Regressió linear sparse per les dades de una llengua electrònica

Xavier De La Fuente Quintana

Curs 2016-2017

Director: Gerard Sanroma, Vicenç Gómez

GRAU EN ENGINYERIA EN SISTEMES AUDIOVISUALS

upf. Universitat Pompeu Fabra Barcelona
Escola Superior Politècnica

Treball de Fi de Grau

Agraïments

Agraeixo aquest treball als meus dos tutors del treball, que m’han ajudat i suportat durant tot el treball.

Resum

Aquest projecte tracta el problema de la regressió lineal amb un nombre petit de mostres i un alt nombre de predictors. Primerament revisa algoritmes recents que prioritzen la *sparsity* de la solució per a solucionar aquest problema com regressió per Lasso i regressió per *L0-norm* i posteriorment presenta els resultats de la seva aplicació a un conjunt de dades de una llengua electrònica.

Resum

This project deals with the problem of linear regression with a small number of samples and a large number of predictors. It first reviews recent algorithms that use a sparsity prior to overcome this problem such as Lasso and L0-norm regression and then presents results of their application to a dataset from the Electronic Tongue

Pròleg

El treball es basa en fer l'anàlisi d'unes dades preses per una llengua electrònica concreta davant una sèrie de concentracions donades per tal de generar un model de regressió capaç d'entendre les dades. La finalitat d'aquest model és que sigui capaç de predir concentracions concretes de certs aminoàcids sense saber la concentració real. L'experiment i les dades preses per la llengua es van dur a terme per part del departament de química de la UAB.

És important trobar algorismes d'anàlisis de dades en camps com l'anàlisis de substàncies químiques, ja que l'alternativa als algorismes és manipular químicament mostra a mostra per tal de poder mesurar finalment els resultats. La presència de una llengua electrònica capaç d'analitzar aquestes substàncies totes de cop es presenta com una eina molt potent.

En aquest treball es treballarà amb dos mètodes de regressió lineal que implementen regularització amb la fi de trobar una solució *sparse* al problema: *Lasso* i *Variational Garrote*. Ens interessa aplicar mètodes de regressió *sparse* perquè no tenim una gran quantitat de mostres (quantitat d'experiments) respecte el nombre de predictors (variables captades cada experiment).

Aquests mètodes s'aplicaran per analitzar dades sintètiques generades per nosaltres i que modificarem nosaltres amb la finalitat de veure en quins casos els mètodes funcionen pitjor. Aquest coneixement ens ajudarà a entendre els resultats alhora d'aplicar els mètodes amb les dades reals, les quals analitzarem en diferents casos: anàlisi segons la distribució que es va fer en l'experiment que ens hem basat, anàlisi amb validació creuada i anàlisis modificant la resposta real.

Índex

1. INTRODUCCIÓ

1.1 Context

Actualment, hi ha una gran demanda alhora de determinar concentracions d'aminoàcids en certs àmbits, com per exemple l'anàlisi clínic, la indústria alimentària o la farmacèutica. S'utilitzem tècniques costoses, lentes o complicades alhora de fer aquest anàlisi, com la cromatografia, que consisteix a separar químicament els components de la solució amb l'aminoàcid per tal de saber la seva concentració; o la electrofèresi capil·lar, una altra tècnica química per separar les concentracions de la solució.

Les llengües electròniques són sensors barats que ens permeten prendre informació sobre l'activitat elèctrica de substàncies. Un mètode alternatiu que permetria reduir temps i costos d'una manera més simple i ràpida.

Una llengua electrònica és un dispositiu que mesura y compara sabors a través d'uns sensors que capten els mateixos compostos, orgànics e inorgànics, que captaria una llengua humana. Els sensors que conformen la llengua no acostumen a ser tots iguals, de la mateixa manera que els receptors gustatius dels humans tampoc ho són. Aquests sensors són més barats i ens donen informació sobre la activitat elèctrica de substàncies. A través de l'anàlisi d'aquesta informació de la activitat elèctrica es pot deduir les seves propietats químiques.

Els materials que conformen els sensors poden ser modificats per tal d'obtenir una major sensibilitat o introduir nous predictors, oferint així la possibilitat d'augmentar l'espectre de aminoàcids (AA) detectables.

No obstant, la llengua electrònica no és més que un conjunt de sensors. És necessari doncs acompanyar aquests amb algoritmes d'anàlisis computacional. Amb l'ús d'algorismes, es pot reduir de manera dràstica la velocitat de càlcul d'un conjunt de concentracions: mentre que amb l'anàlisi químic s'ha de dur a terme per a cada mostra d'un experiment, amb un model analític ben entrenat es podria predir un gran nombre de concentracions correctament a la vegada, necessitant únicament que la llengua prengui les dades necessàries de la substància.

Una bona manera d'analitzar aquestes dades és a través de models de regressió. No obstant, perquè un mètode de regressió funcioni bé es necessitarà un gran nombre de mostres. Per tal de reduir la quantitat d'experiments per dur a terme aquest anàlisi es poden aplicar models de regressió *sparse*.

Aquests models es generen a partir d'un entrenament amb dades conegudes. Les dades seran: la resposta real, que correspon a la concentració real del aminoàcid que s'ha posat a la solució $y \in R^n$ (n experiments) i uns predictors $X = (X_1, X_2, ..., X_p) \in R^{nxp}$ (predictors de p-dimensions per cada experiment). És generarà un model lineal $y \approx X\widehat{w}\ on\ \widehat{w} \in R^p$. La finalitat és trobar un model $\widehat{w}$ que donades unes noves dades $X' = (X'_1, X'_2, ..., X'_q) \in R^{mxq}$ sigui capaç de determinar les seves corresponents $y' \in R^m$ amb una precisió acceptable..

No obstant, amb pocs experiments i molts predictors no es pot aplicar mètodes de regressió lineal directament. En aquests casos la matriu $(X^T X)$ necessària per obtenir el model no és invertible i per tant la solució y no està definida. Els models de regressió *sparse* ofereixen una alternativa en aquests casos en que $(X^T X)$ no és invertible sota l'assumpció que no tots els predictors són rellevants.

En aquest treball s'aplicaran models de regressió *sparse* (Variational Garrote i Lasso) amb l'objectiu d'intentar entendre i posteriorment predir les dades que va mesurar la llengua electrònica en l'experiment, explicat en el següent apartat, amb la finalitat de generar un model que sigui capaç de discernir la concentració d'una substància amb un error acceptable.

1.1.1. Estat de l'art en el processament de dades de llengües electròniques

En aquest document és pretén aplicar models de regressió sobre unes dades preses en un experiment del departament de química de la Universitat Autònoma de Barcelona (UAB). Primerament s'explcarà breument l'experiment a partir de l'informe desenvolupat pels mateixos membres.

Experiment

La finalitat de l'experiment era veure el rendiment que podia oferir una llengua electrònica concreta, juntament amb un model que interpreti les dades amb un mínim de sentit.

L'experiment consistia en mesurar la resposta d'una llengua electrònica davant d'una solució de 0.1 M de monohidroxenfosfat de potassi, 0.1 M de dihifrogenfosfat de potassi i 0.1 M de KCl. A aquesta solució se li va dissoldre una concentració específica de tres aminoàcids: el tryptophan (TRP), el tyrosine (TYR) i la cysteine (CYS). La solució es va ajustar a un pH de 7,5.

Material de l'experiment i mesures:

Llengua electrònica:

La llengua electrònica que van utilitzar en l'experiment consistia en una cèl·lula voltamètrica conformada per un elèctrode de referència, un elèctrode auxiliar i cinc elèctrodes operatius; aquests darrers seran els que prendran les mesures que analitzarem.

Sensors:

Dels cinc sensors que composaven la llengua electrònica, quatre corresponien a sensors compostos de *'bulk-modified epoxy-graphite*' i l'altre a un compost de '*epoxy-graphite*' modificat.

Procés voltamètric:

Les mesures es van prendre a temperatura ambient. Van ser mesurades amb un rang de $-1\ V$, fins a $1,2\ V$, a un *scan rate* de 0.1 $0.1\ V s^{-1}$.

Un cop coneguts el disseny de l'experiment i el seus components, les dades que es van prendre van ser les següents:

1.1.2. Característiques de les dades

Els experiments es van dur a terme un total de 42 vegades: on les 27 primeres corresponen al conjunt d'entrenament i les altres 15 al conjunt de test. Aquest

fet és va elegir a priori i, conseqüentment, es va decidir un conjunt de valors diferents en ambdós casos.

Per al conjunt d'entrenament es va optar per agafar tres valors de concentració específics: $0,0\,M$, $5{\cdot}10^{-5}\,M$ i $3{\cdot}10^{-4}\,M$. Les 27 proves d'entrenament corresponen a totes les possibles combinacions d'aquests tres valors entre ells, i sent cada una d'aquestes combinacions d'un experiment diferent.

Per al conjunt de test, no obstant, els valors de les concentracions no eren fixes, sinó valors aleatoris dins del rang de les concentracions del conjunt d'entrenament ($[0,0\,M,\ 3{\cdot}10^{-4}\,M]$). No obstant, per a aquest conjunt es van trobar valors únics i que no es repeteixen entre experiments o entre concentracions.

La llengua electrònica té una característica a tenir en compte: els sensors són poc selectius amb una especificitat especial per a cada concentració d'un aminoàcid en la solució.

Altrament, s'observa com les dades estan altament correlacionades, la qual cosa proveeix la possibilitat de generar un model de regressió. No obstant, una alta correlació en les dades pot donar colinealitat, fet que pot desencadenar problemes d'estabilitat en el rendiment de l'algorisme.

De la mateixa manera, cadascun d'aquests experiments consisteix en 2780 mostres preses durant l'experiment, que corresponen a 556 mostres per cada sensor. Es veu com el nombre de predictors (2780) supera considerablement el nombre de experiments que utilitzarem per a entrenar (27), fet que resulta problemàtic alhora d'aplicar mètodes simples de regressió, com veurem més endavant.

1.1.3. Motivació per fer servir els mètodes *sparse*

Alhora d'analitzar dades a través de regressió cal tenir en compte la mida de les dades amb les que e treballarà. Donada una resposta $y \in R^n$ (n experiments) i uns predictors $X = (X_1, X_2, ..., X_p) \in R^{nxp}$ (predictors de p-dimensions per cada experiment). Nosaltres volem utilitzar un model lineal $y \approx X\widehat{w}\ on\ \widehat{w} \in R^p$ i la solució al problema és directament $\widehat{w} = (X^T X)^{-1} X^T y$.

Sempre que es pugui invertir $(X^T X)$, el problema tindrà només una solució, i els podrà resoldre a través de l'equació anterior. Aquest cas acostuma a passar quan $n > p$, ja que la solució està ben definida.En els casos en que no es pugui invertir $(X^T X)$, s'haurà d'aplicar altres mètodes amb la finalitat de resoldre el problema. Anomenarem matriu degenerada a la matriu no invertible $(X^T X)$.

El problema es troba quan $n << p$. En aquest cas no és possible invertir la matriu $X^T X$, ja que la matriu no es quadrada. Per poder resoldre el problema en aquest cas, s'assumirà que només uns pocs predictors seran rellevants per poder determinar la variable resposta. Quan apliquem aquesta hipòtesis a priori, s'aconsegueix que el problema estigui ben definit i, per tant, es podrà resoldre.

