Comments for: Rivals, I., & Personnaz L. (2000). Construction of confidence intervals for neural networks based on least squares estimation, ***Neural Networks***, 13, 463–484

Jan Larsen and Lars Kai Hansen

Informatics and Mathematical Modeling

Richard Petersens Plads, Building 321

Technical University of Denmark

DK-2800 Lyngby, Denmark

Phone: +45 4525 3923,3889

Fax: +45 45872588

Email: jl,lkhansen@imm.dtu.dk, Web: eivind.imm.dtu.dk

Rivals and Personnaz (Rivals & Personnaz, 2000) mainly concerns estimation of confidence intervals (or error bars) for neural network prediction models trained by least squares, but also the use of approximate leave-one-out (LOO) cross validation error for model selection is considered.

In (Hansen & Larsen, 1996) "Linear Unlearning for Cross-Validation," *Advances in Computational Mathematics*, 5, 269–280, 1996, we proposed an approximation of the LOO error which in (Rivals & Personnaz, 2000), p. 473, footnote 10, is claimed to be invalid - even in the case of models which are linear in parameters. This is, however, a misrepresentation of our work and incorrect.

In (Hansen & Larsen, 1996) we suggested LOO approximations for general cost functions possibly augmented by a regularization term (e.g., weight decay) for nonlinear as well as linear models. In general we consider models which from the input vector $\boldsymbol{x}$ predict an output y by $\widehat{y} = f(\boldsymbol{x}, \boldsymbol{w})$, where $f(\cdot)$ generally is a nonlinear function of the input $\boldsymbol{x}$ and the parameter vector $\boldsymbol{w}$.

In the case of linear models trained by least squares, the LOO error cf. equation (18) (Hansen & Larsen, 1996) given *exactly* as:

$$E_{\mathrm{LOO}} = \frac{1}{N}\sum_{k=1}^{N} \frac{(y_k - \widehat{\boldsymbol{w}}^\top \boldsymbol{x}_k)^2}{(1 - \boldsymbol{h}_k^\top \boldsymbol{J}^{-1}\boldsymbol{h}_k)^2} \tag{1}$$

where N is the number of data examples, $\widehat{\boldsymbol{w}}$ is the parameter vector which minimizes the least squares cost function (augmented by a regularization term $R(\boldsymbol{w})$), $\boldsymbol{h}_k = \partial f(\boldsymbol{x}_k, \boldsymbol{w})/\partial \boldsymbol{w}|_{\boldsymbol{w}=\widehat{\boldsymbol{w}}}$ which for linear models equals $\boldsymbol{x}$, and $\boldsymbol{J} = \sum_{k=1}^{N} \boldsymbol{h}_k \boldsymbol{h}_k^\top + \partial^2 R(\widehat{\boldsymbol{w}})/\partial \boldsymbol{w} \partial \boldsymbol{w}^\top$. That is, in the case of no regularization ($R(\boldsymbol{w}) = 0$), which is considered in (Rivals & Personnaz, 2000), equation (1) coincides with equations (36) and (38) of (Rivals & Personnaz, 2000). The critique that our approximation is not valid is thus incorrect.

In the case of least squares learning (and other cost functions) for general nonlinear models the LOO error is an $o(1/N)$ approximation[1] according to Theorem 2 in (Hansen & Larsen, 1996). Using the approximation suggested in equations (37) and (38) of (Rivals & Personnaz, 2000) involve terms of higher order[2] $O(1/N^i), i \leq 2$, which is *inconsistent* (see further the proof of Theorem 2 (Hansen & Larsen, 1996)). On the other hand, the approximation

$$\widehat{E}_{\mathrm{LOO}} = \frac{1}{N}\sum_{k=1}^{N}(y_k - f(\boldsymbol{x}_k, \widehat{\boldsymbol{w}})^2 \frac{1 + \boldsymbol{h}_k^\top \boldsymbol{J}^{-1}\boldsymbol{h}_k}{1 - \boldsymbol{h}_k^\top \boldsymbol{J}^{-1}\boldsymbol{h}_k} \tag{2}$$

we suggested in equation (17), (Hansen & Larsen, 1996) is consistent.

