# Locally Weighted Bayesian Regression

**Andrew W. Moore**
The Robotics Institute, Carnegie Mellon University
5000 Forbes Ave, Pittsburgh, PA 15123

**Email:** awm@cs.cmu.edu, **Phone:** 412-268-7599

## 1 The Problem

Suppose we have a dataset with $N$ datapoints. Each datapoint consists of a vector of inputs and a real valued-output, so the dataset is

$$\begin{array}{rcl} \mathbf{x}_0 & , & y_0 \\ \mathbf{x}_1 & , & y_1 \\ & \vdots & \\ \mathbf{x}_{N-1} & , & y_{N-1} \end{array}$$

The inputs need not be real-valued. All we require of them is a distance metric measuring the similarity of a pair of input vectors

$$\text{Dist} : \mathbf{x}, \mathbf{x}' \rightarrow \Re \tag{1}$$

and a set of $M$ basis functions

$$\phi_0 : \mathbf{x} \rightarrow \Re, \phi_1 : \mathbf{x} \rightarrow \Re, \ldots \phi_{M-1} : \mathbf{x} \rightarrow \Re \tag{2}$$

Write $\mathbf{z}(\mathbf{x}) = (\phi_0(\mathbf{x}), \phi_1(\mathbf{x}), \ldots, \phi_{M-1}(\mathbf{x}))$. And then write $\mathbf{z}_i = \mathbf{z}(\mathbf{x}_i)$.

## 2 Regression

Now, suppose we wish to do a weighted linear regression on the basis functions. The $i$th datapoint gets given weight $0 \leq w_i \leq 1$. What is the meaning of a weight?

- **Informally:** A datapoint gets given a large weight if it is believed to be highly relevant to the regression. For example, in a locally weighted regression query, if it is "close" to the query point it gets a weight of 1, and then as the distance to the query is increased the weight changes according to some non-increasing function. Points far from the query may be so irrelevant that they are given a weight of zero.

- **Formally:** We assume that $y_i$ was generated from a univariate gaussian with mean $\boldsymbol{\beta}^T \mathbf{z}_i$ and variance $\sigma^2 / w_i$ where $\boldsymbol{\beta}$ and $\sigma^2$ are unknown random variables to be identified. Note that for datapoints with weight $w_i = 0$ this makes the variance infinite and the datapoint essentially irrelevant for identifying $\boldsymbol{\beta}$ and $\sigma^2$.

We will allow the user to put a prior distribution on $\sigma^2$ and $\boldsymbol{\beta}$. And then the glorious magic of Bayesian statistics will throw up the posteriors for us.

# 3 Priors

The prior on $\sigma^2$ is

$$\sigma^2 \sim \text{Inv-}\chi^2(n_0, \sigma_0^2) \tag{3}$$

See [1] Appendix A for information on getting your hands around the throat of the *scaled* inverse $\chi^2$ distribution. To sample from it in C, see the `damut/minteg.h` code. In the above equation, the prior expected value of $\sigma^2$ is

$$\frac{n_0}{n_0 - 2} \sigma_0^2 \tag{4}$$

and the larger the value of $n_0$ the more confident we are in that prior estimate. $n_0$ can be interpreted as the number of "fake-datapoints-worth" of evidence with which we weight our prior guess.

The prior on $\boldsymbol{\beta}$ is conditional on $\sigma^2$. We'll discuss the implications of this later. $\beta_j$ (for $0 \leq j \leq M-1$) has a prior normal distribution

$$\beta_j \mid \sigma^2 \sim N(\beta_{j0}, \sigma_{\beta_j}^2 \sigma^2) \tag{5}$$

so that we believe apriori that the most likely value of $\beta_j$ is $\beta_{j0}$ and our uncertainty is represented by $\sigma^2 \sigma_{\beta_j}^2$: the larger the more uncertain.

As we will see later, this means our marginal prior distribution on $\beta_j$ is

$$\beta_j \sim t_{n_0}(\beta_{j0}, \sigma^2_{\beta_j}\sigma^2) \tag{6}$$

Notice that the priors on $\beta_i$ and $\beta_j$ are not independent, but *are* conditionally independent given $\sigma^2$.

# 4 Posteriors

We will end up will the full posterior distribution on $\boldsymbol{\beta}$ and $\sigma^2$ specified by a marginal scaled inverse $\chi^2$ distribution on $\sigma^2$ and then a normal distribution on $\boldsymbol{\beta}$ conditional on $\sigma^2$.

What is the number of datapoints used in the regression? It is effectively $N + M$. This, according to [1], is okay even if some or many of our weights are zero. The authors say (Page 260, third paragraph) that "a datapoint with infinite variance has no effect on the inference".

But we begin by defining some intermediate values to be computed.