L'objectiu de buscar una solució *sparse* a un problema de regressió és trobar un model $\widehat{w}$ amb molt coeficients amb valor zero. Per tal d'aplicar aquest coneixement a priori, o aquesta restricció, s'aplica una funció de penalització $\lambda R(f)$ a les dades $V(f(\widehat{xi}),\widehat{yi})$ en forma de $min \sum_{i=1}^{n} V(f(\widehat{xi}),\widehat{yi}) + \lambda R(f)$, on $R(f)$ és la funció de penalització. Les funcions *l-norma* són molt comuns com a funcions de penalització. Dos funcions de penalització molt utilitzades són la norma 1, el qual és coneix amb el nom de lasso, i la norma 0, que s'anomena *Spike-and-Lab*. El Variational Garrote és un algorisme proposat recentment per a solucionar el problema de l'*Spike-and-lab.*

Els resultats existents utilitzen les tècniques de *Partial Least Squares (PLS)* i xarxes neuronals. Les xarxes neuronals no són mètodes de regressió sparse, però el *PLS* si és un mètode de regressió. Consisteix en reduir el nombre de predictors a un altre conjunt més petit de dades descorrelades per poder aplicar-hi *Least Squares Regression*. Així doncs, hem volgut implementar un parell de mètodes de regressió per tal de veure si podíem arribar a entendre les dades amb regressió.

1.2 Objectius

L'objectiu d'aquest Treball de Fi de Grau és analitzar la resposta d'una llengua electrònica concreta (un model en concret) saben les concentracions reals utilitzades i les dades corresponents preses per la llengua electrònica. Els resultats van ser obtinguts en un experiment dut a terme per uns membres de la UAB, dels quals només en tenim les mesures presses.

Així doncs, s'avaluaran els resultats a través de diferents mètodes amb la finalitat de trobar el més eficaç alhora de predir el comportament d'aquest tipus de senyals. El focus estarà en implementar dos mètodes de regressió: el Lasso i el Variational Garrote per tal de veure si les dades es poden aproximar amb un model de regressió. Aquest model ha de ser capaç d'entendre la relació entre les dades captades per la llengua electrònica i la concentració real de l'aminoàcid. Si el comportament front les diverses concentracions mostra un patró prou representatiu i es pot aproximar amb un model, es podrà utilitzar el model per predir concentracions reals d'aminoàcids coneixent només les dades de la llengua electrònica.

La finalitat és generar un model que sigui capaç d'interpretar les dades preses per la llengua electrònica i saber la concentració real de l'aminoàcid a través del model de regressió, sense saber el seu valor real, de la manera més fidedigna possible.

Tenint en compte les dades, en les quals tenim molts predictors pel poc nombre d'experiments, s'han escollit aquests dos mètodes de regressió amb l'objectiu de resoldre el problema presentat. Primerament s'ha elegit la lasso, que ofereix penalització amb norma 1, i, a la vegada, també s'ha elegit el *Variational Garrote*, que és un mètode de regressió que millora a la lasso en termes de interpretabilitat ja que utilitza la norma 0 com a funció de restricció [1]. Per altra banda, en els resultats existents utilitzen *PLS*, un mètode que també es basa en la premissa que no tots els predictors són rellevants alhora d'analitzar les dades, sinó que hi ha predictors que guarden una relació lineal entre ells.

Per tal d'entendre els mètodes aplicats, es generarà un conjunt de dades (entrenament i test) i s'avaluarà per a ambdós mètodes. Això és realitzarà diverses vegades alterant el valor de certs paràmetres que afecten al conjunt d'entrenament o a la resposta a predir. Els diferents resultats es recolliran per a avaluar l'efecte d'aquestes variacions en la informació.

Un cop entesos els mètodes i demostrada la seva validesa, s'observarà si també donen bons resultats quan les dades d'entrada són les de l'experiment. Aquestes dades seran avaluades en diferents casos, conforme venen subdividides en entrenament i test; i després utilitzarem el conjunt d'entrenament únicament per tal de generar un model fent validació creuada (en el qual trobarem el nou conjunt d'entrenament i el nou conjunt de test) per posteriorment validar amb el conjunt de test real.

Altrament, s'intentarà veure si les dades presentades es poden explicar d'alguna altra manera més simple o adequada, de manera que facilitin o agilitzin l'anàlisi de les concentracions.

Per tal d'analitzar totes aquestes dades i dur a terme els algoritmes corresponents al còmput dels resultats, s'utilitzarà Matlab R2016b. Les dades de l'experiment van ser preses per uns membres de la UAB i s'utilitzen les presentades en el corresponent paper [2].

1.3 Planificació del projecte

Alhora de dur a terme el projecte, es defineixen tres fases:

La primera fase se centra en entendre els diferents mètodes a aplicar per tal de poder desenvolupar el corresponents algoritmes d'anàlisis, així com les caracteríques de les dades i el problemes que poden comportar.

En la segona part del projecte s'analitzaran els dos mètodes de regressió amb un conjunt de dades sintètiques que hem generat pròpiament nosaltres. Aquest conjunt serà sotmès a certs canvis en els paràmetres que el generen, de manera que puguem avaluar quan els mètodes ofereixen una bona rendiment i quan, per altra banda, baixa el seu rendiment en termes de interpretabilitat o d'error.

En el tercer bloc, aplicarem els mètodes en les dades reals. Generarem un model de regressió i avaluarem la seu rendiment alhora de predir concentracions d'aminoàcids. Aquest model es generarà altrament en diverses condicions amb el fi d'avaluar si millora els resultats quan alterem en certa manera les dades; es modificarà la resposta real i també es farà validació creuada.

1.4 Estructura de la memòria

Capítol 1: Primerament és fa una explicació del context del projecte, explicant les causes y els objectius per els quals s'ha plantejat resoldre'l. Tanmateix s'explica la metodologia que és seguirà per tal de poder assolir aquests objectius així com els motius pels quals s'ha elegit aplicar els diferents

mètodes. Altrament s'explicarà tant la planificació del projecte, com la estructura que mantindrà aquest document.

Capítol 2: Aquest capítol es centra en la part metodològica sobre el que es basarà els anàlisis de les concentracions i de les dades. Primerament explicarem una mica què és la regressió i quins problemes podem trobar alhora d'aplicar-la, així com solucions a aquest problemes tant comuns en aquest tipus d'anàlisi, on tenim molts predictors en les dades, però poques mostres. Explicarem els diversos mètodes de regressió (Lasso i Variational Garrote) implementats per tal de explicar o interpretar les dades presentades i que han generat els resultat que es discutiran més endavant en l'apartat 4.

Capítol 3: En aquest capítol es descriuran les dades presentades per part del document que resumeix l'experiment, així com característiques que puguin ser interessant, com ara la matriu de correlació. També s'explicarà com s'han generat les dades sintètiques utilitzades per a entendre el comportament dels dos mètodes en posteriors capítols.

Capítol 4: En aquest capítol avaluarem els resultats presentats pels dos mètodes de regressió.En la primera part del capítol veurem el seu comportament davant de diversos conjunts de dades sintètiques que hem generat nosaltres mateixos. En la segona part del capítol avaluarem els resultats que corresponen a l'anàlisi de les dades reals. Avaluarem si els resultats són bons o si per el contrari, el model no prediu de manera satisfactòria i el perquè.

Capítol 5: En aquest capítol trobem les conclusions finals en base als resultats obtinguts i l'anàlisi dut a terme en els apartats anteriors. A més a més, és comentarà què és podria haver canviat durant el projecte per haver obtingut millors resultats. Finalment, s'explicarà possible treballa realitzar a partir dels resultats obtinguts i de les conclusions extretes.

2. ANALISI TEÒRIC

2.1 Aprenentatge automàtic

És important saber què és l'aprenentatge automàtic i perquè ens serveix. Primer de tot, hem de saber que és l'aprenentatge.

L'aprenentatge és un procés que consisteix en extreure informació destinada a la adaptació dels organismes. És un canvi permanent en la conducta com a conseqüència de una experiència. En l'aprenentatge automàtic el que és pretén és que mitjançant un algoritme es forci una experiència en l'ordinador de manera que la màquina evoluciona. Més concretament:

Aprenentatge automàtic: un programa d'ordinador APRÈN de l'experiència E, respecte a una classe de tasques T, i una mesura d'eficiència o rendiment P , si la seva eficiència en les tasques T, tal i com les mesura P, s'incrementen o millora gràcies a l'experiència E [3].

L'aprenentatge automàtic és pot utilitzar per a resoldre problemes molt diversos en molts àmbits del coneixement i l'estadística. Els algoritmes són capaços de generalitzar gràcies al reconeixement de patrons i o a la classificació. L'aprenentatge automàtic està molt relacionant amb la estadística, i s'utilitza en àmbits com el llenguatge natural, biomèdica, algoritmes de cerca, robòtica, computer vision…

Dintre de l'anàlisi de dades, l'aprenentatge automàtic és un mètode utilitzat per a dissenyar models y algoritmes predictius. Comercialment, això és coneix com a anàlisis predictiu. Aquest models permeten produir decisions fiables i repetibles i aprendre a partir de un històric de dades i les seves relacions.

L'aprenentatge automàtic s'ha tornat important ja que trobar patrons és bastant difícil, no sempre es tenen suficients dades d'entrada per tal de poder dur a terme l'anàlisi. Quan això passa, tenim diverses maneres de afrontar-ho, un és la regressió.

2.2 Regressió

2.2.1 Que és la regressió

La regressió és un procés estadístic per entendre les relacions entre certes variables. Donada una matriu de dades X_{nxp} amb p variables independents i n mostres de l'experiment i un vector resposta y_{nx1} de n mostres de variables dependents, la regressió ens permet entendre com afecta un canvi en una variable independent (anomenades predictors) del conjunt X_p a la resposta corresponent d'y mantenint la resta de variables fixes. Per tant, ens permet saber quines variables independents estan relacionades amb la variable dependent. La finalitat de la regressió és generar un model de pesos que ens permetin predir la variable dependent d'una mostra amb les seves corresponents variables independents amb la major precisió possible. La regressió s'utilitza molt en l'aprenentatge automàtic ja que va molt lligat a la predicció de respostes reals a través de les seves corresponents variables.

En regressió intentem resoldre el problema de predir y de la següent manera:

$$y = \alpha + Xv + \xi$$

on ξ és qualsevol possible soroll que pugui haver afectat la mostra.

$$y_i = \alpha + X_{i1}v_2 + X_{i2}v_2 + ... + X_{ip}v_p + \xi_i$$

Els models de regressió es conformen per cinc elements bàsics: α que correspons a *l'intercept*, les incògnites v que corresponen al pendent, les variables independents X, la variable dependent y i el error de predicció ξ.

Si un cop calculats els pesos v, intentem resoldre el problema sense considerar el error ($y = Xv$), trobarem una nova y a la que anomenarem y_{fit}. El *mean squared error* (mse) entre y i y_{fit} es coneix com l'error de predicció ξ.

El que ens interessa es trobar una solució a aquest problema tal que minimitzi aquest error de predicció. Això ho podem fer aplicant diferents criteris o penalitzacions a la funció a minimitzar, aplicant per exemple regularització per

aconseguir un model que encaixi millor tant per explicar les dades actuals com les futures. Per a regressió lineal, trobem la solució de la següent manera:

$$\widehat{v} = (X^T X)^{-1} X^T y$$

No obstant, un dels problemes que afronta la regressió és que aquesta equació no sempre es senzilla de resoldre.

2.2.2 Casos

Donada una matriu d'entrada X_{nxp} amb p variables independents i n mostres de l'experiment, aquesta matriu es pot presentar en tres condicions:

- n > p: En aquest cas, el tamany de les mostres és molt gran i per altra banda tenim pocs predictors. En aquest cas, es diu que el sistema està sobredeterminat. Tenim doncs tenim solució multiple, i podem generar un model que elegeixi la v òptima segons un criteri concret. En aquest cas, els sistemes tenen infinites solucions o cap (el sistema és inconsistent).

- n = p: En aquest cas, el tamany dels predictors és molt gran i per altra banda tenim poques mostres. En aquest cas, es diu que el sistema està determinat. La solució del sistema és única.

