Regressió sparse per les dades de una llengua electrònica

Universitat Pompeu Fabra Barcelona

Xavier De La Fuente Quintana

Índex

Introducció al contingut general del treball

- Context
- Experiment
- Objectius i Metodologia

Anàlisis teòric

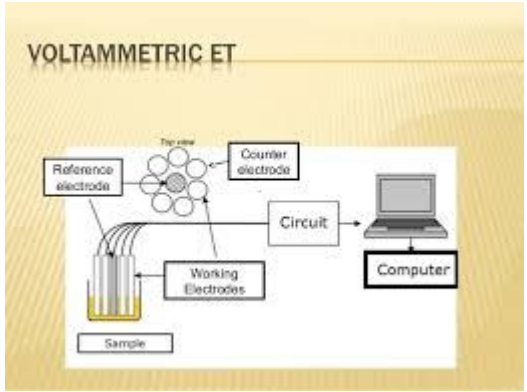
Anàlisis de les dades

Resultats

Conclusions

Introducció al contingut del treball

- Necesitat de mètodes
  - eficaços
  - ràpids
  - gran volum de dades



llengua electrònica + anàlisis dades

Machine learning: Regressió

Sensors: modificar i optimitar

Experiment

Rendiment de una llengua electrònica concreta

- Solució concreta es van disoldre 3 aminoàcids:
  - Tryptophan (Trp)
  - Tyrosine (Tyr)
  - Cysteine (Cys)
- Proces voltamètric: -1 V fins a 1.2 V, a 0.1 $V\,s^{-1}$
- Llengua electrònica: 1è de referència, 1è auxiliar i 5è (4+1)
- Dades: 42 experiments de 556 mostres/experiment (5 sensors)

→ Alta correlació → Mètodes Regressió Sparse

Objectius

- Analitzar les dades a través de mètodes de regressió:

  model predictiu de concentracions de aminoàcids

  - Variational Garrote: l0 norm
  - Lasso: l1 norm

Metodologia

Dades sintètiques

Dades reals de l'experiment

Anàlisi teòric

- Machine learning : Regressió
  - Què és la regressió?

$$y = \alpha + Xv + \xi \quad \Big| \quad \begin{array}{l} X_{nxp} \\ y_{nx1} \\ v_{nx1} \end{array}$$

$$y = Xv \longrightarrow \hat{v} = (X^T X)^{-1} X^T y$$

Casos

- $n > p$

  Sistema sobredeterminat.

  Solució múltiple → Solució óptima

- $n = p$

  Sistema determinat.

  Solució única

- $n < p$

  Sistema subdeterminat.

  Solució inconsistent → No té solució

Problemes

Overfitting

- n<p → Model massa complex
- Sobreactuar a petits cambis com el soroll

Possibles solucions

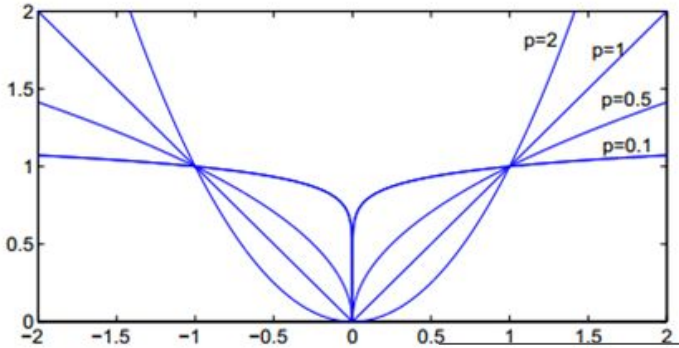
Regularització

$$min \sum_{i=1}^{n} V(f(\hat{x}i), \hat{y}i) + \lambda R(f)$$

λ : Quantitat de regularització

Major error d'entrenament, menor error en test

Validació Creuada



$$l_i = ||x||_p = \sum_{i=1}^{N} (|x_i|^p)^{1/p}$$

Lasso

Donada la funció:

$$y^{\mu} = \sum_{i=1}^{n} \alpha + w_i x_i + \varepsilon^{\mu}$$

Definim les dades D:{x, y} = 1,...,p. La lasso regularitzada serà:

$$min_{\beta,\alpha} \sum_{\mu} (y^{\mu} - \sum_{i=1}^{n} \alpha + w_i x_i) \quad subject\ to \quad ||\beta||_1 \leq t$$

