Model driven machine learning and some applications

Tamás Dózsa [1,2]

[1]Hungarian Research Network
Institute for Computer Science and Control,
Systems and Control Lab

[2]Department of Numerical Analysis,
Faculty of Informatics,
Eötvös Loránd University

Budapest, 2023.11.21

Project no. C1748701 has been implemented with the support provided by the Ministry of Culture and Innovation of Hungary from the National Research, Development and Innovation Fund, financed under the NVKDP-2021 funding scheme.

Outline

Adaptive signal models

Variable subspaces in Hilbert spaces

- Let $(\mathcal{H}, \langle\cdot,\cdot\rangle)$ be a Hilbert space and $f \in \mathcal{H}$ arbitrary.
- Suppose $\varphi_k^{\eta} \in \mathcal{H}$ is a complete and orthogonal basis $(k \in \mathbb{N})$.
- $\boldsymbol{\eta} \in \Gamma \subset \mathbb{R}^L$ $(L \in \mathbb{N})$ determines the system φ_k^{η}.
- For $m \in \mathbb{N}$, $\varphi_0^{\eta}, \ldots, \varphi_{m-1}^{\eta}$ span the m-dimensional closed subspace $\mathcal{U} \subset \mathcal{H}$.
- For a fixed $\boldsymbol{\eta}$, $\exists! \widehat{f} = \sum_{k=0}^{m-1} c_k \varphi_k^{\eta} \in \mathcal{U}$ for which $\|\widehat{f} - f\|$ is minimal and

$$\langle f - \widehat{f}, g\rangle = 0 \quad (\forall g \in \mathcal{U}).$$

Adaptive signal models

Problem statement for applications

- Applications: $\boldsymbol{L_2(\mathbb{R})}$, $H_2(\mathbb{D})$, etc.
- Some apriori information about f is usually known.
- Suppose $m << n$. We want to find a good approximation

$$f_i = f(t_i) \approx \sum_{k=0}^{m-1} c_k \varphi_k^{\eta}(t_i) = (\boldsymbol{\Phi}(\boldsymbol{\eta})\boldsymbol{c})_i$$

$$(i = 1, \ldots, n,\ f \in L_2(\mathbb{R}))$$

- Questions
  1. How to choose the basis φ_k^{η}?
  2. How to determine the optimal $\boldsymbol{\eta}$?

Adaptive signal models

A nonlinear optimization problem

We look for the optimal parameter vector $\boldsymbol{\eta} \in \Gamma \subset \mathbb{R}^L$, for which

$$r_2(\boldsymbol{\eta}) = \|\mathbf{f} - \Phi(\boldsymbol{\eta})\Phi^+(\boldsymbol{\eta})\mathbf{f}\|_2^2 = \|\mathbf{f} - \mathbf{P}_{\Phi(\boldsymbol{\eta})}\mathbf{f}\|_2^2.$$

so-called variable projection functional is minimal.

Properties of variable projection operators

- $\mathbf{P}_{\Phi(\boldsymbol{\eta})}\mathbf{f}$ is the orthogonal projection of $\mathbf{f}$ onto the column space of $\Phi(\boldsymbol{\eta})$.
- The gradient of $r_2(\boldsymbol{\eta})$ can be analytically calculated.[1]
- Minimizing r_2 w.r.t. $\boldsymbol{\eta}$ is known as a **separable nonlinear least squares** (SNLLS) problem.

[1] G., Golub, V., Pereyra. "The differentiation of pseudo-inverses and nonlinear least squares problems whose variables separate." SIAM Journal on numerical analysis (1973)

Adaptive signal models

Signal representation

- Let

$$\boldsymbol{\eta}^* = \arg\min_{\boldsymbol{\eta}\in\mathbb{R}^L} r_2(\boldsymbol{\eta}) = \arg\min_{\boldsymbol{\eta}\in\mathbb{R}^L} \|\mathbf{f} - \Phi(\boldsymbol{\eta})\Phi^+(\boldsymbol{\eta})\mathbf{f}\|_2^2.$$

