HAL
archives-ouvertes.fr

Response to the comments by J. Larsen and L. K. Hansen for : Rivals I. Personnaz L. (2000), Construction of confidence intervals for neural networks based on least squares estimation (Neural Networks 13)

Isabelle Rivals, Léon Personnaz

▶ To cite this version:

Isabelle Rivals, Léon Personnaz. Response to the comments by J. Larsen and L. K. Hansen for : Rivals I.

Personnaz L. (2000), Construction of confidence intervals for neural networks based on least squares estimation (Neural Networks 13). Neural Networks, Elsevier, 2002, 15 (1), pp.141-143. <hal-00798661>

HAL Id: hal-00798661
https://hal-espci.archives-ouvertes.fr/hal-00798661

Submitted on 9 Mar 2013

HAL is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L'archive ouverte pluridisciplinaire **HAL**, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d'enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.

Neural Networks15(1): 141-143.

RESPONSE TO THE COMMENTS BY J. LARSEN AND L. K. HANSEN FOR: RIVALS I. & PERSONNAZ L. (2000). CONSTRUCTION OF CONFIDENCE INTERVALS FOR NEURAL NETWORKS BASED ON LEAST SQUARES ESTIMATION (NEURAL NETWORKS 13)

Isabelle Rivals and Léon Personnaz

Équipe de Statistique Appliquée

École Supérieure de Physique et de Chimie Industrielles,

10 rue Vauquelin, 75231 Paris Cedex 05, France.

Phone: 33 1 40 79 45 45 Fax: 33 1 40 79 44 25

E-mail: Isabelle.Rivals@espci.fr

Abstract

We answer several comments made by Hansen and Larsen (2001) about our paper (Rivals & Personnaz, 2000). In this paper, we dealt with the construction of confidence intervals (CIs) for neural networks based on least squares (LS) estimation and using the linear Taylor expansion of the network output. We also suggested a method for the detection of the possible overfitting of a trained neural network, and an estimate of its leave-one-out (LOO) score that does not necessitate additional trainings. Finally, we showed that the frequentist approach we adopt compares favourably with other analytic approaches such as the conceptually very different Bayesian approach.

Keywords

Approximate leave one out score, confidence intervals, cross validation, least squares estimation, Taylor expansion.

1. On the LOO score in the case of a linear model

An exact expression of the LOO score in the linear case was established long ago (Antoniadis, Berruyer & Carmona, 1992; Efron & Tibshirani, 1993). In Rivals and Personnaz (2000), we have put forward an approximate expression of the LOO score for a nonlinear model. This expression [Eqs. (37) and (38) in Rivals and Personnaz (2000)] *is exact in the case of a linear model* [Eqs. (36) and (38) in Rivals and Personnaz (2000)]. Another approximation of the LOO score had been proposed by Hansen and Larsen [Eq. (17) in Hansen and Larsen (1996), Eq. (2) in Hansen and Larsen (2001)], which *does not coincide with the exact expression for a linear model* [Eq. (18) in Hansen and Larsen (1996)], Eq. (1) in Hansen and Larsen (2001)]. In Rivals and Personnaz (2000), we made reference to this approximation of Hansen and Larsen, and added that “unfortunately, it is not valid even in the linear case”. We meant that its not being exact in the linear case makes it problematic for the nonlinear case. Despite their comments, Hansen and

Larsen's approximation is definitely not exact in the linear case (compare Eqs. (1) and (2) in Hansen & Larsen (2001), see further the Appendix).

2. On the consistency of approximate LOO scores

Another comment of Hansen and Larsen concerns the consistency of both approximations: they claim that, unlike ours, their approximation is consistent. This claim is not justified: the theorem 2 in Hansen and Larsen (1996) merely suggests the consistency of a *theoretical limited expansion* of the squared LOO error, but it does not prove the consistency of the *approximation* of the squared LOO error they present (see the Appendix for a detailed discussion). As a matter of fact, none of the two approximations can be claimed to be $o(1/N)$ or $o(1/N^2)$, due to the Gauss-Newton approximation performed in order to apply the matrix inversion lemma. Thus, the argument of the consistency cannot be retained, and we simply prefer an approximation that coincides with the exact expression in the linear case.

3. On various approaches to the construction of CIs

The fact that CIs for nonlinear models were presented in Seber and Wild (1989) has not escaped our attention, and is made clear to the reader in Rivals and Personnaz (2000): Seber and Wild (1989) is referenced six times!

