Bayesian Multioutput Feedforward Neural Network Comparison: A Conjugate Prior Approach

Vivien Rossi and Jean-Pierre Vila, *Member, IEEE*,

***Abstract*— A Bayesian method for the comparison and selection of multi-output feedforward neural network topology, based on the predictive capability, is proposed [1]. As a measure of the prediction fitness potential, an expected utility criterion is considered which is consistently estimated by a sample-reuse computation. As opposed to classic point-prediction-based cross-validation methods, this expected utility is defined from the logarithmic score of the neural model predictive probability density. It is shown how the advocated choice of a conjugate probability distribution as prior for the parameters of a competing network, allows a consistent approximation of the network posterior predictive density. A comparison of the performances of the proposed method with the performances of usual selection procedures based on classic cross-validation and information-theoretic criteria, is performed first on a simulated case study, and then on a well-known food analysis dataset.**

***Index Terms*— Feedforward neural network, Bayesian model selection, conjugate prior distribution, empirical Bayes methods, expected utility criterion.**

I. Introduction

The issue of selecting a right network topology is one of the most debated in feedforward multilayer neural network modeling. A bias/variance trade-off has to be satisfied [16],[7], to get close to some optimal model complexity (number of layers and neurons) protecting as most as possible from both contradictory effects of overfitting and underfitting. However, the dynamics of the bias and variance errors can in general be only estimated through estimation of a huge amount of varied neural network models and comparison on a test data set, which is hardly feasible in practice. Less time-consuming on-line and off-line strategies have then be proposed, which belong to several classes, heuristically or more statistically oriented. Constructive algorithms [19] come within the first category, while more complex constructive-destructive methods often come within the second [30]. More general methods are the so-called regularization techniques [29], based on implicit structure optimization [28]. They consider a fixed topology and they constrain the network parameters in some way, for example by adding penalty (or weight decay) terms to the cost function, in order to avoid saturation of the units.

Regularization techniques using penalty term addition, can be considered as statistically Bayesian since this penalty can be associated with a prior probability on the weight and bias parameters [9]. The well-known MacKay's Bayesian framework for backpropagation [22], originally designed for single output networks, comes within these approaches. For a given neural model this Bayesian formalism leads to the so-called *evidence*, which estimates how likely the model is given the available dataset, and thus can be used in model selection. In the past recent years the MacKay's paradigm has been used in several applied fields, such as mathematical finance [25], [17], [40], for example.

However these approaches present several critical points such as the imprecise relationship between the generalization performance of the network and the advocated Bayesian selection criteria [22], [36], the treatment of the hyperparameters [9], [39], [23], the usually assumed independence of the weight Gaussian priors, and more crucially, scarcely controllable approximations of the posteriors. The evidence method for example relies on a critical Gaussian approximation of the posterior parameter distribution (in which the hyperparameters, the weight decay terms, are fixed to the values maximizing the evidence). It has been observed that this approximation breaks down when the ratio of the dataset size to the number of network parameters is too small [22], [36]. Markov chain Monte Carlo methods have been proposed to replace this Gaussian approximation [27], requiring now skilled simulation expertise and greater computing.

More general and statistically oriented methods are the information-theoretic model comparison criteria [10] as Akaike's AIC, BIC, Mallow's C_p, RIC and NIC [26], which also combine some measure of fit with a penalty term to account for model complexity. However it has been observed that according to the situation considered, the performance of these criteria is rather sensitive to the type of penalty [35] especially in the case of neural networks [2].

Other statistical tools such as asymptotic inferentials tests, $e.g.$ likelihood ratio, Wald's or Rao's Lagrange multiplier tests [33], can also be used to compare feedforward neural models but they are explicitly restricted to the comparison of nested neural models.

Finally, one of the still most attractive comparison procedure, even if computer-intensive, is cross-validation (CV), because of its genericity and limited probability assumption requirements ($e.g.$ exchangeability assumption). However, it has been shown that CV can be inconsistent (unless appropriate data division is done, [34]), as are the methods asymptotically equivalent to it ($e.g.$ AIC and C_p). Moreover, CV is often too conservative and tends to select unnecessary large models.

To counteract these defects, a CV-like Bayesian nonlinear model comparison procedure, inspired by a classic utility criterion [6], has been developed and adjusted to the issue of the comparison of single-output feedforwad neural networks

The authors are with UMR Analyse des Systèmes et Biométrie, INRA-ENSAM, 2 Place P. Viala, 34060 Montpellier, France

[1]This paper is a full extension of the contributed paper *Multioutput Feedforward Neural Network Selection: A Bayesian approach*, given by the authors to the IEEE-INNS 2003 International Joint Conference on Neural Networks (p.495-500 in the proceedings)

[37]. An extension of this Bayesian approach to multiresponse regression models has been designed recently [31].

The present paper proposes an adaptation of this approach to the comparison of multioutput multilayer neural networks but with a specific recourse to the Bayesian conjugate prior theory and to the so-called empirical Bayes approach. This particular Bayesian framework offers the advantage of allowing to introduce data-respectful priors with the possibility of a complete analytical treatment of the posterior and predictive densities. Moreover, the method allows comparison of networks differing by their respective topologies as well as by their input variable sets. For a given data set, this predictive network performance comparison approach can be used to select among neural nets of varied complexity the net achieving the best compromise between complexity and generalization. As a matter of fact, it could seem that after several valuable works such as [27], [18] for example, the predictive accuracy of a Bayesian network is not sensitive to the number of hidden units (of course given enough units not to underfit) and that there is no need to try optimizing their number and organization, provided sensible priors and adequate posterior approximations are used for the network parameters. However from a practical point of view and from that of the general non Bayesian neural net user, there is still a need of simple efficient standard neural net comparison and selection tools, sufficiently generic to relieve as much as possible of any specialized and controversial issues as hierarchisation of net parameters, relevant choices of parameter and hyperparameter priors, efficient design of posterior approximations and relieving also of the specialized simulating estimation apparatus they involved ($e.g.$ MCMC techniques).

In this spirit, this paper puts the emphasis on a general analytical approach, leading to a well known closed form for a consistent approximation of the neural net predictive density and confining all numerical aspects only to the final evaluation of the proposed utility criterion.

The paper is organized as follows. In Section II the statistical framework of the neural model selection problem is set up. In Section III the building elements of the expected-utility-based criterion are considered. Convergent approximations of the parameter posterior and posterior predictive densities, allowing the sample-reuse calculation of the expected utility, are developed in Section IV. Section V shows how this predictive density estimation procedure can easily be adapted to take full account of the structural multi-modality of the likelihood function of a feedforward neural model. In Section VI this Bayesian procedure is applied to a simulated predictive neural network selection problem and then to a well known bench-mark test in spectroscopy. The performances of the procedure are compared with that of AIC, BIC and classic CV procedures. Appendix I briefly recalls elements of Bayesian theory used in the construction of the expected utility criterion. Appendix II provides the proofs of the convergence of the approximations of the required parameter posterior and posterior predictive densities used in the criterion.

II. Multi-output feedforward neural modeling framework

A multi-ouput feedforward neural network model M is a multi-response nonlinear regression model which under the assumption of Gaussian additive errors, can be written as

$$\text{Model } M: \qquad y_i = f(x_i, \theta) + \varepsilon_i \tag{1}$$

where the nonlinear mapping f results from the network topology with x as input vector and θ as parameter vector (the set of all weights and biases of the network), [29].
For $i \in \{1, \ldots, n\}$: $y_i \in \mathbb{R}^d$, $x_i \in \mathbb{R}^l$, $\theta \in \Theta$ a compact subset of $\mathbb{R}^q$, $\varepsilon_i \sim N_d(0, \Sigma)$ with $\Sigma \in \mathcal{S} \subset \mathbb{R}^{d\times d}$ where $\mathcal{S}$ is the set of all positive definite symmetric matrices of dimension $d \times d$. We shall have to consider more often than the variance-covariance matrix Σ, the precision matrix $\Lambda = \Sigma^{-1}$.
Let us denote

- $Z_n = (x_i, y_i), i = 1, \ldots, n,$ the available data set, made of n $i.i.d$ random data points (x, y).
- $y_{1:n} = (y_1, \ldots, y_n)$ and $x_{1:n} = (x_1, \ldots, x_n)$.
- $\dot{f}_{x_i,\theta} = \left(\frac{\partial f(x_i,\theta)}{\partial \theta_j}\right) \quad j \in \{1, \ldots, q\}$. We shall suppose that these derivatives exist for all the neural networks considered.

Other notations :

- $\mathcal{N}_q(\cdot|\mu, \Sigma)$: the q-dimensional Gaussian probability density with expectation μ and covariance matrix Σ.
- $\mathcal{W}i_d(\cdot|\alpha, \beta)$: the d-dimensional Wishart density with parameters α and β.
- $St_d(\cdot|\mu, \Psi, \alpha)$: the d-dimensional Student density with parameters μ, Ψ and α.
- To alleviate notations, integration with respect to θ and Λ over their whole membership set $\Theta \times \mathcal{S}$, will be denoted throughout the paper by $\int$ instead of $\int_{\Theta\times\mathcal{S}}$.

Given Z_n and a set $\mathcal{M}$ of J feedforward neural models $\{M^j, j = 1, \ldots, J\}$, with $\mathrm{E}(y|M^j, x) = f^j(x, \theta^j)$, the issue of interest is to select the best neural model, M^*, in some predictive sense.

III. The expected-utility-based criterion

To do this selection we follow the $maximum-expected-utility$ approach [6] for which the optimal model choice is M^* such that

$$\bar{u}(M^*|Z_n) = \sup_{M^j\in\mathcal{M}} \bar{u}(M^j|Z_n) \tag{2}$$

where

$$\bar{u}(M^j|Z_n) = \int u(M^j, y, x|Z_n) p\left((x, y)|Z_n\right) \mathrm{d}y\mathrm{d}x \tag{3}$$

in which $u(M^j, y, x|Z_n)$ is a given utility function and $p\left((x, y)|Z_n\right)$ is a probability density representing actual beliefs about (x, y) having observed Z_n.
But $p\left((x, y)|Z_n\right)$ in (3) is generally not available. We then search for a consistent estimate of $\bar{u}(M^j|Z_n)$ for each $M^j \in \mathcal{M}$. Following Bernardo and Smith [6] we consider the

n partitions of Z_n: $Z_n = [Z_{n-1}(i), (x_i, y_i)]$ for $1 \leq i \leq n$, where $Z_{n-1}(i)$ denotes the data set Z_n after withdrawal of the data point (x_i, y_i). If we select k of these data points at random (without replacement), we have by the strong law of large numbers under regular assumptions, as n, k grow to infinity [6], [31]

$$\Big| \tfrac{1}{k} \sum_{i=1}^{k} u(M^j, y_i, x_i | Z_{n-1}(i)) - \int u(M^j, y, x | Z_n) p\left((x,y)|Z_n\right) \mathrm{d}y \mathrm{d}x \Big| \xrightarrow{a.s.} 0$$

The expected utility of model $M^j \in \mathcal{M}$ can then be consistently approximated by

$$U_j = \frac{1}{k} \sum_{i=1}^{k} u(M^j, y_i, x_i | Z_{n-1}(i)) \tag{4}$$

Furthermore as we are interested in comparing models from a predictive distribution point of view, as suggested in [6] we take as utility function the logarithmic score

$$u(M^j, y, x | Z_n) = \log p(y | M^j, x, Z_n) \tag{5}$$

In (5) $p(y|M^j, x, Z_n)$ is the posterior predictive density under model M^j of a response y at x, given the past observations Z_n and an appropriate prior density for the neural network model M^j parameters. Let us note that with this choice for the utility function, (4) is similar to the predictive sample reuse criterion of [14] which considers the product of conditional predictive densities.