- Depending on the application we can represent $\boldsymbol{f}$ by
  1. $\boldsymbol{f} \to \boldsymbol{\eta}^* \in \mathbb{R}^L$ (analysis)
  2. $\boldsymbol{f} \to \boldsymbol{c} = \Phi^+(\boldsymbol{\eta}^*)\mathbf{f} \in \mathbb{R}^m$ (dimension reduction).
  3. $\boldsymbol{f} \to \Phi(\boldsymbol{\eta}^*)\Phi^+(\boldsymbol{\eta}^*)\mathbf{f} \in \mathbb{R}^n$ (low pass filtering).
  4. $\boldsymbol{f} \to \boldsymbol{f} - \Phi(\boldsymbol{\eta}^*)\Phi^+(\boldsymbol{\eta}^*)\mathbf{f} \in \mathbb{R}^n$ (high pass filtering).

System identification

Discrete time SISO-LTI systems

$$\boldsymbol{x} = \boldsymbol{h} * \boldsymbol{u} \stackrel{\mathcal{Z}}{\longmapsto} X(z) = H(z)U(z),$$

where

- $\boldsymbol{u}, \boldsymbol{x}$ input and output sequences,
- $\boldsymbol{h}$ impulse response,
- X, Y, H are the $\mathcal{Z}$-transforms of $\boldsymbol{x}, \boldsymbol{y}, \boldsymbol{h}$.
- Suppose system is causal and BIBO stable $\implies H \in H_\infty(\mathbb{D}) \subset H_2(\mathbb{D})$.
- **Identification task:** Find the (inverse) **poles**/zeros of the transfer function $H(z)$ in $\mathbb{D}$.

System identification

MT functions

- Approximate $H \in H_2(\mathbb{D})$ by a complete and orthogonal basis φ_k^η $(k = 0, \ldots, m)$.
- **Idea:** choose Malmquist-Takenaka (MT) functions as the basis functions:

$$\varphi_k^\eta(z) = R_{a_k} \prod_{j=0}^{k-1} B_{a_j}(z) = \frac{\sqrt{1-|a_k|^2}}{1-\bar{a}_k z} \prod_{j=0}^{k-1} \frac{z-a_j}{1-\bar{a}_j z},$$

  where $\boldsymbol{\eta} := (a_0, \ldots, a_{m-1}) \in \mathbb{D}^m$.
- The functions φ_k^η have poles at $1/\bar{a}_j$ $(j \leq k,\ a \in \mathbb{D})$.
- MT systems are complete and orthonormal in $H^2(\mathbb{D})$, provided $\boldsymbol{\eta}$ satisfiers the *Szász condition*.
- Choosing $\boldsymbol{\eta} = (0, 0, 0, \ldots)$ we get the trigonometric system.

System identification

SNLLS formulation

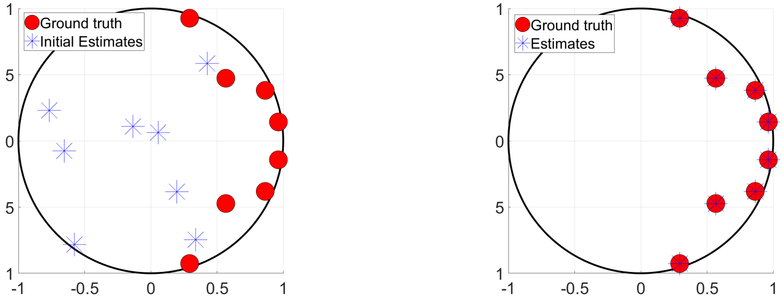
- The frequency response $H|_{\mathbb{T}} \in H_2(\mathbb{T}) \subset L_2(\mathbb{T})$.
- Denote by $\mathbf{h} \in \mathbb{C}^n$ $(n \in \mathbb{N})$ a discrete sampling of $H|_{\mathbb{T}}$.
- SNLLS formulation:

$$r_2(\boldsymbol{\eta}) = \|\Re \boldsymbol{h} - \Phi(\boldsymbol{\eta})\Phi^+(\boldsymbol{\eta})\Re \boldsymbol{h}\|_2^2,$$

  where $\boldsymbol{\eta} = (r_0, \mu_0 \ldots, r_{m-1}, \mu_{m-1}) \subset \mathbb{R}^{2m}$, where $a_k = r_k e^{i\mu_k}$ and the columns of $\Phi(\boldsymbol{\eta})$ contain *real MT-functions*[1] sampled on $\mathbb{T}$.
- It suffices to approximate the real part of $\boldsymbol{h}$ (see e.g. *Titschmarsh theorem*).

