Advanced Regression

Assignment 2

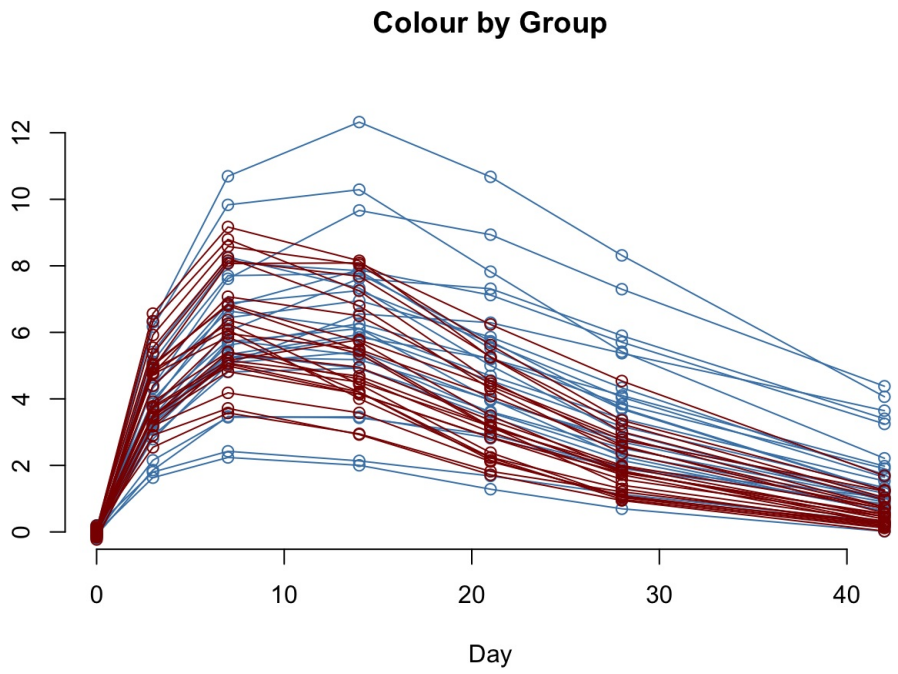
Zach Wolpe

wlpzac001

Fitting Bayesian Regression Splines to non-linear Temporal data.

Question A

First let's examine the raw data, gametocyte freqencies at each measured day, coloured by group.



Consider the Model

/Users/zachwolpe/Desktop/MSc Advanced Analytics/Advanced Regression/Assignments/AR Assignment 2/images

For each of the profiles of group 1 ($i = 1, \ldots, n = 25$) consider fitting the following model (for $j = 1, \ldots, 7$)

$$y_{i,j} = \beta_0 + \sum_{k=1}^{L^*} \tilde{\phi}_k b(|t_j - \xi_k^{(1)}|) + e_{i,j} \text{ for } j = 1, \ldots, 7$$

where $b(x) = x \log(x)$ and $\boldsymbol{\xi}^{(1)} = [0.5, 10, 25]^T$. Assume that $e_{i,j} \overset{iid}{\sim} \mathcal{N}(0, \sigma_e^2)$. Let

$$\boldsymbol{\phi} = \begin{bmatrix} \beta_0 & \tilde{\phi}_1 & \ldots & \tilde{\phi}_3 \end{bmatrix}^T = \begin{bmatrix} \boldsymbol{\beta} \\ \tilde{\boldsymbol{\phi}} \end{bmatrix}$$

such that the prior distribution of $\boldsymbol{\phi}$ is proportional to

$$\boldsymbol{\phi}|\sigma_\phi \propto (\sigma_\phi^2)^{-L^*/2} \exp\left(-\frac{1}{2\sigma_\phi^2}\boldsymbol{\phi}^T \boldsymbol{D}\boldsymbol{\phi}\right) \tag{1}$$

and

$$\boldsymbol{D} = \begin{bmatrix} \mathbf{0}_{1\times 1} & \mathbf{0}_{1\times L^*} \\ \mathbf{0}_{L^*\times 1} & \boldsymbol{I}_{L^*} \end{bmatrix}.$$

Assume that $\sigma_e^2 \sim \mathcal{IG}(i_1, i_2)$ and $\sigma_\phi^2 \sim \mathcal{IG}(i_3, i_4)$. Suggest appropriate values for i_1, i_2, i_3 and i_4. Take note that equation (1) implies:

$$\begin{aligned} [\beta_0] &\propto 1, \\ \tilde{\boldsymbol{\phi}} &\sim \mathcal{N}\left(\mathbf{0},\ \sigma_\phi^2 \boldsymbol{I}_{L^*}\right). \end{aligned}$$

