Regards croisés sur la fouille de données

Objectif : à travers cette rencontre, nous souhaitons réfléchir au couplage entre des concepts de statistiques (apprentissage, méthodes de grande dimension, techniques d'interpolation et de lissage, estimation bayésienne, assimilation de données, compressed sensing) d'une part, et des outils de simulation numérique basés sur la modélisation mathématique (EDP, systèmes dynamiques) de la totalité ou d'une partie du système considéré d'autre part, afin d'obtenir des résultats plus complets pour l'extraction d'informations de bases de données importantes. Il s'agit donc d'enrichir les approches de type boîte noire (input-output) par des modèles réalistes et validés utilisables au travers de la simulation numérique.
Exemples d'interaction données / simulation en modélisation cardiovasculaire

Journée “Regards croisés sur la fouille de données”
LJLL, 16 novembre 2011

Jean-Frédéric Gerbeau

INRIA Paris-Rocquencourt & Université P. & M. Curie (Paris 6)
France
Outline

• Cardiac Electrophysiology
  - data for a “statistical validation” of simulations
  - simulation to generate data (“virtual lab”)
  - simulations to enrich data (“augmented reality”)

• Blood flow
  - data to build a reduced model
  - data assimilation
Electrocardiograms

**Cell scale:**
Hodgkin-Huxley-like models

\[
\begin{align*}
C_m \frac{dV_m}{dt} + I_{ion}(V_m, g) &= 0 \\
\frac{dg}{dt} + G(V_m, g) &= 0
\end{align*}
\]

**Myocardium:**
bidomain models

\[
\begin{align*}
A_m \left( C_m \frac{\partial V_m}{\partial t} + I_{ion}(V_m, g) \right) - \text{div}(\sigma_i \nabla u_i) &= A_m I_{app}, \\
\text{div}(\sigma_e \nabla u_e) + \text{div}(\sigma_i \nabla u_i) &= 0 \\
\frac{\partial g}{\partial t} + G(V_m, g) &= 0
\end{align*}
\]

**Torso:**
Poisson equation

\[
\begin{align*}
\text{div}(\sigma_T \nabla u_T) &= 0 \\
u_T &= u_e, \text{ on } \Sigma \\
\sigma_e \nabla u_e \cdot n &= -\sigma_T \nabla u_T \cdot n_T, \text{ on } \Sigma
\end{align*}
\]
12-lead ECG

- Simulated ECG:

- Example of a real ECG:

Boulakia, Cazeau, Fernández, JFG, Zemzemi, Annals Biomed Engng 2010
Infarct

- **Anterior** infarct: ST elevation
- **Posterior** infarct: ST depression

Example: Anterior infarct

*Simulation: E. Schenone*
Bundle branch blocks

Healthy case

Right bundle branch block

Chapelle, Fernández, JFG, Moireau, Sainte-Marie, Zemzemi, FIMH 2009
Statistical classification
(F. Ieva & AM. Paganoni, Politecnico di Milano)

- **Prometeo** project: 198 real ECGs recorded by Basic Life Support (BLS) in Milan since 2008
- Three steps:
  - Smoothing
  - Registration

- Statistical k-means clustering:
  \[
  \arg \min_{S=\{S_1, \ldots, S_k\}} = \sum_{i=1}^{k} \sum_{x_j \in S_i} d(x_j, \mu_i)
  \]
Statistical classification

(F. Ieva & AM. Paganoni, Politecnico di Milano)

- Confusion matrices:

<table>
<thead>
<tr>
<th></th>
<th>$d_1$</th>
<th></th>
<th>$\tilde{d}_1$</th>
<th></th>
<th>$d_2$</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Normal</td>
<td>RBBB</td>
<td>LBBB</td>
<td>Normal</td>
<td>RBBB</td>
</tr>
<tr>
<td>1</td>
<td>95</td>
<td>7</td>
<td>1</td>
<td>71</td>
<td>12</td>
</tr>
<tr>
<td>2</td>
<td>6</td>
<td>42</td>
<td>3</td>
<td>30</td>
<td>36</td>
</tr>
<tr>
<td>3</td>
<td>0</td>
<td>0</td>
<td>44</td>
<td>0</td>
<td>1</td>
</tr>
</tbody>
</table>

Data used to validate “statistically” computer simulations

- Pilot analysis:
  - Database of 25 normal ECG, 10 LBBB, 13 RBBB
  - Our normal, LBBB and RBBB ecg are correctly classified !!!