[1] F. Schipp, "Racionális ortogonális rendszerek a jel- és képfeldolgozásban", ELTE-IK Jegyzettár (2016)

System identification



(a) Random η inverse poles. **(b)** Optimal η^* parameters.

Figure: A numerical example from[1].

[1] T. Dózsa, M. Szabari, A. Soumelidis, P. Kovács, "Pole identification using discrete Laguerre expansion and variable projection", In. Proc. The 22nd World Congress of the International Federation of Automatic Control (IFAC2023), (2023)

Supervised learning

Problem formulation

- Approximate $G : \mathcal{X} \to \mathcal{Y}$, where $\mathcal{Y}$ is a topological space.
- Common domains for G: $\mathcal{X} \subset \mathbb{R}^n$ or $\mathcal{X} \subset \mathbb{R}^{n \times s}$ $(n, s \in \mathbb{N})$.
- Common ranges for G:
  - $\mathcal{Y} \subset \{1, 2 \ldots n\}$ $(n \in \mathbb{N})$ (**classification**, i.e. $|\mathcal{Y}| = 2$ binary classification).
  - $\mathcal{Y} \subset \mathbb{R}^n$ $(n \in \mathbb{N})$ (**Regression**).

Models

- Identify $G_{\boldsymbol{\theta}} \approx G$, where $\boldsymbol{\theta}$ is a vector of parameters (usually $\boldsymbol{\theta} \in \mathbb{R}^P$, or $\boldsymbol{\theta} \in \mathbb{C}^P$, $P \in \mathbb{N}$).
- **Goal:** find $\boldsymbol{\theta}$ so that $E : \mathcal{Y} \times \mathcal{Y} \to [0, \infty)$

$$E(G(f), G_{\boldsymbol{\theta}}(f)) \quad f \in \mathcal{X}$$

is minimized for *all* $f \in \mathcal{X}$.

Model training and evaluation

Training and test sets

- Suppose that the values $G(f_1), G(f_2), \ldots, G(f_q) \quad (f_k \in \mathcal{T} \subset \mathcal{X},\ k = 1, \ldots, q)$ are known.
- Training and test sets: $\mathcal{T}_{tr} \cup \mathcal{T}_{te} = \mathcal{T}, \quad \mathcal{T}_{tr} \cap \mathcal{T}_{te} = \emptyset.$

Model training and evaluation

- **Training:** solve
$$\min_{\theta \in \mathbb{R}^P} E(G_{\boldsymbol{\theta}}(f), G(f)), \quad (\forall f \in \mathcal{T}_{tr})$$
Denote by $\boldsymbol{\theta}^*$ a solution.
- **Testing:** evaluate
$$\frac{1}{|\mathcal{T}_{te}|} \sum_{f \in \mathcal{T}_{te}} E(G_{\boldsymbol{\theta}^*}(f), G(f)).$$

Classical supervised learning framework

Supervised learning steps

1. Identify $\mathcal{T} \subset \mathcal{X}$ and $G(f)$ ($f \in \mathcal{T}$). Measurements, expert input, data augmentation etc.
2. Feature extraction:
   - Classical approach: transform the dataset $\mathcal{T}$ before training. (e.g. PCA, time frequency representations, etc.)
   - **Feature extraction transformations are incorporated into the model G_θ (e.g. convolutional neural networks)**.
3. Choose model architecture $G_{\boldsymbol{\theta}}$. (e.g. SVM, Neural Networks, etc.).
4. Train model on (possibly transformed) data from $\mathcal{T}_{tr}$ and evaluate on $\mathcal{T}_{te}$.

Open questions

Common issues

- Enough data? Labelled correctly?
- Loss function represents task to solve?
- Was the model architecture chosen correctly?
- Can extracted features be interpreted?
- Can we trust the trained model's predictions?