The main topic of (Rivals & Personnaz, 2000), estimation of confidence intervals, was first discussed in a neural network context by (Buntine & Weigend, 1991) which presented a similar procedure, however, (Buntine & Weigend, 1991) is not referenced. Similar confidence intervals for nonlinear models have also been presented by (Seber & Wild, 1989, p. 193) using linear Taylor expansion, and the general expression for the variance of the prediction error conditioned on $\boldsymbol{x}$ reads:

$$V\{(y - \widehat{y}) \,|\, \boldsymbol{x}\} = \sigma^2 + \boldsymbol{h}(\boldsymbol{x})^\top \langle \delta\boldsymbol{w}\delta\boldsymbol{w}^\top \rangle_D \boldsymbol{h}(\boldsymbol{x}) \tag{3}$$

where σ^2 is the variance of inherent additive error, $\boldsymbol{h}(\boldsymbol{x}) = \partial f(\boldsymbol{x}, \boldsymbol{w})/\partial \boldsymbol{w}|_{\boldsymbol{w}=\widehat{\boldsymbol{w}}}$, $\delta\boldsymbol{w} = \boldsymbol{w} - \widehat{\boldsymbol{w}}$ is the parameter fluctuation and $\langle \delta\boldsymbol{w}\delta\boldsymbol{w}^\top \rangle_D$ the parameter covariance matrix with respect to data sets D of size N.

[1] $o(\cdot)$ is the order function: if $a(N) = o(1/N)$, then $a(N)/N \to 0$ as $N \to \infty$.

[2] $O(\cdot)$ is the Landau order function: if $a(N) = O(1/N)$ then $a(N) = \mathrm{constant}/N$.

The classical asymptotic estimate of the parameter covariance[3] used in (Rivals & Personnaz, 2000) is, $\sigma^2 \boldsymbol{J}^{-1}$, however, the use of LOO for estimating confidence intervals was mentioned in our work (Hansen & Larsen, 1996) and further addressed in (Sørensen, Nørgard, Hansen & Larsen, 1996) in which the parameter covariance is estimated by:

$$\langle \delta \boldsymbol{w} \delta \boldsymbol{w}^\top \rangle_D = \sum_{k=1}^{N} (\Delta \boldsymbol{w}_k - \overline{\Delta \boldsymbol{w}})(\Delta \boldsymbol{w}_k - \overline{\Delta \boldsymbol{w}})^\top, \tag{4}$$

where $\overline{\Delta \boldsymbol{w}} = N^{-1} \sum_{k=1}^{N} \Delta \boldsymbol{w}_k$ and $\Delta \boldsymbol{w}_k = \boldsymbol{J}^{-1} \boldsymbol{h}_k (y_k - f(\boldsymbol{x}_k, \widehat{\boldsymbol{w}}))$.

References

Buntine, W.L., & Weigend, A.S. (1991). Bayesian Back-Propagation. *Complex Systems*, 5, 603–643.

Hansen, L.K., & Larsen, J. (1996). Linear Unlearning for Cross-Validation. *Advances in Computational Mathematics*, 5, 269–280

Larsen, J. (1993). *Design of Neural Network Filters*, Ph.D. Thesis, Electronics Institute, Technical University of Denmark. Available via: `http://eivind.imm.dtu.dk/staff/jlarsen/pubs/thesis-abs.html`

Rivals, I., & Personnaz, L. (2000). Construction of confidence intervals for neural networks based on least squares estimation, *Neural Networks*, 13, 463–484.

[3] (Rivals & Personnaz, 2000) also suggest to use a sandwiched version which is valid for incomplete/biased models, see (Larsen, 1993).

Seber, G.A.F., & Wild, C.J. (1989). *Nonlinear Regression*, New York, New York: John Wiley & Sons.

Sørensen, P., Nørgård, M., Hansen L.K., & Larsen, J. (1996). Cross-Validation with LULOO, in S.I. Amari, L.W. Chan, I. King & K.S. Leung (Eds.), *Proceedings of 1996 International Conference on Neural Information Processing*, ICONIP'96, Hong Kong, (vol. 2, pp. 1305–1310).