Bayesian Linear Regression

Jan Drugowitsch

Apr 2008, last update: May 2010

The Model

The model assumes a linear relation between D-dimensional inputs $\boldsymbol{x}$ and outputs y and constant-variance Gaussian noise, such that the data likelihood is given by

$$p(y|\boldsymbol{x}, \boldsymbol{w}, \tau) = \mathcal{N}(y|\boldsymbol{w}^T\boldsymbol{x}, \tau^{-1}) = \left(\frac{\tau}{2\pi}\right)^{1/2} \exp\left(-\frac{\tau}{2}(y - \boldsymbol{w}^T\boldsymbol{x})^2\right). \tag{1}$$

Given all data $\mathcal{D} = \{\boldsymbol{X}, \boldsymbol{Y}\}$, with $\boldsymbol{X} = \{\boldsymbol{x}_1, \dots, \boldsymbol{x}_N\}$ and $\boldsymbol{Y} = \{y_1, \dots, y_N\}$, the data likelihood is

$$p(\boldsymbol{Y}|\boldsymbol{X}, \boldsymbol{w}, \tau) = \prod_n p(y_n|\boldsymbol{x}_n, \boldsymbol{w}, \tau). \tag{2}$$

The prior on $\boldsymbol{w}$ and τ is conjugate normal inverse-gamma

$$\begin{aligned} p(\boldsymbol{w}, \tau|\alpha) &= \mathcal{N}(\boldsymbol{w}|0, (\tau\alpha)^{-1}\boldsymbol{I})\text{Gam}(\tau|a_0, b_0) \\ &= \left(\frac{\alpha}{2\pi}\right)^{D/2} \frac{b_0^{a_0}}{\Gamma(a_0)} \tau^{D/2+a_0-1} \exp\left(-\frac{\tau}{2}(\alpha\boldsymbol{w}^T\boldsymbol{w} + 2b_0)\right), \end{aligned} \tag{3}$$

parametrised by α. This hyper-parameter is assigned the hyper-prior

$$p(\alpha) = \text{Gam}(\alpha|c_0, d_0) = \frac{1}{\Gamma(c_0)} d_0^{c_0} \alpha^{c_0-1} \exp(-d_0\alpha). \tag{4}$$

Due to the hyper-prior, there is no analytic solution to the posteriors and variational Bayesian inference will be applied.

Variational Bayesian Inference

The variational posteriors are found by maximising the variational bound

$$\mathcal{L}(q) = \iiint q(\boldsymbol{w}, \tau, \alpha) \ln \frac{p(\boldsymbol{Y}|\boldsymbol{X}, \boldsymbol{w}, \tau)p(\boldsymbol{w}, \tau|\alpha)p(\alpha)}{q(\boldsymbol{w}, \tau, \alpha)} \mathrm{d}\boldsymbol{w}\mathrm{d}\tau\mathrm{d}\alpha \leq \ln p(\mathcal{D}), \tag{5}$$

where $p(\mathcal{D})$ is the model evidence, and under the assumption that the variational distribution $q(\boldsymbol{w}, \tau, \alpha)$, which approximates the posterior $p(\boldsymbol{w}, \tau, \alpha|\mathcal{D})$, factors into $q(\boldsymbol{w}, \tau)q(\alpha)$.

The variational posterior for $\boldsymbol{w}, \tau$ is given by

$$\begin{aligned} \ln q^*(\boldsymbol{w}, \tau) &= \ln p(\boldsymbol{Y}|\boldsymbol{X}, \boldsymbol{w}, \tau) + \mathbb{E}_\alpha(\ln p(\boldsymbol{w}, \tau|\alpha)) + \text{const.} && (6) \\ &= \left(\frac{D}{2} + a_0 - 1 + \frac{N}{2}\right) \ln \tau \\ &\quad - \frac{\tau}{2}\left(\boldsymbol{w}^T\left(\mathbb{E}_\alpha(\alpha)\boldsymbol{I} + \sum_n \boldsymbol{x}_n\boldsymbol{x}_n^T\right)\boldsymbol{w} + \sum_n y_n^2 - 2\boldsymbol{w}^T\sum_n \boldsymbol{x}_n y_n + 2b_0\right) + \text{const} && (7) \\ &= \ln \mathcal{N}(\boldsymbol{w}|\boldsymbol{w}_N, \tau^{-1}\boldsymbol{V}_N)\text{Gam}(\tau|a_N, b_N), && (8) \end{aligned}$$

