Machine Learning
SS 2020
Exercise sheet 2

Solution by
Lorenzo Minneci, Daniel Strenger
11939539, 01531211

May 12, 2020

Exercise 1. (Online Bayesian Linear Regression)
Solution. 1.

$$p(t|\boldsymbol{x},\boldsymbol{w}) = \prod_{n=1}^{N} \sqrt{\frac{\beta}{2\pi}} \exp\left(-\beta \frac{(t-\boldsymbol{w}^T\boldsymbol{\phi}(x_n))^2}{2}\right)$$

$$\log(p(\boldsymbol{t}|\boldsymbol{x},\boldsymbol{w})) = \sum_{n=1}^{N} \frac{1}{2}(\ln\beta - \ln 2\pi) + [-\frac{\beta}{2}(t-\boldsymbol{w}^T\boldsymbol{\phi}(x_n))^2] = \frac{N}{2}(\ln\beta - \ln 2\pi) - \beta \underbrace{\frac{1}{2}\sum_{n=1}^{N}(t-\boldsymbol{w}^T\boldsymbol{\phi}(x_n))^2}_{E_D} =$$

$$\frac{N}{2}(\ln\beta - \ln 2\pi) - \beta E_D$$

Solution. 2.
It's a multivariate Gaussian, and again we are going to compute the logarithm. So:

$$\begin{aligned}\ln(p(\boldsymbol{w})) = \ln(\mathcal{N}(\boldsymbol{w}|\boldsymbol{0},\alpha^{-1}I)) = \ln\left(\frac{1}{(2\pi)^{\frac{M+1}{2}}|\alpha^{-1}I|}\exp\left((-\frac{1}{2}\boldsymbol{w}^T(\alpha^{-1}I)\boldsymbol{w})\right)\right) = \\ -\frac{M+1}{2}\ln(2\pi) - \frac{1}{2}\ln(|\alpha^{-1}I|) - \left(\frac{1}{2}\boldsymbol{w}^T(\alpha^{-1}I)\boldsymbol{w}\right) = \\ -\frac{M+1}{2}\ln(2\pi) + \frac{M+1}{2}\ln(\alpha) - \frac{\alpha}{2}\boldsymbol{w}^{\boldsymbol{T}}\boldsymbol{w}\end{aligned}$$

In this setting we notice that only the third term is dependent on w, while the first two terms are constant. We'll use this consideration for the next exercise.

Solution. 3.
Bayes theorem states that:

$$p(\boldsymbol{w}|\boldsymbol{t}) = \frac{p(\boldsymbol{t}|\boldsymbol{w})p(\boldsymbol{w})}{p(\boldsymbol{t})}$$

We assume that $p(\boldsymbol{t})$ is not depending on w, so also its logarithm will be independent.
Let's compute the logarithm of the posterior simply as:

$$\ln(p(\boldsymbol{t}|\boldsymbol{w})) + \ln(p(\boldsymbol{w})) + cost = \underbrace{\frac{N}{2}\left(\ln\beta - \ln(2\pi)\right)}_{\text{cost}} - \frac{\beta}{2}\sum_{n=1}^{N}(t - \boldsymbol{w}^T\boldsymbol{\phi}(x_n))^2 +$$

$$+ \left[\underbrace{-\frac{M+1}{2}\ln(2\pi) - \frac{M+1}{2}\ln(\alpha)}_{\text{cost}} + \left(-\frac{\alpha}{2}\boldsymbol{w^T w}\right)\right] =$$

concluding:

$$\ln(p(\boldsymbol{w}|\boldsymbol{t})) = -\frac{\beta}{2}\sum_{n=1}^{N}(t - \boldsymbol{w}^T\boldsymbol{\phi}(x_n))^2 - \frac{\alpha}{2}\boldsymbol{w}^T\boldsymbol{w} + cost$$

where we grouped in *cost* all the terms constant with respect to w. The maximization of this distribution w.r.t. $\boldsymbol{w}$ is equivalent to the minimization of the sum-of-squares approach with addition of a quadratic regularization term $\lambda = \frac{\alpha}{\beta}$.

Solution. 4.
Let's now maximise the log-posterior w.r.t. w. We will express the elements t and $\phi(x_i)$ as part of $\boldsymbol{t}$ and $\boldsymbol{\Phi}$

$$\frac{\partial}{\partial w_i}\left[\frac{\beta}{2}\sum_{n=1}^{N}(t - \boldsymbol{w}^T\boldsymbol{\phi}(x_n))^2 - \frac{\alpha}{2}\boldsymbol{w}^T\boldsymbol{w} + cost\right] = \frac{\beta}{2}(2\Phi_i^T\Phi\boldsymbol{w} - 2\Phi_i^T\boldsymbol{t}) + \frac{\alpha}{2}2\boldsymbol{w} =$$

$$= \left(\beta\Phi^T\Phi + \alpha I\right)\boldsymbol{w} - \beta\Phi_i^T\boldsymbol{t} \stackrel{!}{=} 0$$

$$\Rightarrow \boldsymbol{w}_{map} = (\Phi^T\Phi + \frac{\alpha}{\beta}I)^{-1}\Phi\boldsymbol{t}$$

That we can see it's equal to $\boldsymbol{m} = \beta\boldsymbol{S}\boldsymbol{\Phi^T}\boldsymbol{t}$, being $\boldsymbol{S} = \left(\alpha\boldsymbol{I} + \beta\boldsymbol{\Phi^T\Phi}\right)^{-1}$.
We took β out of $\boldsymbol{S}$, which is finally a scalar multiplier in $\boldsymbol{m}$.