Model driven ML

Incorporate domain knowledge into supervised learning schemes through the use of interpretable mathematical models.

VP-NET

Variable Projection (VP) layers

- Suppose that for $\boldsymbol{f} \in \mathbb{R}^n$

$$\boldsymbol{f}_k = f(t_k) \quad (f \in L_2(\mathbb{R})).$$

Then, the mappings

1. $G_\eta(\boldsymbol{f}) = \boldsymbol{c} = \Phi^+(\boldsymbol{\eta})\mathbf{f} \in \mathbb{R}^m$,
2. $G_\eta(\boldsymbol{f}) = \Phi(\boldsymbol{\eta})\Phi^+(\boldsymbol{\eta})\mathbf{f} \in \mathbb{R}^n$,
3. $G_\eta(\boldsymbol{f}) = \boldsymbol{f} - \Phi(\boldsymbol{\eta})\Phi^+(\boldsymbol{\eta})\mathbf{f} \in \mathbb{R}^n$,

where $\boldsymbol{\eta} \in \mathbb{R}^P$ and the columns $\Phi(\boldsymbol{\eta}) \in \mathbb{R}^{n\times m}$ contain samplings of an orthogonal basis in $L_2(\mathbb{R})$ are called **variable projection (VP)** layers.

VP-NET

Properties of VP layers

- The gradients $\frac{\partial G_\eta}{\partial \eta}$ can be analytically calculated provided $\frac{\partial \Phi(\eta)}{\partial \eta}$ is known.[1]
- G_η can implemented as a layer in a neural network.
- If the basis functions in the columns of $\Phi(\boldsymbol{\eta})$ are chosen correctly, the parameters $\boldsymbol{\eta}$ can have physical meanings.[2]
- Usually $\boldsymbol{\eta}$ contains less parameters than equivalent convolution layer.

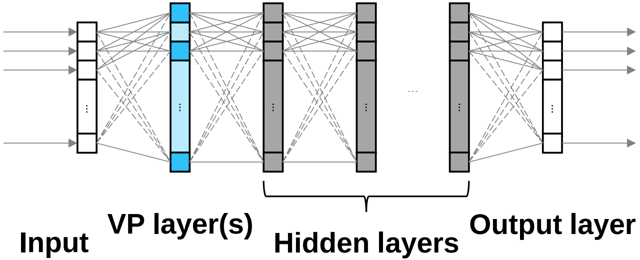
[1] G., Golub, V., Pereyra. "The differentiation of pseudo-inverses and nonlinear least squares problems whose variables separate." SIAM Journal on numerical analysis (1973)

[2] P. Kovács, G. Bognár, C. Huber, and M. Huemer. "VPNET: Variable projection networks." International Journal of Neural Systems (2022)

VP-NET

Usual VP-NET architecture

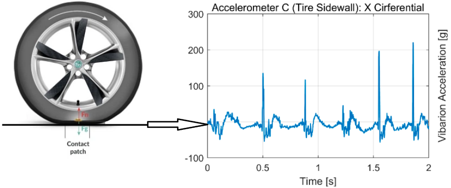
- First layers are VP-layers: these learn an appropriate representation of the data.
- Lower layers are fully connected: these solve the classification/regression task.



Tire sensor signal processing

Problem description

- Sensor: tire implanted 3D force sensors.
- Signals: changes in resistance due to mechanical forces.
- Task: surface abnormality detection.



Collaboration

SZTAKI

ELTE EÖTVÖS LORÁND UNIVERSITY

NANOSENSORS
Centre for Energy Research
Institute of Technical Physics and Materials Science

Tire sensor signal processing

Signal properties

- Quasi periodic, quasi compactly supported.
- Width of support changes with vehicle speed.
- Tire revolutions occurring on abnormal surface lower SNR.
- **IDEA:** construct VP-layer using variable projection and adaptive Hermite functions. Adaptive high pass filtering.