with

$$\boldsymbol{V}_N^{-1} = \mathbb{E}_\alpha(\alpha)\boldsymbol{I} + \sum_n \boldsymbol{x}_n \boldsymbol{x}_n^T, \quad (9)$$

$$\boldsymbol{w}_N = \boldsymbol{V}_N \sum_n \boldsymbol{x}_n y_n, \quad (10)$$

$$a_N = a_0 + \frac{N}{2}, \quad (11)$$

$$\begin{aligned} b_N &= b_0 + \frac{1}{2}\left(\sum_n y_n^2 - \boldsymbol{w}_N^T \boldsymbol{V}_N^{-1} \boldsymbol{w}_N\right) \\ &= b_0 + \frac{1}{2}\left(\sum_n (y_n - \boldsymbol{w}_N^T \boldsymbol{x}_n)^2 + \mathbb{E}_\alpha(\alpha)\boldsymbol{w}_N^T \boldsymbol{w}_N\right). \end{aligned} \quad (12)$$

The variational posterior for α is

$$\ln q^*(\alpha) = \mathbb{E}_{\boldsymbol{w},\tau}(\ln p(\boldsymbol{w}, \tau|\alpha)) + \ln p(\alpha) + \text{const.} \quad (13)$$

$$= \left(c_0 - 1 + \frac{D}{2}\right)\ln\alpha - \alpha\left(d_0 + \frac{1}{2}\mathbb{E}_{\boldsymbol{w},\tau}(\tau \boldsymbol{w}^T \boldsymbol{w})\right) + \text{const.} \quad (14)$$

$$= \ln \text{Gam}(\alpha|c_N, d_N), \quad (15)$$

with

$$c_N = c_0 + \frac{D}{2}, \quad (16)$$

$$d_N = d_0 + \frac{1}{2}\mathbb{E}_{\boldsymbol{w},\tau}(\tau \boldsymbol{w}^T \boldsymbol{w}). \quad (17)$$

The expectations are evaluated with respect to the variational distribution and are given by

$$\mathbb{E}_{\boldsymbol{w},\tau}(\tau \boldsymbol{w}^T \boldsymbol{w}) = \frac{a_N}{b_N}\boldsymbol{w}_N^T \boldsymbol{w}_N + \text{Tr}(\boldsymbol{V}_N), \quad (18)$$

$$\mathbb{E}_\alpha(\alpha) = \frac{c_N}{d_N}. \quad (19)$$

The variational bound itself consists of

$$\begin{aligned} \mathcal{L}(q) &= \mathbb{E}_{\boldsymbol{w},\tau}(\ln p(\boldsymbol{Y}|\boldsymbol{X}, \boldsymbol{w}, \tau)) + \mathbb{E}_{\boldsymbol{w},\tau,\alpha}(\ln p(\boldsymbol{w}, \tau|\alpha)) + \mathbb{E}_\alpha(\ln p(\alpha)) \\ &\quad - \mathbb{E}_{\boldsymbol{w},\tau}(\ln p(\boldsymbol{w}, \tau)) - \mathbb{E}_\alpha(\ln p(\alpha)), \end{aligned} \quad (20)$$

