Case Study -1: An introduction to the Emulatorr package

Danny Scarponi

Contents

Chapter 1

Introduction

This tutorial is an introduction to the main functionality of the emulatorr package, using a synthetic example of an epidemiological model. We start by briefly highlighting what history matching with emulation is and how it works.

Computer models, otherwise known as simulators, have been widely used in almost all fields in science and technology. A computer model is a way of representing the fundamental dynamics of a system. Due to the complexity of the interactions within a system, computer models frequently contain large numbers of parameters.

Before using a model (e.g. for prediction or planning) it is fundamental to explore plausible values for its parameters, calibrating the model to the observed data. This poses a significant problem, considering that it may take several minutes or even hours for the evaluation of a single scenario of a complex model. This difficulty is compounded for stochastic models, when tens to thousands of realisations can be required for each scenario. As a consequence, comprehensive analysis of the entire input space, requiring vast numbers of model evaluations, is often unfeasible. Emulation, combined with history matching, allows us to overcome this issue.

1.1 History Matching

History matching concerns the problem of identifying those parameter sets that would give rise to acceptable matches between the model outputs and the observed data. This part of the input space is referred to as non-implausible, while its complement is known as implausible. History matching proceeds as a series of iterations, or waves, where implausible parameter sets are identified and discarded. Each wave focuses the search for implausible space in the space that was characterized as non-implausible in all previous waves: thus the non-implausible space shrinks with each iteration. To decide whether a parameter set x is implausible we introduce the implausibility measure, which evaluates the difference between the emulator and computer model results, weighted by how uncertain we are at x. If such measure is too high, the parameter set is discarded in the next wave of the process.

Note that history matching as just described still relies on the evaluation of the model at a large number of parameter sets, which is often unfeasible. Here is where emulators play a crucial role.

1.2   Emulators

A long established method for handling computationally expensive models is to first construct an emulator: a fast statistical approximation of the model that can be used as a surrogate. In other words, we can think of an emulator as a way of representing our beliefs about the behaviour of a complex model.

First of all the model is run at a manageable number of parameter sets, to provide training data to build the emulator. Then prior beliefs about the structure of the emulator are set to reflect our prior knowledge. Once this is done, the training data can be exploited to adjust our prior beliefs and obtain a more accurate representation of the model. The trained emulator will then provide an expected value of the model output at any parameter set x, along with a corresponding uncertainty estimate reflecting our beliefs about the uncertainty in the approximation.

Emulators have two advantages. First, they are computationally efficient - typically orders of magnitude faster than the computer models they approximate. Second, they allow for the uncertainty in their approximations to be taken into account. These two properties mean that emulators can be used to make inferences as a surrogate for the model itself. In particular, when going through the history matching process, it is possible to evaluate the implausibility measure at any given parameter set by comparing the observed data to the emulator output, rather than the model output. This speeds up the process and allows for a comprehensive exploration of the input space.

1.3   A simple example of history matching and emulation

To show how history matching and emulation work, we present an example with a one-dimensional emulator.

The model we consider is the deterministic function $f(x) = \sin(0.04\pi x)$. The value of $f(x)$ is considered unknown at all parameters, apart from $x = \{0, 10, 20, 30, 45, 50\}$, where the model is run. These six known points are represented by the black dots in the top graph of figure 1.1 and are used to train the emulator. The blue line is the emulator's posterior mean, and the red lines represent its posterior uncertainty. The three horizontal lines represent the observed data ($z = -0.7$) and the 95% confidence interval (± 0.06) that we use to history match the model.

The next step involves choosing an implausibility measure. Such measure weighs the difference between z and the value suggested by the emulator at a given point, taking into consideration all the uncertainties that are present in the system. The bottom graph of figure 1.1 shows the implausibility for the emulator and observed data. The horizontal green line is an implausibility cut-off, which determines whether an input x is implausible or not. The implausibility plot shows that a match between the model's output and the observed data is unlikely to be found for values of x smaller than 30 and larger than 45.

With the emulator and the implausibility measure at our disposal, we can now sample the non-implausible space to obtain a new set of points that will be used to run the model in the next wave
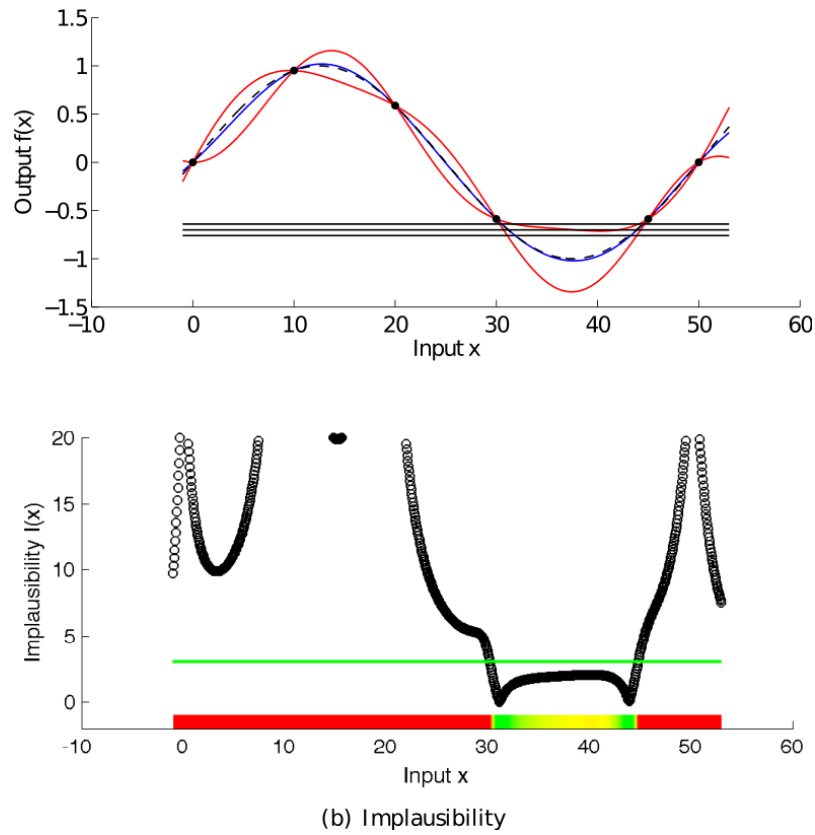
(b) Implausibility

Figure 1.1: First wave of history matching

of the history matching process. Note that at each wave the emulator is only constructed over a smaller region of input space compared to the previous wave: this means that there is a higher density of model runs in the new reduced input space, which leads to improvements in the precision of the emulator.

Figure 1.2 shows the second wave of history matching for our simple sin model. The model was run for the non-implausible value of $x = 36$ and this point was included in the training data. Note how the emulator's posterior variance has decreased in the region of interest. Consequently, the non-implausible region has shrunk dramatically, indicating that a match can only be found for $30.5 < x < 32.5$ and $42.5 < x < 44.5$, where indeed the function $f(x)$ takes values between -0.8 and -0.63.
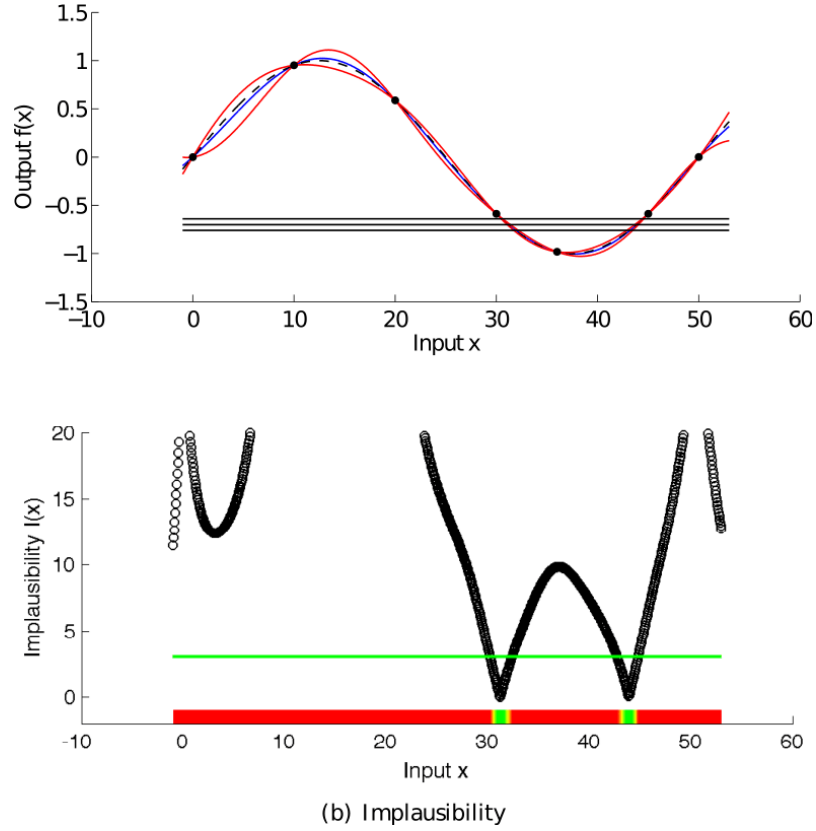

(b) Implausibility

Figure 1.2: Second wave of history matching

Based on the level of accuracy required, one can either stop here or perform more iterations of the process.

Chapter 2

Introduction to the model

The model in question is a stochastic SEIRS model, with four parameters: rate of transmission between each infectious person and each susceptible person β_M; transition rate from exposed to infectious γ_M; recovery rate from infectious to recovered δ_M; and a 'reinfection' rate from recovered to susceptible μ_M.
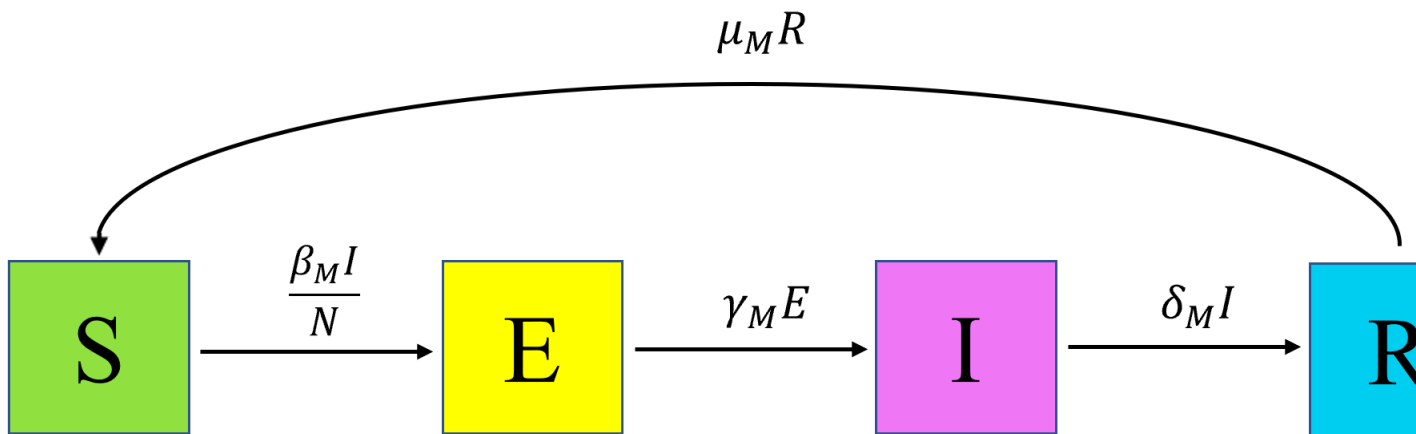

Figure 2.1: SEIRS Diagram

Expressed in terms of differential equations, the transitions are

$$\frac{dS}{dt} = -\frac{\beta_M SI}{N} + \mu_M R \tag{2.1}$$

$$\frac{dE}{dt} = -\gamma_M E + \frac{\beta_M I}{N} \tag{2.2}$$

$$\frac{dI}{dt} = -\delta_M I + \gamma_M E \tag{2.3}$$

$$\frac{dR}{dt} = -\mu_M R + \delta_M I \tag{2.4}$$

where N represents the total population, $N = S + E + I + R$. For simplicity, we consider a closed population, so that N is constant.

