Neural Computation, vol. 16, pp. 419 – 443 (2004)

REPLY TO THE COMMENTS ON "Local Overfitting Control via Leverages" in "Jacobian Conditioning Analysis for Model Validation" by I. Rivals and L. Personnaz

Yacine OUSSAR, Gaétan MONARI, Gérard DREYFUS
ESPCI, Laboratoire d'Électronique
10, rue Vauquelin
F – 75005 PARIS - FRANCE

Abstract

"Jacobian Conditioning Analysis for Model Validation" by Rivals and Personnaz is a comment on Monari and Dreyfus (2002). In the present reply, we disprove their claims. We point to flawed reasoning in the theoretical comments, and to errors and inconsistencies in the numerical examples. Our replies are substantiated by seven counter-examples, inspired from real data, which show that (i) the comments on the accuracy of the computation of the leverages are unsupported, and that (ii) following the approach they advocate leads to discarding valid models, or validating overfitted models.

Neural Computation, vol. 16, pp. 419 – 443 (2004)

REPLY TO THE COMMENTS ON "Local Overfitting Control via Leverages" in "Jacobian Conditioning Analysis for Model Validation" by I. Rivals and L. Personnaz

Yacine OUSSAR, Gaétan MONARI, Gérard DREYFUS
ESPCI, Laboratoire d'Électronique
10, rue Vauquelin
F – 75005 PARIS - FRANCE

1. INTRODUCTION

"Jacobian Conditioning Analysis for Model Validation" by Rivals and Personnaz (this issue) is a detailed comment on Monari and Dreyfus (2002), from the second sentence of their abstract to the last paragraph of their text. The authors claim that we "followed" a previous, controversial (Larsen et al., 2001) paper of theirs (Rivals et al., 2000). In this reply, we disprove all their claims. We point to flawed reasoning in their theoretical comments, and to errors and inconsistencies in the numerical examples. Our replies to the comments are substantiated by seven counter-examples, which show that (i) their comments on the accuracy of the computation of the leverages are unsupported, and that (ii) following their approach leads to making wrong decisions: discarding valid models, or validating overfitted models.

The present paper is organized as follows: in the first section, we disprove the comments in "Jacobian Conditioning Analysis for Model Validation" on the accuracy of the computation of the leverages. In the second section, we show that their comments on model validation are erroneous, and we provide several counter-examples in which decisions made on the basis of the condition-number selection criterion advocated by the authors are wrong. We conclude by summarizing the arguments that disprove all three conclusions of Rivals and Personnaz's comments.

2. REPLY TO THE COMMENTS ON THE ACCURACY OF THE COMPUTATION OF THE LEVERAGES

Section 1 of "Jacobian Conditioning Analysis for Model Validation" is entitled "On the jacobian matrix of a nonlinear model". The authors first recall theoretical results, and they recall the validation method advocated in (Rivals et al., 2000). That part will be replied to in section 3 of this article.

Section 2 of "Jacobian Conditioning Analysis for Model Validation" is entitled "Comment for the proposition of (Monari and Dreyfus, 2002)". The authors claim that relation (A.4) of our paper (relation (15) of their comment) is not accurate enough for the purpose of model validation, and that a more classical computation method should be used instead. In order to substantiate their claims, they exhibit a small handcrafted numerical example (section 3.1 of their paper): we show below that it is irrelevant to model selection, and that it is inconsistent with the claims made by the authors in the previous section of their paper. They show that the traditional method can compute the leverages with an accuracy of 10^{-16}; however, we show in the following that (i) the authors fail to provide evidence that such accuracy is relevant in the context of nonlinear model selection, and that (ii) they fail to provide evidence that our method does not meet the actual accuracy requirements in that context.

The recommended approach to a problem in numerical analysis consists in asking two questions: (1) what numerical accuracy should be achieved in order to get insight into the problem at hand? (2) how can the above accuracy be achieved? Obviously, that approach should be used in discussing the accuracy of the computation of the leverages for model selection. In (Monari et al. 2002), we discussed nonlinear model selection in the context of machine learning; the computation of the leverages of the observations is one of the ingredients of the original validation method that we describe. Therefore, question (1), which is *not* asked in "Jacobian Conditioning Analysis for Model Validation", is: for the purpose of model selection, what is the desired accuracy for the computation of the leverages of models obtained by training from examples? The answer to that question is straightforward: the accuracy should be of the magnitude of the "noise" on the leverages. Since training is generally performed by iterative optimization of a cost function, it is stopped when the two models obtained in two successive iterations are considered "identical". Therefore, the numerical "noise" on the leverages is the difference between the values of the leverages of two models that are considered "identical". Two models are considered identical if some

criterion is met, e.g. if the variation of the cost function is smaller than a prescribed value, or if the variation of the magnitudes of the parameter vector is smaller than a prescribed value, or similar criteria. Rivals and Personnaz do not address the question of estimating the numerical noise on the leverages; they claim that the accuracy should be on the order of 10^{-16}, without providing any evidence that such accuracy is relevant to the problem at hand. Therefore, their criticisms are unsupported in the context of machine learning.

Conversely, we describe, in section 3, examples of nonlinear model selection: neural networks were trained from examples by minimizing the least squares cost function with the Levenberg-Marquardt algorithm. Training was terminated when the magnitude of the relative variation of the cost function between two iterations of the algorithm became smaller than 10^{-10}, which is an extremely conservative stopping criterion (a typical stopping criterion would be a relative difference of 10^{-5} or even more). The root mean square of the variations of the leverages was computed as:

$$\Delta h(k) = \sqrt{\frac{1}{N}\sum_{i=1}^{N}\left(h_{ii}(k+1) - h_{ii}(k)\right)^2}$$

where $h_{ii}(k)$ is the leverage of observation i at iteration k of the training algorithm, and N is the number of examples. Iterations k and k+1 were chosen such that the relative variation of the cost function was smaller that 10^{-10}. The values of Δh were consistently found to be substantially larger than 10^{-10}: clearly, there is no point in computing the leverages with an accuracy of 10^{-16}, while the "noise" on the leverages is actually larger by more than six orders of magnitude, even in unusually conservative conditions. Furthermore, the discrepancy between the leverages computed by the traditional method (advocated in the comment) and by our method was smaller than 10^{-12}: it is smaller, by several orders of magnitude, than the noise on the leverages, hence is not significant.

Actually, in most real-life applications, training will be terminated much earlier; therefore, the root-mean-square difference between the leverages of two models that are considered equivalent will be larger by still many more orders of magnitude.

In order to substantiate their claims, Rivals and Personnaz provide a handcrafted numerical example whose results are shown on Table 1 of their paper. In the following, we show that those results are inconsistent with the authors' claims in "Jacobian Conditioning Analysis", and irrelevant to model selection. The authors consider a jacobian matrix with two columns: the elements of one of them are equal to one, and the second column vector is equal to the

sum of the first column vector and of a "small" vector which is a normally distributed random vector multiplied by a scalar ranging from 10^{-6} to 10^{-15}. Most results presented in that table are inconsistent and irrelevant, for the following reasons:

1. Rivals and Personnaz fail to state that most models used to derive the numerical results presented in that table are actually discarded by the selection criterion that they advocate: out of 10,000 random realizations of the jacobian matrix, *all* models with $a = 10^{-12}$ and $a = 10^{-15}$ are discarded for having a condition number larger than 10^{+8}; 9,988 models out of 10,000 with $a = 10^{-8}$ are discarded for the same reason; conversely, all models with $a = 10^{-6}$ are accepted; therefore, the only results of table 1 that are relevant to model selection, according to the selection criterion advocated by the authors, are the results pertaining to $a = 10^{-6}$ (first row of the table); the other results are irrelevant, since there is no point in discussing the accuracy of the computation of leverages for models that are discarded by the selection criterion advocated by the authors. Moreover, the results reported in the first row show that our method computes the leverages with an accuracy of 10^{-10}, which is on the order of the noise on the leverages, as shown above.
2. The structure of the jacobian matrix indicates that the model has two parameters θ_0 and θ_1, and is of the form

   $y(x) = \theta_0 + f(x, \theta_1)$, with $\frac{\partial f}{\partial \theta_1} = 1 + \varepsilon(x, \theta_1)$

   where the values of $\varepsilon(x, \theta_1)$ can be modeled as realizations of a random variable equal to a normal variable multiplied by a factor of 10^{-6} to 10^{-15}. Therefore, the model is of the form

   $y((x) = \theta_0 + \theta_1 + \gamma(x, \theta_1)$ with $\frac{\partial \gamma}{\partial \theta_1} = \varepsilon(x, \theta_1)$.