- n < p: El cas més problemàtic en regressió és quan el tamany dels predictors és molt gran i per altra banda tenim poques mostres. En aquest cas, es diu que el sistema està sobredeterminat. En aquest cas, els sistemes són més incosistents (no té solució).

En el nostre cas, tenim un sistema subdeterminat, ja que tenim més predictors (2780) que observacions (42). Nosaltres intentarem resoldre el sistema per regressió lineal. Alhora de resoldre el problema, els dos mètodes que utilitzarem seran la lasso i el Variational Garrote, dos mètodes de regressió lineal que fan selecció de variables i regularització, procurant resoldre possibles problemes com el *overfitting*.

2.2.3 Problemes dels sistemes sobredeterminats

2.2.3.1 Overfitting

El *overfitting* es produeix quan un model és excessivament complex, com per exemple, quan tingui masses predictors pel número d'observacions del que es disposa.

Podem dir que un model que presenta *overfitting* pot sobreactuar (overreact) a canvis mínims en les dades de entrada, és a dir, reacciona a petits canvis, com per exemple soroll, en comptes de fer-ho a la relació que guarden les dades cosa que comporta també errors pobres alhora de predir el resultat que correspondria a nous valors d'entrada. El *overfitting* apareix quan el model comença a memoritzar en comptes de aprendre; això és un error perquè quan presentem unes noves dades al model, aquest fallarà completament.

És important llavors trobar una solució el més *sparse* possible, és a dir, que necessiti el mínim de predictors en detriment del error de predicció en entrenament. Contràriament l'error d'entrenament, trobar una solució més *sparse* pot facilitar que l'error de predicció del conjunt de test sigui menor al que hauria ofert un model amb més predictors diferents de zero en la solució.

Una forma de corregir *l'overfitting* és augmentant la quantitat de mostres que tinguem de cada predictor, és a dir, si per la matriu del apartat anterior X_{nxp} estem en el cas en que n<<p, el que intentarem serà augmentar n de manera que cada cop s'aproximi més a p. No obstant, no sempre és fàcil fer més experiments per tal de prendre més mostres i per tant s'ha de buscar alguna alternativa. Una manera fàcil de conseguir més mostres per a un mateix experiment és amb validació creuada. Per altra banda, una manera d'aconseguir una solució més *sparse* seria a través de la regularització.

2.2.3.2 Regularització

La regularització és el procés de introduir informació adicional a les dades amb la fi de evitar l'*overfitting* i el problema d'un problema ben definit (que complís les regles del model de Jaques Hadamard).

En general, una regularització R(f) és introduïda a una funció a minimitzar V:

$$min \sum_{i=1}^{n} V(f(\widehat{xi}), \widehat{yi}) + \lambda R(f)$$

Lambda (λ) controla la importància o la quantitat de regularització que aplicarem de la funció $R(f)$, que normalment són restriccions de *'smoothness'* i la norma del vector.

No obstant, el fet de sumar un terme a la funció a minimitzar li afecta de manera que augmenta el seu error. No obstant, es pot donar el cas on preferim trobar una solució més *sparse* (menys densa) en detriment del error de predicció del model.

Hi ha diferents funcions de regularització que es poden utilitzar, normalment, definim la regularització l_i com:

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

Es important saber quin tipus de $l_i - norm$ volem en el nostre cas, ja que les diferents funcions del model afecten de manera diferent o enfoquen de manera diferent les dades. Per exemple, la $l_0 - norm$ ofereix una solució molt més *sparse* que la $l_1 - norm$ només per la forma que tenen sobre el pla. Vegem un exemple de quina forma es veuen diverses $l_i - norm$ en el pla:

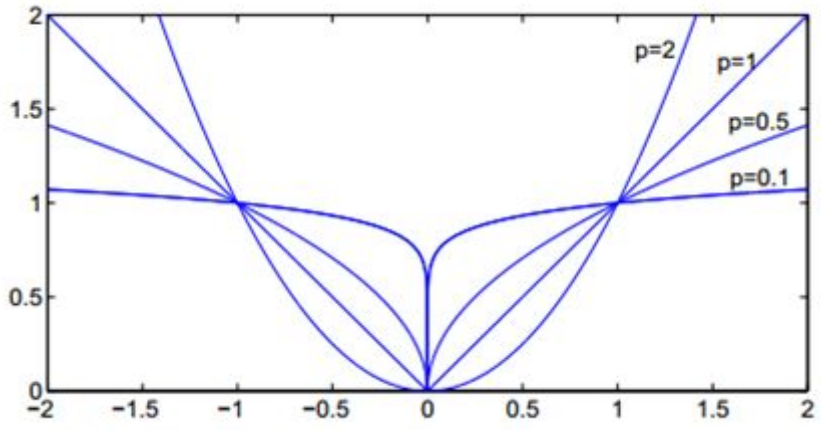


Figura 1: Mostra la gràfica corresponent a diferents normes l_p

Com podem veure si comparem per exemple p=1 i p=2, es la corba de la funció. Mentre que per al segon cas podem trobar la solució en qualsevol punt tangent a la corba, amb p=1 aconseguim que la solució es trobi en la punxa de

la funció, que sempre es trobarà sobre l'eix de coordenades, i per tant reduirà el tamany de la solució (tindrà més coeficients amb valor zero).

Així doncs, tot i que sembli que afegir un terme a la funció a minimitzar hagi de penalitzar l'error de predicció, això passa només per al conjunt d'entrenament. Si es dona el cas de *overfitting*, la regularització presenta un menor error en el conjunt de test que sinó la apliquéssim, ja que s'encarrega de lidiar amb aquells predictors irrellevants o repetits en certa manera.

Utilitzar regularització pot comportar un augment en el error de predicció d'entrenament del model, no obstant, aquest fet s'accepta perquè el que es pretén es trobar una solució que sigui més *sparse*. Altrament, si tenim una bona predicció per al conjunt de test, és a dir un error de predicció baix, podrem dir que no hi haurà *overfitting*.

2.3.3.3 Validació creuada

El que es pretén quan es fa validació creuada és testejar el model en la fase d'entrenament amb la finalitat de prevenir problemes com ara *l'overfitting*. La validació creuada acostuma a compondre's per diferents rondes i sent la mitja d'aquestes iteracions els resultats del model.

Més concisament, quan tenim un conjunt de dades X_{nxp} corresponent al conjunt d'entrenament i un conjunt de dades X_{mxp} corresponent al conjunt de dades de test, per tal de fer validació creuada utilitzarem primerament només el conjunt d'entrenament X_{nxp}. Aquest conjunt es segmentarà en dos nous conjunts complementaris, un amb major mostres i un amb menys. El primer conjunt s'utilitzarà com a conjunt de entrenament, el segon, com a conjunt de test. Aquest procés es durà a terme unes quantes vegades per a conjunts de entrenament i de test diferents, i la mitja serà el model resultant. Aquest nou model serà el que es validarà amb el conjunt de test X_{mxp} que havíem separat al principi. La validació creuada consisteix en múltiples iteracions per tal d'obtenir una millor visió del rendiment del model en termes d'error de predicció.

S'acostuma a fer validació creuada quan no hi ha suficients dades disponibles per a separar en conjunt d'entrenament i de test sense perdre significat de modelat de les dades, és a dir, quan per una matriu X_{nxp}, $n << p$, ja que en aquestes situacions el model acostuma a presentar problemes de *overfitting*.

2.3.4 Diferents aproximacions

Hi ha varies aproximacions per a resoldre aquests problemes, com per exemple *ridge regression*, la *Lasso* o el *Variational Garrote*.

La ridge regression aporta un element nou, la regularització. El terme de regularització ($\frac{1}{2}\lambda\Sigma i\, w_i^2$ sent $\lambda > 0$) que substitueix la matriu de covariança per $X + \lambda I$, obtenint una matriu de rang màxim per tot p. La ridge regression ens aporta més precisió en la predicció que la regressió lineal simple, tot i que no aporta més interpretabilitat de la solució.

La lasso per altra part, utilitza l1 en comptes de l2, el qual implica que la solució tendeix a ser més *sparse*, millorant tant la precisió de la predicció com la interpretabilitat de la solució.

El Variational Garrote, implementa tant la regularització com mètodes per a evitar mínims locals. Altrament, utilitza l0 en comptes de l1 com a penalització, de la mateixa manera que la lasso. Augmentant tant la predicció com la precisió, així com ampliant el casos en que pot resoldre el problema.

2.4 Mètodes de regressió

2.4.1 Lasso

Lasso (*least absolute shrinkage and selection operator*), és un mètode d'anàlisi de regressió. Es caracteritza per implementar tant la selecció de variables com per fer regularització per tal de trobar una solució amb una millor precisió en la predicció i la interpretabilitat del model resultant.

Una diferència que presenta la lasso conforme altres mètodes es que és capaç de assolir 'non-zero coefficients'. La lasso utilitza una regularització amb l1, utilitzant el valor absolut en comptes del mòdul. Això afecta als límits de la forma de la restricció (geomètricament), permetent trobar els valors tangents a la recta formada per l1, i que sempre es troben sobre algun eix de coordenades, sent les coordenades corresponent a la variable del eix 0.

Tot i que en un principi la lasso es va concebre per a resoldre el problema dels mínims quadrats, la seva capacitat permet poder aplicar aquest mètode per a

altre tipus de problemes com models estadístics incloent models lineals generalitzats, equacions d'estimació generalitzades, models de riscs proporcionals, i M-estimadors, d'una manera senzilla.

Donada la funció:

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

Definim les dades com $D : \{x^{\mu}, y^{\mu}\}, \mu = 1, ..., p$. La lasso regularitzada de la estimació 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.

2.4.2 Variational Garrote (VG)

El *Variational Garrote* (VG) és un mètode de regressió lineal. Es caracteritza per tenir una mètode *forward*, on aplica tècniques del *Variational*, i posteriorment fa *backwards* a través regressió a través de MAP. La regressió MAP es caracteritza per treballar *backwards* des de mapes més vells fins als mapes més nous. Comparant directament amb la lasso, considerem el model de regressió [1]:

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

on s_i identifica els predictors rellevants i.

Definim les dades com $D : \{x^{\mu}, y^{\mu}\}, \mu = 1, ..., p$. El *likehood term* es:

$$p(y|x, s, w, \beta) = \sqrt{\frac{\beta}{2\pi}}^{p/2} exp(-\frac{\beta}{2} (y - \sum_{i=1}^{n} w_i s_i x_i)^2)$$

Assumim que cada s_i té idèntica probabilidad a priori:

$$p(s_i|\gamma) = \frac{exp(\gamma s_i)}{1 + exp(\gamma)}$$

amb una γ donada que especifica l'*sparsity* de la solució. Així doncs, la probabilitat a posteriori es:

$$p(s, w, \beta | D, \gamma) = \frac{p(w,\beta)\, p(s|\gamma)\, p(D | s,w\beta)}{p(D | \gamma)}$$

No obstant, el calcular estimació per MAP o estadístiques d'aquest posteriori és difícil. Per això, el Variational Garrote proposa calcular una aproximació Variational a la probabilitat marginal a posteriori $p(w, \beta | D, \gamma) = \sum_s p(s, w, \beta | D, \gamma)$ i posteriorment calculat la solució MAP respecte w, β. Com $p(D | \gamma)$ no depèn de w, β; ho podem ignorar.