To generate runs from this model, we use SimInf, a package that provides a framework to conduct data-driven epidemiological modelling in realistic large scale disease spread simulations. This requires us to define the transitions, the compartments, and the initial population. If we want multiple repetitions for each choice of parameters, we create a data.frame with identical rows, each of which has the same initial population. Here we will choose 50 repetitions per choice of parameters and consider an initial population of 1000 of who 50 are infected. Note that if we were to start with one infectious individual, there would a be a significant probability that some runs of the model would not show an epidemic (since it could happen that the only infectious person recovers before infecting other people). Choosing a relatively high number of initial infectious people helps us circumvent any problems that would come from bimodality and keep the tutorial simple. Bimodality will be dealt in the more advanced case studies.

```
compartments <- c("S","E","I","R")
transitions <- c(
  "S -> beta*I*S/(S+I+E+R) -> E",
  "E -> gamma*E -> I",
  "I -> delta*I -> R",
  "R -> mu*R -> S"
)
nreps <- 50
u0 <- data.frame(
  S = rep(950, nreps),
  E = rep(0, nreps),
  I = rep(50, nreps),
  R = rep(0, nreps)
)
```

We select parameter values and parse the model using the function `mparse`, which takes transitions, compartments, initial values of each compartment, parameter values and the time span to simulate a trajectory. We then run the model and plot the trajectories of interest.

```
params <- c(beta = 0.5, gamma = 0.5, delta = 0.1, mu = 0.1)
model <- mparse(transitions = transitions, compartments = compartments, u0 = u0, gdata = params,
result = run(model)
plot(result, compartments = c("E","I","R"), col=c("#07AF3F", "#FAAE48","#48E2FA"))
```
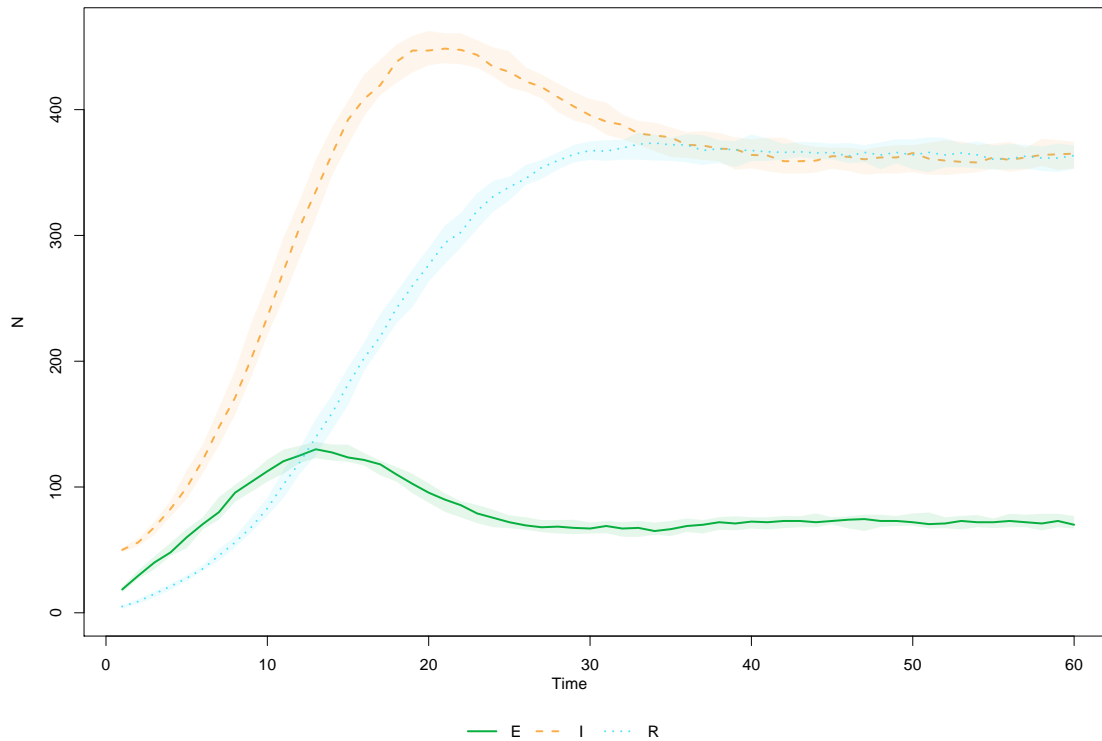
Note that the emulatorr package is code-agnostic. Although we chose SimInf for this case study, the user of emulatorr is completely free to select the package (and programming language) that most suits them to obtain simulations of their computer model.

In order to extract the relevant information from the data provided by the SimInf run, a helper function `getOutputs` has been included in this document. It takes a data.frame of parameter sets, and a list of times, and returns a data.frame of the results. We then create a data.frame `outputs` by binding the parameter values and the results obtained.

```
points <- expand.grid(list(beta = c(0.4, 0.6),
                           gamma = c(0.4, 0.6),
                           delta = c(0.05, 0.15),
                           mu = c(0.05, 0.15)
))
results <- getOutputs(points, seq(10,30,by=5))
outputs <- data.frame(cbind(points, results))
head(outputs)
#>   beta gamma delta   mu    I10    I15    I20    I25    I30      EV10      EV15
#> 1  0.4   0.4  0.05 0.05 215.50 396.98 542.00 574.52 546.36 5.258158 7.709474
#> 2  0.6   0.4  0.05 0.05 363.04 590.78 626.98 573.78 514.00 7.771784 5.396176
```

```
#> 3  0.4    0.6   0.05 0.05 274.94 484.84 591.14 580.04 533.64 9.024302 8.759843
#> 4  0.6    0.6   0.05 0.05 465.08 660.82 631.24 561.88 514.08 7.399687 5.428873
#> 5  0.4    0.4   0.15 0.05 109.00 161.74 206.82 221.50 212.54 4.452561 5.538387
#> 6  0.6    0.4   0.15 0.05 191.50 295.58 312.10 268.90 214.06 6.075497 5.376017
#>       EV20     EV25     EV30
#> 1 6.341308 5.515830 5.381474
#> 2 5.105180 4.071243 4.446290
#> 3 5.450726 4.860454 4.398943
#> 4 4.265616 4.739088 3.583558
#> 5 6.993931 5.403629 4.061191
#> 6 4.488396 4.138197 5.069623
```

Each row of `outputs` corresponds to a parameter set and contains information regarding the number of infectious individuals I for that set. Each row of column $I10$ (resp. $I15, I20, I25, I30$) contains the mean value of I at time 10 (resp. $15, 20, 25, 30$) for the 50 runs of the relative parameter set. Similarly, columns $EV10, EV15, EV20, EV25, EV30$ provide a measure of the ensemble variability for each parameter set, at each desidered time: this is defined here as the standard deviation of the 50 runs, plus 3% of the range of the runs. The trained emulators outputs will be estimates of the means, while the ensemble variability will be used to quantify the uncertainty of such estimates.

Before we tackle the emulation, we need a set of `wave0` data. For this, we define a set of ranges for the parameters, and generate parameter sets using a Latin Hypercube design. We will run the model over 80 parameter sets; 40 of these will be used for training while the other 40 will form the validation set for the emulators.

```
ranges <- list(
  beta = c(0.2, 0.8),
  gamma = c(0.2, 1),
  delta = c(0.1, 0.5),
  mu = c(0.1, 0.5)
)
pts <- 2*(maximinLHS(80, 4)-1/2)
r_centers <- map_dbl(ranges, ~(.[2]+.[1])/2)
r_scales <- map_dbl(ranges, ~(.[2]-.[1])/2)
pts <- data.frame(t(apply(pts, 1, function(x) x*r_scales + r_centers)))
head(pts)
#>        beta     gamma     delta        mu
#> 1 0.4722596 0.5065206 0.2109966 0.2302821
#> 2 0.6695106 0.7461251 0.4244475 0.4043467
#> 3 0.6782483 0.8215380 0.2988149 0.1662161
#> 4 0.5040921 0.4402073 0.3765790 0.3339568
#> 5 0.6595974 0.6816483 0.1633500 0.2398984
#> 6 0.7018773 0.7616248 0.2693328 0.3034066
```

Note that the first time we create `pts`, we get 80 parameter sets where each parameter value is distributed on $[-1, 1]$. This is not exactly what we need, since each parameter has a different range. We therefore define `r_centers` (resp. `r_scales`) which contains the midpoint (resp. the size) of

the range of each parameter. Using these two pieces of information, we re-center and re-scale `pts`.

We obtain the model runs for the parameter sets in `pts` through the `getOutputs` function. We bind the parameter sets in `pts` to the model runs and save everything in the data.frame `wave0`.

```
wave0 <- data.frame(cbind(pts,getOutputs(pts, seq(10,30,by=5)))) %>%
setNames(c(names(ranges), paste0("I",seq(10,30,by=5)),paste0('EV',seq(10,30,by=5))))
head(wave0)
#>        beta    gamma     delta        mu    I10    I15    I20    I25    I30
#> 1 0.4722596 0.5065206 0.2109966 0.2302821 109.18 162.10 208.64 231.36 238.62
#> 2 0.6695106 0.7461251 0.4244475 0.4043467  73.70  95.60 115.86 128.32 136.96
#> 3 0.6782483 0.8215380 0.2988149 0.1662161 173.28 208.54 197.90 180.58 173.78
#> 4 0.5040921 0.4402073 0.3765790 0.3339568  41.58  46.74  53.94  59.14  64.98
#> 5 0.6595974 0.6816483 0.1633500 0.2398984 304.78 414.74 415.94 398.50 393.82
#> 6 0.7018773 0.7616248 0.2693328 0.3034066 213.34 281.18 286.44 280.40 276.96
#>       EV10     EV15     EV20     EV25     EV30
#> 1 6.295766 6.714768 6.233370 5.135637 5.291699
#> 2 4.105439 4.915240 6.099573 4.981953 4.611736
#> 3 6.043500 5.443354 5.163750 4.159341 4.321689
#> 4 3.564083 3.557626 3.004995 3.675507 4.404652
#> 5 7.835108 5.082501 5.248709 4.866771 5.713337
#> 6 6.675898 5.252673 5.061117 4.635021 5.947779
```

Finally, we split `wave0` into two parts: a training set, on which we will train the emulators, and a validation set, which will be used to do diagnostics of the emulators.

```
samples <- sample(nrow(wave0), 40)
train0 <- wave0[samples,1:9]
valid0 <- wave0[!seq_along(wave0[,1])%in%samples,1:9]
```

Chapter 3

Perfoming a full wave of emulation and history matching

In this section we show a simple and direct way of performing a full wave of emulation and history matching (the first wave). This is done by using the function `full_wave`, which needs the following information:

- Training data;

- Validation data;

- A list of ranges for the parameters;

- A list with the names of the model outputs to emulate;

- The targets: for each of the model outputs to emulate, we need a pair (val, sigma) that we will use to evaluate implausibility. The 'val' component represents the mean value of the output and 'sigma' represents our uncertainty about it;

- The number of parameter sets to generate for the next wave;

- The sampling method we want to use.

We already have almost all of these pieces. We just need the model output names

```
output_names <- paste0("I", seq(10,30, by=5))
```

and the targets:

```
targets = list(
  I10 = list(val = 240, sigma = 25.27),
  I15 = list(val = 396, sigma = 40.99),
  I20 = list(val = 453, sigma = 46.48),
  I25 = list(val = 428, sigma = 43.98),
```

```
  I30 = list(val = 392, sigma = 40.30)
)
```

<div style="background:grey">Show: More on how targets were set on P**??**</div>

```
test_full_wave <- full_wave(train0, valid0, ranges, output_names, targets, 120, sample_method =
#> Running diagnostics...
#> Generating new sample points...
#> 497 non-implausible points generated. Applying V-optimality...
#>
```

Here we set the argument `sample_method` to 'importance', to specify that importance sampling should be used when selecting parameter sets for the next wave. The `full_wave` function does the following for us:

- creates `base_emulators`: these are a premilinary version of the emulators necessary to set our priors for the Bayes Linear method;

- creates the `emulators`: these are obtained by adjusting the base emulators through the Bayes Linear update formulae;

- provides us with `new_ranges`, the parameter ranges for the next wave,

- provides us with `next_sample`, the new sample parameter sets, where the model will be run to build the next wave emulators.

To see how the parameter space has changed after the first wave of the process, we use the function `wave_points`, which plots the old and the new set of parameters on the same pair of axis:

```
wave_points(list(wave0, test_full_wave$next_sample), names(ranges))
```

Wave points pairs plot
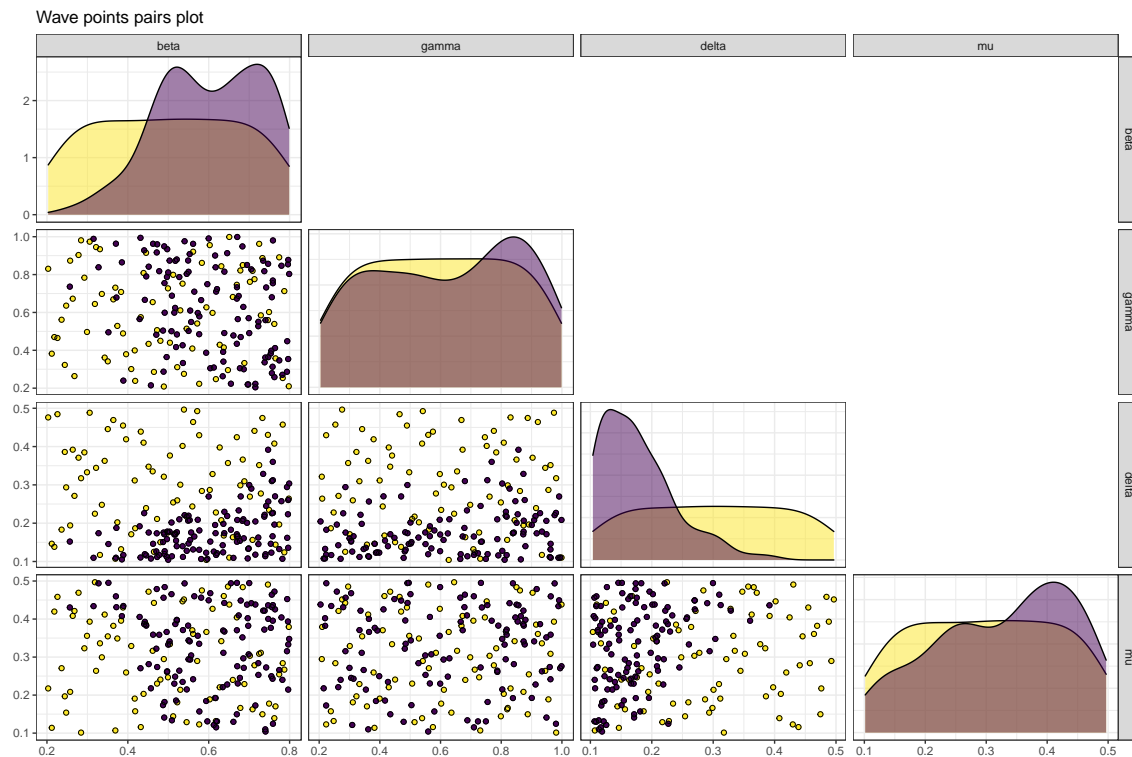
Here `wave0` parameter sets are in yellow and the new sets are in purple. The plots in the main diagonal show the distribution of points in the two sets. Let us take a look at the yellow and purple distributions for the delta parameter. While the distribution of points in `wave0` is rather uniform, the distribution of the new wave peaks at low values of delta and decreases to zero for large values of delta. Similarly, if we look at the delta-mu plot, the yellow points are uniformly distributed, while the purple points are concentrated in the upper left region: this suggests that parameter sets with high values of delta and low values of mu are unlikely to give a good fit for the model.