Tire sensor signal processing

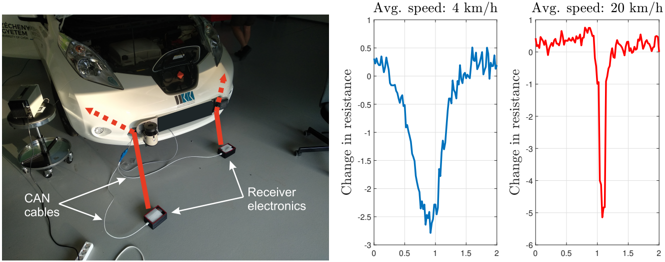


Figure: LEFT: test vehicle and readout electronics **RIGHT**: signal morphologies at different velocities.

Tire sensor signal processing

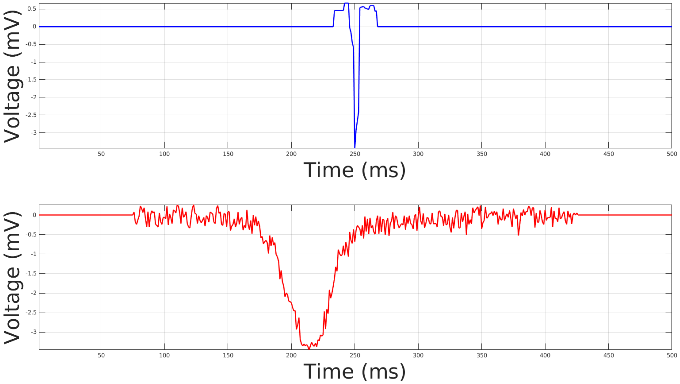


Figure: **TOP:** a tire revolution on normal surface. **BOTTOM:** a tire revolution on abnormal surface.

Tire sensor signal processing

Adaptive Hermite functions

- Classical Hermite polynomials: $\{h_k \mid k \in \mathbb{N}\}$.
- Hermite functions:

$$\varphi_k(t) = h_k(t) \Big/ ||h_k||_2 \cdot \sqrt{w(t)} \qquad (k \in \mathbb{N}),$$

  where $w(t) = e^{-t^2}$.
- Adaptive Hermite functions:[1]

$$\varphi_k^{(\tau,\lambda)}(t) := \sqrt{\lambda}\varphi_k(\lambda(t-\tau)) \qquad (t, \tau \in \mathbb{R}, \lambda > 0).$$

- VP-layer: $G_{(\tau,\lambda)}(\boldsymbol{f}) := \boldsymbol{f} - \Phi(\tau,\lambda)\Phi^{+}(\tau,\lambda)\boldsymbol{f}$

[1] T. Dózsa and P. Kovács, "ECG Signal Compression Using Adaptive Hermite Functions," ADVANCES IN INTELLIGENT SYSTEMS AND COMPUTING, vol. 399, pp. 245–254, 2015.

Tire sensor signal processing

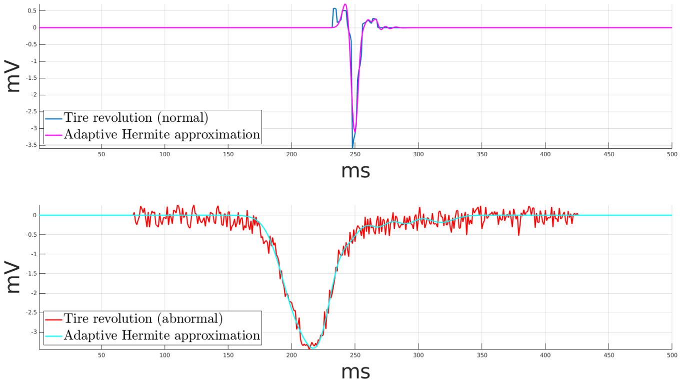


Figure: Adaptive Hermite approximations of tire signals

Tire sensor signal processing

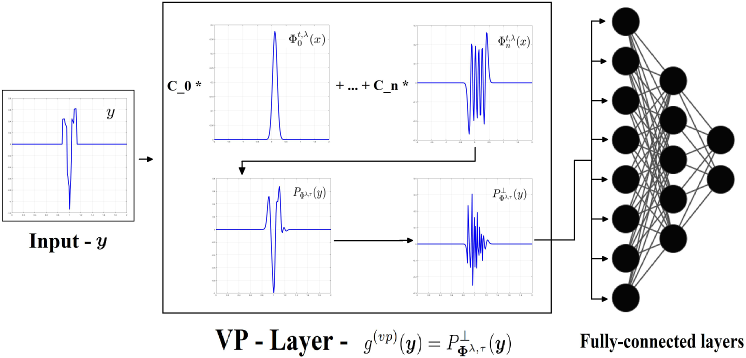


Figure: Proposed VP-NET architecture for road surface abnormality recognition

Tire sensor signal processing

Results

- Sensor: Nanosensors Laboratory, MFA
- Test vehicle: Nissan Leaf, SZTAKI-SCL
- Data: 282 normal and 235 abnormal tire revolutions.

