On Bayesian Model Assessment and Choice Using Cross-Validation Predictive Densities: Appendix

Appendix of Research report B23

Aki Vehtari and Jouko Lampinen
Laboratory of Computational Engineering
Helsinki University of Technology
P.O.Box 9203, FIN-02015, HUT, Finland
{Aki.Vehtari,Jouko.Lampinen}@hut.fi

April 11, 2001

Appendix: Prior and MCMC specification details

Short description of the prior and the MCMC specification details is given here. See (Neal, 1996, 1997, 1999; Lampinen & Vehtari, 2001) and the FBM software manual (Neal, 2000) for additional details.

The MCMC sampling was done with the FBM[1] software and Matlab-code derived from the Netlab[2] toolbox. FBM was used when possible (normal and logistic models) because of its speed. Matlab-based implementation was slower, but it was much easier to implement new features to it.

In the following, we use the notation $r \sim F(a)$ as shorthand for $p(r) = F(r|a)$ where a denotes the parameters of the distribution F, and the random variable argument r is not shown explicitly. $N(\mu, \sigma^2)$ denotes a normal distribution with mean μ and variance σ^2. Parametrization of the Inverse Gamma here is

$$\text{Inv-gamma}(\sigma^2|\sigma_0^2, \nu) \propto (\sigma^2)^{-(\nu/2+1)} \exp\left(-\frac{1}{2}\nu\sigma_0^2\sigma^{-2}\right),$$

which is equal to a scaled inverse chi-square distribution (Gelman et al., 1995, Appendix A). The parameter ν is the number of degrees of freedom and σ_0^2 is a scale parameter. We first describe common details for MLP and GP and then specific details for each case.

MLP

We used one hidden layer MLP with tanh hidden units, which in matrix format can be written as

$$f(x, \theta_w) = b^2 + w^2 \tanh\left(b^1 + w^1 x\right).$$

The θ_w denotes all the parameters w^1, b^1, w^2, b^2, which are the hidden layer weights and biases, and the output layer weights and biases, respectively. The Gaussian priors for the weights were

$$w^1 \sim N(0, \alpha_{w^1})$$
$$b^1 \sim N(0, \alpha_{b^1})$$
$$w^2 \sim N(0, \alpha_{w^2})$$
$$b^2 \sim N(0, \alpha_{b^2})$$

[1] `http://www.cs.toronto.edu/~radford/fbm.software.html`
[2] `http://www.ncrg.aston.ac.uk/netlab/`

where the α's are the variance hyperparameters. The conjugate inverse Gamma hyperprior for α_j's is

$$\alpha_j \sim \text{Inv-gamma}(\alpha_{0,j}, \nu_{\alpha,j})$$

The fixed values for the highest level hyperparameters in the case studies were similar to those used in (Neal, 1996, 1998). The hyperpriors for w^1 and w^2 were scaled according to the number of inputs K and hidden units J. Typical values were

$$\begin{aligned}\nu_{\alpha,w^1} &= 0.5\\ \alpha_{0,w^1} &= (0.05/K^{1/\nu_{\alpha,w^1}})^2.\end{aligned}$$

ARD prior was used for input weights

$$\begin{aligned}w^1_{j,k} &\sim N(0, \alpha_{k,w^1}),\\ \alpha_{k,w^1} &\sim \text{Inv-gamma}(\bar{\alpha}_{w^1}, \nu_{\alpha,w^1})\\ \bar{\alpha}_{w^1} &\sim \text{Inv-gamma}(\bar{\alpha}_{0,w^1}, \nu_{\bar{\alpha},w^1}),\end{aligned}$$

where the average scale of the α_k is determined by the next level hyperparameters. Sampling of the weights was done with HMC and sampling of the hyperparameters was done with Gibbs sampling.

GP

We used a simple covariance function producing smooth functions

$$C_{ij} = \eta^2 \exp\left(-\sum_{u=1}^{p} \rho_u^2 (x_u^{(i)} - x_u^{(j)})^2\right) + \delta_{ij} J^2 + \delta_{ij}\sigma_e^2.$$

Jitter was fixed to $J^2 = 0.01$. Inverse Gamma prior for scale was

$$\eta^2 \sim \text{Inv-gamma}(\alpha_{\eta^2}, \nu_{\eta^2})$$

And ARD prior was used for relevance parameters

$$\begin{aligned}\rho_u^2 &\sim \text{Inv-gamma}(\alpha_{\rho^2}, \nu_{\rho^2})\\ \alpha_{\rho^2} &\sim \text{Inv-gamma}(\alpha_{0,\rho^2}, \nu_{0,\rho^2}),\end{aligned}$$

Sampling of the parameters was done with HMC.