In the following sections we will explain step by step what `full_wave` does behind the scenes. This will not only enhance the reader's overall understanding, but will also provide them with the necessary tools to have more control over the process and customise it through their judgement.

Chapter 4

Constructing the emulators

The first task that the `full_wave` function accomplishes, is to build the emulators based on the training data. We start this section by establishing the structure of the emulators that we want to construct. We then show how to build emulators step by step.

4.1 Background: the structure of an emulator

An emulator is a way of representing our beliefs about the behaviour of an unknown function. In our case, where the model is stochastic, the unknown function is taken to be the mean of each of the model outputs over multiple runs. Given a set of model runs and an estimate of the parameters, we can use the emulator to get a probability distribution for the mean of a model output at any parameter set, without the need to run the model at the chosen set.

In this tutorial, we will construct an emulator for each of the model outputs separately. The general structure of a univariate emulator is as follows:

$$f(x) = g(x)\beta + u(x),$$

where $g(x)$ are the regression functions, β the regression parameters, and $u(x)$ is the correlation structure. We split the correlation structure further into two pieces - for two parameter sets x and x', the structure is:

$$u(x) = \sigma^2 \left[(1 - \delta)c(x, x') + \delta I_{\{x=x'\}} \right].$$

Here σ^2 is the (prior) emulator variance and $c(x, x')$ is a correlation function; the simplest such function is squared-exponential

$$c(x, x') = \exp\left(-\frac{\sum_i (x_{i,A} - x'_{i,A})^2}{\theta_i^2} \right).$$

The subscript A indicates that the correlation function operates only on the active parameters for the emulator: that is, parameters that contribute to the regression surface. To ensure the correlation

structure is well-defined, the 'nugget' term $\delta I_{\{x=x'\}}$ is included: this operates on all the parameters in the input space and represents the proportion of the overall variance due to the ensemble variability. The θ_i hyperparameters are the correlation lengths for the emulator. The size of the correlation lengths determine how close two parameter sets must be in order for the corresponding residual values to be highly correlated. A smaller θ_i value means that we believe that the function is less smooth with respect to parameter i, and thus that the values for the corresponding parameters x_i and x'_i must be closer together in order to be highly correlated. The simplifying assumption that all the correlation length parameters are the same, that is $\theta_i = \theta$ for all i, can be made. In such case, the larger θ is, the smoother the local variations of the emulators will be.

4.2   Constructing the emulators step by step

To construct the emulators, two steps are required:

1) We create a set of 'initial' emulators `ems0` by fitting a regression surface to `train0`. These simple emulators will provide us with estimates for the regression surface parameters and the basic correlation specifications;

2) The emulators `ems0` constitute our prior which we adjust to the training data through the Bayes Linear update formulae. In this way we obtain the final version of our first wave emulators: `ems0_adjusted`.

Let us now go through each step in detail.

4.2.1   Step 1

The function `emulator_from_data` creates the initial emulators for us. We pass `emulator_from_data` the training data, the name of the model outputs we want to emulate, the list of parameter ranges and the ensemble variability . Taken this information, `emulator_from_data` finds both the regression parameters and the active parameters for each of the indicated model outputs. Model selection is performed using stepwise addition or deletion (as appropriate), using the AIC criterion to find the minimal best fit.

```
evs <- apply(wave0[10:ncol(wave0)], 2, mean)
ems0 <- emulator_from_data(train0, output_names, ranges, ev=evs)
ems0[[1]]
#> Parameters and ranges:  beta: c(0.2, 0.8); gamma: c(0.2, 1); delta: c(0.1, 0.5); mu: c(0.1, 0
#> Specifications:
#>        Basis Functions:  (Intercept); beta; gamma; delta; mu; I(beta^2); I(delta^2); beta:g
#>        Active variables:  beta; gamma; delta; mu
#>        Beta Expectation:  78.5001; 97.2587; 39.957; -117.2044; 6.0284; 17.1566; 62.8314; 50
#>        Beta Variance (eigenvalues):  0; 0; 0; 0; 0; 0; 0; 0; 0; 0; 0; 0
#> Correlation Structure:
#>        Variance:  3257.59
#>        Expectation:  0
#>        Correlation length:  1
#>        Nugget term:  0.07553078
#> Mixed covariance:  0 0 0 0 0 0 0 0 0 0 0 0
```

Note that we calculated the ensemble variability `evs` taking the mean of the column *EV* in `wave0`. The function `emulator_from_data` uses `evs` to estimate the delta parameter, i.e. the proportion of the overall variance due to the ensemble variability.

Show: More details about the ems0 objects on P**??**

We can plot the emulators to see how they represent the output space: the `emulator_plot` function does this for emulator expectation, variance, standard deviation, and implausibility (more on which later).

```
for (i in 1:length(ems0)) ems0[[i]]$output_name <- output_names[i]
names(ems0) <- output_names
emulator_plot(ems0)
```
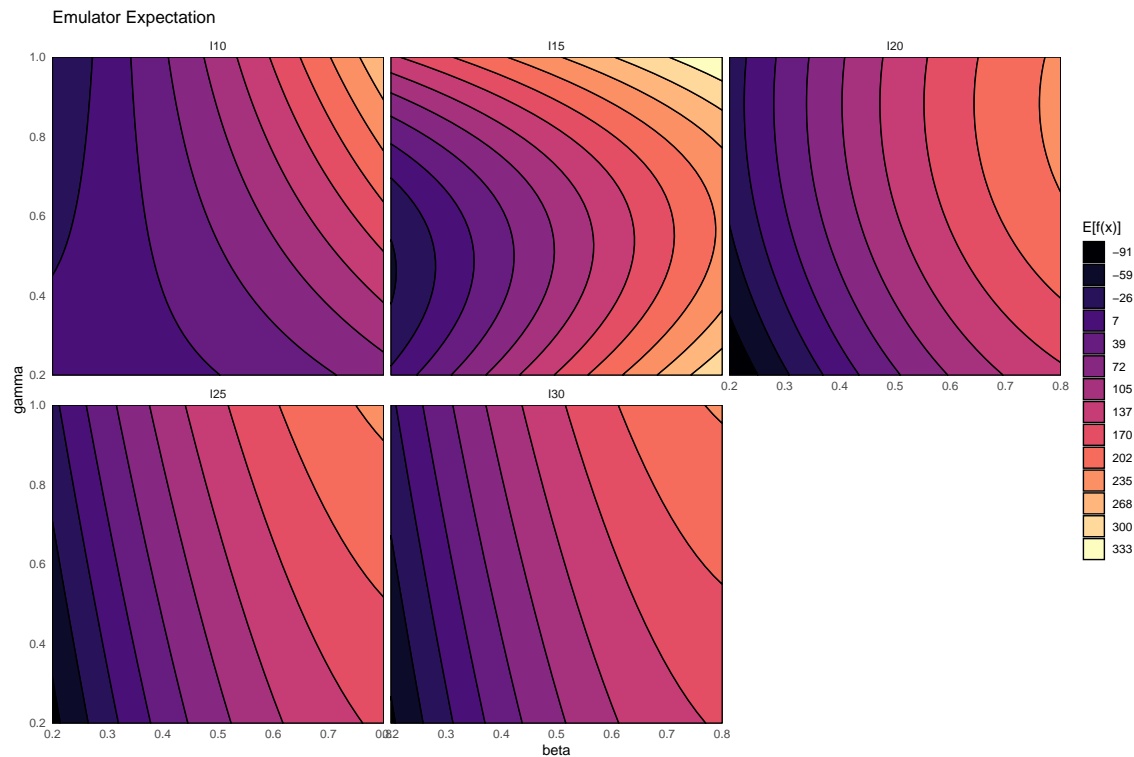

The emulator expectation plots show the structure of the regression surface, which is at most quadratic in its parameters, through a 2D slice of the input space. The two parameters β and γ are selected and for each model output the plot shows the expected value produced by the relative emulator for all possible pairs (β, γ).

To plot the emulators standard deviation we just use `emulator_plot` passing 'sd' as second argument:

```
emulator_plot(ems0, 'sd')
```
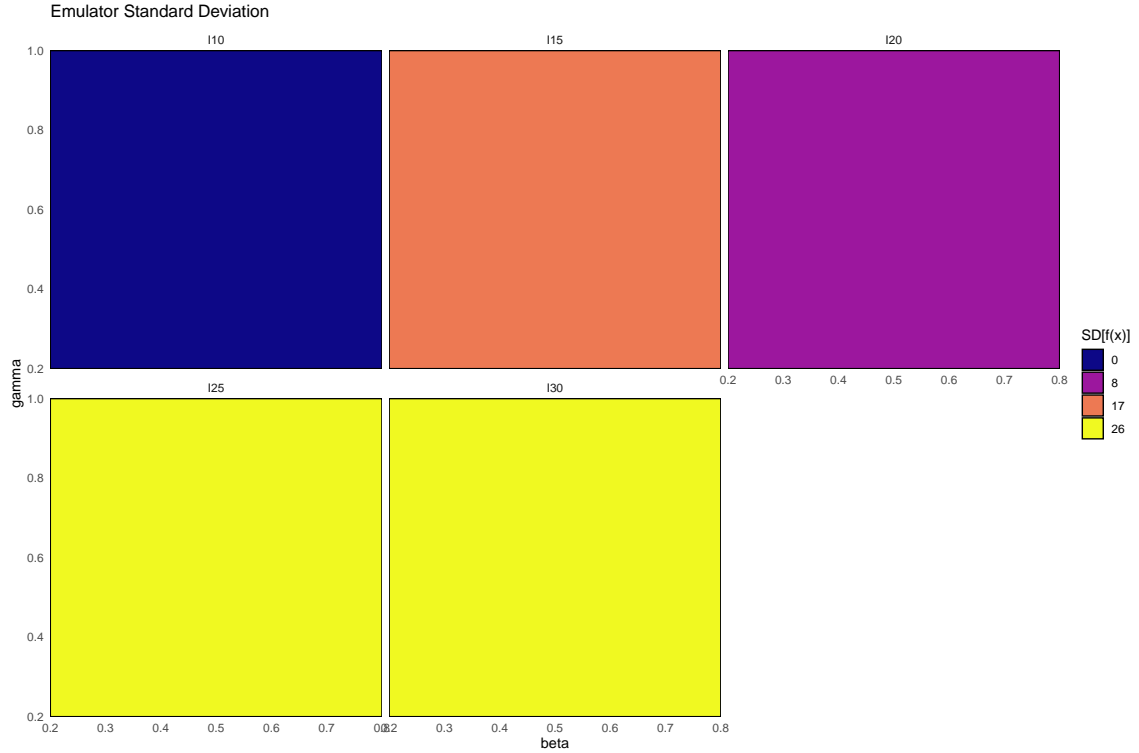

Here we immediately see that the emulator variance (or equivalently, standard deviation) is simply constant across the parameter space for each emulated output. This is not what we want though, since one would expect emulators to be less uncertain around the parameter sets that have been evaluated by the computer model. This will be taken care of in the next step.

