Statistics, Vol. 42, No. 4, August 2008, 291–311

Taylor & Francis
Taylor & Francis Group

Bayesian selection of multiresponse nonlinear regression model

Vivien Rossi[a] and Jean-Pierre Vila[b]*

[a] *UR Dynamique des forêts naturelles, CIRAD, Campus International de Baillarguet, 34398 Montpellier cedex 5, France;* [b] *UMR Analyse des Systèmes et de Biométrie, INRA-ENSAM, 2 Place P. Viala, 34060 Montpellier cedex 1, France*

(*Received 14 August 2006; final version received 22 August 2007*)

A Bayesian method for the selection and ranking of multiresponse nonlinear regression models in a given model set is proposed. It uses an expected utility criterion based on the logarithmic score of a posterior predictive density of each model. Two approaches are proposed to get this posterior. The first is based on a general asymptotically convergent approximation of the model parameter posterior corresponding to a wide class of parameter priors. The second, a numerical one, uses well-known pure or hybrid MCMC methods. Both posterior approximation approaches allow the practical computation of the expected utility criterion on a sample-reuse basis. This leads to a class of Bayesian cross-validation (CV) procedures, aiming at finding the model having the best predictive ability among a set of models. Varied comparisons of the performances of the Bayesian procedures, with that of AIC, BIC and standard CV procedures, are proposed on a simulated and a real model selection problem case studies, of low and high parametric dimensions respectively.

Keywords: multiresponse nonlinear regression; Bayesian model selection; expected utility criterion; MCMC methods

1. Introduction

When it can be reduced to parameter hypothesis testing, nonlinear model selection can be performed through extensions of well-known asymptotic inferential tests as likelihood ratio, Wald's or Rao's Lagrange multiplier tests [1]. When selection has to be done among nonnested models other tools have to be used. Information-theoretic criteria as Akaike's and its various derivatives are then frequently used and offer easy-to-use procedures for selection of parsimonious models [2]. These criteria combine some measure of fit with a penalty term to account for model complexity. It is well known that depending on the kind of penalty used, each of these criteria performs well only in one type of situation. To counter this limitation, [3] proposed recently an adaptive selection procedure which combines the benefit of several of such criteria as Akaike's information criterion (AIC), Bayesian information criterion (BIC), Mallow's C_p and risk inflation criterion (RIC). Many adaptive Bayesian variants of the previous penalized

*Corresponding author. Email: vila@supagro.inra.fr

ISSN 0233-1888 print/ISSN 1029-4910 online

DOI: 10.1080/02331880701739824
http://www.informaworld.com

measure-of-fit criteria have also been considered, especially for variable selection in regression see [4–7] and the references therein. These last approaches, and recent ones based on bootstrapping and hypothesis testing [8], take advantage of the theoretical advances in resampling and Markov chain Monte Carlo methods. Most of these approaches have been developed for single-response models but their extension to the multiresponse case is generally straightforward. When the selection problem does not reduce to variable selection in multiple linear regression or comparison of nested models, the most favoured computer intensive procedure is cross validation (CV) [9], based on model prediction capability. Its reduced probability assumption requirements (as exchangeability assumption) makes classic CV particulary attractive. But CV is known to be inconsistent [10] as are all the methods asymptotically equivalent to it (*e.g.* AIC, C_p, the jackknife and the bootstrap). Moreover, CV is often too conservative (selection of unnecessary large models). Recourse to Bayesian variants of CV can help to reduce this loss of efficiency [9, 11, 12]. Furthermore in addition to point-prediction-based comparisons according to quadratic loss, the Bayesian approach can offer alternatives as comparisons resting upon predictive distributions, according to logarithmic or quadratic scores (see *e.g.* [13]). The greater richness of information attached to predictive distributions makes this last type of approach particularly attractive for model selection. Conjugate-prior-based Bayesian approaches of this type, have for example, been proposed for the comparison of feedforward neural network architectures [14, 15].

This paper is devoted to the study of generalizations of such Bayesian approaches for the selection and ranking of multiresponse nonlinear regression models. It uses a sample-reuse calculation of an expected utility built from the logarithmic score of a posterior predictive density, for each competing model. We consider two quite different approaches to get this posterior. The first is based on a general asymptotically convergent approximation of the model parameter posterior. This posterior parameter approximation itself is valid for a wide class of parameter priors. By so doing the critical issue of the choice of an appropriate prior for the model parameters and that of the calculation of the resulting posterior, are conveniently defused. The second approach to reach the posterior of interest, is a numerical one and uses well-known pure or hybrid MCMC methods from noninformative priors. More precisely, a mixed procedure combining the respective advantages of the Gibbs sampling and of the Metropolis–Hastings algorithm is considered, in addition to the standard Metropolis–Hastings algorithm.

The paper is organized as follows. In Section 2 the statistical framework of the multiresponse nonlinear regression within which the model selection problem is considered, is set up. In Section 3 the building elements of the expected-utility-based criterion are considered. The issue of the posterior predictive density to be used in the computation of the utility criterion is then examined in the next two sections. More precisely, in Section 4 a general convergent analytic approximation of a model parameter posterior which is valid for any usual prior is developed. A convergent analytic approximation of the related posterior predictive density is then built up under the form of a multivariate Student distribution. In Section 5, two numerical counterparts to this analytic approximation are considered through the MCMC procedures mentioned above. In Section 6 the three variants of our Bayesian selection procedure are applied to a simulated problem of multi-response regression model selection and then to an actual selection problem of a multi-output feedforward predictive neural network in a Soil Science study. These two case studies are representative of low and high parametric dimension situations respectively. It is shown how the performances of the Bayesian procedure compare advantageously with that of AIC, BIC and classic CV procedures on the same problems.

Appendix A provides the proofs of all lemmas, propositions and convergence theorems of expected utility criteria, parameter posterior and predictive density approximations.

Finally, Appendix B presents the results of the comparisons of all the model selection criteria considered, on the two case studies.

2. Multiresponse modeling framework

We are interested in multiresponse nonlinear regression models of the form

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

where for $i \in \{1, \dots, n\}$

$$y_i \in I\!R^d \quad x_i \in I\!R^l \quad \theta \in \Theta \subset I\!R^s \quad \varepsilon_i \sim N_d(0, \Sigma) \quad \Sigma \in \mathcal{S} \subset I\!R^{d \times d}$$

where $\mathcal{S}$ is the set of all positive definite symmetric matrices of dimension $d \times d$. In this paper, we shall have to consider more often than the variance–covariance matrix Σ, the precision matrix $\Lambda = \Sigma^{-1}$.

Let us denote

- $Z_n = (x_i, y_i), i = 1, \dots, n$, the data set, made of n *i.i.d.* random data points (x, y).
- $\dot{f}_{x_i,\theta} = \big(\partial f(x_i, \theta)/\partial\theta_j\big) j \in \{1, \dots, q\}$ when these derivatives exist.

Other notations:

- $\mathcal{N}_s(\cdot|\mu, \Sigma)$: the s-dimensional Gaussian distribution with expectation μ and covariance matrix Σ.
- $\mathcal{W}i_d(\cdot|\alpha, \beta)$: the d-dimensional Wishart distribution with parameters α and β.
- $St_d(\cdot|\mu, \Psi, \alpha)$: the d-dimensional Student distribution with parameters μ, Ψ and α.

[1] show, how given Z_n, under model M and under appropriate regularity conditions, consistent maximum likelihood estimates (or equivalent least squares estimates) $\hat{\theta}_n$ and $\widehat{\Lambda}_n$ of θ and Λ respectively can be calculated, such that

$$\hat{\theta}_n = \operatorname{argmin}_\theta \det\left[\sum_{i=1}^{n} \big(y_i - f(x_i, \theta)\big)\big(y_i - f(x_i, \theta)\big)'\right]$$

$$\widehat{\Lambda}_n^{-1} = \hat{\Sigma}_n = \frac{1}{n}\sum_{i=1}^{n} \big(y_i - f(x_i, \hat{\theta}_n)\big)\big(y_i - f(x_i, \hat{\theta}_n)\big)' \tag{2}$$

More generally when model M is incorrect, there still exist values θ_o and Λ_o to which the maximum likelihood estimates $\hat{\theta}_n$ and $\hat{\Lambda}_n$ converge almost surely with n [16, 17]: θ_o and Λ_o 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.

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

3. The expected-utility-based criterion

To do this selection we follow the maximum-expected-utility approach [13] for which the optimal model choice is M^* such that

$$U(M^*|Z_n) = \sup_{M^j \in \mathcal{M}} U(M^j|Z_n) \tag{3}$$

where

$$U(M^j|Z_n) = \int u(M^j, y, x|Z_n)p\big((x, y)|Z_n\big)\mathrm{d}y\,\mathrm{d}x \tag{4}$$

with $u(M^j, y, x|Z_n)$ a given utility function and $p((x, y)|Z_n)$ a probability density representing actual beliefs about (x, y) having observed Z_n.

As we are interested in comparing models from a predictive distribution point of view we take as utility function the logarithmic score

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

where $\log p_j(y|x, Z_n)$ is the posterior predictive density of a response y of model M^j at x, given the past observations Z_n, a prior density $p(\theta, \Lambda)$ for the model M^j parameters and its related posterior $p(\theta, \Lambda|Z_n)$.

With this choice for the utility function

$$U_n^j = U(M^j|Z_n) = \int u(M^j, y, x|Z_n)p\big((x, y)|Z_n\big)\mathrm{d}y\,\mathrm{d}x = E_{x,y}[\log p_j(y|x, Z_n)] \tag{6}$$

Using this criterion implies that it can be computed for every n and every possible data set Z_n. But $p((x, y)|Z_n)$ in Equation (6) is not available. We can only search for an estimate of U_n^j for each $M^j \in \mathcal{M}$. To do this, we consider a well-known approximation through CV.

3.1. *Expected utility approximation through CV*

CV predictive density methods for model comparison have been considered by several authors (see reviews of [18, 19]).

