Regression, Bayesian Regression, and Emperical Bayes Regression

Manish Agarwal

July 2025

1 Regression

The *linear regression* model involves linear combination of the input variables

$$y(\boldsymbol{x}, \boldsymbol{w}) = w_0 + w_1 x_1 + \cdots + w_D x_D.$$

where $\boldsymbol{x} = (x_1, \ldots, x_D)^T$. This can be extended by considering linear combinations of fixed nonlinear functions of the input variables, of the form

$$y(\boldsymbol{x}, \boldsymbol{w}) = w_0 + \sum_{j=1}^{M-1} w_j \phi_j(\boldsymbol{x})$$

where $\phi_j(\boldsymbol{x})$ are known as the *basis functions*. The total number of parameters here is M. w_0 is the fixed offset to the data called *bias* and can be subsumed by defining the dummy basis function $\phi_0(\boldsymbol{x}) = 1$ so that

$$y(\boldsymbol{x}, \boldsymbol{w}) = \sum_{j=0}^{M-1} w_j \phi_j(\boldsymbol{x}) = \boldsymbol{w}^T \boldsymbol{\phi}(\boldsymbol{x})$$

where $\boldsymbol{w} = (w_0, \ldots, w_{M-1})^T$ and $\boldsymbol{\phi} = (\phi_0, \ldots, \phi_{M-1})^T$. Examples of basis functions are polynomial $\phi_j(x) = x^j$, splines, Gaussian $\phi_j(x) = \exp\left(-\frac{(x-\mu_j)^2}{2s^2}\right)$, sigmoid $\phi_j(x) = \sigma\left(\frac{x-\mu_j}{s}\right)$ with $\sigma(a) = \frac{1}{1+e^{-a}}$, Fourier, wavelets.

We assume that the target variable t is given by a deterministic function $y(\boldsymbol{x}, \boldsymbol{w})$ with additive Gaussian noise so that

$$t = y(\boldsymbol{x}, \boldsymbol{w}) + \epsilon$$

where ϵ is a zero mean Gaussian random variable with precision (inverse variance) β. Thus we can write

$$p(t|\boldsymbol{x}, \boldsymbol{w}, \beta) = \mathcal{N}(t|y(\boldsymbol{x}, \boldsymbol{w}), \beta^{-1}).$$

Note that the conditional mean is $\mathbb{E}[t|\boldsymbol{x}] = \int tp(t|\boldsymbol{x})dt = y(\boldsymbol{x}, \boldsymbol{w})$. The Gaussian noise assumption implies the distribution of $t|\boldsymbol{x}$ to be unimodal (if in appropriate for some application, mixtures of conditional Gaussian distributions, which permits multi-modal conditional distributions, could be used).

The data set of inputs $\boldsymbol{X} = \{\boldsymbol{x}_1, \ldots, \boldsymbol{x}_N\}$ has corresponding targets $t_1, \ldots, T_N$. We group the target variables $\{t_n\}$ into a column vector $\boldsymbol{t}$. Making the assumption that these data points are drawn independently from the normal distribution above, we obtain the following expression for the likelihood function (with parameters $\boldsymbol{w}$ and β)

$$p(\boldsymbol{t}|\boldsymbol{X}, \boldsymbol{w}, \beta) = \prod_{n=1}^{N} \mathcal{N}(t_n|\boldsymbol{w}^T \boldsymbol{\phi}(\boldsymbol{x}_n), \beta^{-1}).$$

Since we are not modeling the distribution of input variable in a supervised learning problem, we can drop $\boldsymbol{x}$ from our expressions. Taking the logarithm of the likelihood function we have

$$\ln p(\boldsymbol{t}|\boldsymbol{w}, \beta) = \frac{N}{2} \ln \beta - \frac{N}{2} \ln 2\pi - \beta E_D(\boldsymbol{w})$$

where

$$E_D(\boldsymbol{w}) = \frac{1}{2} \sum_{n=1}^{N} (t_n - \boldsymbol{w}^T \boldsymbol{\phi}(\boldsymbol{x}_n))^2.$$

We can use maximum likelihood to determine $\boldsymbol{w}$ and β. This gives

$$\boldsymbol{w}_{ML} = (\boldsymbol{\Phi}^T\boldsymbol{\Phi})^{-1}\boldsymbol{\Phi}^T\boldsymbol{t} = \boldsymbol{\Phi}^\dagger\boldsymbol{t}$$

and

$$\frac{1}{\beta_{ML}} = \frac{1}{N}\sum_{n=1}^{N}(t_n - \boldsymbol{w}_{ML}^T\boldsymbol{\phi}(\boldsymbol{x}_n))^2.$$

Thus, the inverse of the noise precision is given by the residual variance of the target values around the regression function.

2 Bayesian Regression

The effective model complexity, governed by the number of basis functions, needs to be controlled according to the size of the data set. Adding a regularization term controls this, as it can't be decided by maximizing the likelihood function, because it always leads to excessively complex models and over-fitting. One could use cross-validation to determine it, but it is both computationally expensive and wasteful of the valuable data. Bayesian treatment of linear regression avoids the over-fitting problem and leads to automatic methods of determining model complexity using the training data alone.