Apoximarem la suma del Variational amb la desigualtat de Jensen [1] obtenint la Variational Free Energy ($-F(q, w, \beta)$), on $q(s)$ es la aproximació Variational. La q òptima es troba minimitzant F respecte q.Definirem:

$$q_i(s_i) = m_i s_i + (1 - m_i)(1 - s_i)$$

on q està especificat quan m_i = q_i ($s_i = 1$), el qual anomenem m. Així doncs, ara podrem saber el valors esperats respecte a q:

$$F = \frac{\beta p}{2}\Big(\sum_{i,j}^{n} m_i m_j w_i w_j + \sum_i m_i(1-m_i) w_i{}^2 \chi_{ij} - 2\sum_{i=1}^{n} m_i w_i b_i + \sigma_y^2\Big) -$$
$$- \gamma \sum_{i=1}^{n} m_i + \sum_{i=1}^{n}(m_i log(m_i) + (1-m_i) log(1-m_i)) - \frac{p}{2} log\frac{\beta}{2\pi}$$

La primera línia de l'equació correspon al *likehood term* i la segona correspon a la entropia de $q(s)$ i la prioritat sobre s .La solució a minimitza F respecte m dona les equacions, adicionalment maximitzem $p(w, \beta | D, \gamma)$ respecte w, β.

$$m_i = \sigma(\gamma + \frac{\beta p}{2} w_i^2 x_{ii})$$

$$w = (\chi')^{-1} b \quad ; \qquad \chi_{ij} = \chi_{ij} m_j + (1 - m_j)\chi_{ij}\delta_{ij}$$

$$\frac{1}{\beta} = \sigma_y^2 - \sum_{i=1}^{n} m_i w_i x_i$$

amb $\sigma(x) = (1 + exp(x))^{-1}$, trobem la solució resolent les equacions anteriors

Per la aproximació Variational/MAP, el model predictiu és:

$$y = \sum_i m_i w_i x_i + \varepsilon^\mu$$

La solució es calcula en dos passos, primer calculem w_i utilitzant Ordinary Least Squares (OLS) i després busquem m_i minimitzant:

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

Donada la similitud entre aquest mètode i el Variational, el mètode és diu Variational Garrote. No obstant, aquesta solució no equival a substituir $m_i w_i$ per $w_i s_i$ en la primera equació, ja que utilitzem una nova matriu χ' en comptes de χ.

El *VG*, al igual que qualsevol altre mètode que pretén resoldre un problema no convex, pot presentar problemes de mínims locals. No obstant, al aproximar amb una combinació de *MAP* i del *Variational* juntament amb el '*anhealing procedure'* que aconseguim al augmentar γ, seguit de un 'escalfament' per detectar *'hysteresis'* ajuda a evitar el conflicte de trobar-se en un mínim local.

3. ANÀLISI DE LES DADES

3.1 Dades sintètiques

Abans de començar a treballar amb les dades reals mesurades en l'experiment, hem creat un conjunt sintètic de dades per tal de veure el rendiments del diferents algoritme en casos extrems.

Les dades X són generades a partir de una distribució Gaussiana multivariable amb mitja zero amb una estructura de covariança especificada. La resposta y es genera a partir de la fórmula: $y^{\mu} = \sum_i \hat{w}_i \hat{x}_i^{\mu} + d\,\xi^{\mu}$

Així doncs, els paràmetres que es modificaran respecte a aquesta equació seran: tamany del conjunt d'entrenament $\hat{x}_i^{\mu}$, és a dir, el tamany de i; la *sparsity* del vector $\hat{w}$ ($\frac{número\ w \neq 0}{número\ w}$, que anomenarem δ); el rang del soroll, delimitat per $d = \frac{1}{\sqrt{\beta}}$ i ρ, que és la correlació de les dades $\hat{x}$.

3.2 Concentracions

Les dades presentades corresponent a un experiment on es van mesurar les respostes de una llengua elèctrica concreta a unes concentracions específiques de tres substàncies diferents: 'cysteine' (CYS), 'tyrosine' (TYR) i 'tryptophan' (TRP). Es van fer 42 experiments diferents amb 42 combinacions diferents de concentracions.

Les mesures es van prendre utilitzant una cel·la de sensors: un elèctrode de referència, un elèctrode auxiliar i els 5 sensors actius, on 4 eren del mateix tipus i un de modificat.

Les dades consisteixen de un conjunt de entrenament (27 mostres) i un conjunt de test (15 mostres) de 2780 predictors cadascú. Les concentracions de les mostres d'entrenament poden assolir tres nivells diferents: 0 M, 0.5 ·10-4 M i 3.0 ·10-4 M; combinant aquest tres valors entre les tres concentracions obtenim els 27 experiments corresponents al conjunt d'entrenament.

Per altra banda, trobem les concentracions del conjunt de test. Dintre del rang de valors del conjunt d'entrenament ([0,3]) es van generar una sèrie de valors aleatoris de combinacions de concentracions (cap combinació del conjunt de

test apareix al conjunt d'entrenament) per tal de generar situacions inèdites que es poguessin produir. En la figura de sota es poden apreciar els diferents valors.

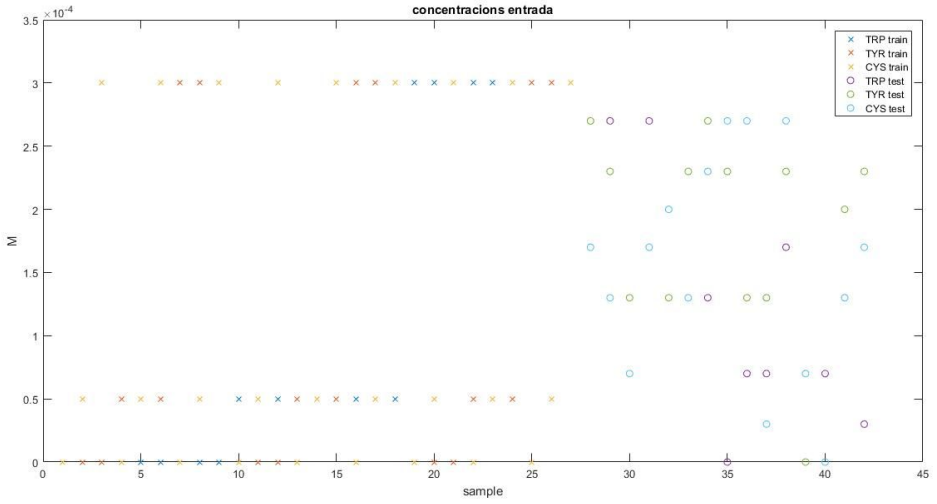


Figura 2: Valors de concentració reals de les dades d'entrada

Per tal de poder predir concentracions futures, s'ha introduït al model de una matriu d'entrada X, on cada fila correspon a la resposta dels cinc sensors concatenada per a una combinació de concentracions concreta (1 mostra). La senyal capturada per cada sensor té un tamany de 556 variables, conformant un total de 2780 variables per mostra. En la següent figura podem veure un exemple de la senyal de una mostra.

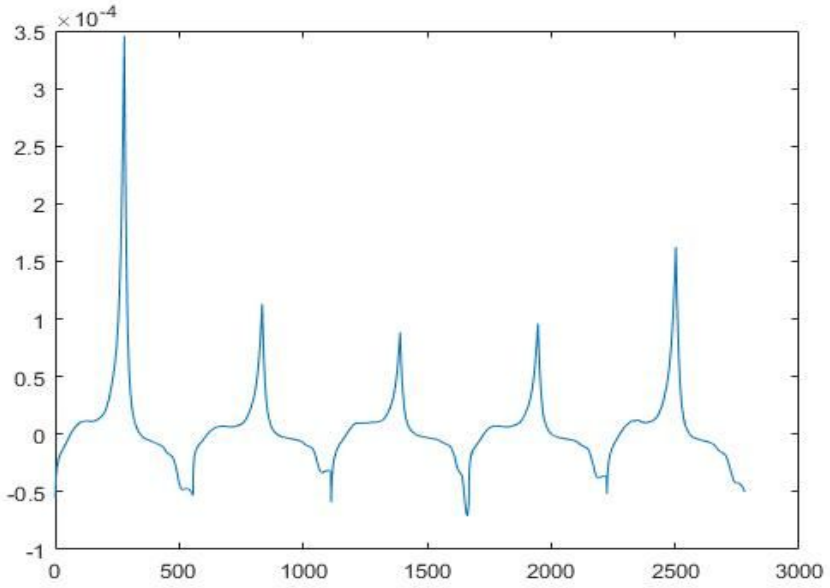


Figura 3: Exemple d'una mostra de l'experiment de la matriu de dades recollides per la llengua electrònica $X_i \in R^p$

En total, de la mateixa manera que per a les concentracions, tenim 42 mostres, les quals hem repartit en test (27) i entrenament (15), conformant una matriu de 42x2780.

Donat aquest volum de dades, abans de aplicar els models de predicció, s'ha volgut analitzar una mica les dades per veure si se'n podia treure informació rellevant, o si més no, reduir el tamany de les dades d'entrada per agilitzar els càlculs.

Altrament, en vista dels resultats obtingut, no sempre s'ha utilitzat tot aquest conjunt com a dades, quan fem validació creuada per exemple, el conjunt de entrenament conforma els nous conjunts que utilitzarem: un nou conjunt de entrenament més petit que 27 i un conjunt de test (els restants dels 27 que no hem agafat per a entrenament) que varia quines concentracions utilitzem per a cada conjunt en cada *k-fold* de la validació creuada. El conjunt real de test es reserva per a la validació.

Quan no fem validació creuada, s'ha utilitzat els conjunt de manera estàndar: el conjunt d'entrenament (27) per a generar un model, i el conjunt de test (15) per a testejar la validesa del model.

3.3 Matriu de correlació

La matriu de correlació és una matriu que ens dóna informació sobre quan correlacionades es troben les diferents variables d'una vector 1xP, és a dir, si hi ha alguna relació entre cada parell de variables del vector, resultant en una matriu PxP on s'especifica la correlació de cada variable amb cadascuna de les altres. Es caracteritza per ser simètrica i per tenir la diagonal formada per 1, ja que es compara la correlació de cada variable amb ella mateixa.

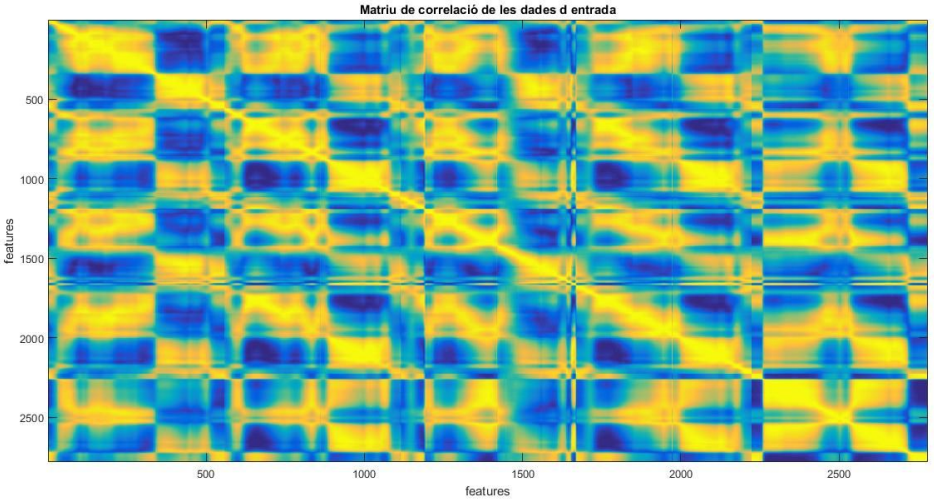
Saber la correlació de les variables a priori és molt important, ja que hi ha mètodes que no són capaços de trobar una solució òptima quan les dades estan altament correlacionades. De la mateixa manera, és molt difícil entendre dades que no tenen gens de correlació entre elles. Calcular la matriu de correlació és important per enfocar l'anàlisi de les dades i entendre possibles explicacions sobre models que generen una solució objectivament dolenta.