Question A.i

Gibbs Sampler

In order to learn the parameter posterior distributions, one ought to derive the appropriate posteriors functional forms. The model requires learning the following parameters:

$$\boldsymbol{\Theta} = \{\boldsymbol{\phi}, \hat{\sigma_e^2}, \hat{\sigma_\phi^2}\}$$

φ Posterior

Derive the posterior distribution of the parameter estimate $\boldsymbol{\phi}$:

$$\begin{aligned} \boldsymbol{\phi} &\propto prior \times likelihood \\ \boldsymbol{\phi} &\propto exp(-\frac{1}{2\sigma_\phi^2}\boldsymbol{\phi}`D\boldsymbol{\phi}) \times exp(-\frac{1}{2\sigma_e^2}(y - X\boldsymbol{\phi})`(y - X\boldsymbol{\phi})) \\ \boldsymbol{\phi} &\propto exp(\frac{1}{\sigma_\phi^2}\boldsymbol{\phi}`D\boldsymbol{\phi} + \frac{1}{\sigma_e^2}y`y - \frac{1}{\sigma_e^2}2y`X\boldsymbol{\phi} + \frac{1}{\sigma_e^2}\boldsymbol{\phi}`X`X\boldsymbol{\phi}) \\ \boldsymbol{\phi} &\propto exp(\boldsymbol{\phi}`[\frac{1}{\sigma_\phi^2}D + X`X\frac{1}{\sigma_e^2}]\boldsymbol{\phi} - \frac{1}{\sigma_e^2}2y`X\boldsymbol{\phi}) + \dots \\ \boldsymbol{\phi} &\propto exp((\boldsymbol{\phi} - \mu_\phi)`\Sigma_\phi^{-1}(\boldsymbol{\phi} - \mu_\phi)) \end{aligned}$$

Where:

$$

$$\begin{aligned} \Sigma_\phi &= [\frac{1}{\sigma_\phi^2}D + X`X\frac{1}{\sigma_e^2}]^{-1} \\ \mu_\phi &= \Sigma_\phi \frac{1}{\sigma_e^2}X`y \end{aligned}$$

$$

This implies the posterior of $\boldsymbol{\phi}$ to be given by a multivariate normal distribution:

$$\pi(\boldsymbol{\phi}|x, \sigma_e^2, \sigma_\phi^2) \sim \mathcal{N}_p(\mu_\phi, \Sigma_\phi)$$

σ variance components posteriors

The variance components are known to have a inverse gamma posterior distributions, thus *Inverse Gamma prior distributions* are set for conjugacy:

$$\begin{aligned} \sigma_e^2 &\sim IG(i_1, i_2) \\ \sigma_\phi^2 &\sim IG(i_3, i_4) \end{aligned}$$

Whilst the priors are identical, the two parameters use different *likelihoods* to compute the posterior - σ_e^2 replies on the likelihood of the model, thus $\mathcal{L}(y)$ - σ_ϕ^2 replies on the likelihood of the ϕ parameter, thus $\mathcal{L}(\phi)$

The posterior of the error variance is then computed as

$$\pi(\sigma_e^2|x) \sim IG(i_1 + \frac{n}{2}, i_2 + \frac{1}{2}(y - X\phi)'(y - X\phi))$$

& similarly the posterior of variance of φ, where d is the rank of ϕ

$$\pi(\sigma_\phi^2|x) \sim IG(i_3 + \frac{d}{2}; i_4 + \frac{1}{2}\phi' D\phi)$$

Define Gibbs Sampler

Now the we have the relevant posterior distributions we can iteratively sample until convergence - implement a Gibbs Sampler - to learn the posterior distributions.

Here we define the Gibbs Sampler for this penalized least squares parameter & variance component estimation.

```
## Loading required package: MASS
```

```
##
## Attaching package: 'MASS'
```

```
## The following object is masked from 'package:dplyr':
##
##     select
```

Compute for Inidividual Profiles

A model is fit for each individual profile. The Gibbs sampler is implemented to learn the distribution of each set of parameters for each profile.