We introduce a prior probability distribution over the model parameters $\boldsymbol{w}$. To start with, we treat β to be a known constant. The conjugate prior to the exponential of the quadratic, as we found before, is a Gaussian distribution of the form

$$p(\boldsymbol{w}) = \mathcal{N}(\boldsymbol{w}|\boldsymbol{m}_0, \boldsymbol{S}_0)$$

having mean $\boldsymbol{m}_0$ and covariance $\boldsymbol{S}_0$. The posterior distribution can be written as

$$p(\boldsymbol{w}|\boldsymbol{t}) = \mathcal{N}(\boldsymbol{w}|\boldsymbol{m}_N, \boldsymbol{S}_N)$$

where

$$\begin{aligned}\boldsymbol{m}_N &= \boldsymbol{S}_N(\boldsymbol{S}_0^{-1}\boldsymbol{m}_0 + \beta\boldsymbol{\Phi}^T\boldsymbol{t}) \\ \boldsymbol{S}_N^{-1} &= \boldsymbol{S}_0^{-1} + \beta\boldsymbol{\Phi}^T\boldsymbol{\Phi}.\end{aligned}$$

Now we consider a particular parsimonious Gaussian prior parametrized by parameter α

$$p(\boldsymbol{w}|\alpha) = \mathcal{N}(\boldsymbol{w}|\boldsymbol{0}, \alpha^{-1}\boldsymbol{I})$$

giving the corresponding posterior distribution over $\boldsymbol{w}$ as

$$\boldsymbol{m}_N = \beta\boldsymbol{S}_N\boldsymbol{\Phi}^T\boldsymbol{t}$$

$$\boldsymbol{S}_N^{-1} = \alpha\boldsymbol{I} + \beta\boldsymbol{\Phi}^T\boldsymbol{\Phi}.$$

The log of the posterior distribution is given by

$$\ln p(\boldsymbol{w}|\boldsymbol{t}) = -\frac{\beta}{2}\sum_{n=1}^{N}(t_n - \boldsymbol{w}^T\boldsymbol{\phi}(\boldsymbol{x}_n))^2 - \frac{\alpha}{2}\boldsymbol{w}^T\boldsymbol{w} + \text{const}.$$

Maximizing this posterior distribution with respect to $\boldsymbol{w}$ is equivalent to minimizing sum-of-squares error with addition of a quadratic regularization term, with $\lambda = \frac{\alpha}{\beta}$.

Other forms of prior over the parameter can be considered. For instance, we can generalize the Gaussian prior to give

$$p(\boldsymbol{w}|\alpha) = \left(\frac{q}{2}\left(\frac{\alpha}{2}\right)^{1/q}\frac{1}{\Gamma(1/q)}\right)^M \exp\left(-\frac{\alpha}{2}\sum_{j=1}^{M}|w_j|^q\right)$$

in which $q = 2$ corresponds to the Gaussian distribution. Finding the maximum of the posterior distribution over $\boldsymbol{w}$ corresponds to minimization of the regularized error function given by:

$$\frac{1}{2}\sum_{n=1}^{N}(t_n - \boldsymbol{w}^T\boldsymbol{\phi}(\boldsymbol{x}_n))^2 + \frac{\lambda}{2}\sum_{j=1}^{M}|w_j|^q.$$

The case of $q = 1$ is known as **lasso** and $q = 2$ as **ridge** regression. Figure 1 shows contours of the regularization function for different values of q.

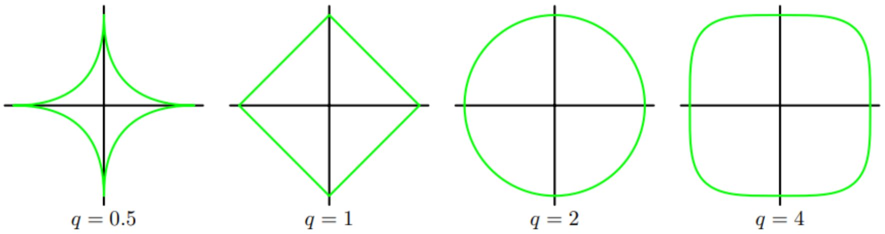

Figure 1: Contours of the regularization term for various values of the parameter q.

In practice we are interested in making predictions of t for new values of $\boldsymbol{x}$. This requires the predictive distribution to be evaluated defined by

$$p(t|\boldsymbol{t}, \alpha, \beta) = \int p(t|\boldsymbol{w}, \beta)p(\boldsymbol{w}|\boldsymbol{t}, \alpha, \beta)d\boldsymbol{w}$$

in with $\boldsymbol{t}$ is the vector of target values from the training set, and we have omitted the corresponding input vectors from the right-hand side of the conditioning statements to simplify the notation. The prediction distribution takes the form

$$p(t|\boldsymbol{x}, \boldsymbol{t}, \alpha, \beta) = \mathcal{N}(t|\boldsymbol{m}_N^T\boldsymbol{\phi}(\boldsymbol{x}), \sigma_N^2(\boldsymbol{x})) \qquad \sigma_N^2(\boldsymbol{x}) = \frac{1}{\beta} + \boldsymbol{\phi}(\boldsymbol{x})^T\boldsymbol{S}_N\boldsymbol{\phi}(\boldsymbol{x}).$$

The first term represents the noise on the data whereas the second term reflects the uncertainty associated with the parameters $\boldsymbol{w}$, and are additive in variance because the noise processes are independent Gaussians. In the limit of $N \to \infty$ the second term goes to 0.