Diem que hi ha correlació entre dues variables quan els valors de una d'aquestes varia sistemàticament respecte als homònims valors de l'altra. La correlació és un valor que pot oscil·lar de -1 a 1 (si calculéssim la correlació del valor absolut de les variables, aquesta només podria trobar-se entre 0 y 1).

Direm que si la correlació està entre 0.5 i 1 (-0.5 i -1), les variables estan altament correlades, si per altra part la correlació es troba entre 0.5 i 0.3 (-0.5 i -0.3), direm que estan mitjanament correlades; finalment, si la correlació entre les variables es troba entre 0.3 i 0.1 (-0.3 i -0.1), direm que aquestes variables estan poc correlades. Valors de 1 i -1 no s'acostumen a trobar en matrius de dades, així com correlació 0, que significa que les variables son totalment independents.

De la mateixa manera, si tenim N mostres d'aquestes P variables, la matriu de correlació resultarà de comparar cada columna amb la resta de les columnes que conformen la matriu, resultant també en una matriu PxP.

Per tal de saber les característiques de les nostres dades, hem calculat la seva matriu de correlació. Alhora de fer regressió, el que intentem es trobar una funció que sigui capaç de representar o entendre les dades. Si totes les dades estan altament correlacionades, caldrà fer un anàlisis per saber quines són més rellevants i quines poden ser despreciades ja que el seu significat està ja present en una altra variable correlacionada amb aquesta.



Figrua 4: Matriu de correlació de les dades de la matriu d'entrada X.

En la imatge anterior podem veure com hi ha cinc diagonals ben marcades i com hi ha zones ben fosques, donant a entendre una forta correlació negativa. Tanmateix, podem apreciar cinc diagonals ben marcades, el qual denota que les 5 senyals presses pels 5 sensors estan altament correlades. Més concretament, podem veure un alta correlació positiva de 278 posicions seguida de una correlació negativa de 278 posicions. Això és degut a que les primeres es corresponent a la pujada de la senyal i les altres a la baixada.

Per entendre-ho millor, graficarem el valor absolut de la anterior imatge, per remarcar els punts d'alta correlació:

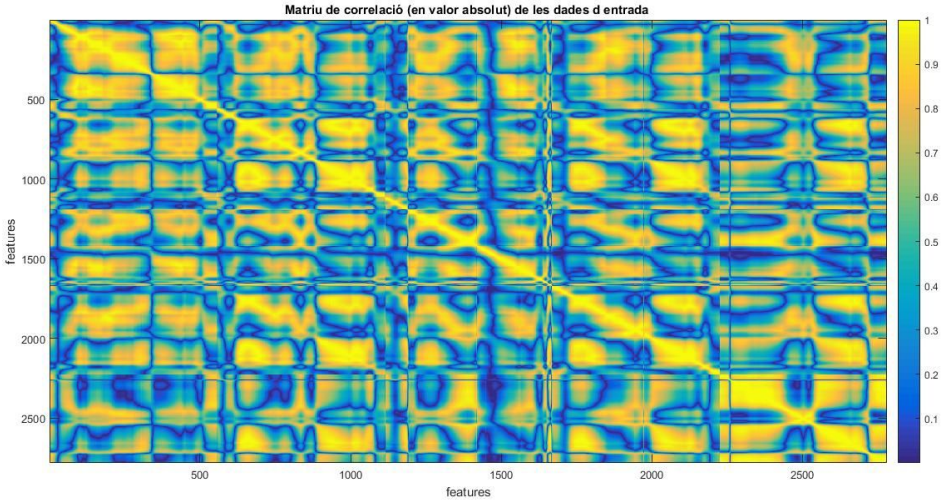


Figura 5: Valor absolut de la matriu de correlació de les dades d'entrada X

Com podem veure la imatge de sobre, la majoria dels coeficients de la matriu estan per sobre del 0.5 (fins i tot el 0.7) de correlació, sent ben poc els coeficients que tenen correlació blava entre ells (poca).

Així doncs, com que les dades estan correlacionades, podrem analitzar-les a través d'una funció. No obstant, si les dades estan molt correlacionades, pot passar que el mínim soroll afecti altament al model que es generarà, presentant un pitjor rendiment.

4. RESULTATS

4.1 Dades sintètiques

Els experiments amb les dades sintètiques s'han repetit 100 cops per a cada cas, és a dir, per a cada data set em generat 100 models de Lasso i 100 models de Variational Garrote. Cadascun dels paràmetres que es mostren son el resultat de la mitja d'aquestes 100 sent el resultat mostrat per cada paràmetre, la mitja aritmètica $\bar{x}$ i la desviació estàndard σ.

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

Els resultats que avaluem dels mètodes corresponen a:

- MSE Train: mean squared error $(MSE = \frac{1}{N}\sum_{i=1}^{N}(\widehat{Y}_i - y_i)^2)$ calculat entre la y_{pred} i la y real per al conjunt d'entrenament .
- MSE Test: mean squared error calculat entre la y_{pred} i la y real per al conjunt de test.
- NºNonZeros: correspon al número de valors diferents a zero en el la solució v.
- ΣNonZeros: correspon a la suma dels valors diferents a zero en el la solució v.
- $\lambda\,(o\,\gamma)$:hyperparametre de la regularització.

4.1.1 Predir un model ideal

En aquest apartat, el que avaluarem serà el rendiment dels dos mètodes presentats sobre un conjunt de dades sintètiques. En aquest cas hem generat una resposta a predir a través d'un model $y^{\mu} = \sum_{i=1}^{n} w_i x_i$, és a dir, la resposta és directament els valors d'entrada pels seus pesos corresponents; no hi ha presència de soroll que pugui molestar alhora de predir.

Els valors de la gaussiana són : $\beta = 2;\ \delta = 10;\ n = 100;\ n_{test} = 50;\ \mu = 0.7$. Les dades s'han calculat fent la mitja de calcular 100 models de cada tipus per al conjunt de dades generades.

Els paràmetres que s'han extret per tal d'analitzar els diferents casos han sigut la mitja aritmètica entre tots els models i la seva desviació estàndard.

Model sense soroll
$(y = \alpha + Xv)$

Model sense soroll
$(y = \alpha + Xv)$

	Lasso		Variational Garrote	
	$\bar{x}$	σ	$\bar{x}$	σ
MSE_train	1.1316e-05	0	2.4064	1.9068
MSE_test	2.2935e-05	0	2.5487	1.9196
NºnonZeros	12	0	10.11	1.1000
sum(v)	1.9970	0	2.0336	0.2405
lambda	5.229,5	17331	-17.3379	7.3911

Model amb soroll
$(y = \alpha + Xv + \xi)$

	Lasso		Variational Garrote	
	$\bar{x}$	σ	$\bar{x}$	σ
MSE_train	0.5109	0.0726	3.0192	2.1563
MSE_test	0.3543	0.0417	3.1091	2.1043
NºnonZeros	14.1800	1.9663	10	0
sum(v)	1.9701	0.0426	2.0892	0.2896
Gamma	5229.5	17331	-16.7868	7.4112

Quan les dades d'entrada no presenten soroll, el més significatiu és que el garrote encerta sempre el nombre de non zeros en el vector mentre que la lasso troba més coeficients dels que realment hi ha. De la mateixa manera, quan afegim soroll en la resposta predita, aquest fet és manté: el Variational Garrote sempre ofereix una

solució més *sparse* que la lasso alhora de predir un model de dades generat (o mesurat) sense soroll.

Respecte l'error de predicció, la lasso ofereix una solució que aproxima millor les dades de test. El garrote per altra banda no presenta malts resultats respecte a l'error, encara que sí que són pitjors als de la lasso.

4.1.2 Predir un model amb soroll

Ara que hem vist que ambdós models són vàlids alhora de generar un model de regressió sobre un conjunt de dades, veurem com es comporten quan aquestes dades sintètiques es generen d'una manera més real. En aquest experiment el que farem serà afegir un soroll a les respostes a predir pel model. D'aquesta manera simulem condicions reals que es poden donar per molts factors, com soroll en l'ambient de mesura, la precisió dels instruments de mesura, arrodonir els càlculs,...

El model utilitzat per a generar la resposta a les dades en aquest experiment és el següent: $y = \alpha + Xv + \xi$. Aquest nou terme ($\xi = \frac{1}{\sqrt{\beta}} * random$) és el soroll que hem generat nosaltres. És important la elecció d'aquest paràmetre $sigma = \frac{1}{\sqrt{\beta}}$, ja que, com veurem en experiments posteriors, un soroll molt alt distorsiona massa la senyal per a que pugui ser aproximada amb prou precisió.

Els valors de la gaussiana són : $\beta = 2;\ \delta = 10;\ n_{train} = 100;\ n_{test} = 50;\ \mu = 0.7$. Les dades s'han calculat fent la mitja de calcular 100 models de cada tipus per al conjunt de dades generades.

<u>Aproximació sense soroll ($y = \alpha + Xv$)</u>

	LASSO		Variational Garrote	
	$\bar{x}$	σ	$\bar{x}$	σ
MSE_train	0.1757	0	0.4145	0.0539
MSE_test	1.7447	0	0.6755	0.1702
NºZeros	56	0	24.2100	21.2252
Sum(v)	-4.3230	0	-5.8315	0.1856
Lambda (o Gamma)	5229.5	17331	-8.1416	3.9880

Aproximació amb soroll ($y = \alpha + Xv + \xi$)

	LASSO		Variational Garrote	
	$\bar{x}$	S'està carregant...	S'està carregant...	S'està carregant...
MSE_train	0.6371	0.1112	0.9235	0.1222
MSE_test	2.0123	0.1854	1.1843	0.2543
NºZeros	65.8400	13.0033	17.3200	18.1302
Sum(v)	-4.0763	0.3410	-5.8167	0.2260
Lambda (o Gamma)	5229.5	17331	-13.3412	6.4476

Avaluem un conjunt de dades una mica més real, un conjunt amb soroll: afegirem soroll a la resposta real a predir i avaluarem si es millor predir aquesta resposta considerant un soroll aleatori afegit a la resposta predita o no.

Si comparem per a ambdós mètodes els resultats, podem veure com afegir soroll disminueix la *sparsity* de la resposta; trobem més coeficients diferents de zero al aproximar aquest model que en el apartat anterior. Aquest fet no és sorprenent, ja que la quantitat de soroll respecte la senyal sempre ha sigut un problema alhora de fer regressió. Si per altra banda comparem els errors, tots dos mètodes empitjoren tant en error de predicció com en error d'entrenament.

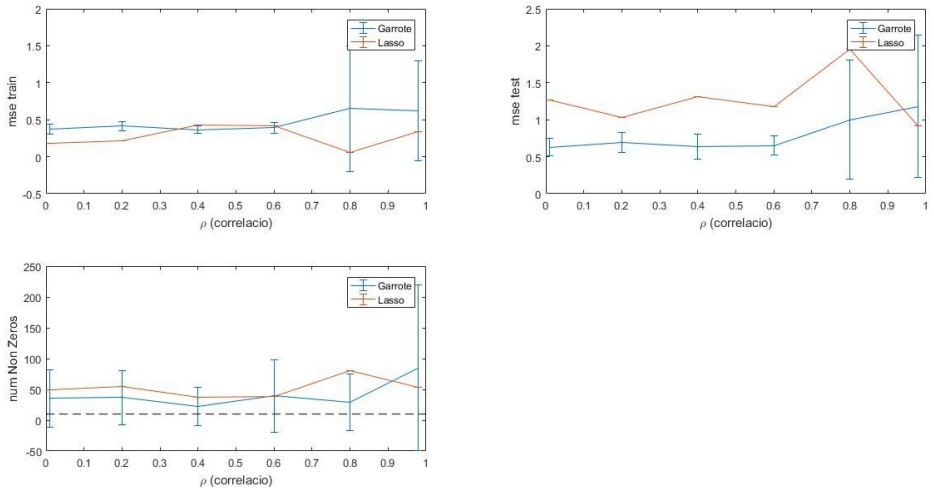
Si hem de comparar els dos mètodes entre ells, podem dir que el *Variational Garrote* presenta un major error d'entrenament però obté un menor error de predicció en test. Respecte a la *sparsity*, el *Variational Garrote* presenta un numero menor de valors diferents de zero en el vector S'està carregant.... Així doncs, a diferència de l'apartat anterior on es troba un cas ideal, en aquest cas el *Variational Garrote* presentat tant una solució més *sparse* com un error en test menor. Per tant, podem dir que el *Variational Garrote* hauria de ofereix millors resultats que la Lasso en situacions reals amb soroll.