$$\begin{aligned} \mathbb{E}_{\boldsymbol{w},\tau}(\ln p(\boldsymbol{Y}|\boldsymbol{X}, \boldsymbol{w}, \tau)) &= \frac{N}{2}\left(\psi(a_N) - \ln b_N - \ln 2\pi\right) \\ &\quad - \frac{1}{2}\sum_n \left(\frac{a_N}{b_N}(y_n - \boldsymbol{w}_N^T \boldsymbol{x}_n)^2 + \boldsymbol{x}_n^T \boldsymbol{V}_N \boldsymbol{x}_n\right), \end{aligned} \quad (21)$$

$$\begin{aligned} \mathbb{E}_{\boldsymbol{w},\tau,\alpha}(\ln p(\boldsymbol{w}, \tau|\alpha)) &= \frac{D}{2}\left(\psi(a_N) - \ln b_N + \psi(c_N) - \ln d_N - \ln 2\pi\right) \\ &\quad - \frac{1}{2}\frac{c_N}{d_N}\left(\frac{a_N}{b_N}\boldsymbol{w}_N^T \boldsymbol{w}_N + \text{Tr}(\boldsymbol{V}_N)\right) \\ &\quad - \ln\Gamma(a_0) + a_0 \ln b_0 + (a_0 - 1)(\psi(a_N) - \ln b_N) - b_0 \frac{a_N}{b_N}, \end{aligned} \quad (22)$$

$$\mathbb{E}_\alpha(\ln p(\alpha)) = -\ln\Gamma(c_0) + d_0 \ln c_0 + (c_0 - 1)(\psi(c_N) - \ln d_N) - d_0 \frac{c_N}{d_N}, \quad (23)$$

$$\begin{aligned} \mathbb{E}_{\boldsymbol{w},\tau}(\ln q(\boldsymbol{w}, \tau)) &= \frac{D}{2}(\psi(a_N) - \ln b_N - \ln 2\pi - 1) - \frac{1}{2}\ln|\boldsymbol{V}_N| \\ &\quad - \ln\Gamma(a_N) + a_N \ln b_N + (a_N - 1)(\psi(a_N) - \ln b_N) - a_N, \end{aligned} \quad (24)$$

$$\mathbb{E}_\alpha(\ln q(\alpha)) = -\ln\Gamma(c_N) + (c_N - 1)\psi(c_N) + \ln d_N - c_N. \quad (25)$$

In combination, that gives

$$\begin{aligned}\mathcal{L}(q) \quad = \quad & -\frac{N}{2}\ln 2\pi - \frac{1}{2}\sum_n \left(\frac{a_N}{b_N}(y_n - \boldsymbol{w}_N^T\boldsymbol{x}_n)^2 + \boldsymbol{x}_n^T\boldsymbol{V}_N\boldsymbol{x}_n\right) + \frac{1}{2}\ln|\boldsymbol{V}_N| + \frac{D}{2} \\ & -\ln\Gamma(a_0) + a_0\ln b_0 - b_0\frac{a_N}{b_N} + \ln\Gamma(a_N) - a_N\ln b_N + a_N \\ & -\ln\Gamma(c_0) + c_0\ln d_0 + \ln\Gamma(c_N) - c_N\ln d_N \end{aligned} \tag{26}$$

This bound is maximised by iterating over the updates for $\boldsymbol{V}_N$, $\boldsymbol{w}_N$, a_N, b_N, c_N, and d_N until $\mathcal{L}(q)$ reaches a plateau.

Predictive Density

The predictive density is evaluated by approximating the posterior $p(\boldsymbol{w},\tau|\mathcal{D})$ by its variational counterpart $q(\boldsymbol{w},\tau)$, to get

$$p(y|\boldsymbol{x},\mathcal{D}) = \iint p(y|\boldsymbol{x},\boldsymbol{w},\tau)p(\boldsymbol{w},\tau|data)\mathrm{d}\boldsymbol{w}\mathrm{d}\tau \tag{27}$$

$$\approx \iint p(y|\boldsymbol{x},\boldsymbol{w},\tau)q(\boldsymbol{w},\tau)\mathrm{d}\boldsymbol{w}\mathrm{d}\tau \tag{28}$$