Let us consider the n partitions of Z_n: $Z_n = [Z_{n-1}[i], (x_i, y_i)]$ for $1 \le i \le n$, where $Z_{n-1}[i]$ denotes the data set Z_n after withdrawal of the data point (x_i, y_i). Let us denote $u_{i,n}^j = \log p_j(y_i|x_i, Z_{n-1}[i])$. The $\{u_{i,n}^j, i = 1, \ldots, n\}$ constitute a collection of leave-one-out CV predictive densities. Let

$$\hat{U}_n^j = \frac{1}{n}\sum_{i=1}^{n} u_{i,n}^j \tag{7}$$

[13] proposed to use $\hat{U}_n^j$ as a good approximation of U_n^j the expected utility of model M^j. More recently, [20, 21] confirmed this recommendation and estimated its probability distribution.

By looking for the model M^j which maximizes (7), the 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 favours the likelihood of each observation with respect to the others.

The criterion (7) is based on the posterior predictive density $p_j(y|x, Z_n)$, where

$$p_j(y|x, Z_n) = \int p_j(y|x, \theta, \Lambda)p_j(\theta, \Lambda|Z_n)\mathrm{d}\theta\,\mathrm{d}\Lambda \tag{8}$$

in which $p_j(\theta, \Lambda|Z_n)$ is the parameter posterior induced by a given parameter prior $p_j(\theta, \Lambda)$. However, the choice of a relevant prior and the exact calculation of its posterior can be difficult, if not untractable. On these bases, the two next sections are devoted to convergent analytic and numerical approximations respectively, of $p_j(y|x, Z_n)$.

4. L_1-convergent posterior predictive density approximations

The first step of our approach is to show how a general L_1-convergent approximation of the parameter posterior density which is valid for a wide class of parameter priors, can be built up for each of the J competing models $\{M^j, j = 1, \ldots, J\}$ (the index j will be dropped to alleviate notations in all the following).

4.1. *A L_1-convergent approximation of the parameter posterior density*

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

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

H_2 There exist consistent estimators of θ_o and Λ_o, model M^j parameter values such as defined in Section 2. When M^j is the true model, θ_o and Λ_o are its parameter values.

H_3 The model function $f_j(x, \theta)$ is of class C^1 both in x and θ.

H_4 There exists a compact subset $C \subset \Theta \times \mathcal{S}$ including an open neighbourhood of (θ_o, Λ_o), such that $p(\theta, \Lambda)$ is bounded and strictly positive on C.

Let $p(\theta, \Lambda)$ be a given parameter prior. Now, by definition of a parameter posterior density for model (1)

$$p(\theta, \Lambda | Z_n) = K_n |\Lambda|^{n/2} \exp\left\{ -\frac{1}{2} \sum_{i=1}^{n} \|y_i - f(x_i, \theta)\|^2_\Lambda \right\} p(\theta, \Lambda) \tag{9}$$

where K_n is a normalizing constant.

We consider an easily tractable approximate:

$$\text{let } \hat{p}(\theta, \Lambda | Z_n) = \mathcal{N}_s\left(\theta \middle| \hat{\theta}_n, \left(\sum_{i=1}^{n} \dot{f}'_{x_i, \hat{\theta}_n} \hat{\Lambda}_n \dot{f}_{x_i, \hat{\theta}_n} \right)^{-1} \right)$$
$$\times \mathcal{W}i_d\left(\Lambda \middle| \frac{n+d+1}{2}, \frac{1}{2} \sum_{i=1}^{n} (y_i - f(x_i, \hat{\theta}_n))(y_i - f(x_i, \hat{\theta}_n))' \right) \tag{10}$$

Let $\hat{P}(.)$ be the related probability measure over $\mathcal{F}$ the sigma-algebra associated with $\Theta \times \mathcal{S}$. The following lemma ensures the consistency of $\hat{p}(\theta, \Lambda | Z_n)$ to θ_o, Λ_o.

LEMMA 4.1 *Suppose assumptions $\mathcal{H}$ are satisfied. Let A be any measurable subset of $\Theta \times \mathcal{S}$ including an open neighbourhood of (θ_o, Λ_o). Then $\lim_{n \to \infty} \hat{P}(A) = 1$ a.s.*

To ensure that $\hat{p}(\theta, \Lambda | Z_n)$ is a consistent estimate of $p(\theta, \Lambda | Z_n)$, let us consider the following prior-free approximation of $p(\theta, \Lambda | Z_n)$

$$\tilde{p}(\theta, \Lambda | Z_n) = \tilde{K}_n |\Lambda|^{n/2} \exp\left\{ -\frac{1}{2} \sum_{i=1}^{n} \|y_i - f(x_i, \theta)\|^2_\Lambda \right\} \times I_{C_\mathcal{S}}(\theta, \Lambda) \tag{11}$$

where $\tilde{K}_n$ is a normalizing constant and $I_{C_\mathcal{S}}$ is the indicator function of $C_\mathcal{S}$, with $C_\mathcal{S} = \text{Proj}_\Theta(C) \times \mathcal{S}$, where $\text{Proj}_\Theta(C)$ denotes the projection of C upon Θ.

With respect to Equation (9) that amounts to take an improper prior with density equal to one on $C_\mathcal{S}$ and zero elsewhere.

This prior-free approximation (11) of $p(\theta, \Lambda | Z_n)$ allows to characterize the asymptotic behaviour of the L_1 distance over C, between $\hat{p}(\theta, \Lambda | Z_n)$ and $p(\theta, \Lambda | Z_n)$ as the number of data n grows to infinity. This is done through the following three propositions.

PROPOSITION 4.2 *Under assumptions $\mathcal{H}$*

$$\lim_{n \to \infty} \int_C |\tilde{p}(\theta, \Lambda | Z_n) - p(\theta, \Lambda | Z_n)| \mathrm{d}\theta \, \mathrm{d}\Lambda = 0 \ \ a.s. \tag{12}$$

PROPOSITION 4.3 *Under assumptions $\mathcal{H}$*

$$\lim_{n\to\infty}\int_C |\hat{p}(\theta,\Lambda|Z_n)-\tilde{p}(\theta,\Lambda|Z_n)|\mathrm{d}\theta\,\mathrm{d}\Lambda = 0 \ a.s. \tag{13}$$

PROPOSITION 4.4 *Under assumptions $\mathcal{H}$*

$$\lim_{n\to\infty}\int_C |\hat{p}(\theta,\Lambda|Z_n)-p(\theta,\Lambda|Z_n)|\mathrm{d}\theta\,\mathrm{d}\Lambda = 0 \ a.s. \tag{14}$$

Proposition 4.4 is an immediate consequence of Proposition 4.2 and Proposition 4.3, by noting that

$$|\hat{p}-p| \le |\hat{p}-\tilde{p}| + |\tilde{p}-p|$$

Proposition 4.4 makes it possible to get the L_1 consistency of $\hat{p}(\theta,\Lambda|Z_n)$ as n grows to infinity:

THEOREM 4.5 *Under assumptions $\mathcal{H}$*

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

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

$$\begin{aligned}\int |\hat{p}(\theta,\Lambda|Z_n)-p(\theta,\Lambda|Z_n)|\mathrm{d}\theta\,\mathrm{d}\Lambda &= \int_C |\hat{p}-p| + \int_{C^c}|\hat{p}-p| \\ &\le \int_C |\hat{p}-p| + \int_{C^c}\hat{p} + \int_{C^c} p \end{aligned} \tag{15}$$

By Proposition 4.4 the first integral in Equation (15) tends to zero as n tends to ∞. By Lemma 4.1 the second integral in Equation (15) tends to zero as n tends to ∞ since (θ_o,Λ_o) does not belong to C^c. The same is true for the third integral in Equation (15) since p is consistent [17, 22, 23] and (θ_o,Λ_o) does not belong to C^c. ■

4.2. *Application to L_1-approximations of the posterior predictive density*

Thanks to the consistent estimate of $p(\theta,\Lambda|Z_n)$ provided in the previous section, it is now possible to get a consistent estimate of the density of interest $p(y|x,Z_n)$.

By definition

$$p(y|x,Z_n) = \int p(y|x,\theta,\Lambda)p(\theta,\Lambda|Z_n)\mathrm{d}\theta\,\mathrm{d}\Lambda \tag{16}$$

THEOREM 4.6 *Let $\hat{p}(\theta,\Lambda|Z_n)$ be a general L_1-convergent approximation of the parameter posterior density $p(\theta,\Lambda|Z_n)$ and let*

$$\hat{p}(y|x,Z_n) = \int \hat{p}(y|x,\theta,\Lambda)\hat{p}(\theta,\Lambda|Z_n)\mathrm{d}\theta\,\mathrm{d}\Lambda \tag{17}$$

with

$$\hat{p}(y|x,\theta,\Lambda) = \frac{|\Lambda|^{1/2}}{(2\pi)^{d/2}}\exp\left\{-\frac{1}{2}\|y-f(x,\hat{\theta}_n)\|^2_\Lambda\right\}. \tag{18}$$

Then

$$\lim_{n\to\infty}\int \big|\hat{p}(y|x,Z_n)-p(y|x,Z_n)\big|\mathrm{d}y = 0 \quad a.s. \tag{19}$$

With the help of both Theorem 4.5 and Theorem 4.6 it is possible to get a more practical L_1-convergent approximation of the posterior predictive density under the form of a multivariate-Student distribution.

COROLLARY 4.7 *Under assumptions $\mathcal{H}$, let*

$$\hat{p}_n(y|x, Z_n) = St_d\left(y|f(x, \hat{\theta}_n), \frac{n+2}{n}\hat{\Lambda}_n, n+2\right) \tag{20}$$

then

$$\lim_{n\to\infty}\int \left|\hat{p}_n(y|x, Z_n) - p(y|x, Z_n)\right|\mathrm{d}y = 0 \quad a.s. \tag{21}$$

Proof Bringing $\hat{p}(\theta, \Lambda|Z_n)$ as given by Equation (10) into (17), easily leads to $\hat{p}_n(y|x, Z_n)$ as given by Equation (20). ■

5. Numerical posterior predictive density approximations by an MCMC approach

Under the assumption $\varepsilon_i \sim N_d(0, \Sigma)$ as in Equation (1), it is possible to use an MCMC method to approximate the posterior predictive density. Remember that

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

Let $\{(\theta^1, \Lambda^1), \ldots, (\theta^m, \Lambda^m)\}$ be a sample from $p(\theta, \Lambda|Z_{n-1}[i])$. By the strong law of large numbers we have

$$\lim_{m\to\infty}\frac{1}{m}\sum_{k=1}^{m} p\big(y_i|x_i, \theta^k, \Lambda^k\big) = p(y_i|x_i, Z_{n-1}[i]) \quad a.s.$$

The density $p(\theta, \Lambda|Z_n)$ as in Equation (9), is only known up to a normalizing constant, and cannot directly provide a sample $\{\theta^k, \Lambda^k\}$. However, one can get an approximately $p(\theta, \Lambda|Z_{n-1}[i])$-distributed sample thanks to an MCMC method as the Metropolis–Hastings algorithm [24] or, as will be shown, a hybrid Gibbs-Hastings algorithm [25, 26] under certain conditions for the parameter prior $p(\theta, \Lambda)$. Nevertheless, such MCMC integrations can suffer from instability [19] and heavy computing. The interpretation of the utility criterion approximated by such MCMC methods can thus be delicate. However, in order to compare these now well-known approaches with the analytic ones developed previously, we implemented them, when allowed by the application considered (see Section 6).