On penalitzem β i α pot prendre qualsevol valor

Garrote

Donada la funció:

$$y^{\mu} = \sum_{i=1}^{n} w_i s_i x_i + \varepsilon^{\mu} \qquad \sum_{i=1}^{n} s_i \leq t \quad \text{on } s_i = 0, 1$$

El model predictiu és: $y = \sum_i m_i w_i x_i + \varepsilon^{\mu}$

Definim les dades D:{x, y} = 1,...,p. El Variational garrote serà:

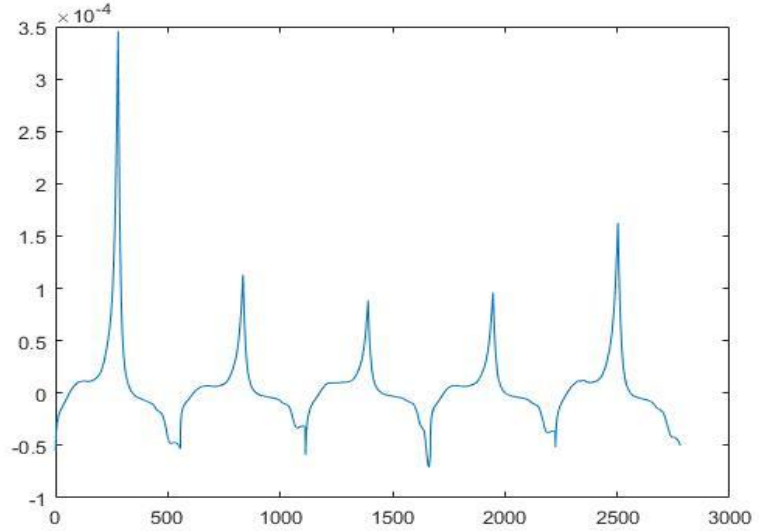
$$min \sum_{\mu} (y^{\mu} - \sum_{i=1}^{n} m_i w_i x_i) \quad subject\ to \quad \sum_i m_i \leq t$$

Anàlisi de les dades

- Dades sintètiques

$$y^{\mu} = \sum_i \hat{w}_i \hat{x}_i^{\mu} + d\,\xi^{\mu}$$

- Concentracions reals



	Train set (10^{-4} M)																				
#	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	21
TRP	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.5	0.5	0.5	0.5	0.5	0.5	0.5	0.5	0.5	3.0	3.0	3.0
TYR	0.0	0.0	0.0	0.5	0.5	0.5	3.0	3.0	3.0	0.0	0.0	0.0	0.5	0.5	0.5	3.0	3.0	3.0	0.0	0.0	0.0
CYS	0.0	0.5	3.0	0.0	0.5	3.0	0.0	0.5	3.0	0.0	0.5	3.0	0.0	0.5	3.0	0.0	0.5	3.0	0.0	0.5	3.0

