theta}), \boldsymbol{\theta})$ can be very challenging to numerically maximize.
  - straight Newton-Raphson typically unsuccessful
  - A combined strategy such as Nelder-Mead with Newton-Raphson works reasonably well
- Other estimators of $\hat{\sigma}^2$, $\widehat{\boldsymbol{\theta}}$ are possible:
  - REML (Restricted Maximum Likelihood)

† Li and Sudjianto: Analysis of computer experiments using penalized likelihood in Gaussian Kriging models (2005).

EBLUP Properties

- Still interpolates the data
- $l(\widehat{\beta}(\boldsymbol{\theta}), \hat{\sigma}^2(\boldsymbol{\theta}), \boldsymbol{\theta})$ can be very challenging to numerically maximize.
    - straight Newton-Raphson typically unsuccessful
    - A combined strategy such as Nelder-Mead with Newton-Raphson works reasonably well
- Other estimators of $\hat{\sigma}^2$, $\widehat{\boldsymbol{\theta}}$ are possible:
    - REML (Restricted Maximum Likelihood)
    - Penalized MLE†

† Li and Sudjianto: Analysis of computer experiments using penalized likelihood in Gaussian Kriging models (2005).

EBLUP Properties

- Still interpolates the data
- $l(\widehat{\beta}(\boldsymbol{\theta}), \hat{\sigma}^2(\boldsymbol{\theta}), \boldsymbol{\theta})$ can be very challenging to numerically maximize.
  - straight Newton-Raphson typically unsuccessful
  - A combined strategy such as Nelder-Mead with Newton-Raphson works reasonably well
- Other estimators of $\hat{\sigma}^2$, $\widehat{\boldsymbol{\theta}}$ are possible:
  - REML (Restricted Maximum Likelihood)
  - Penalized MLE†
  - Cross-validation

† Li and Sudjianto: Analysis of computer experiments using penalized likelihood in Gaussian Kriging models (2005).

EBLUP Properties

- Still interpolates the data
- $l(\widehat{\beta}(\boldsymbol{\theta}), \hat{\sigma}^2(\boldsymbol{\theta}), \boldsymbol{\theta})$ can be very challenging to numerically maximize.
  - straight Newton-Raphson typically unsuccessful
  - A combined strategy such as Nelder-Mead with Newton-Raphson works reasonably well
- Other estimators of $\hat{\sigma}^2$, $\widehat{\boldsymbol{\theta}}$ are possible:
  - REML (Restricted Maximum Likelihood)
  - Penalized MLE†
  - Cross-validation
- The EMPSE is a point-wise measure of predictive uncertainty, *not* a path-wise measure of uncertainty.

† Li and Sudjianto: Analysis of computer experiments using penalized likelihood in Gaussian Kriging models (2005).

EBLUP Properties

- Still interpolates the data
- $l(\widehat{\beta}(\boldsymbol{\theta}), \hat{\sigma}^2(\boldsymbol{\theta}), \boldsymbol{\theta})$ can be very challenging to numerically maximize.
  - straight Newton-Raphson typically unsuccessful
  - A combined strategy such as Nelder-Mead with Newton-Raphson works reasonably well
- Other estimators of $\hat{\sigma}^2, \quad \widehat{\boldsymbol{\theta}}$ are possible:
  - REML (Restricted Maximum Likelihood)
  - Penalized MLE†
  - Cross-validation
- The EMPSE is a point-wise measure of predictive uncertainty, *not* a path-wise measure of uncertainty.
- Typically use $\hat{y}(\mathbf{x}_0) \pm 1.96\hat{s}^2(\mathbf{x}_0)$ for a 95% interval.

† Li and Sudjianto: Analysis of computer experiments using penalized likelihood in Gaussian Kriging models (2005).

Restricted/Residual MLE (REML)

- Want to produce estimates of $\sigma^2, \boldsymbol{\theta}$ that are less biased than MLE's.

Restricted/Residual MLE (REML)

- Want to produce estimates of $\sigma^2, \boldsymbol{\theta}$ that are less biased than MLE's.
- Idea: maximize a REML likelihood

Restricted/Residual MLE (REML)

- Want to produce estimates of $\sigma^2, \boldsymbol{\theta}$ that are less biased than MLE's.
- Idea: maximize a REML likelihood
  - simply the likelihood of some transformation of the original data.