4.2.2   Step 2

We now use the `adjust` method on our emulators to obtain the final Bayes Linear version of our `wave0` emulators:

```
ems0_adjusted <- map(seq_along(ems0), ~ems0[[.]]$adjust(train0, output_names[[.]]))
```

Note that the `adjust` method works with the data in `train0` exactly as the function `emulator_from_data` did: it performs Bayes Linear adjustment, given the data. This function creates a new emulator object with the adjusted expectation and variance of beta as the primitive specifications, and supplies the values for the new emulator to compute the adjusted expectation and variance of $u(x)$, and the adjusted $Cov[\beta, u(x)]$. Due to the update formulae, the correlation structure now depends on where in the input space it is evaluated.

```r
names(ems0_adjusted) <- output_names
emulator_plot(ems0_adjusted)
```


Emulator Expectation

```r
emulator_plot(ems0_adjusted, var = 'sd')
```

Emulator Standard Deviation
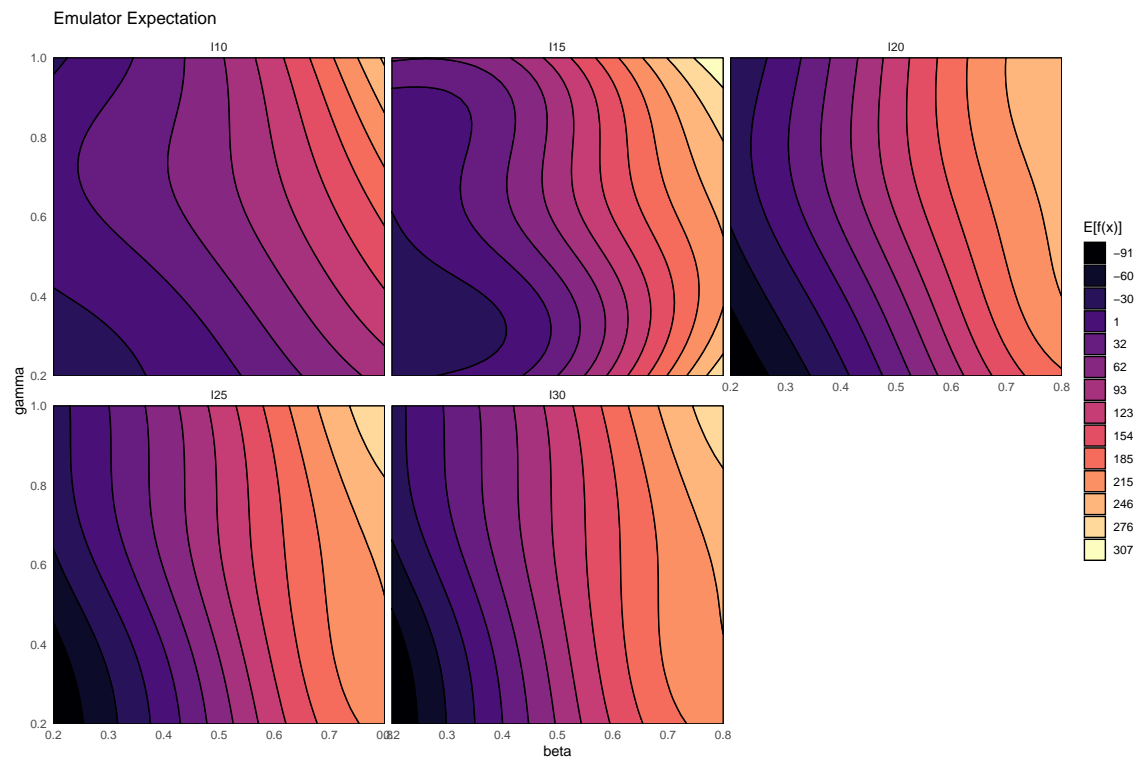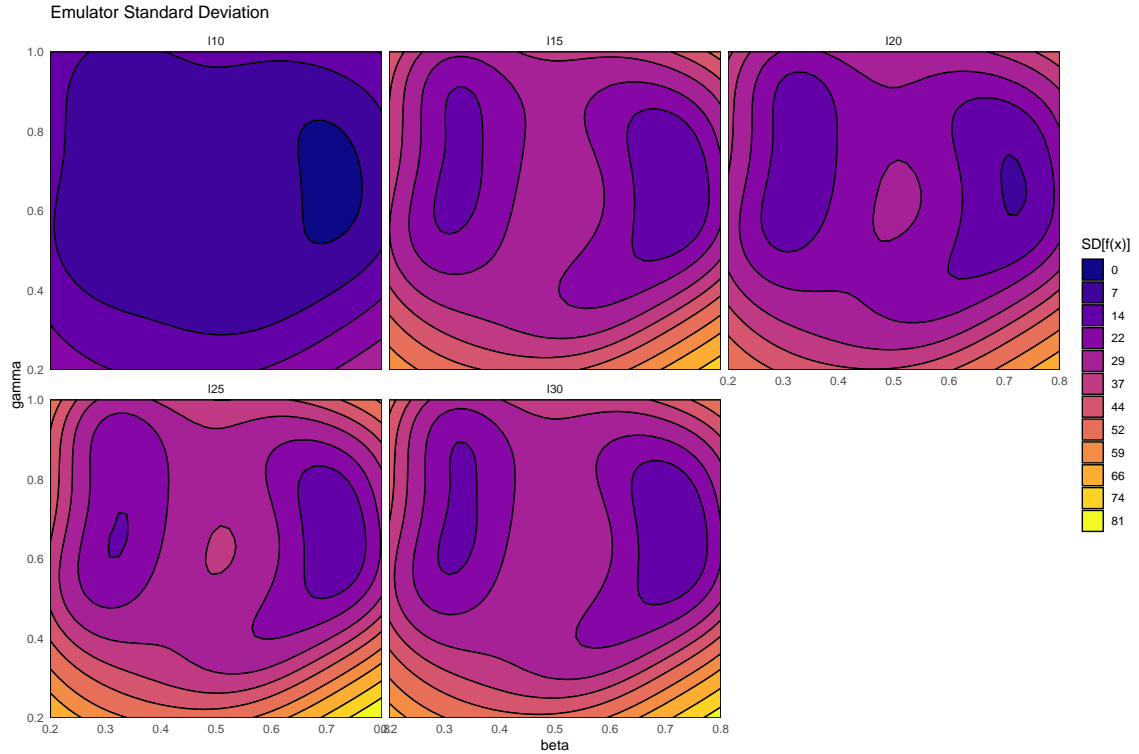
We can see that the adjusted emulators more reasonably show the structure of the model. The
variance has been updated: the closer the evaluation point is to a training point, the lower the
variance (as it 'knows' the value at this point). In fact, evaluating these emulators at parameter sets
in the training data demonstrates this fact:

```
em_evals <- ems0_adjusted$I10$get_exp(train0[,names(ranges)])
all(abs(em_evals - train0$I10) < 10^(-12))
#> [1] TRUE
all(ems0_adjusted$I10$get_cov(train0[,names(ranges)]) < 10^(-12))
#> [1] FALSE
```

Note the comparative speeds of evaluation, here. The initial 80 parameter sets we generated from the
model took around 45 seconds on a relatively powerful laptop; evaluating the emulator expectation
over a 40×40 grid takes less than 5 seconds; evaluating the emulator variance over the same grid
takes 30 seconds.

We now need to consider whether these emulators are actually performing as we would expect them
to. For this, we need to consider emulator diagnostics.

Chapter 5

Emulator diagnostics

For a given set of emulators, we want to assess how accurately they reflect the model outputs over the input space. To do so, we can consider a number of diagnostic tests.

This section is divided in four parts:

1) First we introduce the concept of implausibility;

2) We show how to visualise the emulator's implausibility for a given model output;

3) Three standard emulator diagnostics are introduced together with functions that help us visualise them;

4) We then conclude by analysing parameter sets that fail diagnostics.

5.1   Background: the implausibility measure

For a given model output and a given target, the implausibility is defined as the difference between the emulator output and the target, taking into account all sources of uncertainty; for a parameter set x, the schematic form for the implausibility $I(x)$ is

$$I(x) = \frac{|f(x) - z|}{\sqrt{\sum \sigma^2}},$$

where $f(x)$ is the emulator output, z the target, and the σ terms represent uncertainties.

The uncertainty that goes into the denominator of the emulator implausibility comprises the sigma values in the `targets` list and the emulator variance at the given parameter set. Note that if our targets were not synthetic, we would also include the observation uncertainty, to represent the finite accuracy with which real data is observed, and the model discrepancy, accounting for the fact that no model perfectly represents reality.

An important aspect to consider is the choice of cut-off for the implausibility measure. The implausibility is a metric for evaluating how far out from being a good fit any parameter set is:

there is no hard-and-fast rule for deciding at what point a parameter set is too implausible. Indeed, there are two things to consider when we have multiple univariate emulators.

First of all: what cut-off should we impose? A rough rule of thumb loosely follows Pukelsheim's 3σ rule, which states that any unimodal distribution can be treated normally, in the sense that a 5% confidence interval corresponds to 3σ around the mean. This is only the case for a single such distribution; for multiple univariate emulators it is slightly more involved. However a rough starting cut-off m, for confidence interval $1 - \alpha$ and N emulators, would be

$$m = \Phi^{-1} \left(\frac{1 + (1 - \alpha^{1/N})}{2} \right)$$

where Φ^{-1} is the inverse of the normal distribution CDF.

Second: given multiple emulators, how do we measure overall implausibility? We want a single measure for the implausibility at a given parameter set, but for each emulator we obtain an individual value for I. The simplest way to combine them is to consider maximum implausibility at each parameter set:

$$I_M(x) = \max_{i=1,\dots,N} I_i(x),$$

where $I_i(x)$ is the implausibility at x coming from the ith emulator. For large collections of emulators, it may be useful to instead consider the second-, or third-maximum implausibility. Where some model outputs are deemed more important than others (for instance, putting greater weight on emulation of the peak of an epidemic), we may instead take a weighted average across the implausibity measures.

5.2   Implausibility visualisations

To calculate the implausibility we will use list `targets`, which represents our observations.

The default behaviour of the diagnostics and plots we will see here is to take a cut-off of 3 (following Pukelsheim's 3σ rule), and take maximum implausibility across the emulated outputs. For instance, to find the emulator implausibility for the first output we use the `emulator_plot` function specifying 'imp' for implausibility and passing it the target for the first output:

```
emulator_plot(ems0_adjusted[[1]], 'imp', targets = targets[[1]], cb=TRUE)
```

Emulator Implausibility
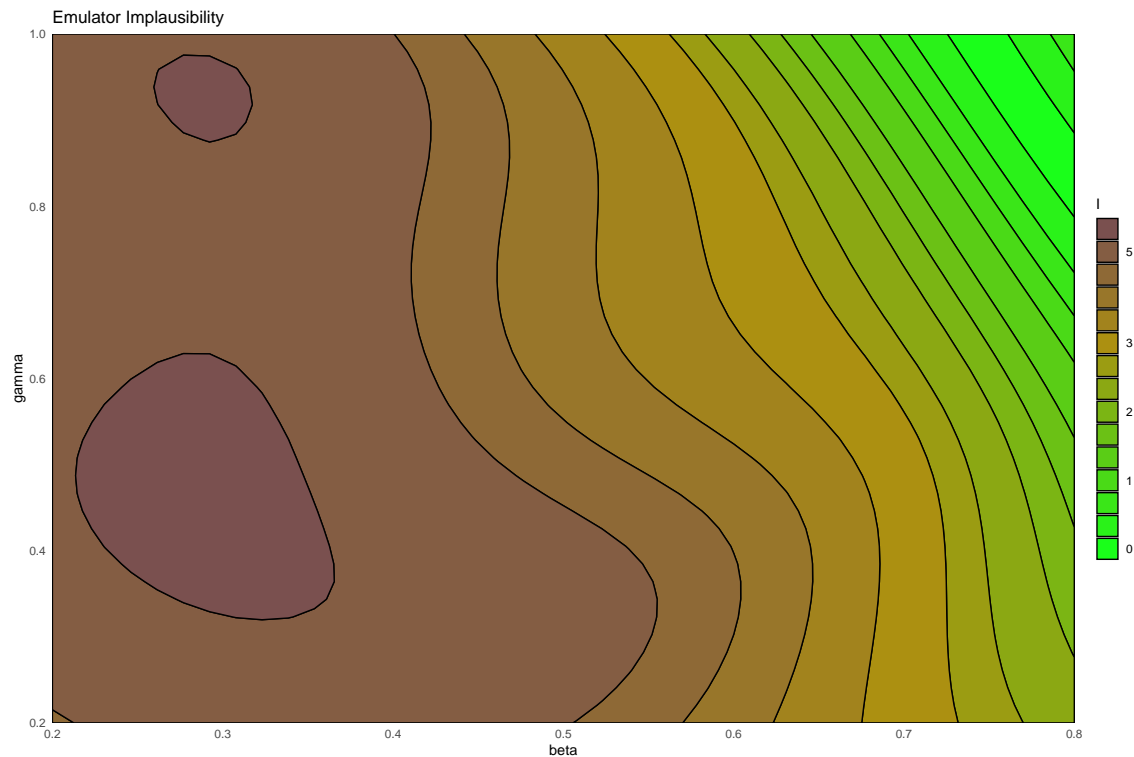
This is a 2D slice through the input space: for a chosen pair $(\bar{\beta}, \bar{\gamma})$, the plot shows the implausibility of the parameter set $(\bar{\beta}, \bar{\gamma}, \delta_M, \mu_M)$, where δ_M denotes the mid-range value of the delta parameter and similarly for μ_M. Parameter sets with a high implausibility (orange region) are highly unlikely to give a good fit and will be discarded when forming the parameters sets for the next wave.

5.3 Three emulator diagnostics

The first three diagnostics are relatively straightforward, and can be presented together. For a given validation set, we can consider the following:

- Within uncertainties, does the emulator output accurately represent the equivalent model output?

- What are the standard errors of the emulator outputs in light of the model outputs?

- Does the emulator adequately classify parameter sets as implausible or non-implausible?