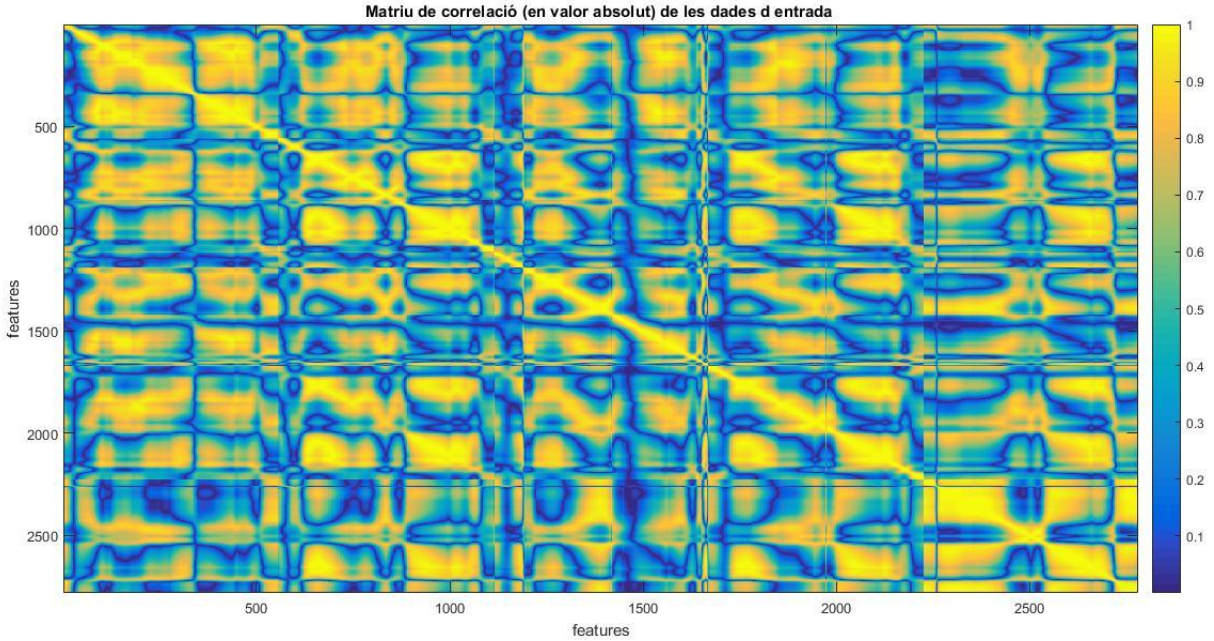
	Train set (10^{-4} M)						Test set (10^{-4} M)														
#	22	23	24	25	26	27	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15
TRP	3.0	3.0	3.0	3.0	3.0	3.0	2.7	2.7	1.3	2.7	1.3	1.3	1.3	0.0	0.7	0.7	1.7	0.7	0.7	1.3	0.3
TYR	0.5	0.5	0.5	3.0	3.0	3.0	2.7	2.3	1.3	1.7	1.3	2.3	2.7	2.3	1.3	1.3	2.3	0.0	0.0	2.0	2.3
CYS	0.0	0.5	3.0	0.0	0.5	3.0	1.7	1.3	0.7	1.7	2.0	1.3	2.3	2.7	2.7	0.3	2.7	0.7	0.0	1.3	1.7

Matriu de correlació

Valor absolut

Alta correlació

>0.5, >0.7



Resultats

Els paràmetres que s'han extret per tal de analitzar els resultats són:

- MSE entrenament: $MSE = \frac{1}{N}\sum_{i=1}^{N}(\hat{Y}_i - y_i)^2$
- MSE test:
- Sparsity de la solució: $\text{count}\left(\frac{\text{número } w \neq 0}{\text{número } w}\right)$
- Suma de la solució: $\text{sum}\left(\frac{\text{número } w \neq 0}{\text{número } w}\right)$
- Hiperaràmetre: $\lambda \ o \ \gamma$

- Mitja aritmètica:

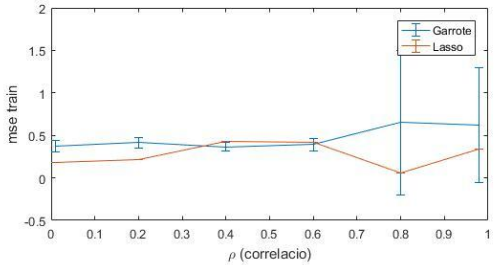
$$\bar{x} = \frac{1}{N}\sum_{i=1}^{N} x_i$$

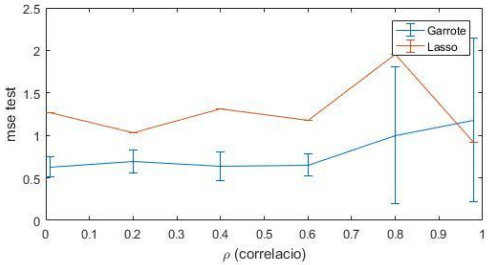
- Desviació estàndard:

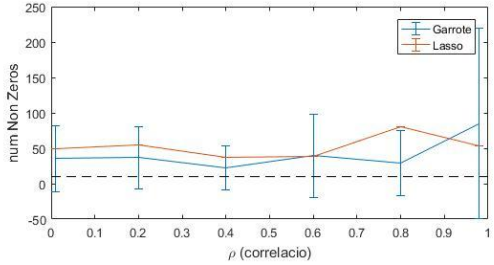
$$\sigma = \sqrt{\frac{\sum_{i=1}^{N}(x_i - \bar{x})^2}{N-1}}$$

Dades sintètiques

Com afecta la correlació als nostres mètodes?