   No modeler would ever design a model with such a functional form: an obvious re-parameterization of the model consists in defining a new parameter $\theta_2 = \theta_0 + \theta_1$, so that the first column is still made of 1's (derivative of y with respect to θ_2) and the second column is made of the values of $\varepsilon(x, \theta_1)$, modeled as small random Gaussian variables; thus the jacobian matrix is much better conditioned. If the same experiment is performed as in table 1, after re-parameterization: (i) for $a = 10^{-6}$, the difference between the number of parameters and the sum of the leverages computed by our method becomes $2.\ 10^{-16}$, so that the traditional method is not more accurate than ours; (ii) for $a = 10^{-8}$, about 25% of the models are accepted by the selection criterion

advocated by Rivals and Personnaz, and, as above, the accuracy of the computation of the leverages by our method is on the order of 10^{-16}, i.e. is comparable to that of the traditional method.

Therefore, the example displayed in Table 1 of "Jacobian Conditioning Analysis" is handcrafted to prove that the traditional method of computing the leverages is more accurate than ours in that specific case; however, that case is irrelevant to model selection because (1) there is no point in computing accurately the leverages of models that, according to the selection criterion advocated in the same paper, would actually be discarded, and because (2) no knowledgeable modeler would design a model having a jacobian matrix of that form.

To summarize, in "Jacobian Conditioning Analysis", Rivals and Personnaz do not provide any evidence that an accuracy of 10^{-16} for the computation of the leverages is desirable in the context of machine learning, nor do they provide any evidence that our method does not achieve the accuracy that is required in realistic conditions. Therefore, the claim that the traditional method is superior to ours in the machine learning context is unsupported.

In the next sections, additional counter-examples further support the above conclusions.

Furthermore, it should be noted that the issue of numerical accuracy was far from central in our paper, relation A4 being presented in an appendix. Therefore, we did not find it necessary to elaborate on that question, since it can be answered in a straightforward fashion, as shown above.

3. REPLY TO THE COMMENTS ON MODEL VALIDATION

In the previous section, we showed that the second conclusion stated in section 4 of "Jacobian Conditioning Analysis" is unsupported. In the present section, we disprove the other two conclusions of their comment, together with claims made in (Rivals et al., 2000). We show that, contrary to the statement of the authors, we did not follow the approach advocated in that paper, since it is not correct from a numerical analysis point of view, and leads to making wrong decisions for model validation.

3.1. Reply to section 1.1 of (Rivals et al., 2003)

The use of the jacobian matrix (relations (2) and (3) of (Rivals et al., 2003)) for the analysis of the identifiability of nonlinear models is classical in statistics. It can be traced back to 1956.

Relation 1.4 of (Rivals et al., 2003) describes one of several standard confidence interval estimates for nonlinear models. It can be found in textbooks on nonlinear regression (see for instance (Seber et al., 1989)). In view of the contents of section 3.2 of the present reply, it is relevant to note that many other confidence interval estimation methods for nonlinear models can be found in the literature (Tibshirani, 1996).

Relation 1.5 of "Jacobian Conditioning Analysis" is an unfounded claim. The authors claim that, in (Rivals et al., 2000), they established an upper bound of the leave-one-out error. Nothing of that kind can be found in that paper, in which, following (Hansen et al., 1996) and (Sorensen et al., 1996), they derived an approximation of the leave-one-out error under the assumption of the validity of a first-order Taylor approximation, in parameter space, of the output of the model. We provided a more rigorous proof in (Monari, 1999) and in (Monari et al., 2000).

Appendix 1 is standard textbook material.

3.2. Reply to section 1.2 of "Jacobian Conditioning Analysis"

The contents of section 1.2 of "Jacobian Conditioning Analysis", entitled "Numerical considerations", is erroneous in two respects:

1. Model selection aims at finding the statistical model that generalizes best, among different candidate models. Poor generalization may be due either to poor training, which is very easy to detect, or to overfitting. Overfitting occurs when the model has too many adjustable parameters in view of the complexity of the problem. Therefore, model selection and model validation aim at detecting, and discarding, models that are likely to exhibit overfitting. It has long been known in statistics that a preliminary screening can be performed by ascertaining that *the jacobian matrix of the model has full rank*. Rivals and Personnaz dismiss that method and state that "the notion of the rank of Z is not usable in order to take the decision to discard a model". They claim that the criterion should be "whether or not the inverse of $Z^T Z$ needed for the estimation of a confidence interval can be estimated accurately". That shift of focus to a completely different issue is a scientific reasoning flaw: the fact that the confidence intervals advocated by the authors cannot be computed "accurately" (see next paragraph for a discussion of that issue) for a given model does not mean that the

model should be discarded: it means that the *confidence interval estimation method* should be discarded. Another confidence interval estimate should be used instead, that does not rely on matrix inversion (e.g. bootstrap methods). Thus, the comments on our approach to model validation are unfounded, and the discussion of model selection presented in (Rivals et al., 2000) is essentially irrelevant. That will be further substantiated by six counter-examples in section 3.4.

2. Rivals and Personnaz state that "a decision [of discarding a model] should be taken on the basis of whether or not the inverse of $Z^T Z$ needed for the estimation of a confidence interval can be estimated accurately". We have just shown that statement to be erroneous; nevertheless, let us discuss that statement from a numerical point of view. Rivals and Personnaz claim that the confidence intervals should be estimated accurately, but they do not explain what "accurately" means. More specifically, they do not ask the relevant question: "how accurately should the confidence interval be estimated in order to get insight into the problem of model validation?". A part of the answer is the following: since the estimation of the confidence interval is derived from a first-order Taylor expansion of the output of the model, in parameter space, around a minimum of the cost function, there is no point in computing the confidence interval, hence the inverse of $Z^T Z$, with a numerical accuracy that is better than the accuracy of the Taylor development. Despite the fact that many authors investigated that issue (see for instance (Bates and Watts, 1998)), Rivals and Personnaz claim that the condition number of the jacobian matrix should be smaller than 10^8, *irrespective of the problem and of the model*[1]. That cannot be true, since (i) the inaccuracy due to the first-order Taylor expansion may be much higher than the numerical accuracy required by that criterion, and since (ii) the accuracy of the Taylor expansion is problem-dependent. Therefore, that "universal" condition-number selection criterion can be expected to lead to discarding perfectly valid models, as shown below with simple counter-examples.

To summarize our reply to the theoretical part of the comments concerning model validation:

[1] The same statement was made recently by the same authors on the same subject in (Rivals et al., 2003b), reporting results obtained in 2000.

1. from a basic point a view, there is a flaw in the scientific reasoning on which those comments are based: instead of discarding models that do not generalize well, the selection method that Rivals and Personnaz advocate discards models for which the authors' favorite confidence interval estimation method fails; actually, the authors should blame their confidence interval estimates for not being accurate, instead of blaming the model for not lending itself to their confidence interval estimation method;
2. from a numerical point of view, their selection criterion does not take into account the accuracy of the approximations on which the estimation of the confidence intervals is based, so that the criterion may be much more stringent than actually required.

The first misconception may lead to accepting models that are invalid; the second may lead to rejecting models that are valid. Both situations will be exemplified below by seven counter-examples.

3.3. Reply to the numerical example

As a further criticism to our article, section 3.2 of "Jacobian Conditioning Analysis", entitled "Neural modeling", considers the simplest example that was presented in our article, namely, the neural modeling of a $\sin x / x$ function with added noise. They consider, just as we did, models with one to four neurons. They discuss that problem in much more detail than we did, and they come to the conclusion that 2-neuron architectures are appropriate, *which is exactly the conclusion stated in our paper*. Therefore, their example can hardly be considered as supporting their criticisms.

