

# Continuous monitoring of dynamic systems: Application to monitoring of dialyzed patients.

Laurent JEANPIERRE, and François CHARPILLET.

**Abstract--** We detail in this article a diagnosing architecture for critical dynamic systems where the interaction with a human expert is necessary. We addressed this problem by enforcing strong semantics into a Hidden Markov Model so that its results and parameters may be interpreted directly by the expert. In this situation, adapting the model to the evolution of the modeled system is challenging because of the lack of data. We developed a specific method which allows the expert to direct the learning algorithm by locally correcting the computed diagnosis. Using a gradient method for optimizing a compromise between the expert's corrections and statistical criteria, the model can be adapted efficiently. This architecture has been applied to the remote monitoring of dialyzed patients for 3 years with a great evaluation by the dialysis center's physicians.

**Index Terms--** Adaptive systems, Biomedical signal analysis, Gradient methods, Hidden Markov models.

## I. INTRODUCTION

THE automated monitoring of critical systems is becoming more and more necessary as the number of human experts becomes insufficient in some domains. This is particularly true in the medical monitoring of chronic diseases, where patients live longer and diseases are diagnosed earlier. This increase in healthcare demand does not match with the actual number of physicians.

In such a situation, the automated monitoring can be very useful, monitoring several patients simultaneously while drawing the physicians' attention to the patients who need some treatment. However, this is not limited to medical applications, and several domains can benefit from this approach. Therefore, we will consider in this paper a generic dynamic system (*the system* in the following).

Such problems have been considered in the past, aiming at the proper detection of specific conditions in a controlled environment like Intensive Care Units [1]. Several models have been designed to translate the experts' knowledge into a rule-based expert system [2] or to match specific patterns in the clinical data [3].

However, the first approach is rather limited due to the increasing number of rules necessary for implementing such knowledge, and because of the poor results obtained in noisy environments. The second approach focuses on the data, modeling them very precisely and handling the noise properly. However, this kind of model usually requires very complex models which can only be understood by their makers.

We chose an alternative approach, focusing on the data with an expert point of view. By implementing a stochastic model anchored in the expert's domain, we were able to obtain a diagnosis architecture which is able to deal with noise and uncertainty while allowing the expert to intuitively understand the reason for each parameter.

However, regarding the long term monitoring of specific systems, the adaptation of the model must be considered. This is particularly true with medicine, where patients evolve when they age, and where pathologies evolve with time. In such conditions, where the model must be adapted quickly, we require the algorithm for learning new parameters from few data while conserving the semantics of the model.

Unsupervised learning cannot be applied to such situations because of the lack of a proper learning corpus and the necessity for keeping the model semantics. Supervised learning is usually difficult to use, requiring the expert to provide lots and lots of information. Therefore, we designed a semi-supervised learning method that uses partial directives from the expert and statistical criteria for adapting the model parameters. The expert's indications allow for the semantics to be kept during the process, while the numerical criteria allow for reaching a durable model. Additionally, the synergy of these two aspects makes the architecture tolerant with respect to mistakes from the expert.

In a first part, we will describe the expert-compatible stochastic model. Next, we will detail the learning algorithm. Finally, results from the clinical study of the automatic monitoring of dialyzed patients will be presented, illustrating the successful collaboration of the computer with physicians.

## II. A FUZZY MARKOV MODEL

### A. Hidden Markov models

We chose hidden Markov models (HMM) for modeling the dynamic system because they are able to deal with noise and uncertainty very efficiently. Moreover, their discrete versions have been studied for several years, and very efficient algorithms are available.

Manuscript received October 30, 2004.

L. Jeanpierre was with the University Henry Poincaré, Nancy 1, FRANCE. He is now with the University Nancy 2, FRANCE (telephone: 33 3 83593025, e-mail: Laurent.jeanpierre1@free.fr).

F. Charpillet is with INRIA Lorraine, FRANCE (telephone: 33 383592085, e-mail: charp@loria.fr).

As opposed to their classical use in speech recognition, for example, we chose to use a single Markov model for modeling the whole system. This allows us for designing an efficient model which is able to deal with the dynamics of all the possible diagnoses, but also with the dynamics of the evolution from one condition to another. Additionally, it allows for the easy diagnosis of hybrid conditions.