If both $\boldsymbol{w}$ and β are treated as unknown, we can introduce a conjugate prior distribution $p(\boldsymbol{w}, \beta)$, a Gaussian-Gamma distribution, making the predictive distribution a Sutdent's t-distribution.

3 Empirical Bayes Regression

We introduce prior distributions over the hyperparameter α and β and make predictions by marginalizing with respect to these hyperparameters as well as with respect to the parameters $\boldsymbol{w}$.

$$p(t|\boldsymbol{t}) = \int\int\int p(t|\boldsymbol{w}, \beta)p(\boldsymbol{w}|\boldsymbol{t}, \alpha, \beta)p(\alpha, \beta|\boldsymbol{t})\ d\boldsymbol{w}\ d\alpha\ d\beta$$

where

$$p(\alpha, \beta|\boldsymbol{t}) \propto p(\boldsymbol{t}|\alpha, \beta)p(\alpha, \beta).$$

But this is analytically intractable. Instead, we set the hyperparameters to specific values determined by maximizing the marginal likelihood function obtained by first integrating over the parameter $\boldsymbol{w}$. This is called **empirical Bayes**. This will allow use to determine values for these hyperparameters from the training data alone, without recourse to cross-validation.

Integrating over the weight parameter $\boldsymbol{w}$ we get the marginal likelihood function (evidence function)

$$p(\boldsymbol{t}|\alpha, \beta) = \int p(\boldsymbol{t}|\boldsymbol{w}, \beta)p(\boldsymbol{w}|\alpha)\ d\boldsymbol{w} = \left(\frac{\beta}{2\pi}\right)^{N/2}\left(\frac{\alpha}{2\pi}\right)^{M/2}\int e^{-E(\boldsymbol{w})}\ d\boldsymbol{w}$$

where M is the dimensionality of $\boldsymbol{w}$, and we have defined

$$E(\boldsymbol{w}) = \beta E_D(\boldsymbol{w}) + \alpha E_W(\boldsymbol{w}) = \frac{\beta}{2}\|\boldsymbol{t} - \boldsymbol{\Phi}\boldsymbol{w}\|^2 + \frac{\alpha}{2}\boldsymbol{w}^T\boldsymbol{w}.$$

We complete the square over $\boldsymbol{w}$ to get

$$E(\boldsymbol{w}) = E(\boldsymbol{m}_N) + \frac{1}{2}(\boldsymbol{w} - \boldsymbol{m}_N)^T\boldsymbol{A}(\boldsymbol{w} - \boldsymbol{m}_N)$$

where $\boldsymbol{A} = \alpha \boldsymbol{I} + \beta \boldsymbol{\Phi}^T \boldsymbol{\Phi} = \nabla^2 E(\boldsymbol{w}) = \boldsymbol{S}_N^{-1}$ (Hessian matrix), $\boldsymbol{m}_N = \beta \boldsymbol{A}^{-1} \boldsymbol{\Phi}^T \boldsymbol{t}$ (mean of the posterior distribution) and $E(\boldsymbol{m}_N) = \frac{\beta}{2} \|\boldsymbol{t} - \boldsymbol{\Phi} \boldsymbol{m}_N\|^2 + \frac{\alpha}{2} \boldsymbol{m}_N^T \boldsymbol{m}_N$. We can now integrate over $\boldsymbol{w}$ to get

$$\int e^{-E(\boldsymbol{w})} \, d\boldsymbol{w} = e^{-E(\boldsymbol{m}_N)} \int e^{-\frac{1}{2}(\boldsymbol{w}-\boldsymbol{m}_N)^T \boldsymbol{A} (\boldsymbol{w}-\boldsymbol{m}_N)} \, d\boldsymbol{w} = e^{-E(\boldsymbol{m}_N)} (2\pi)^{M/2} |\boldsymbol{A}|^{-1/2}.$$

The log of marginal likelihood is

$$\ln p(\boldsymbol{t}|\alpha, \beta) = \frac{M}{2} \ln \alpha + \frac{N}{2} \ln \beta - E(\boldsymbol{m}_N) - \frac{1}{2} \ln |\boldsymbol{A}| - \frac{N}{2} \ln(2\pi).$$

Solving for $\boldsymbol{\alpha}$: We define the following eigenvector equation

$$(\beta \boldsymbol{\Phi}^T \boldsymbol{\Phi}) \boldsymbol{u}_i = \lambda_i \boldsymbol{u}_i.$$