Moreover, their numerical example contains errors and inconsistencies:

1. Table 2 of "Jacobian Conditioning Analysis" displays numerical results related to that example. In that table as well as in the rest of the paper, the authors use notations that are different from the notations of the paper they are commenting upon, which makes things confusing to the non-specialist reader. Moreover, they do not provide any definition of the quantities that they use. The quantity dubbed *ALOOS* seems to be the square of the estimated leave-one-out score denoted by E_p in our paper. The quantity called *MSTE* is the square of the training root mean square error *TMSE* of our paper.

The quantity termed *MSPE* is the square of the validation root mean square error denoted by *VMSE* in our paper. Those quantities are defined as:

$$TMSE = \sqrt{\frac{1}{N}\sum_{i=1}^{N} R_i^2}\ ,\ MSTE = TMSE^2,\ E_p = \sqrt{\frac{1}{N}\sum_{i=1}^{N}\left(\frac{R_i}{1-h_{ii}}\right)^2}\ ,$$

$$ALOOS = E_p^2,\ VMSE = \sqrt{\frac{1}{N_V}\sum_{i=1}^{N_V} R_i^2}\ ,\ MSPE = VMSE^2,$$

where R_i is the modeling error on example i of the relevant set, h_{ii} is the leverage of example i, N is the number of examples in the training set, and N_V is the number of examples in the validation set.

The last line of the table reports a *MSTE* of 1.9 10^{-3} and an *ALOOS* of 5.3 10^{-7}. That cannot be true: since $0 < h_{ii} < 1$, each term of the sum in E_p is larger than the corresponding term in *TMSE*, so that one has $E_p > TMSE$, or, equivalently, $ALOOS > MSTE$. Therefore, the result reported for 4 hidden neurons is wrong by at least four orders of magnitude.

2. As mentioned above, the authors of (Rivals et al., 2003) agree with us that a 2-hidden-neuron architecture is the most appropriate for the problem under investigation. Although our paper does not discuss the models with 3 hidden neurons, they claim that our method would have accepted models with three hidden neurons, whereas they dismiss that architecture. That is worth investigating. For architectures with 3 hidden neurons, we performed 100 trainings with different initial values of the parameters, in the conditions that Rivals and Personnaz describe. That generates a relatively small number of significantly different models. Taking advantage of the fact that the "true" regression function f is known, the mean square distance D between the model and the true regression function was computed as:

$$D = \sqrt{\frac{1}{N_D}\sum_{k=1}^{N_D}\left[f\left(x_k\right) - g\left(x_k\right)\right]^2}$$

where N_D = 5,000 (drawn from a uniform distribution), $f(x)$ = sinc $[10(x + 1)/\pi]$, and $g(x)$ is the output of the model. D is the best estimate of the theoretical risk, i.e. of the generalization ability of the model. Table 1 shows the *MSPE*, the distance D, the

condition number *K* of the jacobian matrix, and the number of occurrences of the cost function minimum.

Minimum 1			Minimum 2		
MSPE	D^2	*K*	*MSPE*	D^2	*K*
$2.9\ 10^{-3}$	$1.2\ 10^{-3}$	2 to $7\ 10^{+9}$	$3.8\ 10^{-3}$	$1.6\ 10^{-3}$	$4.0\ 10^{+6}$
16			67		

Minimum 3			Other minima		
MSPE	D^2	*K*	*MSPE*	D^2	*K*
$4.3\ 10^{-3}$	$2.1\ 10^{-3}$	10^{+16} to 10^{+17}	$> 8\ 10^{-3}$	$> 5\ 10^{-3}$	$> 10^{+8}$
12			5		

Table 1

The best model (model with the smallest D, which also has the smallest MSPE) is discarded by the condition-number selection criterion advocated in (Rivals et al., 2003). It is also worth noting that models that correspond to the same minimum of the cost function have widely varying condition numbers, even though the MSPE's agree to four decimal places.

Other examples of similar situations, where the condition-number selection criterion rejects valid models, are exhibited below.

3. The authors of (Rivals et al., 2003) claim that they generated a training set by adding noise to the function sinc $[10(x + 1)]$. From a cursory look at Figures 1 and 2 of their paper, it is clear that such is not the case. It seems that, actually, sinc $[10(x + 1)/\pi]$ was implemented.

To summarize: in order to prove that their selection method is superior to ours, Rivals and Personnaz investigated one of the examples presented in our paper. Their conclusion is exactly the same as ours, so that their example cannot be considered as evidence of the superiority of their approach. Moreover, we showed that their example contains a result that is erroneous by at least four orders of magnitude; in addition, their validation method leads to discarding valid models, for reasons that were explained above, in section 3.2.

3.4. Additional counter-examples

The following counter-examples disprove the claim that the stated limit on the condition number of the jacobian matrix is an appropriate screening criterion, which invalidates the first conclusion in "Jacobian Conditioning Analysis", and a large part of the contents of (Rivals et al., 2000). The counter-examples show additionally that, as we have already demonstrated, the accuracy of the computation of the leverages is not critical for model selection; that invalidates the second point of the conclusion of the comment, as well as another part of (Rivals et al., 2000).

The problem that we address here is inspired from real data: the quantity to be modeled is a thermodynamic parameter (the liquidus temperature) of industrial glasses, as a function of the oxide contents of the latter. The estimation of the liquidus temperature is important for glass manufacturing processes; a detailed description of the application can be found in (Dreyfus et al., 2003). Figure 1 shows the simplest instance of real data on that problem. It is interesting because the singular points actually have a physical meaning, related to phase transitions. That application prompted us to investigate the modeling of data generated from function

$$y = |\sin x| + \cos x \qquad (1)$$

which is plotted on Figure 2.

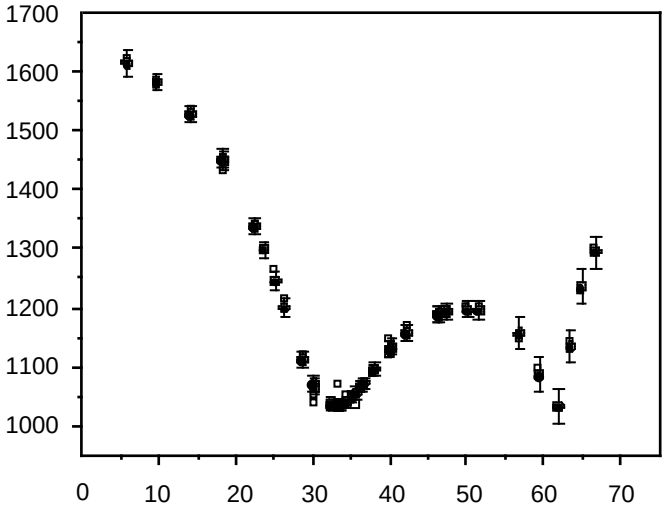

Figure 1
Liquidus temperature vs. lithium oxide concentration.

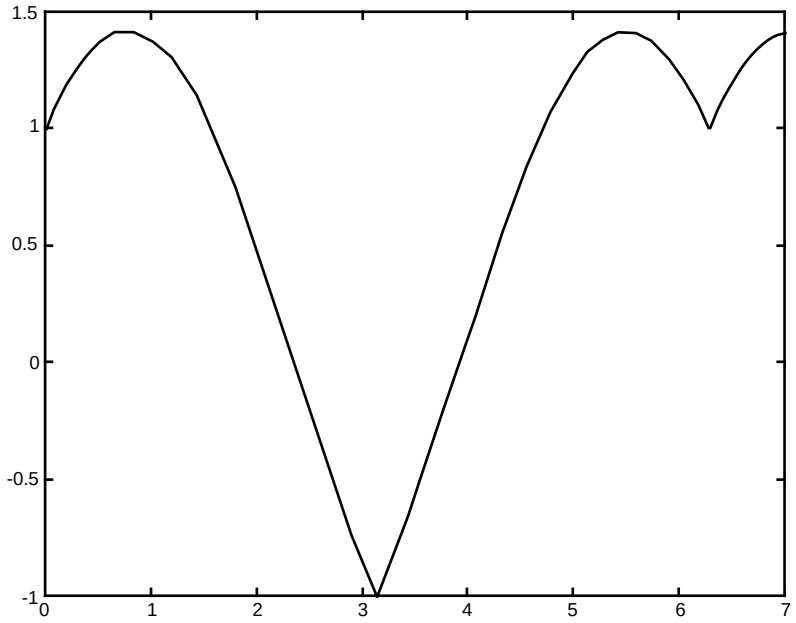

Figure 2
Academic example inspired from Figure 1.

