Computing Energy Scores

Oliver Robinson

March 2023

1 Computing the Predictive Distribution without Integrating

Suppose we have some observed data $x_1, \ldots x_n$. The joint distribution of (y, θ) is given by the product of the likelihood (with $\eta = 1$) and the posterior:

$$\begin{aligned} p(y, \theta) &\propto l(y \mid \theta)\pi(\theta \mid x_1, \ldots, x_n) \\ &\propto l(y \mid \theta)\Pi_{i=1}^{n} l(x_i \mid \theta)\pi(\theta) \end{aligned}$$

We can then use MCMC to produce m samples $(y_1, \theta_1), \ldots, (y_m, \theta_m)$ from this joint distribution. Then, we can marginalise these samples to consider the y_i and θ_i values separately.

We have $y_1, \ldots y_m \overset{\text{iid}}{\sim} f^{\eta=1}(y)$, that is, these are i.i.d. samples from the (unnormalised) predictive distribution with $\eta = 1$. We then need to be able to compute the unnormalised predictive distribution $f^{\eta}(y)$ for a generic η, which is given by:

$$f^{\eta}(y) = \int l^{\eta}(y \mid \theta)\pi(\theta \mid x_1, \ldots, x_n)d\theta \tag{1}$$

$$\tilde{f}_{\eta}(y) \simeq \frac{1}{m}\sum_{i=1}^{m} l^{\eta}(y \mid \theta_i) \tag{2}$$

where $\theta_i \sim \pi(\theta_i \mid x_1, \ldots, x_n)$, i.e. we are using our generated θ_i samples to compute an estimate of this integral (hence removing the need to perform this integration).

2 Calculating the Energy Scores

We can now calculate an energy score for each (η, x) pair, hence giving us a way of computing the optimal η for a given x; by minimising this energy score:

$$\begin{aligned} \eta(x) &= \arg\min_{\eta} S_E(P^{\eta}, x) \\ &= \arg\min_{\eta} 2\mathbb{E}_{X \sim P^{\eta}}||X - x||_2^{\beta} - \mathbb{E}_{X, X' \sim P^{\eta}}||X - X^{'}||_2^{\beta} \\ &= \arg\min_{\eta} 2\sum_{i=1}^{m} w(y_i)||y_i - x||_2^{\beta} - \sum_{i \neq j} w(y_i, y_j)||y_i - y_j||_2^{\beta} \end{aligned}$$

where $w(y_i) = \frac{\frac{\tilde{f_\eta}(y_i)}{f_1(y_i)}}{\sum_{i=1}^{n}\frac{\tilde{f_\eta}(y_i)}{f_1(y_i)}}$ and $w(y_i, y_j) = \frac{\frac{\tilde{f_\eta}(y_i)\tilde{f_\eta}(y_j)}{f_1(y_i)f_1(y_j)}}{\sum_{i \neq j}\frac{\tilde{f_\eta}(y_i)\tilde{f_\eta}(y_j)}{f_1(y_i)f_1(y_j)}}$

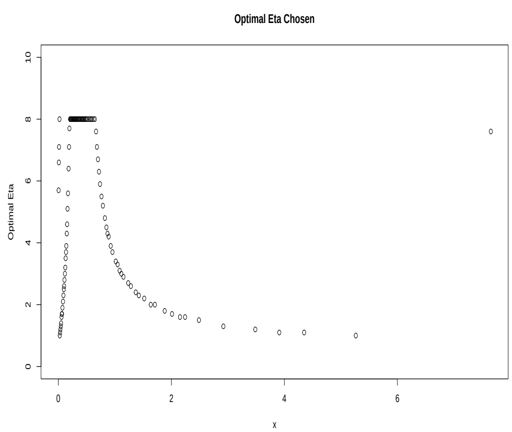
We can then optimise the function $\eta(x)$ using the Bayesian Optimisation with Gaussian Processes approach previously implemented.

3 Pareto/Log-Normal Example

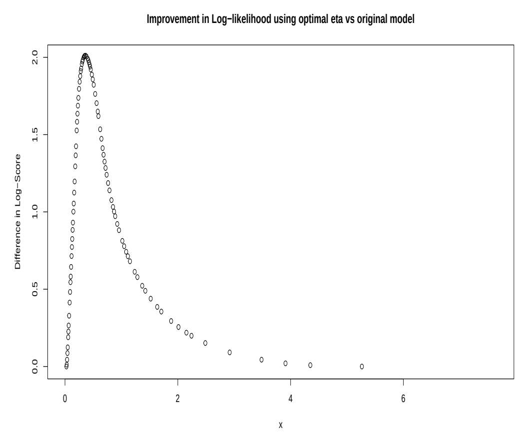
Our first example is to consider some data drawn from a Pareto distribution, which is then modelled by a Log-Normal.

3.1 Using Log-Likelihood to Compute η

One approach to computing the optimal η is to choose the η value which maximises the log-likelihood of the η-generalised predictive distribution for a given x. This can then be done for a range of x values, as shown in Figure 1a below. We can then use a Gaussian process to fit a function through these points which gives us an optimal η for any given value of x.