We then decide to take as M^* the model $M^j \in \mathcal{M}$ maximizing

$$U_j = \frac{1}{k} \sum_{i=1}^{k} \log p\left(y_i | M^j, x_i, Z_{n-1}(i)\right) \tag{6}$$

This procedure selects on a sample-reuse basis, the model under which the data set Z_n achieves the highest level of some internal consistency: the best model is that which on the whole, most favors the likelihood of each observation with respect to the others.

The next section will be devoted to the calculation of a convergent approximation $\hat{p}$ of the parameter posterior predictive density p, for each neural network model M^j, leading to the practical criterion

$$\widehat{U}_j = \frac{1}{k} \sum_{i=1}^{k} \log \hat{p}\left(y_i | M^j, x_i, Z_{n-1}(i)\right) \tag{7}$$

such that, given k

$$\widehat{U}_j \overset{n \to \infty}{\longrightarrow} U_j \quad a.s. \tag{8}$$

IV. Posterior predictive densities: a consistent approximation

In order to compute (6) we need a posterior predictive density for the response at a given x, under model M^j, conditional to the training set Z_n, for each $M^j \in \mathcal{M}$.

For a given network M as in (1), such a posterior is defined by

$$p(y|x, \mathcal{T}, Z_n) = \int p(y|x, \theta, \Lambda) p(\theta, \Lambda | \mathcal{T}, Z_n) \mathrm{d}\theta \mathrm{d}\Lambda \tag{9}$$

In (9) $p(y|x, \theta, \Lambda)$ is given by model (1) and $p(\theta, \Lambda | \mathcal{T}, Z_n)$ is a (θ, Λ) posterior probability density with $\mathcal{T}$ a vector of hyperparameters. This (θ, Λ) posterior density is obtained by Bayes' theorem from a given (θ, Λ) prior density $p(\theta, \Lambda | \mathcal{T})$:

$$p(\theta, \Lambda | \mathcal{T}, Z_n) = \frac{p(Z_n | \theta, \Lambda) p(\theta, \Lambda | \mathcal{T})}{\int p(Z_n | \theta, \Lambda) p(\theta, \Lambda | \mathcal{T}) \mathrm{d}\theta \mathrm{d}\Lambda} \tag{10}$$

For a given (θ, Λ) prior, the computation of the posterior (10) is generally untractable. One possible approach to consistenly estimate (9) is to use a technique of Bayesian learning for neural network. These techniques are based on (θ, Λ) sampling from an MCMC $p(\theta, \Lambda | \mathcal{T}, Z_n)$ posterior density estimation (see for example [18], [12], [20]). However, such MCMC integrations frequently suffer from instability [15] which can impair the relevance of the final utility criterion estimation. In addition, another major and preliminary difficulty of this Bayesian training approach is of course the (θ, Λ) prior choice itself, which in spite of several attractive approaches [21], [27], remains a critical issue lacking from a general response easy to handle especially for non Bayesians. These difficulties led us to consider an analytical treatment of the parameter posterior and predictive posterior densities estimations, from a well known class of parameter priors.

A. (θ, Λ) prior density

Let us note that under the assumptions of model (1) the probability density of $(y_{1:n} | x_{1:n}, \theta, \Lambda)$ belongs to the exponential family:

$$\begin{aligned} & p(y_{1:n} | x_{1:n}, \theta, \Lambda) \\ &= \frac{|\Lambda|^{n/2}}{(2\pi)^{nd/2}} \exp\Big\{ -\frac{1}{2} \sum_{i=1}^{n} \|y_i - f(x_i, \theta)\|_{\Lambda}^2 \Big\} \\ &= c \times g(\theta, \Lambda) \times \exp\Big\{ -\frac{1}{2} tr\Big[\big(\sum_{i=1}^{n} y_i y_i' \big) \Lambda \Big] \\ &\quad + \sum_{i=1}^{n} f(x_i, \theta)' \Lambda y_i \Big\} \end{aligned} \tag{11}$$

with $c = \frac{1}{(2\pi)^{-nd/2}}$ and $g(\theta, \Lambda) = |\Lambda|^{n/2} \exp\Big\{ -\frac{1}{2} \sum_{i=1}^{n} \|f(x_i, \theta)\|_{\Lambda}^2 \Big\}$.

This suggests to take as (θ, Λ) prior density the *c*onjugate density with respect to the likelihood $p(y_{1:n} | x_{1:n}, \theta, \Lambda)$, thus ensuring tractability of the related posterior. Actually, the fundamental advantage of a conjugate prior density is to provide very easily the related posterior density since, because of a closure property, both densities belong to the same family of probability distribution [3].

From (11) and by definition of conjugate families for regular exponential families of probability distributions, we have easily

$$
\begin{aligned}
& p(\theta,\Lambda|\mathcal{T}) \\
&= K[\mathcal{T}]^{-1}[g(\theta,\Lambda)]^{\tau_0}\exp\Big\{-\frac{1}{2}tr\Big[\mathcal{T}_1\,\Lambda\Big] \\
&\quad +\sum_{i=1}^{n} f(x_i,\theta)'\Lambda\mathcal{T}_2^i\Big\} \\
&= K[\mathcal{T}]^{-1}|\Lambda|^{\tau_0 n/2}\exp\Big\{-\frac{\tau_0}{2}\sum_{i=1}^{n}\|f(x_i,\theta)\|^2_\Lambda \\
&\quad +\sum_{i=1}^{n} f(x_i,\theta)'\Lambda\mathcal{T}_2^i-\frac{1}{2}tr\Big[\mathcal{T}_1\,\Lambda\Big]\Big\} \qquad (12)
\end{aligned}
$$

with $\mathcal{T}_1$ a $d\times d$ symmetric matrix, $\mathcal{T}_2^i$ a $\mathbb{R}^d$ vector for $i=1,\ldots,n$, $K[\mathcal{T}]^{-1}$ a normalizing constant and $\mathcal{T}=(\tau_0,\ \mathcal{T}_1,\ \mathcal{T}_2^1,\ldots,\mathcal{T}_2^n)$, a set of hyperparameters.
Interpretation of $p(\theta,\Lambda|\mathcal{T})$: see final remark of § IV.C.

B. (θ,Λ) posterior density

Under model M the parameter posterior density associated to the prior $p(\theta,\Lambda|\mathcal{T})$ is then given by (see Appendix I):

$$
p(\theta,\Lambda|Z_n,\mathcal{T}) = p(\theta,\Lambda|\mathcal{T}+t(y_{1:n})) \qquad (13)
$$

with $\mathcal{T}+t(y_{1:n})=(\tau_0+1,\mathcal{T}_1+\sum_{i=1}^n y_i y_i',\ \mathcal{T}_2^1+y_1,\ldots,\mathcal{T}_2^n+y_n)$
From (12) and (13) we have

$$
\begin{aligned}
& p(\theta,\Lambda|Z_n,\mathcal{T}) \\
&= K[\mathcal{T}+t(y_{1:n}]^{-1}|\Lambda|^{(\tau_0+1)n/2} \\
&\quad\times\exp\Big\{-\frac{\tau_0+1}{2}\sum_{i=1}^{n}\|f(x_i,\theta)\|^2_\Lambda \\
&\quad +\sum_{i=1}^{n} f(x_i,\theta)'\Lambda(\mathcal{T}_2^i+y_i) \\
&\quad -\frac{1}{2}tr\Big[(\mathcal{T}_1+\sum_{i=1}^{n} y_i y_i')\,\Lambda\Big]\Big\} \qquad (14)
\end{aligned}
$$

At this point, we have to decide how to treat the hyperparameters $\mathcal{T}$: we could try to integrate them out but under the problematic choice of a second level prior and other possible drawbacks [23]. In the present case, a more tractable and natural approach is to optimize them by maximizing the prior density of the observations themselves:

$$
p(y_{1:n}|x_{1:n},\mathcal{T}) = \int p(y_{1:n}|x_{1:n},\theta,\Lambda)p(\theta,\Lambda|\mathcal{T})\mathrm{d}\theta\mathrm{d}\Lambda \qquad (15)
$$

It can be shown that

$$
\begin{aligned}
p(y_{1:n}|x_{1:n},\mathcal{T}) &= \Pi_{i=1}^n p(y_i|x_i,\mathcal{T}) \\
&\overset{n\to\infty}{\asymp} \Pi_{i=1}^n \mathcal{N}_d\Big(y_i|f(x_i,\theta_0),\frac{a}{2}\beta(\mathcal{T})^{-1}\Big) \qquad (16)
\end{aligned}
$$

where the ith factor in the right-hand side is the value at y_i of the d-dimensional normal density with mean $f(x_i,\theta_0)$ and inverse covariance matrix $\frac{a}{2}\beta(\mathcal{T})^{-1}$, with

- $a=n\tau_0+2$
- $\beta(\mathcal{T})=\frac{1}{2}\sum_{i=1}^n\Big(\tau_0 f(x_i,\theta_0)f(x_i,\theta_0)'-f(x_i,\theta_0)\mathcal{T}_2^{i\prime}-\mathcal{T}_2^i f(x_i,\theta_0)'\Big)+\frac{1}{2}\mathcal{T}_1$
- $\theta_0=\mathrm{argmin}_\theta\det\Big[\sum_{i=1}^n\Big(\mathcal{T}_2^i/\tau_0-f(x_i,\theta)\Big)\Big(\mathcal{T}_2^i/\tau_0-f(x_i,\theta)\Big)'\Big]$

$p(y_{1:n}|x_{1:n},\mathcal{T})$ is then asymptotically maximized by

$$
\tau_0=1,\ \mathcal{T}_1=\sum_{i=1}^n y_i y_i',\ \mathcal{T}_2^i=y_i,\ i=1,\ldots,n \qquad (17)
$$

a setting under which θ_0 and $\beta(\mathcal{T})$ are equal to $\hat{\theta}_n$ and $\frac{n}{2}\widehat{\Lambda}_n^{-1}$, where $\hat{\theta}_n$ and $\widehat{\Lambda}_n$ are the maximum likelihood estimates of θ and Λ and given by [33]:

$$
\begin{aligned}
\widehat{\theta}_n &= \mathrm{argmin}_\theta\det\Big[\sum_{i=1}^n\Big(y_i-f(x_i,\theta)\Big)\Big(y_i-f(x_i,\theta)\Big)'\Big] \\
\widehat{\Lambda}_n^{-1} &= \widehat{\Sigma}_n \\
&= \frac{1}{n}\sum_{i=1}^n\Big(y_i-f(x_i,\hat{\theta}_n)\Big)\Big(y_i-f(x_i,\hat{\theta}_n)\Big)' \qquad (18)
\end{aligned}
$$