5.1. *Metropolis–Hastings approximate sampling from $p(\boldsymbol{\theta}, \boldsymbol{\Lambda}|\boldsymbol{Z_n})$*

First, let us introduce briefly the principle of the Metropolis–Hastings algorithm in our context

$$(i)\ (\tilde{\theta}, \tilde{\Lambda}) \sim q(\theta, \Lambda|\theta^t, \Lambda^t)$$

$$(ii)\ (\theta^{t+1}, \Lambda^{t+1}) = \begin{cases} (\theta^t, \Lambda^t) & \text{with probability } 1-\rho \\ (\tilde{\theta}, \tilde{\Lambda}) & \text{with probability } \rho \end{cases}$$

with

$$\rho = 1 \wedge \frac{p\big(\tilde{\theta}, \tilde{\Lambda}|Z_n\big)}{p\big(\theta^t, \Lambda^t|Z_n\big)} \frac{q\big(\theta^t, \Lambda^t|\tilde{\theta}, \tilde{\Lambda}\big)}{q\big(\tilde{\theta}, \tilde{\Lambda}|\theta^t, \Lambda^t\big)} \tag{22}$$

The density $q(\theta, \Lambda|\theta^t, \Lambda^t)$ is generally called the proposal distribution. The Markov chain $\{(\theta^t, \Lambda^t), t \in \mathbb{N}\}$ thus generated, admits $p(\theta, \Lambda|Z_n)$ as stationary distribution if the support of q contains the support of $p(\theta, \Lambda|Z_n)$. See [25] or [26] for more details and for the theoretical study of this algorithm.

In our context, the density $p(\theta, \Lambda|Z_n)$ is only known up to a normalizing constant

$$\begin{aligned} p(\theta, \Lambda|Z_n) &= \frac{p(Z_n|\theta, \Lambda)p(\theta, \Lambda)}{\int p(Z_n|\theta, \Lambda)p(\theta, \Lambda)\mathrm{d}\theta\, \mathrm{d}\Lambda} \\ &\propto p(Z_n|\theta, \Lambda)p(\theta, \Lambda) \end{aligned} \tag{23}$$

which does not impede the use of the Metropolis–Hastings algorithm, since only the ratio $p(\tilde{\theta}, \tilde{\Lambda}|Z_n)/p(\theta^t, \Lambda^t|Z_n)$ is considered, in Equation (22).

For computational efficiency, the distribution q should be chosen so that it can be easily evaluated and sampled. As the convergence is faster when q is close to $p(\theta, \Lambda|Z_n)$ (see [26]), a good candidate for the proposal distribution in our case is the following

$$q(\theta^{t+1}, \Lambda^{t+1}|\theta^t, \Lambda^t) = \mathcal{N}_s\left(\theta^{t+1}|\theta^t, V^t\right) \mathcal{W}i_d\left(\Lambda^{t+1}\left|\frac{n+d+1}{2}, \frac{n}{2}(\Lambda^t)^{-1}\right.\right) \tag{24}$$

with $V^t = (\sum_{i=1}^n \dot{f}'_{x_i,\theta^t}\Lambda^t \dot{f}_{x_i,\theta^t})^{-1}$. Such a proposal distribution is close to $\hat{p}(\theta, \Lambda|Z_n)$ in Equation (10), which is itself a convergent approximation of $p(\theta, \Lambda|Z_n)$.

The choice of the prior distribution, $p(\theta, \Lambda)$ is delicate. For instance Jeffrey's priors, which are invariant to parameter transformation, are not adapted to our context. First, Jeffrey's principle is controversial for multiparameter models [27]. Second, Jeffrey's noninformative prior density, $p(\theta, \Lambda) \propto \left|J(\theta, \Lambda)\right|^{1/2}$, where $J(\theta, \Lambda)$ is the Fisher information matrix, is most often untractable because of the general nonlinearity with respect to θ. It reduces to $p(\theta, \Lambda) \propto \left|\Lambda\right|^{(d+1)/2}$ when θ is discarded but it has nevertheless to be proscribed since by (23) $p(\theta, \Lambda)$ must be such that $\int p(Z_n|\theta, \Lambda)p(\theta, \Lambda)\mathrm{d}\theta\, \mathrm{d}\Lambda < \infty$.

A prior which preferentially weights a theoretical neighbourhood of the true parameter values is of particular interest. A good candidate is

$$\begin{aligned} p(\theta, \Lambda) = \mathcal{N}_s &\left(\theta|\hat{\theta}_n, \left(\sum_{i=1}^n \dot{f}'_{x_i,\hat{\theta}_n}\hat{\Lambda}_n \dot{f}_{x_i,\hat{\theta}_n}\right)^{-1}\right) \\ &\times \mathcal{W}i_d\left(\Lambda\left|\frac{n+d+1}{2}, \frac{1}{2}\sum_{i=1}^n (y_i - f(x_i, \hat{\theta}_n))(y_i - f(x_i, \hat{\theta}_n))'\right.\right) \end{aligned} \tag{25}$$

which is equal to the density $\hat{p}(\theta, \Lambda|Z_n)$ of 4.1, which is itself consistent by Lemma 4.1. By Proposition 4.4, this density is also a convergent approximation of the posterior density of the parameters.

Let us note that the Λ part of this density is Wishart. Due to the assumed Gaussian likelihood of the observations $y_1, \ldots, y_n$, this density is then a conjugate prior for Λ [27]. As a consequence, conditionally on θ, the posterior density of Λ is still a Wishart density. This allows the use of a more powerful MCMC algorithm [25, 26]: a hybrid version of the Gibbs and Hastings algorithms. More details are given in the following section.

5.2. *Hybrid Gibbs–Hastings approximate sampling from $p(\theta, \Lambda|Z_n)$*

Let the parameter prior be of the more general form

$$p(\theta, \Lambda) = h(\theta)\mathcal{W}i_d(\Lambda|\alpha, \beta) \tag{26}$$

Then

$$\begin{aligned} p(\theta, \Lambda|Z_n) &\propto p(Z_n|\theta, \Lambda)p(\theta, \Lambda) \\ &\propto \prod_{i=1}^{n} p(y_i, x_i|\theta, \Lambda)h(\theta)\mathcal{W}i_d(\Lambda|\alpha, \beta) \\ &\propto \prod_{i=1}^{n} p(y_i|x_i, \theta, \Lambda)p(x_i|\theta, \Lambda)h(\theta)\mathcal{W}i_d(\Lambda|\alpha, \beta) \\ &\propto \prod_{i=1}^{n} p(y_i|x_i, \theta, \Lambda)h(\theta)\mathcal{W}i_d(\Lambda|\alpha, \beta) \\ &\propto |\Lambda|^{n/2+\alpha-(d+1)/2} \exp\left\{-\frac{1}{2}\sum_{i=1}^{n} \|y_i - f(x_i, \theta)\|_{\Lambda}^2 - tr[\beta\Lambda]\right\} h(\theta) \end{aligned} \tag{27}$$

where $h(\theta)$ is supposed to be such that $p(\theta, \Lambda|Z_n)$ exists.

(27) shows that conditionally on θ, the marginal Λ posterior is then still a Wishart density:

$$p(\Lambda|\theta, Z_n) = \mathcal{W}i_d\left(\Lambda\middle|\frac{n}{2} + \alpha, \beta + \frac{1}{2}\sum_{i=1}^{n}[y_i - f(x_i, \theta)][y_i - f(x_i, \theta)]'\right) \tag{28}$$

By contrast, the conditional marginal θ posterior $p(\theta|\Lambda, Z_n)$ is not reachable due to the model non-linearity with respect to θ and is only known up to a normalizing constant. Gibbs' algorithm cannot then be used. However, the following hybrid version from [26] or [25] can then be considered

$$\begin{aligned} &(i)\ \Lambda^{t+1} \sim p(\Lambda|\theta^t, Z_n) \\ &(ii)\ \tilde{\theta} \sim q(\theta|\theta^t, \Lambda^{t+1}) \\ &\qquad \theta^{t+1} = \begin{cases} \theta^t & \text{with probability } 1 - \rho \\ \tilde{\theta} & \text{with probability } \rho \end{cases} \end{aligned}$$

where q is a proposal distribution for θ, and

$$\rho = 1 \wedge \left\{ \frac{p\left(\tilde{\theta}|\Lambda^{t+1}, Z_n\right)}{q\left(\tilde{\theta}|\theta^t, \Lambda^{t+1}\right)} \Big/ \frac{p\left(\theta^t|\Lambda^{t+1}, Z_n\right)}{q\left(\theta^t|\theta^t, \Lambda^{t+1}\right)} \right\}$$

in which the ratio $p(\tilde{\theta}|\Lambda^{t+1}, Z_n)/p(\theta^t|\Lambda^{t+1}, Z_n)$ can still be evaluated.

This hybrid algorithm has been shown to combine the advantages of both algorithms of Metropolis–Hastings and Gibbs sampling [25, 26].

Now from a practical point of view we can choose again for $p(\theta, \Lambda)$ the prior given by Equation (25) which satisfies Equation (26), and for the proposal distribution we can choose the restriction to θ of the distribution q as given by Equation (24):

$$q\left(\theta^{t+1}|\theta^t, \Lambda^{t+1}\right) = \mathcal{N}_s\left(\theta^{t+1}|\theta^t, V^t\right) \tag{29}$$

with $V^t = \left(\sum_{i=1}^{n} \dot{f}'_{x_i,\theta^t}\Lambda^{t+1}\dot{f}_{x_i,\theta^t}\right)^{-1}$.