Solution. 5, 6, 7, 8.
In Fig. 1 $K = 33$ was chosen, $\sigma_r^2 = 6$, $\alpha = 0.0005$, $\beta = 0.0009$. We can see here that as we increase σ_r^2 the curve gets slightly smoother while interpolating between the data points. On the other hand, given the same parameters, if we decrease σ_r^2 (e.g. $\sigma_r^2 = 0.1$) the curve present peaks, since we're assigning a high likelihood to each data sample. We're then facing overfitting, and our function learns very particular behaviours from our data and try to fits it in a over complicated model. In this case, each data sample influences very slightly its neighbours. We can also notice how abruptly our prediction for $x > K$ flattens on the mean value.

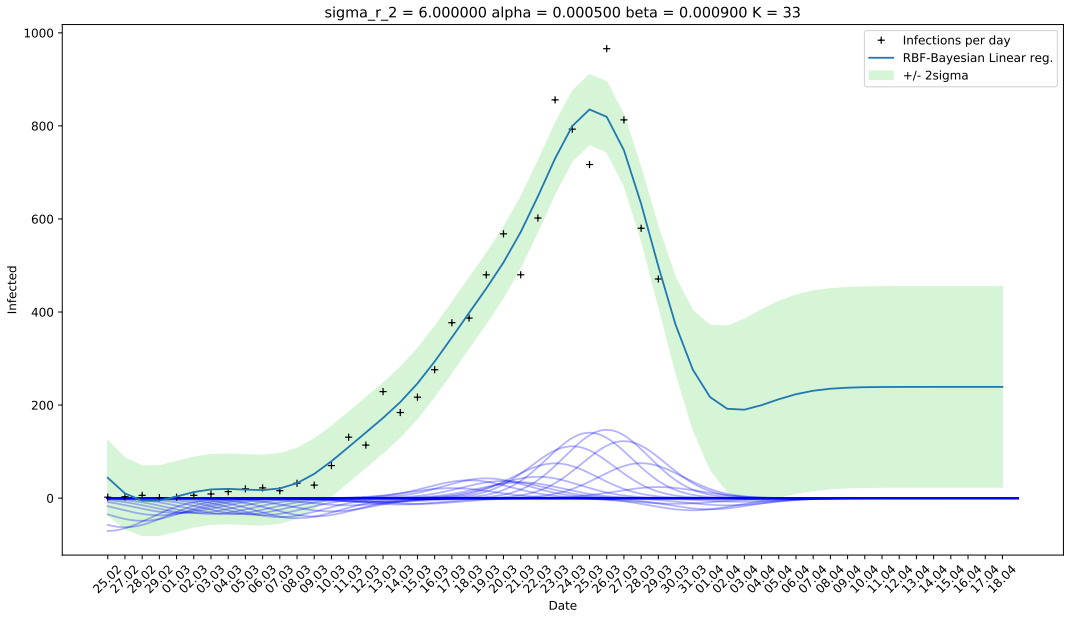

Figure 1: $\sigma_r^2 = 6$

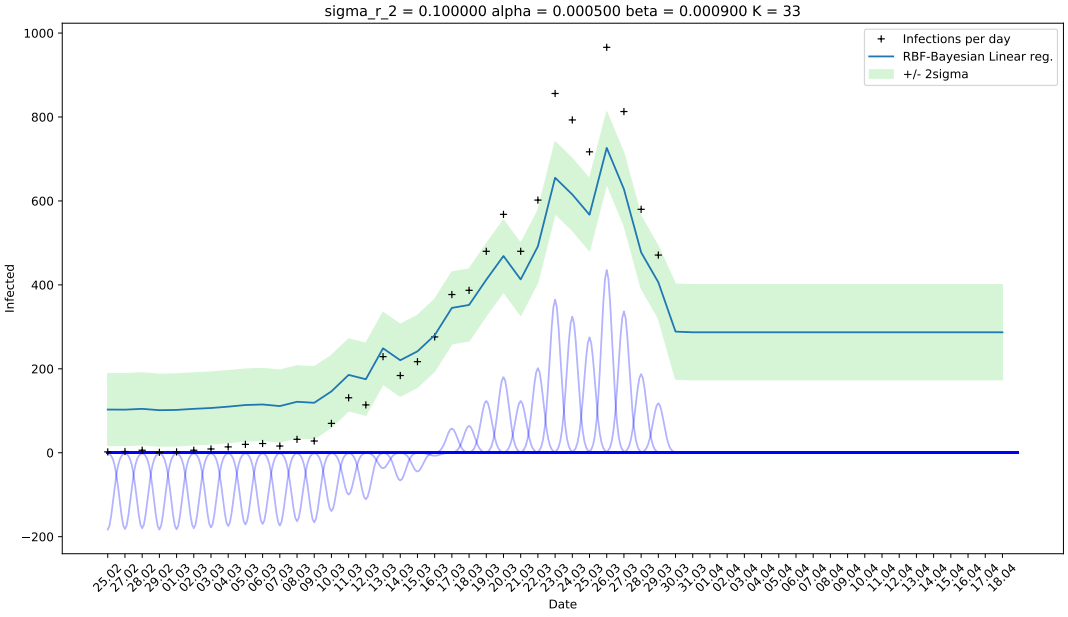

Figure 2: $\sigma_r^2 = 0.1$

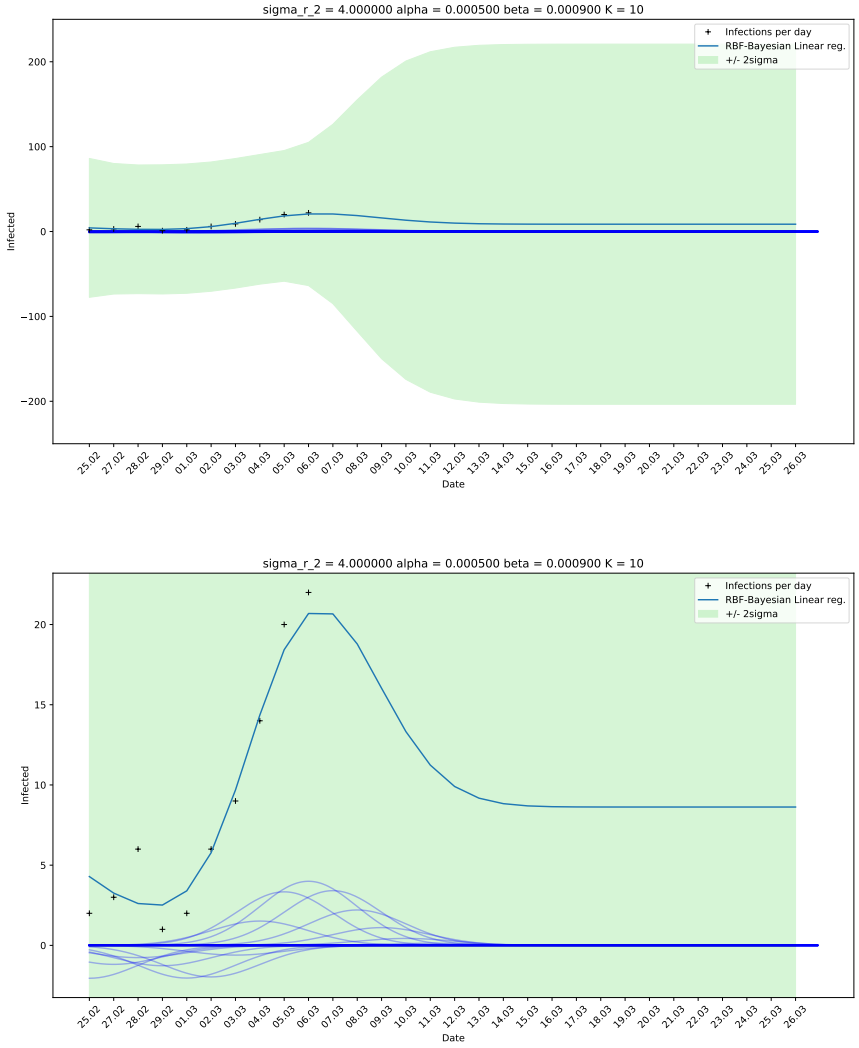

Figure 3: Prediction P=20 days for K=10 data samples

As we can see from the bottom Fig. 3, the regression curve for K=10 fits quite decently the data, but the uncertainty increases as soon as we try to predict a bigger number of data points. It's also not surprising that we have smaller values of σ in the region in which the data points lie. As the number of the data point points increases (Fig.4, 5), the green $+/-2\sigma$ region gets smaller with respect to the data range.
In Fig. 6 we highlighted the MAP weights as well. In Fig. 7, 8 setting $\alpha = 0$ we obtained an unregularized least-squares plot, since $\lambda = \frac{\alpha}{\beta}$. In Fig. 9, 10, we set different (extreme) ratios for $\frac{\alpha}{\beta}$, with $0.1 < \alpha, \beta < 10$ In this last two Figures we notice how an high value for of $\frac{\alpha}{\beta}$ encourages the coefficients to decay. Finally, in Figure 11 we see how decreasing β (precision parameter) increases σ

