Exercise 5

Machine Learning I

|  |  |
| --- | --- |
|  | 4A-1. |

We know:

Using Bayes’ theorem and rearranging terms we get:

The approximately “” is valid because most density is concentrated around .  
As we saw in the lecture slides, the evidence approximation is useful to make inferences about model complexity and fit.

Let us see what happens in if we vary the input :

As the likelihood is strongly related to the fit, acquires information about the model fit. As fit is related to model complexity, a better fit usually entails a more complex model as well. Additionally, because of our also receives information about model complexity and plausibility.   
  
If is too high compared to our belief , then goes down. Think about what this means. A very high posterior around leads to a point estimate not unlike the maximum likelihood method. This may lead to little bias in the input set **.** As we know from the bias/variance decomposition, this may increase variance for different datasets . Which means we could have learned too much from our data.

Because goes down in this case, it helps us combat overfitting.

On the other hand, if our fit is good ( high), yet we have a general enough model ( low), we receive a good value for .

The optimal value for is thereby a balance of model complexity, belief () and fit ().  
  
See more on page 163, Bishop PRML.

|  |  |
| --- | --- |
|  | 4A-2. |

|  |
| --- |
| Prerequisites |

|  |
| --- |
| Let be invertible, arbitrary and be the identity matrix.  We then have:  The proof of above can be seen here:  <http://www0.cs.ucl.ac.uk/staff/gridgway/mil/mil.pdf>  Also, as seen in the last lecture slides, the posterior of Bayesian linear regression with normal prior is: |

Let us have old and “new” datapoints.  
Because the predictive distribution is dependent on the posterior, it makes sense to calculate the new posterior first.

As seen in the task, we have:

This leads to:

Note: We have because the predictions are independent from each other. Focusing on the exponent:

This allows us to complete the square above and we get the updated posterior:

with

Note: When we get the same result as in the slides for the prior .

Now we use for the covariance:

where

Now we just plug our new covariance matrices into :

The last equality holds because

is at least positive semidefinite (for why, look up the properties of positive semidefinite matrices, especially in regard to and .