5.3. *Stopping time of the MCMC generations*

Deciding when to stop the generation of the $\{\theta^k, \Lambda^k\}$ Markov chain is an important practical issue. For the following applications, the approach used to monitor the convergence to stationarity is based on comparing parallel simulated sequences. Among the many ways to compare parallel sequences, we employed the approach proposed by [28]. In the applications to follow, 200 trajectories corresponding to different starting points were generated, and stationarity of the chain was assumed when the potential scale reduction $\sqrt{\hat{R}}$ estimate was below 1.1, a usual critical value. The quantity $\hat{R}$ is computed from the between and within-sequence variances, see [28] or [26] for more details.

6. Comparison of criteria performances on case studies

The analytic posterior predictive density approximation (20) and the two MCMC approximations (Sections 5.1 and 5.2) are used to estimate the expected utility criterion through CV as in Equation (7), on a simulated and a real life model selection problem successively.

These three computing versions of the U criterion are also compared to the standard CV, AIC and BIC criteria on the same case studies:

Two versions of the CV criterion are considered: the classic one $\mathrm{CV}_I = \sum_{i=1}^{n} \|y_i - f(x_i, \hat{\theta}_{n-1}[i])\|^2_{I_{d\times d}}$ and a normalized one $\mathrm{CV}_Q = \sum_{i=1}^{n} \|y_i - f(x_i, \hat{\theta}_{n-1}[i])\|^2_{Q^{-1}}$ to take account of the correlations between the components of the response vector, where $\hat{\theta}_{n-1}[i]$ are the maximum likelihood estimates 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 [2]: 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 parameters.

We consider first a simulated model selection problem with competing models of low parametric dimension. Secondly, we consider an actual model selection problem from a true dataset. In this second case study, in soil sciences, the competing models are neural networks of large parametric dimension.

6.1. *A simulated model selection problem*

Let us consider the following five two-response regression models:

$$f_1(x,\theta) = \begin{cases} \theta_1\sqrt{x} + \theta_2 \\ \log x + \theta_3 \end{cases} \qquad f_2(x,\theta) = \begin{cases} \dfrac{x}{1+\theta_1 x} + \theta_2 \\ \log x + \theta_3 \end{cases} \qquad f_3(x,\theta) = \begin{cases} \dfrac{\theta_1 x}{\theta_2 + x} + \theta_3 \\ \sqrt{x} + \theta_4 \end{cases}$$

$$f_4(x,\theta) = \begin{cases} \theta_1 \log x + \theta_2 \\ \dfrac{x}{1+\theta_3 x} + \theta_4 \end{cases} \qquad f_5(x,\theta) = \begin{cases} \theta_1 \log x + \theta_2 \\ \dfrac{\theta_3 x}{\theta_4 + x} + \theta_5 \end{cases}$$

Data were simulated from model f_3 with additive centred Gaussian noise of covariance matrix $\Lambda^{-1} = \left[\begin{smallmatrix} 0.5 & 0.5 \\ 0.5 & 1 \end{smallmatrix}\right]$ and the parameter values $\theta_1 = 2., \theta_2 = 1.5, \theta_3 = 1., \theta_4 = 1$. Three sets of comparison were carried out with respectively 50, 100 and 200 data points for which the x values were sampled from $\mathcal{U}[1,\ 10]$ the uniform distribution on $[1, 10]$. Let us note that on this subset of $\mathbb{R}$, the five model functions considered have a comparable behaviour, for a large set of parameter ranges.

The U criterion (7) was first applied according to the three Bayesian CV approaches presented in this paper based on the analytic, the Metropolis–Hastings and the hybrid Gibbs–Hastings posterior density approximations respectively. Tables B1, B2 and B3 (Appendix B.1) display the scores reached by the five models (winning scores, which maximize the criterion, are in bold). For the other criteria, AIC, BIC and CV, Tables B4, B5, B6 and B7 (Appendix B.1) display the scores reached by each model (winning scores, which minimize the criterion, are in bold).

Save for the dataset size $n = 50$ the analytic and the hybrid Gibbs–Hastings approximations of the U criterion, always sharply select the true model. The behaviours of AIC and CV are not satisfactory because they end in failure for the largest dataset and offer low contrasted values. That of BIC and that of the Metropolis–Hastings approximation of the U criterion, are completely wrong with no right selection.

The good results of the hybrid Gibbs–Hastings approximation on this example illustrate its superiority over the Metropolis–Hastings one. These two approximations suffer from a rather high variability, the more so as the number of unknown parameters is large. Moreover, a full week of computation with Matlab® on a 3Ghz Pentium4 was necessary to obtain each of both MCMC approximations of the criterion for each of the five models considered. With these MCMC estimations, the computing time increases dramatically with the number of unknown parameters and the number of observations.

6.2. *An actual model selection problem*

In this second case study, the high parametric dimension of each of the competing models considered (feedforward neural networks) prevented the computation of both posterior predictive density MCMC approximations (Section 5) to estimate the U criterion through CV. Consequently, the computation of the approximation (7) of the U criterion was based on the analytic convergent predictive density approximation (20), for each competing model in each comparison test.

A total of 370 soil layers of Languedoc Plain (south of France) were sampled to study their water retention properties [29]. A soil clod of about 30 cm^3 was taken each time. The water contents of the clods at six metric water potentials (3, 10, 30, 100, 300 and 1500 kPa) were measured. Concurrently, nine basic soil variables were also measured for each clod (the bulk density, the proportions of several classes of silt, sand and clay particles and the organic carbon content). A predictive multi-response neural network model with the six water contents as outputs and the nine soil variables as inputs was investigated.

Five quite different feedforward fully connected neural structures, with nine inputs and six outputs, were compared (which were just a few among other possible feedforward neural network structures):

NN1: one hidden layer of 7 neurons (118 parameters).
NN2: two hidden layers of 5 and 4 neurons respectively (104 parameters).
NN3: two hidden layers of 7 and 2 neurons respectively (104 parameters).
NN4: one hidden layer of only 1 neuron (the simplest possible structure, 22 parameters).
NN5: one hidden layer of 10 neurons (166 parameters).

Two types of model comparison were carried out, from the 370 data points:

- In the first one, the U, CV$_Q$, AIC and BIC criteria were computed for the five neural models using all the 370 data points, each according to its own rule.
- In the second one, the six models were first fitted on a common subset of 320 data points (learning basis) randomly chosen among the 370 of the initial data set. After that, the mean-squared error of prediction (MSEP) of each of the five models was computed on the remaining 50 data points (test basis).

This last procedure was repeated with different randomly sampled 320-point learning bases (and complementary test bases). This MSEP comparison can be considered as a reference for all other types of criterion.

Considering the high number of parameters for each neural model and then the risk of local minima trapping of the least squares parameter estimations, all the necessary fittings were systematically repeated from 50 different starting values randomly chosen in the space of the parameters.

Table B8 (Appendix B.2) shows the score reached by each of the five neural models for each of the four criteria on the full 370-point data set.

Table B9 (Appendix B.2) displays the MSEP values respectively reached by the five models on a typical 50-point test basis.

Both the U and the CV_Q criteria selected the same neural model, NN1, which coincides with the model of smallest MSEP, whereas both the AIC and BIC criteria wrongly selected the same model, NN5, which presents one of the highest MSEP, but unsurprisingly the lowest number of parameters.

Let us note moreover, that NN1 is not the model with the highest number of parameters, which shows the ability of the U criterion to trade-off between complexity and over fitting.

All the other model comparisons done with other randomly sampled learning bases and complementary test bases, led to the same relative behaviours of the respective criteria.

7. Conclusion

The advantage of predictive distributions in comparing several models in the light of current data has been advocated for long (see [30] for example). Predictive distributions allow diagnostic functions of the local adequacy of the entertained model, to be built from a subsample of the data population and lead to utility functions defined on the whole population [13]. In this framework, this paper shows how a convergent approximation of a predictive distribution of a multiresponse regression model can result from an analytic approach or a numerical MCMC-based one. An approximation of the model expected utility itself is then available which can be computed on a CV-like basis and can be used for ranking competing models. The analytic approximation approach has two main advantages over the numerical approximation ones: on the one hand it is free of any parameter prior density choice and on the other hand it presents very little variability in comparison with that of the numerical MCMC-based approximations. MCMC-based approximations are dependent on the choice of a parameter prior and a proposal distribution, while needing furthermore dramatically heavy computing times even for models of reasonable parametric dimension. It thus seems preferable to use the analytic approach which is in addition much more easier to implement. Finally, comparisons with CV, AIC and BIC criteria on simulated and real data sets of different sizes, on competing models of varied complexity and parameter dimension, show that the performance of the expected utility criterion analytic approximation is much less sensitive to these factors.

References

[1] G.A.F. Seber and C.F. Wild, *Nonlinear Regression*, Wiley, NewYork, 1989.

[2] K.P. Burnham and D.R. Anderson, *Model Selection and Inference*, Springer, NewYork, 1998.

[3] X. Shen and J. Ye, *Adaptive model selection*, J. Amer. Statist. Assoc. 97 (2002), pp. 210–221.

[4] E.I. George and R. McCullogh, *Variable selection via Gibbs sampling,* J. Amer. Statist. Assoc. 88 (1993), pp. 881–889.

[5] E.I. George and R. McCullogh, *Stochastic search variable selection*, in *Practical Markov Chain Monte Carlo in Practice,* W.R. Gilks, S. Richardson, and D.J. Spiegelhater, eds., Chapman and Hall, London, 1995, pp. 203–214.

[6] E.L. George and R. McCullogh, *Approaches for Bayesian variable selection*, Statist. Sinica 7 (1997), pp. 339–373.

[7] E.I. George and D.P. Foster, *The risk inflation criterion for multiple regression*, Ann. Statist. 22 (1994), pp. 1947–1975.

[8] J. Rao, *Bootstrap choice of cost complexity for better subset selection*, Statist. Sinica 9 (1999), pp. 273–288.

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

[10] J. Shao, *Linear model selection by cross-validation*, J. Amer. Statist. Assoc. 88 (1993), pp. 486–494.