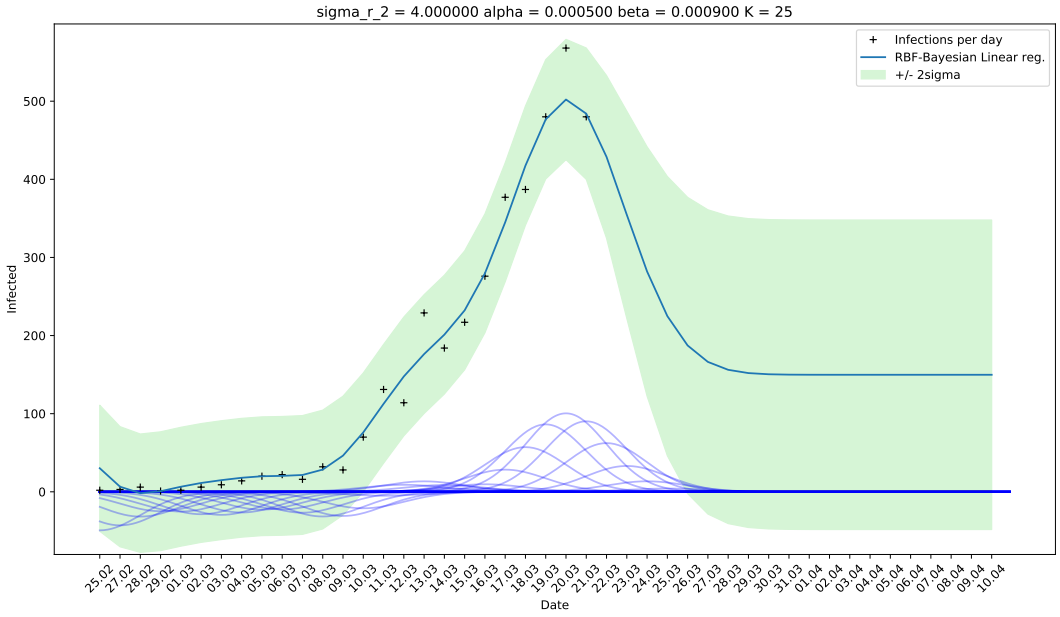

Figure 4: K = 25

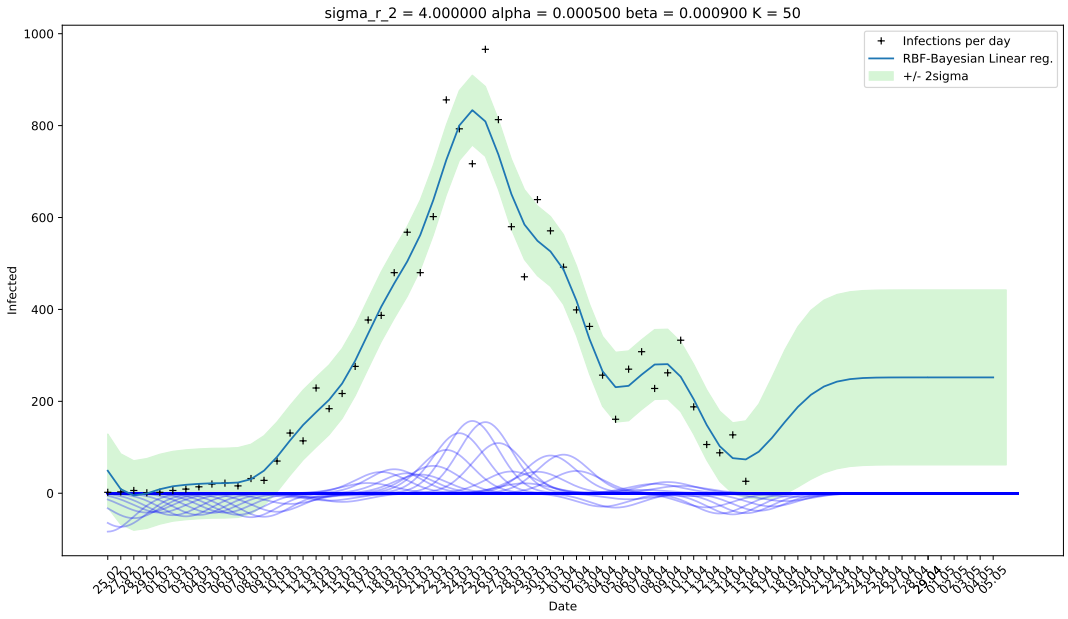

Figure 5: K = N = 50

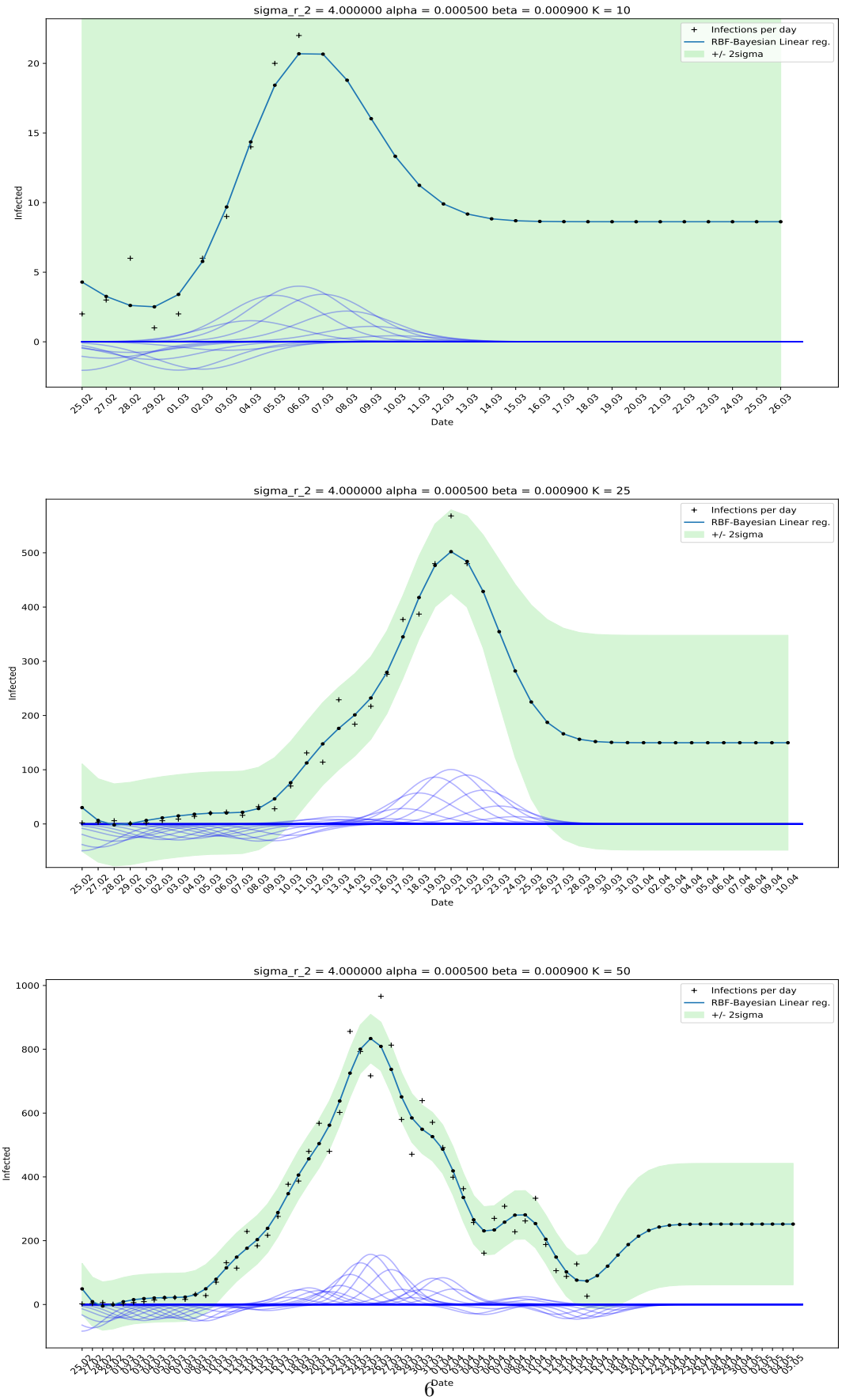

Figure 6: Complete predictive distribution and MAP solution

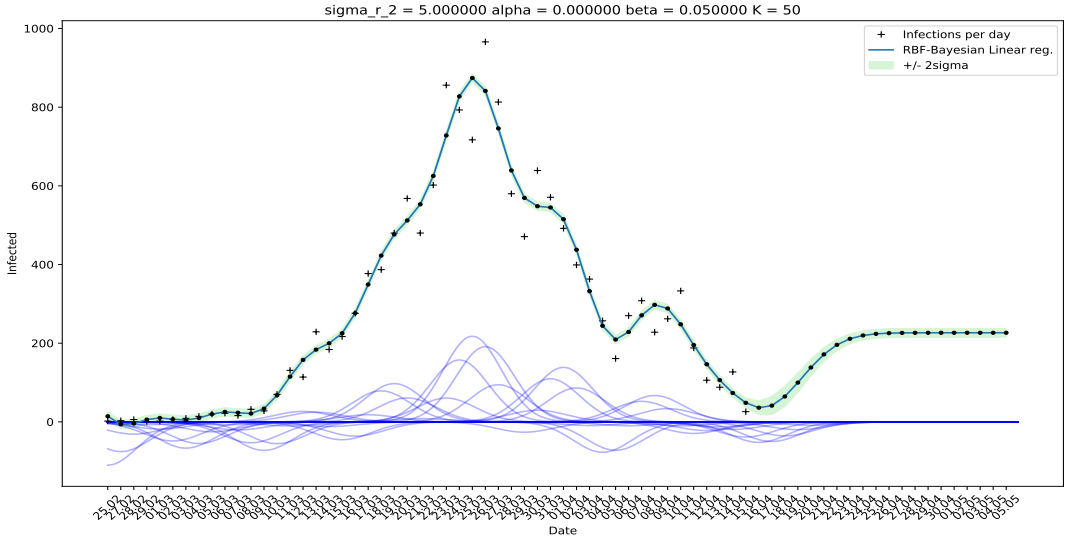

Figure 7: $\alpha = 0$, $\sigma_r^2 = 5$

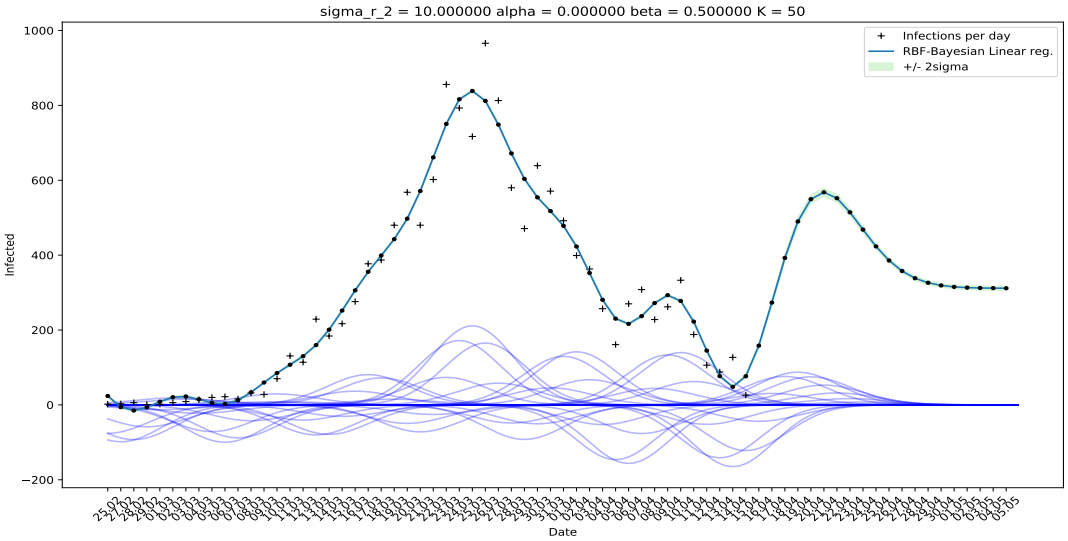

Figure 8: $\alpha = 0$, $\sigma_r^2 = 10$

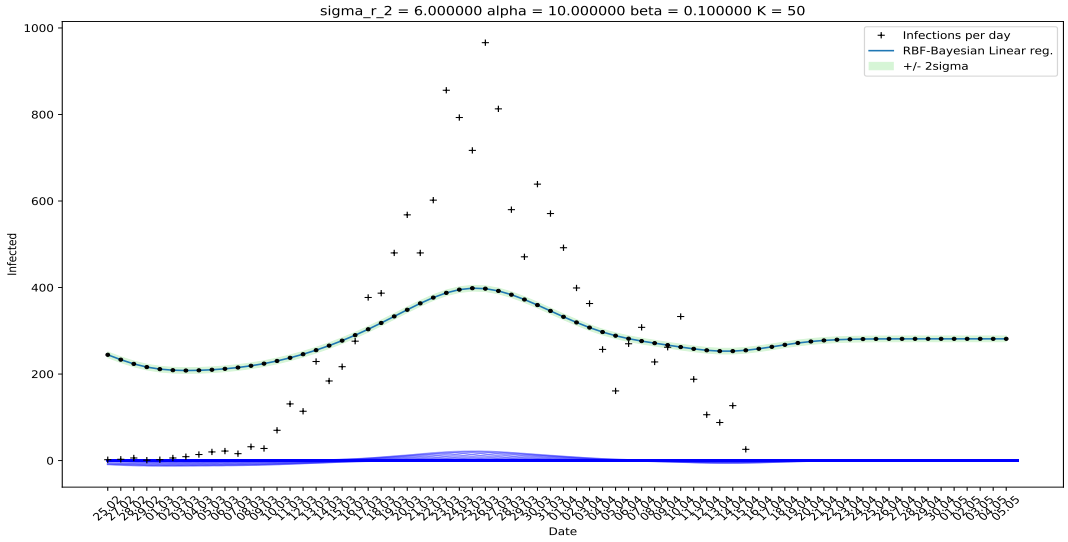

Figure 9: $\frac{\alpha}{\beta} = 100$, $0.1 < \alpha, \beta < 10$

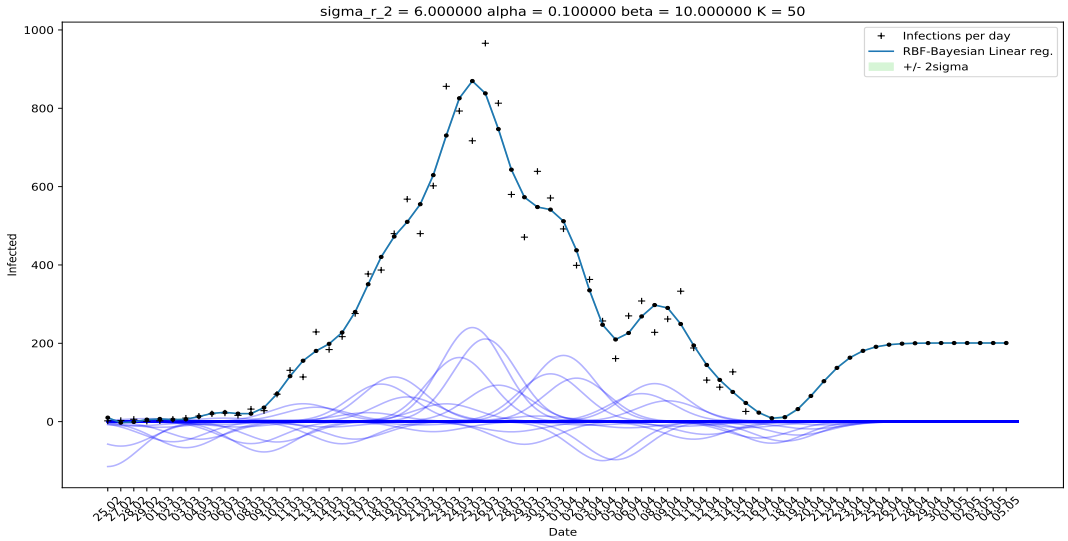

Figure 10: $\frac{\alpha}{\beta} = 0.01$, $0.1 < \alpha, \beta < 10$

