Support Vector Regression

A) Consider the 1-dimensional SVR problem with two datapoints shown in Figure 1.

1. Estimate the value of b (use equation given in class) and draw the regressive line found by ϵ-SVR, for a kernel width of 0.1 and an $\epsilon = 0.1$.
2. Draw the effect of increasing ϵ to 0.3, 0.5 and 1.
3. Draw the effect of increasing the kernel width.

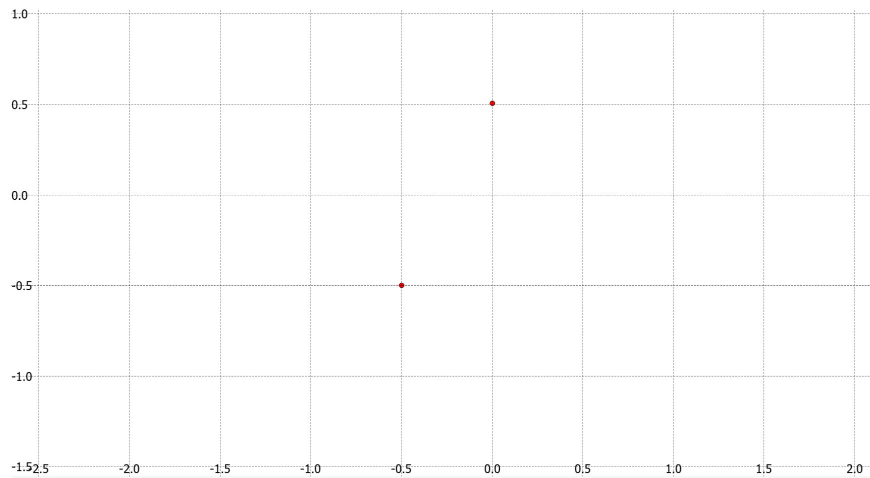


Figure 1: ϵ-SVR, case a

B) Consider the 1-dimensional SVR problem shown in Figure 2, where 2 data points are superimposed on x (i.e. they have the same value for x but a different value for y) .
Discuss the effect of the ϵ parameter on the regression function in this case.

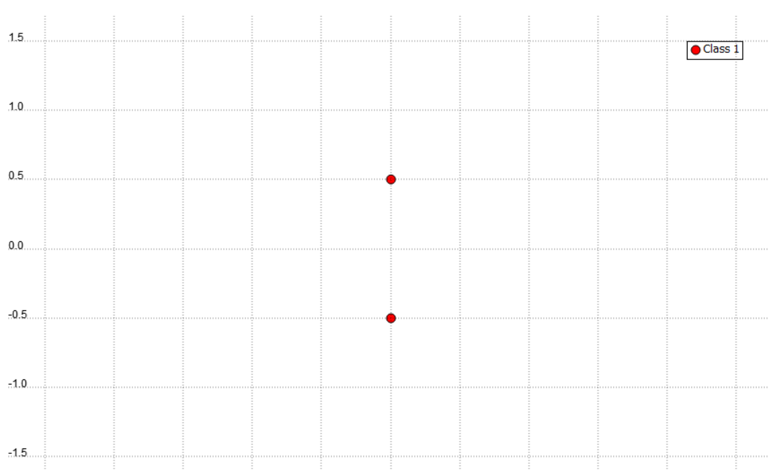


Figure 2: ϵ-SVR, case b

Solutions

A)

1. We have $\alpha_1^* = \alpha_2$ and b is simply equal to the average of the y_i as the effect of the datapoints cancel each other out.

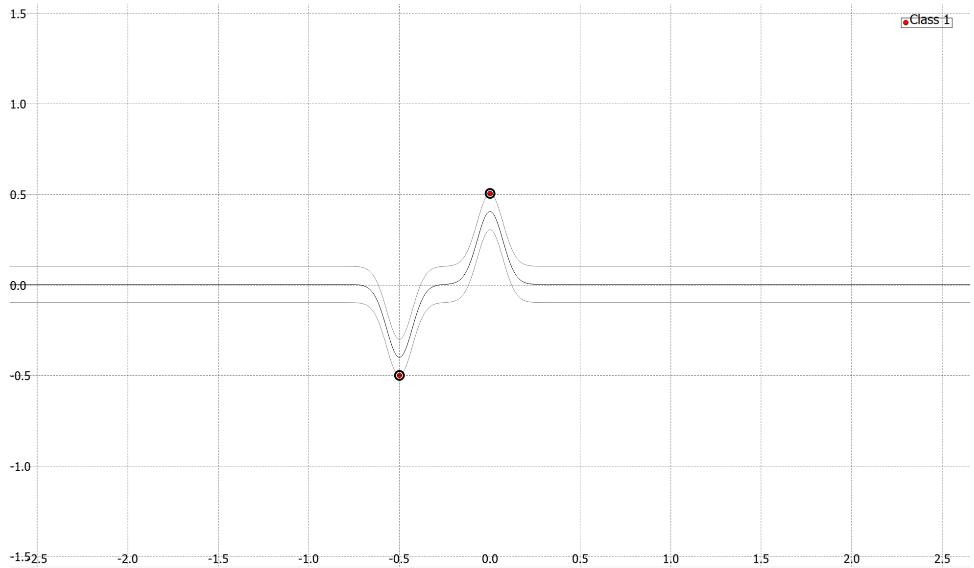


Figure 3: ϵ-SVR, case a

2. For $\epsilon > 0.5$, the "ϵ-tube" is too wide and as a result both the data points fall inside it and hence they cannot be support vectors (see Figure 4). In this case, the solution will be a function that evaluates everywhere to b. As no point is considered as support vector, b simply amounts to the mean of the data, i.e.:

$$b = \frac{1}{M} \sum_{j=1}^{M} y^j$$

When $\epsilon \leq 0.5$, both data points become support vector simultaneously and the regression function starts showing nonlinear behavior near the data points, ultimately going back to the value of b far away from them. The notion of "near" and "far" is decided by the kernel width.

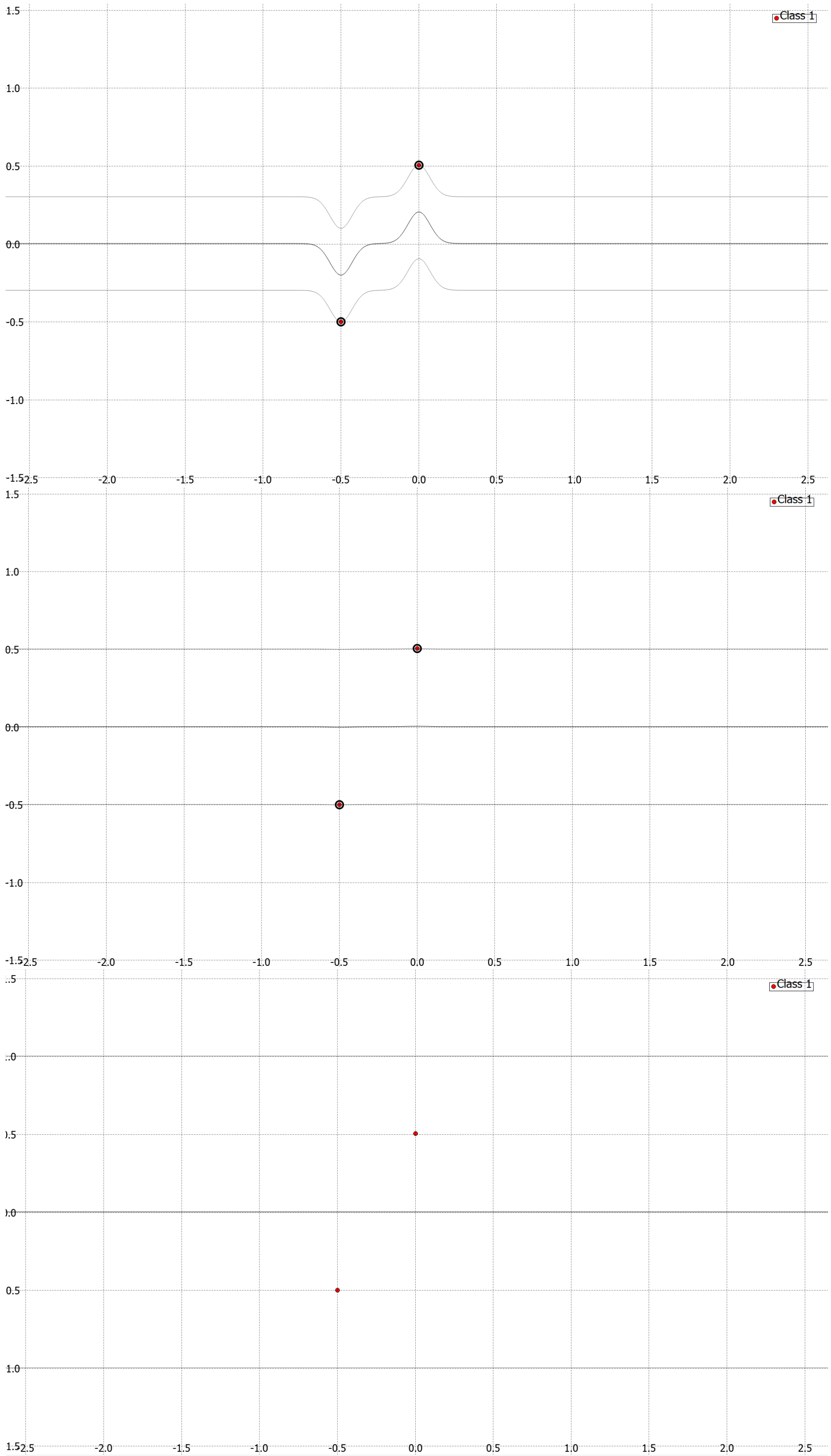


Figure 4: From top to bottom, ϵ-SVR with $\epsilon = 0.3$, $\epsilon = 0.5$, $\epsilon = 1.0$, and fixed $\sigma = 0.1$.

3. Increasing the kernel width lead the regressor to overshoot the datapoints and predict the trend generated by the two datapoints for some time until it goes back to the standard value of b.