To allow the expert for understanding the model, we based its states on the situations to diagnose. Since we consider long term monitoring situations for detecting anomalies, there exists a central condition where everything is right. This will be the *Normal* state. From this point, possible anomalies are modeled by one or more additional states, depending on possible evolutions.

Therefore, the probability indicated by the model for a state can be interpreted directly as the probability of some anomaly to occur. Similarly, the probability for evolving from one state to another can be interpreted as the evolution from one anomaly to another.

### B. Fuzzy observations

Each state of a HMM must be associated with an effect on the observation of the system. This is direct knowledge, which is very easy to obtain, compared with a diagnosis procedure. Information like "infections generally imply fever" can be translated as a strong probability for observing *fever* in the state *infection*.

However, most of the observations are continuous in almost all the real situations. The trouble is that these continuous signals must be mapped to some probability for each state. The 2 common approaches are not really well suited to our situation: using discrete intervals to model a continuous signal implies some threshold effect when the observed value is near the boundary of 2 intervals. To avoid generating too much noise while discretizing the signal, the intervals must be small enough for the model to be able to cope with this threshold effect. However, this implies generating an observation probability for each interval of each state. Therefore, the number of parameters quickly grows out of control, and the influence of one parameter on the final diagnosis is not obvious anymore.

The second approach consists of using a weighted sum of normal probability distributions, which allows for generating any bounded continuous function. However, even if this solves the above threshold effect, the number of parameters necessary for generating each probability function is really big: Each normal density function requires two parameters (its mean and its variance) for modeling a single "bump", even if the signals are supposed independent. Generating other kinds of functions requires adding several of them. Finally, the exact influence of a specific parameter on the final diagnosis is almost impossible to describe easily.

This is why we chose to express our model's observation probabilities through fuzzy intervals, as recommended in [4]. Since we are bound to diagnosing anomalies, most of the signals have a set of values that should be observed normally.

Therefore, we defined three fuzzy values representing the signal being *weak*, *normal*, or *strong*. The fuzzy nature of these sets rubs out the threshold effect, because we do not observe one symbol at a time. Instead, the model receive a confidence form each signal for being weak, normal or strong. Finally, the semantics of a given signal being *strong* in a given state is obvious. These parameters can be interpreted with no difficulty.

Depending on the signals' configuration with respect to a specific problem, the probability of observing a given signal may be considered as independent from the other signals or not. In the first situation, as shown in [5], we can merge all the confidences through the formula

$$P(O_t|d) = \prod_{vs_i \in \text{signals}} \sum_{\sigma \in V(vs_i)} P(vs_{t,i} = \sigma) P(vs_i = \sigma | d) \quad (1)$$

where the influence of diagnosis  $d$  on the observation  $O$  measured at time  $t$  depends on each signal  $vs$ , which is seen as a fuzzy set of symbols  $V(vs)$ .

When the signals are not independent, we have to consider a covariance matrix, which expresses the probability for observing joint signals. This makes the parameters set grow dramatically and we suggest avoiding this situation. An alternative method would be to generate several states, expressing the various signals dependencies. For example, if two signals must be similar for a given diagnosis, we can generate three states, both signals being respectively weak, normal, or strong.

Fig. 1 shows the observation of the Diatelic [6] system. In this case, the computer has to monitor a population of home dialyzed patients who send their medical signals through Internet daily. The model has to detect hydration and ideal weight evolutions, so that physicians may be alerted early enough to avoid hospitalization. The HMM is based on 5 states and 4 signals modeled through 3 fuzzy values each. Fig. 1 shows that interpreting the probabilities is straightforward: each curve plots the probability for observing the row's signal in the column state.

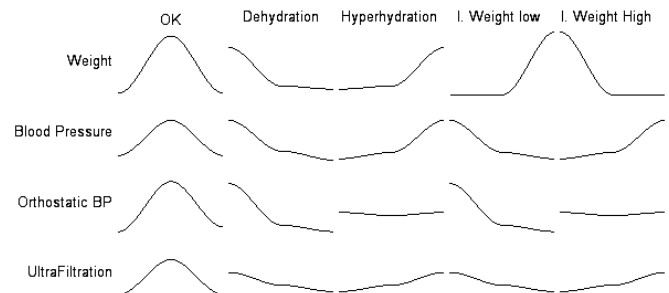


Fig. 1. Graphical representation of the DIATELIC observation function; each column is associated with a specific state/diagnosis while each row regards a medical signal. Plots are the sum of 3 components (*weak*, *normal*, and *strong*) weighted by the probability for observing each symbol in each state. The higher the plot, the more probable the observation is, in the associated state.