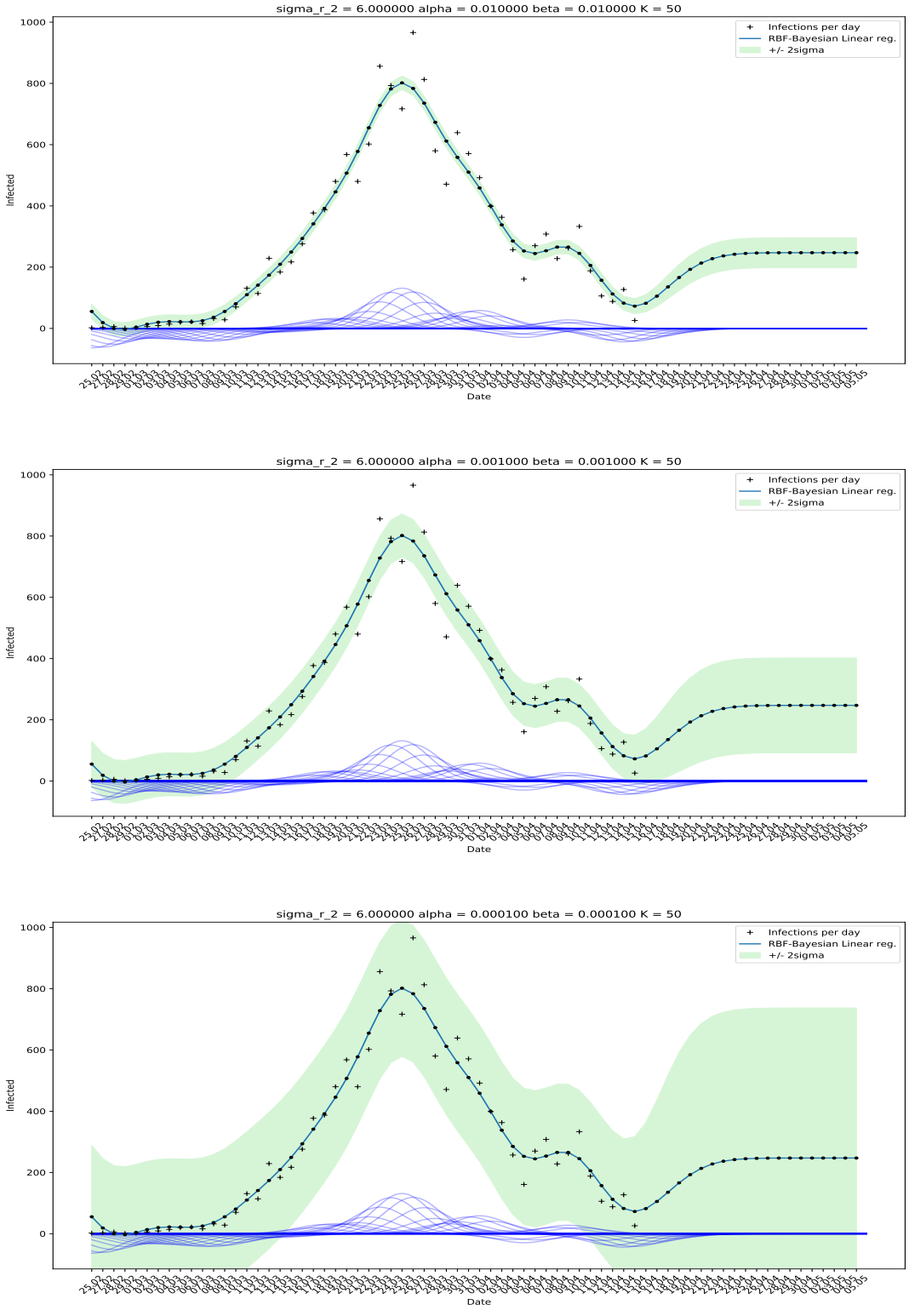

Figure 11: $\frac{\alpha}{\beta} = 1$ for decreasing values of α or β

Exercise 2. (Logistic Regression)

Solution. 1. We have $p(t=1|\tilde{\boldsymbol{x}}) = \sigma(\tilde{\boldsymbol{w}}^t\tilde{\boldsymbol{x}}) = \sigma(1 \cdot \boldsymbol{w}^t\tilde{\boldsymbol{x}})$ and

$$\begin{aligned} p(t=-1|\tilde{\boldsymbol{x}}) = 1 - \sigma(\tilde{\boldsymbol{w}}^t\tilde{\boldsymbol{x}}) = 1 - \frac{1}{1+e^{-\tilde{\boldsymbol{w}}^t\tilde{\boldsymbol{x}}}} = \frac{1+e^{-\tilde{\boldsymbol{w}}^t\tilde{\boldsymbol{x}}}}{1+e^{-\tilde{\boldsymbol{w}}^t\tilde{\boldsymbol{x}}}} - \frac{1}{1+e^{-\tilde{\boldsymbol{w}}^t\tilde{\boldsymbol{x}}}} \\ = \frac{e^{-\tilde{\boldsymbol{w}}^t\tilde{\boldsymbol{x}}}}{1+e^{-\tilde{\boldsymbol{w}}^t\tilde{\boldsymbol{x}}}} = \frac{1}{e^{\tilde{\boldsymbol{w}}^t\tilde{\boldsymbol{x}}}+1} = \sigma(-\tilde{\boldsymbol{w}}^t\tilde{\boldsymbol{x}}) \end{aligned}$$

2.

$$\begin{aligned} -\log(p(\tilde{\boldsymbol{w}}|\boldsymbol{t},\boldsymbol{x})) = -\log(p(\boldsymbol{w})) - \sum_{n=1}^{N}\log(\sigma(t_n\tilde{\boldsymbol{w}}^{\boldsymbol{t}}\tilde{\boldsymbol{x}}_{\boldsymbol{n}})) = \\ -\log(p(\boldsymbol{w})) - \sum_{n=1}^{N}\log(\frac{1}{1+e^{-t_n\tilde{\boldsymbol{w}}^{\boldsymbol{t}}\tilde{\boldsymbol{x}}_{\boldsymbol{n}}}}) \\ = -\log(p(\boldsymbol{w})) + \sum_{n=1}^{N}\log(1+e^{-t_n\tilde{\boldsymbol{w}}^{\boldsymbol{t}}\tilde{\boldsymbol{x}}_{\boldsymbol{n}}}) = \\ -\log\left(\frac{1}{(2\pi S^2)^{N/2}}\exp\left(-\frac{1}{2S^2}\sum_{n=1}^{N}w_n^2\right)\right) + \sum_{n=1}^{N}\log(1+e^{-t_n\tilde{\boldsymbol{w}}^{\boldsymbol{t}}\tilde{\boldsymbol{x}}_{\boldsymbol{n}}}) \\ = -\log\left(\frac{1}{(2\pi S^2)^{N/2}}\right) + \frac{1}{2S^2}\sum_{n=1}^{N}w_n^2 + \sum_{n=1}^{N}\log(1+e^{-t_n\tilde{\boldsymbol{w}}^{\boldsymbol{t}}\tilde{\boldsymbol{x}}_{\boldsymbol{n}}}) \end{aligned}$$

3.