Thus, $\boldsymbol{A}$ has eigenvalues $\alpha + \lambda_i$. Now, $\frac{d}{d\alpha} \ln |\boldsymbol{A}| = \frac{d}{d\alpha} \ln \prod_i (\lambda_i + \alpha) = \frac{d}{d\alpha} \sum_i \ln(\lambda_i + \alpha) = \sum_i \frac{1}{\lambda_i + \alpha}$. Thus the stationary point with respect to α satisfies

$$0 = \frac{M}{2\alpha} - \frac{1}{2} \boldsymbol{m}_N^T \boldsymbol{m}_N - \frac{1}{2} \sum_i \frac{1}{\lambda_i + \alpha} \implies \alpha \boldsymbol{m}_N^T \boldsymbol{m}_N = M - \alpha \sum_i \frac{1}{\lambda_i + \alpha} = \sum_i \frac{\lambda_i}{\lambda_i + \alpha} = \gamma.$$

Since $\boldsymbol{m}_N$ and γ both depend on α we get the implicit solution obtained by iterations

$$\alpha = \frac{\gamma}{\boldsymbol{m}_N^T \boldsymbol{m}_N}.$$

Solving for $\boldsymbol{\beta}$: We first note $\frac{d\lambda_i}{d\beta} = \frac{\lambda_i}{\beta}$ (because λ_i is proportional to β) giving $\frac{d}{d\beta} \ln |\boldsymbol{A}| = \frac{d}{d\beta} \sum_i \ln(\lambda_i + \alpha) = \frac{1}{\beta} \sum_i \frac{\lambda_i}{\lambda_i + \alpha} = \frac{\gamma}{\beta}$. The stationary point satisfies

$$0 = \frac{N}{2\beta} - \frac{1}{2} \sum_{n=1}^{N} (t_n - \boldsymbol{m}_N^T \boldsymbol{\phi}(\boldsymbol{x}_n))^2 - \frac{\gamma}{2\beta} \implies \frac{1}{\beta} = \frac{1}{N - \gamma} \sum_{n=1}^{N} (t_n - \boldsymbol{m}_N^T \boldsymbol{\phi}(\boldsymbol{x}_n))^2.$$

This again is an implicit solution for β and is solved by iterations. If both α and β are to be determined from the data, then their values can be re-estimated together after each update of γ.

Since $\beta \boldsymbol{\Phi}^T \boldsymbol{\Phi}$ is positive definite the values of $\frac{\lambda_i}{\lambda_i + \alpha}$ lie between 0 and 1, and consequently $0 \leq \gamma \leq M$. For directions in which $\lambda)i \gg \alpha$, the corresponding parameter w_i will be close to its maximum likelihood value, and the ratio $\frac{\lambda_i}{\lambda_i + \alpha}$ will be close to 1, which is a *well determined* parameters as it aligns with the data. Conversely, for directions $\lambda_i \ll \alpha$, the corresponding w_i will be close to zero, as will the ratio $\frac{\lambda_i}{\lambda_i + \alpha}$. In these directions the likelihood is relatively insensitive to the parameter value. The quantity γ therefore measures the *effective total number of well determined parameters*. The expression for β express the inverse precision as an average of the squared differences between the targets and the model predictions. The average by $N - \gamma$ again reflects the effective number of parameters bias correction.

3.1 Relevance Vector Machines

We now impose a sparse kernel prior called **relevance vector machines**. Here we introduce a separate hyperparameter α_i for each of the weight parameter w_i instead of a single shared hyperparameter. Thus the weight prior takes the form

$$p(\boldsymbol{w}|\boldsymbol{\alpha}) = \prod_{i=1}^{N} \mathcal{N}(w_i|0, \alpha_i^{-1})$$

where $\boldsymbol{\alpha} = (\alpha_1, \ldots, \alpha_M)^T$. We shall see that, a significant proportion of α_i go to infinity, rendering the corresponding weights zero and induce sparsity. To recap, the posterior distribution (for linear regression) for the weights is again Gaussian of the form

$$p(\boldsymbol{w}|\boldsymbol{t}, \boldsymbol{X}, \boldsymbol{\alpha}, \beta) = \mathcal{N}(\boldsymbol{w}|\boldsymbol{m}, \boldsymbol{\Sigma})$$

where the mean and covariance are

$$\boldsymbol{m} = \beta \boldsymbol{\Sigma} \boldsymbol{\Phi}^T \boldsymbol{t} \qquad \boldsymbol{\Sigma} = (\boldsymbol{A} + \beta \boldsymbol{\Phi}^T \boldsymbol{\Phi})^{-1}$$

where $\boldsymbol{\Phi}$ is a $N \times M$ design matrix with elements $\Phi_{ni} = \phi_i(\boldsymbol{x}_n)$, and $\boldsymbol{A} = diag(\alpha_i)$. The values of $\boldsymbol{\alpha}$ and β are determined using evidence approximation (empirical Bayes) as before, in which we maximize the marginal likelihood function obtained by integrating out the weight parameters

$$p(\boldsymbol{t}|\boldsymbol{X}, \boldsymbol{\alpha}, \beta) = \int p(\boldsymbol{t}|\boldsymbol{X}, \boldsymbol{w}, \beta) p(\boldsymbol{w}|\boldsymbol{\alpha}) d\boldsymbol{w}.$$