In the present section, we exhibit three pairs of counter-examples: in each pair, one counter-example shows the condition-number selection criterion discarding a valid model, and the other counter-example shows that criterion validating a model that is overfitted.
Note that, in all the following, the values of the leverages computed by our method and by the traditional method advocated in "Jacobian Conditioning Analysis" are in excellent agreement (e.g. agree within nine decimal places for counter-example 1 and eleven decimal places for counter-example 2). Therefore, those examples do not provide any support to the claim that the traditional method is superior to ours in the machine learning context.

In all the following, neural network training was performed by the Levenberg-Marquardt algorithm. For a given number of hidden neurons, 100 trainings were performed with different parameter initializations. 7,000 equally spaced examples were generated as a validation set. Experiments were performed under Matlab on a standard PC. Distance *D* (defined in section 3.3) was also computed from 7,000 equally spaced points.

Following the notations of our paper (Monari et al. 2002), we denote by *TMSE* the root mean square error on the training set (as defined in section 3.3), and by *VMSE* the equivalent quantity computed on the validation set of 7,000 examples (an excellent estimate of the generalization error).

Before discussing the counter-examples, it may be useful to emphasize the main point of (Monari et al., 2002): we claim that overfitting can be efficiently monitored by checking the distribution of the leverages (hence the title of our paper "Local overfitting control via leverages"). The leverages obey the following relations:

$0 \le h_{ii} \le 1 \quad \forall i,$

$$\sum_{i=1}^{N} h_{ii} = q \ .$$

Since the sum of the leverages is equal to the number of parameters q, the leverage of example i can be interpreted as the fraction of the degrees of freedom of the model that is devoted to fitting the model to observation i. Therefore, ideally, all examples should have essentially the same leverages, equal to q/N: if one or several points have leverages close to 1, the model has devoted a large fraction of its degrees of freedom to fitting those points, hence may have fitted the noise on those examples very accurately. In other words, *the more peaked the distribution of the leverages around q/N, the less prone to overfitting the model*. Rivals and Personnaz were unaware of that point in their previous papers, as evidenced by the fact that they never used the word *leverage* prior to "Jacobian Conditioning Analysis".

In order to give a quantitative assessment of the "distance" of the model to a model where all leverages are equal to q/N, we defined (Monari et al., 2002) the parameter $\mu = \frac{1}{N} \sum_{i=1}^{N} \sqrt{\frac{N}{q} h_{ii}}$: the closer μ to 1, the more peaked the distribution of the leverages around its mean q/N; $\mu = 1$ if all leverages are equal to q/N. Hence, the closer μ to 1, the less overfitted the model.

Alternatively, one may use the normalized standard deviation σ_n of the leverages, defined as:

$$\sigma_n = \sqrt{\frac{N}{q\,(N-q)} \sum_{i=1}^{N} \left(h_{ii} - \frac{q}{N} \right)^2} \ .$$

$\sigma_n = 0$ if all leverages are equal to q/N, and $\sigma_n = 1$ in the worst case of overfitting, where q leverages are equal to 1 and $(N-q)$ leverages are equal to zero. Hence, the smaller σ_n, the less prone to overfitting the model. Both quantities are computed in the following counter-examples.

3.4.1. Counter-examples 1 and 2

In a first set of experiments, 35 equally spaced points were generated from relation (1) to serve as a training set. Uniform noise was added, with standard deviation 0.1

Figure 3 shows the generating function, the training data and the output of a model with 5 hidden neurons, together with the values of the leverages. The relevant figures for that model are summarized in the first row of Table 2.

	TMSE	*VMSE*	μ	σ_n	Distance *D*	*K*	Leverages > 0.95
Counter-example 1	0.078	0.117	0.984	0.38	0.062	$1.5\ 10^9$	3
Counter-example 2	0.066	0.123	0.979	0.56	0.072	$2.9\ 10^6$	11

Table 2
TMSE, VMSE, μ, σ_n and D as defined in the text; K: condition number of the jacobian matrix; last column: number of leverages that are larger than 0.95

Its condition number exceeds the limit stated in "Jacobian Conditioning Analysis" by more than one order of magnitude, so that it should be discarded according to those comments. Actually it is a valid model: its generalization error is small since its *VMSE*, computed from 7,000 examples, is slightly larger than the standard deviation of the noise, and the distance *D* is even smaller. μ is close to 1 and σ_n is far from 1. Finally, the leverage values computed by the traditional method and by ours (relation A.4 of (Monari et al., 2002)) are in excellent agreement: the root mean square of the differences between the leverages computed by our method and by the traditional method is on the order of $3.7\ 10^{-10}$.

The three points with high leverages are located at the boundaries of the input range, as expected. The W-shape of the leverage graph indicates that the points located in the vicinity of the main minimum are influential, as expected.

Thus, counter-example 1 is an example of the condition-number selection criterion stated in "Jacobian Conditioning Analysis" discarding a valid model.

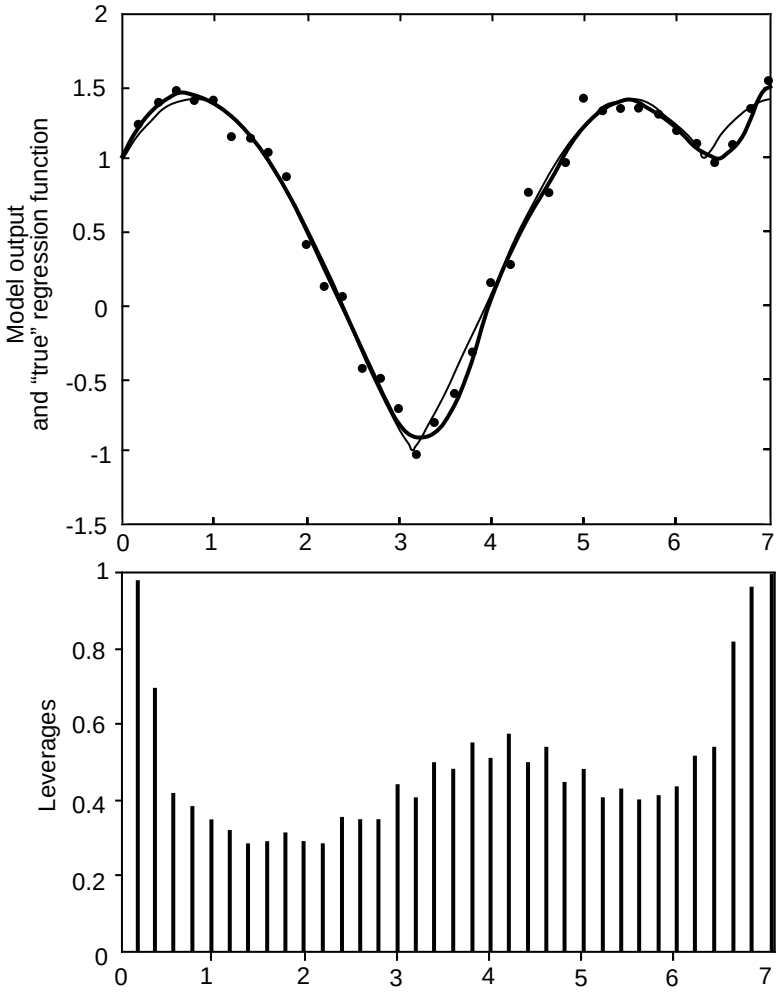

Figure 3
Counter-example 1: the condition-number selection criterion discards a valid model.

Counter-example 2 (Figure 4) is a model with 7 hidden neurons, trained in the same conditions and with the same data as counter-example 1. The characteristics of the model are shown on the second row of Table 2. The root mean square of the differences between the leverages computed by our method and by the traditional method advocated by Rivals and Personnaz is on the order of $1.7\ 10^{-12}$.