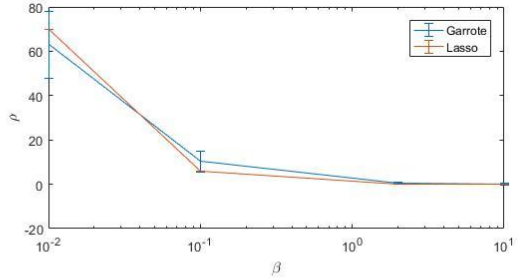


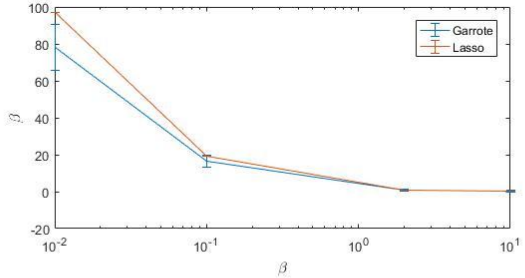


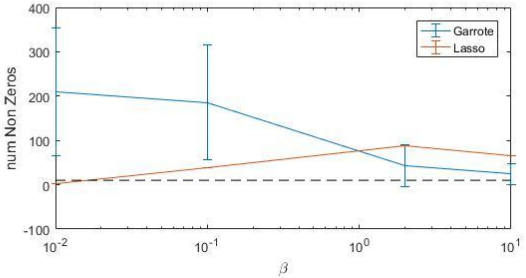
$\beta = 2;\ \delta = 10;\ n_{train} = 100;\ n_{test} = 50$

$\rho = 0.01,\ 0.2,\ 0.4,\ 0.8,\ 0.98$

I el soroll?



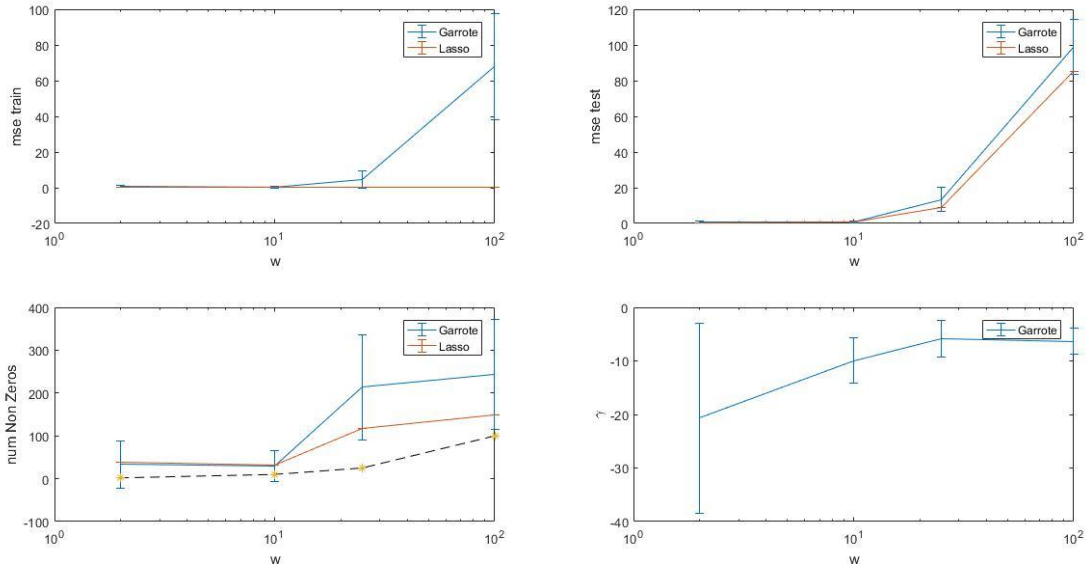




$\delta = 10;\ n_{train} = 100;\ n_{test} = 50;\ \mu = 0.5$

$\beta = 10,\ 2,\ 0.1,\ 0.01$

I la densitat de la solució que busquem?