This is readily evaluated to give the log marginal likelihood in the form

$$\ln p(\boldsymbol{t}|\boldsymbol{X}, \boldsymbol{\alpha}, \beta) = \ln \mathcal{N}(\boldsymbol{t}|\boldsymbol{0}, \boldsymbol{C}) = -\frac{1}{2}(N \ln 2\pi + \ln |\boldsymbol{C}| + \boldsymbol{t}^T \boldsymbol{C}^{-1} \boldsymbol{t})$$

where $\boldsymbol{t} = (t_1, \dots, t_N)^T$ and we defined the $N \times N$ matrix

$$\boldsymbol{C} = \beta^{-1} \boldsymbol{I} + \boldsymbol{\Phi} \boldsymbol{A}^{-1} \boldsymbol{\Phi}^T.$$

We set the required derivatives of the marginal likelihood to zero and obtain the following re-estimation equations

$$\alpha_i = \frac{\gamma_i}{m_i^2}$$
$$\frac{1}{\beta} = \frac{\|\boldsymbol{t} - \boldsymbol{\Phi} \boldsymbol{m}\|^2}{N - \sum_i \gamma_i}$$

where m_i is the i-th component of the posterior mean $\boldsymbol{m}$. The quantity γ_i measures how well the corresponding parameter w_i is determined by the data and is defined by

$$\gamma_i = 1 - \alpha_i \Sigma_{ii}$$

where Σ_{ii} is the i-th diagonal component of the posterior covariance $\boldsymbol{\Sigma}$. Iteration until convergence would yield the desired values. A proportion of the hyperparameters $\{\alpha_i\}$ are driven to large values, and so the weight parameters w_i have posterior distributions with mean and variance both zero. Thus those parameters, and the corresponding basis functions $\phi_i(\boldsymbol{x})$, are removed from the model and play no role in making predictions for new inputs. The inputs corresponding to remaining nonzero weights are called *relevance vectors*.

Having found values $\boldsymbol{\alpha}^*$ and β^* for the hyperparameters that maximize the marginal likelihood, we can evaluate the predictive distribution over t for a new input $\boldsymbol{x}$, given by

$$p(t|\boldsymbol{x}, \boldsymbol{X}, \boldsymbol{t}, \boldsymbol{\alpha}^*, \beta^*) = \int p(t|\boldsymbol{x}, \boldsymbol{w}, \beta^*) p(\boldsymbol{w}|\boldsymbol{X}, \boldsymbol{t}, \boldsymbol{\alpha}^*, \beta^* \; d\boldsymbol{w} = \mathcal{N}(t|\boldsymbol{m}^T \boldsymbol{\phi}(\boldsymbol{x}), \sigma^2(\boldsymbol{x})).$$

Thus the predictive mean is given by the same formula above with $\boldsymbol{w}$ set to the posterior mean $\boldsymbol{m}$, and the variance of the predictive distribution is given by $\sigma^2(\boldsymbol{x}) = \frac{1}{\beta^*} + \boldsymbol{\phi}(\boldsymbol{x})^T \boldsymbol{\Sigma} \boldsymbol{\phi}(\boldsymbol{x})$, where $\boldsymbol{\Sigma}$. For localized basis functions, the predictive variance for linear regression models becomes small in regions of input space where there are no basis functions, that is the model will become increasingly certain of its predictions when extrapolating outside the domain of the data, which of course is undesirable (Gaussian process does not suffer from this problem but is computationally expensive).

Sparsity arise because any finite value of α will always assign a lower probability to the data if the direction of $\phi(\boldsymbol{x})$ is misaligned with training data vector $\boldsymbol{t}$, provided that β is set to its optimal value. For M basis functions, we first make explicit all the dependence of the marginal likelihood on a particular α_i and then determine its stationary points explicitly. We pull out the contribution from α_i in the matrix $\boldsymbol{C}$ to give

$$\boldsymbol{C} = \beta^{-1} \boldsymbol{I} + \sum_{j \neq i} \alpha_j^{-1} \boldsymbol{\varphi}_j \boldsymbol{\varphi}_j^T + \alpha_i^{-1} \boldsymbol{\varphi}_i \boldsymbol{\varphi}_i^T = \boldsymbol{C}_{-i} + \alpha_i^{-1} \boldsymbol{\varphi}_i \boldsymbol{\varphi}_i^T,$$

where $\boldsymbol{\varphi}_i$ denotes the i-th column of $\boldsymbol{\Phi}$, which is a N dimensional vector with elements $(\phi_i(\boldsymbol{x}_i), \dots, \phi_i(\boldsymbol{x}_N))$, in contrast to ϕ_n which denotes the n-th row of $\boldsymbol{\Phi}$. The matrix $\boldsymbol{C}_{-i}$ represents the matrix $\boldsymbol{C}$ with the contribution from basis function i removed. We can write

$$|\boldsymbol{C}| = |\boldsymbol{C}_{-i}||1 + \alpha_i^{-1} \boldsymbol{\varphi}_i^T \boldsymbol{C}_{-i}^{-1} \boldsymbol{\varphi}_i|$$

$$\boldsymbol{C}^{-1} = \boldsymbol{C}_{-i}^{-1} - \frac{\boldsymbol{C}_{-i}^{-1}\boldsymbol{\varphi}_i\boldsymbol{\varphi}_i^T\boldsymbol{C}_{-i}^{-1}}{\alpha_i + \boldsymbol{\varphi}_i^T\boldsymbol{C}_{-i}^{-1}\boldsymbol{\varphi}}$$