On the other hand, the paper "Bayesian back-propagation" (Buntine & Weigend, 1991) cannot be considered to present a similar procedure. In the paper, we stress the important conceptual difference between the frequentist approach we adopt, and the Bayesian approach. We take explicitly position with respect to the latter, by stating that we are interested in CIs for the regression. The work by MacKay (1992a,b) on the construction of CIs in the Bayesian framework is abundantly referenced, as well as Bishop's chapter on the subject (Bishop, 1995).

Finally, our paper is not concerned with prediction intervals: thus, the comments about Eqs. (3) and (4) of Hansen and Larsen (2001) are not to the point.

Appendix. Derivation of the approximate LOO scores

We deal with static modeling problems in the case of a noise free n-input vector $\mathbf{x} = [x_1\ x_2\ \dots\ x_n]^T$, and a noisy scalar output y_p. We consider a family of nonlinear models $\{f(\mathbf{x}, \boldsymbol{\theta}), \mathbf{x} \in \mathbb{R}^n, \boldsymbol{\theta} \in \mathbb{R}^q\}$, and a data set $\{\mathbf{x}^k, y_p^k\}_{k=1 \text{ to } N}$. A LS parameter estimate $\boldsymbol{\theta}_{LS}$ minimizes the quadratic cost function (we consider the case with no regularization):

$$J(\boldsymbol{\theta}) = \frac{1}{2} \sum_{k=1}^{N} (y_p^k - f(\mathbf{x}^k, \boldsymbol{\theta}))^2 = \frac{1}{2} (\mathbf{y}_p - \mathbf{f}(x, \boldsymbol{\theta}))^T (\mathbf{y}_p - \mathbf{f}(x, \boldsymbol{\theta}))$$

where $x = [x^1\ x^2\ \dots\ x^N]^T$ is the (N,n) input matrix. We denote by $\boldsymbol{\theta}_{LS}^{(k)}$ the LS parameter estimate on the k-th LOO set $\{\mathbf{x}^i, y_p^i\}_{i=1 \text{ to } N,\ i \neq k}$. The k-th residual r^k and the k-th LOO error e^k are defined by:

$$\begin{cases} r^k = y_p^k - f(\mathbf{x}^k, \boldsymbol{\theta}_{LS}) \\ e^k = y_p^k - f(\mathbf{x}^k, \boldsymbol{\theta}_{LS}^{(k)}) \end{cases}$$

A.1. Linear case

In the case of a linear model , we have:

$$\mathbf{f}(x, \boldsymbol{\theta}) = x\, \boldsymbol{\theta}$$

The LOO error e^k is a function of r^k and of the k-th diagonal element of the orthogonal projection matrix $p_x = x (x^T x)^{-1} x^T$ on the range of x:

$$e^k = \frac{r^k}{1 - [p_x]_{kk}}$$

Hence:

$$(e^k)^2 = \frac{(r^k)^2}{(1 - [p_x]_{kk})^2} \qquad \text{(A1)}$$

The diagonal terms of p_x verify $0 \leq [p_x]_{kk} \leq 1$; we assume $[p_x]_{kk} < 1$.

A.2. Nonlinear case

In the nonlinear case, let us consider limited expansions of $(e^k)^2$ in $\Delta\boldsymbol{\theta}^{(k)} = \boldsymbol{\theta}_{LS}^{(k)} - \boldsymbol{\theta}_{LS}$.

First order expansion of $(e^k)^2$ in $\Delta\boldsymbol{\theta}^{(k)}$:

$$(r^k)^2 + \frac{\partial (r^k)^2}{\partial \boldsymbol{\theta}^T}\bigg|_{\boldsymbol{\theta}_{LS}} \Delta\boldsymbol{\theta}^{(k)} = (r^k)^2 + 2\, r^k \frac{\partial r^k}{\partial \boldsymbol{\theta}^T}\bigg|_{\boldsymbol{\theta}_{LS}} \Delta\boldsymbol{\theta}^{(k)} \qquad \text{(A2)}$$