Algorithm	Accuracy (on the test set)
SVM	94.23%
CNN	97.12%
FCNN	97.11%
VPNet (proposed)	**98.08%**

Table: Road surface abnormality recognition[1]

[1] T. Dózsa et. al., "Road Abnormality Detection Using Piezoresistive Force Sensors and Adaptive Signal Models," in IEEE Transactions on Instrumentation and Measurement, (2022)

Current research directions

Theoretical considerations

- Variable projection operators for other ML models (e.g. SVM[1] [2] and spiking networks[3]).
- Interpretable transformations using other new frameworks (e.g. hyperbolic convolution operators).

Applications

- Real-time road surface abnormality recognition.
- Wheel force estimation based on tire sensor signals.

---

[1] T. Dózsa and P. Kovács, Variable projection support vector machines, Proc. 4th International Conference on Advances in Signal Processing and Artificial In- telligence (ASPAI), (2022)

[2] T. Dózsa F. Deuschle, B. Cornelis, P. Kovács, "Variable projection support vector machines and some applications using adaptive Hermite expansions", International Journal of Neural Systems (2023) (Under review)

[3] P. Kovács, and K. Samiee. "Arrhythmia Detection Using Spiking Variable Projection Neural Networks." Computing in Cardiology (CinC) (2022)

VP-SVM

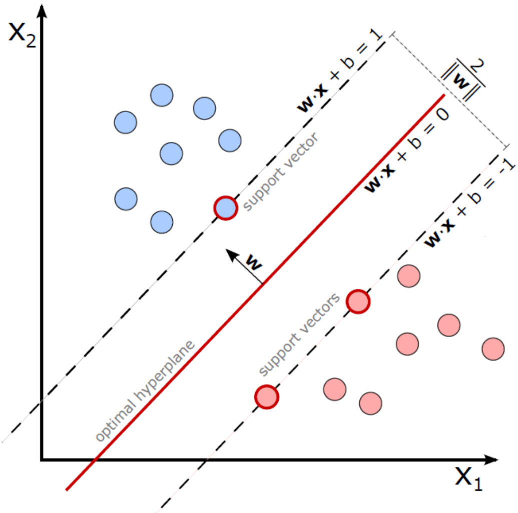
Support vector classification

- Suppose $G : \mathbb{R}^n \supset \mathcal{X} \to \mathcal{Y} := \{-1, 1\}$.
- A Support Vector Machine (SVM) aims to identify an optimal hyperplane separating the examples in $\mathcal{X}$.
- $G_\theta(\boldsymbol{f}) := \mathrm{sgn}(\boldsymbol{w}^T \boldsymbol{f} + b) \quad (\theta := [\boldsymbol{w}, b] \in \mathbb{R}^{n+1}, \boldsymbol{f} \in \mathcal{X} \subset \mathbb{R}^n)$.
- Training (soft-margin SVC):

$$
\begin{aligned}
\min_{\boldsymbol{w}\in\mathbb{R}^n, b\in\mathbb{R}} \quad & \frac{1}{2}\|\boldsymbol{w}\|^2 + \sum_{j=1}^{q} \xi_j, \\
\text{subject to} \quad & y_k(\boldsymbol{w}^T x_k + b) \geq 1 - \xi_k, \\
& \xi_k \geq 0 \quad (k = 1, \ldots, q).
\end{aligned}
$$

- Convex optimization can be used to solve for $\boldsymbol{w}$ and b.