Robot Arm

The Normal residual model with Inverse Gamma prior for variance was used

$$\begin{aligned}e &\sim N(0, \sigma^2)\\ \sigma^2 &\sim \text{Inv-gamma}(\sigma_0^2, \nu_\sigma).\end{aligned}$$

Prior specification for MLP model specified in FBM notation was

```
net-spec log 2 8 2 / - x0.05:0.5:1 0.05:0.5 - x0.05:0.5 - 1
model-spec log real 0.05:0.5
```

These commands set hyperparameters of parameter priors to values: $\nu_{\alpha,w^1} = 0.5$, $\bar{\alpha}_{0,w^1} = (0.05/2^{1/\nu_{\alpha,w^1}})^2$, $\nu_{\bar{\alpha},w^1} = 1$, $\nu_{\alpha,b^1} = 0.5$, $\alpha_{0,b^1} = 0.05^2$, $\nu_{\alpha,w^2} = 0.5$, $\alpha_{0,w^2} = (0.05/8^{1/\nu_{\alpha,w^2}})^2$, $\alpha_{b^2} = 1$, and Normal residual model parameters to $\sigma_0 = 0.05$, and $\nu_\sigma = 0.5$. MCMC specification for MLP model in FBM notation was

```
net-gen log fix 0.1 0.5
mc-spec log repeat 10 heatbath hybrid 10 0.05
net-mc log 1
mc-spec log repeat 50 sample-sigmas heatbath hybrid 10 0.15
net-mc log 2
mc-spec log repeat 50 sample-sigmas heatbath 0.9 hybrid 10 0.2 negate
net-mc log 3
mc-spec log repeat 500 sample-sigmas heatbath 0.9 hybrid 10 0.4 negate
net-mc log 1000
```

`net-gen` intializes weights to zero and following paramteres as $\alpha_{k,w^1} = 0.1$, $\alpha_{b^1} = 0.1$, $\alpha_{w^2} = 0.5$, and $\sigma = 0.1$. First three `mc-spec` specifications are used to improve convergence and last `mc-spec` specification specifies the main run (`net-mc` does the sampling). The last mc-spec specification means: alternate Gibbs sampling and HMC with chain length 10, step size 0.4, and persistence 0.9, save every 500th iteration and last net-mc does iterations until 1000 iterations are saved. See FBM manual for full details. All samples were used to estimate convergence and autocorrelations. For predictive densities 100 samples were used (every ninth sample starting from sample 100).

Prior specification for GP model specified in FBM notation

```
gp-spec log 2 2 - - 0.01 / 0.05:0.5 0.05:0.5:1
model-spec log real 0.05:0.5
```

These commands set hyperparameters of parameter priors to values: $J = 0.01$, $\alpha_{\eta^2} = 0.05^2$, $\nu_{\eta^2} = 0.5$, $\nu_{\rho^2} = 0.5$, $\alpha_{0,\rho^2} = 0.05^2$, $\nu_{0,\rho^2} = 1$, and Normal residual model parameters to $\sigma_0 = 0.05$, and $\nu_\sigma = 0.5$. MCMC specification for GP model in FBM notation

```
gp-gen log fix 0.2 0.1
mc-spec log repeat 5 heatbath hybrid 10 0.1
gp-mc log 1
mc-spec log repeat 5 heatbath hybrid 10 0.4 sample-variances
gp-mc log 1000
```

`gp-gen` intializes following paramteres as $\eta^2 = 0.2$, $\rho_u^2 = 0.1$ and $\sigma = 0.2$. First `mc-spec` specification is used to improve convergence (per-case variances are not yet sampled) and second `mc-spec` specification specifies the main run (`gp-mc` does the sampling). The last mc-spec specification means: aletranate HMC with chain length 10, step size 0.4, no persistence and sampling per-case variances with Gibbs sampling, save every 5th iteration and last gp-mc does iterations until 1000 iterations are saved. See the FBM manual for full details. All samples were used to estimate convergence and autocorrelations. For predictive densities 100 samples were used (every ninth sample starting from sample 100).

Concrete Quality Estimation

Prior specifications for MLPs and GPs were same as in robot arm case except for parameters of alternative residual models.