These are encapsulated in the `validation_diagnostics` function.

```
which_invalid <- validation_diagnostics(ems0_adjusted, valid0, output_names, targets = targets)
```
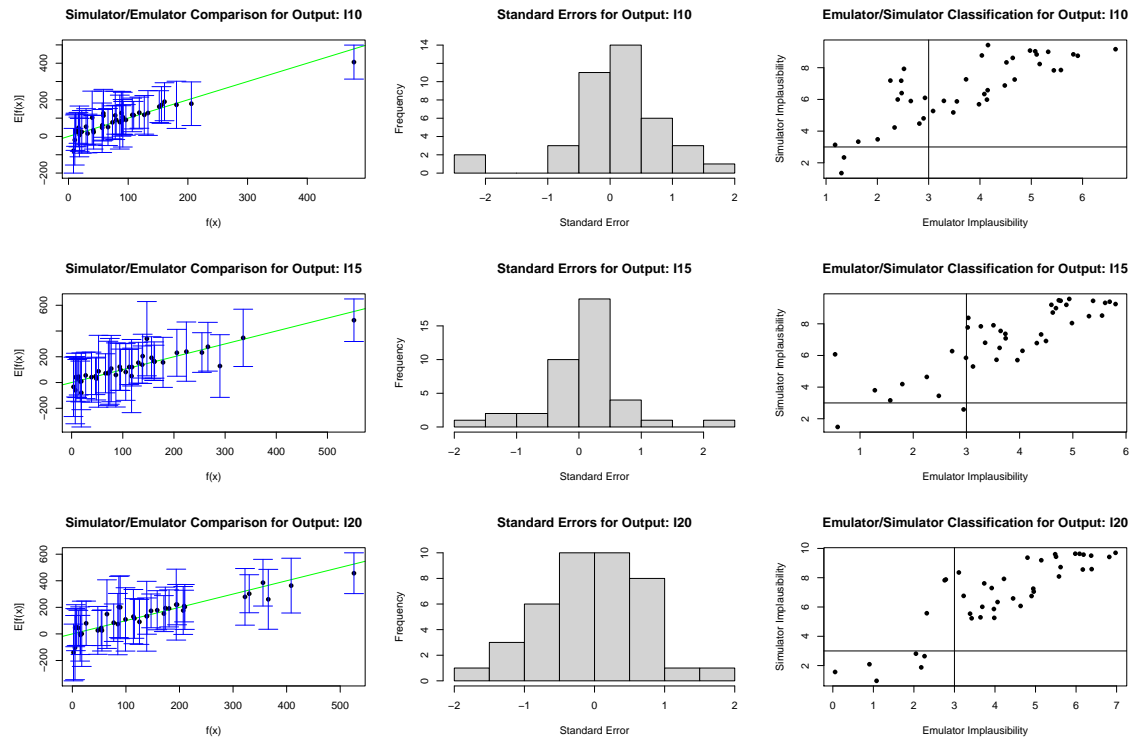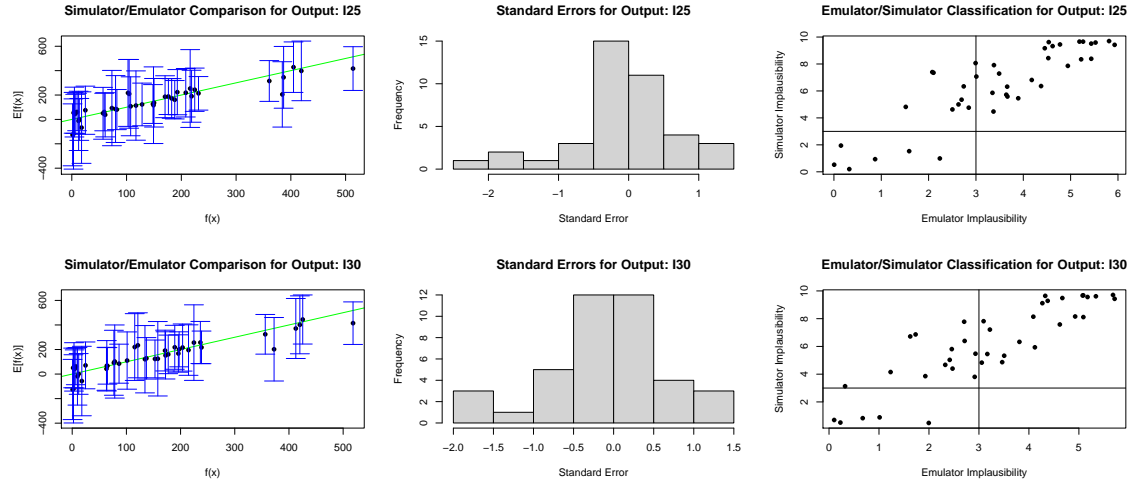
The first column of plots gives an indication of the emulator outputs against the model outputs: the emulator outputs are plotted against the model outputs with a 3σ confidence interval overlaid. An 'ideal' emulator would exactly reproduce the model results: this behaviour is represented by the green line $f(x) = E[f(x)]$. Any parameter set whose emulated prediction lies more than 3σ away from the model output is highlighted in red.

The second column gives the standard errors normalised by the standard deviation. When dealing with large samples, a 95 confidence interval for the mean can be estimated by adding and subtracting $1.96 \ddot{\text{O}}(\text{Standard Error})$ to the sample mean. From this we can deduce the following rule: if an emulator approximates the relative model output well, then the normalised standard errors should fall between -2 and 2 in 95 of the times (where we replaced 1.96 with 2).

Finally, the third column compares the emulator implausibility at the parameter sets to the equivalent model implausibility (i.e. the implausibility calculated replacing the emulator output with the model output). There are three cases to consider:

- The emulator and model both classify a set as implausible/non-implausible: this is fine. Both are giving the same classification for the parameter set.

- The emulator classifies a set as non-implausible, while the model rules it out: this is also fine. The emulator should not be expected to shrink the parameter space as much as the model does, at least not on a single wave. Parameter sets classified in this way will survive this wave, but may be removed on subsequent waves as the emulators grow more accurate on a reduced

parameter space.

- The emulator rules out a set, but the model does not: these are the problem sets, suggesting that the emulator is ruling out parts of the parameter space that it should not be ruling out.

The function `validation_diagnostic`, along with producing the plots, also returns a data.frame consisting of those parameters sets which failed one or more diagnostic tests.

```
which_invalid
#> [1] beta  gamma delta mu
#> <0 rows> (or 0-length row.names)
```

It is often worth considering these parameter sets, particularly if they lie close to the boundary of the space: having a few parameter sets which fail diagnostics is not the end of the world, but we should at least consider whether the emulator is failing in parts of the space we would want it to be performing well on.

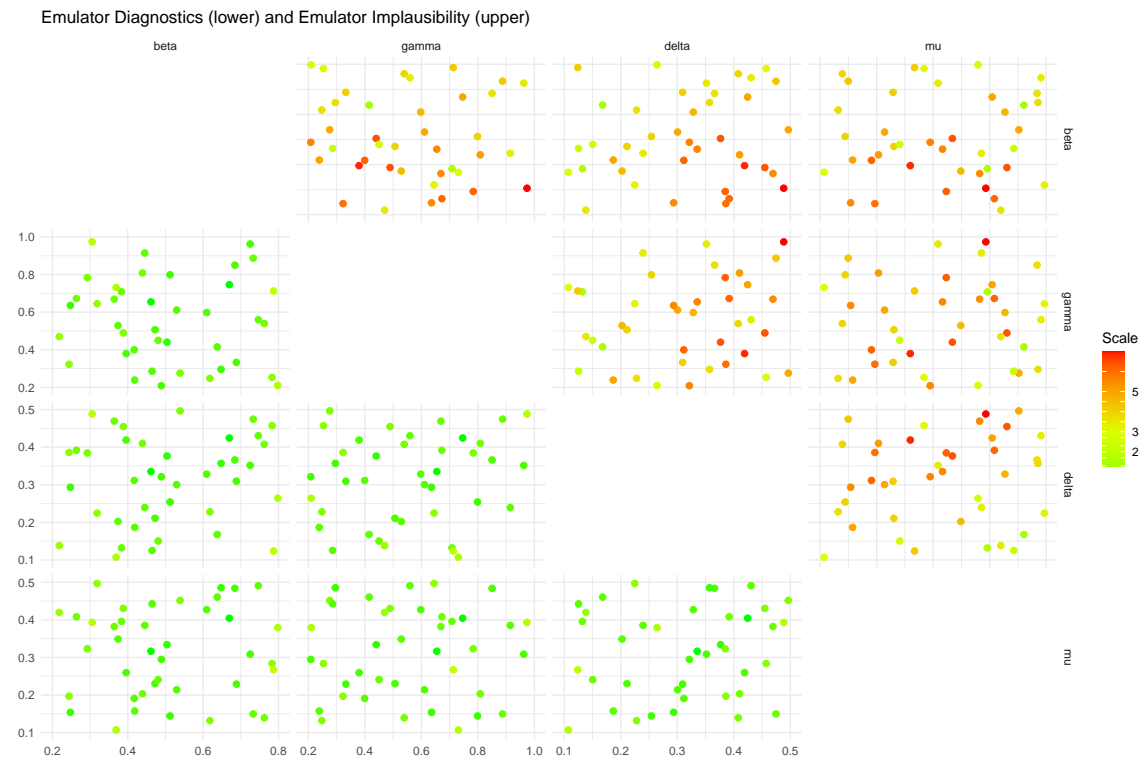## 5.4   Parameter sets failing diagnostics

5.4.1   Visualisation

A helper for visualising problematic parameter sets is provided in the function `validation_pairs`: this gives pairs plots of the parameter sets in the validation data, colouring them by their diagnostic success (bottom left) and predicted implausibility (top right). The diagnostics part gives the maximum standard error at each point: the standard error is

$$\frac{|\text{emulator value} - \text{model value}|}{\sqrt{\text{emulator variance}}}$$

for each emulated output and we maximise over the outputs.

```
vp <- validation_pairs(ems0_adjusted, valid0, targets)
```

Emulator Diagnostics (lower) and Emulator Implausibility (upper)

We can see that the parameter sets that are struggling with diagnostics are indeed on the boundaries of the space, particularly on the boundary of the (δ, μ) space. Examination of the upper half of this plot shows that a large proportion of such parameter sets will be classified as implausible, so they lie in parts of the parameter space that will have no impact on the overall history matching process.

> Show: More options for the validation_pairs function on P??

5.4.2 Space removed function

One way we can get a feel for what cut-off value is reasonable is via the `space_removed` function, which for a given set of emulators will determine how much of the space will be removed by a particular implausibility cut-off. By default, `space_removed` shows the percentage of space that is removed by a specific wave when:

- the ensemble variability is exactly the value provided by the modeller,

- the ensemble variability is 80% (resp. 90%, 110%, 120%) of the value provided by the modeller.

```
sp1 <- space_removed(ems0_adjusted, valid0, targets)
```

Space removed as a function of implausibility cut–off and structural discrepancy
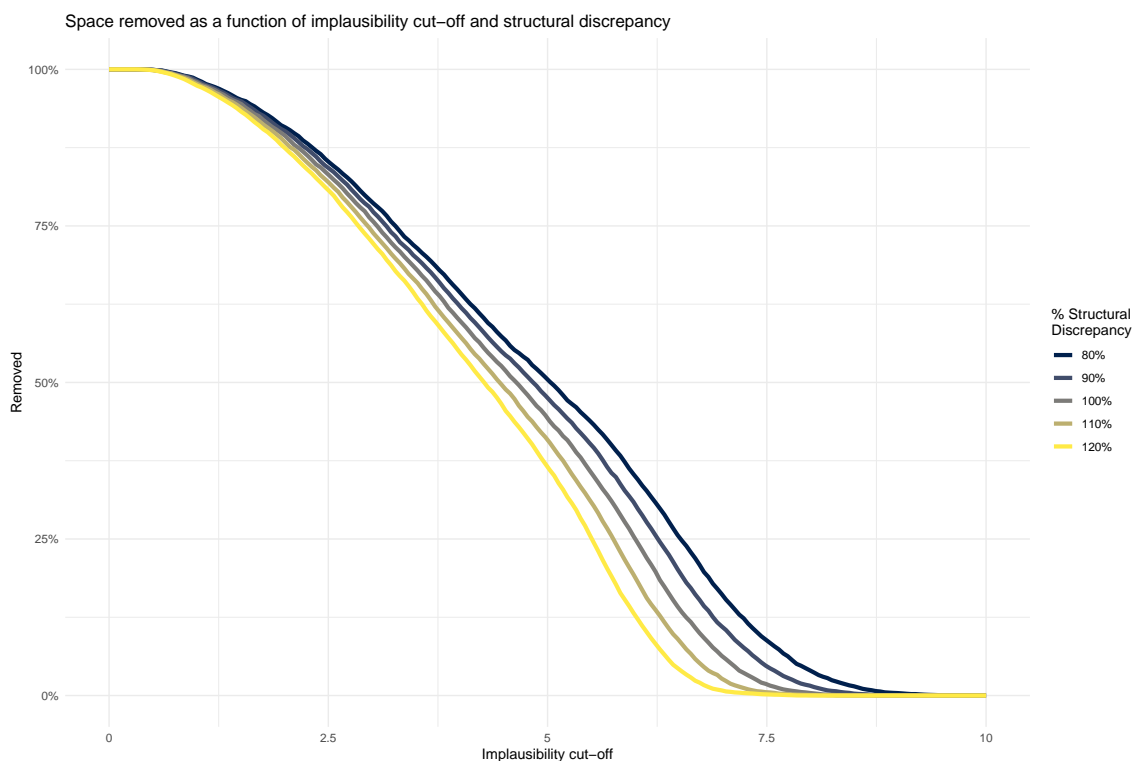
A cut-off of 3 here, using maximum implausibility, would be sufficient to remove around 75% of the current parameter space. This is a reasonable level of removal for a first wave: however, if the expected amount of removal was much lower we could consider whether it is sensible to reduce the cut-off (a companion plot that shows how many diagnostic failures would result from a particular cutoff value is in the pipeline).

Show: More details on the space_removed function on P**??**

The diagnostics here give an indication of the suitability of the emulators in emulating the outputs at this wave. If there are particular model outputs for which the emulators do not give a good fit, then we can modify the specifications for that emulator directly (for example, modifying the correlation length, the variance, or the regression surface) and re-train; if the emulator simply cannot provide a good fit to a model output, we can choose not to emulate this output for the wave in question.

Chapter 6

Point Generation

Having generated emulators based on `wave0` data, evaluated their suitability, and considered a means by which to rule out parameter sets, we can now produce a new set of parameter sets to pass to the model.

This section is divided in two parts:

1) We first see how to generate new sets of parameters (that will be used to train wave 1 emulators);

2) We then compare the performance of the initial parameter sets with the new parameter sets. In other words, we ask: do the model outputs at the new parameter sets match the observations better than the model outputs at the initial parameter sets?