$\beta = 2;\ n_{train} = 100;\ n_{test} = 50;\ \mu = 0.5$

$\delta = 2,\ 10,\ 25,\ 100$

Concentracions reals: Entrenament i test paper

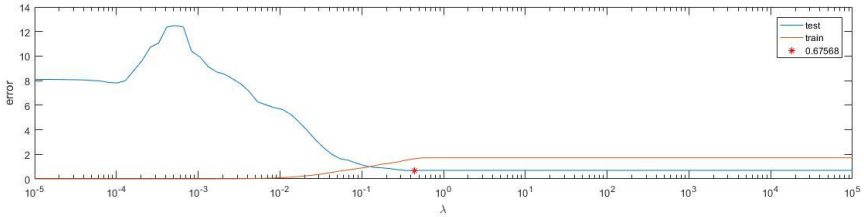
Lasso

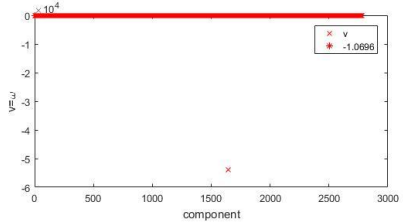
	TRP		TYR		CYS	
	mitja	d.e.	mitja	d.e.	mitja	d.e.
MSE_train	1.6421	0	1.7222	0	1.3646	0
MSE_test	0.6757	0	0.9931	0	0.7188	0
NºZeros	1	0	0	0	7	0
sum(v)	-53995	0	0	0	1376900	0
Gamma	3212.7	13040	3212.8	13040	3212.7	13040

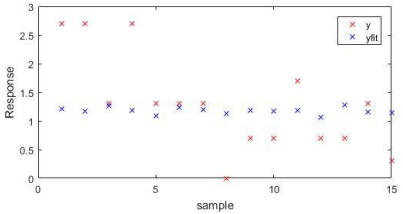
Variational Garrote

	TRP		TYR		CYS	
	mitja	d.e.	mitja	d.e.	mitja	d.e.
MSE_train	1.3361	0.6470	1.4181	0.3962	1.0000	0.3800
MSE_test	1.2886	0.3596	1.2785	0.3051	0.9973	0.1626
NºZeros	730.5500	1031.1	368.9474	497.5114	904.6000	730.0285
sum(v)	-0.3001	0.6015	-0.4801	0.5271	0.7605	0.8959
Gamma	-8.3655	6.3533	-9.3200	5.8160	-6.4765	6.3826

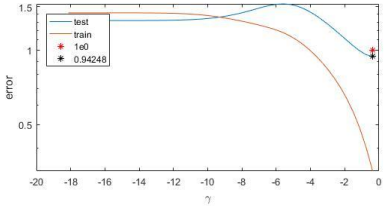
Concentracions reals: Entrenament i test paper