Second order expansion of $(e^k)^2$ in $\Delta\boldsymbol{\theta}^{(k)}$:

$$\begin{aligned}
&(r^k)^2 + 2\, r^k \left.\frac{\partial r^k}{\partial \boldsymbol{\theta}^T}\right|_{\boldsymbol{\theta}_{LS}} \Delta\boldsymbol{\theta}^{(k)} + \frac{1}{2}\left(\Delta\boldsymbol{\theta}^{(k)}\right)^T \left.\frac{\partial^2 (r^k)^2}{\partial\boldsymbol{\theta}\partial\boldsymbol{\theta}^T}\right|_{\boldsymbol{\theta}_{LS}} \Delta\boldsymbol{\theta}^{(k)} \\
&= (r^k)^2 + 2\, r^k \left.\frac{\partial r^k}{\partial \boldsymbol{\theta}^T}\right|_{\boldsymbol{\theta}_{LS}} \Delta\boldsymbol{\theta}^{(k)} + \left(\Delta\boldsymbol{\theta}^{(k)}\right)^T \left(\left.\frac{\partial r^k}{\partial \boldsymbol{\theta}}\right|_{\boldsymbol{\theta}_{LS}} \left.\frac{\partial r^k}{\partial \boldsymbol{\theta}^T}\right|_{\boldsymbol{\theta}_{LS}} + \left.\frac{\partial^2 r^k}{\partial\boldsymbol{\theta}\partial\boldsymbol{\theta}^T}\right|_{\boldsymbol{\theta}_{LS}}\right) \Delta\boldsymbol{\theta}^{(k)} \\
&= (r^k)^2 + 2\, r^k \left.\frac{\partial r^k}{\partial \boldsymbol{\theta}^T}\right|_{\boldsymbol{\theta}_{LS}} \Delta\boldsymbol{\theta}^{(k)} + \left(\left.\frac{\partial r^k}{\partial \boldsymbol{\theta}^T}\right|_{\boldsymbol{\theta}_{LS}} \Delta\boldsymbol{\theta}^{(k)}\right)^2 + \left(\Delta\boldsymbol{\theta}^{(k)}\right)^T \left.\frac{\partial^2 r^k}{\partial\boldsymbol{\theta}\partial\boldsymbol{\theta}^T}\right|_{\boldsymbol{\theta}_{LS}} \Delta\boldsymbol{\theta}^{(k)} \\
&= \left(r^k + \left.\frac{\partial r^k}{\partial \boldsymbol{\theta}^T}\right|_{\boldsymbol{\theta}_{LS}} \Delta\boldsymbol{\theta}^{(k)}\right)^2 + \left(\Delta\boldsymbol{\theta}^{(k)}\right)^T \left.\frac{\partial^2 r^k}{\partial\boldsymbol{\theta}\partial\boldsymbol{\theta}^T}\right|_{\boldsymbol{\theta}_{LS}} \Delta\boldsymbol{\theta}^{(k)}
\end{aligned} \tag{A3}$$

However, an exact expression of $\Delta\boldsymbol{\theta}^{(k)}$ is not available. In Rivals and Personnaz (2000) (Eq. (B7)) as well as in Hansen and Larsen (1996) [Eqs. (8), (15) and (16) for a non regularized LS cost function], the following approximation of $\Delta\boldsymbol{\theta}^{(k)}$ is used:

$$\widehat{\Delta\boldsymbol{\theta}^{(k)}} = -\left(z^T z\right)^{-1} \mathbf{z}^k \frac{r^k}{1 - [p_z]_{kk}} \tag{A4}$$

where $\mathbf{z}^k = \partial f(\mathbf{x}^k, \boldsymbol{\theta})/\partial\boldsymbol{\theta}\,|_{\boldsymbol{\theta}=\boldsymbol{\theta}_{LS}}$ and $p_z = z\left(z^T z\right)^{-1} z^T$ denotes the orthogonal projection matrix on the range of the (N,n) Jacobian matrix $z = \left[\mathbf{z}^1\ \mathbf{z}^2\ \ldots\ \mathbf{z}^N\right]^T$. This result is obtained when approximating the Hessian with the squared Jacobian (Gauss-Newton approximation):

$$h = \left.\frac{\partial^2 J}{\partial\boldsymbol{\theta}\ \partial\boldsymbol{\theta}^T}\right|_{\boldsymbol{\theta}_{LS}} = \sum_{k=1}^{N} \mathbf{z}^k \left(\mathbf{z}^k\right)^T + \sum_{k=1}^{N} r^k \left.\frac{\partial^2 r^k}{\partial\boldsymbol{\theta}\ \partial\boldsymbol{\theta}^T}\right|_{\boldsymbol{\theta}_{LS}} \approx \sum_{k=1}^{N} \mathbf{z}^k \left(\mathbf{z}^k\right)^T = z^T z \tag{A5}$$

This approximation cannot be characterized in terms of an order function unless the model is true, i.e. unless there exists a value $\boldsymbol{\theta}_p$ of $\boldsymbol{\theta}$ such that $f(\mathbf{x}, \boldsymbol{\theta}_p) = E(Y|\mathbf{x})$; in that case, $E(H) = z^T z$. Naturally, this assumption cannot be made for any candidate model.