We can now write the log marginal likelihood in the form

$$L(\boldsymbol{\alpha}) = L(\boldsymbol{\alpha}_{-i}) + \lambda(\alpha_i) = L(\boldsymbol{\alpha}_{-i}) + \frac{1}{2}\left(\ln\alpha_i - \ln(\alpha_i + s_i) + \frac{q_i^2}{\alpha_i + s_i}\right)$$

where $s_i = \boldsymbol{\varphi}_i^T\boldsymbol{C}_{-i}^{-1}\boldsymbol{\varphi}_i$ (called sparsity of $\boldsymbol{\varphi}_i$) and $q_i = \boldsymbol{\varphi}_i^T\boldsymbol{C}_{-i}^{-1}\boldsymbol{t}$ (quality of $\boldsymbol{\varphi}_i$). The sparsity measures the extent to which basis function $\boldsymbol{\varphi}_i$ overlaps with the other basis vectors in the model,and the quality represents a measure of the alignment of the basis vector $\boldsymbol{\varphi}_n$ with the error between the training set values $\boldsymbol{t}$ and the vector $\boldsymbol{y}_{-i}$ of predictions that would result from the model with the vector $\boldsymbol{\varphi}_i$ excluded. The stationary points of the marginal likelihood with respect to α_i occur when the derivative is zero

$$\frac{d\lambda_i}{d\alpha_i} = \frac{\alpha_i^{-1}s_i^2 - (q_i^2 - s_i)}{2(\alpha_i + s_i)^2} = 0$$

Recall $\alpha_i \geq 0$, we see that if $q_i^2 < s_i$, then $\alpha_i \to \infty$ provides a solution. Conversely if $q_i^2 > s_i$, we can solve to get $\alpha_i = \frac{s_i^2}{q_i^2 - s_i}$. Second derivative analysis confirm that these are unique maxima of $\lambda(\alpha_i)$. This results in the following efficient algorithm:

- Initialize β.
- Initialize using one basis function $\boldsymbol{\varphi}_1$, with hyperparameter $\alpha_1 = \frac{s_1^2}{1_1^2 - s_1}$ with the remaining hyperparameters α_j for $j \neq i$ initialized to infinity, so that only $\boldsymbol{\varphi}_1$ is included in the model.
- Evaluate $\boldsymbol{\Sigma}$ and $\boldsymbol{m}$ along with q_i and s_i for all basis functions.
- Select a candidate basis function $\boldsymbol{\varphi}_i$.
- if $q_i^2 > s_i$, and $\alpha_i < \infty$, so that the basis vector $\boldsymbol{\varphi}_i$ is already included in the model, then update α_i.
- if $q_i^2 > s_i$, and $\alpha_i = \infty$, then add $\boldsymbol{\varphi}_i$ to the model, and evaluate hyperparameter α_i.
- if $q_i^2 \leq s_i$, and $\alpha_i < \infty$ then remove basis function $\boldsymbol{\varphi}_i$ from the model, and set $\alpha_i = \infty$.
- Update β.
- if converged terminate, otherwise go to step 3.

In practice it is convenient to evaluate the quantities $Q_i = \boldsymbol{\varphi}_i^T\boldsymbol{C}^{-1}\boldsymbol{t}$ and $S_i = \boldsymbol{\varphi}_i^T\boldsymbol{C}^{-1}\boldsymbol{\varphi}_i$. The quality and sparseness variables are then expressed as $q_i = \frac{\alpha_i Q_i}{\alpha_i - S_i}$ and $s_i = \frac{\alpha_i S_i}{\alpha_i - S_i}$. When $\alpha_i = \infty$, we have $q_i = Q_i$ and $s_i = S_i$. We can then write

$$Q_i = \beta\boldsymbol{\varphi}_i^T\boldsymbol{t} - \beta^2\boldsymbol{\varphi}_i^T\boldsymbol{\Phi}\boldsymbol{\Sigma}\boldsymbol{\Phi}^T\boldsymbol{t}$$

$$S_i = \beta\boldsymbol{\varphi}_i^T\boldsymbol{\varphi}_i - \beta^2\boldsymbol{\varphi}_i^T\boldsymbol{\Phi}\boldsymbol{\Sigma}\boldsymbol{\Phi}^T\boldsymbol{\varphi}_i$$

where $\boldsymbol{\Phi}$ and $\boldsymbol{\Sigma}$ involve only those basis vectors that corresponds to finite hyperparameters α_i. At each stage the required computations therefore scale like $\mathcal{O}(M^3)$, where M is the number of active basis vectors in the model and is typically much smaller than number N of training examples.

4 Mathematical Results

4.1 Matrix Identities

A matrix A has elements A_{ij} where i indexes the rows, and j indexes the columns.