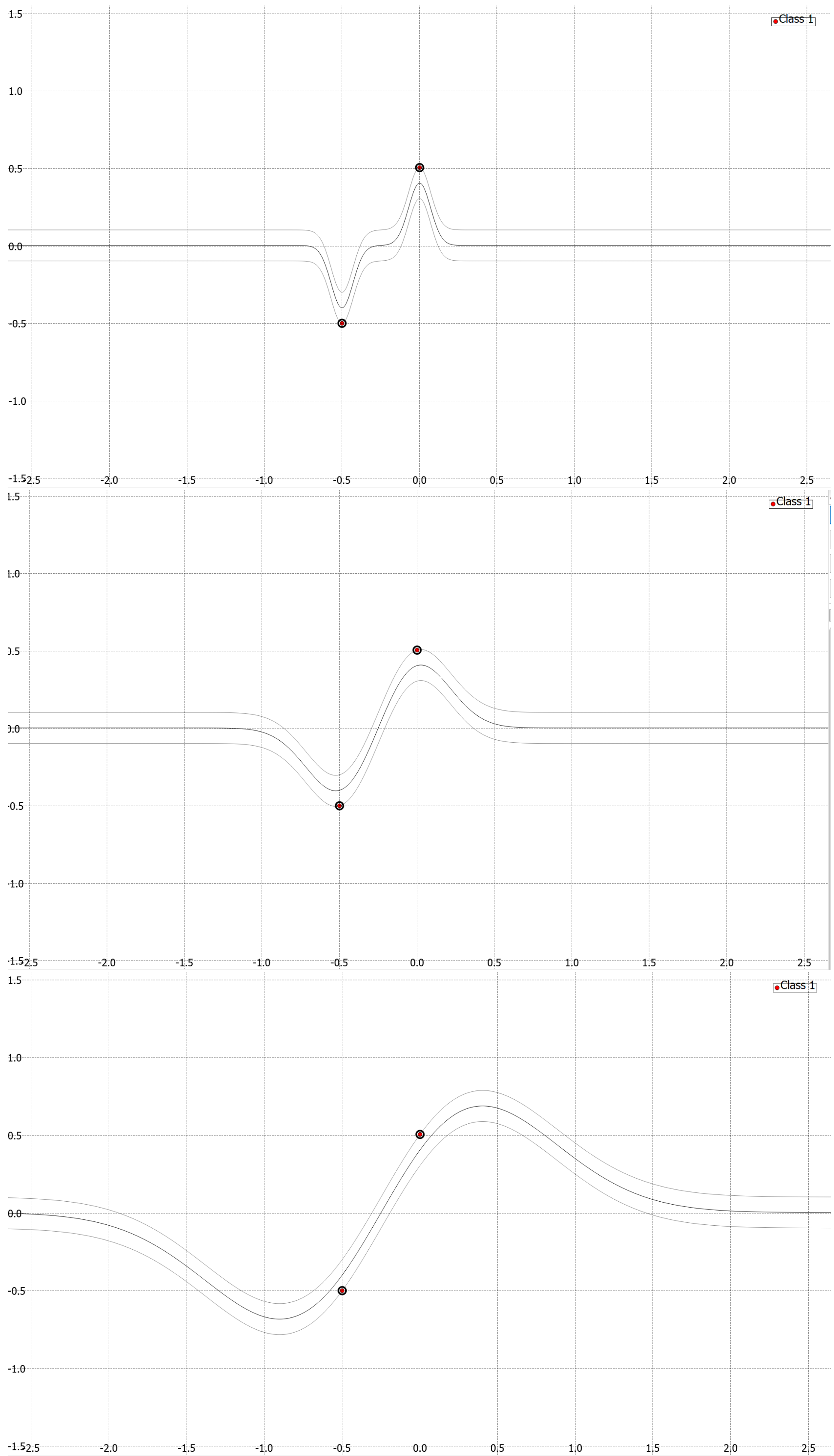


Figure 5: From top to bottom, ϵ-SVR with kernel width 0.1, 0.5 and 0.9, and fixed $\epsilon = 0.1$

B) The SVR regressor is

$$y = f(x) = \sum_{i=i}^{M} (\alpha_i^* - \alpha_i) k(x^i, x) + b \tag{1}$$

and the following constraint holds:

$$\sum_{i=1}^{M} (\alpha_i^* - \alpha_i) = 0 \tag{2}$$

which in our 2-point problem yields

$$(\alpha_1^* - \alpha_1) = -(\alpha_2^* - \alpha_2) \tag{3}$$

As the two points have the same value for x, we have

$$k(x^1, x) = k(x^2, x) \quad \forall x \tag{4}$$

Plugging (3) and (4) in (1) yields $y = b \quad \forall x$. So the regression function is a horizontal line. This is true whether or not the points are support vectors. So the parameter ϵ does not matter in this case.