4.1.3 Variar paràmetres de les dades sintètiques

En aquest apartat, en el que ens centrarem serà en canviar diferents factors que afectaran al conjunt de dades, tant a les dades que utilitzarem per a predir la resposta, com el tamany dels pesos S'està carregant..., o per altra banda paràmetres com el soroll, que afectaran a la resposta real directament.

4.1.3.1 Variar la correlació en les dades

Els valors de la gaussiana són : S'està carregant.... Les dades s'han calculat fent la mitja de calcular 100 models de cada tipus per al conjunt de dades generades



Figra 6: La figura de dalt mostra per als dos mètodes. A dalt a l'esquerra el mse d'entrenament, a dalt a la dreta el mse de test, a l'esquerra el número de valors en S'està carregant...diferents de zero. Anàlisi per als diferents valors de rho: 0.01, 0.2, 0.4, 0.6, 0.8 i 0.98.

Els dos mètodes funcionen millor quan es troben amb conjunts de dades amb cert grau de correlació. No obstant, conforme augmentem la correlació ambdós mètodes empitjoren en rendiment. Quan les dades no estan correlacionades, els mètodes no presenten carències aparents respecte majors nivells de correlació.

Alhora de predir dades, la correlació que presentin aquestes pot afectar el rendiment del model que intenta predir les dades o trobar-hi patrons. Per part de la Lasso, podem veure que conforme baixa la correlació de les dades baixa dràsticament l'error d'entrenament, no obstant, l'error de test sembla que s'estanca. Respecte *l'sparsity*, la solució és més *sparse* per valors mitjos de correlació i menys *sparse* per dades o molt correlades, o molt poc.

Tanmateix, el *Variational Garrote* es comporta de manera similar, conforme baixa la correlació, l'error d'entrenament disminueix, l'error de test per altra banda, s'estanca per la majoria de casos, exceptuant els casos d'alta correlació on veiem que l'error en test també augmenta així com *l'sparsity*. No obstant, el

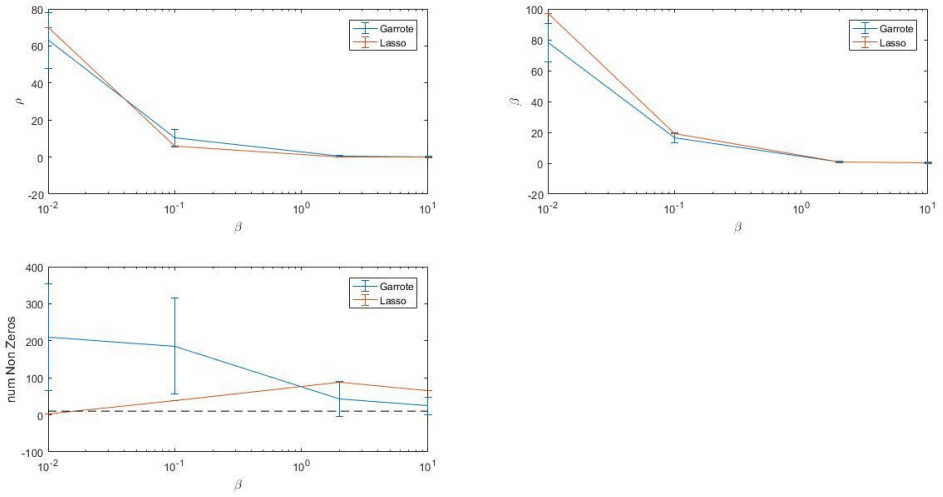
Variational Garrote aguanta més correlació que la lasso, sobretot alhora de trobar una solució *sparse*.

Es pot observar com es manté la conclusió a la que habiem arribat abans, el garrote ofereix un millor error de test, així com una solució més *sparse* que la Lasso. En els casos extrems això no es compleix del tot, on tots dos baixen en rendiment.

4.1.3.2 Variar la quantitat de soroll

El soroll és molt important alhora de predir una senyal, si és soroll que interfereix en una senyal és massa gran en relació al valor d'aquesta, no hi haurà manera de predir aquesta senyal. No obstant, es interessant veure fins a quin punt els dos models toleren la presència de soroll, i a més a més, en quina quantitat.

S'està carregant...



Figra 7:La figura de dalt mostra per als dos mètodes. A dalt a l'esquerra el mse d'entrenament, a dalt a la dreta el mse de test, a l'esquerra el número de valors en S'està carregant...diferents de zero. Anàlisi per als diferents valors de beta: 10, 2, 0.1 i 0.01..

En aquest cas el que analitzem és l'efecte del soroll en la resposta que intentem predir. Normalment, predir una senyal amb soroll és més difícil que predir una senyal sense soroll. Així doncs, podem veure que conforme disminueix beta (augmenta el rang de soroll) els dos mètodes mostren pitjors resultats.

Els casos que es presenten representen molt poc soroll, un ~10% del soroll, un ~30% del soroll i un ~90% del soroll. A partir d'un terç del soroll podem veure que els resultats empitjoren considerablement; no obstant, fins a un 10% del soroll, els error de test encara són prou bons, la solució, per altra banda, no es molt *sparse*.

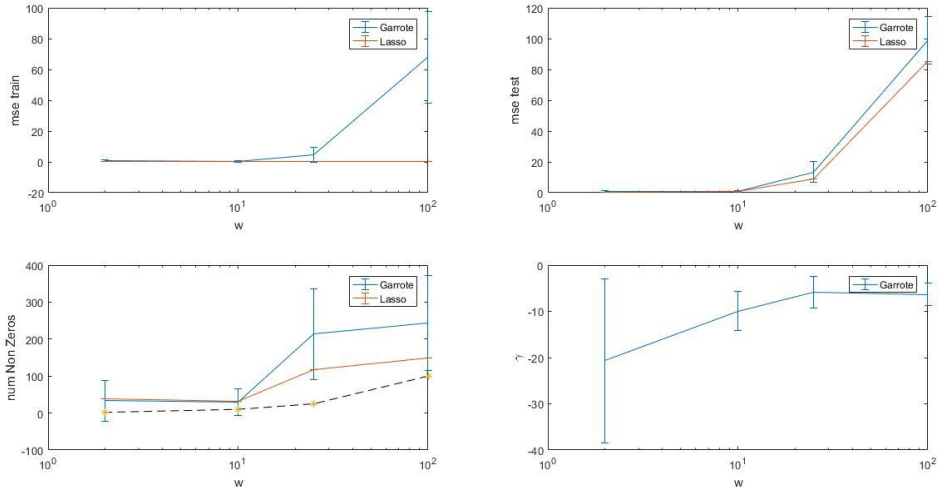
És interessant veure com quan hi ha molt soroll i els dos mètodes no són capaços de trobar resultats satisfactoris; es comporten diferent com podem veure la figura la figura corresponent al número de valors diferents de zero de la solució. Podem veure com la Lasso ofereix una solució plena de zeros, el que implica que totes les respostes seran predites amb un valor constant. Per altra banda, el Variational Garrote troba la S'està carregant... més propera a zero, però presenta una solució molt densa.

Així doncs, si tenim la resposta a predir presenta molt soroll serà molt difícil definir un model adequat.

4.1.3.3 Variar la *sparsity* del vector S'està carregant...

En aquest apartat el que volem avaluar és el rendiment dels algoritmes alhora de tenir que trobar solucions que no són *sparse*; és a dir, variarem la densitat del vector de pesos que generarà la resposta real. És interessant aquest experiment ja que ambdós mètodes apliquen regularització amb la finalitat de trobar una solució més *sparse* en detriment de l'error, així que seria interessant veure si els mètodes ofereixen una solució més *sparse* que la solució real. Els valors de la gaussiana són :

S'està carregant...



Figra 9: La figura de dalt mostra per als dos mètodes. A dalt a l'esquerra el mse d'entrenament, a dalt a la dreta el mse de test, a l'esquerra el número de valors en S'està carregant...diferents de zero i la figura de sota a la dreta correspon a l'hiperparàmetre del Variational Garrote S'està carregant.... Anàlisi per als diferents nombre de valors diferents de zero en la solució: 2, 10, 25 i 100..

Ambdós algoritmes han demostrat que les solucions menys *sparse* s'aproximen amb un nombre de coeficients menys encertat que en casos com quan hi ha 100 coeficients diferents de zero. Això no vol dir que a solucions molt *sparse* les solucions no siguin bones. Com podem veure en el cas del hiperparàmetre S'està carregant... del Variational Garrote, la elecció del hiperparàmetre s'ajusta a la *sparsity* de la solució real.

Entre tots dos algoritmes, podem dir que el Variational Garrote troba normalment solucions més *sparse* excepte en el cas extrem de *sparsity*, en el que respecta a la *sparsity* de la solució. Respecte el error, la lasso presenta menor error que el Variational Garrote.

4.2 Concentracions

Saben ara quins son els comportaments dels algoritmes donades diferents condicions en les dades d'entrada, ens veiem en forces d'encarar unes dades reals i extreure'n que hi passa en aquestes. Així doncs, generarem un model Lasso i un model Variational Garrote de la mateixa manera que amb les dades sintètiques i analitzarem els resultats que se'n puguin extreure.

Les dades consisteixen en tres concentracions, per tant, els resultats els tindrem tres cops (1 per cada concentració). Les dades que s'han utilitzat en els

següents experiments són les que corresponen únicament les d'entrenament del paper.

4.2.1 Anàlisi estàndard:

En primer lloc, hem desenvolupat el anàlisi amb les dades tal com es relaten en el document, assignant els corresponents conjunts d'entrenament (27) i test (15):

Lasso

	TRP		TYR		CYS	
	S'està carregant..	S'està carregant..	S'està carregant..	S'està carregant..	S'està carregant..	S'està carregant..
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

Variatonal Garrote

	TRP		TYR		CYS	
	S'està carregant...	S'està carregan	S'està carregant...	S'està carregant.	S'està carregant...	S'està carregant..
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

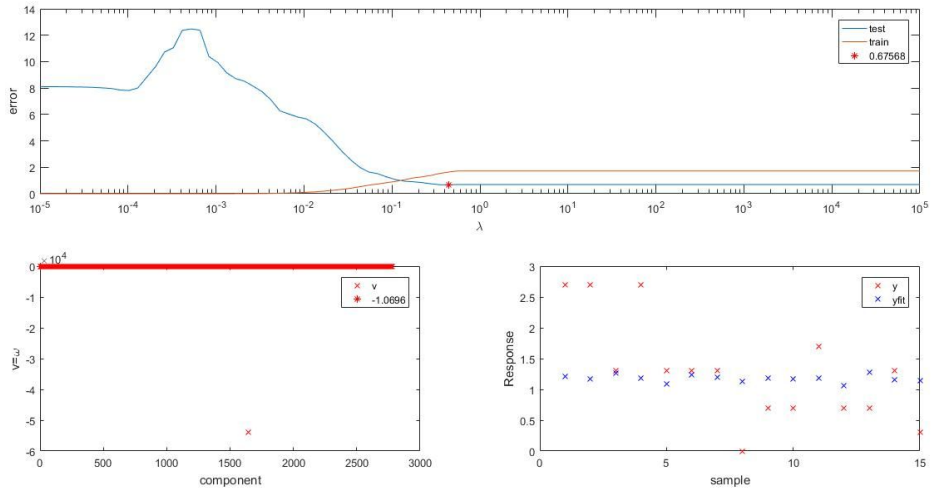
En la taula anterior es presenta el resultats que presenta els dos models que hem implementat vers les dades reals conforme les tenim. Cada mètode s'ha fet 20 cops i la mitja d'aquests 20 models, juntament amb la desviació estàndard, són les dades que es presenten a la taula.