### III. ADAPTING THE MODEL

#### A. Learning conditions

Considering the continuous monitoring of a critical system, the adaptation period should be as short as possible. It is therefore impossible to gather sufficient data for building a state-of-the-art corpus. This means that the model will be learnt from a small corpus, which is certainly biased.

When the model is to ensure that the monitored system is safe all the time, in particular, recently gathered data will concern mainly the *normal* state. It is clearly improbable that all the possible diagnoses occur in a short period of time.

Therefore, Expectation-Maximization algorithms [7] like Baum & Welch are not applicable. They would overfit the data and spoil the model semantics.

To prevent the model semantics from drifting, we chose to optimize the parameters with respect to the computed diagnosis rather than the observation probability. The trouble is that the real state of the system is unknown. Therefore we have to rely on the Expert's diagnosis to bias the process: the algorithm has to learn from the Expert's directives.

To simplify the operator's task, the computer first displays its current diagnosis. Since this diagnosis was relatively good until recently, the correct diagnosis is expected to be relatively close from it. From there, one can move probability curves to their correct position, providing the algorithm with directions.

From a computer's point of view, the corrected diagnosis is nothing more than an array  $c_t(d)$  specifying the probability the state  $d$  should have at time  $t$ . Therefore, the main objective should be to minimize the differences between the computed diagnosis and the corrected one.

#### B. Algorithm choice

When the corrected diagnosis is known, adapting the model parameters simply reduces to a minimization problem. Thus, a simple gradient descent algorithm has been chosen to minimize the error function (2).

$$\sum_{d \in \text{states}} \sum_{t=1}^T (b_t(d) - c_t(d))^2 \quad (2)$$

with  $b_t(d)$  being the computed probability the diagnosis  $d$  has at time  $t$ . Obviously, the array  $b_t(s)$  is computed by the Forward procedure [8] whose core formula of this procedure is

$$b_t(s) = \sum_{s' \in \text{States}} b_{t-1}(s') \cdot P(Q_t = s \mid Q_{t-1} = s') \cdot P(O_t \mid Q_t = s) \quad (3)$$

with  $O_t$  the observation and  $Q_t$  the real state at time  $t$ . Obviously, these equations are differentiable, but computing all their derivatives is clearly a long task since (3) is recursive. For this reason, a derivative-free method has been chosen for

optimizing it. More precisely, we implemented a bracketing minimization, along with a relaxation method based on [9].

In the remaining of this section, the bracketing search in one-dimension and the relaxation we propose will be explained before considering some possible generalizations.

#### C. Bracketing minimization

The main idea for Bracketing Optimization in one dimension is to progressively reduce the search interval around the optimum. Therefore, it is necessary to obtain first an interval containing the minimum. Since all the parameters are probabilities, this is trivial.

The real problem is to reduce this interval while keeping the minimum in it. To achieve this goal, we need a third point between the terminals. There are two main situations depending on this new point being better than the others or not. When the best point is one of the interval terminals, the problem is easy. Supposing the function is smooth enough, its minimum will be between the best point and the middle one: the search interval is halved.

When the middle point is the best one, the problem is harder: two more points are needed to figure out on which side the minimum is. Having five points, the best one has two neighbors since no terminal is the best (this was the above situation). Supposing the function is smooth enough, the minimum should be somewhere between these two neighbors.

This algorithm is simple but efficient. It has almost no condition on the error function (it must be continuous), and is easy to implement. However, there is no guaranty it will find the global optimum, like all the local algorithms. Overcoming this limitation will be considered in the last subsection.

#### D. Relaxation to multidimensional functions

The principle of relaxation methods is that an optimal point in a multidimensional space has to be optimal in any possible projection of it. Therefore, the optimal point is not reached as long as there are non-optimal projections.

The naive relaxation consists in considering each dimension in turn: all the parameters but one are fixed to their current value. When all the parameters have been optimized once, it is necessary to consider them again: actually, each parameter is optimized with respect to a given configuration of the other parameters. Later, these parameters are optimized, their values change, and previously optimized parameter may no longer be optimal. These cycles only stop when no significant enhancement is achieved for any parameter. Then, all these projections are locally optimal: the objective is reached.