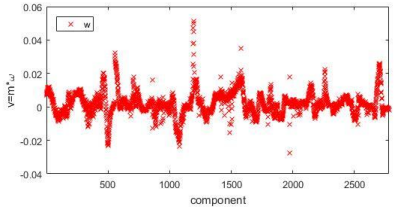




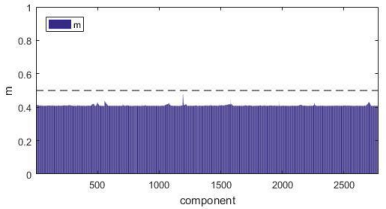


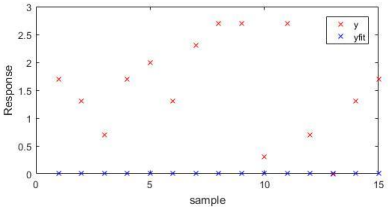
Lasso





Variational Garrote





Concentracions reals: Entrenament i testeig del conjunt d'entrenament

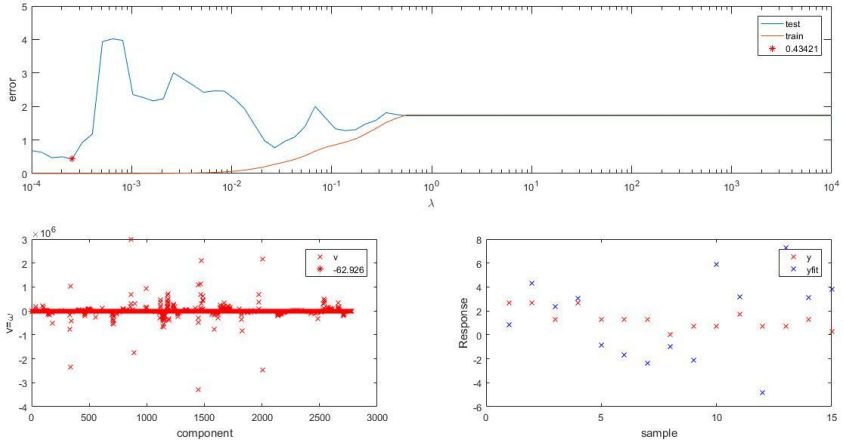
Lasso

	TRP		TYR		CYS	
	mitja	d.e.	mitja	d.e.	mitja	d.e.
MSE_train	1.6641	0	1.4112	0	1.7222	0
MSE_test	2.4609	0	2.5182	0	1.7222	0
NºZeros	0	0	3	0	0	0
sum(v)	0	0	-4.0761e+05	0	0	0
Gamma	506.2241	16110	506.1681	16110	506.2241	1611

Variational Garrote

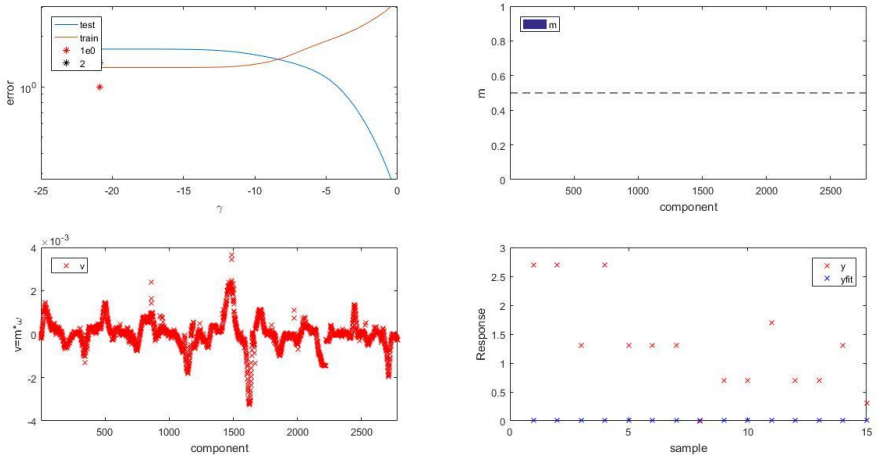
	mitja	d.e.	mitja	d.e.	mitja	d.e.
MSE_train	1.3099	0.1621	1.2717	0.1821	1.0665	0.3283
MSE_test	0.8011	0.2475	0.9824	0.2609	0.5699	0.2432
NºZeros	614.1704	266.8825	690.6593	318.0390	1035.1	459.4942
sum(v)	-0.1638	0.1093	-0.7460	0.3096	0.6496	0.3266
Gamma	0.1621	1.9896	-9.2414	2.6956	-7.3952	3.1205

Concentracions reals: Entrenament i testeig del conjunt d'entrenament



Lasso

Variational Garrote



Concentracions reals: Validació Creuada

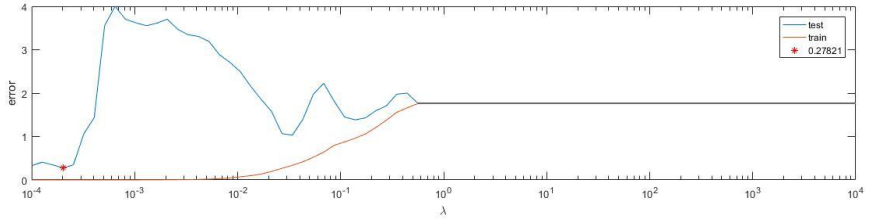
Lasso

	TRP		TYR		CYS	
	mitja	d.e.	mitja	d.e.	mitja	d.e.
MSE_train	9.7119	3.5226	9.7451	4.0346	9.9006	3.3771
MSE_test	11.6773	4.3039	12.6444	3.9357	10.9502	3.3298
NºZeros	0.6667	1.9149	2.8667	5.7055	1.8000	3.0284
sum(v)	404160	14026000	-403570	2055000	306440	155050
Gamma	506.3613	1611	506.3266	1611	506.3446	1611