Restricted/Residual MLE (REML)

- Want to produce estimates of $\sigma^2, \boldsymbol{\theta}$ that are less biased than MLE's.
- Idea: maximize a REML likelihood
  - simply the likelihood of some transformation of the original data.
  - this transformation is usually a linear combination of the data such that these linear combinations of orthogonal to $\mathbf{F}\beta$.

Restricted/Residual MLE (REML)

- Want to produce estimates of $\sigma^2, \boldsymbol{\theta}$ that are less biased than MLE's.
- Idea: maximize a REML likelihood
  - simply the likelihood of some transformation of the original data.
  - this transformation is usually a linear combination of the data such that these linear combinations of orthogonal to $\mathbf{F}\beta$.
- Assume $\mathbf{F}$ is of full rank ($n \times p$ for $p \leq n$ so $\text{rank}(\mathbf{F}) = p$.)

Restricted/Residual MLE (REML)

- Want to produce estimates of $\sigma^2, \boldsymbol{\theta}$ that are less biased than MLE's.
- Idea: maximize a REML likelihood
  - simply the likelihood of some transformation of the original data.
  - this transformation is usually a linear combination of the data such that these linear combinations of orthogonal to $\mathbf{F}\beta$.
- Assume $\mathbf{F}$ is of full rank ($n \times p$ for $p \leq n$ so $\text{rank}(\mathbf{F}) = p$.)
- Choose an orthogonal matrix $\mathbf{C}$ with rank $n - p$ s.t. $\mathbf{CF} = \mathbf{0}$.

Restricted/Residual MLE (REML)

- Want to produce estimates of $\sigma^2, \boldsymbol{\theta}$ that are less biased than MLE's.
- Idea: maximize a REML likelihood
  - simply the likelihood of some transformation of the original data.
  - this transformation is usually a linear combination of the data such that these linear combinations of orthogonal to $\mathbf{F}\beta$.
- Assume $\mathbf{F}$ is of full rank ($n \times p$ for $p \leq n$ so $\text{rank}(\mathbf{F}) = p$.)
- Choose an orthogonal matrix $\mathbf{C}$ with rank $n - p$ s.t. $\mathbf{CF} = \mathbf{0}$.
- Then, $\widetilde{\mathbf{Y}} = \mathbf{CY} \sim N\left(\mathbf{0}, \sigma^2 \mathbf{CRC}^\mathsf{T}\right)$

Restricted/Residual MLE (REML)

- Want to produce estimates of $\sigma^2, \boldsymbol{\theta}$ that are less biased than MLE's.
- Idea: maximize a REML likelihood
  - simply the likelihood of some transformation of the original data.
  - this transformation is usually a linear combination of the data such that these linear combinations of orthogonal to $\mathbf{F}\beta$.
- Assume $\mathbf{F}$ is of full rank ($n \times p$ for $p \leq n$ so rank$(\mathbf{F}) = p$.)
- Choose an orthogonal matrix $\mathbf{C}$ with rank $n - p$ s.t. $\mathbf{CF} = \mathbf{0}$.
- Then, $\widetilde{\mathbf{Y}} = \mathbf{CY} \sim N\left(\mathbf{0}, \sigma^2 \mathbf{CRC^T}\right)$
- And $L_{REML}(\sigma^2, \boldsymbol{\theta}) = \frac{1}{(2\pi)^{\frac{n-p}{2}} |\sigma^2 \mathbf{CRC^T}|^{-\frac{1}{2}}} exp(-\frac{1}{2\sigma^2} \widetilde{\mathbf{Y}}^T (\mathbf{CRC}^T)^{-1} \widetilde{\mathbf{Y}})$.