This method works very well; however, it is generally inefficient due to a stairs-like behavior [9]: each step slightly moves the optimum point in some direction, affecting the others' optima. Since this decreases as cycles accumulate (because each step gradually enhances the solution), the algorithm eventually converges. Such a situation is represented on Fig. 2.

To overcome this behavior, Powell proposed replacing one of the parameters by a combination of the previous cycle's

optimization. This allows for short-cutting efficiently those stairs but it also has some drawbacks. First, since one parameter is replaced by some combination at the end of each optimization cycle, there exists a risk that this new vector is almost collinear with another one. When this occurs, one of the problem dimensions disappears, and a whole family of potential solutions is lost. The second drawback is related to constrained parameters. Since a new parameter is generated, its constraints must be defined. This is why Powell has considered only unconstrained variables.

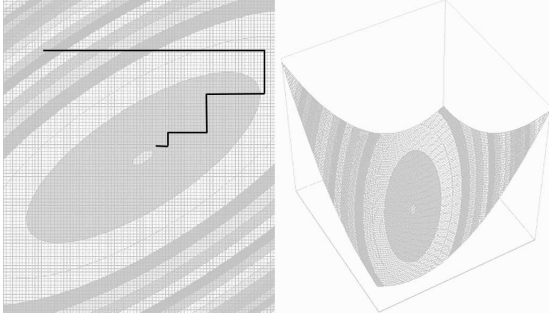


Fig. 2. Illustration of the stairs-like gradient descent: the process is clearly inefficient. Left: 2D view with the optimization superimposed. Right: 3D view of the same function.

To solve the first problem, [10] proposed to reset the optimization vectors after a while. We chose an alternative solution: a new temporary parameter is created instead of replacing one. This way, the optimization can be done along this vector, but there is no possibility of losing solutions. Its drawback is that these optimized directions cannot be combined into a better one on later cycles. Therefore, it may converge slower than Powell's method.

The second enhancement is related to constrained parameters. Considering our diagnosis model, which is stochastic, we focused on probabilistic parameters. Obviously, each parameter is limited to legal probabilities, but the real concern is that probability distributions (like the Markov state transitions, for example) add shared constraints: all the related probabilities must sum to 100 %.

This is a perfectly defined linear problem. There is a unique solution and, if one parameter is defined by the others, the solution space becomes convex. The constraints generated by the  $n$  parameters  $P_i$  of a given distribution are shown in equation (4).

$$\begin{cases} 0 \leq P_i \leq 1 - \sum_{t=1}^{n-1} P_t & \forall i \in [1..n-1] \\ P_n = 1 - \sum_{t=1}^{n-1} P_t \end{cases} \quad (4)$$

During the extra optimization step, we will define a parameter  $\lambda$  that will influence the others depending on the previous cycle's optimization offsets  $\Delta P_i$ :

$$P_i^* = P_i + \lambda \cdot \Delta P_i \quad \forall i \in [1..n] \quad (5)$$

Resolving the resulting system requires two temporary values  $\zeta$  and  $\eta$ , summarizing respectively the global influence of the cycle's modifications and the current parameter configuration. The final constraints are shown in (6):

$$\begin{aligned} \zeta &= \sum_{j=1}^{n-1} \Delta P_j \text{ and } \eta = \sum_{j=1}^{n-1} P_j \\ \begin{cases} \text{if } \Delta P_i > 0 & -P_i / \Delta P_i \leq \lambda \\ \text{if } \Delta P_i < 0 & -P_i / \Delta P_i \geq \lambda \end{cases} \\ \begin{cases} \text{if } \zeta > 0 & 1 - \eta / \zeta \leq \lambda \\ \text{if } \zeta < 0 & 1 - \eta / \zeta \geq \lambda \end{cases} \end{aligned} \quad (6)$$

Finally, merging these constraints for all the parameters of all the distributions produces the validity interval for the new parameter  $\lambda$ . Within these bounds, the implied parameters are within their validity interval (by construction of  $\lambda$ 's bounds), and all the probability distributions constraints are satisfied. Considering the bracketing algorithm, these bounds are automatically enforced; there is no need for checking the validity of any parameter at any time.

The proof and all the solving of (6) are available in [11].