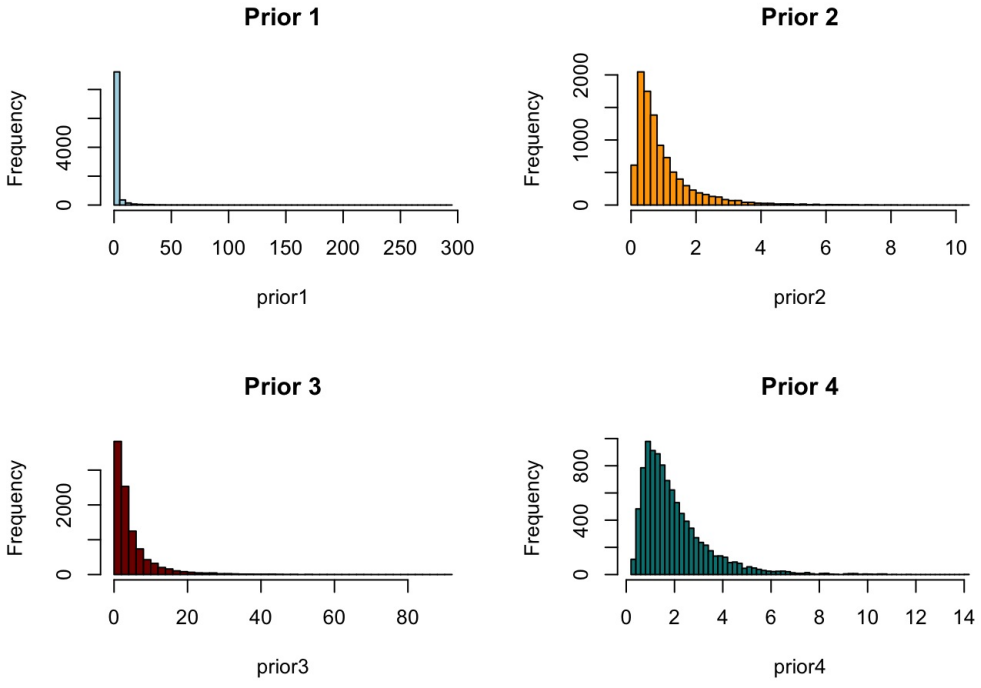
The Sampler returns a list $phis: the parameter estimates $sig_e: sig_e $sig_phi: sig_phi

Each model has the same number of parameters to estimate, as such we will store the results in a list of list (each sub-list containing these parameter estimates).

Prior specification

Here we examine 4 candidate priors, we have no assumptions of the true posterior & thus havent a way of setting the prior intelligently: - $i_1 = 2;\ i_2 = 0.1$ - $i_1 = 1;\ i_2 = 1$ - $i_1 = 5;\ i_2 = 3$ - $i_1 = 2;\ i_2 = 4$

Here is a visualization of these prior distributions:



Prior 3 appears favourable, it's decreasing in x through the entire range (as in prior 1 but unlike prior's 2 & 4) & still covers a large spectrum of possible values (unlike 1).

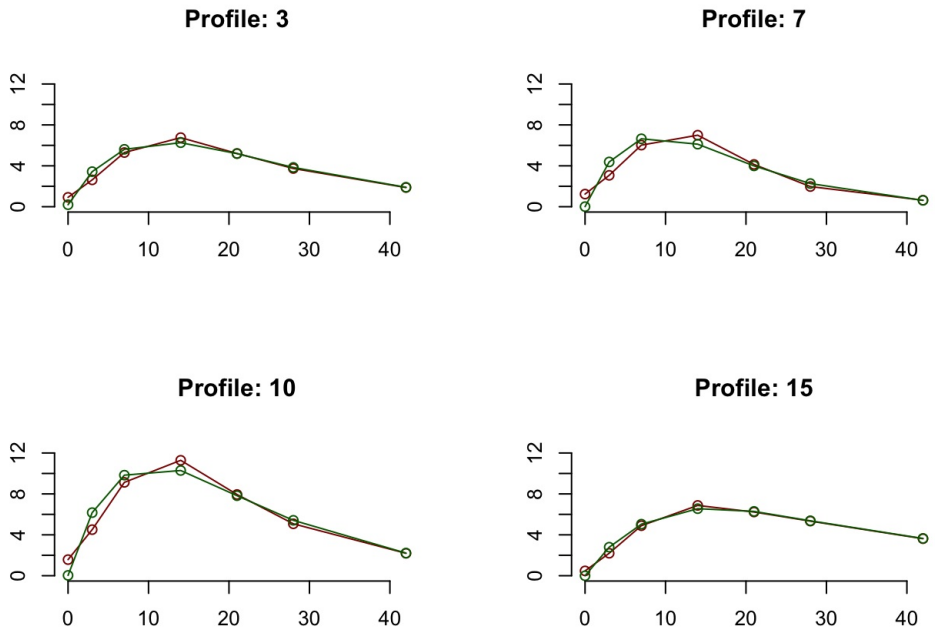
We will use prior 3 for both inverse gamma prior values.