$$\begin{aligned} \frac{\partial}{\partial b}E(\tilde{\boldsymbol{w}}) = \sum_{n=1}^{N}\frac{\partial}{\partial b}\log(1+e^{-t_n\tilde{\boldsymbol{w}}^{\boldsymbol{t}}\tilde{\boldsymbol{x}}_{\boldsymbol{n}}}) = \sum_{n=1}^{N}\frac{-t_ne^{-t_n\tilde{\boldsymbol{w}}^{\boldsymbol{t}}\tilde{\boldsymbol{x}}_{\boldsymbol{n}}}}{1+e^{-t_n\tilde{\boldsymbol{w}}^{\boldsymbol{t}}\tilde{\boldsymbol{x}}_{\boldsymbol{n}}}}, \\ \frac{\partial}{\partial w_i}E(\tilde{\boldsymbol{w}}) = \frac{1}{2S^2}\sum_{n=1}^{N}\frac{\partial}{\partial w_i}w_n^2 + \sum_{n=1}^{N}\frac{\partial}{\partial w_i}\log(1+e^{-t_n\tilde{\boldsymbol{w}}^{\boldsymbol{t}}\tilde{\boldsymbol{x}}_{\boldsymbol{n}}}) \\ = \frac{w_i}{S^2} + \sum_{n=1}^{N}\frac{-t_n\tilde{x}_n^ie^{-t_n\tilde{\boldsymbol{w}}^{\boldsymbol{t}}\tilde{\boldsymbol{x}}_{\boldsymbol{n}}}}{1+e^{-t_n\tilde{\boldsymbol{w}}^{\boldsymbol{t}}\tilde{\boldsymbol{x}}_{\boldsymbol{n}}}} \end{aligned}$$

5. b)

e,f) The decision boundary is the curve

$$p(t=1|\boldsymbol{x}) = 0.5 \Leftrightarrow \frac{1}{1+e^{\tilde{\boldsymbol{w}}^{\boldsymbol{t}}\tilde{\boldsymbol{x}}}} = 0.5 \Leftrightarrow 1+e^{\tilde{\boldsymbol{w}}^{\boldsymbol{t}}\tilde{\boldsymbol{x}}} = 2 \Leftrightarrow \tilde{\boldsymbol{w}}^{\boldsymbol{t}}\tilde{\boldsymbol{x}} = \log(1) = 0,$$

so it is the line

$$y = -\frac{w_1}{w_2}x - \frac{b}{w_2}.$$

In this case the computed decision boundary really seperates the red and blue training points (for both values of S^2), although in some cases the points may not be linearly separable in two dimensions, which also happened in some tests. The probability around the decision boundary is about 0.5, which is clear, because the boundary was chosen for this reason. If a high variance for $\boldsymbol{w}$ is assumed, the probability decreases rather slowly when moving away from the decision boundary. For higher S^2 the probability decreases much faster and

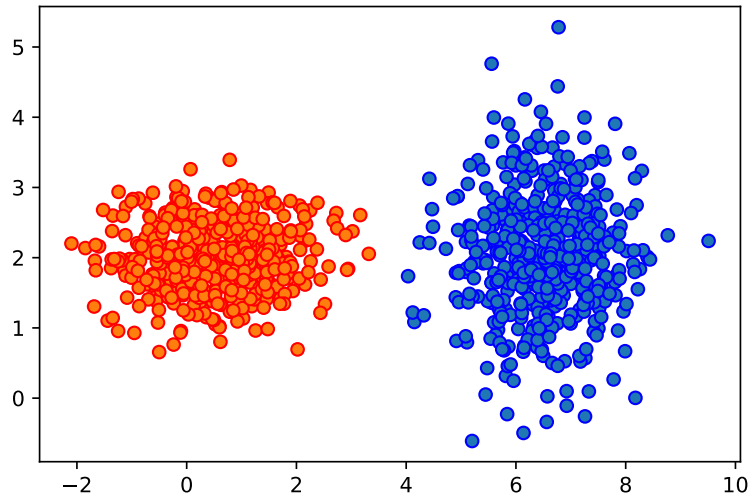

Figure 12: The generated points

the predictions can be made with more certainty.

6. The accuracy for variing S^2 is

It fits the previous observation that for higher S^2 the predictions can be made with more certainty, but this effect seems to be limited by $S^2 \approx 0.1$, after that no real improvement is achieved.

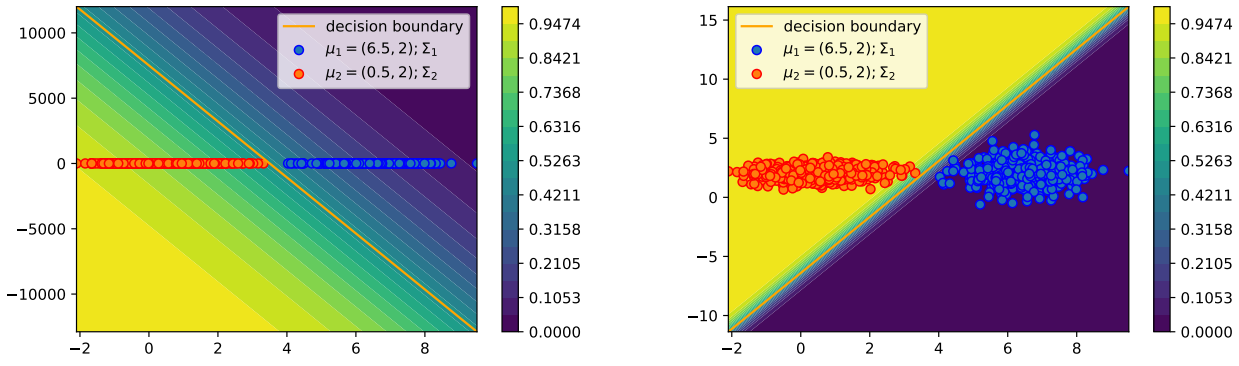

Figure 13: decision boundary and probabilities for $S^2 = 10^{-3}, 10^4$

S^2	training	validation
10^{-4}	0.6088	0.6029
10^{-3}	0.8435	0.8377
10^{-2}	0.9003	0.8826
10^{-1}	0.9193	0.9123
1	0.9165	0.9152
10^{1}	0.9169	0.9109
10^{2}	0.9165	0.9145
10^{3}	0.9143	0.9101
10^{4}	0.9165	0.9138