$$= \iint \mathcal{N}(y|\boldsymbol{w}^T\boldsymbol{x},\tau^{-1})\mathcal{N}(\boldsymbol{w}|\boldsymbol{w}_N,\tau^{-1}\boldsymbol{V}_N)\mathrm{Gam}(\tau|a_N,b_N)\mathrm{d}\boldsymbol{w}\mathrm{d}\tau \tag{29}$$

$$= \int \mathcal{N}(y|\boldsymbol{w}^T\boldsymbol{x},\tau^{-1}(1+\boldsymbol{x}^T\boldsymbol{V}_N\boldsymbol{x}))\mathrm{Gam}(\tau|a_N,b_N)\mathrm{d}\tau \tag{30}$$

$$= \mathrm{St}\left(y|\boldsymbol{w}^T\boldsymbol{x},(1+\boldsymbol{x}^T\boldsymbol{V}_N\boldsymbol{x})^{-1}\frac{a_N}{b_N},2a_N\right), \tag{31}$$

where standard results of convolving Gaussians with other Gaussians and Gamma distributions where used. The resulting distribution is a Student's t distribution with mean $\boldsymbol{w}^T\boldsymbol{x}$, precision $(1+\boldsymbol{x}^T\boldsymbol{V}_N\boldsymbol{x})^{-1}a_N/b_N$, and $2a_N$ degrees of freedom, which has a variance of $(1+\boldsymbol{x}^T\boldsymbol{V}_N\boldsymbol{x})b_N/(a_N-1)$.

Using Automatic Relevance Determination

Automatic Relevance Determination (ARD) determines the relevance of the elements of the input to determine the output by assigning a separate shrinkage prior to each element of the weight vector, which is in turn adjusted by a hyper-prior. While the data likelihood remains unchanged, the prior on $\boldsymbol{w},\tau$ is modified to be

$$\begin{aligned}p(\boldsymbol{w},\tau|\boldsymbol{\alpha}) &= \mathcal{N}(\boldsymbol{w}|\boldsymbol{0},(\tau\boldsymbol{A})^{-1})\mathrm{Gam}(\tau|a_0,b_0) \\ &= \frac{|\boldsymbol{A}|^{1/2}}{\sqrt{2\pi}^D}\frac{b_0^{a_0}}{\Gamma(a_0)}\tau^{D/2+a_0-1}\exp\left(-\frac{\tau}{2}(\boldsymbol{w}^T\boldsymbol{A}\boldsymbol{w}+2b_0)\right),\end{aligned} \tag{32}$$

where the vector $\boldsymbol{\alpha} = (\alpha_1,\dots,\alpha_D)^T$ forms the diagonal of $\boldsymbol{A}$. All of the α's are independent, such that the hyper-prior is given by

$$p(\boldsymbol{\alpha}) = \prod_i \mathrm{Gam}(\alpha_i|c_0,d_0) = \prod_i \frac{1}{\Gamma(c_0)}d_0^{c_0}\alpha_i^{c_0-1}\exp(-d_0\alpha_i). \tag{33}$$

Variational Bayesian inference is performed as before, to get the variational posteriors

$$q^*(\boldsymbol{w},\tau) = \mathcal{N}(\boldsymbol{w}|\boldsymbol{w}_N,\tau^{-1}\boldsymbol{V}_N)\mathrm{Gam}(\tau|a_N,b_N), \qquad q^*(\boldsymbol{\alpha}) = \prod_i \mathrm{Gam}(\alpha_i|c_N,d_{Ni}), \tag{34}$$