An intuitive idea of this optimal setting can be reached from (15) by seeing that $p(y_{1:n}|x_{1:n},\mathcal{T})\le p(y_{1:n}|x_{1:n},\hat{\theta}_n,\hat{\Lambda}_n)$. The maximization of $p(y_{1:n}|x_{1:n},\mathcal{T})$ will be favored, as n grows to infinity, by choosing a setting for $\mathcal{T}$ such that the prior density $p(\theta,\Lambda|\mathcal{T})$ loads more and more in priority a neighborhood of $(\hat{\theta}_n,\hat{\Lambda}_n)$. A simple look at (12) and (11) shows that this will be achieved by the setting (17). Let us note that this setting is related to the so-called empirical Bayes approach [24].
From now on, we shall only consider the setting (17) for the hyperparameters and thus $\mathcal{T}$ will not appear any more in the expression of the prior and posterior densities of the parameters. We then have from (14)

$$
p(\theta,\Lambda|Z_n)=K_n|\Lambda|^n\exp\Big\{-\sum_{i=1}^n\|y_i-f(x_i,\theta)\|^2_\Lambda\Big\} \qquad (19)
$$

where $K_n=K^{-1}[\mathcal{T}+t(y_{1:n})]$ is the normalizing constant. However with a parameter posterior as (19) the computation of the posterior predictive density (9) will be intractable for a general neural model f. Let us consider then a convergent approximation of $p(\theta,\Lambda|Z_n)$ allowing the computation of a convergent approximation of $p(y|x,Z_n)$ under model M.

C. A L_1-convergent approximation of the parameter posterior density

Let $\mathcal{H}$ be the following set of assumptions for model M:

H_1 $x_i\in\mathcal{X}$ a compact subset of $\mathbb{R}^l$, $i=1,\ldots,n$.

H_2 The model function $f(x,\theta)$ is of class C^1 both in x and θ (this assumption is satisfied by usual networks with differentiable transfert function in their units).

Let $\hat{p}(\theta,\Lambda|Z_n)=$

$$
\mathcal{N}_q\Big(\theta|\hat{\theta}_n,V_\theta\Big)\mathcal{W}i_d\Big(\Lambda|n+\frac{d+1}{2},V_\Lambda\Big) \qquad (20)
$$

with

$$V_\theta = \Big(2\sum_{i=1}^n \dot{f}'_{x_i,\hat{\theta}_n}\hat{\Lambda}_n\dot{f}_{x_i,\hat{\theta}_n}\Big)^{-1}$$
$$V_\Lambda = \sum_{i=1}^n (y_i - f(x_i,\hat{\theta}_n))(y_i - f(x_i,\hat{\theta}_n))'$$

Let us recall now that under general conditions there exist limit values θ_* and Λ_* to which the maximum likelihood estimates under model M, $\hat{\theta}_n$ and $\hat{\Lambda}_n$, converge almost surely with n [38], [1]. These values are the true parameter values when model M is the correct one. When model M is incorrect (which is always the case for neural network modelling of actual data), θ_* and Λ_* are the parameter values minimizing the Kullback-Leibler information criterion between the true (x,y) data distribution and the (x,y) distribution induced by model M. Moreover the parameter posterior distribution concentrates around these limit values θ_*, Λ_* (see [4], [5] and especially [1] for details).
The following lemma extends this concentration property to the distribution of density $\hat{p}(\theta,\Lambda|Z_n)$.

Lemma 1: Suppose assumptions $\mathcal{H}$ are satisfied. Let A be a measurable set of $\Theta\times\mathcal{S}$ which contains an open neighborhood of the limit parameter values (θ_*,Λ_*). Then

$$\lim_{n\to\infty}\hat{P}(A) = 1 \quad a.s.$$

where $\hat{P}$ is the probability measure associated with the density $\hat{p}(\theta,\Lambda|Z_n)$.

This lemma ensures the consistency of $\hat{p}(\theta,\Lambda|Z_n)$, *i.e.* its asymptotic concentration at (θ_*,Λ_*).

Theorem 1: Under assumptions $\mathcal{H}$

$$\lim_{n\to\infty}\int |\hat{p}(\theta,\Lambda|Z_n) - p(\theta,\Lambda|Z_n)|\mathrm{d}\theta\mathrm{d}\Lambda = 0 \qquad a.s. \quad (21)$$

Remark: In the same way it could have been shown that

$$\int \Big|\hat{p}(\theta,\Lambda) - p(\theta,\Lambda)\Big|\mathrm{d}\theta\mathrm{d}\Lambda \overset{n\to\infty}{\longrightarrow} 0 \quad a.s. \qquad (22)$$

with

$$\hat{p}(\theta,\Lambda) = \mathcal{N}_q\Big(\theta|\hat{\theta}_n, 2V_\theta\Big)\mathcal{W}i_d\Big(\Lambda|\frac{1}{2}(n+d+1), \frac{1}{2}V_\Lambda\Big) \quad (23)$$

(23) shows that unsurprisingly the conjugate prior $p(\theta,\Lambda|\mathcal{T})$ with the setting (17) takes the form of a "data-respectful" distribution for n sufficiently large. Most remarkable is that the form of this prior approximation and that of the posterior (20) also respect the usual Bayesian choices for this kinds of parameters, confirming thus the interest of this conjugate approach.

D. A L_1-convergent approximation of the posterior predictive density

By definition

$$p(y|x,Z_n) = \int p(y|x,\theta,\Lambda)p(\theta,\Lambda|Z_n)\mathrm{d}\theta\mathrm{d}\Lambda \qquad (24)$$

Let

$$\hat{p}(y|x,Z_n) = \int \hat{p}(y|x,\theta,\Lambda)\tilde{p}(\theta,\Lambda|Z_n)\mathrm{d}\theta\mathrm{d}\Lambda \qquad (25)$$

with $\tilde{p}(\theta,\Lambda|Z_n)$ a L_1-convergent approximation of the parameter (θ,Λ) posterior density.
and

$$\hat{p}(y|x,\theta,\Lambda) = \frac{|\Lambda|^{1/2}}{(2\pi)^{d/2}}\exp\Big\{-\frac{1}{2}\|y - f(x,\hat{\theta}_n)\|^2_\Lambda\Big\} \quad (26)$$

Theorem 2: Under assumptions $\mathcal{H}$

$$\lim_{n\to\infty}\int \Big|\hat{p}(y|x,Z_n) - p(y|x,Z_n)\Big|\mathrm{d}y = 0 \qquad a.s. \quad (27)$$

Now take $\tilde{p}(\theta,\Lambda|Z_n)$ as equal to $\hat{p}(\theta,\Lambda|Z_n)$ as given by (20), and let

$$\hat{p}_n(y|x,Z_n) = \mathrm{St}_d(y|f(x,\hat{\theta}_n), \frac{n+1}{n}\hat{\Lambda}_n, 2n+2) \qquad (28)$$

Corollary 1: Under assumption $\mathcal{H}$

$$\lim_{n\to\infty}\int \Big|\hat{p}_n(y|x,Z_n) - p(y|x,Z_n)\Big|\mathrm{d}y = 0 \qquad a.s. \quad (29)$$

Proof: bringing $\tilde{p}(\theta,\Lambda|Z_n) \equiv \hat{p}(\theta,\Lambda|Z_n)$ into (25) with (26), leads easily to (28).
A tractable convergent approximation $\hat{p}(y|x,Z_n)$ of the posterior predictive density $p(y|x,Z_n)$ under model M is now available, which can be applied to each model $M^j \in \mathcal{M}$. According to (7), a consistent approximation $\widehat{U}_j$ of the expected utility of model M^j can now be computed, for $j = 1,\ldots,J$.

V. Managing the neural model likelihood multi-modality

The posterior predictive density approximation proposed in the previous section to compute the expected utility approximation $\widehat{U}$ of a given neural model M, assumes that $\hat{\theta}_n$ in (18) is the argument of the minimum of the quadratic cost function $\det\Big[\sum_{i=1}^n \Big(y_i - f(x_i,\theta)\Big)\Big(y_i - f(x_i,\theta)\Big)'\Big]$ or equivalently the argument of the maximum of the related likelihood. It has been shown that for a general likelihood function the uniqueness of this optimum is ultimately satisfied under regularity conditions as the data set size n increases [13]. But for a multilayer perceptron model there are always several families of equivalent local optima. These families are connected with two types of symmetry transformation corresponding to parameter-sign changes and neuron interchanges [11]. These transformations lead to equivalent network input-output mappings. More precisely, for a H-hidden-layer network with m_h neurons on layer h, the overall symmetry factor is $\mathrm{SF} = \Pi_{h=1}^{h=H} m_h! 2^{m_h}$ [37]. This shows that each local mode of the likelihood function (or local minima of the sum of squares surface) belongs to a class of SF equivalent optima. The total number TNC of such classes can hardly be analytically determined in general. But a reasonable exploration of the network parameter space can reveal the NC most attractive of such classes. The missing remaining classes, of lower attractiveness and lower

contribution to the topology of the likelihood surface, will not have much consequence for n sufficiently large.

Let $\hat{\theta}_{c,s}$ be the location of the s^{th} local likelihood optimum within the c^{th} class, with $1 \leq c \leq$ NC and $1 \leq s \leq$ SF. Let $\hat{p}(\theta, \Lambda|\hat{\theta}_{c,s})$ and $\hat{p}(\theta, \Lambda|Z_n, \hat{\theta}_{c,s})$ be the parameter prior and parameter posterior approximations computed respectively from (23) and (20), for $\hat{\theta}_n = \hat{\theta}_{c,s}$.

Under the assumption that the overlap between all the prior densities $\hat{p}(\theta, \Lambda|\hat{\theta}_{c,s})$ is negligible and that the NC $\times$ SF local optima have all the same probability of being reached by the parameter estimation procedure, it can be shown that a reliable approximation of the neural network posterior predictive density is instead of (28) given by

$$\hat{p}(y|x, Z_n) = \left(1 \Big/ \sum_{c=1}^{NC} K_c\right) \sum_{c=1}^{NC} K_c \hat{p}(y|x, Z_n, \hat{\theta}_c) \qquad (30)$$

where

- $\hat{p}(y|x, Z_n, \hat{\theta}_c)$ is given by (28) in which $\hat{\theta}_c$ can be taken as any of the SF equivalent local optimal arguments $\hat{\theta}_{c,s},\ 1 \leq s \leq$ SF.

$$\bullet\ K_c = \int p(y_{1:n}|x_{1:n}, \theta, \Lambda)\hat{p}(\theta, \Lambda|\hat{\theta}_c)\mathrm{d}\theta\mathrm{d}\Lambda = \frac{(2\pi)^{\frac{q}{2}} \pi^{\frac{d(d-1)}{4}} \prod_{j=1}^{d} \Gamma(\alpha + \frac{1-j}{2})}{|v|^{\alpha}} \qquad (31)$$

with $\alpha = n + \frac{d+1}{2}$ and $v = \sum_{i=1}^{n} \Big(y_i - f(x_i, \hat{\theta}_c)\Big)\Big(y_i - f(x_i, \hat{\theta}_c)\Big)' = n\widehat{\Lambda}_n^{-1}$.