1) Hansen and Larsen approximate the first order expansion of $(e^k)^2$ [Eq. (A2)] by replacing $\Delta\boldsymbol{\theta}^{(k)}$ with $\widehat{\Delta\boldsymbol{\theta}^{(k)}}$:

$$\widehat{(e^k)^2}_{H\&L} = (r^k)^2 - 2\, r^k \left(\mathbf{z}_k\right)^T \widehat{\Delta\boldsymbol{\theta}^{(k)}} = (r^k)^2 \left(1 + \frac{2\,[p_z]_{kk}}{1 - [p_z]_{kk}}\right) = (r^k)^2 \left(\frac{1 + [p_z]_{kk}}{1 - [p_z]_{kk}}\right) \tag{A6}$$

Note that, in the linear case, $z = x$ and $p_z = p_x$ and hence $\widehat{(e^k)^2}_{H\&L} \neq (e^k)^2$ [Eq. (A6)] does not coincide with Eq. (A1)). It is necessary to expand $(e^k)^2$ up to the order two for the expansion to be exact in the linear case.

2) We use the square of the first order expansion of e^k [Eq. (2)], and replace $\Delta\boldsymbol{\theta}^{(k)}$ with $\widehat{\Delta\boldsymbol{\theta}^{(k)}}$:

$$\widehat{(e^k)^2}_{R\&P} = \left(r^k + \left. \frac{\partial r^k}{\partial \boldsymbol{\theta}^T} \right|_{\boldsymbol{\theta}_{LS}} \widehat{\Delta\boldsymbol{\theta}^{(k)}} \right)^2 = (r^k)^2 \frac{1}{(1 - [p_z]_{kk})^2} \qquad \text{(A7)}$$

This is an approximation of the second order expansion of $(e^k)^2$ [Eq. (A3)] which neglects the term $(\Delta\boldsymbol{\theta}^{(k)})^T \, \partial^2 r^k / \partial\boldsymbol{\theta}\partial\boldsymbol{\theta}^T \big|_{\boldsymbol{\theta}_{LS}} \Delta\boldsymbol{\theta}^{(k)}$.

Note that, in the linear case, $\widehat{(e^k)^2}_{R\&P} = (e^k)^2$ (Eq. (A7) coincides with Eq. (A1)).

To conclude, $\widehat{(e^k)^2}_{H\&L}$ is an approximation of the first order expansion of $(e^k)^2$ in $\Delta\boldsymbol{\theta}^{(k)}$, and $\widehat{(e^k)^2}_{R\&P}$ is an approximation of the second order expansion of $(e^k)^2$ in $\Delta\boldsymbol{\theta}^{(k)}$. Both approximations replace $\Delta\boldsymbol{\theta}^{(k)}$ par $\widehat{\Delta\boldsymbol{\theta}^{(k)}}$ due to the Gauss-Newton approximation, and in $\widehat{(e^k)^2}_{R\&P}$, a part of the second order term is neglected due to a similar approximation. Since $\widehat{\Delta\boldsymbol{\theta}^{(k)}}$ cannot be characterized in terms of any order function (the object of the Theorem 2 in Hansen and Larsen (1996) is not $\Delta\boldsymbol{\theta}^{(k)}$ but the theoretical unknown $\Delta\boldsymbol{\theta}^{(k)}$[1]), it is impossible to state that either of these approximations is consistent. We simply prefer an approximation that coincides with the exact expression in the linear case.

References

Antoniadis A., Berruyer J. & Carmona R. (1992). Régression non linéaire et applications. Paris: Economica.

Bishop M. (1995). Neural Networks for Pattern Recognition. Oxford: Clarendon Press.

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

Efron B. & Tibshirani R. J. (1993). An introduction to the bootstrap. New York: Chapman & Hall.

Hansen L. K. & Larsen J. (1996). Linear unlearning for cross validation. Advances in computational mathematics **5**, 296-280.

---

[1] Note that the proof of Theorem 2 in Hansen and Larsen (1996) is rather elliptic.

Hansen L. K. & Larsen J. (2001). Comments for: Rivals I. & Personnaz L. (2000). Construction of confidence intervals for neural networks based on least squares estimation (Neural Networks **13**). Neural Networks ?, ?-?.

MacKay D. J. C. (1992a). Bayesian interpolation. Neural computation **4**, 415-447.

MacKay D. J. C. (1992b). A practical Bayesian framework for backprop networks. Neural computation **4**, 448-472.

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

Seber G. A. F. & Wild C. (1989). Nonlinear regression. New York: Wiley.