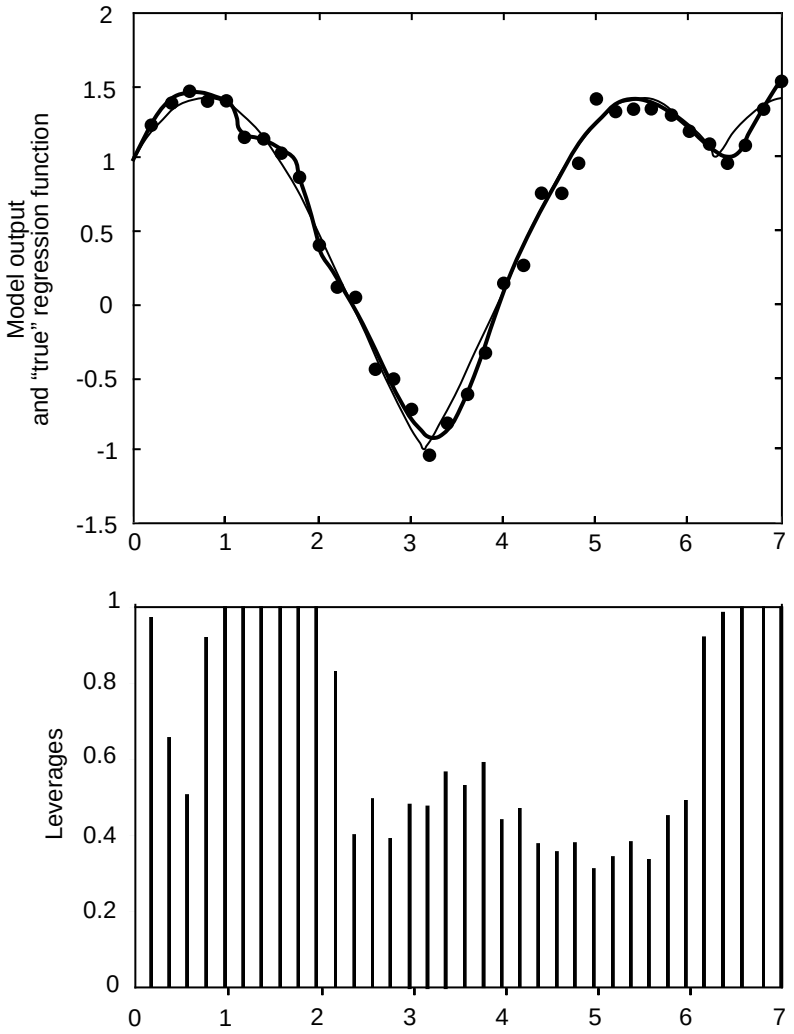

Figure 4
Counter-example 2: the condition-number selection criterion fails to detect overfitting.

The condition number of the jacobian matrix of the model is way below the limit stated by Rivals and Personnaz. Hence, according to their comment, the model should be accepted. However, it is clear, from Figure 4, that the model is strongly overfitted. That is further substantiated by three facts

1. the validation error *VMSE* is twice as large as the training error *TMSE*,
2. 11 leverages (almost 1 out of 3) are larger than 0.95, and that 13 of them are larger than 0.90. The high leverages are located between $x = 1$ and $x = 2$, where overfitting is clearly apparent on Figure 4, and also at the boundaries of the input range, as usual;
3. μ is substantially smaller than 1, or, equivalently, the standard deviation of the leverages σ_n is larger than that of counter-example 1.

The selection method that Rivals and Personnaz advocate nevertheless fails to detect that gross overfitting. As usual, computing the leverages by the traditional method and by relation A.4 of our paper does not make any significant difference.

The difference between the two models is also clear from Figure 5, which shows the histograms of the leverages for the model that is discarded by the condition-number selection criterion (top figure) and for the model that is accepted by that criterion (bottom figure). As explained above, the distribution of the leverages should be as peaked as possible around q/N. Clearly, the leverage distribution for the model accepted by the suggested condition-number selection criterion is extremely far from complying with that condition, having a very large number of leverages that are close to 1.

To summarize, counter-example 2 shows an example of the condition-number selection criterion failing to discard an overfitted model.

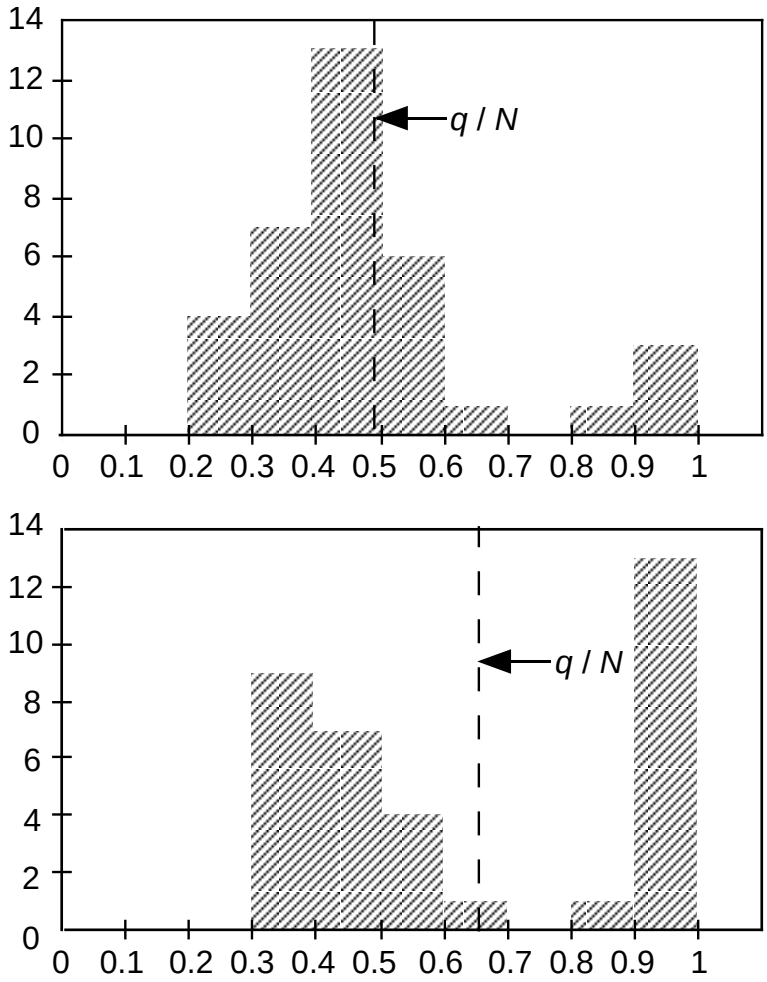

Figure 5
Leverage distributions of counter-example 1 (top) and counter-example 2 (bottom).

3.4.2. Counter-examples 3 and 4

The second set of experiments was performed as follows. 100 neural networks were first trained with a large number of points (350) generated by the regression function (1), *without added noise*. Then the values of the parameters of those models were used as initial values for a re-training with only 35 points with added noise. Since the latter training starts with initial parameters of excellent models, it may be expected that the resulting models are the best models that one can hope to obtain, given the limited number of points and the noise in the training set.

Figure 6 shows a good model with 5 hidden neurons, whose characteristics are reported in the first row of Table 3. Despite its performance, the model is rejected by the condition-number selection criterion, since its condition number is larger, by one order of magnitude, than the rejection limit specified its authors.

	TMSE	*VMSE*	μ	σ_n	Distance D	K	Leverages > 0.95
Counter-example 3	0.078	0.12	0.99	0.36	0.063	10^9	2
Counter-example 4	0.068	0.13	0.95	0.62	0.079	$1.7\ 10^7$	7

Table 3
TMSE, VMSE, μ, σ_n and D as defined in the text; K: condition number of the jacobian matrix; last column: number of leverages that are larger than 0.95

Figure 7 shows the behavior of a model that also has five hidden neurons; its characteristics are summarized in the second row of Table 3. That model is much worse than counter-example 3: its VMSE is higher while its TMSE is smaller (a clear sign of overfitting), the distance between the model and the regression function is higher, and 20% of the leverages are larger than 0.95. μ is substantially smaller than for counter-example 3, and σ_n is almost twice as large as that of counter-example 3. Nevertheless, the condition number is one order of magnitude below the rejection limit, so that the condition-number selection criterion fails to detect that gross overfitting. Furthermore, the leverages computed by the traditional method, and by our method, agree within 10^{-11}.