A primera vista es pot apreciar una diferencia bastant significativa entre els dos mètodes: per una part la Lasso ofereix solucions molt *sparse* (o fins i tot sense cap valor en la solució diferent de zero) mentre que el Variational Garrote,

contràriament al que hem vist en les dades sintètiques, mostra solucions molt poc *sparse*.

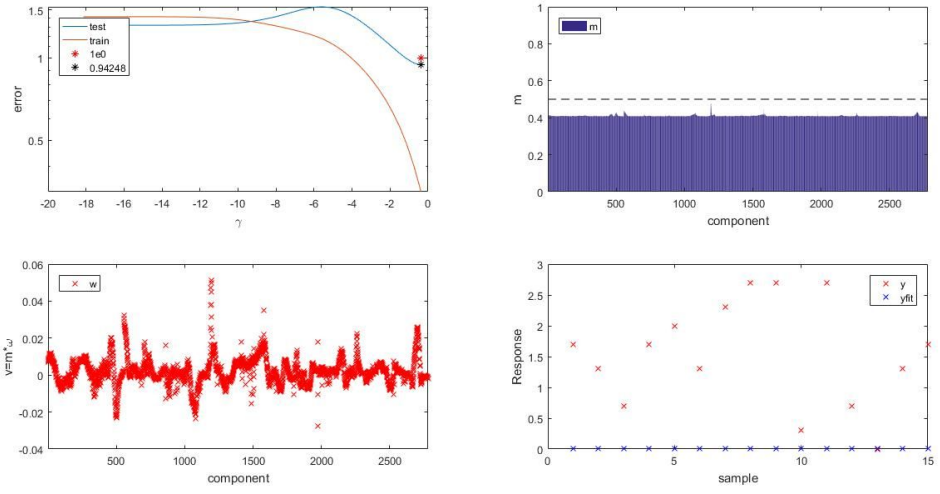
Per tal de poder mostra millor què és el que passa realment a les dades, hem generat la següent figura, on es mostra un exemple de un model calculat per la lasso i un model calculat pel *Variational Garrote* per un aminoàcid concret, en aquest cas el TYR.

Lasso:



Garrote:

Figura 10: La figura de sobre repre



senta el error d'entrenament i l'error de test, així com la seva S'està carregant...òptima. La figura de sota a l'esquerra correspon als coeficients de la solució S'està carregant..., que en el cas de la lasso S'està carregant...i en el cas del Variational GarroteS'està carregant.... La figura de sota a l'esquerra correspon a les respostes reals del conjunt d'entrenament vers les respostes del conjunt d'entrenament predites.

En les figures anteriors es mostren tres gràfiques equivalents entre el Variational Garrote i la Lasso. Si ens fixem en l'error veiem com es comporta en funció de la regularització, podem veure com conforme augmentem la regularització augmenta l'error d'entrenament però per altra banda disminueix l'error de test, per a ambdós casos. El més intrigant però es pot observar en l'última figura, on comparem la resposta real amb la resposta predita (S'està carregant...). Podem veure que en ambdós casos la resposta predita és el mateix valor indistintament del valor resposta que intentem aproximar. Això no es d'estranyar mirant la figura de la seva esquerra, on per la Lasso veiem que no hi ha cap pes diferent de zero en la solució i per part del Variational Garrote veiem que, tot i trobar solució amb uns pesos diferents de zero, aquests no compleixen amb el criteri per a ser activats per m, i per tant, la solució en ambdós casos correspon al que seria *l'intercept* calculat alhora de generar el model.

Aquests resultats ens recorden molt al cas del apartat anterior on el soroll era massa important alhora de predir un model, ja que, com podem veure, la Lasso ofereix una solució plena de zeros , només en el cas de la concentració TYR trobem una resposta predita no constant. El Variational Garrote per altra banda presenta solucions molt denses. L'únic cas amb dades sintètiques on el Variational Garrote i la Lasso mostren aquests resultats és quan no troben solució.

Com podem veure en les taules anteriors i els diferents histogrames que es van treure sobre aquestes dades, els resultats presentats no són gaire bons. Creiem que part de la culpa d'aquest resultats és la distribució dels valors de les concentracions, ja que els valors que utilitzem per a testejar són diferents als que utilitzem per a entrenar.

Per això implementarem la validació creuada. A través d'aquesta, serem capaços de veure si els diferents valors entre entrenament i test són gaires rellevants. Això ho podrem saber perquè utilitzarem el conjunt d'entrenament tant per generar els models com per validar-los amb un subconjunt que no hagin utilitzat per a entrenament. Un cop s'hagi generat i validat el model amb aquest valors, s'avaluarà amb el conjunt de test.

4.2.2 Anàlisi amb validació creuada

En vista dels resultats obtinguts en l'últim anàlisis, hem decidit provar de fer validació creuada al conjunt d'entrenament. Així doncs, del conjunt d'entrenament de 27 mostres el dividirem en K particions i en separarem una; aquesta correspondrà al conjunt de test del *kfold*. Aquest procés és farà tants cops com particions tinguem, i sempre totes les particions han sigut conjunt de test un cop.

Per tal de dur a terme els càlculs amb un conjunt d'entrenament suficient gran i un conjunt de validació rellevant, s'ha elegit 9 com a *K-fold*. A partir d'ara, el que farem serà variar els conjunts d'entrenament i de validació. Des d'ara, deixarem de costat el conjunt de test i utilitzarem únicament el conjunt d'entrenament, que consisteix en 27 mostres. D'aquestes 27 mostres en separarem 3 i conformaran el conjunt de validació; les 24 restants conformaran el conjunt d'entrenament.

Un cop calculat aquests valors, es calcula la seva mitja i s'elegeix la solució òptima: en el cas de la lasso es la S'està carregant...l'error de test sigui mínim, mentre que en el cas de la lasso, elegim la S'està carregant...que retorna el mètode.

Un cop conegudes les S'està carregant..., es recalcularà el model amb un dels conjunts utilitzat i es calcula l'error i els altres paràmetres de la taula. Aquest experiment s'ha dut a terme 15 cops i la mitja i la desviació estàndard entre aquests experiments és el que es mostra en la següent taula:

Lasso

	TRP		TYR		CYS	
	S'està carregant..	S'està carregant..	S'està carregant..	S'està carregant..	S'està carregant..	S'està carregant..
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

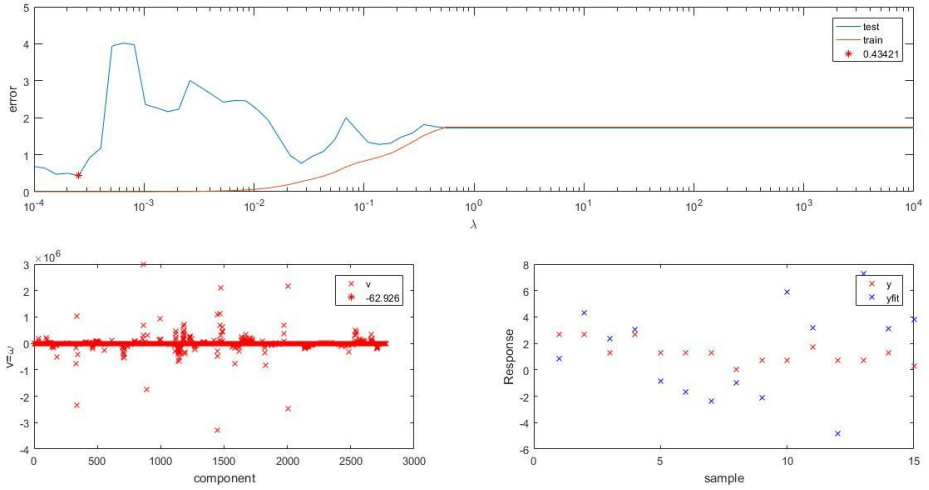
Variational Garrote

	TRP		TYR		CYS	
	S'està carregant..	S'està carregant..	S'està carregant..	S'està carregant..	S'està carregant..	S'està carregant..
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

Com podem veure en la taula anterior, els resultats no han millorat gaire, en el cas de la lasso, trobem solucions que no contenen valors diferents a zero i uns errors de predicció de test molt alts, això ens dóna a entendre que en general, el mètode no acaba de funciona gaire bé. El Variational Garrote també mostra el mateix problema amb el número de valors diferents de zero en la solució, no obstant, els seus error de test semblen ser millors contràriament al que esperaríem.

Per tal de poder mostra millor què és el que passa realment a les dades, hem generat la següent figura, on es mostra un exemple de un model calculat per la lasso i un model calculat pel Variational Garrote per un aminoàcid concret, en aquest cas el TYR.

Lasso:



Variational Garrote:

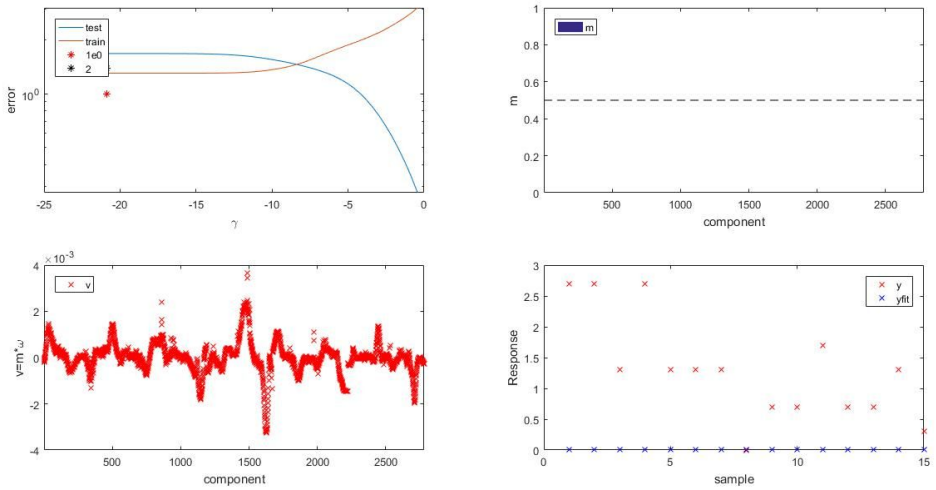


Figura 12: La figura de sobre representa el error d'entrenament i l'error de test, així com la seva S'està carregant... òptima. La figura de sota a l'esquerra correspon als coeficients de la solució S'està carregant..., que en el cas de la lasso S'està carregant...i en el cas del Variational Garrote S'està carregant.... La figura de sota a l'esquerra correspon a les respostes reals del conjunt d'entrenament vers les respostes del conjunt d'entrenament predites. En el cas del Variational Garrote, la figura de dalt a la dreta correspon a les m, podem veure com cap supera el ground truth.

Les gràfiques que podem observar a sobre són el resultat de fer validació creuada amb un aminoàcid concret. Primerament s'ha fet l'anàlisi tal com s'ha explicat anteriorment, les figures mostren la mitja de tots els *folds* de la validació creuada.

