Feature selection in models described by ODEs and PDEs

Jean-Frédéric GERBEAU [1,2], **Damiano LOMBARDI** [1,2], **Eliott TIXIER** [2,1]

[1] Centre INRIA de Paris, France
[2] Laboratoire Jacques-Louis Lions, UPMC Sorbonne Universités, France

eliott.tixier@inria.fr

Introduction

- Aim: select a **small set** of features designed to **easily solve** inverse and classification problems.
- Each model parameter associated with feature (**numerical biomarker**), solution of optimization problem.
- Method is applied to biological models in realistic scenarios.
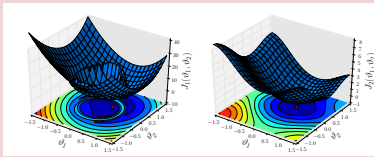

Figure 1: Cost functions associated with an inverse problem: using the whole signal (left) and using the numerical biomarkers (right).

Methods

- Uncertain parameters: $\boldsymbol{\theta} = (\theta_1, \ldots, \theta_p)$.
- **Dictionary of features** extracted from model outputs: $\{b_1, \ldots, b_m\}$.
- Numerical biomarker: linear combination of given dictionary entries: $y_h(\boldsymbol{\theta}) = \sum_{j=1}^m w_j^{(h)} b_j(\boldsymbol{\theta}), \ 1 \le h \le p$.
- Objective: find numerical biomarker **maximally correlated** with parameter and **minimally correlated** with all the other parameters.
- This translates into:

$$\forall h \in \{1, \ldots, p\}, \quad \begin{cases} \max_{y_h} & \text{cov}(y_h(\boldsymbol{\theta}), \theta_h) \\ \min_{y_h} & |\text{cov}(y_h(\boldsymbol{\theta}), \theta_k)|, \quad \forall k \ne h \\ \text{s.t.} & \text{var}(y_h(\boldsymbol{\theta})) = 1 \end{cases}$$

- Weights are **sparse** for interpretability and to reduce overfitting.
- Weights $\mathbf{w}^{(h)}$ minimize following cost function:

$$2\mathcal{J}_h(\mathbf{w}^{(h)}) = \sum_{k=1}^p \left[\mathbb{E}(y_h \theta_k) - \delta_{kh}\right]^2 + \xi \left[\mathbb{E}\left(y_h^2\right) - 1\right]^2 + \lambda_h \|\mathbf{w}^{(h)}\|_1.$$

- Problem discretized by sampling parameters and evaluating model for each sample.
  $\mathbb{E}[\cdot]$ operator approximated using Monte-Carlo quadrature.
- Cost function minimized with Nesterov accelerated gradient [1].
- Sparsity-promoting λ_h calibrated with "L-curve" criterion.

Use of numerical biomarkers

Inverse Problems

- 2-distance in numerical biomarker space is (up to perturbation), equivalent to 2-distance in parameter space.
- Parameter estimation problem becomes **easier to solve** (Fig. 2).
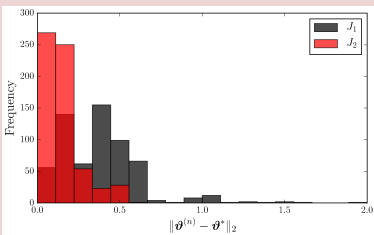- **Gradient-free minimization** algorithm [3] with successive local quadratic approximations is used.

Figure 2: Histograms of the parameter estimation error (2-distance to true solution in parameter space) after performing 625 different inverse problems with the action potential model (see Applications section): using the whole signal (black) and using the numerical biomarkers (red) in the cost function.

Classification

- Numerical biomarkers used as features in classification problems.
- Support Vector Classification with Gaussian kernel [2] algorithm.

Application to biological models described by ODEs or PDEs

Electrophysiology: single cell action potential

- Heart cell 0-D action potential model: 29 coupled non-linear ODEs.
- Need for fast and robust parameter estimation to assess effect of drugs onto heart muscle cells.
- Observable: function of time from which features are extracted: duration, amplitude, etc. (Fig. 3).
- Dictionary enriched with derivatives, area under curve, etc.
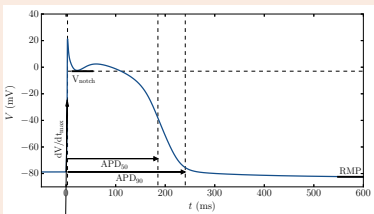- Numerical biomarkers are corrections of known biomarkers (Fig. 4).