with

$$\begin{aligned}
\boldsymbol{V}_N^{-1} &= \mathbb{E}_\alpha(\boldsymbol{A}) + \sum_n \boldsymbol{x}_n \boldsymbol{x}_n^T, && (35)\\
\boldsymbol{w}_N &= \boldsymbol{V}_N \sum_n \boldsymbol{x}_n y_n, && (36)\\
a_N &= a_0 + \frac{N}{2} && (37)\\
b_N &= b_0 + \frac{1}{2}\left(\sum_n y_n^2 - \boldsymbol{w}_N^T \boldsymbol{V}_N^{-1} \boldsymbol{w}_N\right) \\
&= b_0 + \frac{1}{2}\left(\sum_n (\boldsymbol{w}_N^T \boldsymbol{x}_N - y_n)^2 + \boldsymbol{w}_N^T \mathbb{E}_\alpha(\boldsymbol{A}) \boldsymbol{w}_N\right), && (38)\\
c_N &= c_0 + \frac{1}{2}, && (39)\\
d_{Ni} &= d_0 + \frac{1}{2}\mathbb{E}_{\boldsymbol{w},\alpha}(\tau w_i^2), && (40)
\end{aligned}$$

with expectations $\mathbb{E}_{\boldsymbol{w},\alpha}(\tau w_i^2) = w_{Ni}^2 a_N / b_N + (\boldsymbol{V}_N)_{ii}$, and $\mathbb{E}_\alpha(\boldsymbol{A}) = \boldsymbol{A}_N$ is a diagonal matrix with elements $\mathbb{E}_\alpha(\alpha_i) = c_N / d_{Ni}$.

The variational bound changes to

$$\begin{aligned}
\mathcal{L}(q) &= -\frac{N}{2}\ln 2\pi - \frac{1}{2}\sum_n \left(\frac{a_N}{b_N}(y_n - \boldsymbol{w}_N^T \boldsymbol{x}_n)^2 + \boldsymbol{x}_n^T \boldsymbol{V}_N \boldsymbol{x}_n\right) + \frac{1}{2}\ln|\boldsymbol{V}_N| + \frac{D}{2} \\
&\quad - \ln\Gamma(a_0) + a_0 \ln b_0 - b_0 \frac{a_N}{b_N} + \ln\Gamma(a_N) - a_N \ln b_N + a_N \\
&\quad + \sum_i \left(-\ln\Gamma(c_0) + c_0 \ln d_0 + \ln\Gamma(c_N) - c_N \ln d_{Ni}\right). && (41)
\end{aligned}$$

The predictive distribution remains unchanged, as the prior does not appear in the expression for the variational posterior $p(\boldsymbol{w}, \tau)$.

Implementation

The scripts bayes_linear_fit .m and bayes_linear_fit_ard .m are straightforward implementations that compute the posterior parameters, without and with ARD, respectively. They operate by iteratively updating the parameters of $q^*(\boldsymbol{w}, \tau)$ and $q^*(\alpha)$, while monitoring $\mathcal{L}(q)$. The scripts stop as soon as either the change in $\mathcal{L}(q)$ between two consecutive iterations drops below 0.001% or the number of iterations exceeds 100.

The code is vectorised in order to speed up computation. In particular, the input X is assumed to be an $N \times D$ matrix, with $\boldsymbol{x}_n^T$ as its rows. y is a column vector, containing all y_n's. This allows for several vectorised operations, such as

$$\begin{aligned}
\sum_n \boldsymbol{x}_n \boldsymbol{x}_n^T &= \mathsf{X}' * \mathsf{X}, && (42)\\
\sum_n \boldsymbol{x}_n y_n &= \mathsf{X}' * \mathsf{y}, && (43)\\
\boldsymbol{w}_N^T \boldsymbol{x}_n &= (\mathsf{X} * \mathsf{w})_n, && (44)\\
\sum_n \boldsymbol{x}_n^T \boldsymbol{V}_N \boldsymbol{x}_n &= \textbf{sum}(\textbf{sum}(\mathsf{X} .* (\mathsf{X} * \mathsf{V}))). && (45)
\end{aligned}$$

The rest of the code should be self-explanatory. The only mayor difference between bayes_linear_fit .m and bayes_linear_fit_ard .m is that in the latter version, the variables E_a, dn, and E_t are vectors, and all operations on these variables are vectorised.

The script bayes_linear_post.m computes the predictive density parameters for a set of input vectors, again given in matrix form X.