- To go further:
  - How to reproduce the variability of real ECG?
5-8 potential measures in the heart (EGM) are available “for free” in the pacemaker

Purpose: use these data to reconstruct a “rough ECG”, useful for a quick clinical examination

For a patient, ELA has a pair EGM/ECG at “t=0”. They use it to define a “transfer function”

Question: how robust is the transfer function with respect to a change in the patient condition?

Computer simulations used to generate data
The algorithm can reproduce what has been learnt!
• Learning: healthy case.

• Reconstruction: “pathological” case (“RBBB” 10 ms)

• If the conditions change: the algorithm fails...
Our approach

- Enrich clinical data with simulation data
- Machine learning techniques (Kernel ridge regression):

$$\min_{f \in \mathcal{H}} \frac{1}{n} \sum_{i=1}^{n} (y_i - f(x_i))^2 + \lambda \|f\|_{\mathcal{H}}$$

where $\mathcal{H}$ is the RKHS associated to the kernel

$$K(x, y) = \exp(-d(x, y)^2 / 2\sigma^2)$$

and $\forall x \in \mathcal{X}, f(x) = \sum_{i=1}^{n} \alpha_i K(x_i, x)$
RBBB 50 ms  
Training: RBBB 50 ms

RBBB 50 ms  
Training: RBBB 10 ms

RBBB 50 ms  
Training: 
RBBB 10, 20, 40, 60, 70, 80, 90, 100, 120, 150, 200 ms

Computer simulations used to enrich the data

Ebrard, Fernandez, JFG, Rossi, Zemzemi, FIMH 2009
**Real data:**

- Patient with very frequent extrasystole
- Training on the first 10 heart beats
- Test from heart beat 40 to 50
The Reduced Basis Element Method:
Application to Fluid Flow and Maxwell’s Equations

Yvon Maday

1Laboratoire Jacques-Louis Lions
Université Pierre et Marie Curie, Paris, France
and Brown Univ.

EU Regional School 2010 in Computational Engineering Science
Collaboration with

- Y. Chen, UMass Dartmouth
- J. Hesthaven, Brown
- E. Løvgren, Simula
- E. M. Rønquist, Trondheim
Motivation

Among the most prominent applications in medical applications is the analysis of internal flows

- blood flows in arteries
- air flow in the lung

**Figure:** Reconstructed geometries of Willis complex (Thiriet) and of the upper part of the lung (Fetita-Prêteux)
Another motivation comes from the design of mechanical parts, e.g.

- air conditioning system
- wave guides ...
Motivation

- In this range of applications, the challenge of the simulations comes more from the complexity of the geometry.
- There is some repetitiveness or similarities in the behavior of the flow that allows for the definition of reduced model strategies.
The reduced basis element method

- Reduced basis approximation
- Domain decomposition
The reduced basis method relies on few basic concepts (1/2)

- We are looking for the solution to a family of parametrized problems expressed as PDE (stationary or evolution),

- A standard approximation (finite element, finite differences, finite volumes, spectral methods...) is available and can be used to compute some instances of the solution,

- The complexity of the set of all solutions is rather small; as a consequence it is possible to choose some parameters so that the associated solutions constitute an alternative discretization basis: the reduced basis,
The reduced basis method relies on few basic concepts (2/2)

- We are ready to perform some costly computations (off line, e.g., last week) provided this allows today to be more efficient,
- The problem should be written as an affine problem (or at least can be approximated as such),
- There exists an a posteriori estimator that allows to validate the RB computations.
Let us consider a family of parameter dependant problems (PDE) : Find $u$ solution of

$$\mathcal{F}(u, \mu) = 0$$

where the parameter $\mu$ belongs to, e.g. $\mathbb{R}^d$.

This can be an inverse type problem or an optimisation one, or even a control problem.

We are able to provide an approximation (accurate enough) $u(\mu)$ of each problem (1) : we denote as $u_\delta(\mu)$ these approximated solutions.
Even if these computations are feasible, the approximation of \( u(\mu) \) for many values of the parameter \( \mu \) is way too expensive.