VP-SVM


X2
X1
w·x + b = 1
w·x + b = 0
w·x + b = -1
2/‖w‖
w
support vector
support vectors
optimal hyperplane

VP-SVM

Unconstrained formulation

- Unconstrained SVM training formulation for a linearly separable $\mathcal{X}$:

$$\min_{\boldsymbol{w}\in\mathbb{R}^n, b\in\mathbb{R}} C\cdot\sum_{i=1}^{q}\max(0, 1-y_i(\boldsymbol{w}^T\boldsymbol{f}_i+b))+\|\boldsymbol{w}\|_2^2, \ (C\in\mathbb{R}).$$

- Can solve for $\boldsymbol{w}$ and b using (sub)gradient methods.
- Other efficient optimization algorithms exist.[1]

[1] J. Shawe-Taylor, and S. Shiliang. "A review of optimization methodologies in support vector machines." Neurocomputing 74.17 (2011)

VP-SVM

Variable Projection Support Vector Machines

- Suppose

$$G_\theta(\boldsymbol{f}) := \operatorname{sgn}(\boldsymbol{w}^T(\Phi(\boldsymbol{\eta})^+\boldsymbol{f}) + b)$$
$$(\theta := \{\boldsymbol{w}, \boldsymbol{\eta}, b\}, \boldsymbol{w} \in \mathbb{R}^m, \boldsymbol{\eta} \in \mathbb{R}^P, b \in \mathbb{R}).$$

- VP-SVM training objective:

$$C \sum_{i=1}^{q} \max(0, 1 - y_i(\boldsymbol{w}^T(\Phi^+(\boldsymbol{\eta})\boldsymbol{f}_i) + b) +$$
$$\|\boldsymbol{w}\|_2^2 + R(\boldsymbol{\eta}),$$

where $\boldsymbol{w} \in \mathbb{R}^n$ and $R(\boldsymbol{\eta})$ is an added regulatory term:

$$R(\boldsymbol{\eta}) = \frac{\alpha}{q} \sum_{i=1}^{q} \frac{\|\boldsymbol{f}_i - \Phi(\boldsymbol{\eta})\Phi(\boldsymbol{\eta})^+\boldsymbol{f}_i\|_2^2}{\|\boldsymbol{f}_i\|_2^2} \quad (\alpha \in \mathbb{R}).$$

VP-SVM

VP-SVM properties

- (Sub)gradient based methods can be used for training.
- $R(\eta)$ prevents the problem of vanishing gradients.
- More suitable for light-weight applications (less parameters than VP-NET).
- Can be used with Mercer kernels as well.[1] [2]

[1] T. Dózsa and P. Kovács, Variable projection support vector machines, Proc. 4th International Conference on Advances in Signal Processing and Artificial In- telligence (ASPAI), (2022)

[2] T. Dózsa F. Deuschle, B. Cornelis, P. Kovács, "Variable projection support vector machines and some applications using adaptive Hermite expansions", International Journal of Neural Systems (2023) (Accepted)

Sensor fault detection

Problem description

- **GOAL:** identify peaks in accelerometer measurements appearing due to hardware failure.
- Sudden peaks can appear due to so-called shock events. These have similar morphology to sensor peaks.
- Peaks due to sensor failure and physical phenomena may overlap.