REML: Example

- $\mathbf{Y} = (y_1, \ldots, y_n)^T \sim N(\mathbf{1}\beta_0, \sigma^2\mathbf{R})$ and let

$$\mathbf{C} = \begin{bmatrix} 1 & -1 & 0 & \ldots & 0 \\ 1 & 0 & -1 & \ldots & 0 \\ \ldots & 0 & 0 & -1 & \ldots \\ \vdots & \vdots & \vdots & \vdots & \vdots \\ 1 & 0 & \ldots & 0 & -1 \end{bmatrix}$$

REML: Example

- $\mathbf{Y} = (y_1, \ldots, y_n)^T \sim N(\mathbf{1}\beta_0, \sigma^2\mathbf{R})$ and let

$$\mathbf{C} = \begin{bmatrix} 1 & -1 & 0 & \ldots & 0 \\ 1 & 0 & -1 & \ldots & 0 \\ \ldots & 0 & 0 & -1 & \ldots \\ \vdots & \vdots & \vdots & \vdots & \vdots \\ 1 & 0 & \ldots & 0 & -1 \end{bmatrix}$$

- Then $\widetilde{\mathbf{Y}} = \begin{pmatrix} y_1 - y_2 \\ y_1 - y_3 \\ \vdots \\ y_1 - y_n \end{pmatrix} \sim N\left(\mathbf{0}, \sigma^2\mathbf{CRC}^T\right)$

Penalized MLE

- Given some $\lambda > 0$, this method finds $\sigma^2, \boldsymbol{\theta}$ that maximizes

$$L_p(\sigma^2, \boldsymbol{\theta}|\mathbf{y}, \lambda) = L(\widehat{\beta}(\boldsymbol{\theta}), \sigma^2, \boldsymbol{\theta}|\mathbf{y}) - n \sum_{k=1}^{d} p_\lambda(\theta_k)$$

where $\widehat{\beta}(\boldsymbol{\theta})$ is the usual GLS solution and the penalty $p_\lambda(\boldsymbol{\theta})$ grows as $\boldsymbol{\theta}$ grows.

† Fan and Li: Variable selection via nonconcave penalized likelihood and its oracle properties (2001)

Penalized MLE

- Given some $\lambda > 0$, this method finds $\sigma^2, \boldsymbol{\theta}$ that maximizes

$$L_p(\sigma^2, \boldsymbol{\theta}|\mathbf{y}, \lambda) = L(\widehat{\beta}(\boldsymbol{\theta}), \sigma^2, \boldsymbol{\theta}|\mathbf{y}) - n \sum_{k=1}^{d} p_\lambda(\theta_k)$$

  where $\widehat{\beta}(\boldsymbol{\theta})$ is the usual GLS solution and the penalty $p_\lambda(\boldsymbol{\theta})$ grows as $\boldsymbol{\theta}$ grows.
- Eg:

† Fan and Li: Variable selection via nonconcave penalized likelihood and its oracle properties (2001)

Penalized MLE

- Given some $\lambda > 0$, this method finds $\sigma^2, \boldsymbol{\theta}$ that maximizes

$$L_p(\sigma^2, \boldsymbol{\theta}|\mathbf{y}, \lambda) = L(\widehat{\beta}(\boldsymbol{\theta}), \sigma^2, \boldsymbol{\theta}|\mathbf{y}) - n \sum_{k=1}^{d} p_\lambda(\theta_k)$$

  where $\widehat{\beta}(\boldsymbol{\theta})$ is the usual GLS solution and the penalty $p_\lambda(\boldsymbol{\theta})$ grows as $\boldsymbol{\theta}$ grows.
- Eg:
  - $p_\lambda(\theta) = \lambda|\theta|$ (linear penalty)

† Fan and Li: Variable selection via nonconcave penalized likelihood and its oracle properties (2001)

Penalized MLE

- Given some $\lambda > 0$, this method finds $\sigma^2, \boldsymbol{\theta}$ that maximizes

$$L_p(\sigma^2, \boldsymbol{\theta}|\mathbf{y}, \lambda) = L(\widehat{\beta}(\boldsymbol{\theta}), \sigma^2, \boldsymbol{\theta}|\mathbf{y}) - n \sum_{k=1}^{d} p_\lambda(\theta_k)$$

  where $\widehat{\beta}(\boldsymbol{\theta})$ is the usual GLS solution and the penalty $p_\lambda(\boldsymbol{\theta})$ grows as $\boldsymbol{\theta}$ grows.
- Eg:
  - $p_\lambda(\theta) = \lambda|\theta|$ (linear penalty)
  - $p_\lambda(\theta) = \lambda\theta^2/2$ (quadratic penalty)

† Fan and Li: Variable selection via nonconcave penalized likelihood and its oracle properties (2001)