It is assumed that the problem is regular in \( \mu \) implying a regular behavior of \( u_\delta(\mu) \) as a function of \( \mu \),

in the sense that the set of all solutions \( u_\delta(\mu) \) when \( \mu \) varies is of small size, in a sense to be made precise.
Let us consider the set of all discrete solutions $X_\delta = \text{Vect}\{u_\delta(\mu), \mu \in \mathcal{D}\}$.

- We are looking for a way to approximate $X_\delta$ by a (very) small dimension space: one criterion can be PCA (actually can be used in a feasibility step but generally way too expensive...)

**Important:** note that, in order to be precise enough, the standard approximation (symbolised by $\delta$) relies on very high number of dof (e.g., $10^6$ degrees of freedom) ... in this context, small denotes something $<< 100$
Of course the previous hypothesis requires some properties on the set of all solutions (regularity..)
that is relative to the smoothness of the set \( \{ u(\mu), \mu \in \mathcal{D} \} \).
This smoothness can be characterized by the notion of \( n \)-width following Kolmogorov
We are looking for a way to approximate $X_\delta$ by a (very) small dimension space.

We are able to choose particular values for $\mu_k$, $k = 1, \ldots, N$, and we approximate by $X_\delta$ par $X_N = \text{Vect}\{u_\delta(\mu_k), \ k = 1, \ldots, N\}$, hence, as a linear combination of $u_\delta(\mu_k)$

$$u(\mu) \approx \sum_{k=1}^{N} \alpha_k(\mu) u_\delta(\mu_k),$$

A Galerkin type method is proposed to approximate $u(\mu)$ ... that reveals much smarter than the inter- or extra-polation.
We are interested in solving
\[-\Delta u + \mu_1 \frac{e^{\mu_2 u} - 1}{\mu_2} = f\]

the results with the interpolation process are

<table>
<thead>
<tr>
<th>$N$</th>
<th>4</th>
<th>8</th>
<th>12</th>
<th>16</th>
<th>20</th>
</tr>
</thead>
<tbody>
<tr>
<td>erreur sur $u$</td>
<td>6.53 E-03</td>
<td>1.05 E-03</td>
<td>7.34 E-05</td>
<td>1.30 E-05</td>
<td>5.05 E-06</td>
</tr>
</tbody>
</table>
We are interested in solving

\[
\frac{\partial u}{\partial t} - \Delta u + \mu_1 \frac{e^{\mu_2 u} - 1}{\mu_2} = f
\]

the reduced basis considers 3 parameters \(\mu_1, \mu_2, t\)... results are similar

<table>
<thead>
<tr>
<th>(N)</th>
<th>1</th>
<th>10</th>
<th>20</th>
<th>30</th>
<th>40</th>
</tr>
</thead>
<tbody>
<tr>
<td>erreur sur (u)</td>
<td>(3.82 \times 10^{-1})</td>
<td>(1.62 \times 10^{-3})</td>
<td>(1.46 \times 10^{-4})</td>
<td>(1.88 \times 10^{-5})</td>
<td>(4.94 \times 10^{-6})</td>
</tr>
</tbody>
</table>
Reduced element method

The domain of interest is first decomposed into several subdomains,

\[ \overline{\Omega} = \bigcup_{k=1}^{K} \overline{\Omega}_{k}^{bb} \]

where each “building block” \( \Omega_{k}^{bb} \) is assumed to be the image of a reference \( \hat{\Omega} \).

The mapping \( \varphi_{k} \) between \( \hat{\Omega} \) and \( \Omega_{k}^{bb} \) will be assume to be piecewise affine (and obviously continuous) so that

\[ \Omega_{k}^{bb} = \varphi_{k}[\hat{\Omega}] \]
Figure: A first geometry.
Figure: A second geometry.
Figure: A third geometry.
Figure: A fourth geometry.
Reduced element method

Figure: A fourth geometry.
Reduced element method

As a precomputation, the problem of interest is solved over various deformations of each reference building block and stored, after mapping, on the reference building block. This gives basis functions \( \hat{\zeta}_1, \hat{\zeta}_2, \ldots, \hat{\zeta}_N \), supposed to be linearly independent. These basic solutions are mapped over each \( \Omega_{bb}^k \) through \( \varphi_k \). The solution corresponding to an unknown, deformed geometry is then represented as a linear combination of these mapped solutions

\[
Y_N = \{ v_N \in L^2(\Omega) \mid v_N|_{\Omega_{bb}^k} \circ \varphi_k \in \text{span}\{\hat{\zeta}_1, \hat{\zeta}_2, \ldots, \hat{\zeta}_N\} \}.
\]