(a) Optimal η values chosen for the η-generalised Log-Normal distribution



(b) Difference in Log-Likelihood between $\eta = 1$ and the optimal η chosen above

Figure 1: Pareto/Log-Normal Example

3.2 Using Energy Scores to Compute η (with exact predictive distribution)

Another method for computing the optimal η is to compute the energy scores as outlined above in Section 2. In our example, we are able to analytically calculate the exact form (up to a normalising constant) of $f^\eta(y)$, i.e. it is possible to compute the integration in (1) (with a fair bit of effort!). This then allows us compute the weights $w(y_i)$ and $w(y_i, w_j)$ exactly for use in the energy score. Implementing this approach for this example yields the optimal η values shown in Figure 2 below.

3.3 Using Energy Scores to Compute η (with approximate predictive distribution)

Often in many examples it will not be possible to compute the integration in (1) analytically as we did in the previous section. It may therefore be necessary to approximate this integration in order to give us an estimate of the predictive distribution $\tilde{f}_\eta(y)$, as shown in (2). This approximate predictive distribution can then be used in the same way as the exact version to compute the weights $w(y_i)$ and $w(y_i, y_j)$ for the energy scores, which we then use to compute the optimal η values.

I have tried to implement this approach, currently it is selecting the same η value for all values of x. Needs more work to implement this correctly, but should be doable.

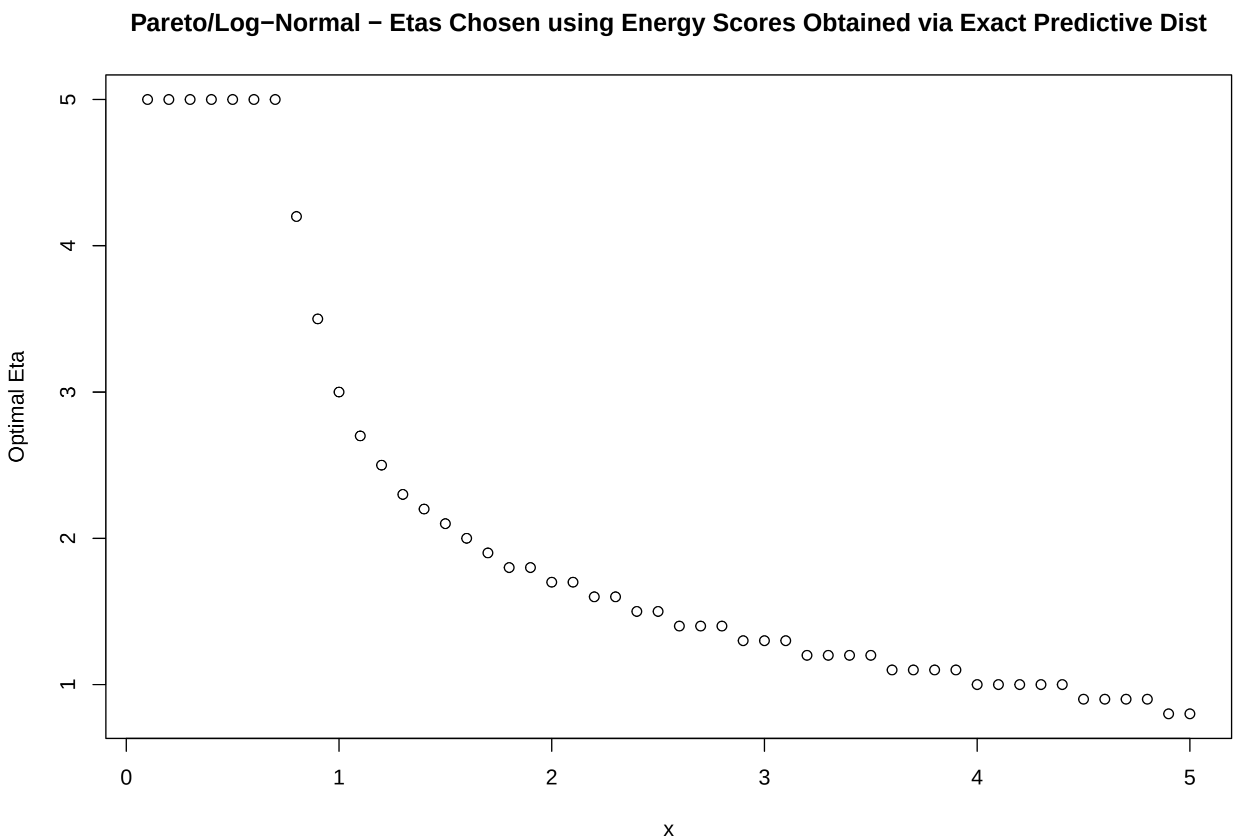


Figure 2: Optimal η values chosen using energy scores computed with exact predictive distribution