[11] S. Geisser and W.F. Eddy, *A predictive approach to model selection*, J. Amer. Statist. Assoc. 74 (1979), pp. 153–160.

[12] A.E. Gelfand, D.K. Dey, and H. Chang, *Model determination using predictive distributions with implementation via sampling-based methods*, in *Bayesian Statistics 4*, J.M. Bernardo, J.O. Berger, A.P. Dawid, and A.F.M. Smith, eds., Oxford University Press, NewYork, 1992, pp. 147–167 (with discussion).

[13] J.M. Bernardo and A.F.M. Smith, *Bayesian Theory,* Springer, NewYork, 1994.

[14] J.P. Vila, V. Wagner, and P. Neveu, *Bayesian nonlinear model selection and neural networks:A conjugate prior approach,* IEEE Trans. Neural Netw. 11 (2000), pp. 265–278.

[15] V. Rossi and J.P. Vila, *Bayesian multioutput feedforward neural networks comparison: A conjugate prior approach,* IEEE Trans. Neural Netw. 17 (2006), pp. 35–47.

[16] H. White, *Maximum likelihood estimation of misspecified models*, Econometrica 50 (1982), pp. 1–25.

[17] C. Abraham and B. Cadre, *Asymptotic properties of posterior distributions derived from misspecified models*, Comptes Rendus Académie des Sciences Paris, Series I 335 (2002), pp. 495–498.

[18] A.E. Gelfand and D.K. Dey, *Bayesian Model Choice: asymptotics and exact calculations*, J. R. Statist. Soc. B 56 (1994), pp. 501–514.

[19] A.E. Gelfand, *Model determination using sampling-based methods,* in *Markov Chain Monte Carlo in Practice*, W.R. Gilks, S. Richardson, and D.J. Spiegelhalter, eds., Chapman and Hall, London, 1996, pp. 145–162.

[20] A. Vehtari and J. Lampinen, *Bayesian model assessment and comparison using cross-validation predictive densities*, Neur. Comput. 14 (2002), pp. 2439–2468.

[21] A. Vehtari and J. Lampinen, *Expected utility estimation via cross-validation*, in *Bayesian Statistics 7*, J.M. Bernardo *et al.*, eds., Oxford University Press, 2003, pp. 701–710.

[22] R.H. Berk, Limiting behavior of posterior distribution when the model is incorrect, Ann. Math. Stat. 37 (1966), pp. 51–58.

[23] R.H. Berk, *Consistency a posteriori*, Ann. Math. Stat. 41 (1970), pp. 894–906.

[24] W.R. Gilks, S. Richardson, and D.J. Spiegelhalter, *Markov Chain Monte Carlo in Practice*, Chapman and Hall, London, 1996.

[25] L. Tierney, *Markov chains for exploring posterior distributions*, Ann. Stat. 22 (1994), pp. 1701–1762 (with discussion).

[26] Robert, C. and Casella, G., *Monte Carlo Statistical Methods,* Springer-Verlag, NewYork, 1999.

[27] A. Gelman, J.B. Carlin, H.S. Stern, and D.B. Rubin, *Bayesian Data analysis*, Chapman and Hall, London, 1995.

[28] A. Gelman, *Inference and monitoring convergence,* in W.R. Gilks, S. Richardson, and D.J. Spiegelhalter, eds., Chapman and Hall, London, 1996, pp. 131–143.

[29] D. Leenhardt, M. Voltz, M. Bornand, and R. Webster, *Evaluating soil maps for prediction of soil water properties*, Eur. J. Soil Sci. 45(3) (1994), pp. 293–301.

[30] G.E.P. Box, *Sampling and Bayes inference in scientific modeling,* J. R. Statist. Soc. B 26 (1980), pp. 211–252 (with discussion).

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

Appendix A: Proofs

A.1 *Proof of Lemma 4.1*

Let us study first the asymptotic behaviour of (θ, Λ) according to $\hat{p}(\theta, \Lambda|Z_n)$.

(i) Expectations: from Equation (10) and assumption H_2 we have

$$\mathrm{E}_{\hat{p}}[\theta, \Lambda|Z_n] = \left(\hat{\theta}_n, \frac{n+d+1}{n}\hat{\Lambda}_n\right) \overset{n\to\infty}{\longrightarrow} (\theta_o, \Lambda_o) \ a.s. \tag{A1}$$

(ii) Variances:

- Let β_n be the inverse of the θ variance. From Equation (10) $\beta_n = V(\theta)^{-1} = \sum_{i=1}^n \dot{f}'_{x_i,\hat{\theta}_n} \hat{\Lambda}_n \dot{f}_{x_i,\hat{\theta}_n}$.

  Let us show that some terms of β_n grow to infinity with n:

  Let $\dot{\beta}_n = 1/n \sum_{i=1}^n \dot{f}'_{x_i,\theta_o} \Lambda_o \dot{f}_{x_i,\theta_o}$, $\tilde{\beta} = \mathrm{E}_x\left[\dot{f}'_{x,\theta_o} \Lambda_o \dot{f}_{x,\theta_o}\right]$ and $\tilde{\beta}_n = 1/n \sum_{i=1}^n \dot{f}'_{x_i,\hat{\theta}_n} \hat{\Lambda}_n \dot{f}_{x_i,\hat{\theta}_n}$.

  As the $\{x_i\}$ are i.i.d. we have by the strong law of large numbers: $\lim_{n\to\infty} \dot{\beta}_n = \tilde{\beta}$.

  As $(\hat{\theta}_n, \hat{\Lambda}_n)$ converge a.s. to (θ_o, Λ_o) on the probability space $(\Omega, \mathcal{U}, P)$, there exists a P-negligible subset D of $\mathcal{U}$ such that for all $\omega \in \Omega \backslash D$ one has $\lim_{n\to\infty}(\hat{\theta}_n(\omega), \hat{\Lambda}_n(\omega)) = (\theta_o, \Lambda_o)$. Let us consider the event $\omega \in \Omega \backslash D$.

As the $\{x_i\}$ belong to the compact set $\mathcal{X}$ and f is C^1, for all $\epsilon > 0$ there exists N_ϵ such that for all $n > N_\epsilon$ and for all $x \in \mathcal{X}$ one has $\|\dot{f}'_{x,\hat{\theta}_n(\omega)}\hat{\Lambda}_n(\omega)\dot{f}_{x,\hat{\theta}_n(\omega)} - \dot{f}'_{x,\theta_o}\Lambda_o\dot{f}_{x,\theta_o}\| < \epsilon$. Then for all $\epsilon > 0$ there exists N_ϵ such that for all $n > N_\epsilon$, $\|\dot{\beta}_n - \tilde{\beta}_n(\omega)\| < \epsilon$. This implies that

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

Then

$$\lim_{n\to\infty} \beta_n - n\tilde{\beta} = 0 \ \ a.s.$$

and the variance of θ according to $\hat{p}_n$ tends to zero as n tends to infinity.

- Let λ_{ij} be the ij^{th} term of Λ, which follows the Wishart distribution introduced in Equation (10).

According to [31], $\lambda_{ii} \sim ((1/n)\hat{\Lambda}_n)_{ii}\chi^2_{(n+d+1)/2}$. Then

$$V(\lambda_{ii}) = \frac{(\hat{\Lambda}_n)^2_{ii}(n+d+1)}{n^2} \overset{n\to\infty}{\longrightarrow} 0 \ \ a.s.$$

Let $l_{i,j}$ be the d-vector with 1 at the i^{th} and j^{th} components and zero elsewhere. According to [31], $\lambda_{ii} + \lambda_{jj} + 2\lambda_{ij} \sim l_{i,j}((1/n)\hat{\Lambda}_n)l'_{i,j}\chi^2_{(n+d+1)/2}$. Then

$$V(\lambda_{ii} + \lambda_{jj} + 2\lambda_{ij}) = \frac{(l_{i,j}\hat{\Lambda}_n l'_{i,j})^2\,(n+d+1)}{n^2} \overset{n\to\infty}{\longrightarrow} 0 \ \ a.s.$$

and then

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

Now let A be any measurable subset of $\Theta \times \mathcal{S}$ including an open neighbourhood of (θ_o, Λ_o). There exists ϵ such that $B_\epsilon(\theta_o, \Lambda_o) \subset A$, where $B_\epsilon(\theta_o, \Lambda_o)$ is the closed parallelotope of side ϵ centred at (θ_o, Λ_o). Moreover, due to the consistency of $(\hat{\theta}_n, \hat{\Lambda}_n)$ there exists $N_\epsilon \in \mathbb{N}$, such that for all $n > N_\epsilon$, $B_{\epsilon/2}(\hat{\theta}_n, \hat{\Lambda}_n) \subset B_\epsilon(\theta_o, \Lambda_o)$ $a.s.$, and

$$\hat{P}(B_{\epsilon/2}(\hat{\theta}_n, \hat{\Lambda}_n)) \le \hat{P}(B_\epsilon(\theta_o, \Lambda_o)) \le \hat{P}(A) \ \ a.s. \tag{A2}$$

Let η_i, $i = 1, \ldots, s, s+1, \ldots, s + d(d+1)/2$, uniformly denote the s components of θ and the $d(d+1)/2$ components of Λ, and $n_p = s + d(d+1)/2$. Let $\hat{\eta}^n_i = \mathrm{E}_{\hat{P}}[\eta_i]$.

$$\hat{P}(B_{\epsilon/2}(\hat{\theta}_n, \hat{\Lambda}_n)) = 1 - \hat{P}(\overline{B_{\epsilon/2}(\hat{\theta}_n, \hat{\Lambda}_n)}) = 1 - \hat{P}\left(\left\{\eta : \max_{i=1,\ldots,n_p} |\eta_i - \hat{\eta}^n_i| > \frac{\epsilon}{2}\right\}\right)$$

For all $i = 1, \ldots, n_p$, from Markov inequality: $\hat{P}(|\eta_i - \hat{\eta}^n_i| > \epsilon/2) \le (4V(\eta_i)/\epsilon^2)$.

Moreover, for all $i = 1, \ldots, n_p$, $\lim_{n\to\infty} V(\eta_i) = 0$, by ii). Then for all $\epsilon > 0$ there exists $N_i \in \mathbb{N}$ such that $(4V(\eta_i)/\epsilon^2) \le \varepsilon$ for $n > N_i$.