Note that \( Y_N \) is not an acceptable discretization space for \( H^1(\Omega) \), the matching between the different subdomains is ensured through the use of Lagrange multipliers.
Reduced element method

We now define $X_N$ to this purpose by \textit{gluing} the functions of $Y_N$ across the interfaces $\gamma_{k,\ell}^{bb}$ between two stages

$\rightarrow$ Lagrange multipliers

$$X_N = \{ v \in Y^1_N, \quad \forall k, \ell, \quad \forall \psi \in W_{k,\ell}, \quad \int_{\Gamma^e_{k,\ell}} (v^+ - v^-) \psi \, ds = 0 \},$$

where $W_{k,\ell}$ is some well chosen space over $\Gamma^e_{k,\ell}$ $\rightarrow$ nonconforming approximation.

The discrete problem then reads: Find $u_N$ in $X_N$ such that

$$a(u_N, v_N) = f(v_N), \quad \forall v_N \in X_N.$$
Fluid flows

Figure: Domain decomposition.

Y. Maday (UPMC & Brown)
Figure: Error distribution for a new configuration $N_P = 15$, $N_B = 15$ error plot for the pressure error $\max = 3.10^{-2}$, for the velocity error $\sim 3.10^{-3}$. 
Figure: Error distribution for a new configuration $N_P = 15$, $N_B = 30$ error plot for the pressure max=$6.10^{-3}$, for the velocity error $\sim 4.10^{-4}$, size problem.
Figure: A stenosis problem with $N_P = 15, N_B = 15$. 
Figure: A stenosis problem with $N_P = 15$, $N_B = 30$. 
Table: Steady Stokes solution on a multi-block bypass with three pipe blocks and two bifurcation blocks. Here, $N = 3N_1 + 2N_2$. 

| $N$ | $N_1$ | $N_2$ | $|u_N - u|_{H^1}$ | $||p_N - p||_{L^2}$ |
|-----|-------|-------|------------------|-------------------|
| 45  | 9     | 9     | $9.3 \cdot 10^{-3}$ | $3.3 \cdot 10$    |
| 55  | 11    | 11    | $3.1 \cdot 10^{-3}$ | $5.3 \cdot 10^{-1}$|
| 65  | 13    | 13    | $2.3 \cdot 10^{-3}$ | $9.0 \cdot 10^{-2}$|
| 75  | 15    | 15    | $1.4 \cdot 10^{-3}$ | $5.3 \cdot 10^{-2}$|
| 105 | 15    | 30    | $5.4 \cdot 10^{-4}$ | $3.0 \cdot 10^{-2}$|
Certified Reduced Basis Model Characterization: a Frequentistic Uncertainty Framework

D.B.P. Huynh, D.J. Knezevic, A.T. Patera

aDepartment of Mechanical Engineering and Center for Computational Engineering, Massachusetts Institute of Technology, Cambridge, MA, 02139 USA
Figure 1: The transient thermal conduction problem. The origin of the $x = (x_1, x_2)$ coordinate system is the lower left corner.
Figure 2: Possibility regions for $\epsilon = 0.005$, $\gamma = 0.95$, $K = 10$, $M = 101$, $I = 0$ and (a) $N = 10$ and (b) $N = 40$; $\kappa^* = (4,1)$ is indicated in each plot with a green cross.
Figure 3: Possibility regions for white noise with $\epsilon = 0.005$. We set $\gamma = 0.95$, $I = 0$, $N = 40$ and (a) $K = 10$, $M = 101$; (b) $K = 2$, $M = 101$; (c) $K = 2$, $M = 6$. 
Quelques idées venues depuis cette journée
Dans un modèle multi-domaine : utiliser les données acquises pour reconstruire la solution...

- avec des « magic points » ou des « magic moments »

ex sur le tableau
Dans un système où la modélisation par EDP et la simulation ont été validées, un système de bases réduites est construit :

On utilise la base réduite et les « magic points », ou « magic moments » pour assurer des cohérences entre différentes données → données statistiques améliorées