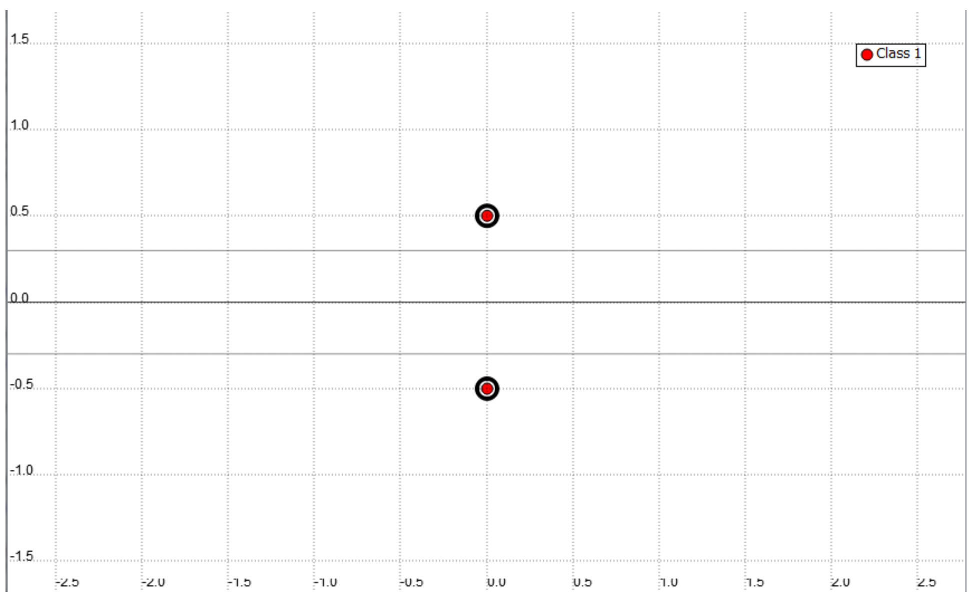


Figure 6: ϵ-SVR, case b

GMR

A) For each of these 2D distributions fitted with Gaussian Mixture Models (GMM), draw approximately the expected result of Gaussian Mixture Regression (i.e. the regressive signal given by GMR) on top of the datapoints (see Figure 7). Consider that, as we did in class, we compute $\mathbb{E}[p(y|x)]$ where x is the horizontal axis and y is the vertical axis. Assume an equal prior for each of the Gaussians:

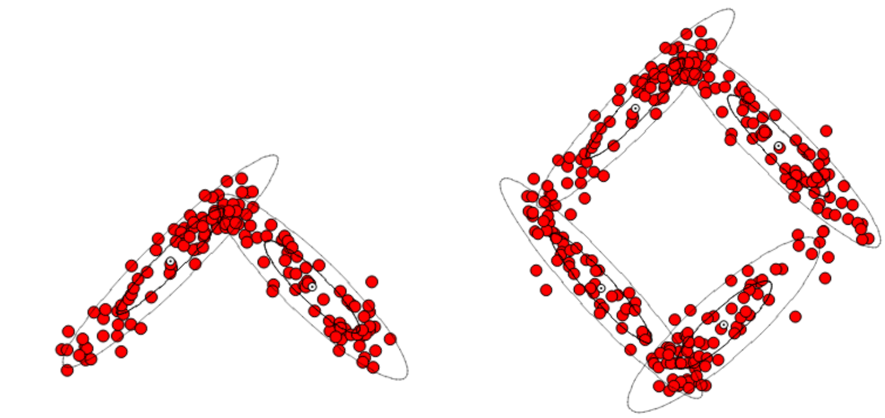


Figure 7: 2D distributions fitted with 2 Gaussians (left) and 4 Gaussians (right). The priors are assumed identical for all Gaussians.

B) As in the previous question, draw approximately the expected result of GMR. However, take this time into account the priors of each Gaussian (see Figure 8). Note that the priors are arbitrary and may not correspond to the real priors found by GMM on this dataset:

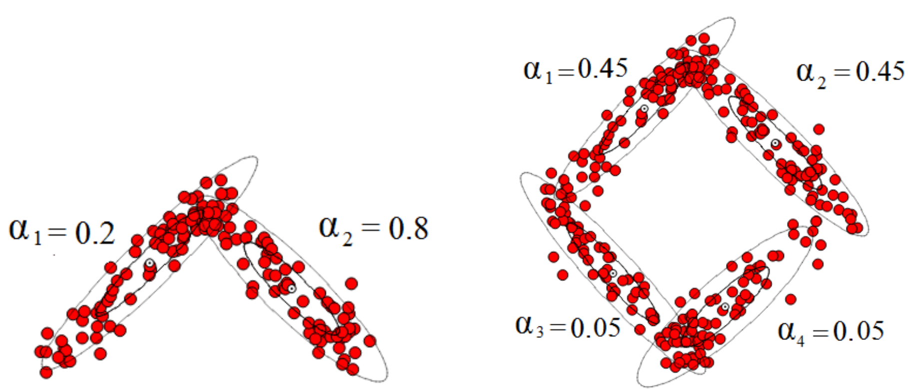
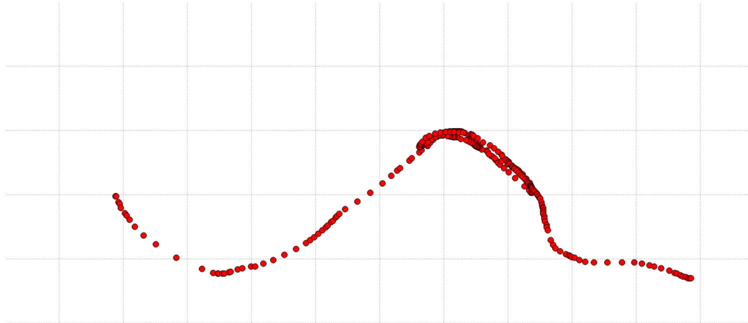
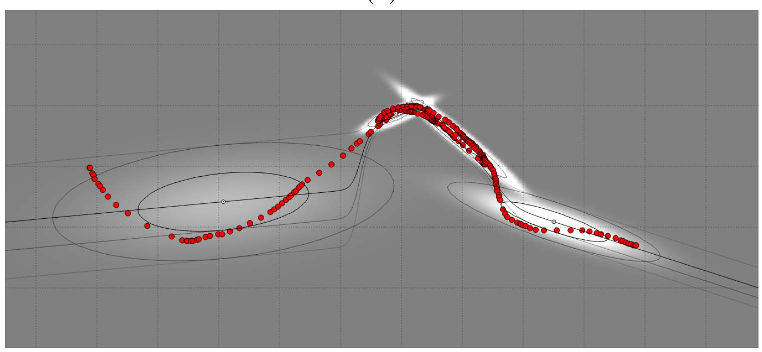