Let $N = \max N_i$, $i = 1, \ldots, n_p$. Then for all $n > N$ and all $i = 1, \ldots, n_p$

$$\hat{P}\left(|\eta_i - \hat{\eta}^n_i| > \frac{\epsilon}{2}\right) \le \epsilon \text{ and then } \hat{P}\left(\max_{i=1,\ldots,n_p} |\eta_i - \hat{\eta}^n_i| > \frac{\epsilon}{2}\right) \le \epsilon.$$

Finally, by (A2), for all $\epsilon > 0$ there exists $N \in \mathbb{N}$ such that for $n > N$ it holds

$$\hat{P}(A) \ge 1 - \varepsilon \quad a.s. \tag{A3}$$

A.2 *Proof of Proposition 4.2*

Let us first introduce the following lemma which is necessary to the proof:

LEMMA A.1 *Let f, g be functions $\mathbb{R}^d \to \mathbb{R}$. Let K be a connected compact of $\mathbb{R}^d$. If $\int_K f(x)\mathrm{d}x \neq 0$, if g is continuous, then there exists $x_0 \in K$ such that*

$$\int_K f(x)g(x)\mathrm{d}x = g(x_0)\int_K f(x)\mathrm{d}x$$

Proof Let $a = (\int_K f(x)g(x)\mathrm{d}x / \int_K f(x)\mathrm{d}x)$. As K is compact and g is continuous, there exists m and M such that for all $x \in K$, it holds $m \le g(x) \le M$. The connexity of K ensures therefore that $g(K) = [m, M]$ and then there exists $x_o \in K$ such that $g(x_o) = a$. ■

Let us note that $C \subset C_{\mathcal{S}} = \mathrm{Proj}_\Theta(C) \times \mathcal{S}$.

Let us show that the Kullback–Leibler distance between $\tilde{p}(\theta, \Lambda|Z_n)$ and $p(\theta, \Lambda|Z_n)$ over C tends to zero as n tends to infinity. From Equations (9) and (11) we easily have:

$$\mathcal{K}_C(p, \tilde{p}) = \int_C p(\theta, \Lambda|Z_n) \log \frac{p(\theta, \Lambda|Z_n)}{\tilde{p}(\theta, \Lambda|Z_n)} \mathrm{d}\theta\, \mathrm{d}\Lambda = \log \frac{K_n}{\tilde{K}_n} P(C) + E_{p,C}[\log p(\theta, \Lambda)] \tag{A4}$$

where $P(C) = \int_C p(\theta, \Lambda|Z_n)\mathrm{d}\theta\, \mathrm{d}\Lambda$ and $E_{p,C}[\log p(\theta, \Lambda)] = \int_C \log p(\theta, \Lambda) p(\theta, \Lambda|Z_n)$ $\mathrm{d}\theta\, \mathrm{d}\Lambda$.

- Let us study the behaviour of $E_{p,C}[\log p(\theta, \Lambda)]$:
  For all $\varepsilon > 0$ let $V_\varepsilon = B_\varepsilon(\theta_o, \Lambda_o)$, the closed ball of radius ε and centred at (θ_o, Λ_o) as defined in Section 2. For ε sufficiently small it holds $V_\varepsilon \subset C$. Therefore,

$$\begin{aligned} E_{p,C}[\log p(\theta, \Lambda)] &= \int_C \log p(\theta, \Lambda) p(\theta, \Lambda|Z_n)\mathrm{d}\theta\, \mathrm{d}\Lambda \\ &= \int_{C\setminus V_\varepsilon} \log p(\theta, \Lambda) p(\theta, \Lambda|Z_n)\mathrm{d}\theta\, \mathrm{d}\Lambda \\ &\quad + \int_{V_\varepsilon} \log p(\theta, \Lambda) p(\theta, \Lambda|Z_n)\mathrm{d}\theta\, \mathrm{d}\Lambda \end{aligned}$$

Hence

$$\left| E_{p,C}[\log p(\theta, \Lambda)] - \int_{V_\varepsilon} \log p(\theta, \Lambda) p(\theta, \Lambda|Z_n)\mathrm{d}\theta\, \mathrm{d}\Lambda \right|$$

$$\leq \sup_{\theta, \Lambda \in C} \left| \log p(\theta, \Lambda) \right| \int_{C\setminus V_\varepsilon} p(\theta, \Lambda|Z_n)\mathrm{d}\theta\, \mathrm{d}\Lambda$$

Let us first note that by H_4 there exists a finite positive constant $Q(C)$ such that $Q(C) = \sup_{\theta, \Lambda \in C} |\log p(\theta, \Lambda)|$.

We know by [17, 22, 23] that under H_2, $p(\theta, \Lambda|Z_n)$ is consistent as n tends to infinity: it concentrates around (θ_o, Λ_o), the true parameter values when model M is the correct one, and more generally the parameter values minimizing the Kullback–Leibler criterion between the true (x, y) data distribution and that induced by model M when it is incorrect.

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

$$\int_{C\setminus V_\varepsilon} p(\theta, \Lambda|Z_n)\mathrm{d}\theta\, \mathrm{d}\Lambda < \varepsilon \tag{A5}$$

and

$$0 \leq P(C) - \int_{V_\varepsilon} p(\theta, \Lambda|Z_n)\mathrm{d}\theta\, \mathrm{d}\Lambda < \varepsilon \tag{A6}$$

Moreover by Lemma A1, for all n there exists $(\tilde{\theta}, \tilde{\Lambda}) \in V_\varepsilon$ such that

$$\int_{V_\varepsilon} \log p(\theta, \Lambda) p(\theta, \Lambda|Z_n)\mathrm{d}\theta\, \mathrm{d}\Lambda = \log p(\tilde{\theta}, \tilde{\Lambda}) \int_{V_\varepsilon} p(\theta, \Lambda|Z_n)\mathrm{d}\theta\, \mathrm{d}\Lambda \tag{A7}$$

By (A6) there exists $\tilde{\varepsilon} : 0 \leq \tilde{\varepsilon} \leq \varepsilon$ such that $\int_{V_\varepsilon} p(\theta, \Lambda|Z_n)\mathrm{d}\theta\, \mathrm{d}\Lambda + \tilde{\varepsilon} = P(C)$.

Then

$$\left| E_{p,C}[\log p(\theta, \Lambda)] - \log p(\tilde{\theta}, \tilde{\Lambda})(P(C) - \tilde{\varepsilon}) \right| \leq \varepsilon Q(C) \tag{A8}$$

But $\lim_{\varepsilon \to 0} \log p(\tilde{\theta}, \tilde{\Lambda})(P(C) - \tilde{\varepsilon}) = \log p(\theta_o, \Lambda_o) P(C)$. Therefore,

$$\lim_{n \to \infty} E_{p,C}[\log p(\theta, \Lambda)] = \log p(\theta_o, \Lambda_o) P(C) \tag{A9}$$

- Let us study now the behaviour of $\log(K_n/\tilde{K}_n)$:
  For brevity's sake let us denote $f_n(\theta, \Lambda) = |\Lambda|^{n/2} \exp\{-1/2 \sum_{i=1}^n \|y_i - f(x_i, \theta)\|_\Lambda^2\}$.

  From Equations (9) and (11): $K_n^{-1} = \int f_n(\theta, \Lambda) p(\theta, \Lambda)\mathrm{d}\theta\, \mathrm{d}\Lambda$ and $\tilde{K}_n^{-1} = \int_{C_S} f_n(\theta, \Lambda)\mathrm{d}\theta\, \mathrm{d}\Lambda$.

  Let $\epsilon > 0$ and $V_\epsilon = B_\epsilon(\theta_o, \Lambda_o)$. As $p(\theta, \Lambda|Z_n)$ and $\tilde{p}(\theta, \Lambda|Z_n)$ are consistent, for all ϵ there exist N_ϵ and $\tilde{N}_\epsilon$ such that for all $n > N_\epsilon$

$$0 \leq 1 - K_n \int_{V_\epsilon} f_n(\theta, \Lambda) p(\theta, \Lambda)\mathrm{d}\theta\, \mathrm{d}\Lambda \leq \epsilon \tag{A10}$$

and for all $n > \tilde{N}_\epsilon$

$$0 \leq 1 - \tilde{K}_n \int_{V_\epsilon} f_n(\theta, \Lambda)\mathrm{d}\theta\, \mathrm{d}\Lambda \leq \epsilon \tag{A11}$$

From Equation (A11) there exists $\tilde{\epsilon} \leq \epsilon$ such that $\int_{V_\epsilon} f_n(\theta, \Lambda)\mathrm{d}\theta\, \mathrm{d}\Lambda = (1 - \tilde{\epsilon})/\tilde{K}_n$.

Moreover according to Lemma A.1 there exist $(\tilde{\theta}, \tilde{\Lambda}) \in V_\epsilon$ such that

$$\begin{aligned}\int_{V_\epsilon} f_n(\theta, \Lambda) p(\theta, \Lambda) \mathrm{d}\theta\, \mathrm{d}\Lambda &= p(\tilde{\theta}, \tilde{\Lambda}) \int_{V_\epsilon} f_n(\theta, \Lambda) \mathrm{d}\theta\, \mathrm{d}\Lambda \\ &= p(\tilde{\theta}, \tilde{\Lambda}) \frac{1-\tilde{\epsilon}}{\tilde{K}_n} \end{aligned} \tag{A12}$$

which with Equation (A10) gives

$$\left| 1 - p(\tilde{\theta}, \tilde{\Lambda}) \frac{K_n}{\tilde{K}_n} (1-\tilde{\epsilon}) \right| \le \epsilon \tag{A13}$$

By assumption H_4 $p(\theta, \Lambda)$ is strictly positive and bounded over $V_\epsilon = B_\epsilon(\theta_o, \Lambda_o)$. We deduce that

$$\lim_{n\to\infty} \log \frac{K_n}{\tilde{K}_n} = -\log p(\theta_o, \Lambda_o)$$

By Equations (A9) and (A4) we finally get

$$\lim_{n\to\infty} \mathcal{K}_C(p, \tilde{p}) = 0 \tag{A14}$$

which leads to Equation (12) and Proposition 4.2, since L_1-convergence is ensured by Kullback convergence.