Let us note that under the same assumption the minimum squared error loss prediction of the neural network at x given Z_n, is given by the mean of the posterior predictive density (30)

$$\hat{y}_{|x,Z_n} = \left(1 \Big/ \sum_{c=1}^{NC} K_c\right) \sum_{c=1}^{NC} K_c f(x, \hat{\theta}_c) \qquad (32)$$

VI. Case studies

The U-criterion as given by (7) has been compared with usual model selection criteria able to deal with correlated multioutput responses, on a simulated and on an actual neural network selection problem. In each of the following case study, N is the size of the available dataset, from which n data points are sampled at random to compute the U-criterion (7) with $k = n$ (which is of course the best choice for k but also the most costly) and the CV, AIC and BIC criteria. The MSEP (mean squared error of prediction) on the remaining $N - n$ data points is also considered but as a reference criteria.

For a given neural model M:

The CV criterion is defined as CV $= \sum_{i=1}^{n} \|y_i - f(x_i, \hat{\theta}_{n-1}[i])\|^2_{Q^{-1}}$ where $\hat{\theta}_{n-1}[i]$ is the maximum likelihood estimate of θ on $Z_{n-1}[i]$ and Q is the empirical variance-covariance matrix of the $\{y_i\}_{i=1,\ldots,n}$.

For the AIC and BIC criteria usual forms are considered [10]: AIC $= -2\log\mathcal{L}(\hat{\theta}_n, \hat{\Lambda}_n) + 2K$ and BIC $= -2\log\mathcal{L}(\hat{\theta}_n, \hat{\Lambda}_n) + K\log n$, where K is the total number of the neural model parameters and $\mathcal{L}(\hat{\theta}_n, \hat{\Lambda}_n)$ is the neural model maximum likelihood.

The MSEP is defined as MSEP $= \sum_{i=n+1}^{N} \Big(y_i - f(x_i, \hat{\theta}_n)\Big)' Q^{-1} \Big(y_i - f(x_i, \hat{\theta}_n)\Big)$, after the network has been trained on the first data subset of size n.

A. A simulated case study

Let us consider the following nine feedforward fully connected neural structures with three inputs x^1, x^2, x^3 and two outputs y^1, y^2:

- NNi: one hidden layer of i neurons (6i+2 parameters), $i = 1, \cdots, 6$.
- N6N3: two hidden layers of 6 and 3 neurons respectively (53 parameters).
- N6N8: two hidden layers of 6 and 8 neurons respectively (98 parameters).
- N7N10: two hidden layers of 7 and 10 neurons respectively (130 parameters).

$N = 1000$ data points were independently and identically simulated from network NN5 for a given set of parameter values, with $x^1 \sim \mathcal{U}[-10,\ 10]$, $x^2 \sim \mathcal{N}(3, 5^2)$, $x^3 \sim \mathcal{U}[-1,\ 7]$ and an additive Gaussian noise ε on the two outputs, $\varepsilon \sim \mathcal{N}(0, \Sigma)$ with $\Sigma = \begin{bmatrix} 1.75 & 0.8 \\ 0.8 & 2.5 \end{bmatrix}$.

The first $n = 500$ data points were used to compute the scores reached by the nine networks according to the U, CV, AIC and BIC criteria respectively. The remaining 500 data points were used to compute the MSEP of each network on this test data subset. All the results are shown in Table I (winning scores are in bold. Note that the U-criterion has to be maximized and the other three ones and the MSEP have to be minimized).

The U and CV-criteria select the right network, NN5, as does the MSEP on the test data. However, one can note that the score reached by NN5 contrasts with those of the other eight networks more sharply according to the U-criterion than according to the CV-criterion and even than according to the MSEP. On the other hand, AIC and BIC behave very badly, by simply ranking the networks according to their growing complexity. Table II concisely sums up these behaviours through the pairwise Wilcoxon rank correlation coefficients of the criteria and the MSEP on the test data set.

TABLE I

Scorings of the nine neural models according to the four criteria and the MSEP - Case study A

Networks	U	CV	AIC	BIC	MSEP
NN1	-8.5176	39.164	**435.9818**	**462.3683**	45.122
NN2	-7.9224	23.715	446.8827	493.0592	37.946
NN3	-5.0229	5.9486	452.7242	518.6906	10.707
NN4	-2.2254	1.2371	459.5458	545.3020	6.3251
NN5	**0.6812**	**1.0973**	464.6480	570.1941	**2.0014**
NN6	-0.4401	1.1437	475.6331	600.9692	2.8412
N6N3	-2.4138	1.1932	501.6511	628.2354	6.5689
N6N8	-2.2699	1.3934	601.6449	924.8800	6.2608
N7N10	-2.6198	1.3334	664.0071	1092.8	7.7527

TABLE II
WILCOXON RANK CORRELATIONS OF THE CRITERIA SCORINGS
CASE STUDY A

	U	CV	AIC-BIC	MSEP
U	1	0.9167	-0.4834	**1**
CV		1	-0.5334	**0.9167**
AIC-BIC			1	**-0.4834**

B. The spectroscopic Tecator data

The previous Bayesian approach (U-criterion) was applied to the selection of a 2-output feedforward neural network for the Tecator meat data [8], [36]. The data recorded by a Tecator spectrometer (the Infratec Food and Feed Analyser) are available in the Statlib, by courtesy of the Tecator Company and H. H. Thodberg (http://lib.stat.cmu.edu/datasets/tecator). In [37], the single-output version of the proposed approach was applied to the selection of a multilayer perceptron for the prediction of the fat content of a meat sample on the basis of its near infrared absorbance spectrum as available in the Tecator data set. The results were compared to that of the MacKay's Bayesian evidence method [22] used by Thodberg [36]. The goal is now to select a network which best predicts both the fat and protein meat contents.

1) The data: Following Thodberg recommendations the first $n = 172$ samples of the Tecator data set are used for computing the four selection criteria for each competing model. The 43 next ones are used to compute the MSEP of each model. The input variables are 13 preprocessed principal components of the spectra. The 2 output variables are the fat and protein meat contents.

2) The competing networks: 7 feedforward neural models with 13 inputs and 2 outputs are considered. These models were derived from the single-output network f_p with 3 neurons on a single hidden layer, previously selected by the U-criterion for the fat prediction problem [37]. These 7 competing neural models, f_1, f_2, f_3, f_4, f_5, f_6, f_7, are made of a single hidden layer with 1, 2, 3, 4, 5, 6 and 7 fully connected neurons respectively. Table III shows the score reached by each of the 7 models for each of the 4 criteria U, CV, AIC, BIC, on the 172 first samples of the data set and the MSEP of the related networks on the 43 next samples. One can note that the U-ranking of the networks is much closer to the MSEP-ranking, than are the other three criteria rankings. The two best networks according to the U-criterion, f_4, f_5, are also the two best ones according to the MSEP. Idem for the two worst ones, f_1, f_2. With regard to the small size of the training data set with respect to the average parameter number of the competing networks, the performance of the U-criterion is rather satisfying. That of the CV-criterion is not so good, because of the conservative trend of CV which tends to favor unduly complex structures (CV has ranked the seven networks according to their decreasing complexity). The respective AIC and BIC-rankings are even much more unsatisfying, since unsurprisingly, these two criteria have penalised too much the network complexity and have simply ranked the seven networks according to their growing complexity (in contrast to CV). The pairwise Wilcoxon rank correlation coefficients displayed by Table IV express strikingly these respective performances and confirm the quite satisfying behaviour of the U-criterion with respect to the MSEP.

TABLE III
SCORINGS OF THE SEVEN NEURAL MODELS ACCORDING TO THE FOUR CRITERIA AND THE MSEP - CASE STUDY B

Networks	U	CV	AIC	BIC	MSEP
f_1	-3.7639	1.9333	**135.1183**	**166.8199**	6.7284
f_2	-2.4911	1.1056	172.6568	232.5376	5.6680
f_3	-1.7563	0.6947	196.1045	284.1645	2.4688
f_4	**-1.5830**	0.6162	229.3208	345.5600	2.1241
f_5	-1.7277	0.6105	259.3319	403.7504	**2.0931**
f_6	-1.7796	0.6048	290.8199	463.4175	2.2661
f_7	-1.7510	**0.5678**	322.3335	523.1103	2.8895

TABLE IV
WILCOXON RANK CORRELATIONS OF THE CRITERIA SCORINGS
CASE STUDY B

	U	CV	AIC-BIC	MSEP
U	1	0.5714	-0.5714	**0.8214**
CV		1	-1	**0.5357**
AIC-BIC			1	**-0.5357**

VII. CONCLUSION

This paper shows how the richness of information and the robustness attached to predictive probability distributions can benefit to the right selection of a multi-output feedforward neural net topology. The proposed Bayesian method relies upon a convergent approximation, built from a conjugate parameter prior density, of the neural net predictive probability distribution. This predictive distribution is used to define an expected utility criterion which can be consistently estimated on a sample-reuse basis. For a given data set this criterion detects the neural model in a given set, which on the whole most favors the likelihood of each observations with respect to the others. As compared to the evidence approach, which could be readily extended to multioutput networks, our posterior density approximations are normal and Wishart rather than normal, leading to multivariate Student approximations for the predictive densities. The behaviour of the criterion is compared, on simulated and actual neural model selection problems, with the behaviours of classic model selection criteria as point-prediction-based cross validation criterion and information-based AIC and BIC criteria. Both comparisons reveal the satisfactory trade-off reached by this Bayesian criterion between fitness induced by structural neural complexity and generalization capability offered by simpler structures. Moreover the greater small-data-set robustness of the criterion with respect to that of the classic point-wise cross-validation criterion is also evidenced. Finally, because of its analytic basis, the computing cost of such a utility criterion is comparable to that of the standard cross-validation criterion and generally lower than that of the criteria based on MCMC Bayesian

learning and without the problem of efficient stopping rules met by these last criteria.