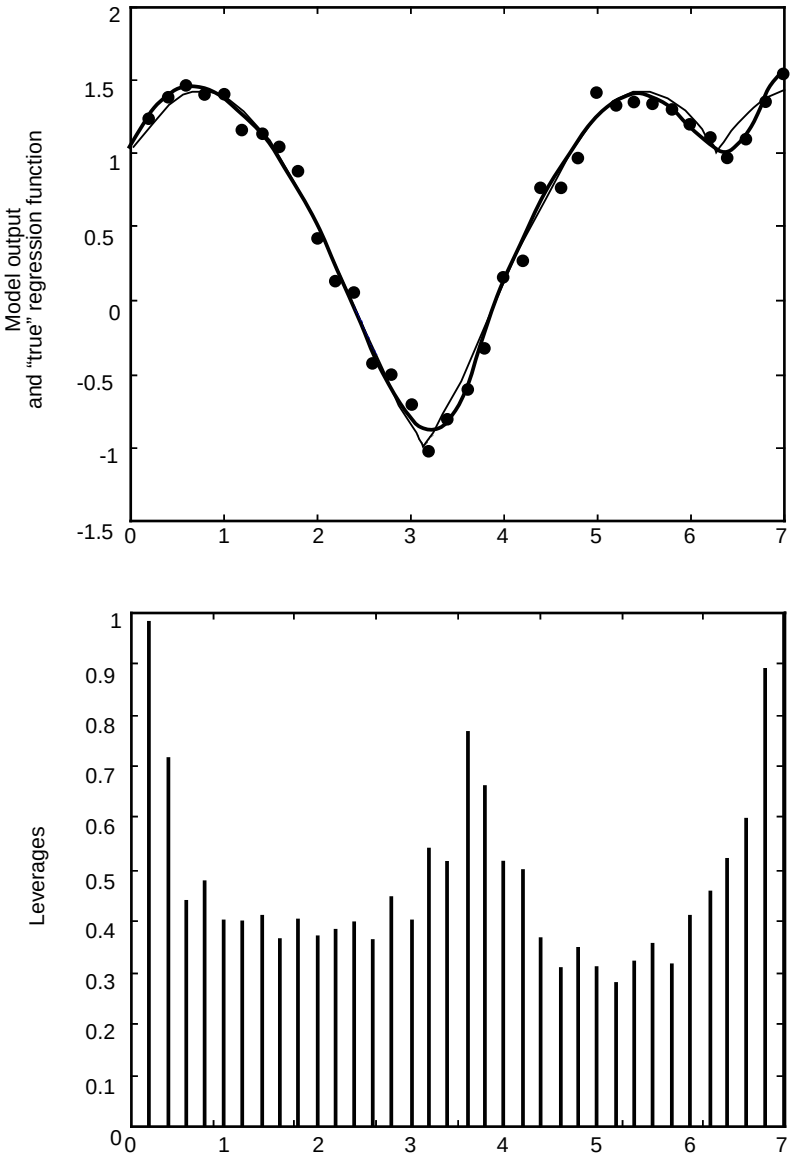

Figure 6
Counter-example 3: the condition-number selection criterion discards a valid model.

3.4.3. Counter-examples 5 and 6

In a final set of numerical experiments, the training set was constructed with a larger number of examples in the vicinity of the singular points, the total number of points being kept constant, equal to 35. Training was performed with random parameter initialization, as in counter-examples 1 and 2.

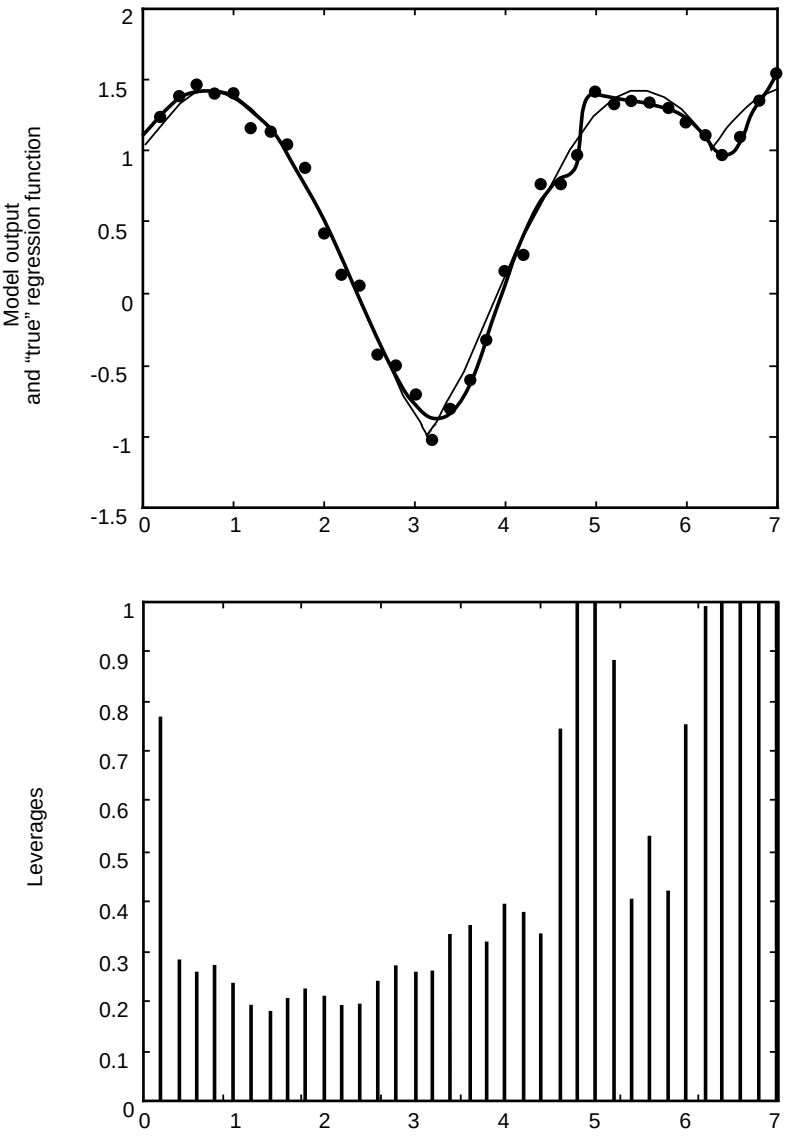

Figure 7
Counter-example 4: the condition-number selection criterion fails to discard an overfitted model.

Figure 8 shows the behavior of a model with 5 hidden neurons, whose characteristics are summarized in Table 4. As expected, no overfitting occurs in the vicinity of the minima of the function, but, since the total number of points was kept constant, leverages become higher between the minima. Nevertheless, this is a very reasonable model given the training data. It is discarded by the condition-number selection criterion, since its condition number is larger than the rejection limit by six orders of magnitude.