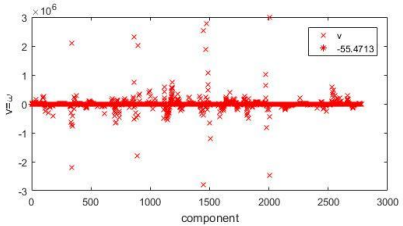
Variational Garrote

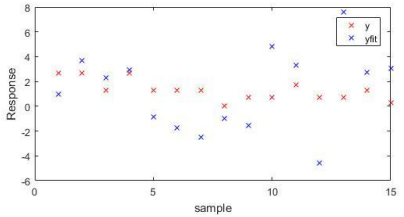
	TRP		TYR		CYS	
	mitja	d.e.	mitja	d.e.	mitja	d.e.
MSE_train	1.0694	0.1382	1.3558	0.2579	1.3645	0.2371
MSE_test	0.7374	0.2342	0.8733	0.2800	0.9787	0.2597
NºZeros	1.0559e+03	230.0969	713.1676	430.1101	669.0102	341.1572
sum(v)	0.5208	0.2324	-0.7036	0.3794	-0.2170	0.1723
Gamma	-6.9758	1.5908	-9.2647	2.9963	-9.4159	2.8607

Concentracions reals: Validació Creuada

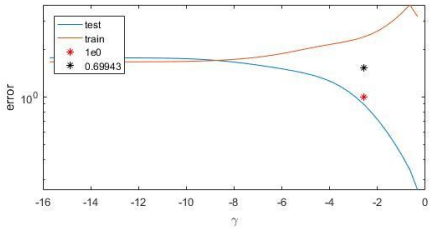


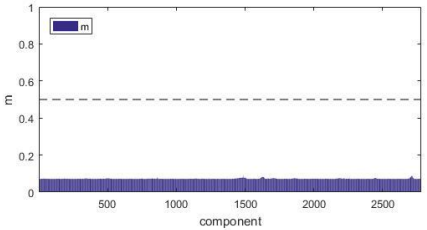
Lasso

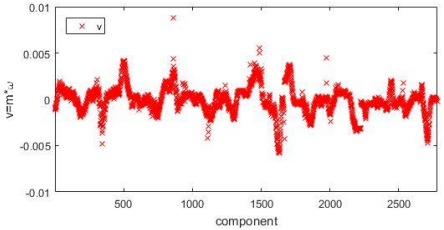


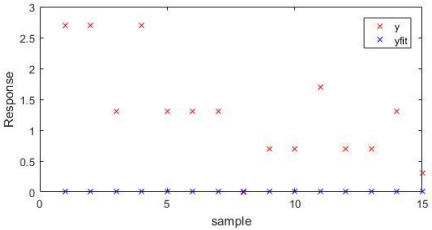


Variational Garrote









Conclusions

- Anàlisis de les dades sintètiques ✓
- Anàlisis de les dades reals ✗
  - Dades ‘estàndard’ ✗
  - Dades conjunt d’entrenament ✗
  - Dades validació creuada ✗

Perquè?

- Pròpia naturalesa de les dades
- Llengua ofereix unes mesures no molt bones
- Alta correlació en les dades → nealitat
- Rang de valors del conjunt d'entrenament vers el conjunt de test

Conclusió:

El rendiment dels mètodes degenera amb les dades reals

Future work

- Entendre bé les característiques de les dades
  - ❏ Principal Component Analysis (PCA)

- Sensors redundants
- Modificar algoritmes per per buscar solucions més denses
- Aplicar un altre tipus d'anàlisis o model:
  - ❏ Classificació

Gràcies