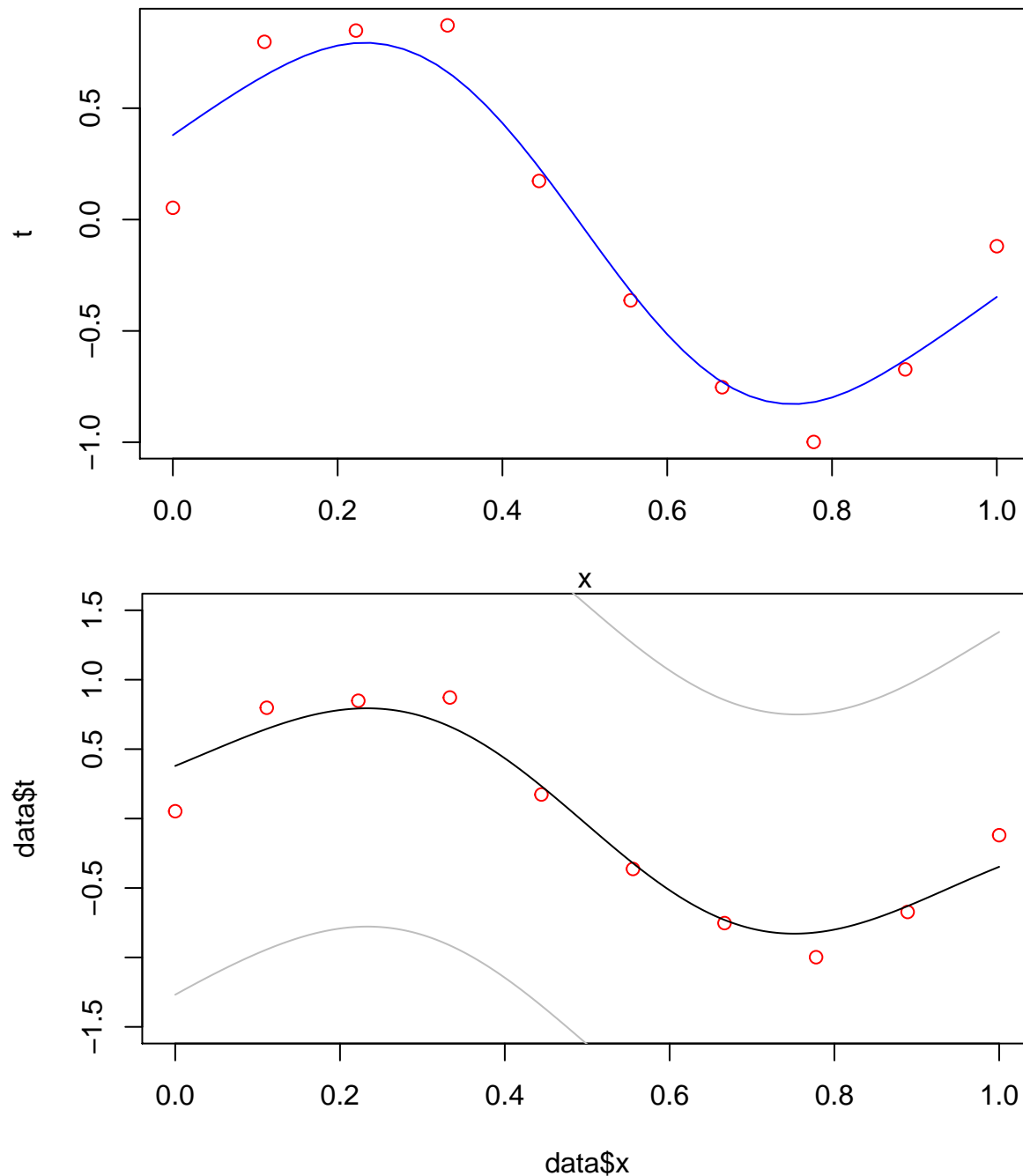
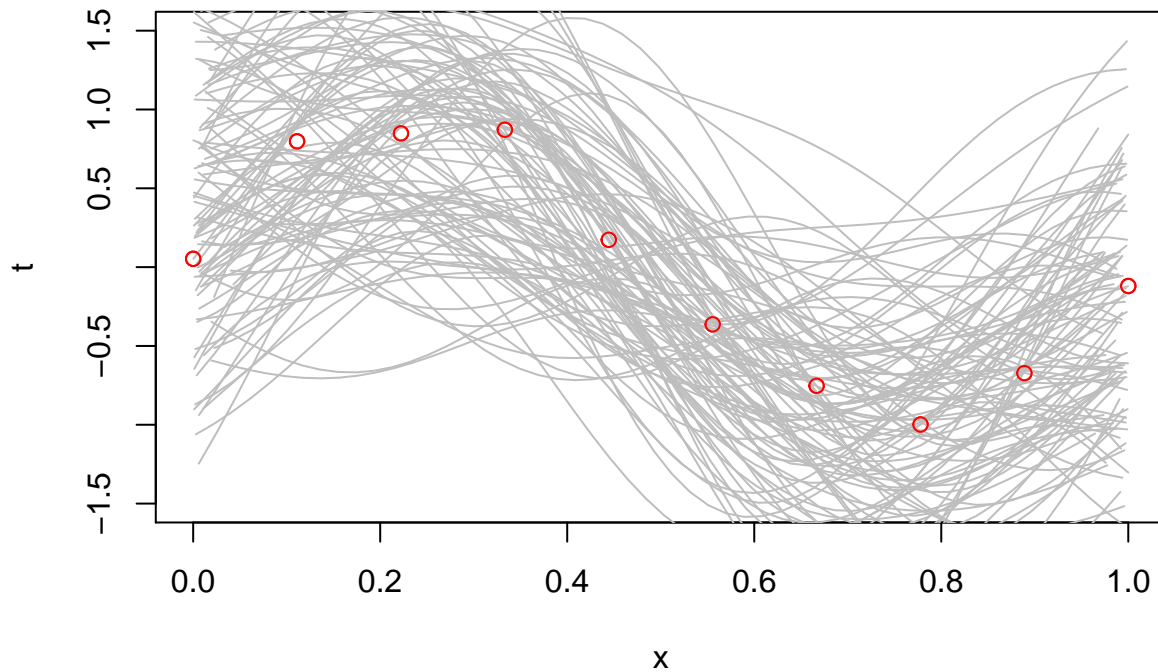