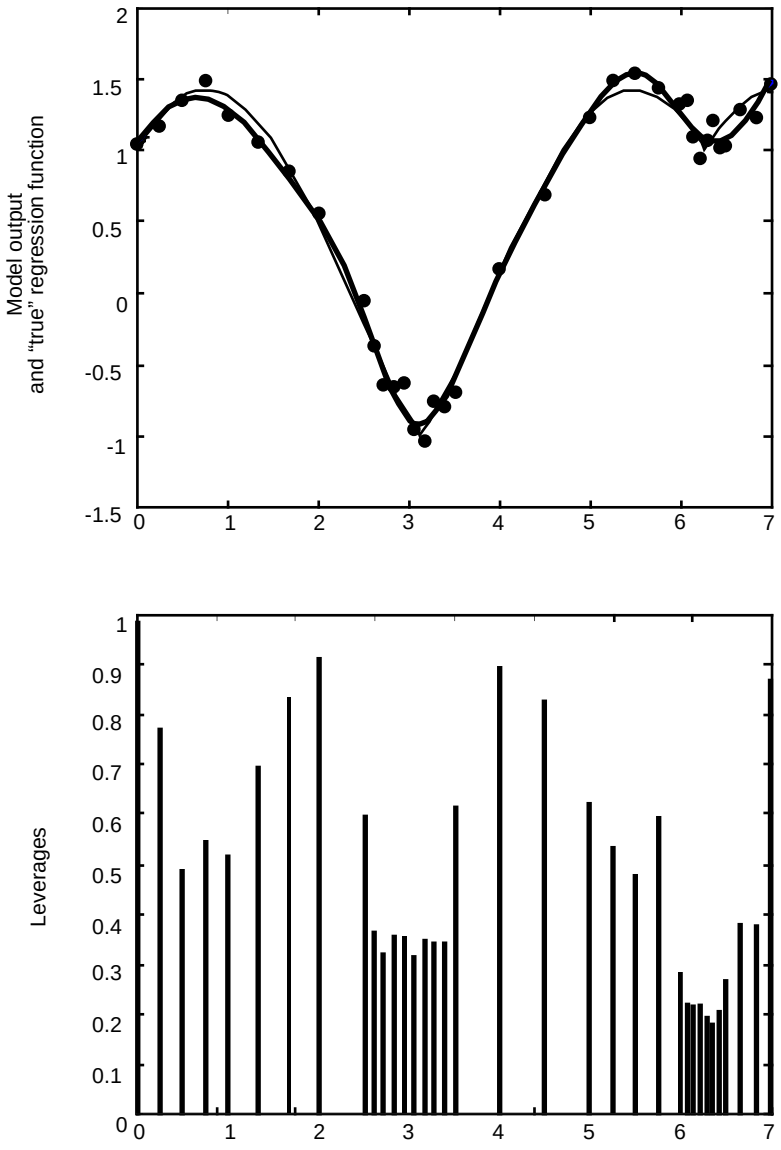

Figure 8
Counter-example 5: the condition-number selection criterion discards a valid model.

	TMSE	*VMSE*	μ	σ_n	Distance D	K	Leverages > 0.95
Counter-example 5	0.082	0.12	0.97	0.47	0.069	$4.2\ 10^{14}$	1
Counter-example 6	0.068	0.14	0.95	0.62	0.091	$1.3\ 10^{5}$	6

Table 4
TMSE, VMSE, μ, σ_n and D as defined in the text; K: condition number of the jacobian matrix; last column: number of leverages that are larger than 0.95

By contrast, Figure 9 shows a model with 5 hidden neurons, whose characteristics are summarized in the second row of Table 4. This is again an overfitted model, whose distance

to the regression function, and *VMSE*, are much poorer than those of counter-example 5; nevertheless, it is accepted by the condition-number selection criterion.

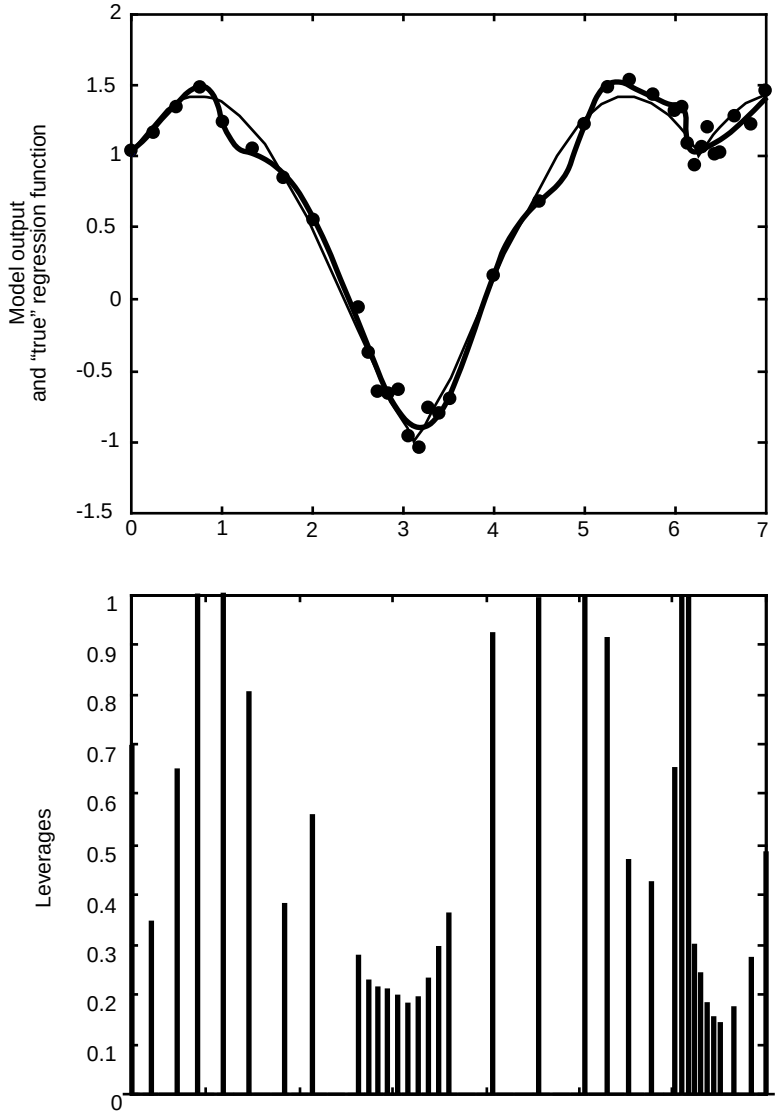

Figure 9
Counter-example 6: the condition-number selection criterion fails to discard an overfitted model.

3.5. The relevance of the condition-number selection criterion to overfitting

In the previous section, we showed several examples of the condition-number selection criterion accepting overfitted models or discarding valid models. The above counter-examples are just a selection among many more similar counter-examples, so that it is natural to wonder how frequently such situations will occur. More specifically, one can ask the following question: what is the probability that the parameters of a network with one input and five hidden neurons can be estimated reliably from 35 equally spaced points (counter-examples 1, 2, 3 and 4)? In order to gain some insight into that question, the following numerical

experiment was performed: 10,000 different neural networks with one input and five hidden neurons were generated, with random parameters, uniformly distributed with variance 10. For each model, the leverages of 35 equally spaced points between –1 and +1, and the normalized standard deviation σ_n of their distribution, were computed. The condition number K of the jacobian matrix was also computed.

Figure 10 shows σ_n as a function of K: each network is shown as a dot; dots lying on the x-axis are networks whose jacobian matrix is rank deficient. For models whose jacobian matrix has full rank, no trend can be found in that graph: thus, the condition number has essentially nothing to do with the distribution of the leverages, hence is essentially irrelevant to overfitting. Any model located to the right of the vertical line would be discarded by the condition-number selection criterion, despite the fact that some of them have excellent leverage distributions, hence are functions whose parameters can legitimately be estimated from data pertaining to 35 equally spaced points. Actually, the "best" network (network with the most peaked leverage distribution, i.e. with smallest σ_n) is discarded, whereas several poor networks (with large values of σ_n), including the network with the largest σ_n, are accepted. Similarly, no clear trend can be found in the graph of σ_n *vs*. K when data is more abundant.

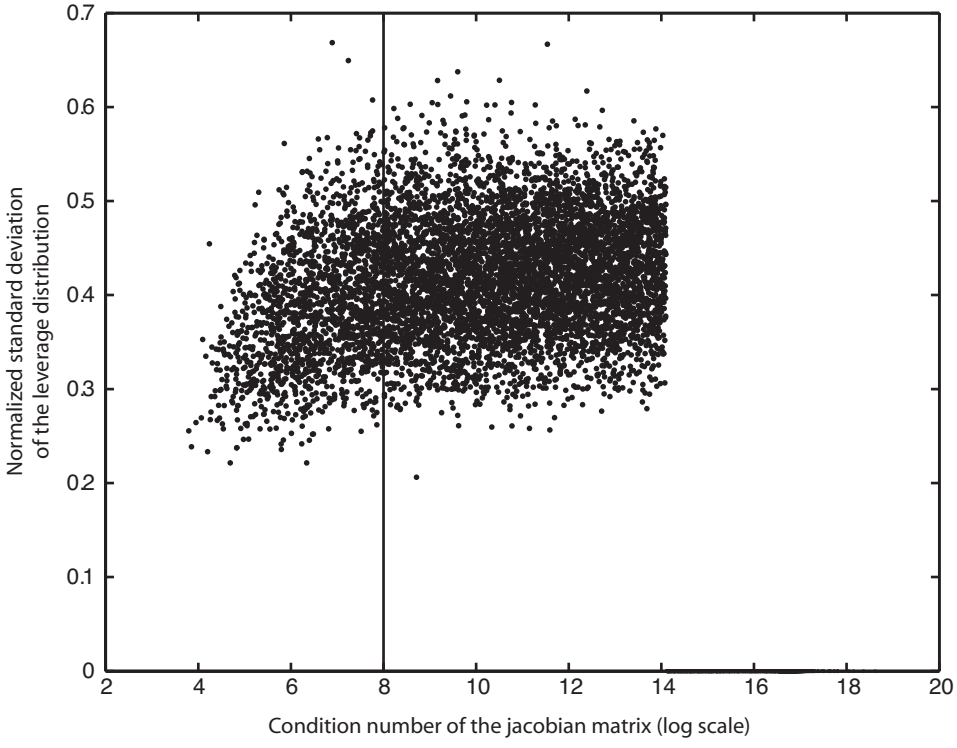

Figure 10
Normalized standard deviation of the distribution of the leverages of 35 equally spaced points, and jacobian matrix condition number, for 1,000 neural networks with five hidden neurons.