## 6.1 Generating sets of parameters for the next wave

The function `generate_new_runs` is designed to generate new sets of parameters; its default behaviour is as follows.

- If no prior parameter sets are provided, a set is generated using a Latin Hypercube Design, rejecting implausible parameter sets;
- Pairs of parameter sets are selected at random and more sets are sampled from lines connecting them, with particular importance given to those that are close to the non-implausible boundary;
- Using these as seeding points, more parameter sets are generated using importance sampling to attempt to fully cover the non-implausible region.

All of these steps can be overridden or modified, but the default behaviour allows for a good rudimentary search of the non-implausible space.

```
new_points <- generate_new_runs(ems0_adjusted, ranges, n_points = 120, z = targets)
#> [1] "Performing Latin hypercube sampling with rejection..."
#> [1] "Performing line sampling..."
```
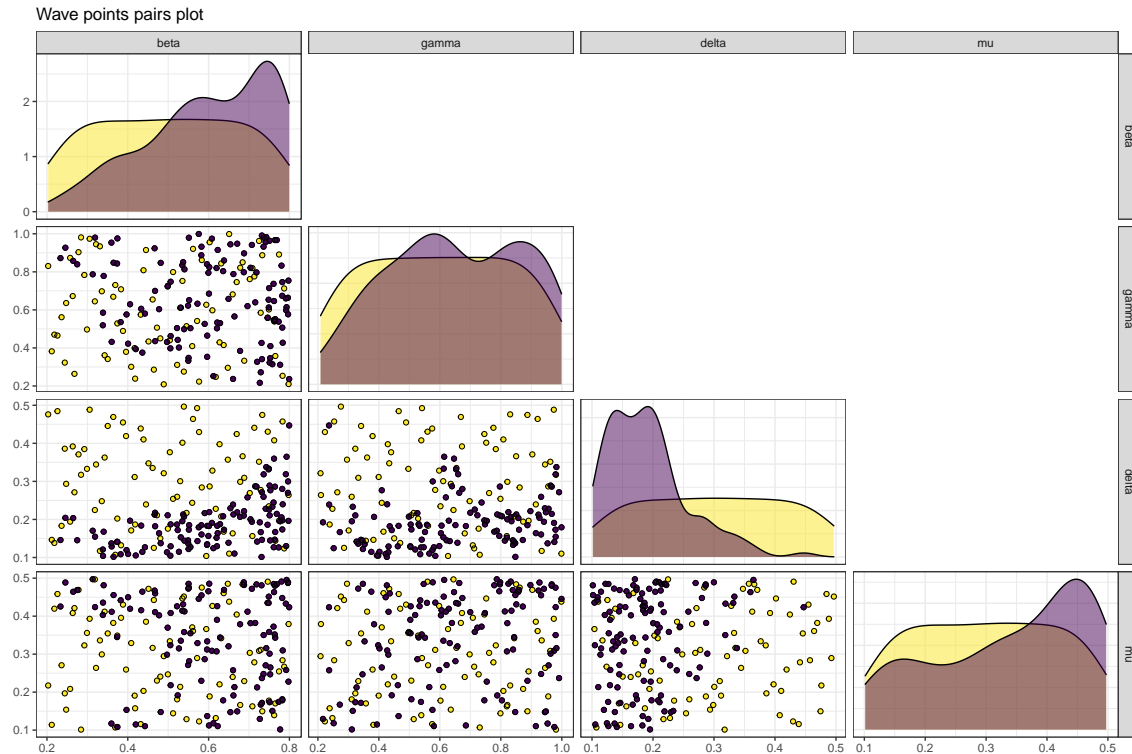
```
#> [1] "Performing importance sampling..."
plot(new_points, pch = 16, cex = 0.5)
```
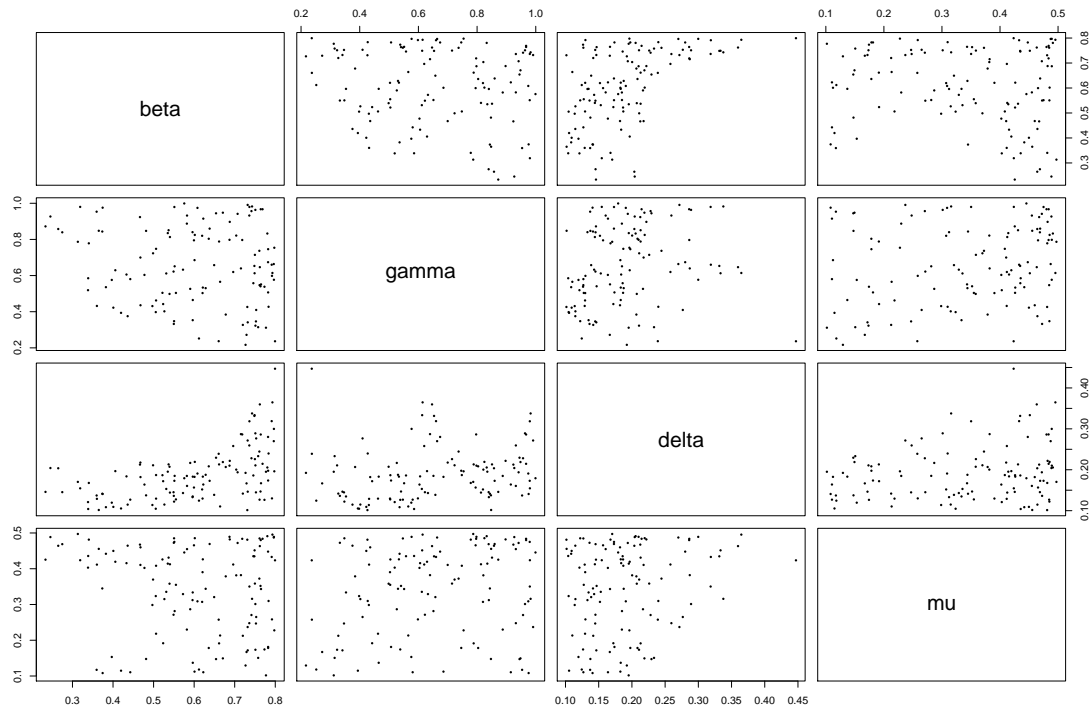

We can start to see the structure of the non-implausible region, here. The `wave_points` function provides a better indication of the difference between the two sets of wave data.

```
wave_points(list(wave0, new_points), in_names = names(ranges))
```

Here `wave0` parameter sets are in yellow and `new_points` (i.e. new parameter sets) are in purple. The plots in the main diagonal show the distribution of parameter sets in `wave0` and that of `new_points`.

6.2   Comparing next and old parameter sets

Now we can put `new_points` into the model and obtain the model outputs:

```
next_wave <- getOutputs(new_points, seq(10,30,by=5))
```

Binding together `new_points` and `next_wave` we obtain `wave1`, the full data for the next wave.

We can see how much better the `wave1` parameter sets perform compared to the original `wave0` parameter sets using `simulator_plot`.

```
wave1 <- data.frame(cbind(new_points,next_wave))%>%
  setNames(c(names(ranges),paste0("I",seq(10,30,by=5)),   paste0("EV",seq(10,30,by=5))))
all_points <- list(wave0[1:9], wave1[1:9])
simulator_plot(all_points, targets)
```

Simulator evaluations at wave points
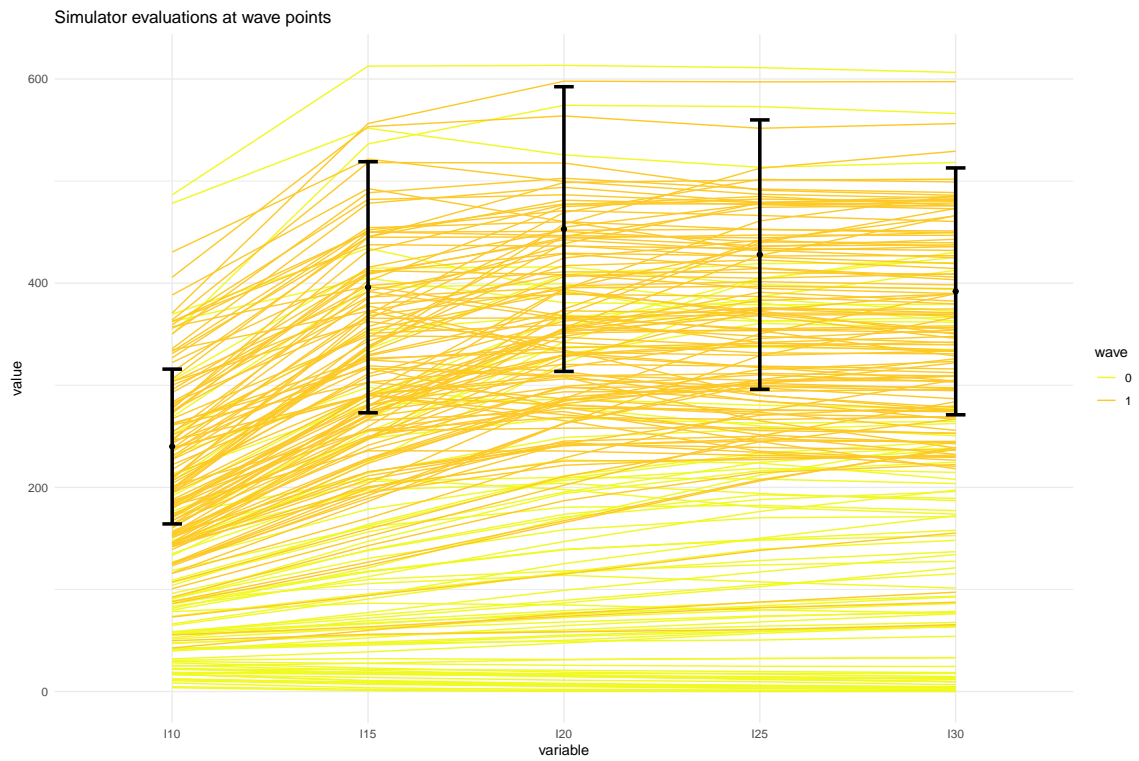
We can see that, compared to the space-filling random parameter sets used to train the first emulators, the new parameter sets are in much closer agreement with our targets. Subsequent waves, trained on these new parameter sets, will be more confident in the new non-implausible region and will therefore refine the region in light of the greater certainty.

Chapter 7

Further waves

We follow the same procedure for subsequent waves, with a couple of caveats.

7.1 Next wave: wave 1

### 7.1.1 Training wave 1 emulators

First of all we train a new set of emulators, in the same way we did for `ems0`:

```r
sampling <- sample(nrow(wave1), 40)
train1 <- wave1[sampling,1:9]
valid1 <- wave1[!seq_along(wave1[,1])%in%sampling,1:9]
new_ranges <- map(names(ranges), ~c(min(wave1[,.]), max(wave1[,.]))) %>% setNames(names(ranges))
ems1 <- emulator_from_data(train1, output_names, new_ranges, quadratic = T)
deltas <- apply(wave1[,10:14], 2, mean)/map_dbl(ems1, ~.$u_sigma)
ems1 <- emulator_from_data(train1, output_names, new_ranges, deltas = deltas, quadratic = TRUE)
for (i in 1:length(ems1)) ems1[[i]]$output_name <- output_names[i]
ems1_adjusted <- map(seq_along(ems1), ~ems1[[.]]$adjust(train1, output_names[[.]]))
names(ems1_adjusted) <- output_names
```

7.1.2 Evaluating implausibility across all waves

We can apply diagnostics to this as before, using `valid1` as the validation set. Assuming the diagnostics are acceptable, we then proceed to consider implausibility - however, we need the implausibility over the whole input space, and the new emulators have only been trained on a subset thereof. We must therefore consider implausibility across all waves, rather than just the wave under consideration at the time.

```r
all_waves <- c(ems0_adjusted, ems1_adjusted)
all_targets <- c(targets, targets)
emulator_plot(all_waves, var = 'maximp', targets = all_targets, cb=TRUE)
```
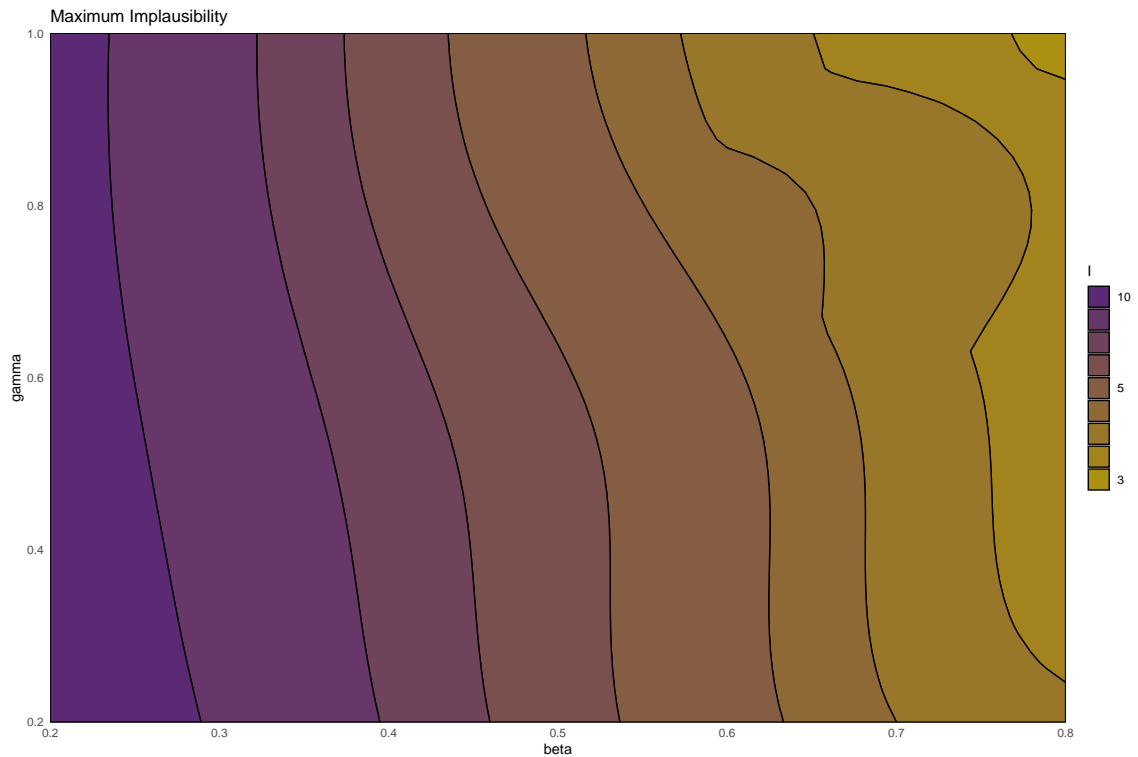
This may seem an unwieldy way to approach this (and it is, at present); however, it is important to remember that the number of emulators at each wave may not be the same; for example, if we have had to remove a model output at wave 1, then the targets would be accordingly changed. In this illustration case, we did not have to worry about doing so since we have assumed that all targets can be emulated.