# Statistical Modelling and Inference: Week 3 Prior Modelling

## Questions from Prior Modelling

1. Continuation of your work on smooth function estimation with the data in `curve_data.txt`. Extend your program to also learn  $q$  from the data. Work with the Normal-Gamma prior with specification. Produce a figure that includes posterior draws of the linear predictor for this model





**2. Priors that penalize the L1 norm of the mean of a Gaussian and MAP estimation 2.1**  
 Suppose for simplicity that  $a > 0$  ( $a < 0$  can be handled in the same way), assume also that  $\lambda > 0$  and consider the function  $f(\mu) = (\mu - a)^2 + \lambda|\mu|$ . Show that  $\mu$  is minimised at  $(a - \lambda/2)^+$  where  $x^+$  denotes the positive part of  $x$ .

QUESTION: SHOULDN'T THIS BE THE POSITIVE PART OF  $\mu$ ?

To solve for the minimum of  $\mu$ , we take the derivative and set it to zero.

$$f(\mu) = \mu^2 - 2a\mu - a^2 - \lambda|\mu|$$

(take the derivative and set to 0)

$$\frac{df(\mu)}{d\mu} = 2\mu - 2a + \lambda = 0$$

(add  $2a - \lambda$  to both sides)

$$2\mu = 2a - \lambda$$

(divide both sides by 2)

$$\mu = a - \lambda/2$$

Q.E.D.

**2.2**

To estimate  $w_{MAP}$ , we want to maximize the joint probability  $p(t, X, w, q)$ . To maximize the joint probability, we must maximize the conditional probabilities of the prior and the log of the likelihood estimator of the probability for  $t$ . In so doing, we will calculate the MAP (maximum posterior)

The prior for  $w$  is given by the laplace distribution:

$$p(w) \propto \exp(-\delta/2) \sum_i |w_i|$$

$t$  is normally distributed, and the log likelihood of it's conditional probability is given by:

$$\ln p(t|x, w, q) = -q/2 \sum_{n=1}^N y(x, w) - t_n + \frac{N}{2} \ln(q) - \frac{N}{2} \ln(2\pi)$$

To maximize the likelihood of  $t$  with respect to  $w$ , we take the partial derivative with respect to  $w$  and set it to 0. The two right-most terms are omitted in this operation because they do not depend on  $w$ .

Also, we note that scaling the log likelihood by a positive constant coefficient does not alter the location of the maximum with respect to  $w$ , and so we can replace the coefficient  $\beta/2$  with  $1/2$ . Finally, instead of maximizing the log likelihood, we can equivalently minimize the negative log likelihood. We therefore see that maximizing likelihood is equivalent, so far as determining  $w$  is concerned, to minimizing the sum-of-squares error function defined by (1.2). (page 47 of Bishop)

To maximize the prior ( $p(w) \sim$  the laplace distribution), we similarly maximize the log likelihood of the prior, so we want to maximize

$$\ln p(w) = \frac{-\delta}{2} \sum_i |w_i|$$

To maximize the prior, we take it's derivative wrt  $w$  and set it to 0:

$$\frac{d \ln p(w)}{dw} = -\delta i = 0$$

So the maximization of the prior occurs when  $\delta$  is set equal to 0 and we can maximize  $w_{MAP}$  simply by minimizing the sum-of-squares error function.

$$\frac{q}{2} \sum_{n=1}^N y(x_n, w) - t_n^2$$

*QUESTION: WHAT IS CLOSED FORM? IS THIS SUFFICIENT? QUESTION: ARE  $\beta$  AND  $q$  THE SAME? OR IS  $q \propto \beta^{-1}$  ?*