#### IV. EXPERIMENTAL RESULTS

This diagnosing architecture has been experimented in various situations, ranging from locating mobile robots to anesthesia monitoring. The most advanced experiment was Diatelic [6], because all parts of it have been stressed during a 3 years prospective clinical study managed by physicians from the ALTIR. Therefore, we will focus on this particular experiment.

The fuzzy observation function has been presented in section II (see Fig. 1), along with the model's states. The transition has been defined for 2 actions: *observe* and *forget*. The first action is used when the patient is at home. It implements mainly the persistence of the health level: We assume that a patient evolves slowly. A state of *dehydration* cannot evolve into a *hyperhydration* in one day.

The *forget* action is used when the patient comes at the hospital or when a physician changes the patient's profile. In these conditions, the expected evolution is brutal. The model should forget any previous situation.

The experiment included 30 patients, half of them being monitored by our system while the others were monitored the standard way. The two groups are statistically comparable. Significance is evaluated through a standard ANOVA (analysis of variance) statistical test.

Table I shows the compared characteristics of both groups. These characteristics show no significant variation. More than these numerical values, medical causes for the renal insufficiency and residual kidney functions are also statistically comparable across the two groups. Therefore, the obtained results are really meaningful.

TABLE I.  
GROUP CHARACTERISTICS.

Group	Diatelic	Reference
Sex (Men / Women)	8 / 7	9 / 6
Average age (standard deviation)	69.8 ( $\pm 14.8$ )	70.7 ( $\pm 12.4$ )
Diabetic patients	5	4
Comorbidity (Charlson Index <sup>a</sup> )	5.7	4.8
Distance from the center (in km)	52	52.5

<sup>a</sup> The Charlson Index is an evaluation of the illness severity based on the age and pathologies, ranked with respect to their mortality.

Two interesting facts have been revealed by this experiment after 3 years: The number of spontaneous visits and the average blood pressure of patients monitored by Diatelic show a significant drop, compared to the other patients.

Whichever their group, each patient comes and visit their nephrologists once a month. This period can be adapted by the physician, so that ill patients come more often than healthy ones. However, any patient may consult at the hospital if something goes wrong.

Table II shows the same results, but classified group per group. Moreover, the total number of days spent per year at the hospital has been added. These statistics clearly show that patients monitored by Diatelic consult less than the others. In particular, an ANOVA test shows ( $P < 0.0066$ ) that the number of spontaneous visits is significantly lower for the patients who are monitored by our system. The number of expected visits is almost the same, while the number of hospitalizations seems to decrease. However, the high standard deviations of both groups prevent us from gaining significant conclusions.

TABLE II.  
YEARLY MEDICAL INTERVENTIONS FOR DIALYZED PATIENTS.

medical actions per year	Diatelic	Reference
Programmed visits	10.6 $\pm$ 2.3	11.0 $\pm$ 2.3
Spontaneous visits	2.8 $\pm$ 2.1	5.3 $\pm$ 2.7
Days at the hospital	11 $\pm$ 14.5	20.5 $\pm$ 36.1

The second point of interest is the decrease of blood pressure of Diatelic-monitored patients. As shown on the Table III, while classical patients keep their blood pressure constant, patients monitored by our module decrease theirs. Since all patients where in a slight hyperhydration state before the experiment, this is really beneficent. An ANOVA test confirms this, with  $P < 0.03$ . During the experiment, the drugs consumption of the patients shows a similar evolution, but the ANOVA reports only a  $P < 0.061$ ; therefore, this is not a significant result for physicians (the threshold is  $P < 0.05$ ).

The patients' weight has a similar behavior: whereas the Diatelic patients suffer from no significant weight increase, the others show a dramatic augmentation: Wilcoxon's test shows a significant tendency with  $P < 0.026$ .

TABLE III.  
GLOBAL EVOLUTION OF MEDICAL SIGNALS OVER A TWO YEAR PERIOD.

Variations of	Diatelic	Reference
Weight (Kg)	+0.413 ( $\pm 4.3$ )	+2.631 ( $\pm 3.9$ )
Blood Pressure (mm Hg)	-1.177 ( $\pm 1.133$ )	-0.023 ( $\pm 1.582$ )
Drugs (kinds)	-0.2 ( $\pm 0.561$ )	+0.333 ( $\pm 0.9$ )

Finally, evaluating the learning algorithm is quite difficult. First, the real condition of the patients is, in general, unknown for everyone. Even more, since we have very little data for training the model, and because they are temporally organized, it is not possible to train the algorithm of part of the data, while estimating its robustness on the remainder. Therefore, we relied totally on the physicians' own diagnosis to judge the system results.