|  |  |
| --- | --- |
|  | 4A-3. |

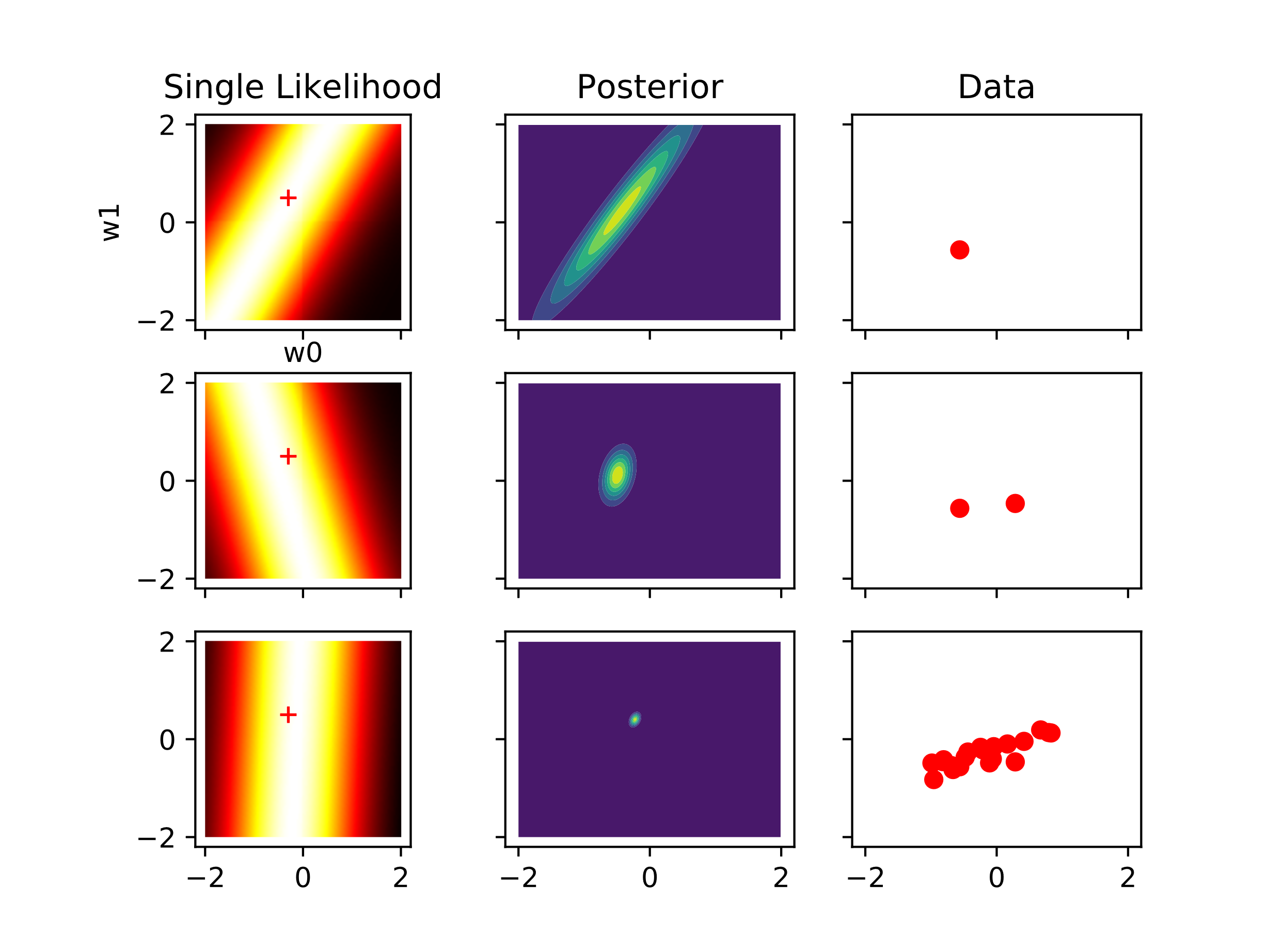


Figure 1 Replication of Bishop’s Figure 3.7 in Python. The Likelihood is of a single data point

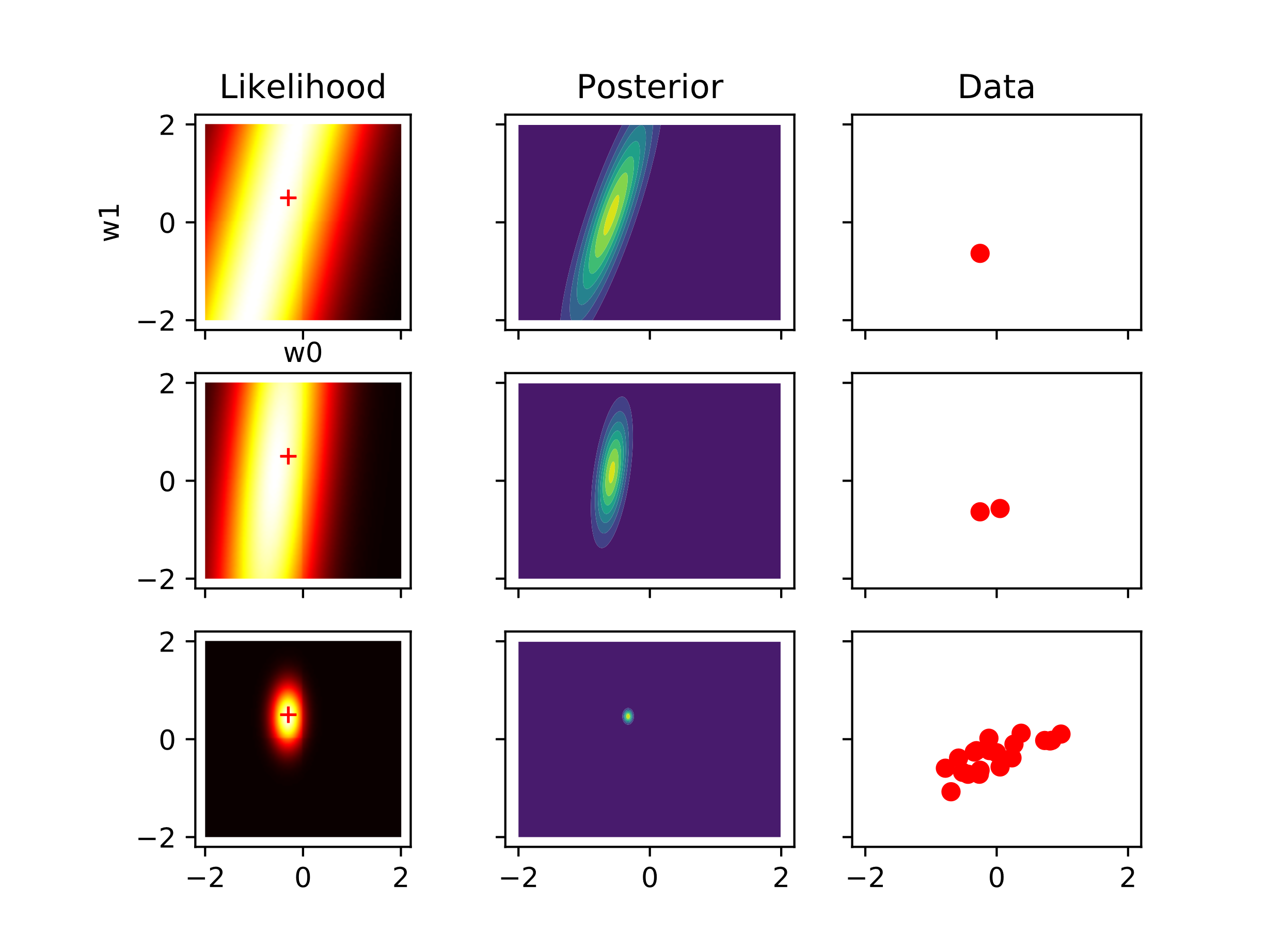


Figure 2 Replication of Bishop’s Figure 3.7 in Python this time with joint likelihood of the samples.