APPENDIX I
BAYESIAN PRELIMINARIES

Proposition 1 (Bernardo and Smith [6]): Let $Z = (z_1, \ldots, z_\ell)$ be a random sample from a w-dimensional regular exponential family distribution. Its likelihood is given by

$$p(Z|\phi) = \Big(\prod_{j=1}^{\ell} s(z_j)\Big) g(\phi)^{\ell} \exp\Big\{\sum_{i=1}^{w} c_i \psi_i(\phi) \sum_{j=1}^{\ell} h_i(z_j)\Big\} \tag{33}$$

then the conjugate prior density of the parameter vector ϕ has the form

$$p(\phi|\mathcal{T}) = K[\mathcal{T}]^{-1} g(\phi)^{\tau_0} \exp\Big\{\sum_{i=1}^{w} c_i \psi_i(\phi) \tau_i\Big\}, \quad \phi \in \Phi \tag{34}$$

where $\mathcal{T} = (\tau_0, \tau_1, \ldots, \tau_w)$, vector of hyperparameters, is such that

$$K[\mathcal{T}] = \int_{\Phi} g(\phi)^{\tau_0} \exp\Big\{\sum_{i=1}^{w} c_i \psi_i(\phi) \tau_i\Big\} d\phi < \infty \tag{35}$$

Proposition 2 (Bernardo and Smith [6]): Under the assumptions of Proposition 1
(i) the posterior density for ϕ is

$$p(\phi|Z, \mathcal{T}) = p(\phi|\mathcal{T} + t_\ell(Z)) \tag{36}$$

where

$$\mathcal{T} + t_\ell(Z) = \big(\tau_0 + \ell, \tau_1 + \sum_{j=1}^{\ell} h_1(z_j), \ldots, \tau_w + \sum_{j=1}^{\ell} h_w(z_j)\big)$$

(ii) the predictive density for future observations $\bar{Z} = (\bar{z}_1, \ldots, \bar{z}_m)$ is

$$\begin{aligned} p(\bar{Z}|Z, \mathcal{T}) &= p(\bar{Z}|\mathcal{T} + t_\ell(Z)) \\ &= \prod_{j=1}^{m} s(\bar{z}_j) \frac{K\big(\mathcal{T} + t_\ell(Z) + t_m(\bar{Z})\big)}{K\big(\mathcal{T} + t_\ell(Z)\big)} \end{aligned} \tag{37}$$

where

$$t_m(\bar{Z}) = \big(m, \sum_{j=1}^{m} h_1(\bar{z}_j), \ldots, \sum_{j=1}^{m} h_w(\bar{z}_j)\big)$$

The adaptation of these results to the context of multiresponse nonlinear regression introduced in Section 2 is straightforward. In this context, $\ell = 1, z_1 = y_{1:n}$ and $\dim(z_1) = nd$.

APPENDIX II
PROOFS

A. Proof of Lemma 1

Let us first show that the expectation of the probability distribution of density $\hat{p}(\theta, \Lambda|Z_n)$ converges to (θ_*, Λ_*) as defined in IV-C.
By definition of the normal and the Wishart probability distributions and by the almost sure convergence of the maximum likelihood estimators $(\hat{\theta}_n, \hat{\Lambda}_n)$ to (θ_*, Λ_*), it comes

$$\mathrm{E}_{\hat{p}}\Big(\theta, \Lambda|Z_n\Big) = \Big(\hat{\theta}_n, \frac{2n + d + 1}{2n}\hat{\Lambda}_n\Big) \overset{n\to\infty}{\longrightarrow} \Big(\theta_*, \Lambda_*\Big) \quad a.s.$$

Let us show now that the variance of the probability distribution of density $\hat{p}(\theta, \Lambda|Z_n)$ tends to zero as n grows to infinity.

• Let β_n, be the inverse of the variance-covariance matrix of θ

$$V_\theta^{-1} = \Big(2\sum_{i=1}^{n} \dot{f}'_{x_i,\hat{\theta}_n} \hat{\Lambda}_n \dot{f}_{x_i,\hat{\theta}_n}\Big)^{-1} = \beta_n$$

Let us show that β_n grows to infinity with n:

$$\begin{aligned} \beta_n &= 2\sum_{i=1}^{n} \dot{f}'_{x_i,\hat{\theta}_n} \hat{\Lambda}_n \dot{f}_{x_i,\hat{\theta}_n} \\ &= 2n\frac{1}{n}\sum_{i=1}^{n} \dot{f}'_{x_i,\hat{\theta}_n} \hat{\Lambda}_n \dot{f}_{x_i,\hat{\theta}_n} \\ &= 2n\tilde{\beta}_n \end{aligned}$$

Let

$$\dot{\beta}_n = \frac{1}{n}\sum_{i=1}^{n} \dot{f}'_{x_i,\theta_*} \Lambda_* \dot{f}_{x_i,\theta_*}$$

According to the strong law of large numbers, as the x_i are i.i.d., one has

$$\lim_{n\to\infty} \dot{\beta}_n = \mathrm{E}_x[\dot{f}'_{x,\theta_*} \Lambda_* \dot{f}_{x,\theta_*}] = \tilde{\beta} \quad a.s.$$

As the $\{x_i\}$ belong to a compact set and f is C^1, we can deduce that

$$\lim_{n\to\infty} \frac{1}{n}\sum_{i=1}^{n} \dot{f}'_{x_i,\hat{\theta}_n} \hat{\Lambda}_n \dot{f}_{x_i,\hat{\theta}_n} = \tilde{\beta} \qquad a.s.$$

Let us show that $\tilde{\beta}$ is positive definite: For all $u \in \mathbb{R}^q$, for all n,
$u'\tilde{\beta}_n u \geq 0$ and then $u'\tilde{\beta} u \geq 0$. If u is such that $u'\tilde{\beta}u = 0$, we have for all $i \in \mathbb{N}$

$$\lim_{n\to\infty} \|\dot{f}_{x_i,\hat{\theta}_n} u\|_{\hat{\Lambda}_n} = \|\dot{f}_{x_i,\theta_*} u\|_{\Lambda_*} = 0 \quad a.s.$$

$u \neq 0$, would imply that f does not depend on all the parameters θ, which contradicts the definition of f. u must then be equal to zero and $\tilde{\beta}$ is positive definite. Then, $V_\theta^{-1} \overset{n\to\infty}{\approx} n\tilde{\beta}$ and $V_\theta \overset{n\to\infty}{\longrightarrow} 0$.
• Let us study the variance of Λ :
Let λ_{ij} be the ij^{th} term of the matrix Λ which follows the Wishart distribution included in (20).

According to [32] about the Wishart distribution: $\lambda_{ii} \sim \left(\frac{2}{n}\widehat{\Lambda}_n\right)_{ii}\chi^2_{2n+d+1}$

then

$$V(\lambda_{ii}) = \frac{8\left(\widehat{\Sigma}_n\right)^2_{ii}(2n+d+1)}{n^2} \overset{n\to\infty}{\longrightarrow} 0. \quad (38)$$

Let $l_{i,j}$ be the d-dimensional vector with the i^{th} and j^{th} components equal to 1 and the others equal to 0. According to [32] $\lambda_{ii}+\lambda_{jj}+2\lambda_{ij} \sim l_{i,j}\left(\frac{2}{n}\widehat{\Lambda}_n\right)l'_{i,j}\chi^2_{2n+d+1}$ and $V(\lambda_{ii}+\lambda_{jj}+2\lambda_{ij}) = \frac{8(l_{i,j}\left(\widehat{\Lambda}_n\right)l'_{i,j})^2(2n+d+1)}{n^2} \overset{n\to\infty}{\longrightarrow} 0.$

Then $V(\lambda_{ij}) \overset{n\to\infty}{\longrightarrow} 0$.

Let A be a measurable subset of $\Theta \times \mathcal{S}$ including an open neighbourhood of (θ_*, Λ_*). There exists ϵ such that $B_\epsilon(\theta_*, \Lambda_*) \subset A$, where $B_\epsilon(\theta_*, \Lambda_*)$ is the closed parallelotope of side ϵ centered at (θ_*, Λ_*). As $(\hat{\theta}_n, \hat{\Lambda}_n)$ converge $a.s.$ to (θ_*, Λ_*), there exists $N_\epsilon \in \mathbb{N}$, such that for all $n > N_\epsilon$, $B_{\epsilon/2}(\hat{\theta}_n, \hat{\Lambda}_n) \subset B_\epsilon(\theta_*, \Lambda_*)$ $a.s.$

Then

$$\hat{P}(B_{\epsilon/2}(\hat{\theta}_n, \hat{\Lambda}_n)) \leq \hat{P}(B_\epsilon(\theta_*, \Lambda_*)) \leq \hat{P}(A) \quad a.s.$$

Let η_i, $i = 1, \ldots, q, q+1, \ldots, q+d(d+1)/2$ denote the q components of θ and the $d(d+1)/2$ components of Λ. Let $\hat{\eta}_i^n = \mathrm{E}_{\hat{p}_n}(\eta_i)$. Let $K = q + d(d+1)/2$ be the total number of network model parameters .

$$\begin{aligned}\hat{P}(B_{\epsilon/2}(\hat{\theta}_n, \hat{\Lambda}_n)) &= 1 - \hat{P}(\overline{B_{\epsilon/2}(\hat{\theta}_n, \hat{\Lambda}_n)}) \\ &= 1 - \hat{P}(\{\eta : \max_{i=1,\ldots,K} |\eta_i - \hat{\eta}_i^n| > \frac{\epsilon}{2}\})\end{aligned}$$

According to the Markov inequality, for all $i = 1, \ldots, K$, one has

$$\hat{P}(|\eta_i - \hat{\eta}_i^n| > \epsilon/2) \leq \frac{4V(\eta_i)}{\epsilon^2}$$

As we shown previously, for all $i = 1, \ldots, K$, $\lim_{n\to\infty} V(\eta_i) = 0$. Hence, for all $\varepsilon > 0$ there exists $N_i \in \mathbb{N}$ such that $\frac{4V(\eta_i)}{\epsilon^2} \leq \varepsilon$ for $n > N_i$. Let $N = \max\{N_i, i = 1, \ldots, K\}$. For all $i = 1, \ldots, K$ and all $n > N$, one has

$$\hat{P}(|\eta_i - \hat{\eta}_i^n| > \epsilon/2) \leq \epsilon$$

and then

$$\hat{P}(\max_{i=1,\ldots,K} |\eta_i - \hat{\eta}_i^n| > \epsilon/2) \leq \epsilon.$$

Finally, for all $\varepsilon > 0$, there exists $N \in \mathbb{N}$ such that for all $n > N$

$$\hat{P}(A) \geq 1 - \varepsilon \quad a.s.$$

B. Proof of Theorem 1

Let $\beta_n = \left(2\sum_{i=1}^n \dot{f}'_{x_i,\hat{\theta}_n}\hat{\Lambda}_n\dot{f}_{x_i,\hat{\theta}_n}\right)^{-1}$.

Let C be a compact subset of $\Theta \times \mathcal{S}$ including an open neighbourhood of (θ_*, Λ_*).