Their evaluation is very positive. From their opinion, a model which has been trained properly can monitor a patient for several weeks, without requiring anymore tuning.

Regarding a particular situation, a patient has been hospitalized because of some hyperhydration, whereas the model indicated he was all right. By correcting the diagnosis and having the model learnt from it, we observed that the adapted model looked like the standard model for cardiac insufficient patients. The physician peeked into the patient's record and saw this patient previously had a heart insufficiency. However, a recovery had been diagnosed a few weeks before, and the model had been adapted accordingly.

Therefore, this anecdote clearly shows that the learning algorithm was able to identify the real sources of the anomaly.

## V. CONCLUSION

The diagnosing architecture we show in this paper allow for expressing the model itself in a human-understandable form. This is particularly valuable when the software is to interact with human specialists like physicians, because they are able to understand the semantics of each part of the model. For example, any parameter has a precise meaning, and its modification has the expected effects on the computation.

However, even if we care for minimizing the number of parameters, a given model may gather tens of them. Thus, setting all of them manually can be really laborious and prone to errors. To address this situation, we describe a learning algorithm that uses directions from the human expert to adapt the model while keeping its semantics intact.

Because of the low constraints on the value function the algorithm optimizes, a compromise between the expert's desires and statistical criteria can be reached. This guarantees a stable diagnosis, and allows the expert for focusing on the most important mistakes solely.

Further enhancements of this architecture may include a finer model. More precisely, the discrete nature of the state space can be responsible for some strange diagnosis behaviors sometimes.

Therefore, evolving toward a continuous model could be very useful. However, expressing such a model so that the expert could identify its semantics may be quite difficult.

#### ACKNOWLEDGMENT

We would like to thank the physicians from the ALTIR, and more particularly Dr. Durand and Dr Mario, who helped in developing this methodology. Their feedback and their investment in the clinical study have been essential to the success of the project.

#### REFERENCES

- [1] C. Huang, "Intelligent Alarms: Allocating Attention Among Concurrent Processes," *PhD thesis Stanford University*, 1999.
- [2] B.G. Buchanan, and E.H. Shortliffe, "The MYCIN Experiments of the Stanford Heuristic Programming Project". *Rule-Based Expert Systems*, Addison-Wesley, Reading, MA. 1984
- [3] N. Lavrac, E. Keravnou, B. Zupan, "Intelligent Data Analysis in Medicine," *Encyclopedia of Computer Science and Technology*, Vol.42 pp 113-157, Dekker, New York, 2000.
- [4] F. Steinmann, "On the usefulness of fuzzy sets in medical AI." *Artificial Intelligence in Medicine* 21: 131-137. 820J, 2001.
- [5] S. Koenig, and R.G. Simmons, "Unsupervised learning of probabilistic models for robot navigation", *International Conference on Robotics and Automation*, Minneapolis, 1996.
- [6] R. Hery, L. Romary, F. Charpillet, J-M Pierrel, J-P Thomesse, E. Petitjean, L. Jeanpierre, P-Y Durand, and J. Chanliau, "System for the remote monitoring of patients", *Patent* WO 0154571, 2001.
- [7] A.P. Dempster, N.M. Laird, and D.B. Rubin, "Maximum likelihood from incomplete data via the EM algorithm." *Journal of the Royal Statistical Society*, Series B, Vol. 34, 1977.
- [8] G.D. Forney, "The forward-backward algorithm." *Allerton conference on communication, control and computing*, pp. 432-446, 1997.
- [9] M.J.D. Powell, "Numerical Methods for Nonlinear Algebraic Equations," *Gordon and Breach, New York*, 1970.
- [10] R.P. Brent, "Algorithms for minimization without Derivatives," *Prentice-Hall*, 1973.
- [11] L.Jeanpierre, "Apprentissage et Adaptation pour la Modélisation Stochastique de Systèmes Dynamiques Réels". *PhD thesis*, Université Henri Poincaré Nancy-1, France, 2002.