Run Computation

Now that priors are selected we can run the model:

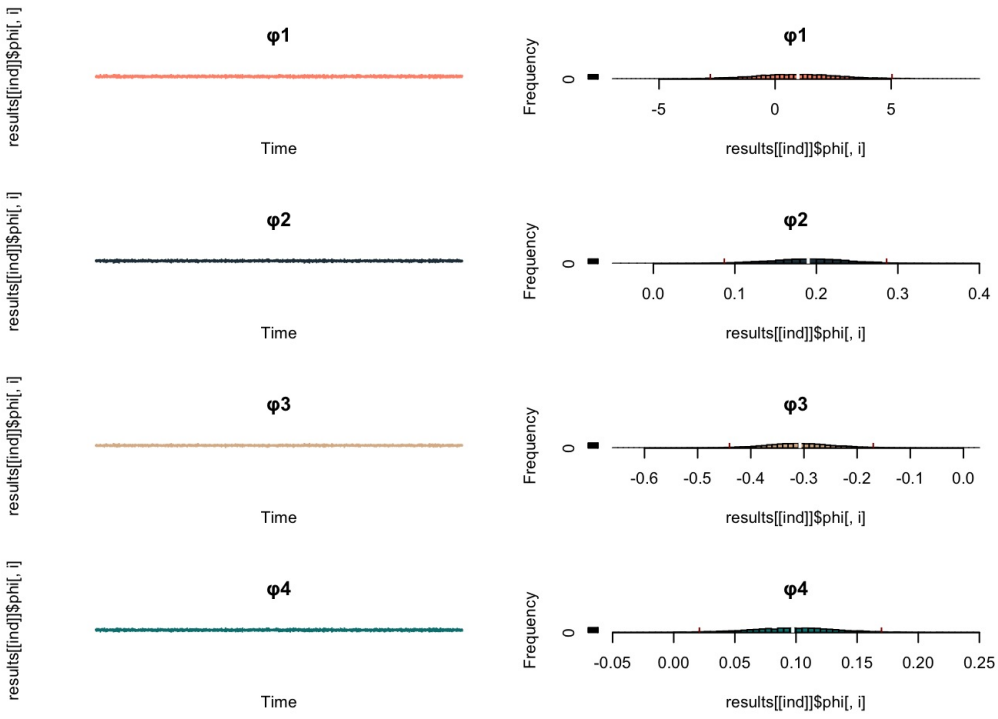
A.ii Examine Results

Now that the model has been fit for each profile, we wish to examine each fit. Examining all 25 fits is cumbersome & tedious, instead the we sample 4 fits at random & thereafter visualize the profile's results.



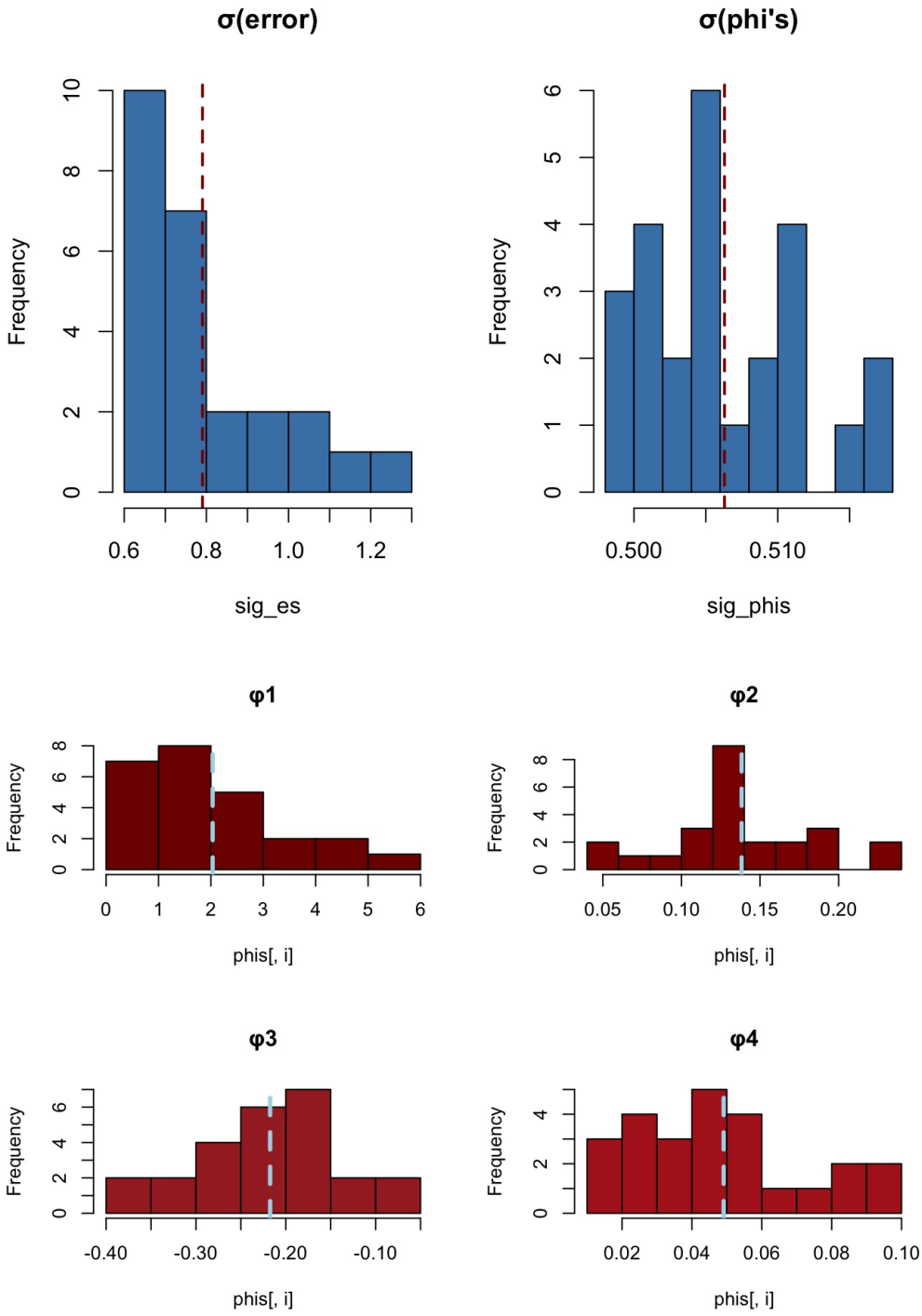
The true values is given in green whilst the fitted value is given in red. Problematically, the model's appear to overfit the data. Given the ratio of the number of parameters to number of samples $(4:7)$ the model is nearly saturated. As such the fit is extremely tight to the given profiles & probably won't generalize well.