If we compare the implausibility plot we just obtained with the implausibility plot from the previous wave, we see that the red area has increased significantly: this shows that wave 1 is shrinking down the non-implausible space, exactly as we expected.

The remainder of the analysis proceeds much as in the first wave. In generating new parameter sets, we would of course provide all_waves to the point generation function.

```
new_new_points <- generate_new_runs(all_waves, ranges, n_points = 120, z = all_targets)
#> [1] "Performing Latin hypercube sampling with rejection..."
#> [1] "Performing line sampling..."
#> [1] "Performing importance sampling..."
plot(new_new_points, pch = 16, cex = 0.5)
```
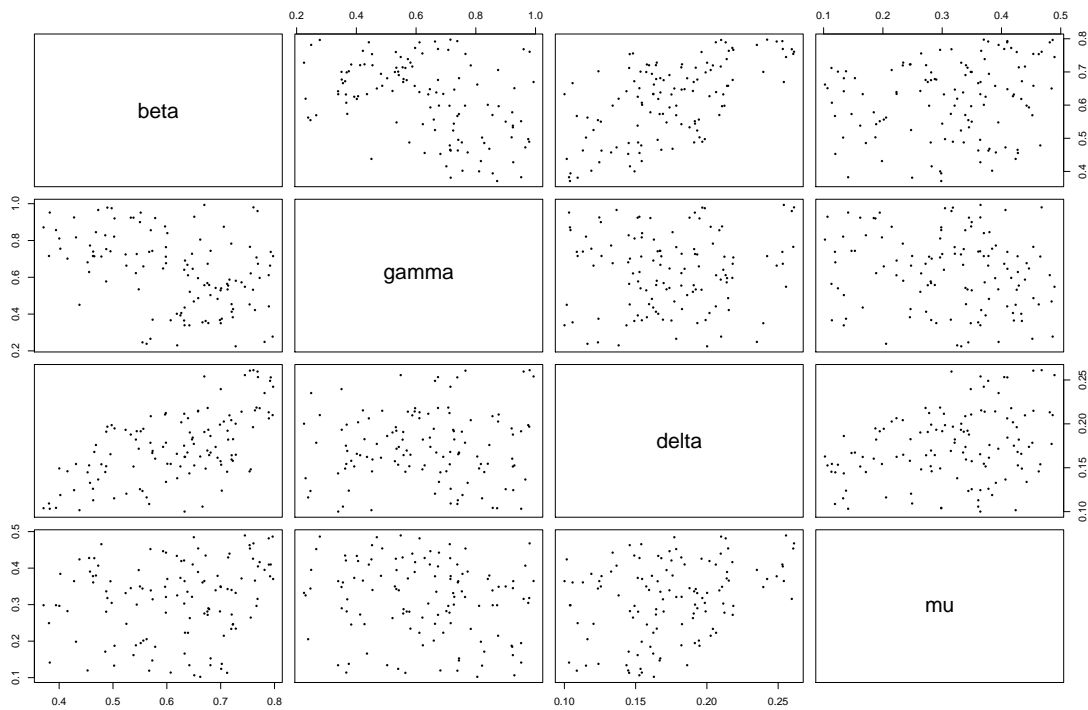
We can compare the distribution of parameter sets at the end of `wave0` with that of parameter sets at the end of `wave1`:

```
wave_points(list(wave1, new_new_points), in_names = names(ranges))
```

Wave points pairs plot
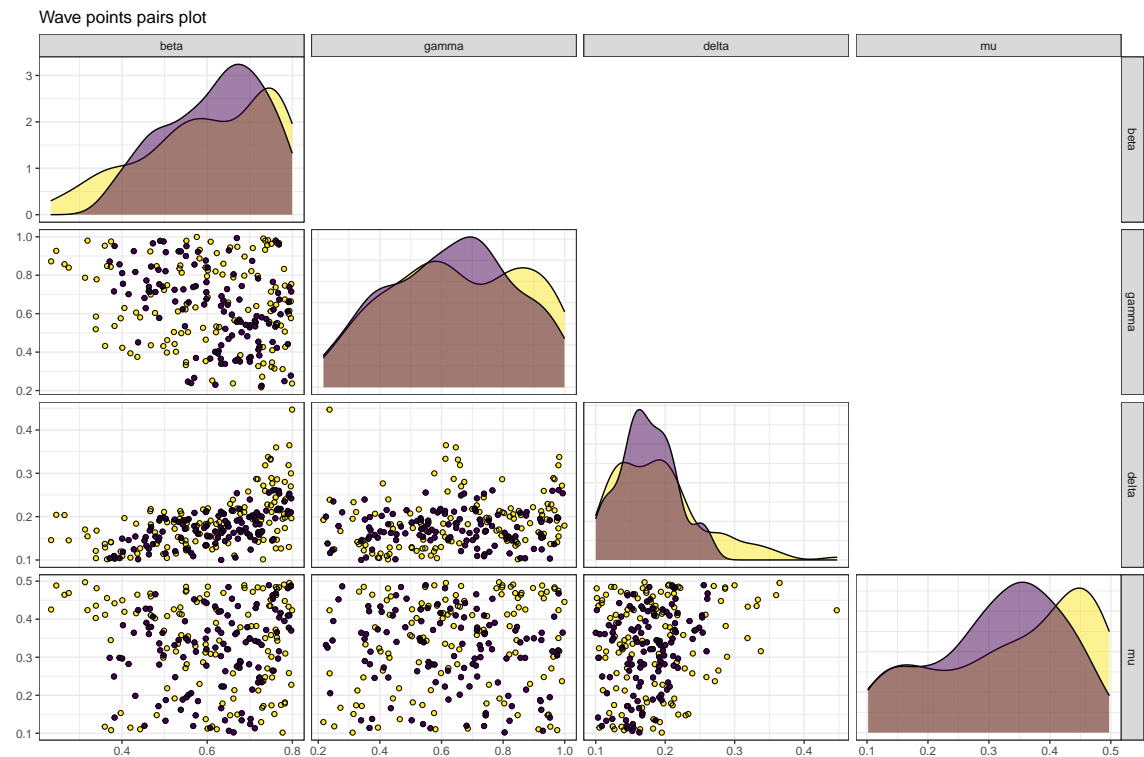
The last step is to create `wave2`, that will be used to train `wave2` emulators.

```r
next_next_wave <- getOutputs(new_new_points, seq(10,30,by=5))
wave2 <- data.frame(cbind(new_new_points,next_next_wave))%>%
  setNames(c(names(ranges),paste0("I",seq(10,30,by=5)),   paste0("EV",seq(10,30,by=5)))))
```

Through the `simulator_plot` function we see how much better the `wave2` parameter sets perform compared to `wave1` and `wave0` parameter sets.

```r
all_points <- list(wave0[1:9], wave1[1:9], wave2[1:9])
simulator_plot(all_points, targets, palette=c("#F9F920", "#FA8816","#C98CFF"))
```
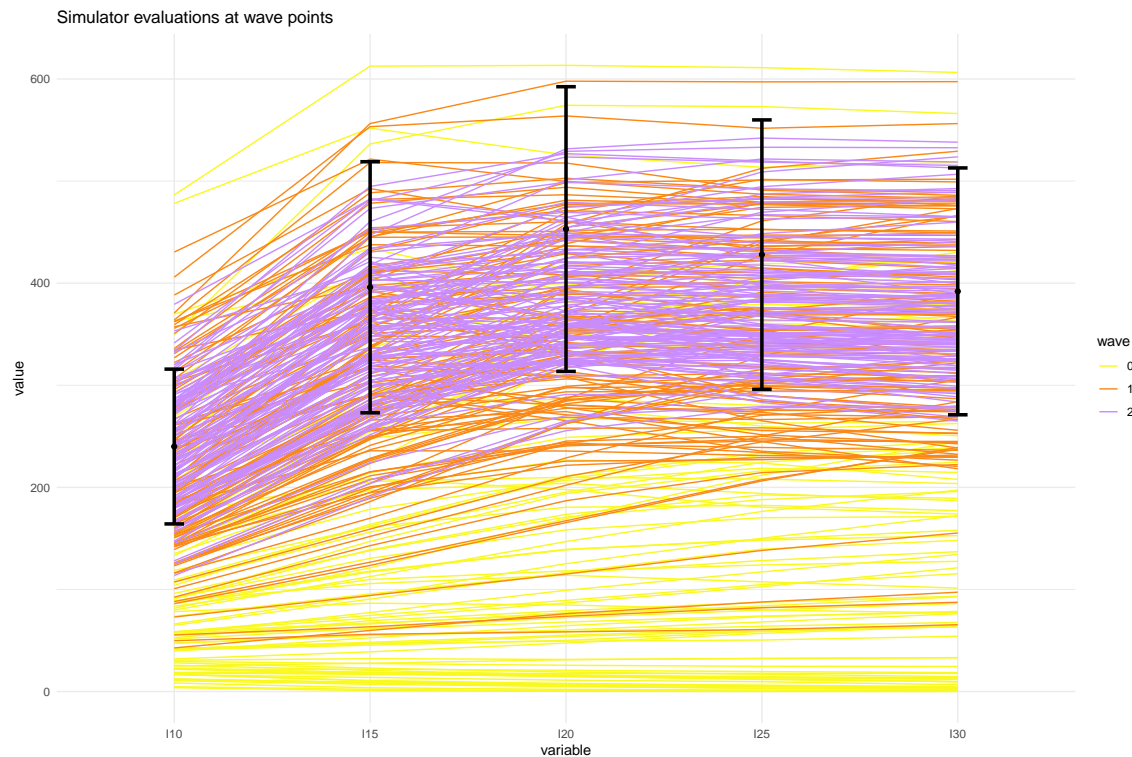
Next waves of the process can be produced simply repeating all the steps in section 7.1.

7.2 Next wave: wave 2

7.2.1 Training wave 2 emulators

```
sampling <- sample(nrow(wave2), 40)
train2 <- wave2[sampling,1:9]
valid2 <- wave2[!seq_along(wave2[,1])%in%sampling,1:9]
new_new_ranges <- map(names(ranges), ~c(min(wave2[,.]), max(wave2[,.]))) %>% setNames(names(ranges))
ems2 <- emulator_from_data(train2, output_names, new_new_ranges, quadratic = T)
deltas <- apply(wave2[,10:14], 2, mean)/map_dbl(ems2, ~.$u_sigma)
ems2 <- emulator_from_data(train2, output_names, new_new_ranges, deltas = deltas, quadratic = TRUE)
for (i in 1:length(ems2)) ems2[[i]]$output_name <- output_names[i]
ems2_adjusted <- map(seq_along(ems2), ~ems2[[.]]$adjust(train2, output_names[[.]]))
names(ems2_adjusted) <- output_names
```