Figure 8: 2D distributions fitted with 2 Gaussians (left) and 4 Gaussians (right). The priors α_i are displayed next to each Gaussian.

C) Observe the original dataset below (Figure 9a) and the solution found by GMR (Figure 9b). How do you explain the poor fit with the two large Gauss functions on both end of the datasets? How could you fix the problem?



(a)



(b)

Figure 9: 2D nonlinear regression problem. (a) Original dataset. (b) Results from the application of GMR. The color shading corresponds to the uncertainty measure (likelihood) of the model.

D) Consider the dataset shown in Figure 10:

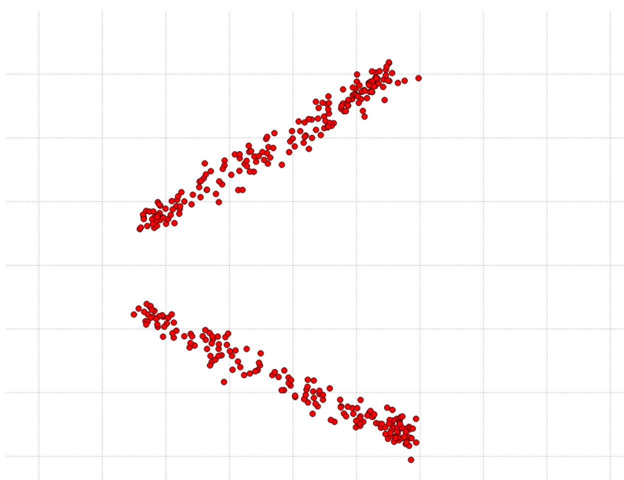


Figure 10: 2D regression problem.

1. Discuss the effect of using one or two components in GMR?
2. GMM models the joint distribution $p(x, y)$ of a dataset. In GMR, one takes the expectation $\hat{y} = \mathbb{E}[p(y|x)]$ of the conditional distribution $p(y|x)$ as the output of the regression. Considering

that you use a mixture of two gaussians to model the distribution in Figure 10, do you have suggestions for a different way of obtaining the output from the conditional distribution?

E) Draw a 2D example where two different GMMs give the same GMR output (only draw the Gaussians, you do not have to draw a dataset). Assume equal priors.

F) What shape of regression function can you achieve with only one Gaussian in GMR? Does it depend on the covariance matrix of the Gaussian? (Optional: prove your answer by computing the output of a GMR with a diagonal Gaussian).

Solutions

A) For one Gaussian, the expectation corresponds to the mean. The regression line thus passes through the first axis of each Gaussian (which corresponds to the mean of the conditional $p(y|x)$, since other Gaussians have extremely low weight). In the middle of the image, each component influences the regression output and there is a smooth interpolation between the two locally linear sections. For the second distribution, the distribution is symmetric, so is the conditional all along the input, therefore the expectation and mean is always at the line of symmetry (see Figure 11).

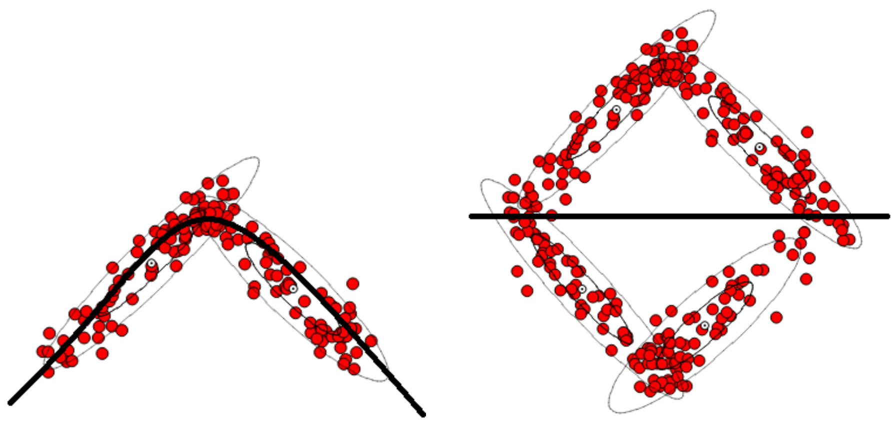


Figure 11: 2D distributions fitted with 2 Gaussians (left) and 4 Gaussians (right). The priors are assumed identical for all Gaussians. The black line corresponds to the resulting regression function.