Figure 3: Action potential as a function of time. Highlighted features: action potential duration (APD), upstroke derivative (dV/dt$_{max}$), resting potential (RMP), notch potential (V$_{notch}$).
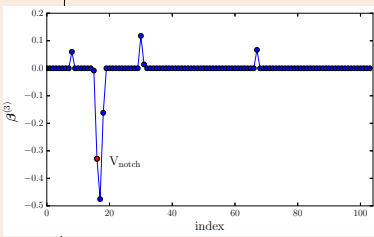

Figure 4: Numerical biomarker associated with 4th parameter. It is a correction of V$_{notch}$, classically associated with this parameter.

Electrophysiology: tissue field potential

- 2-D model at tissue scale. Field potential is function of time (Fig. 5) and is available at 9 locations in space (electrodes).
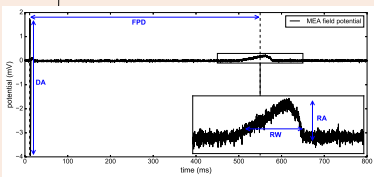- Need for robust features to use in a drug classification context.

Figure 5: Field potential as a function of time. Highlighted features: field potential duration (FPD), repolarization amplitude (RA), repolarization width (RW).

Score	Classical biomarkers		Numerical biomarkers	
	mean	std.	mean	std.
Cohen's kappa	0.18	0.15	0.51	0.27
θ_1 AUC	0.65	0.10	0.86	0.11
θ_2 AUC	0.92	0.09	1.00	0.00
θ_3 AUC	0.36	0.11	0.82	0.10
averaged AUC	0.65	-	0.89	-

Table 1: Classification scores for MEA measurements comparing classical and numerical biomarkers. Cohen's kappa is a score ranging from -1 (worst) to 1 (best), 0 being coin flip. AUC is the area under curve of Receiver Operating Characteristic (ROC).

- Classification results more accurate using numerical biomarkers.

Cardiovascular: full-body hemodynamics

- Network of 55 1-D arteries.
- Need for a feature to monitor arterial stiffness, independently of natural daily fluctuations of pressure wave velocity (PWV).
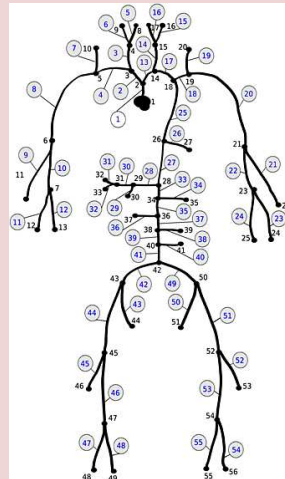- Numerical biomarker is correction of PWV (Fig. 7)

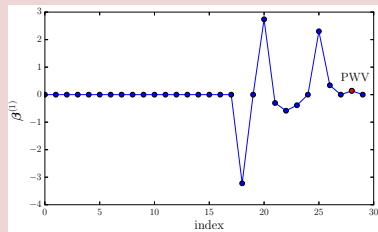**Figure 6:** Full body 1-D haemodynamics model of 55 arteries [4].

Figure 7: Numerical biomarker weights associated with 2nd parameter (arterial stiffness). Combination of PWV (used in practice) and mechanical powers.

Links

A Python implementation of the numerical biomarkers design algorithm is available on our **GitHub repository**:
https://github.com/eltix/numbio

Figure 8: Flash this QR-code to access our GitHub repository.

References

[1] Brendan O'Donoghue and Emmanuel Candès. Adaptive restart for accelerated gradient schemes. *Foundations of computational mathematics*, 15(3):715–732, 2015.

[2] F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, and E. Duchesnay. Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research*, 12:2825–2830, 2011.

[3] Michael JD Powell. The bobyqa algorithm for bound constrained optimization without derivatives. *Cambridge NA Report NA2009/06, University of Cambridge, Cambridge*, 2009.

[4] Philippe Reymond, Fabrice Merenda, Fabienne Perren, Daniel Rüfenacht, and Nikos Stergiopulos. Validation of a one-dimensional model of the systemic arterial tree. *American Journal of Physiology-Heart and Circulatory Physiology*, 297(1):H208–H222, 2009.