Let C^c be the subset of $\Theta \times \mathcal{S}$ complementary to C.

$$\begin{aligned}&\int |\hat{p}(\theta,\Lambda|Z_n) - p(\theta,\Lambda|Z_n)|\mathrm{d}\theta\mathrm{d}\Lambda \\ &= \int_C |\hat{p} - p| + \int_{C^c} |\hat{p} - p| \\ &\leq \int_C |\hat{p} - p| + \int_{C^c} \hat{p} + \int_{C^c} p \quad (39)\end{aligned}$$

By Lemma 1 $\int_{C^c} \hat{p}(\theta,\Lambda|Z_n)\mathrm{d}\theta\mathrm{d}\Lambda \overset{n\to\infty}{\longrightarrow} 0$ $a.s.$ Moreover, by consistency of posterior densities [1] $\int_{C^c} p(\theta,\Lambda|Z_n)\mathrm{d}\theta\mathrm{d}\Lambda \overset{n\to\infty}{\longrightarrow} 0$ $a.s.$ To prove the theorem it remains to show that $\int_C |\hat{p}(\theta,\Lambda|Z_n) - p(\theta,\Lambda|Z_n)|\mathrm{d}\theta\mathrm{d}\Lambda \overset{n\to\infty}{\longrightarrow} 0$ $a.s.$

Let $C_{\mathcal{S}} = \mathrm{Proj}_\Theta(C) \times \mathcal{S}$, where $\mathrm{Proj}_\Theta(C)$ denotes the projection of C upon Θ, and let us note that $C \subset C_{\mathcal{S}}$. We are going to show the stronger result

$$\lim_{n\to\infty} \int_{C_{\mathcal{S}}} |\hat{p}(\theta,\Lambda|Z_n) - p(\theta,\Lambda|Z_n)|\mathrm{d}\theta\mathrm{d}\Lambda = 0 \qquad a.s.$$

From (19)

$$p(\theta,\Lambda|Z_n) = K_n|\Lambda|^n \exp\Big\{-\sum_{i=1}^n \|y_i - f(x_i,\theta)\|^2_\Lambda\Big\} \quad (40)$$

and from (20)

$$\begin{aligned}\hat{p}(\theta,\Lambda|Z_n) = \hat{K}_n|\Lambda|^n \exp\Big\{&-\frac{1}{2}\|\theta - \hat{\theta}_n\|^2_{\beta_n} \\ &-\sum_{i=1}^n \|y_i - f(x_i,\hat{\theta}_n)\|^2_\Lambda\Big\} \quad (41)\end{aligned}$$

where $\hat{K}_n$ is a normalizing constant.

Let us denote $\mathrm{E}_{\hat{p},C_{\mathcal{S}}}[.] = \int_{C_{\mathcal{S}}} [.]\hat{p}(\theta,\Lambda|Z_n)\mathrm{d}\theta\mathrm{d}\Lambda$.

Let us show that the Kullback-Leibler distance between the distributions p and $\hat{p}$ over $C_{\mathcal{S}}$ tends almost surely to 0 as n grows to infinity. This will result in their convergence in L_1-norm.

$$\begin{aligned}&\mathcal{K}_{C_{\mathcal{S}}}(\hat{p},p) \\ &= \int_{C_{\mathcal{S}}} \hat{p}(\theta,\Lambda|Z_n)\log\frac{\hat{p}(\theta,\Lambda|Z_n)}{p(\theta,\Lambda|Z_n)}\mathrm{d}\theta\mathrm{d}\Lambda \\ &= \log\frac{\hat{K}_n}{K_n} + \mathrm{E}_{\hat{p},C_{\mathcal{S}}}\Big[-\frac{1}{2}\|\theta - \hat{\theta}_n\|^2_{\beta_n} \\ &\quad -\sum_{i=1}^n \|y_i - f(x_i,\hat{\theta}_n)\|^2_\Lambda + \sum_{i=1}^n \|y_i - f(x_i,\theta)\|^2_\Lambda\Big] \quad (42)\end{aligned}$$

let us consider the successive terms of $\mathcal{K}_{C_{\mathcal{S}}}(\hat{p},p)$:

- Let $\mathcal{E}^n_{\mathcal{C}_{\mathcal{S}}} = \mathrm{E}_{\hat{p},C_{\mathcal{S}}}[-\frac{1}{2}\|\theta - \hat{\theta}_n\|^2_{\beta_n}]$, which is finite. Since $\|\theta - \hat{\theta}_n\|^2_{\beta_n} \sim \chi^2_q$ under $\hat{p}$, it comes immediatly
$$0 > \mathcal{E}^n_{\mathcal{C}_{\mathcal{S}}} > \mathrm{E}_{\hat{p}}[-\frac{1}{2}\|\theta - \hat{\theta}_n\|^2_{\beta_n}] = -\frac{q}{2}$$
- For all $i = 1, \ldots, n$, in a neighbourhood of $\hat{\theta}_n$ it holds

$$
\begin{aligned}
&\|y_i - f(x_i,\theta)\|^2_\Lambda \\
&\quad = \|y_i - f(x_i,\hat\theta_n) + \dot f_{x_i,\hat\theta_n}(\theta-\hat\theta_n)\|^2_\Lambda + o(\|\theta-\hat\theta_n\|^2) \\
&\quad = \|y_i - f(x_i,\hat\theta_n)\|^2_\Lambda + \|\dot f_{x_i,\hat\theta_n}(\theta-\hat\theta_n)\|^2_\Lambda \\
&\qquad +2 < y_i - f(x_i,\hat\theta_n), \dot f_{x_i,\hat\theta_n}(\theta-\hat\theta_n) >_\Lambda \\
&\qquad +o(\|\theta-\hat\theta_n\|^2) \qquad (43)
\end{aligned}
$$

and

$$
\begin{aligned}
&-\sum_{i=1}^n \|y_i - f(x_i,\hat\theta_n)\|^2_\Lambda + \sum_{i=1}^n \|y_i - f(x_i,\theta)\|^2_\Lambda \\
&\quad = \sum_{i=1}^n \|\dot f_{x_i,\hat\theta_n}(\theta-\hat\theta_n)\|^2_\Lambda \\
&\qquad +2\sum_{i=1}^n < y_i - f(x_i,\hat\theta_n), \dot f_{x_i,\hat\theta_n}(\theta-\hat\theta_n) >_\Lambda \\
&\qquad +n\ o(\|\theta-\hat\theta_n\|_\Lambda) \qquad (44)
\end{aligned}
$$

According to the definition of the Wishart probability distribution, $\mathrm{E}_{\hat p}[\Lambda] = \dfrac{2n+d+1}{2n}\hat\Lambda_n$. Then

$$
\begin{aligned}
&\mathrm{E}_{\hat p, C_{\mathcal S}}[\sum_{i=1}^n \|\dot f_{x_i,\hat\theta_n}(\theta-\hat\theta_n)\|^2_\Lambda] \\
&\quad = \mathrm{E}_{\hat p, C_{\mathcal S}}[\mathrm{E}_{\hat p}[\sum_{i=1}^n \|\dot f_{x_i,\hat\theta_n}(\theta-\hat\theta_n)\|^2_\Lambda |\theta]] \\
&\quad = \mathrm{E}_{\hat p, C_{\mathcal S}}[\frac{2n+d+1}{2n}\sum_{i=1}^n \|\dot f_{x_i,\hat\theta_n}(\theta-\hat\theta_n)\|^2_{\hat\Lambda_n}| \\
&\quad = \frac{2n+d+1}{2n}\mathrm{E}_{\hat p, C_{\mathcal S}}[\frac{1}{2}\|\theta-\hat\theta_n\|^2_{\beta_n}] \\
&\quad = -\frac{2n+d+1}{2n}\mathcal{E}^n_{C_{\mathcal S}} \qquad (45)
\end{aligned}
$$

In the same way

$$
\begin{aligned}
&\mathrm{E}_{\hat p, C_{\mathcal S}}[\sum_{i=1}^n < y_i - f(x_i,\hat\theta_n), \dot f_{x_i,\hat\theta_n}(\theta-\hat\theta_n) >_\Lambda] \\
&\quad = \mathrm{E}_{\hat p, C_{\mathcal S}}\Big[\\
&\qquad \mathrm{E}_{\hat p}[\sum_{i=1}^n < y_i - f(x_i,\hat\theta_n), \dot f_{x_i,\hat\theta_n}(\theta-\hat\theta_n) >_\Lambda |\theta]\Big] \\
&\quad = \frac{2n+d+1}{2n}\mathrm{E}_{\hat p, C_{\mathcal S}}\Big[\\
&\qquad \sum_{i=1}^n < y_i - f(x_i,\hat\theta_n), \dot f_{x_i,\hat\theta_n}(\theta-\hat\theta_n) >_{\hat\Lambda_n} \Big] \\
&\quad = \frac{2n+d+1}{2n}\mathrm{E}_{\hat p, C_{\mathcal S}}[0] \\
&\quad = 0 \qquad (46)
\end{aligned}
$$

since $(\hat\theta_n, \hat\Lambda_n)$ are the least square estimators of (θ,Λ).

The Kullback-Leibler distance between p and $\hat p$ over $C_{\mathcal S}$ then becomes:

$$
\begin{aligned}
\mathcal{K}_{C_{\mathcal S}}(\hat p, p) &= \log\tfrac{\hat K_n}{K_n} + \mathcal{E}^n_{C_{\mathcal S}} - \tfrac{2n+d+1}{2n}\mathcal{E}^n_{C_{\mathcal S}} \\
&\quad +\mathrm{E}_{\hat p, C_{\mathcal S}}[n\ o(\|\theta-\hat\theta_n\|^2)] \qquad (47)
\end{aligned}
$$

Let us show now that $\lim_{n\to\infty} \mathrm{E}_{\hat p, C_{\mathcal S}}[n\ o(\|\theta-\hat\theta_n\|^2)] = 0$:
It was shown in Section B.1 that $\beta_n \sim 2n\tilde\beta$ as n grows to infinity, with $\tilde\beta$ a positive definite matrix.
By the equivalence of norms on $\mathbb{R}^d$ there exist α_1 and α_2 positive, such that

$$\alpha_1\|\theta-\hat\theta_n\|^2_{\tilde\beta} \le \|\theta-\hat\theta_n\|^2 \le \alpha_2\|\theta-\hat\theta_n\|^2_{\tilde\beta}$$

$$\alpha_1 n\|\theta-\hat\theta_n\|^2_{\tilde\beta} \le n\|\theta-\hat\theta_n\|^2 \le n\alpha_2\|\theta-\hat\theta_n\|^2_{\tilde\beta}$$

and then for n large,

$$\frac{1}{2}\alpha_1\|\theta-\hat\theta_n\|^2_{\beta_n} \le n\|\theta-\hat\theta_n\|^2 \le \frac{1}{2}\alpha_2\|\theta-\hat\theta_n\|^2_{\beta_n}$$

As $\|\theta-\hat\theta_n\|^2_{\beta_n} \sim \chi^2_q$ under $\hat p$, it comes

$$\frac{1}{2}\alpha_1 q \le \mathrm{E}_{\hat p}[n\|\theta-\hat\theta_n\|^2] \le \frac{1}{2}\alpha_2 q$$

There exist then $\tilde\alpha_1$ and $\tilde\alpha_2$ positive such that

$$\tilde\alpha_1 \le \mathrm{E}_{\hat p, C_{\mathcal S}}[n\|\theta-\hat\theta_n\|^2] \le \tilde\alpha_2$$