B) For the first distribution, the right Gaussian has more influence than the left Gaussian, and hence influences more the regression in the center. For the second one, the Gaussians on the top take advantage over the ones at the bottom, the result tends towards the above Gaussians (see Figure 12).

C) The reason for the poor fit comes from the fact that the central region of the dataset has much more datapoints than the two ends of the dataset. Since GMM chooses to put Gauss functions so as to maximize the likelihood, it will make more effort to fit well regions with high density of datapoints, since a poor fit of this region would lead to a larger decrease of the likelihood. Having different density of datapoints can happen typically if you gather data from monitoring displacement of a robot. In regions where the robot moves slowly, you end up with more datapoints than in regions where the robot moves fast, since the sampling rate of your sensor remains the same.

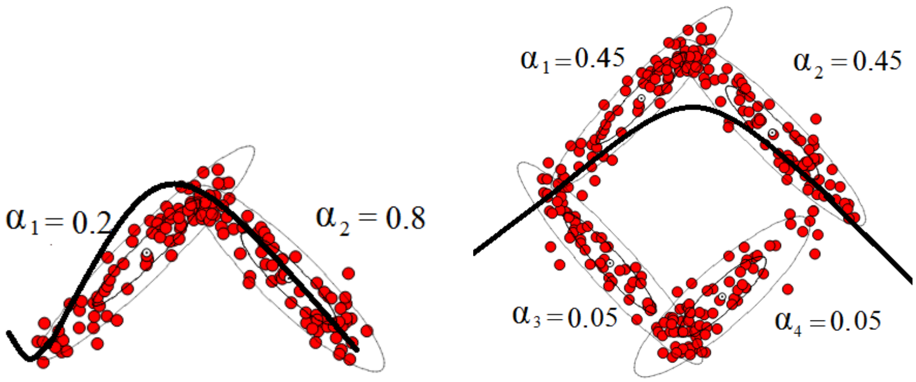
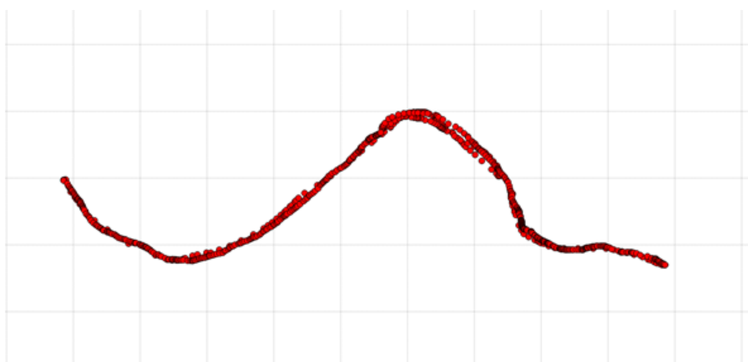
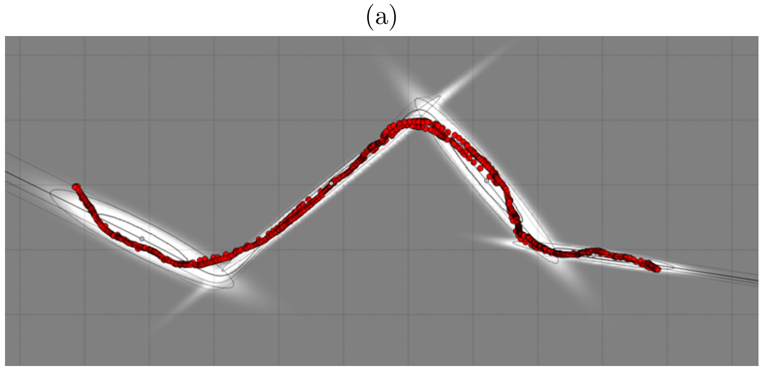


Figure 12: 2D distributions fitted with 2 Gaussians (left) and 4 Gaussians (right). The priors α_i are displayed next to each Gaussian. The black line corresponds to the resulting regression function.

You could either add more datapoints artificially in regions with low density (by duplicating existing datapoints and adding small noise). The result of doing this is shown in Figure 13. As you can see the distribution of the Gauss functions is now more even and the resulting fit of the data is closer to all datpoints in the curve.