Now to exmamine the learnt parameters. Again, examining all profiles is cumbersome & a random will be chosen to be examined.



A random sample of coefficents appear to be roughly normally distributed and have converged to stationary samples. The parameters appear to have been fit correctly.

Suppose we want to examine the range of parameter estimates over all inidividual samples, to assess the variablitly of parameter coefficients across profiles. Here we examine the range of *mean* parameter estimates over the 25 profiles.



Conculsion & Discussion

The D penalty matrix keeps all ϕ parameter fits small, whist the intercept is free to take on larger values. Parameter estimates may converge to normally distributed if we had more data.

Variance components maintain the *inverse gamma* structure.

Question B

Consider:

Use **all of the data** (both group 1 and group 2 simultaneously) and fit the following model (for $j = 1, \ldots, 7$):

$$y_{i,j} = \beta_0 + \beta_1 \text{group}_i + \sum_{k=1}^{L^*} \tilde{\phi}_{i,k} b(|t_j - \xi_k^{(1)}|) + e_{i,j} \text{ for } i = 1, \ldots, 25, \tag{2}$$

$$y_{i,j} = \beta_0 + \beta_1 \text{group}_i + \sum_{k=1}^{L^*} \tilde{\phi}_{i,k} b(|t_j - \xi_k^{(2)}|) + e_{i,j} \text{ for } i = 26, \ldots, 50 \tag{3}$$

where $\boldsymbol{\xi}^{(1)} = [0.5, 10, 25]^T$ and $\boldsymbol{\xi}^{(2)} = [0.5, 5, 25]^T$. The variable 'group' is an indicator variable which is coded 1 for group 1 and 0 for group 2. Take note that β_0 and β_1 are the same for both groups although the spline regression coefficients are allowed to be different for each profile. Assume that both σ_e^2 and σ_ϕ^2 are common for both groups (i.e. $\sigma_e^2 \sim \mathcal{IG}(i_1, i_2)$ and $\sigma_\phi^2 \sim \mathcal{IG}(i_3, i_4)$). Further assume that each $e_{i,j}$ are independently and identically distributed (for all i, j).