4.3 *Proof of Proposition 4.3*

By definition

$$\tilde{p}(\theta, \Lambda | Z_n) = \tilde{K}_n |\Lambda|^{n/2} \exp\left\{ -\frac{1}{2} \sum_{i=1}^n \|y_i - f(x_i, \theta)\|^2_\Lambda \right\} \times I_{C_\mathcal{S}}(\theta, \Lambda) \tag{A15}$$

and

$$\hat{p}(\theta, \Lambda | Z_n) = \hat{K}_n |\Lambda|^{n/2} \exp\left\{ -\frac{1}{2} \|\theta - \hat{\theta}_n\|^2_{\beta_n} - \frac{1}{2} \sum_{i=1}^n \|y_i - f(x_i, \hat{\theta}_n)\|^2_\Lambda \right\} \tag{A16}$$

where $\beta_n = \sum_{i=1}^n \dot{f}'_{x_i, \hat{\theta}_n} \hat{\Lambda}_n \dot{f}_{x_i, \hat{\theta}_n}$.

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

Let us consider the Kullback–Leibler distance between $\tilde{p}$ and $\hat{p}$ over $C_\mathcal{S}$

$$\begin{aligned}\mathcal{K}_{C_\mathcal{S}}(\hat{p}, \tilde{p}) &= \int_{C_\mathcal{S}} \hat{p}(\theta, \Lambda | Z_n) \log \frac{\hat{p}(\theta, \Lambda | Z_n)}{\tilde{p}(\theta, \Lambda | Z_n)} \mathrm{d}\theta\, \mathrm{d}\Lambda \\ &= \log \frac{\hat{K}_n}{\tilde{K}_n} + E_{\hat{p}, C_\mathcal{S}} \left[-\frac{1}{2} \|\theta - \hat{\theta}_n\|^2_{\beta_n} - \frac{1}{2} \sum_{i=1}^n \|y_i - f(x_i, \hat{\theta}_n)\|^2_\Lambda \right] \\ &\quad + E_{\hat{p}, C_\mathcal{S}} \left[\frac{1}{2} \sum_{i=1}^n \|y_i - f(x_i, \theta)\|^2_\Lambda \right] \end{aligned} \tag{A17}$$

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

- Let $\mathcal{E}_{C_\mathcal{S}, n} = E_{\hat{p}, C_\mathcal{S}}[-(1/2)\|\theta - \hat{\theta}_n\|^2_{\beta_n}]$.

  By definition of $\hat{p}(\theta, \Lambda)$ in Equation (10): $0 > \mathcal{E}_{C_\mathcal{S}, n} \ge E_{\hat{p}}[-(1/2)\|\theta - \hat{\theta}_n\|^2_{\beta_n}] = -s/2$.
- For all $i = 1, \ldots, n$, in a neighbourhood of $\hat{\theta}_n$ it holds

$$\begin{aligned}\|y_i - f(x_i, \theta)\|^2_\Lambda &= \|y_i - f(x_i, \hat{\theta}_n) + \dot{f}_{x_i, \hat{\theta}_n}(\theta - \hat{\theta}_n)\|^2_\Lambda + o(\|\theta - \hat{\theta}_n\|^2) \\ &= \|y_i - f(x_i, \hat{\theta}_n)\|^2_\Lambda + \|\dot{f}_{x_i, \hat{\theta}_n}(\theta - \hat{\theta}_n)\|^2_\Lambda \\ &\quad + 2\langle y_i - f(x_i, \hat{\theta}_n), \dot{f}_{x_i, \hat{\theta}_n}(\theta - \hat{\theta}_n)\rangle_\Lambda + o(\|\theta - \hat{\theta}_n\|^2) \end{aligned} \tag{A18}$$

Therefore,

$$\begin{aligned}&\frac{1}{2} \sum_{i=1}^n \|y_i - f(x_i, \theta)\|^2_\Lambda - \frac{1}{2} \sum_{i=1}^n \|y_i - f(x_i, \hat{\theta}_n)\|^2_\Lambda \\ &= \frac{1}{2} \sum_{i=1}^n \|\dot{f}_{x_i, \hat{\theta}_n}(\theta - \hat{\theta}_n)\|^2_\Lambda + \sum_{i=1}^n \langle y_i - f(x_i, \hat{\theta}_n), \dot{f}_{x_i, \hat{\theta}_n}(\theta - \hat{\theta}_n)\rangle_\Lambda + n\, o\big(\|\theta - \hat{\theta}_n\|^2\big) \end{aligned} \tag{A19}$$

By Equation (2) and by definition of $\hat{p}(\theta, \Lambda | Z_n)$ in Equation (10): $E_{\hat{p}}[\Lambda] = ((n + d + 1)/n)\hat{\Lambda}_n$.

Therefore,

$$E_{\hat{p},C_{\mathcal{S}}}\left[\sum_{i=1}^{n}\|\dot{f}_{x_i,\hat{\theta}_n}(\theta-\hat{\theta}_n)\|_{\Lambda}^2\right]=E_{\hat{p},C_{\mathcal{S}}}\left[E_{\hat{p}}\left[\sum_{i=1}^{n}\|\dot{f}_{x_i,\hat{\theta}_n}(\theta-\hat{\theta}_n)\|_{\Lambda}^2|\theta\right]\right]$$

$$=E_{\hat{p},C_{\mathcal{S}}}\left[\frac{n+d+1}{n}\sum_{i=1}^{n}\|\dot{f}_{x_i,\hat{\theta}_n}(\theta-\hat{\theta}_n)\|_{\hat{\Lambda}_n}^2\,\right|$$

$$=\frac{n+d+1}{n}E_{\hat{p},C_{\mathcal{S}}}\left[\|\theta-\hat{\theta}_n\|_{\beta_n}^2\right]$$

$$=-2\frac{n+d+1}{n}\mathcal{E}_{C_{\mathcal{S}},n} \tag{A20}$$

Similarly

$$E_{\hat{p},C_{\mathcal{S}}}\left[\sum_{i=1}^{n}\langle y_i-f(x_i,\hat{\theta}_n),\,\dot{f}_{x_i,\hat{\theta}_n}(\theta-\hat{\theta}_n)\rangle_{\Lambda}\right]$$

$$=E_{\hat{p},C_{\mathcal{S}}}\left[E_{\hat{p}}\left[\sum_{i=1}^{n}\langle y_i-f(x_i,\hat{\theta}_n),\,\dot{f}_{x_i,\hat{\theta}_n}(\theta-\hat{\theta}_n)\rangle_{\Lambda}|\theta\right]\right]$$

$$=\frac{n+d+1}{n}E_{\hat{p},C_{\mathcal{S}}}\left[\sum_{i=1}^{n}\langle y_i-f(x_i,\hat{\theta}_n),\,\dot{f}_{x_i,\hat{\theta}_n}(\theta-\hat{\theta}_n)\rangle_{\hat{\Lambda}_n}\right]$$

$$=\frac{n+d+1}{n}E_{\hat{p},C_{\mathcal{S}}}[0]$$

$$=0 \tag{A21}$$

as $(\hat{\theta}_n,\hat{\Lambda}_n)$ are least squares estimators.

By Equations (A20) and (A21) the Kullback–Leibler distance between $\tilde{p}$ and $\hat{p}$ comes to

$$\mathcal{K}_{C_{\mathcal{S}}}(\hat{p},\tilde{p})=\log\frac{\hat{K}_n}{\tilde{K}_n}+\mathcal{E}_{C_{\mathcal{S}},n}-\frac{n+d+1}{n}\mathcal{E}_{C_{\mathcal{S}},n}+E_{\hat{p},C_{\mathcal{S}}}\left[n\,o(\|\theta-\hat{\theta}_n\|^2)\right] \tag{A22}$$

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

We saw previously in Section A1 that there exists a matrix $\tilde{\beta}$ such that $\beta_n=V(\theta)^{-1}\overset{n\to\infty}{\simeq}n\tilde{\beta}$. As all norms on $\mathbb{R}^d$ are equivalent, there exists positive scalars α_1 and α_2 such that

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

or

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

and for sufficiently great n

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

According to the $\hat{p}$ distribution (10), $\|\theta-\hat{\theta}_n\|_{\beta_n}^2$ is distributed as a χ_s^2 variable.

Taking expectations

$$\alpha_1 q\le E_{\hat{p}}[n\|\theta-\hat{\theta}_n\|^2]\le\alpha_2 q$$

There exists then two positive scalars $\tilde{\alpha}_1$ and $\tilde{\alpha}_2$ such that

$$\tilde{\alpha}_1\le E_{\hat{p},C_{\mathcal{S}}}\left[n\|\theta-\hat{\theta}_n\|^2\right]\le\tilde{\alpha}_2 \tag{A23}$$

Let us come back to the term $o(\|\theta-\hat{\theta}_n\|^2)$ in Equation (A18):

For all couples (x_i,y_i) let us denote $g_n^i(\theta)=o(\|\theta-\hat{\theta}_n\|^2)$. Then

$$\sum_{i=1}^{n}g_n^i(\theta)=n\,o\big(\|\theta-\hat{\theta}_n\|^2\big)$$

Let $\tilde{g}_n^i(\theta)=(g_n^i(\theta)/\|\theta-\hat{\theta}_n\|^2)$ and $\tilde{g}_n(\theta)=\frac{1}{n}\sum_{i=1}^{n}\tilde{g}_n^i(\theta)$. Then $\lim_{\theta\to\hat{\theta}_n}\tilde{g}_n(\theta)=0$.

As $\lim_{n\to\infty}\hat{\theta}_n = \theta_o$ a.s., for all $i \in \mathbb{N}^*$ we have $\lim_{n\to\infty}\tilde{g}_n^i(\theta_o) = 0$. 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:

$$\forall\varepsilon > 0, \exists N_\varepsilon, \forall x \in \mathcal{X}, \forall n > N_\varepsilon, |\tilde{g}_n^i(\theta_v)| < \varepsilon$$

and

$$\forall\varepsilon > 0, \exists N_\varepsilon, \forall n > N_\varepsilon, |\tilde{g}_n(\theta_o)| < \varepsilon$$

then

$$\lim_{n\to\infty}\tilde{g}_n(\theta_o) = 0$$

Now

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

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

$$\begin{aligned} E_{\hat{p},C_{\mathcal{S}}}[n\ o(\|\theta - \hat{\theta}_n\|^2)] &= \int_{C_{\mathcal{S}}\setminus V_\varepsilon} \tilde{g}_n(\theta)n\|\theta - \hat{\theta}_n\|^2\hat{p}(\theta, \Lambda|Z_n)\mathrm{d}\theta\,\mathrm{d}\Lambda \\ &\quad + \int_{V_\varepsilon} \tilde{g}_n(\theta)n\|\theta - \hat{\theta}_n\|^2\hat{p}(\theta, \Lambda|Z_n)\mathrm{d}\theta\,\mathrm{d}\Lambda \end{aligned} \tag{A24}$$

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 $n > N_\varepsilon$

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

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 $\{\theta_o\}$. As n tends to ∞, then $\sup_{V_\varepsilon}\tilde{g}_n(\theta)$ tends to $\lim_{n\to\infty}\tilde{g}_n(\theta_o) = 0$.