(a)

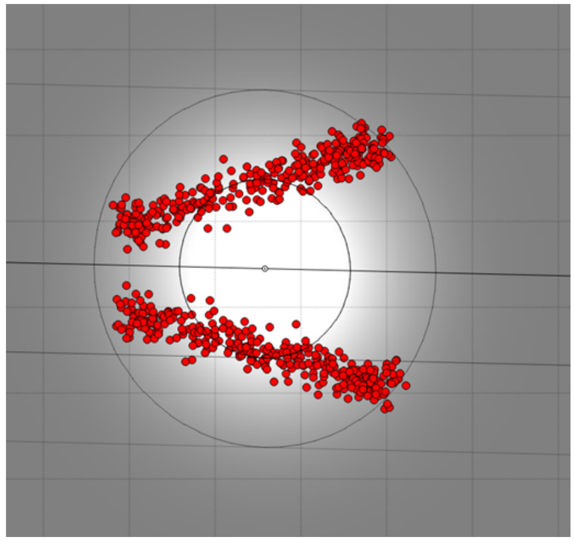


(b)

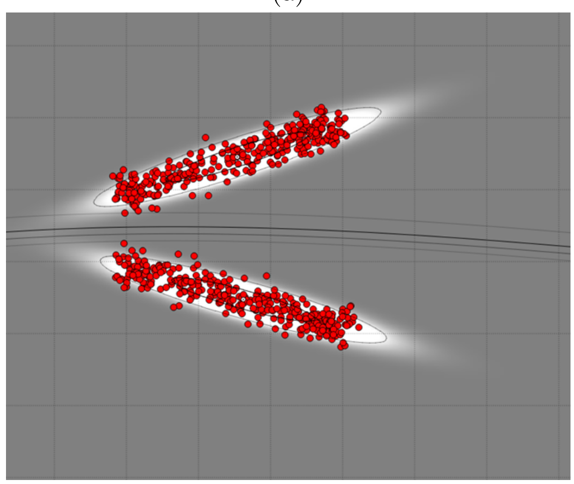
Figure 13: 2D nonlinear regression problem. (a) Preprocessed dataset. (b) Results from the application of GMR. The color shading corresponds to the uncertainty measure (likelihood) of the model.

D)

1. With one component (one Gauss function) the fit is a straight line going through the two groups of datapoints, see Figure 14a. With two Gauss functions, the regressive line remains a quasi straight fit in-between the Gauss functions, see Figure 14b.



(a)



(b)

Figure 14: Results from the application of GMR on the dataset shown in Figure 10 with 1 Gaussian (a) and 2 Gaussians (b). The color shading corresponds to the uncertainty measure (likelihood) of the model.

2. In the problem at hand, we do not have a function, but a bimodal distribution with two equally likely values for y given a value of x. Hence in place of using the expected value, one could climb the gradient of the likelihood to find either of the two modes.

E) A possible solution is drawn in Figure 15.

Figure 15: Two Gaussian Mixture Models (GMM) that would yield the same regression function.

F) With one Gaussian, the regression function can at best be a line. This line is constrained to be horizontal if the covariance matrix is diagonal or spherical. The regression function at a query point x^* for one Gaussian is:

$$y(x^*) = \mathbb{E}[p(y|x)] = \mu_y + \Sigma_{yx}\Sigma_{xx}^{-1}(x - \mu_x)$$

However, since the covariance matrix does not have off-diagonal terms, $\Sigma_{yx} = 0$ and the output y is a constant μ_y, passing through the center of the Gaussian. Otherwise, this is the equation of a line.

GMR & SVR

Consider the distribution of datapoints shown in Figure 16 and assume that you use either:

- SVR with tiny kernel width and very small ϵ
- GMR with a large number of Gauss functions