Derive the Posterior Distributions

In order to learn the parameter posterior distributions, one ought to derive the appropriate posteriors functional forms. The model requires learning the following parameters:

$$\boldsymbol{\Theta} = \{\phi, \hat{\sigma_e^2}, \hat{\sigma_\phi^2}\}$$

φ Posterior

Derive the posterior of the parameter estimate ϕ:

$$\begin{aligned}
\phi &\propto prior \times likelihood \\
\phi &\propto exp(-\frac{1}{2\sigma_\phi^2}\phi`D\phi) \times exp(-\frac{1}{2\sigma_e^2}(y - X\phi)`(y - X\phi)) \\
\phi &\propto exp(\frac{1}{\sigma_\phi^2}\phi`D\phi + \frac{1}{\sigma_e^2}y`y - \frac{1}{\sigma_e^2}2y`X\phi + \frac{1}{\sigma_e^2}\phi`X`X\phi) \\
\phi &\propto exp(\phi`[\frac{1}{\sigma_\phi^2}D + X`X\frac{1}{\sigma_e^2}]\phi - \frac{1}{\sigma_e^2}2y`X\phi) + \ldots \\
\phi &\propto exp((\phi - \mu_\phi)`\Sigma_\phi^{-1}(\phi - \mu_\phi))
\end{aligned}$$

Where:

$$

$$\begin{aligned}
\Sigma_\phi &= [\frac{1}{\sigma_\phi^2}D + X`X\frac{1}{\sigma_e^2}]^{-1} \\
\mu_\phi &= \Sigma_\phi \frac{1}{\sigma_e^2}X`y
\end{aligned}$$

$$

This implies the posterior of ϕ to be given by:

$$\pi(\phi | x, \sigma_e^2, \sigma_\phi^2) \sim \mathcal{N}(\mu_\phi, \Sigma_\phi)$$

σ cvariance components posteriors

The variance components have known to to *Inverse Gamma prior distributions*:

$$\begin{aligned}
\sigma_e^2 &\sim IG(i_1, i_2) \\
\sigma_\phi^2 &\sim IG(i_3, i_4)
\end{aligned}$$

Whilst the priors are identical, the two parameters use different *likelihoods* to compute the posterior - σ_e^2 replies on the likelihood of the model, thus $\mathcal{L}(y)$ - σ_ϕ^2 replies on the likelihood of the ϕ parameter, thus $\mathcal{L}(\phi)$

The posterior of the error variance is then computed as

$$[\sigma_e^2|x,\phi,\sigma_\phi^2] \propto \sigma_e^{2^{-(i_1+1)}} e^{-i_2/\sigma_e^2} \sigma_e^{2^{-n/2}} exp(-\frac{1}{2\sigma_e^2}(y-X\phi)`(y-X\phi))$$

similiarly:

$$[\sigma_\phi^2|x,\phi,\sigma_e^2] \propto \sigma_\phi^{2^{-(i_1+1)}} e^{-i_2/\sigma_\phi^2} \sigma_\phi^{2^{-n/2}} exp(-\frac{1}{2\sigma_\phi^2}\phi`D\phi)$$

These results yeild the following posteriors:

$$\pi(\sigma_e^2|x) \sim IG(i_1+\frac{n}{2}, i_2+\frac{1}{2}(y-X\phi)`(y-X\phi))$$

& similarly the posterior of variance of φ, where d is the rank of ϕ

$$\pi(\sigma_\phi^2|x) \sim IG(i_3+\frac{d}{2}; i_4+\frac{1}{2}\phi`D\phi)$$

Penalty Matrix: D

D defines a penalty/shrinkage applied to the parameter coefficients. Though the global parameters β_0 & β_1 are not penalized, local parameters $\phi_{i,k}$ are penalized with an L2 penalty (sum of the square coefficents is zero).

Design Matrix: X

The design matrix should be specified such that the β parameters are global (shared by each profile) & the ϕ parameters are local (unique to each profile). To achieve this a sparse design matrix of dimensions $(7 \times 50) \times (7 \times 3 + 2)$ is computed in which the first two columns are a columns of $1's$ for the intercept & a binary dummy variable $\in \{0,1\}$ representing a group variable. The remain rows & columns are set to the function $b(|t_j - \xi_k^{group}|)$ over a column & row range for each profile & zero elsewhere.

Fit the model

$Group$ variable is coded as:

$$Group: \ \{group1 = 0; \ \ group2 = 1\}$$

Here we compute the design matrix X & penalty matrix D.

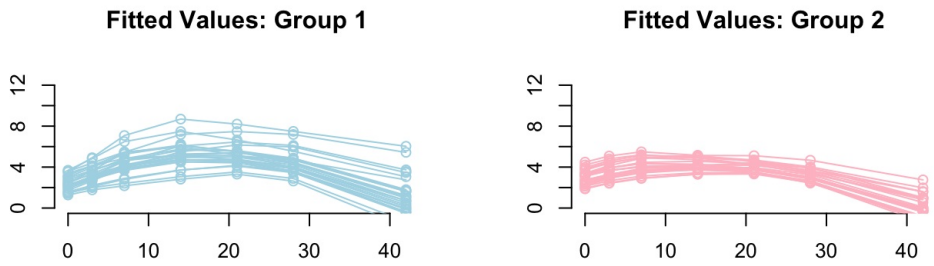
Estimate parameters with Gibbs Sampler, assuming a prior structure:

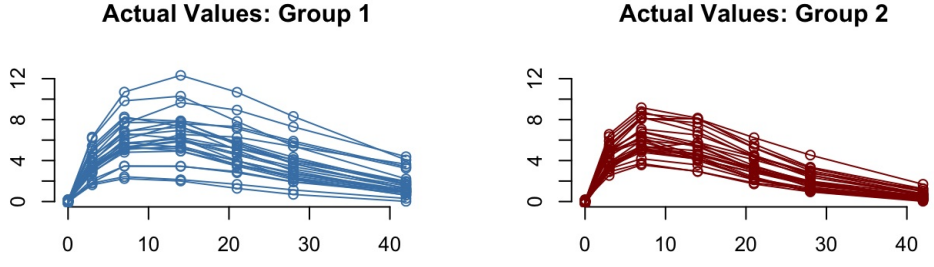
$$covariance\ hyper-parameters: \ \{i_1 = 2; \ i_2 = 0.1; \ i_3 = 5, \ i_4 = 0.1\}$$

Visualize the results, we only examine the first few ϕ parameter values:

Parameter estimates appear to be sampling from stationary distributions. Covariance estimates appear slightly skewed, as expected, towards an inverse gamma form. $95\%\ credibility\ intervals$ and mean results are also provided.

Plot the Fitted Mødel:





Interestly, if 1000 iterations of the gibbs sampler are used, the parameter estimates don't appear stationary, however the fitted values appear far closer to the true values. Whereas if $2000 >$ iterations are sampled, parameter estimates are more stable however the fitted values do not, as closely, mimick the actual profiles. I'm not sure why this could be?

Conclusion & Discussion

The model appears to fit the data very well, parameter estimates have converged to normally distributed & are stationary. The fitted values appear to repesent the actual values well.

One caveat is the senstivity to variance component prior hyperparameters. The prior specification yeilds a major influence on the posterior results.

Here we run the model for 10 differnt prior specifications, capturing variance relationships between hyperparameters to assess the impact of the prior choice.

The follow prior variants were used: 1. $i_1 = 1;\ i_2 = 1;\ i_3 = 1,\ i_4 = 1$ 2. $i_1 = 3;\ i_2 = 1;\ i_3 = 3,\ i_4 = 1$ 3. $i_1 = 3;\ i_2 = 1;\ i_3 = 1,\ i_4 = 1$ 4. $i_1 = 1;\ i_2 = 1;\ i_3 = 3,\ i_4 = 1$ 5. $i_1 = 0.1;\ i_2 = 2;\ i_3 = 0.1,\ i_4 = 2$ 6. $i_1 = 0.1;\ i_2 = 5;\ i_3 = 0.1,\ i_4 = 0.1$ 7. $i_1 = 0.1;\ i_2 = 5;\ i_3 = 5,\ i_4 = 0.1$ 8. $i_1 = 0.1;\ i_2 = 0.1;\ i_3 = 5,\ i_4 = 5$ 9. $i_1 = 2;\ i_2 = 0.1;\ i_3 = 5,\ i_4 = 0.1$ 10. $i_1 = 5;\ i_2 = 5;\ i_3 = 5,\ i_4 = 5$

Note:

$$\lambda = \frac{\sigma_\phi^2}{\sigma_e^2}$$

and that GCV is given by:

$$\mathrm{GCV}(\hat{f}) = \frac{1}{N}\sum_{i=1}^{N}\left[\frac{y_i - \hat{f}(x_i)}{1 - \mathrm{trace}(\mathbf{S})/N}\right]^2$$