Amb la mitja dels pesos S'està carregant... s'ha calculat la resposta predita per al conjunt de validació. No obstant, com hem anat apreciant al llarg del document amb la excepció de la lasso, en l'experiment de l'apartat anterior, cap dels dos mètodes genera una solució prou bona alhora de predir la resposta, ja que, no és que la solució sigui massa o massa poc *sparse*, es que *l'intercept* sembla ser el únic valor rellevant alhora de predir els nous resultats, no obstant, aquest no és prou bo alhora de predir respostes noves, simplement és bo trobant un valor constant que minimitza el error en general.

Podem veure com la validació creuada no sembla aportar millores en el rendiment de l'algorisme. No obstant, això pot ser degut a les dades amb les que estem treballant. Veient que els mètodes no donen resultats gaire satisfactoris encara que treballem tenint en compte el tamany de les nostres dades, podem començar a pensar que el problema pot estar present en les dades d'entrada. Per tal de veure si hi ha certs problemes en les dades a predir, hem reescalat la resposta real de manera que puguem veure si el rendiment millora.

4.2.3 Variar paràmetres de la resposta:

En aquest apartat hem considerat modificar la resposta real al conjunt d'entrenament per tal de millorar la seva predicció. Els dos casos que hem considerat han sigut afegir soroll (S'està carregant...) i el cas en que centrem les dades S'està carregant..., on S'està carregant... es la mitja de S'està carregant.... La taula següent resumeix un experiment dut a terme en ambdós casos alhora (ja que en principi ambdós mètodes centren la resposta en el cas que no ho estigui):

Lasso

	TRP		TYR		CYS	
	S'està carregant..	S'està carregant..	S'està carregant..	S'està carregant..	S'està carregant..	S'està carregant..
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

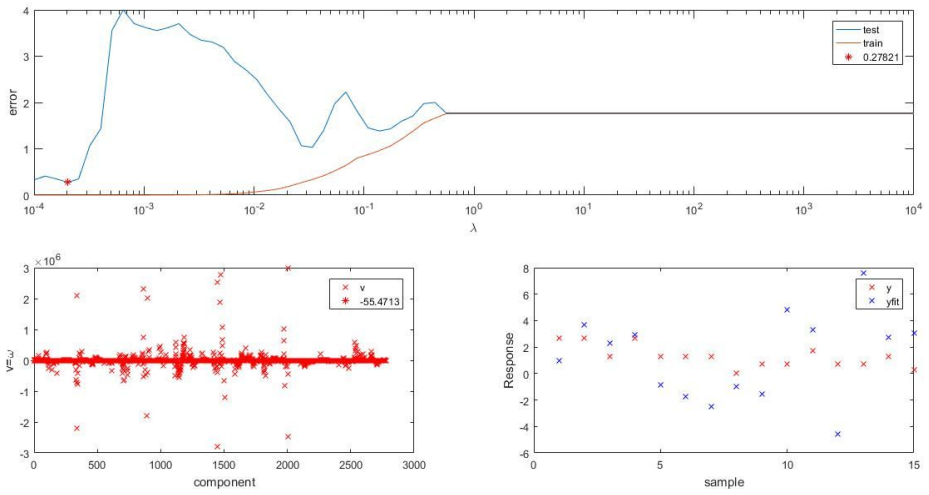
	TRP		TYR		CYS	
	S'està carregant..	S'està carregant..	S'està carregant..	S'està carregant..	S'està carregant..	S'està carregant..
MSE_train	1.0694	0.1382	1.3558	0.2579	1.3645	0.2371
MSE_test	0.7374	0.2342	0.8733	0.2800	0.9787	0.2597
NºZeros	1.0559e+03	230.0969	713.1676	430.1101	669.0102	341.1572
sum(v)	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

Com podem veure en la taula anterior, els canvis que hem fet en les dades no sembla haver millorat la rendiment dels mètodes. Si ens fixem en la *sparsity* de la solució, ambdós mètodes ofereixen solucions amb molts zeros, arribant a ser fins i tot solucions sense cap valor diferent a zero.

És interessant veure en les diferents repeticions d'un mateix experiment com a vegades la solució té algun valor diferent de zero. És en aquests casos on podem veure que la resposta predita ofereix valors no constants. Aquest casos són els més interessants. Respecte els errors de predicció del conjunt de test, podem veure que en aquest cas la lasso ofereix resultats bastant pitjor a l'experiment anterior, mentre que el Variational Garrote únicament perd una mica de rendiment.

Per tal de poder mostra millor què és el que passa realment a les dades, hem generat la següent figura, on es mostra un exemple de un model calculat per la lasso i un model calculat pel Variational Garrote per un aminoàcid concret, en aquest cas el TYR.

Lasso:



Garrote:

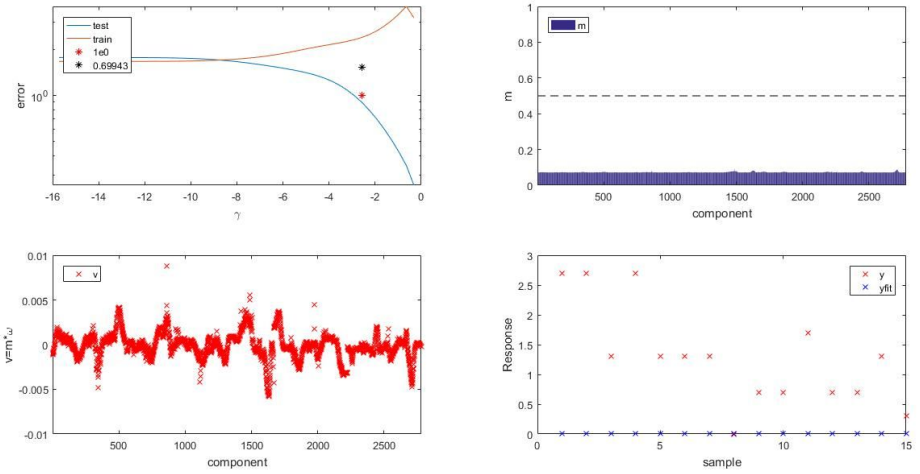


Figura 13: La figura de sobre representa el error d'entrenament i l'error de test, així com la seva S'està carregant... òptima. La figura de sota a l'esquerra correspon als coeficients de la solució S'està carregant..., que en el cas de la lasso S'està carregant...i en el cas del Variational Garrote S'està carregant.... La figura de sota a l'esquerra correspon a les respostes reals del conjunt d'entrenament vers les respostes del conjunt d'entrenament predites.

Com podem veure en les figures, ambdós mètodes ofereixen uns resultats bastant dolents. No obstant, si ens fixem en el gràfic de la lasso, podem veure que, quan afegim regularització, trobem solucions amb algun valor diferent de zero en la solució, i en aquests casos, la resposta predita no es constant. S'ha utilitzat la concentració TYR per a les gràfiques perquè era la més significativa.

En base als resultats que hem anat trobant, no podem assegurar que les dades siguin estimables amb models de regressió amb un rendiment prou bo. Sinó que, contràriament, sembla que els models no siguin capaços d'entendre de cap manera aquestes dades.

5. CONCLUSIONS

5.1 Conclusions

Durant aquest document hem desenvolupat un anàlisi de regressió a través de dos mètodes (Varitional Garrote i Lasso) a les dades recollides amb una llengua electrònica en un experiment de la Universitat Autònoma de Barcelona.

Per tal d'entendre aquests mètodes i constatar el seu rendiment s'han generat unes dades sintètiques en diferents condicions: diferents condicions de soroll en la resposta, diferents condicions de correlació en les dades i diferents graus de sparsity de la solució. Tant Lasso com Variational Garrote han ofert un bon rendiment en casos no extrems.

Posteriorment, s'ha aplicat els mètodes a les dades de concentacions reals. Primerament hem analitzat les dades conformant els conjunt tal com s'informava al document de l'experiment amb resultats molt dolents. Així doncs, s'ha decidit aplicat validació creuada per tal de millorar el model generat per a entendre les dades; no obstant els resultats no han millorat. Finalment, s'ha optat per modificar la senyal real de manera que estigués centrada i tingués una mica de soroll afegit per nosaltres; de la mateixa manera que en els casos anteriors, els resultats no han sigut gaire bons. En general, podem veure com els dos mètodes no troben solució a aquest problema de regressió: per una banda la lasso troba solucions on la majoria (sinó tots) els coeficients tenen valor zero; el Variational Garrote per altra banda, troba solucions molt denses però que no superen el ground truth del mètode. Cal destacar que els conjunt d'entrenament i de test mai han sigut sempre independents en un mateix experiment; cap mostra del conjunt que s'ha utilitzat d'entrenament s'ha utilitzat posteriorment per a testejar.

Un dels possibles motius pels quals no s'ha generat bons resultats pot ser per la pròpia naturalesa de les dades. Encara que aquest és el cas menys probable, pot ser que realment no hi hagi relació entre les dades captades per la llengua electrònica i la resposta real.

Una altre possible motiu seria que la llengua que van voler provar en aquest experiment (un nou model) no oferís unes dades molt fidedignes, amb un soroll considerable, fet que implicaria que aquests models degeneressin en rendiment.

Altrament, hem pogut veure en la matriu de correlació com les dades presentaven una alta correlació entre elles, fet que pot evocar collinealitat en les dades. Aquest fet, com hem vist en el cas cas de les dades sintètiques també empitjorava els resultats dels dos mètodes.

Un altre factor que pot haver afectat el rendiment és el format de les dades de l'experiment. Mentre que les dades d'entrenament només tenen tres valors: 0, 0,5 i 3; els valors per als experiments de test són molt més precisos. És a dir, pot ser amb aquestes dades d'entrenament només som capaços de generar models capaços de discernir entre res, poc i el màxim, i no siguin capaços de predir precisions més concretes.

És un fet que els dos mètodes mètodes degeneren molt en el seu rendiment alhora de tractar amb les dades reals en contrast amb el rendiment que ha mostrat durant els experiments amb les dades sintètiques. S'hauria de veure realment quin és el problema per tal de poder afrontar el problema d'una manera més correcta.

5.2 Future Work

Alhora de continuar amb l'anàlisi d'aquestes dades a través de regressió, seria interessant fer un anàlisi previ més exhaustiu de les dades per tal d'entendre quin problemes en poden poder estar causant. És podria treballar amb les dades per tal de intentar representar-les d'una altra manera: és podria intentar fer Principal Component Analysis (PCA) per tal de veure si podem representar els dades d'una manera més fàcil d'entendre per part dels algorismes.

Tanmateix, es podria considerar que els sensors ens aporten la mateixa informació tots. Sabem que quatre sensors són iguals i que un és diferents. Seria interessant veure com funcionaria el anàlisi si només utilitzéssim les dades captades per dos sensors diferents únicament. Aquest fet eliminaria correlació (i segurament redundància) en les dades i possiblement milloraria els resultats.

En últimes instàncies, es podrien modificar els algoritmes de manera que en comptes de prioritzar les solucions més *sparse*, trobés solucions una mica més densa, però que generessin models amb prediccions de respostes precises.

Bibliografia [arial 14 punts]

[1] Hilbert J. Kappen, Vicenç Gómez **"The Variational Garrote"** https://arxiv.org/pdf/1109.0486.pdf (2014)

[2] Georgina Faura, Andreu González-Calabuig, Manel del Valle **"Analysis of Amino Acid Mixtures by Voltametric Electronic Tongues and Artificial Neural Networks"**

[3] Tom M. Mitchell,1997 **"Machine Learning"**

[4] Hervé Abdi **"Partial Least Squares (PLS) regression"**

[5] Robert Tibshirani **"Regression Shrinkage and selection via the lasso"** https://statweb.stanford.edu/~tibs/lasso/lasso.pdf (1996)

[6] Xin Yan **"Linear Regression Analysis: Theory and Computing"** (2009)