- $(AB)^T = B^TA^T$, $(AB)^{-1} = B^{-1}A^{-1}$, $(A^T)^{-1} = (A^{-1})^T$.
- $(I + AB)^{-1}A = A(I + BA)^{-1}$

- Woodbury identity: $(A+BD^{-1}C)^{-1} = A^{-1} - A^{-1}B(D+CA^{-1}B)^{-1}CA^{-1}$. Useful when A is large and diagonal and easy to invert, which B has many rows but few columns and conversely for C.
- Cyclic property of traces: $Tr(ABC) = Tr(CBA) = Tr(BCA)$.
- If A and B are matrices of size $N \times M$ then $|I_N + AB^T| = |I_M + A^TB|$.

4.2 Gaussian Marginalization

We have

$$p(\boldsymbol{x}) = \mathcal{N}(\boldsymbol{x}|\boldsymbol{\mu}, \boldsymbol{\Lambda}^{-1})$$
$$p(\boldsymbol{y}|\boldsymbol{x}) = \mathcal{N}(\boldsymbol{y}|\boldsymbol{A}\boldsymbol{x}+\boldsymbol{b}, \boldsymbol{L}^{-1}).$$

Then for the marginal distribution of $\boldsymbol{y}$ we have

$$\mathbb{E}[\boldsymbol{y}] = \boldsymbol{A}\boldsymbol{\mu} + \boldsymbol{b}, \quad cov[\boldsymbol{y}] = \boldsymbol{L}^{-1} + \boldsymbol{A}\boldsymbol{\Lambda}^{-1}\boldsymbol{A}^T.$$

For the conditional we have

$$\mathbb{E}[\boldsymbol{x}|\boldsymbol{y}] = (\boldsymbol{\Lambda} + \boldsymbol{A}^T\boldsymbol{L}\boldsymbol{A})^{-1}(\boldsymbol{A}^T\boldsymbol{L}(\boldsymbol{y}-\boldsymbol{b}) + \boldsymbol{\Lambda}\boldsymbol{\mu}) \quad cov[\boldsymbol{x}|\boldsymbol{y}] = (\boldsymbol{\Lambda} + \boldsymbol{A}^T\boldsymbol{L}\boldsymbol{A})^{-1}.$$

This is same as Bayes' theorem.

4.3 Bayesian inference for the Gaussian

For Gaussian variable x with N observations $\boldsymbol{X} = \{x_1, \ldots, x_N\}$, we have

- Mean μ is unknown, variance σ^2 is known: Conjugate prior on mean is Gaussian $p(\mu) = \mathcal{N}(\mu|\mu_0, \sigma_0^2)$. The posterior $p(\mu|\boldsymbol{X}) = \mathcal{N}(\mu|\mu_N, \sigma_N^2)$ with $\mu_N = \frac{\sigma^2}{N\sigma_0^2+\sigma^2}\mu_0 + \frac{N\sigma_0^2}{N\sigma_0^2+\sigma^2}\mu_{ML}$ with $\mu_{ML} = \frac{1}{N}\sum_{n=1}^{N} x_n$, and $\frac{1}{\sigma_N^2} = \frac{1}{\sigma_0^2} + \frac{N}{\sigma^2}$. This is same for multivariate case $\mathcal{N}(\boldsymbol{x}|\boldsymbol{\mu}, \boldsymbol{\Lambda}^{-1})$ with D-dimensional variable $\boldsymbol{x}$.
- Mean μ is known, precision $\lambda = \frac{1}{\sigma^2}$ is unknown: Conjugate prior on precision is Gamma distribution $p(\lambda|a_0, b_0) = \frac{1}{\Gamma(a_0)} b_0^{a_0} \lambda^{a_0-1} e^{-b_0\lambda}$, with mean a_0/b_0 and variance a_0/b_0^2 for $a_0 > 0$. The posterior $p(\lambda|\boldsymbol{X}) = Gam(\lambda|a_N, b_N)$ with $a_N = a_0 + \frac{N}{2}$, and $b_N = b_0 + \frac{1}{2}\sum_{n=1}^{N}(x_n - \mu)^2 = b_0 + \frac{N}{2}\sigma_{ML}^2$. For multivariate case the conjugate prior is a Wishart distribution $\mathcal{W}(\boldsymbol{\Lambda}|\boldsymbol{W}, \nu) = B|\boldsymbol{\Lambda}|^{(\nu-D-1)/2} e^{-\frac{1}{2}Tr(\boldsymbol{W}^{-1}\boldsymbol{\Lambda})}$ where ν is called the number of degree of freedom of the distribution, $\boldsymbol{W}$ is a $D \times D$ scale matrix and normalization constant $B(\boldsymbol{W}, \nu) = |\boldsymbol{W}|^{-\nu/2}\left(2^{\nu D/2}\pi^{D(D-1)/4}\prod_{i=1}^{D}\Gamma\left(\frac{\nu+1-i}{2}\right)\right)^{-1}$.
- If both mean and precision are unknown, we use the Gaussian-Gamma distribution as the conjugate prior $p(\mu, \lambda) = \mathcal{N}(\mu|\mu_0, (\beta\lambda)^{-1})Gam(\lambda|a, b)$, with the coupling evident in precision of mean being a linear function of λ. For multi-variate case the conjugate prior is the Gaussian-Wishart distribution $p(\boldsymbol{\mu}, \boldsymbol{\Lambda}|\boldsymbol{\mu}_0, \beta, \boldsymbol{W}, \nu) = \mathcal{N}(\boldsymbol{\mu}|\boldsymbol{\mu}_0, (\beta\boldsymbol{\Lambda})^{-1})\mathcal{W}(\boldsymbol{\Lambda}|\boldsymbol{W}, \nu)$.