Penalized MLE

- Given some $\lambda > 0$, this method finds $\sigma^2, \boldsymbol{\theta}$ that maximizes

$$L_p(\sigma^2, \boldsymbol{\theta}|\mathbf{y}, \lambda) = L(\widehat{\beta}(\boldsymbol{\theta}), \sigma^2, \boldsymbol{\theta}|\mathbf{y}) - n\sum_{k=1}^{d} p_\lambda(\theta_k)$$

  where $\widehat{\beta}(\boldsymbol{\theta})$ is the usual GLS solution and the penalty $p_\lambda(\boldsymbol{\theta})$ grows as $\boldsymbol{\theta}$ grows.
- Eg:
  - $p_\lambda(\theta) = \lambda|\theta|$ (linear penalty)
  - $p_\lambda(\theta) = \lambda\theta^2/2$ (quadratic penalty)
  - "smooth clipped absolute deviation"†

† Fan and Li: Variable selection via nonconcave penalized likelihood and its oracle properties (2001)

Penalized MLE

- Given some $\lambda > 0$, this method finds $\sigma^2, \boldsymbol{\theta}$ that maximizes

$$L_p(\sigma^2, \boldsymbol{\theta} | \mathbf{y}, \lambda) = L(\widehat{\beta}(\boldsymbol{\theta}), \sigma^2, \boldsymbol{\theta} | \mathbf{y}) - n \sum_{k=1}^{d} p_\lambda(\theta_k)$$

  where $\widehat{\beta}(\boldsymbol{\theta})$ is the usual GLS solution and the penalty $p_\lambda(\boldsymbol{\theta})$ grows as $\boldsymbol{\theta}$ grows.
- Eg:
  - $p_\lambda(\theta) = \lambda|\theta|$ (linear penalty)
  - $p_\lambda(\theta) = \lambda\theta^2/2$ (quadratic penalty)
  - "smooth clipped absolute deviation"†
- Choose λ by cross-validation

† Fan and Li: Variable selection via nonconcave penalized likelihood and its oracle properties (2001)

Cross-Validation Fitting

- Given $\boldsymbol{\theta}$ and $\hat{y}_{-i}(\boldsymbol{\theta}) :=$ BLUP of $y(\mathbf{x}_i)$ with correlation $\boldsymbol{\theta}$ based on the data $\{\mathbf{x}_j, y(\mathbf{x}_j)\}_{j \neq i}$,

$$\hat{y}_{-i}(\boldsymbol{\theta}) = \mathbf{f}^T(\mathbf{x}_i)\widehat{\beta}(\boldsymbol{\theta}) + \mathbf{r}_0^T(\mathbf{x}_i)\mathbf{R}^{-1}(\mathbf{y}_{-i} - \mathbf{F}\widehat{\beta})$$

Cross-Validation Fitting

- Given $\boldsymbol{\theta}$ and $\hat{y}_{-i}(\boldsymbol{\theta}) :=$ BLUP of $y(\mathbf{x}_i)$ with correlation $\boldsymbol{\theta}$ based on the data $\{\mathbf{x}_j, y(\mathbf{x}_j)\}_{j \neq i}$,

$$\hat{y}_{-i}(\boldsymbol{\theta}) = \mathbf{f}^T(\mathbf{x}_i)\widehat{\beta}(\boldsymbol{\theta}) + \mathbf{r}_0^T(\mathbf{x}_i)\mathbf{R}^{-1}(\mathbf{y}_{-i} - \mathbf{F}\widehat{\beta})$$

- Choose $\widehat{\boldsymbol{\theta}}$ as

$$\widehat{\boldsymbol{\theta}}_{CV} = \arg\min_{\boldsymbol{\theta}} \sum_{i=1}^{n} \left(y(\mathbf{x}_i) - \hat{y}_{-i}(\mathbf{x}_i, \boldsymbol{\theta})\right)^2$$

and

$$\widehat{\sigma}_{cv}^2 = \frac{1}{n}\left(\mathbf{y}_n - \mathbf{F}\widehat{\beta}_{CV}\right)^T \widehat{\mathbf{R}}^{-1}(\widehat{\boldsymbol{\theta}}_{CV})\left(\mathbf{y}_n - \mathbf{F}\widehat{\beta}_{CV}\right)$$

where $\widehat{\beta}_{CV} = \widehat{\beta}(\widehat{\boldsymbol{\theta}}_{CV})$.