- Define $Z$ to be an N-row, M-column matrix whose $i$th row is $\mathbf{z}_i$.
- Define $W = \text{Diag}(w_0, w_1, \ldots w_{N-1})$.
- Define $\mathbf{y} = (y_0, y_1, \ldots y_{N-1})^T$.
- Define $P = \text{Diag}\left(\frac{1}{\sigma^2_{\beta_0}}, \frac{1}{\sigma^2_{\beta_1}}, \ldots \frac{1}{\sigma^2_{\beta_{M-1}}}\right)$

The weighted covariance matrix, supplimented by the effect of the $\boldsymbol{\beta}$ priors is

$$\text{Cov}_W = Z^T W Z + P \tag{7}$$

The inverse of this matrix is

$$V_\beta = (\text{Cov}_W)^{-1} = (Z^T W Z + P)^{-1} \tag{8}$$

The mean (and modal) value of the posterior distribution on $\boldsymbol{\beta}$ is

$$\hat{\boldsymbol{\beta}} = V_\beta(Z^T W \mathbf{y} + P\boldsymbol{\beta}_0) \tag{9}$$

Then we define the all-important $s^2$ statistic:

$$s^2 = \frac{(\mathbf{y} - Z\hat{\boldsymbol{\beta}})^T W (\mathbf{y} - Z\hat{\boldsymbol{\beta}}) + (\boldsymbol{\beta}_0 - \hat{\boldsymbol{\beta}})^T P (\boldsymbol{\beta}_0 - \hat{\boldsymbol{\beta}})}{N} \tag{10}$$

Why is denominator $N$? It should be the familiar "effective number of datapoints minus number of parameters." Here, the number of datapoints is $N+M$ and the number of parameters is $M$.

Now we have defined all the intermediate values for our regression posteriors. Write

$$\sigma_n^2 = \frac{n_0 \sigma_0^2 + N s^2}{n_0 + N} \tag{11}$$

Then

$$\sigma^2 \mid \mathbf{Data} \sim \text{Inv-}\chi^2(n_0 + N, \sigma_n^2) \tag{12}$$

and

$$\boldsymbol{\beta} \mid (\sigma^2, \mathbf{Data}) \sim N(\hat{\boldsymbol{\beta}}, V_\beta \sigma^2) \tag{13}$$

# 5 Computational Issues

Notice that the $(\mathbf{y} - Z\hat{\boldsymbol{\beta}})^T W (\mathbf{y} - Z\hat{\boldsymbol{\beta}})$ term from Equation 10 can be computed in terms of regression matrix $Z^T W Z$, vector $Z^T W \mathbf{y}$ and value $\mathbf{y}^T W \mathbf{y}$ thusly

$$(\mathbf{y} - Z\hat{\boldsymbol{\beta}})^T W (\mathbf{y} - Z\hat{\boldsymbol{\beta}}) = \mathbf{y}^T W \mathbf{y} + \hat{\boldsymbol{\beta}}^T (Z^T W \mathbf{y}) \hat{\boldsymbol{\beta}} - 2\hat{\boldsymbol{\beta}} Z^T W \mathbf{y} \tag{14}$$

This is numerically very dangerous indeed, because we're subtracting two possibly equal-sized values from each other. But it *is* very useful to be able to define the whole regression in terms of of $Z^T W Z$, $Z^T W \mathbf{y}$ and $\mathbf{y}^T W \mathbf{y}$ without having to pass in the dataset itself.

# 6 Useful facts

The following three facts are all given explicitly on pages 480 and 481 of [1].

$$X \sim \chi_\nu^2 \leftrightarrow X \sim \text{Gamma}(\alpha = \nu/2, \beta = 1/2) \tag{15}$$

$$X \sim \chi_\nu^2 \leftrightarrow \nu s^2 / X \sim \text{Inv-}\chi^2(\nu, s^2) \tag{16}$$

$$X \sim \chi_\nu^2 \text{ and } \; Z \mid X \sim N(\mu, \omega^2 \nu / X) \leftrightarrow Z \sim t_\nu(\mu, \omega^2) \tag{17}$$

From Equations 16 and 17 we can thus see that

$$Y \sim \text{Inv-}\chi^2(\nu, s^2) \text{ and } Z \mid y \sim N(\mu, \omega^2 Y) \leftrightarrow Z \sim t_\nu(\mu, \omega^2 s^2) \tag{18}$$

(This is shown by writing $W = \nu s^2 / Y$ and plugging through).

# 7 Marginal distribution on the coefficients

The above stuff easily reveals that the marginal distribution on one of the coefficients $\beta_j$ is

$$\beta_j \mid \textbf{Data} \sim t_{(\nu=N+n_0)}(\hat{\boldsymbol{\beta}}_j, V_{\beta_{jj}} \sigma_N^2) \tag{19}$$

# 8 Posterior distribution on mean output

The above stuff also easily reveals that the marginal distribution on the mean predicted output for input $\mathbf{x}$ is

$$\bar{y}(\mathbf{x}) \mid \textbf{Data} \sim t_{(\nu=N+n_0)} \left( \hat{\boldsymbol{\beta}}^T \mathbf{z}, \mathbf{z}^T V_\beta \mathbf{z} \sigma_N^2 \right) \tag{20}$$

where $\mathbf{z} = \mathbf{z}(\mathbf{x})$ is the vector of basis functions of $\mathbf{x}$.

# 9 Posterior distribution on predicted datapoints

The above stuff also easily reveals that the marginal distribution on an actual output for input $\mathbf{x}$ is...

(I haven't worked this out but it should be easy. We add an extra lump of normal noise onto our mean output and normal plus normal gives normal. I bet it's the same as above with $\mathbf{z}^T V_\beta \mathbf{z}$ replaced by $\mathbf{z}^T V_\beta \mathbf{z} + 1$).

# 10 What are good priors?

- Things are probably much safer if things are scaled so that each term is scaled to lie in the range $-1 \leq z_{ij} \leq 1$.

- I believe that if we simply set the prior mean on $\boldsymbol{\beta}$ to zero we'll be in trouble, if, say, the range on some of the $Z$ is very different. This is because of the independence-of-coefficients assumption.

# References

[1] A. Gelman, J. B. Carlin, H. S. Stern, and D. B. Rubin. *Bayesian Data Analysis*. Chapman and Hall, 1995.