Where $S = (X`X + \lambda D)^{-1}X`$

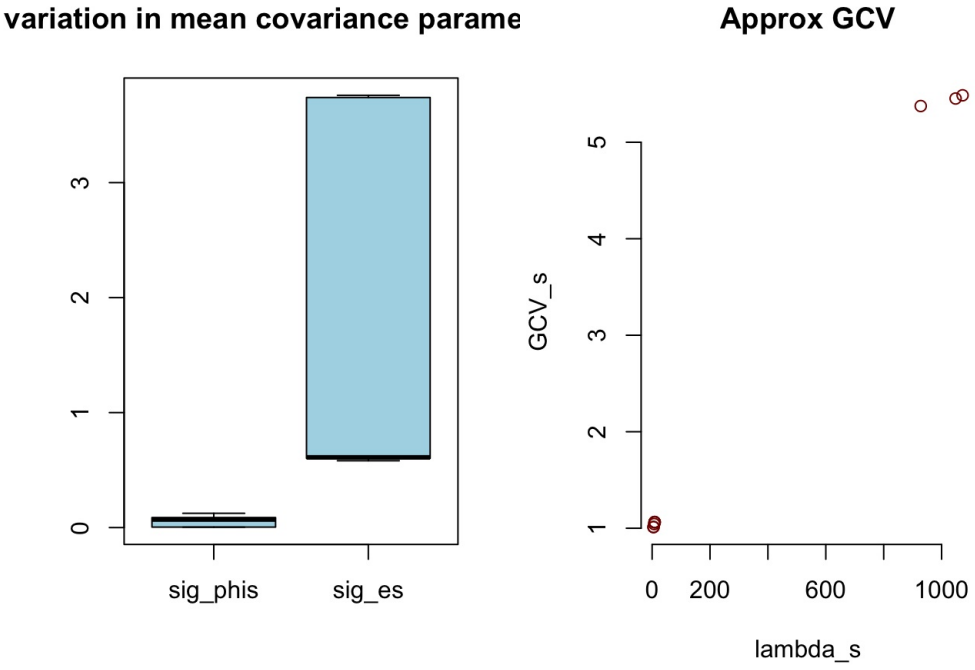
Given the formulation of GCV, the hat matrix (S - effective degrees of freedom) is needed in order to compute GCV - a function of $trace(S)$. This isn't strictly correct in this sense, as the model is not fit by the projection matrix, but rather by learning the parameters via the Gibbs sampler.

Nonetheless, I see no reason this hat matrix cannot be used as an effective degrees of freedom, as the $\lambda = \frac{\sigma_\phi^2}{\sigma_e^2}$ captures the degrees of freedom in the model.

Thus here we fit the predicted values $\hat{y}$ usign the parameter estimates from the Gibbs Sampler, however use the projection matrix to compute $trace(S)$ for a reasonable GCV estimate.

Now that we've run all those specification, lets examine the variation between mean sample variances - essentially estimating the sensitivity of the model to the priors.

We can also examine the approximate GCV metric for the range of limited lamda values.



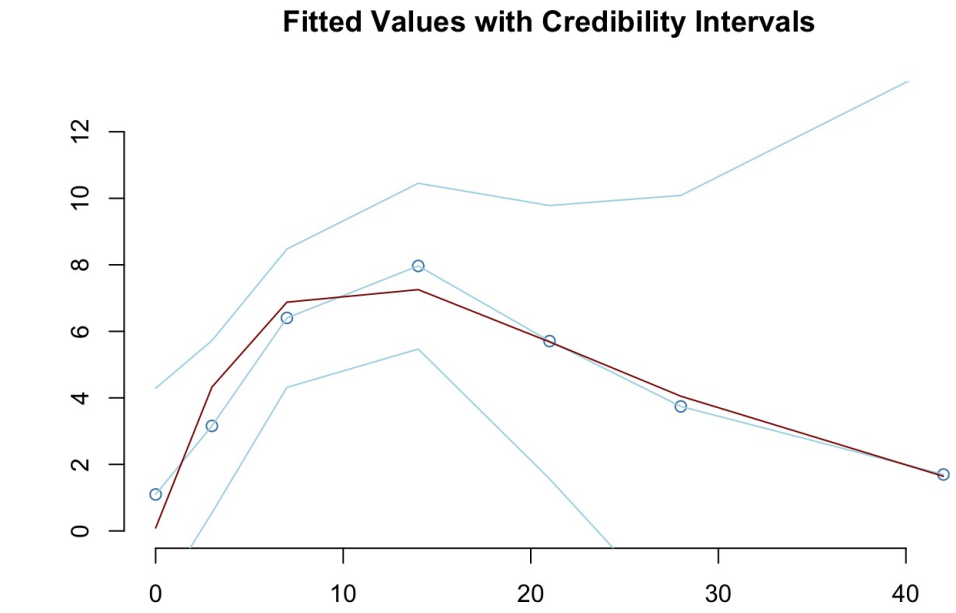
Estimates for σ^2_ϕ don't vary much with the change in priors, σ^2_e appears to have a much larger variance over the hyperparameters.

Selecting hyperparameters may be done with approximate GCV in mind. Some combinations are indistinguishable, though a few parameter specifications yeild exceedingly large λ values & degrees of freedom.

Credibility Intervals

Finally we used one of these better priors (one with the lowest GCV) to fit the model & visualize a randomly selected fit with credibility intervals.

```
## [1] "minimum GCV is at index: 8"
```



Notice the lack of confidence towards the end as these are unique ϕ values for each profile.