Let us come back to the study of the $o(\|\theta-\hat\theta_n\|^2)$, in (43):
For all couple (x_i, y_i) let us note $g^i_n(\theta) = o(\|\theta-\hat\theta_n\|^2)$. Then

$$\sum_{i=1}^n g^i_n(\theta) = n\ o(\|\theta-\hat\theta_n\|^2)$$

Let $\tilde g^i_n(\theta) = \dfrac{g^i_n(\theta)}{\|\theta-\hat\theta_n\|^2}$ and $\tilde g_n(\theta) = \dfrac{1}{n}\sum_{i=1}^n \tilde g^i_n(\theta)$.
Then $\lim_{\theta\to\hat\theta_n} \tilde g_n(\theta) = 0$.
As $\lim_{n\to\infty}\hat\theta_n = \theta_*$ a.s., for all $i \in \mathbb{N}^*$ we have $\lim_{n\to\infty}\tilde g^i_n(\theta_*) = 0$. Moreover as the $\{x_i\}$ belong to the compact subset $\mathcal{X}$ and the model function f is C^1 with respect to x and θ, the last convergence is uniform with respect to x:

$$\forall\ \varepsilon>0,\ \exists\ N_\varepsilon :\ \forall\ x\in\mathcal{X},\ \forall\ n > N_\varepsilon,\ |\tilde g^i_n(\theta_*)| < \varepsilon$$

and

$$\forall\ \varepsilon>0,\ \exists\ N_\varepsilon :\ \forall\ n > N_\varepsilon,\ |\tilde g_n(\theta_*)| < \varepsilon$$

then

$$\lim_{n\to\infty}\tilde g_n(\theta_*) = 0$$

Now let us introduce $\tilde g_n$ in the expectation of interest :

$$\mathrm{E}_{\hat{p},C_{\mathcal{S}}}[n\ o(\|\theta-\hat{\theta}_n\|^2)] = \mathrm{E}_{\hat{p},C_{\mathcal{S}}}[\tilde{g}_n(\theta) n\|\theta-\hat{\theta}_n\|^2]$$

For all $\varepsilon > 0$, let V_ε be the ball of radius ε centred at θ_*, Λ_*. Let us choose ε sufficiently small such that $V_\varepsilon \subset C_{\mathcal{S}}$. Then

$$\begin{aligned}\mathrm{E}_{\hat{p},C_{\mathcal{S}}}[n\ o(\|\theta-\hat{\theta}_n\|^2)] = \\ \int_{C_{\mathcal{S}}\setminus V_\varepsilon} \tilde{g}_n(\theta) n\|\theta-\hat{\theta}_n\|^2 \hat{p}(\theta,\Lambda|Z_n)\mathrm{d}\theta\mathrm{d}\Lambda \\ + \int_{V_\varepsilon} \tilde{g}_n(\theta) n\|\theta-\hat{\theta}_n\|^2 \hat{p}(\theta,\Lambda|Z_n)\mathrm{d}\theta\mathrm{d}\Lambda \quad (48)\end{aligned}$$

As $\hat{p}$ is consistent, for all $\varepsilon > 0$ such that $V_\varepsilon \subset C_{\mathcal{S}}$ there exists N_ε such that for all $n > N_\varepsilon$

$$\int_{C_{\mathcal{S}}\setminus V_\varepsilon} n\|\theta-\hat{\theta}_n\|^2 \hat{p}(\theta,\Lambda|Z_n)\mathrm{d}\theta\mathrm{d}\Lambda < \varepsilon$$

Then for all $n > N_\varepsilon$

$$\left|\mathrm{E}_{\hat{p},C_{\mathcal{S}}}[n\ o(\|\theta-\hat{\theta}_n\|^2)]\right| \leq \sup_{C_{\mathcal{S}}} \tilde{g}_n(\theta)\varepsilon + \sup_{V_\varepsilon} \tilde{g}_n(\theta)\tilde{\alpha}_2$$

As $\sup_{C_{\mathcal{S}}} \tilde{g}_n(\theta)$ is bounded, the first term tends to zero with ε. But as ε tends to zero, V_ε tends to $\{\theta_*\}$. As n grows to infinity, then $\sup_{V_\varepsilon} \tilde{g}_n(\theta)$ tends to $\lim_{n\to\infty} \tilde{g}_n(\theta_*) = 0$.
We conclude that

$$\lim_{n\to\infty} \mathrm{E}_{\hat{p},C_{\mathcal{S}}}[n\ o(\|\theta-\hat{\theta}_n\|^2)] = 0 \quad (49)$$

- Finally let us show that $\lim_{n\to\infty} \log \dfrac{\hat{K}_n}{K_n} = 0$:
To alleviate notations let us denote, from (40) and (41):

$$p \ = \ K_n \times h_n \qquad \text{and} \qquad \hat{p} \ = \ \hat{K}_n \times \hat{h}_n.$$

Let us follow a *reductio ad absurdum* by assuming that $\lim_{n\to\infty} \hat{K}_n/K_n \neq 1$.
Because of (44), (45) and (46), for $n \to \infty$

$$\mathrm{E}_{\hat{p},C_{\mathcal{S}}}\Big[\log \frac{\hat{h}_n}{h_n}\Big] \to 0 \ \text{ and } \ \mathrm{E}_{\hat{p},C_{\mathcal{S}}}\Big[\log \frac{h_n}{\hat{h}_n}\Big] \to 0 \ \text{ a.s.} \quad (50)$$

i) Suppose first that $\lim_{n\to\infty} \dfrac{\hat{K}_n}{K_n} < 1$:
then

$$\begin{aligned}\lim_{n\to\infty} \mathrm{E}_{\hat{p},C_{\mathcal{S}}}\Big[\frac{h_n}{\hat{h}_n}\Big] &= \lim_{n\to\infty} \hat{K}_n \int_{C_{\mathcal{S}}} h_n \mathrm{d}\theta\mathrm{d}\Lambda \\ &< \lim_{n\to\infty} K_n \int_{C_{\mathcal{S}}} h_n \mathrm{d}\theta\mathrm{d}\Lambda \\ &\leq \lim_{n\to\infty} \int K_n h_n \mathrm{d}\theta\mathrm{d}\Lambda \\ &= 1 \quad (51)\end{aligned}$$

Due to the convexity of the exponential function and the consistency of $\hat{p}(\theta,\Lambda|Z_n)$ there exists N such that for $n > N$, Jensen inequality can be applied to $\mathrm{E}_{\hat{p},C_{\mathcal{S}}}[\frac{h_n}{\hat{h}_n}]$, and gives

$$\exp \mathrm{E}_{\hat{p},C_{\mathcal{S}}}\Big[\log \frac{h_n}{\hat{h}_n}\Big] \leq \mathrm{E}_{\hat{p},C_{\mathcal{S}}}\Big[\frac{h_n}{\hat{h}_n}\Big]$$

but By (50)

$$\exp \mathrm{E}_{\hat{p},C_{\mathcal{S}}}\Big[\log \frac{h_n}{\hat{h}_n}\Big] \ \overset{n\to\infty}{\longrightarrow} \ 1$$

then $\lim_{n\to\infty} \mathrm{E}_{\hat{p},C_{\mathcal{S}}}\Big[\dfrac{h_n}{\hat{h}_n}\Big] \geq 1$, which contradicts (51).

ii) Suppose now that $\lim_{n\to\infty} \dfrac{\hat{K}_n}{K_n} > 1$:
By a similar reasoning and since $p(\theta,\Lambda|Z_n)$ is consistent

$$\lim_{n\to\infty} \mathrm{E}_{\hat{p},C_{\mathcal{S}}}\Big[\frac{h_n}{\hat{h}_n}\Big] \ > \ \lim_{n\to\infty} \int_{C_{\mathcal{S}}} K_n h_n \mathrm{d}\theta\mathrm{d}\Lambda \ = \ 1$$

which implies

$$\lim_{n\to\infty} \log \mathrm{E}_{\hat{p},C_{\mathcal{S}}}\Big[\frac{\hat{h}_n}{h_n}\Big] \ < \ 0. \quad (52)$$

Due to the convexity of the -log function and the possibility to apply the Jensen inequality to $\mathrm{E}_{\hat{p},C_{\mathcal{S}}}\Big[\dfrac{\hat{h}_n}{h_n}\Big]$, for sufficiently great n, we have by (50)

$$-\log \mathrm{E}_{\hat{p},C_{\mathcal{S}}}\Big[\frac{\hat{h}_n}{h_n}\Big] \leq \mathrm{E}_{\hat{p},C_{\mathcal{S}}}\Big[\log \frac{h_n}{\hat{h}_n}\Big] \ \overset{n\to\infty}{\longrightarrow} \ 0$$

which contradicts (52).

From i) and ii) we can deduce that $\lim_{n\to\infty} \log(\hat{K}_n/K_n) = 0$, and then, from (47) and (49) that

$$\mathcal{K}_{C_{\mathcal{S}}}(\hat{p},p) \ \overset{n\to\infty}{\longrightarrow} \ 0 \quad (53)$$

This completes the proof of Theorem 1, since Kullback convergence dominates L_1 convergence over $C_{\mathcal{S}}$ and then over C.

C. Proof of Theorem 2

$$\begin{aligned}D &= \int \Big|\ \hat{p}(y|x,Z_n) - p(y|x,Z_n)\ \Big|\ \mathrm{d}y \\ &= \int \Big|\ \int \hat{p}(y|\theta,\Lambda,x)\hat{p}(\theta,\Lambda|Z_n)\mathrm{d}\theta\mathrm{d}\Lambda \\ &\quad - \int p(y|\theta,\Lambda,x)p(\theta,\Lambda|Z_n)\mathrm{d}\theta\mathrm{d}\Lambda\ \Big|\ \mathrm{d}y \\ &= \int \Big|\ \int p(y|\theta,\Lambda,x)\Big[\hat{p}(\theta,\Lambda|Z_n) - p(\theta,\Lambda|Z_n)\Big]\mathrm{d}\theta\mathrm{d}\Lambda \\ &\quad + \int \hat{p}(\theta,\Lambda|Z_n)\Big[\hat{p}(y|\theta,\Lambda,x) - p(y|\theta,\Lambda,x)\Big]\mathrm{d}\theta\mathrm{d}\Lambda\ \Big|\ \mathrm{d}y \\ &\leq \int p(y|\theta,\Lambda,x)\ \Big|\ \hat{p}(\theta,\Lambda|Z_n) - p(\theta,\Lambda|Z_n)\ \Big|\ \mathrm{d}\theta\mathrm{d}\Lambda\mathrm{d}y \\ &\quad + \int \hat{p}(\theta,\Lambda|Z_n)\ \Big|\ \hat{p}(y|\theta,\Lambda,x) - p(y|\theta,\Lambda,x)\ \Big|\ \mathrm{d}\theta\mathrm{d}\Lambda\mathrm{d}y\end{aligned}$$

By Fubini's theorem

$$\begin{aligned} D \leq & \int \left| \hat{p}(\theta,\Lambda|Z_n) - p(\theta,\Lambda|Z_n) \right| \mathrm{d}\theta\mathrm{d}\Lambda \\ & + \int \hat{p}(\theta,\Lambda|Z_n) \int \left| \hat{p}(y|\theta,\Lambda,x) - p(y|\theta,\Lambda,x) \right| \mathrm{d}y\mathrm{d}\theta\mathrm{d}\Lambda \\ & = T_1 + T_2 \end{aligned}$$

As $\hat{p}(\theta,\Lambda|Z_n)$ is assumed to be a L_1-convergent approximation of $p(\theta,\Lambda|Z_n)$, T_1 tends to zero as n grows to infinity. Let us show that the same is true for T_2.