We conclude that

$$\lim_{n\to\infty} E_{\hat{p},C_{\mathcal{S}}}\left[n\ o\left(\|\theta - \hat{\theta}_n\|^2\right)\right] = 0 \tag{A25}$$

- Finally let us show that $\lim_{n\to\infty}\log(\hat{K}_n/\tilde{K}_n) = 0$.

  To alleviate notations let us denote, from Equations (A15) and (A16):

$$\tilde{p} = \tilde{K}_n \times \tilde{f}_n \quad \text{and} \quad \hat{p} = \hat{K}_n \times \hat{f}_n.$$

Let us follow a *reductio ad absurdum* by assuming that $\lim_{n\to\infty}\hat{K}_n/\tilde{K}_n \neq 1$.

Because of Equations (A19)–(A21),

$$E_{\hat{p},C_{\mathcal{S}}}\left[\log\frac{\hat{f}_n}{\tilde{f}_n}\right] \longrightarrow 0 \quad \text{and} \quad E_{\hat{p},C_{\mathcal{S}}}\left[\log\frac{\tilde{f}_n}{\hat{f}_n}\right] \longrightarrow 0 \quad \text{as } n \longrightarrow \infty \tag{A26}$$

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

$$\begin{aligned} \lim_{n\to\infty} E_{\hat{p},C_{\mathcal{S}}}\left[\frac{\tilde{f}_n}{\hat{f}_n}\right] &= \lim_{n\to\infty}\hat{K}_n\int_{C_{\mathcal{S}}}\tilde{f}_n\mathrm{d}\theta\,\mathrm{d}\Lambda \\ &< \lim_{n\to\infty}\tilde{K}_n\int_{C_{\mathcal{S}}}\tilde{f}_n\mathrm{d}\theta\,\mathrm{d}\Lambda \\ &= \lim_{n\to\infty}\int_{C_{\mathcal{S}}}\tilde{K}_n\tilde{f}_n\mathrm{d}\theta\,\mathrm{d}\Lambda = 1 \end{aligned} \tag{A27}$$

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 $E_{\hat{p},C_{\mathcal{S}}}[(\tilde{f}_n/\hat{f}_n)]$ and gives

$$\exp E_{\hat{p},C_{\mathcal{S}}}\left[\log\frac{\tilde{f}_n}{\hat{f}_n}\right] \le E_{\hat{p},C_{\mathcal{S}}}\left[\frac{\tilde{f}_n}{\hat{f}_n}\right]$$

but by Equation (A26)

$$\exp E_{\hat{p},C_{\mathcal{S}}}\left[\log\frac{\tilde{f}_n}{\hat{f}_n}\right] \overset{n\to\infty}{\longrightarrow} 1$$

then $\lim_{n\to\infty} E_{\hat{p},C_{\mathcal{S}}}\left[\tilde{f}_n/\hat{f}_n\right] \ge 1$, which contradicts (A27).

(ii) Suppose now that $\lim_{n\to\infty}(\hat{K}_n/\tilde{K}_n) > 1$:

By a similar reasoning and since $\tilde{p}(\theta, \Lambda|Z_n)$ is consistent

$$\lim_{n\to\infty} E_{\hat{p},C_{\mathcal{S}}}\left[\frac{\tilde{f}_n}{\hat{f}_n}\right] > \lim_{n\to\infty}\int_{C_{\mathcal{S}}} \tilde{K}_n \tilde{f}_n \mathrm{d}\theta\, \mathrm{d}\Lambda = 1$$

which implies

$$\lim_{n\to\infty} \log E_{\hat{p},C_{\mathcal{S}}}\left[\frac{\hat{f}_n}{\tilde{f}_n}\right] < 0. \tag{A28}$$

Due to the convexity of the log function and the possibility to apply the Jensen inequality to $E_{\hat{p},C_{\mathcal{S}}}[(\hat{f}_n/\tilde{f}_n)]$, for sufficiently great n, we have by Equation (A26)

$$-\log E_{\hat{p},C_{\mathcal{S}}}\left[\frac{\hat{f}_n}{\tilde{f}_n}\right] \leq E_{\hat{p},C_{\mathcal{S}}}\left[\log\frac{\tilde{f}_n}{\hat{f}_n}\right] \overset{n\to\infty}{\longrightarrow} 0$$

which contradicts (A28).

From Equation (i) and (ii) we can deduce that $\lim_{n\to\infty}\log(\hat{K}_n/\tilde{K}_n) = 0$, and then, from (A22) and (A25) that

$$\mathcal{K}_{C_{\mathcal{S}}}(\hat{p}, \tilde{p}) \overset{n\to\infty}{\longrightarrow} 0 \tag{A29}$$

This completes the proof of Proposition 4.3, since Kullback convergence dominates L_1 convergence over $C_{\mathcal{S}}$ and then over C.

A.4 *Proof of Theorem 4.6*

Let

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

By Fubini's theorem

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

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 tends to ∞. Let us show that the same is true for T_2.

Let $h(\theta, \hat{\theta}_n) = \int|\hat{p}(y|\theta, \Lambda, x) - p(y|\theta, \Lambda, x)|\mathrm{d}y$. 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 $\lim_{n\to\infty}(\hat{\theta}_n, \hat{\Lambda}_n) = (\theta_o, \Lambda_o)$ *a.s.*, we deduce that $\lim_{n\to\infty} h(\theta_o, \hat{\theta}_n) = 0$. Moreover, for all $\varepsilon > 0$ there exists a neighbourhood of (θ_o, Λ_o), 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:

$$T_2 = \int \hat{p}(\theta, \Lambda | Z_n) h(\theta, \hat{\theta}_n) \mathrm{d}\theta\, \mathrm{d}\Lambda$$

$$T_2 = \int_{V_\varepsilon} + \int_{V_\varepsilon^c}$$

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

$$T_2 \leq 2 \int_{V_\varepsilon^c} \hat{p}(\theta, \Lambda | Z_n) + \varepsilon/2$$

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

It follows that D tends to zero as n tends to ∞.

Appendix B: Case studies result tables

B.1 *Simulated selection problem*

Table B1. U criterion scoring: analytic approximation.

Nb obs	f_1	f_2	$\boldsymbol{f_3}$	f_4	f_5
50	−1.5823	−1.4324	−1.3177	−1.3614	**−1.1165**
100	−1.6620	−1.5919	**−1.4265**	−1.6600	−1.4819
200	−1.7355	−1.6668	**−1.4998**	−1.8800	−1.5703

Table B2. U criterion scoring: Hastings approximation.

Nb obs	f_1	f_2	$\boldsymbol{f_3}$	f_4	f_5
50	−1.6814	−1.7920	−1.5949	−1.9814	**−1.3313**
100	−1.7231	−1.9211	−1.5654	−2.0792	**−1.5409**
200	−1.7681	−1.7771	−1.6188	−2.1698	**−1.5897**

Table B3. U criterion scoring: hybrid G-H approximation.

Nb obs	f_1	f_2	$\boldsymbol{f_3}$	f_4	f_5
50	−1.7472	−1.5171	−1.5274	−1.7583	**−1.4801**
100	−1.7949	−1.6341	**−1.5702**	−1.9610	−1.6757
200	−1.8846	−1.7844	**−1.6377**	−2.0694	−1.8160

Table B4. AIC criterion scoring ($\times 10^{-3}$).

Nb obs	f_1	f_2	$\boldsymbol{f_3}$	f_4	f_5
50	0.2131	**0.2094**	0.2097	0.2113	0.2107
100	0.4332	0.4290	**0.4289**	0.4315	0.4311
200	0.8735	0.8662	0.8639	0.8677	**0.8628**

Table B5. BIC criterion scoring ($\times 10^{-3}$).

Nb obs	f_1	f_2	$\boldsymbol{f_3}$	f_4	f_5
50	0.2306	**0.2268**	0.2330	0.2346	0.2399
100	0.4549	**0.4507**	0.4577	0.4604	0.4672
200	0.8993	**0.8920**	0.8983	0.9021	0.9058

Table B6. CV_I criterion scoring.

Nb obs	f_1	f_2	$\boldsymbol{f_3}$	f_4	f_5
50	1.2309	1.2344	**1.2072**	1.2334	1.2084
100	1.4553	**1.4481**	1.4486	1.4638	1.4725
200	1.5382	1.5351	1.5288	1.5286	**1.5219**

Table B7. CV_Q criterion scoring.

Nb obs	f_1	f_2	$\boldsymbol{f_3}$	f_4	f_5
50	1.7756	**1.6422**	1.6797	1.7100	1.6611
100	1.7283	1.6847	**1.6820**	1.7035	1.6916
200	1.7478	1.7121	1.7006	1.7222	**1.6900**

B.2 *Actual model selection problem in soil science*

Table B8. Scorings of the five neural models according to the four criteria.

Neural model	NN1	NN2	NN3	NN4	NN5
U	**18.4202**	17.8137	16.4986	15.0374	17.0372
CV_Q	**1.2272**	1.3144	1.2456	2.1509	3.1021
AIC $\times 10^{-3}$	2.2468	2.3577	2.2575	**2.2335**	2.5105
BIC $\times 10^{-3}$	2.7086	2.7647	2.6645	**2.3196**	3.1601

Table B9. A typical MSEP scoring of the neural models.

Neural model	NN1	NN2	NN3	NN4	NN5
MSEP $\times 10^{-2}$	**0.5227**	0.5869	0.6126	0.6967	0.8384