If we integrate out precision, we obtain the marginal distribution of x in the form

$$p(x|\mu, a, b) = \int_0^\infty \mathcal{N}(x|\mu, \tau^{-1})Gam(\tau|a, b)d\tau = \frac{b^a}{\gamma(a)}\left(\frac{1}{2\pi}\right)^{1/2}\left[b + \frac{(x-\mu)^2}{2}\right]^{-a-1/2}\Gamma(a+1/2)$$

Defining new variables $\nu = 2a$ and $\lambda = a/b$ the distribution becomes a Student's t-distribution.

$$St(x|\mu, \lambda, \nu) = \frac{\Gamma(\nu/2+1/2)}{\Gamma(\nu/2)}\left(\frac{\lambda}{\pi\nu}\right)^{1/2}\left[1 + \frac{\lambda(x-\mu)^2}{\nu}\right]^{-\nu/2-1/2}.$$

For $\nu = 1$ this reduces to Cauchy distribution and with $\nu \to \infty$ it becomes a Gaussian $\mathcal{N}(x|\mu, \lambda^{-1})$. We see that Student's t-distribution is obtained by adding up an infinite number of Gaussian distributions having the same mean but different precisions (mixture of Gaussians) resulting in longer tails than a Gaussian, making this a robust distribution to outliers. The multivariate version is given by

$$St(\boldsymbol{x}|\boldsymbol{\mu}, \boldsymbol{\Lambda}, \nu) = \frac{\Gamma(\nu/2 + D/2)}{\Gamma(\nu/2)} \frac{|\boldsymbol{\Lambda}|^{1/2}}{(\pi\nu)^{D/2}} \left[1 + \frac{\Delta^2}{\nu}\right]^{-\nu/2-D/2}$$

where D is the dimensionality of $\boldsymbol{x}$, and $\Delta^2 = (\boldsymbol{x} - \boldsymbol{\mu})^T \boldsymbol{\Lambda}(\boldsymbol{x} - \boldsymbol{\mu})$ is the squared Mahalanobis distance. This distribution has a mean $\boldsymbol{\mu}$ if $\nu > 1$, covariance $\frac{\nu}{\nu-2}\boldsymbol{\Lambda}^{-1}$ if $\nu > 2$ and mode $\boldsymbol{\mu}$.

4.4 Bayesian Model comparison

Over-fitting associated with maximum likelihood can be avoided by marginalizing over the model parameters instead of making point estimates of their values. Suppose we wish to compare a set of L models $\{\mathcal{M}_i\}$ where $i = 1, \ldots, L$. The model refers to a probability distribution over the observed data $\mathcal{D}$. We shall suppose that the data is generated from one of these models but we are uncertain which one. Our uncertainty is expressed through a prior probability distribution $p(\mathcal{M}_i)$. Given a training set $\mathcal{D}$, we wish to evaluate the posterior

$$\underbrace{p(\mathcal{M}_i|\mathcal{D})}_{\text{posterior}} \propto \underbrace{p(\mathcal{M}_i)}_{\text{prior}} \underbrace{p(\mathcal{D}|\mathcal{M}_i)}_{\text{model evidence}} .$$

The prior allows use to express a preference for different models. Let us assume all models are equally probable. Model evidence (also called marginal likelihood) expresses the preference shown by the data. For two models the **Bayes factor** is defined as $p(\mathcal{D}|\mathcal{M}_i)/p(\mathcal{D}|\mathcal{M}_j)$. The predictive distribution is given by

$$p(t|\boldsymbol{x}, \mathcal{D}) = \sum_{i=1}^{L} p(t|\boldsymbol{x}, \mathcal{M}_i, \mathcal{D}) p(\mathcal{M}_i|\mathcal{D}).$$

This is an example of a *mixture distribution* where overall prediction is a weighted average distribution of those model distributions. A simple approximation to model averaging is to use the single most probable model, called *model selection*.

For a model governed by a set of parameters $\boldsymbol{w}$, the model evidence is given by

$$p(\mathcal{D}|\mathcal{M}_i) = \int p(\mathcal{D}|\boldsymbol{w}, \mathcal{M}_i) p(\boldsymbol{w}|\mathcal{M}_i)\ d\boldsymbol{w}.$$

It is the probability of generating the data $\mathcal{D}$ from a model whose parameters are sampled at random from the prior. Implicit in the Bayesian model comparison framework is the assumption that the true distribution from which the data are generated is contained within the set of models under consideration. In a practical application, therefore, it will be wise to keep aside an independent test set of data on which to evaluate the overall performance of the final system.