Let $h(\theta,\hat{\theta}_n) = \int \left| \hat{p}(y|\theta,\Lambda,x) - p(y|\theta,\Lambda,x) \right| \mathrm{d}y$. Obviously $0 \leq h(\cdot,\cdot) \leq 2$. The mapping h is continuous and $h(\hat{\theta}_n,\hat{\theta}_n) = 0$ for all $n \in \mathbb{N}^*$. As $\lim_{n\to\infty}(\hat{\theta}_n,\hat{\Lambda}_n) = (\theta_*,\Lambda_*)$ $a.s.$, we deduce that $\lim_{n\to\infty} h(\theta_*,\hat{\theta}_n) = 0$. Moreover, for all $\varepsilon > 0$ there exists a neighbourhood of (θ_*,Λ_*), V_ε, and an integer N_1 such that for almost all $(\theta,\Lambda) \in V_\varepsilon$ and all $n > N_1$ we have $h(\theta,\hat{\theta}_n) < \varepsilon/2$.

Let us now split T_2 according to V_ε :

$$\begin{aligned} T_2 &= \int \hat{p}(\theta,\Lambda|Z_n) h(\theta,\hat{\theta}_n) \mathrm{d}\theta\mathrm{d}\Lambda \\ T_2 &= \int_{V_\varepsilon} + \int_{V_\varepsilon^c} \\ T_2 &\leq \int_{V_\varepsilon^c} \hat{p}(\theta,\Lambda|Z_n) h(\theta,\hat{\theta}_n) \mathrm{d}\theta\mathrm{d}\Lambda \ + \ \varepsilon/2 \\ T_2 &\leq 2 \int_{V_\varepsilon^c} \hat{p}(\theta,\Lambda|Z_n) \mathrm{d}\theta\mathrm{d}\Lambda \ + \ \varepsilon/2 \end{aligned}$$

Due to the consistency of $\hat{p}(\theta,\Lambda|Z_n)$ as n grows to infinity, there exists an integer N_2 such that for all $n > N_2$ we have $\int_{V_\varepsilon^c} \hat{p}(\theta,\Lambda|Z_n)\mathrm{d}\theta\mathrm{d}\Lambda < \varepsilon/4$ and then $T_2 < \varepsilon$.

It follows that D tends to zero as n grows to infinity.

ACKNOWLEDGMENT

The authors would like to thank the Associate Editor and the three anonymous referees for detailed comments which helped to improve the presentation of the paper.

REFERENCES

[1] C. Abraham and B. Cadre, "Asymptotic properties of posterior distributions derived from misspecified models," *C.R. Acad. Sci. Paris*, Ser. I, 335, pp. 495-498, 2002.

[2] U. Anders and O. Korn, "Model selection in neural networks," *Neural Networks*, vol. 12, pp. 309-323, 1999.

[3] J.O. Berger, *Statistical Decision Theory and Bayesian Analysis*. New york: Springer, 1985.

[4] R.H. Berk, "Limiting behavior of posterior distributions when the model is incorrect." *Ann. Math. Statist.*, vol. 37, pp. 51-58, 1966.

[5] R.H. Berk, "Consistency a posteriori," *Ann. Math. Statist.*, vol. 41, pp. 894-906, 1970.

[6] J.M. Bernardo., A.F.M. Smith, *Bayesian Theory*. New York: Springer, 2000.

[7] C.M. Bishop, *Neural Networks for Pattern Recognition*. Oxford: Clarendon Press,1995.

[8] C. Borggard, H. H. Thodberg, "Optimal minimal neural interpretation of spectra," *Anal. Chemistry*, vol. 6, pp. 545-551, 1992.

[9] W.L. Buntine, A.S. Weigend, "Bayesian backpropagation", *Complex Syst.*, vol. 5, pp. 603-643, 1991.

[10] K.P. Burnham, D.R. Anderson, *Model Selection and Inference*. New York: Springer, 1998.

[11] A.M. Chen, H. Lu, R. Hecht-Nielsen, "On the geometry of feedforward neural-network error surfaces", *Neural Comput.*, vol. 5, pp. 910-927, 1993.

[12] J. F. G. De Freitas, M. A. Niranjan, A. H. Gee, A. Doucet, "Sequential Monte Carlo methods to train neural network models", *Neural Computation*, vol 12, 4, pp. 955-993, 2000.

[13] R.V. Foutz, "On the unique consistent solution to the likelihood equations", *J. Amer. Statist. Ass.*, vol. 72, pp. 147-148, 1977.

[14] S. Geisser, W.F. Eddy, "A predictive approach to model selection", *J. Amer. Statist. Ass.*, vol. 74, pp. 153-160, 1979.

[15] A.E. Gelfand, "Models determination using sampling-based methods", in *Markov Chain Monte Carlo in practice*, Eds: W.R. Gilks, S. Richarson, D.J. Spiegelhalter, London: Chapman and Hall, 1996.

[16] S. Geman, E. Bienenstock, R. Doursat, "Neural network and bias/variance dilemma," *Neural Computation*, vol. 4, pp. 1-58, 1992.

[17] R. Gencay and M. Qi, "Pricing and hedging derivative securities with neural networks: Bayesian regularization, early stopping and bagging", *IEEE Trans. Neural Networks*, vol. 12, pp. 726-734, 2001.

[18] D. Husmeier, W.D. Penny, S.J. Roberts, "An Empirical evaluation of Bayesian sampling with hybrid Monte Carlo for training neural network classifiers", *Neural Networks*, vol. 12, pp. 677-705, 1999.

[19] T. Y. Kwok, D.Y. Yeung, "Constructive algorithms for structure learning in feedforward neural networks for regression problems," *IEEE Trans. Neural Networks*, vol 8, pp. 448-472, 1997.

[20] J. Lampinen, A. Vehtari, "Bayesian approach for neural networks–review and case studies", *Neural Networks*, vol. 14, pp. 257-274, 2001.

[21] H.K.H. Lee, "A noninformative prior for neural networks", *Machine Learning*, vol. 50, pp. 197-212, 2003.

[22] D.J.C. MacKay, "A practical Bayesian framework for backpropagation networks," *Neural Comput.*, vol 4, pp. 448-472, 1992.

[23] D.J.C. MacKay, "Hyperparameters: optimize or integrate out?", in *Maximum Entropy and Bayesian Methods*, The Netherlands: Kluwer, 1995.

[24] J.S. Maritz, Lwin, T., *Empirical Bayes Methods*, 2nd ed. London: Chapman and Hall, 1989.

[25] M.C. Medeiros, A. Veiga and C.E. Pedreira, "Modeling exchange rates: smooth transitions, neural networks and linear models," *IEEE Trans. Neural Networks*, vol. 12, pp. 755-764, 2001.

[26] N. Murata, S. Yoshizawa and S. Amari, "A criterion for determining the number of parameters in an artificial neural network model," in *Artificial Neural Networks*. Proceeding of ICANN-91, eds T. Kohonen, K. Mäkisara, O. Simula and J.Kangas, vol. 1, pp. 9-14. Amsterdam: North Holland, 1991.

[27] R. M. Neal, *Bayesian Learning for Neural Network*, Lecture Notes in Statistics 118, New York: Springer, 1996.

[28] P. Nelles, *Nonlinear System Identification*, New York: Springer, 2001.

[29] B.D. Ripley, *Pattern Recognition and Neural Networks*, Cambridge MA: Cambridge Univ. Press, 1996.

[30] I. R. Rivals and L. Personnaz, "Neural-network construction and seclection in nonlinear modeling", *IEEE Trans. Neural Networks*, vol. 14, pp. 804-819, 2003.

[31] V. Rossi, J.P. Vila, "Bayesian selection of multiresponse nonlinear regression model," Rapport de Recherche No 04-01, Groupe de Biostatistique et d'Analyse des Systèmes, ENSA.M-INRA-UMII, 36p., 2004.

[32] G.A.F. Seber, *Multivariate Observations*. New York: Wiley, 1984.

[33] G.A.F. Seber, C.F. Wild, *Nonlinear Regression*. New York: Wiley, 1989.

[34] J. Shao, "Linear model selection by cross-validation," *J. Amer. Statist. Ass.*, vol. 88, pp. 486-494, 1993.

[35] M. Stone, "Cross-validatory choice and assessment of statistical predictions," *J. Royal Statist. Soc., B*, vol. 36, pp. 11-147 (with discussion), 1974.

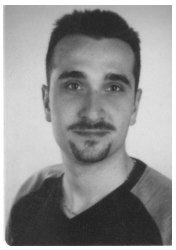
[36] H.H. Thodberg, "A review of Bayesian neural networks with an application to near infrared spectroscopy", *IEEE Trans. Neural Networks*, vol. 7, pp. 56-72, 1996.

[37] J.P. Vila, V. Wagner, P. Neveu, P. "Bayesian nonlinear model selection and neural networks: a conjugate prior approach," *IEEE Trans. on Neural Networks*, vol. 11, pp. 265-278, 2000.

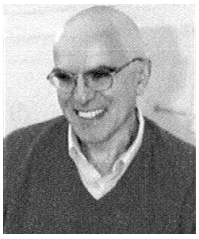
[38] H. White, "Maximum likelihood estimation of misspecified models," *Econometrica*, vol. 50, pp. 1-25, 1982.

[39] D.H. Wolpert, "On the use of evidence neural networks," in *Advances in Neural Information Processing System 5*. San Mateo, CA: Morgan Kaufmann, pp. 539-546, 1993.

[40] B.L. Zhang, R. Coggins, M.A. Jabri, D. Dersch and B. Flower, "Multiresolution forecasting for future trading using wavelet decomposition," *IEEE Trans. Neural Networks*, vol. 12, pp. 765-775, 2001.



Vivien Rossi was born in Montpellier, France, in 1976. He received the Ph.D. degree in Statistics from the Ecole Nationale Supérieure Agronomique de Montpellier in 2004.
He is currently with the Probability and Statistics Team, Institute of Mathematics and Modelisation of Montpellier, University of Montpellier II.
His main research interests include Bayesian statistics, nonlinear filtering especially particle methods, nonparametric estimation and neural networks.



Jean-Pierre Vila (M'98) received the Docteur-ingénieur degree in mathematical statistics from the University of Paris XI-Orsay in 1985.
He is currently Research Director at the Department of Applied Mathematics and Computer Science of INRA, the French National Agronomical Research Institute. His main research interests include statistics and control theory of nonlinear dynamical systems, with special regards to nonparametric estimation, neural networks, filtering theory, and their applications in life sciences.