4. CONCLUSION

"Jacobian Conditioning Analysis for Model Validation" is a comment on our paper (Monari et al. 2002). In the present reply, we disproved all comments made in “Jacobian Conditioning Analysis”.

The first conclusion of Rivals and Personnaz states that the condition number of the jacobian matrix of the model should be used as a criterion for model validation: a model with a condition number larger than 10^{+8} should be discarded; the same statement was made in (Rivals et al., 2000). We proved that statement wrong in two respects:

1. the models that are discarded by that criterion are models for which a particular type of confidence intervals, obtained by a specific estimation method, cannot be computed accurately; this does not mean that the *model* should be discarded: it means that the *confidence interval estimation method* should be discarded; that misconception may lead to discarding valid models.
2. Rivals and Personnaz’s comment on the accuracy required to compute the confidence intervals; those comments are erroneous, because the authors overlook the fact that the confidence intervals stem from a first-order Taylor expansion of the model output: therefore, there is no point in computing the confidence intervals with an accuracy that is better than the accuracy of that first-order approximation. Therefore, the accuracy requested for the computation of the confidence interval is completely problem-dependent: the "universal" criterion exhibited by the authors cannot be valid.

In addition to disproving, on theoretical grounds, the comments made by Rivals and Personnaz, we presented six counter-examples: three of them are instances of the authors’ selection criterion discarding valid models; the other three are instances of the authors’ criterion accepting overfitted models. In short, the condition-number selection criterion, advocated in the first conclusion of Rivals and Personnaz is at best useless, and very frequently leads to making wrong decisions. The claim that we "followed" that approach is unsupported.

In the second paragraph of their conclusion, Rivals and Personnaz state that the numerical method for computing the leverages, which was indicated in an appendix of our paper (Monari et al. 2002), is not accurate enough. We proved that their statement is unsupported. They claim that their method can reach an accuracy of 10^{-16}; however, they do not provide any example, in the context of machine learning, where such accuracy is required, i.e. where the

"noise" on the estimation of the leverages is on the order of 10^{-16}. Conversely, we provided examples where the difference between the leverages of two models that are obtained at two successive iterations after convergence of the training algorithm, i.e. between two models that are considered "identical", exceeds 10^{-16} by several orders of magnitude. Therefore, the accuracy of the method advocated by Rivals and Personnaz is irrelevant in such situations, and the latter did not provide any example where it might be relevant. In addition to disproving their point theoretically, we have shown that, even in the very example that Rivals and Personnaz proffered, the differences between the leverages computed by the traditional method and those computed by our method are negligibly small.

In order to substantiate their claims, Rivals and Personnaz study the numerical example that we investigated in (Monari et al., 2002). They come exactly to the conclusion that was reached in our paper, so that their example cannot be considered a convincing counter-example. However, they reach that conclusion by faulty reasoning and computing. We pointed to inconsistencies in their presentation, and we provided a proof that one of their numerical results is wrong by at least four orders of magnitude.

The third paragraph of the conclusion of "Jacobian Conditioning Analysis" is: "For candidates whose condition number is small enough, and for which the leverages have been computed as accurately as possible according to (12), one may check *additionally* if none of the leverage values is close to one, as already proposed in (Rivals and Personnaz, 1998)". That statement is not acceptable for three reasons:

1. we showed that Rivals and Personnaz provide no evidence that using (12) is necessary, or that our method for the computation of the leverages is inappropriate for model validation;
2. we showed that models with a "small enough" condition number, i.e. a condition number below the limit stated by Rivals and Personnaz, may have several leverages close to 1, and, hence, exhibit strong overfitting (counter-examples 2, 4 and 6), and that, conversely, models with high condition numbers may have reasonable leverages, hence be acceptable (counter-examples 1, 3 and 5);
3. Rivals and Personnaz did *not* state in (Rivals et al., 1998) that leverages should be checked for values close to 1. They made a different suggestion: checking that the sum of the leverages is equal to the number of parameters, and that all leverages are smaller than 1. Actually, they did not realize the significance of leverages close to 1 before reading our paper (Monari et al., 2002), as evidenced

by the fact that the very word *leverage* is used neither in (Rivals et al., 1998), nor in (Rivals et al., 2000).

By contrast, the last sentence of the conclusion of "Jacobian Conditioning Analysis" ("Leverage values close to, but not necessarily larger than one are indeed the symptom of overfitted examples, or of isolated examples at the border of the input domain delimited by the training set") is unquestionable: it is actually the very central idea of our work (Monari et al., 2002).

To summarize, in this article, we disprove all comments of Rivals and Personnaz in "Jacobian Conditioning Analysis" on our previous article (Monari et al., 2002)[2]. Additionally, our replies invalidate a substantial part of the contents of other papers by the same authors on the same subject.

References

Bates, D.M., Watts D.G. (1998) *Nonlinear Regression Analysis and its Applications*. John Wiley and Sons.

Dreyfus, C., Dreyfus, G. (2003). A Machine Learning Approach to the Estimation of the Liquidus Temperature of Glassforming Oxide Blends. *Journal of Non-Crystalline Solids* 318, 63-78.

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

Larsen, J., Hansen, L.K. (2001). Comments for: Rivals I., Personnaz L., Construction of confidence intervals for neural networks based on least squares estimation. *Neural Networks* 15, 141-142.

[2] Rivals and Personnaz add, *after their conclusion*, a section 5 entitled "Other comment for (Monari and Dreyfus, 2002)". In view of the fact that all the other claims of Rivals and Personnaz were disproved, we will not discuss that ultimate comment.

Monari, G. (1999). *Sélection de modèles non linéaires par leave-one-out : étude théorique et application des réseaux de neurones au procédé de soudage par points*. Thèse de l'Université Pierre et Marie Curie, Paris.

Monari, G., Dreyfus, G. (2000). Withdrawing an Example from the Training Set: an Analytic Estimation of it Effect on a Nonlinear Parameterized Model. *Neurocomputing* 35, 195-201

Monari, G., Dreyfus, G. (2002). Local Overfitting Control via Leverages. *Neural Computation* 14, 1481-1506.

Rivals, I., Personnaz, L. (1998). Construction of Confidence Intervals in Neural Modeling Using a Linear Taylor Expansion. *Proceedings of the International Workshop on Advanced Black-Box Techniques for Nonlinear Modeling*.

Rivals, I., Personnaz, L. (2000). Construction of Confidence Intervals for Neural Networks Based on Least Squares Estimation. *Neural Networks* 13, 463-484.

Rivals, I., Personnaz, L. (2003). Jacobian Conditioning Analysis for Model Validation. *Neural Computation,* this issue.

Rivals, I., Personnaz, L. (2003b). MLPs (Mono-Layer Polynomials & Multi-Layer Perceptrons) for nonlinear modeling. Journal of Machine Learning Research **3**: 1383-1398.

Seber, G.A.F., and Wild, C.J. (1989). *Nonlinear regression*. Wiley Series in Probability and Mathematical Statistics. New York, New York: John Wiley & Sons.

Sorensen, P.H., Norgard, M., Hansen, L.K. and Larsen, J. (1996). Cross-Validation with LULOO. *Proceedings of the International Conference on Neural Information Processing - ICONIP '96*.

Tibshirani, R.J. (1996). A Comparison of Some Error Estimates for Neural Models. *Neural Computation* 8, 152-163.