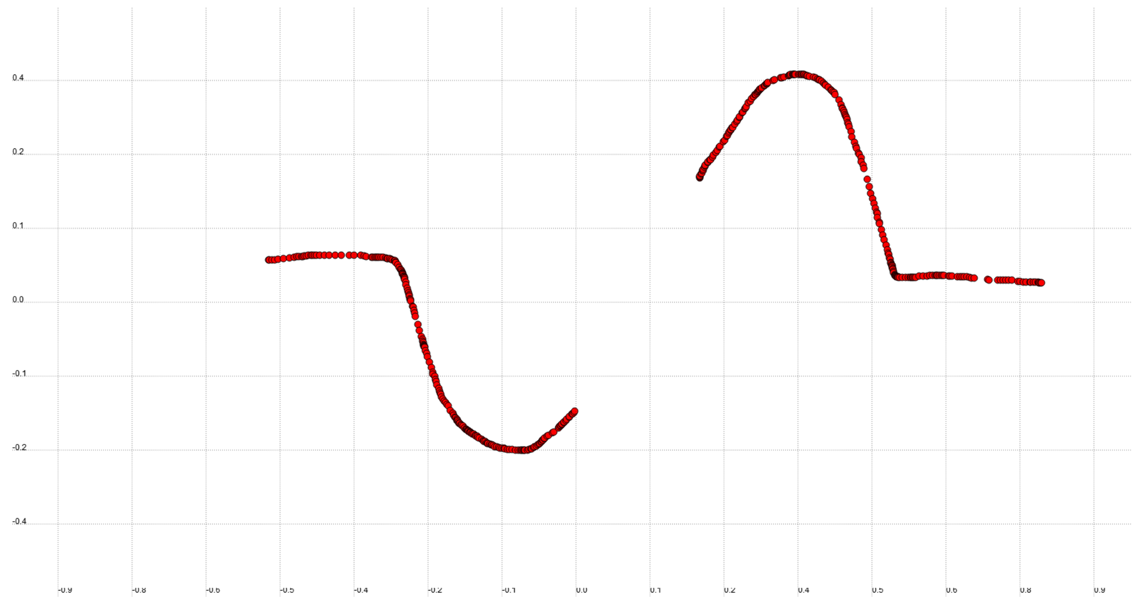
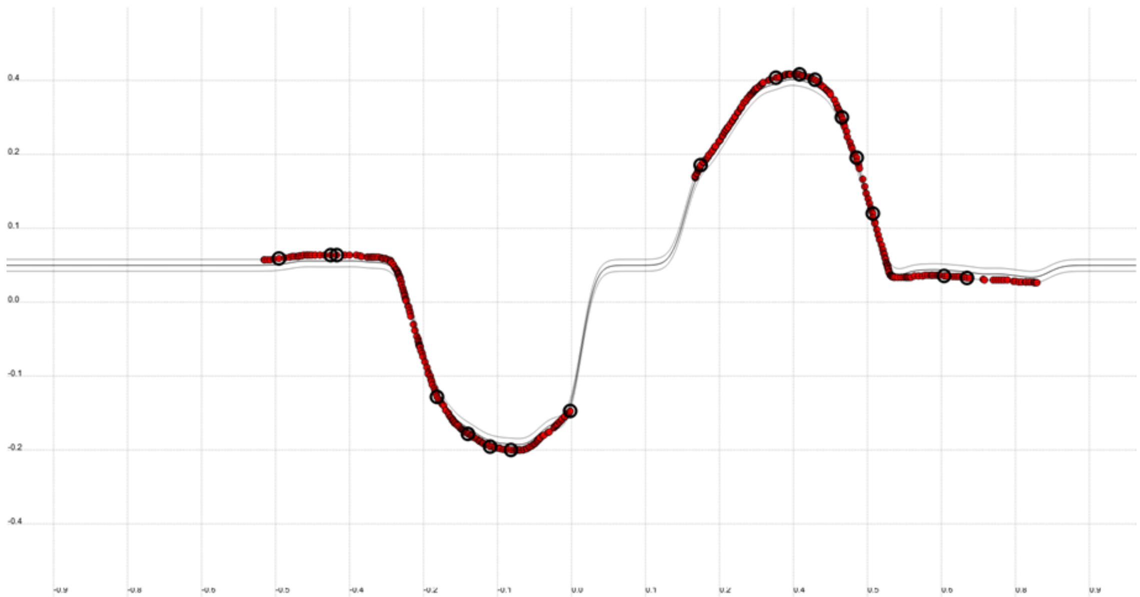


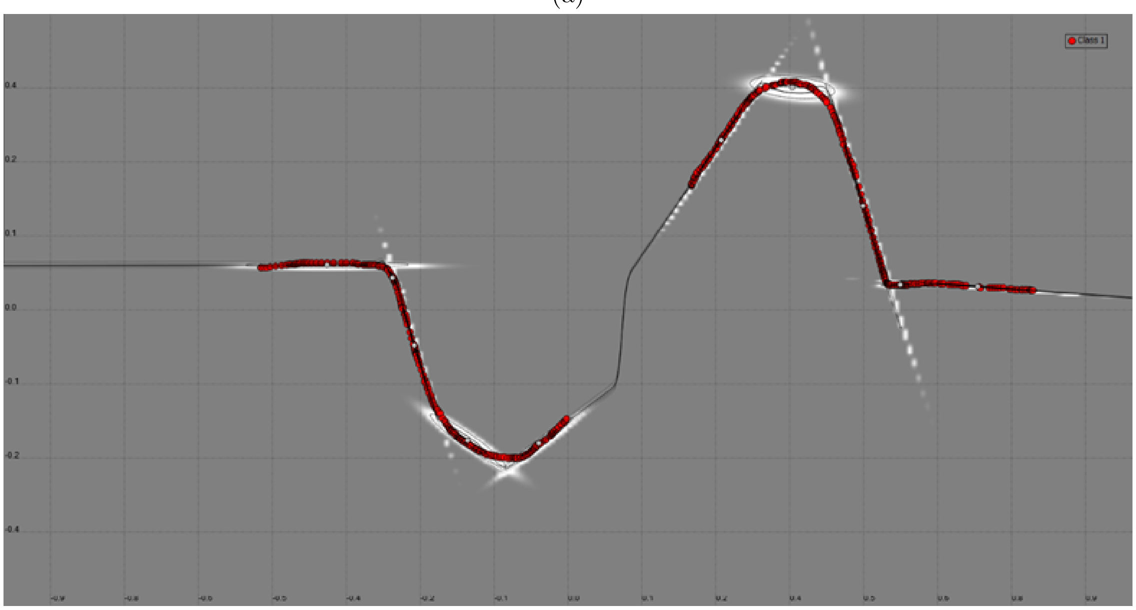
Figure 16: 2D regression problem featuring a "hole" in the dataset

Solutions

The SVR (Figure 17a) and GMR (Figure 17b) solutions are shown below. The interpolation far from the data on both ends of the curves is correct in both cases. For SVR, this is by chance as the extreme parts converge to the mean of the data, i.e. to b. The interpolation in the center where data are missing is poor for SVR, as the kernel width is too small and it interpolates to b for a small interval. GMR also yields a non linear fit due to the fact that Gauss functions are too small and hence have only limited influence.



(a)



(b)

Figure 17