Collaboration

SZTAKI

ELTE
EÖTVÖS LORÁND
UNIVERSITY

SIEMENS

Sensor fault detection

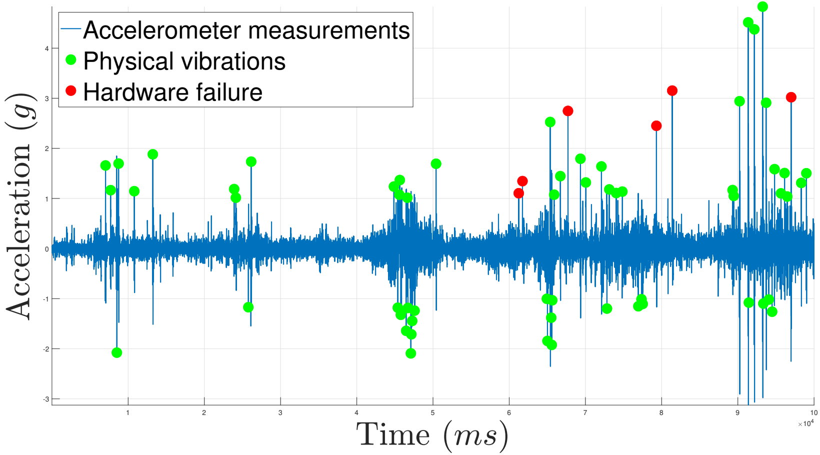


Figure: Accelerometer data to be classified

Sensor fault detection

Classification with VP-SVM

- **Difficulties and observations:**
  - Normal/abnormal examples have similar morphology.
  - Highly unbalanced dataset.
  - Examples have compact support, can be modelled efficiently with adaptive Hermite functions.
- **Methodology and preprocessing**
  - Transform measurement by truncated scalogram using complex Morlet wavelets.
  - Downsample normal examples: small training set, large test set.
  - Use VP-SVM with Gaussian kernel and adaptive Hermite functions.

Sensor fault detection

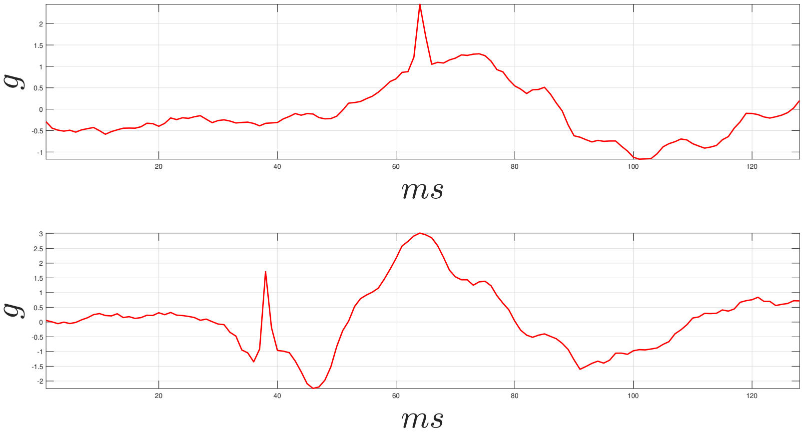


Figure: Peaks due to sensor failure can appear near vibration induced peaks

Sensor fault detection

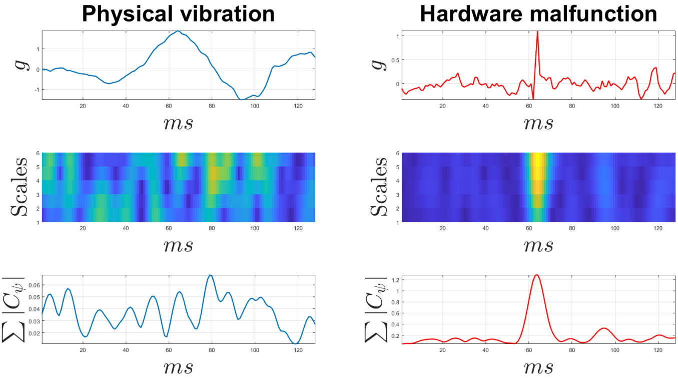


Figure: Preprocessing steps before classification with VP-SVM

Sensor fault detection

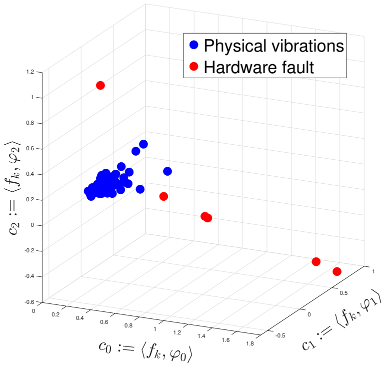


Figure: Output of the optimized adaptive Hermite transformation in the trained VP-SVM.

Conclusion

Summary

- Discussed variable projections in Hilbert spaces.
- Extension of neural networks with variable projection layers for interpretable feature extraction.
- Generalization of framework to other ML algorithms (e.g. SVM)
- Example applications:
  - Road abnormality detection using tire sensor signals.
  - Sensor fault detection in accelerometer measurements.

Conclusion

Thank you for your attention