For MLP, we used Student's t-distribution, where the tails can be controlled by choosing the number of degrees of freedom ν in the distribution. The integration over the degrees of freedom was done by Gibbs sampling for discretized values of ν, so that the residual model is

$$
\begin{aligned}
e &\sim t_\nu(0, \sigma^2) \\
\nu &= V[i] \\
i &\sim U_d(1, K) \\
V[1:K] &= [2,\ 2.3,\ 2.6,\ 3,\ 3.5,\ 4,\ 4.5,\ 5:1:10,\ 12:2:20,\ 25:5:50] \\
\sigma^2 &\sim \text{Inv-gamma}(\sigma_0, \nu_\sigma)
\end{aligned}
$$

where $[a:s:b]$ denotes the set of values from a to b with step s, and $U_d(a, b)$ is a uniform distribution of integer values between a and b.

For GP instead of Student's t-distribution, we used per-case residual variance model, with Inverse-Gamma prior having unknown degrees of freedom

$$\begin{aligned} e^n &\sim N(0, (\sigma^2)^n) \\ (\sigma^2)^n &\sim \text{Inv-gamma}(\sigma_{\text{ave}}, \nu_\sigma) \\ \nu_\sigma &\sim \text{Inv-gamma}(\sigma_\nu, \nu_\nu) \\ \sigma_{\text{ave}} &\sim \text{Inv-gamma}(\sigma_0, \nu_{\sigma,ave}), \end{aligned}$$

where the fixed hyperparameters are σ_ν, ν_ν, σ_0 and $\nu_{\sigma,ave}$. This is equal to the t_ν-distribution residual model with unknown degrees of freedom ν_σ (Geweke, 1993), but easier to implement for GP (Neal, 1999).

We also used input dependent Normal and t_{nu} residual models. Three inputs, which were zero/one variables indicating use of additives, divide the data to six groups. For each group, we had own residual model parameters with common hyperprior. For example for input dependent Normal model (in.dep.-N) we had

$$\begin{aligned} e_i &\sim N(0, \sigma^2_{l(i)}) \\ \sigma^2_l &\sim \text{Inv-gamma}(\sigma^2, \nu_{\sigma_l}) \\ \sigma^2 &\sim \text{Inv-gamma}(\sigma^2_0, \nu_\sigma), \end{aligned}$$

where l is index of group, and $l(i)$ is index of group to which ith data point belongs.

For MLP the main HMC parameters were: the length of individual chains was 100, step size 0.5 with Neal's heuristic step size adjustment, persistence parameter 0.9, and window length in windowing 5. The burn-in stage contained 16 000 iterations and the actual sampling 80 000 iterations, from which 100 samples were stored for the analysis.

For GP the main HMC parameters were: the length of individual chains was 100, step size 0.4 with Neal's heuristic step size adjustment and persistence parameter 0.9. The burn-in stage contained 200 iterations and the actual sampling 1 000 iterations, from which 100 samples were stored for the analysis.

Forest scene classification

Prior specifications for MLP was same as in robot arm case except $\bar{\alpha}_{0,w^1} = (0.2/2^{1/\nu_{\alpha,w^1}})^2$ and logistic transformation was used to compute the probability that a binary-valued target, y, has value 1

$$p(y = 1|x, \theta_w, M) = [1 + \exp(-f(x, \theta_w))]^{-1}.$$

The main HMC parameters were: the length of individual chains was 10, step size 0.2 with Neal's heuristic step size adjustment, persistence parameter 0.95. The burn-in stage contained 5000 iterations and the actual sampling 45 000 iterations, from which 100 samples were stored for the analysis.

Bibliography

Gelman, A., Carlin, J. B., Stern, H. S., and Rubin, D. R. (1995). *Bayesian Data Analysis*. Chapman & Hall.

Geweke, J. (1993). Bayesian treatment of the independent Student-t linear model. *Journal of Applied Econometrics*, 8(Supplement):S19–S40.

Lampinen, J. and Vehtari, A. (2001). Bayesian approach for neural networks – review and case studies. *Neural Networks*, 14(3):7–24.

Neal, R. M. (1996). *Bayesian Learning for Neural Networks*. Springer-Verlag.

Neal, R. M. (1997). Monte Carlo implementation of Gaussian process models for Bayesian regression and classification. Technical Report 9702, Dept. of Statistics, University of Toronto.

Neal, R. M. (1998). Assessing relevance determination methods using DELVE. In Bishop, C. M., editor, *Neural Networks and Machine Learning*, pp. 97–129. Springer-Verlag.

Neal, R. M. (1999). Regression and classification using Gaussian process priors (with discussion). In Bernardo, J. M., Berger, J. O., Dawid, A. P., and Smith, A. F. M., editors, *Bayesian Statistics 6*, pp. 475–501. Oxford University Press.

Neal, R. M. (2000). Software for Flexible Bayesian Modeling [online]. Available at `http://www.cs.toronto.edu/~radford/fbm.software.html`.