### 7.2.2 Evaluating implausibility across all waves

As before, we need to consider implausibility across all waves, rather than just the wave under consideration at the time.

```
all_waves <- c(ems0_adjusted, ems1_adjusted, ems2_adjusted)
all_targets <- c(targets, targets, targets)
emulator_plot(all_waves, var = 'maximp', targets = all_targets, cb=TRUE)
```
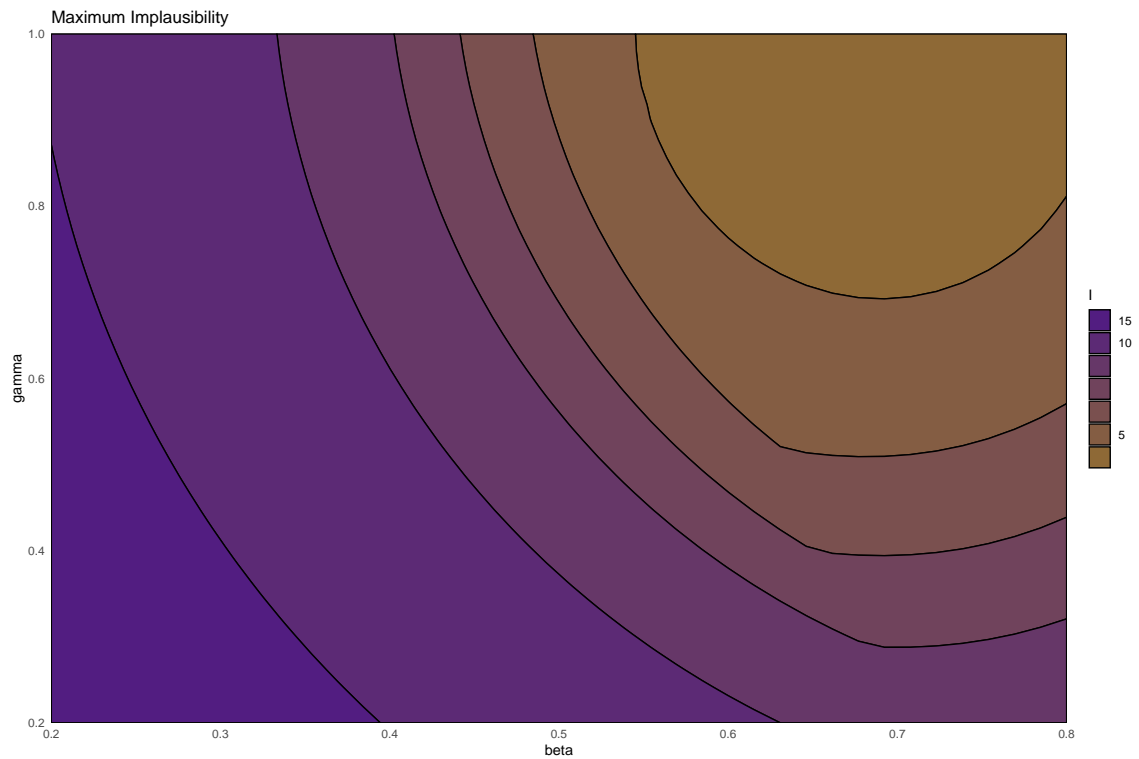

To generate new parameter sets:

```
new_new_new_points <- generate_new_runs(all_waves, ranges, n_points = 120, z = all_targets)
#> [1] "Performing Latin hypercube sampling with rejection..."
#> [1] "Performing line sampling..."
#> [1] "Performing importance sampling..."
plot(new_new_new_points, pch = 16, cex = 0.5)
```
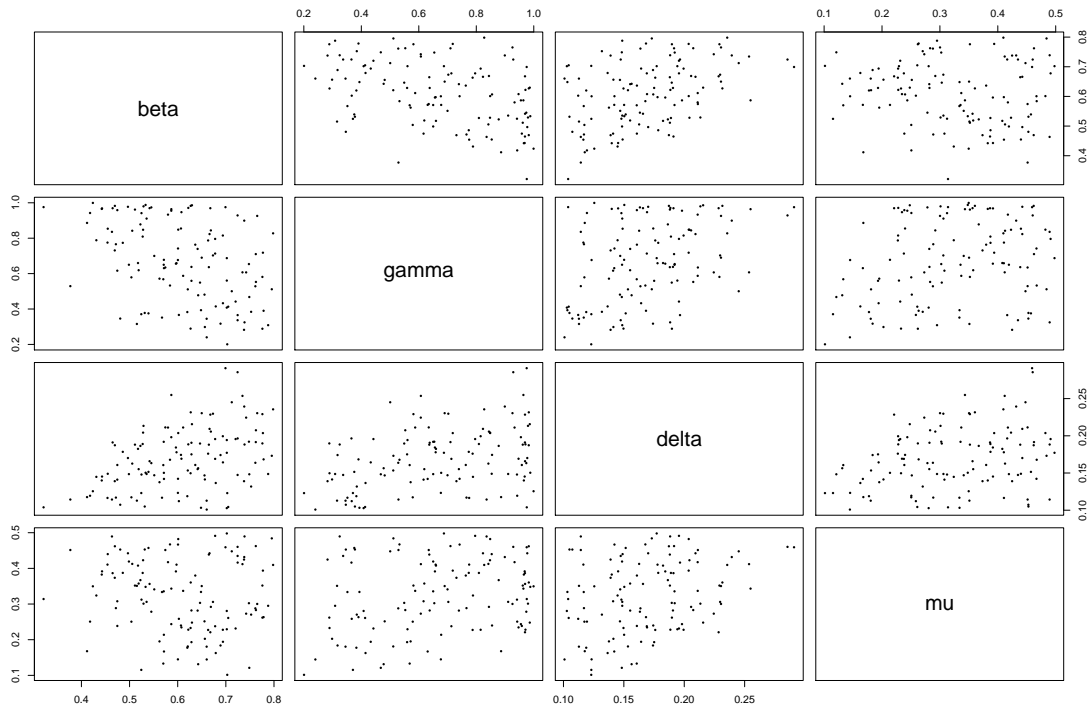
We now create `wave3`:

```
next_next_next_wave <- getOutputs(new_new_new_points, seq(10,30,by=5))
wave3 <- data.frame(cbind(new_new_new_points,next_next_next_wave))%>%
  setNames(c(names(ranges),paste0("I",seq(10,30,by=5)),   paste0("EV",seq(10,30,by=5)))))
```

Through the `simulator_plot` function we check how much better the `wave3` parameter sets perform compared to the original `wave2` parameter sets.

```
all_points <- list(wave2[1:9], wave3[1:9])
simulator_plot(all_points, targets, zero_in=FALSE)
```

Simulator evaluations at wave points
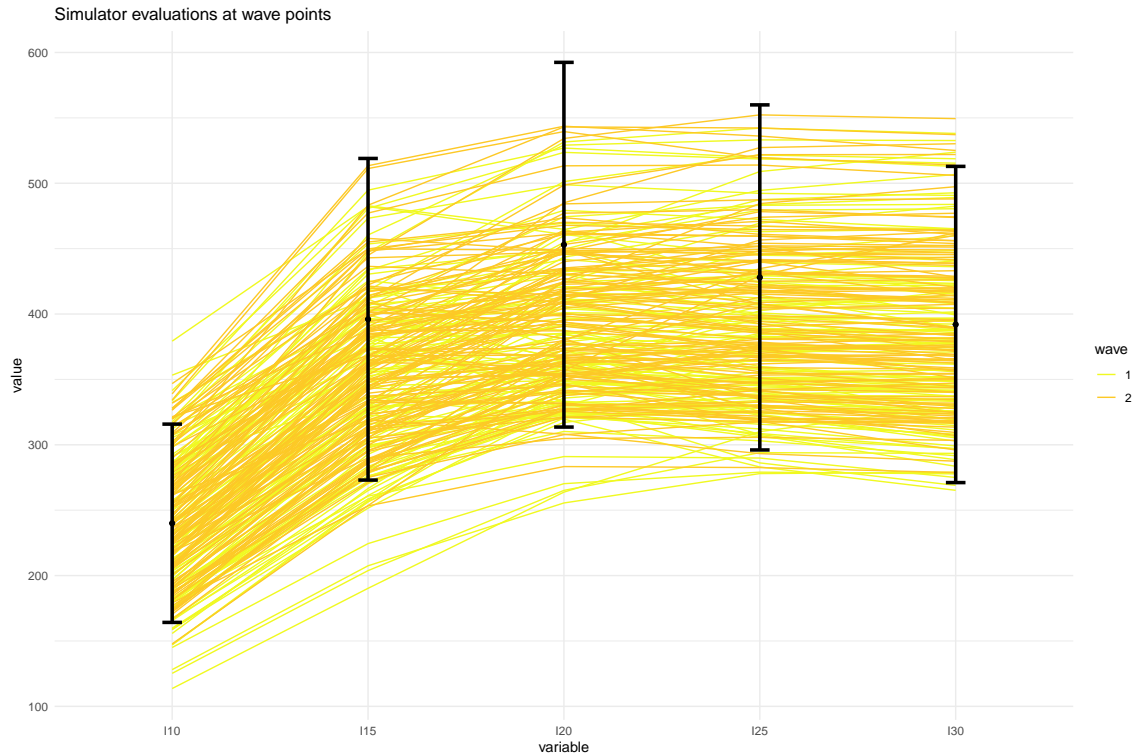
The graph does not show a clear improvement in the performance of `wave3` parameter sets compared to that of `wave2` parameter sets. To understand the reason behind this, let us compare the variability of the model outputs we are emulating with the emulators uncertainty. Below we show the ensemble variability and the uncertainty for `ems0` and `ems2` for each of the emulated outputs.

```
targets$I10$sigma
#> [1] 25.27
targets$I15$sigma
#> [1] 40.99
targets$I20$sigma
#> [1] 46.48
targets$I25$sigma
#> [1] 43.98
targets$I30$sigma
#> [1] 40.3
sigmas0 <- map_dbl(ems0, ~.$u_sigma)
sigmas0
#>      I10      I15      I20      I25      I30
#>  57.0753 112.2360 103.7517 124.6955 120.2566
sigmas2 <- map_dbl(ems2, ~.$u_sigma)
sigmas2
#>      I10      I15      I20      I25      I30
```

```
#> 31.86318 15.87364 28.85938 30.41292 23.41641
```

We see that while `ems0` uncertainties (`sigmas0`) are larger than the ensemble variabilities, `ems2` uncertainties (`sigmas2`) are similar or smaller than the ensemble variabilities. Since the emulators variance is smaller than the uncertainty inherent to the model, the non-implausible space already contains acceptable matches and is unlikely to decrease in size in the next iteration. For this reason, we conclude here the iterating process and therefore the tutorial. Note that if we could revise the uncertainties present in our model and decrease them, we would then be able to perform other waves of the history matching process and shrink the non-implausible space down further.

Chapter 8

Glossary

Beliefs: In Bayesian statistics, probability expresses a degree of belief in an event. Such belief can be based either on prior knowledge or on personal beliefs about the event.

Correlation lengths: The correlation lengths are hyperparameters that appear in the structure of the emulators. They determine how close two parameter sets must be in order for the corresponding residual values to be highly correlated. Large values for the correlation lenghts are chosen if the model is believed to be a smooth function of the parameters.

Emulator output: The data produced by executing an emulator. In each wave of the history matching process, a subset of the model outputs is selected to be emulated.

Ensemble variability: The variability resulting from the stochasticity of the model.

Implausibility: A measure which evaluates the distance between the targets and the model output/emulator output at any given parameter set.

Input space: The set of all possible combinations of parameters.

Model output: Any data produced by executing a model. An example of model output in this case study is the number of infectious individuals at time 10 (or $15, 20, 25, 30$).

Observed data: The data we fit our model to, which usually comes from empirical observations. Since in this case study we work with a synthetic dataset, we prefer to use the word 'targets' rather than 'observed data'.

Parameter set: An element of the input space.

Points: Another word for 'parameter set'.

SEIRS model: A model consisting of four compartments: Susceptible individuals (S), Exposed individuals (E) i.e. people that are infected but not infectious yet, Infectious individuals (I) and Recovered individuals (R). In the model four transitions are allowed: S to E, when a susceptible individual becomes infected, E to I, when an infected individual becomes infectious, I to R, when an infectious individual recovers, and R to S, when a recovered individual becomes susceptible again.

Targets: A list of pairs of the form (val, sigma), one per emulated output, used to evaluate implausibility. The 'val' component represents the mean value of the output and 'sigma' represents our uncertainty about it.

Training data: The data used to train the emulators. It is obtained by running the model at a given number of parameter sets.

Validation data: The data used to validate the emulators. It is obtained by running the model at a given number of parameter sets (different from those used for the training data).

Wave: An iteration of the history matching process.

Appendix A

Bayes Linear Emulation

In the emulatorr package we adopt a Bayes Linear approach to build emulators. While a full Bayesian analysis requires specification of a full joint prior probability distribution to reflect beliefs about uncertain quantities, in the Bayes linear approach expectations are taken as a primitive and only first and second order specifications are needed when defining the prior. Operationally, this means that one just sets prior mean vectors and covariance matrices for the uncertain quantities, without having to decide exactly which distribution is responsible for the chosen mean and covariance. A Bayes Linear analysis may therefore be viewed as a pragmatic approach to a full Bayesian analysis, where the task of specifying beliefs has been simplified. As in any Bayesian approach, our priors (mean vectors and covariance matrices) are then adjusted to the observed data.

The Bayes linear approach to statistical inference takes expectation as primitive. Suppose that there are two collections of random quantities, $B = (B_1, \ldots, B_r)$ and $D = (1, D_1, \ldots D_s)$. Bayes linear analysis involves updating subjective beliefs about B given observation of D. In order to do so, prior mean vectors and covariance matrices for B and D (that is $E[B], E[D], Var[B]$ and $Var[D]$), along with a covariance matrix between B and D (that is $Cov[B, D]$), must be specified.

The Bayes linear update formulae for a vector B given a vector D are:

$$E_D[B] = E[B] + Cov[B, D]Var[D]^{-1}(D - E[D]) \tag{A.1}$$
$$Var_D[B] = Var[B] - Cov[B, D]Var[D]^{-1}Cov[D, B] \tag{A.2}$$
$$Cov_D[B_1, B_2] = Cov[B_1, B_2] - Cov[B_1, D]Var[D]^{-1}Cov[D, B_2]. \tag{A.3}$$

$E_D[B]$ and $Var_D[B]$ are termed the adjusted expectation and variance of B given D. $Cov_D[B_1, B_2]$ is termed the adjusted covariance of B_1 and B_2 given D, where B_1 and B_2 are subcollections of B. The formula given for $E_D(B)$ represents the best linear fit for B given D in terms of minimising the expected squared loss functions $E[(B_k - a_k^T D)^2]$ over choices of a_k for each quantity in $B; k = 1, \ldots, r$, that is, the linear combination of D most informative for B.

\end{comment}