Disseny i implementació d'un mòdul de preprocessament automàtic per l'entrada d'un model de DL

Gerard Atienza Reig

9 de febrer de 2026

Resum– Aquest treball presenta el disseny i la implementació del sistema *PTTN* (Pertorbació, Truncament, Transformació i Normalització) per automatitzar el preprocessament de variables d'entrada en models de *deep learning*. Es proposa un marc teòric basat en moments estadístics que formalitza criteris quasi-òptims de qualitat distribucional i defineix una mesura de bondat $G(X)$ per seleccionar de manera automàtica paràmetres de truncament i transformació, mantenint la normalització com a pas final d'estandardització. El sistema combina inferència automàtica i intervenció manual, assegurant la coherència del *pipeline* mitjançant actualitzacions dinàmiques i una interfície d'escriptori que facilita l'exploració i la validació visual dels resultats, reduint els recursos humans emprats actualment. A més, es presenten experiments que mostren com la mètrica de bondat tendeix a afavorir distribucions centrades, simètriques i amb cues moderades, i una avaluació qualitativa que suggereix que el preprocessament no empitjora el rendiment i pot aportar millores lleugeres. Finalment, la validació també inclou un rànquing de distribucions estadístiques i una avaluació preliminar amb xarxes neuronals (*MLP*), on s'ha demostrat que el sistema accelera la convergència, millora la generalització i redueix el temps d'entrenament. Tot i ser tests de caràcter qualitatiu, estableixen la base per a la validació a escala industrial ja acordada amb el departament d'Analítica de Negoci (AN) de GCO.

Paraules clau– preprocessament de dades; *deep learning*; moments estadístics; truncament; transformacions; normalització; pertorbació; automatització.

Abstract– This work presents the design and implementation of the *PTTN* (Perturbation, Truncation, Transformation, and Normalization) system to automate the preprocessing of input variables in *deep learning* models. A theoretical framework based on statistical moments is proposed, formalizing quasi-optimal criteria for distributional quality and defining a goodness measure $G(X)$ to automatically select truncation and transformation parameters, while preserving normalization as the final standardization step. The system combines automatic inference and manual intervention, ensuring pipeline consistency through dynamic updates and a desktop interface that facilitates visual exploration and validation of the results, reducing the human resources currently employed. Additionally, experiments are presented showing that the goodness metric tends to favor centered, symmetric distributions with moderate tails, along with a qualitative evaluation suggesting that the preprocessing does not degrade performance and may yield slight improvements. Finally, the validation also includes a ranking of statistical distributions and a preliminary evaluation using neural networks (*MLP*), where the system has been shown to accelerate convergence, improve generalization, and reduce training time. Although these tests are qualitative in nature, they establish the foundation for industrial-scale validation already agreed upon with the Business Analytics department (AN) at GCO.

Keywords– data preprocessing; deep learning; statistical moments; truncation; transformations; normalization; perturbation; automation.

• E-mail de contacte: 1636435@uab.cat
• Menció: Computació
• Treball tutoritzat per: Jordi Serra Ruiz
• Curs 2025/26

1 INTRODUCCIÓ I OBJECTIUS

El preprocessament de dades constitueix una de les etapes més determinants en el rendiment i l'eficiència dels models de *deep learning*, tot i que sovint és abordat mitjançant plantejaments poc sistemàtics o insuficientment fonamentats. En aquest context, el departament d'Analítica de Negoci (que d'ara endavant denominarem AN) del Grup Catalana Occident, va traslladar al departament d'Arquitectura de *machine learning* **la necessitat de millorar la metodologia actual de preprocessament**. Sent AN el responsable de l'aplicació de tècniques de ciència de dades i aprenentatge profund per predir, analitzar i optimitzar diversos processos de negoci.

Concretament, es va identificar que el procés vigent estava fortament limitat per procediments manuals, una manca de consens metodològic i una elevada dependència de la intuïció i l'experiència individual de l'analista. En l'estat actual, es realitza un pas de pertorbació opcional per a totes les variables, seguit d'un flux addicional de truncament, transformació i normalització (*TTN*) aplicat exclusivament a les variables numèriques.

Gràcies a **entrevistes i correspondència amb el departament**, s'han identificat diverses ineficiències del procés usat actualment. Per a la definició dels paràmetres associats a aquests passos, els científics de dades d'AN usen *scripts* en *R* i *Python* que s'apliquen variable a variable. En aquests *scripts*, els paràmetres de preprocessament s'ajusten mitjançant un procés iteratiu basat en la intuïció i el refinament empíric, usant com a guia qualitativa diverses representacions gràfiques generades durant l'execució. Aquest enfocament, a banda de no disposar d'estàndards quantitatius que permetin justificar i consensuar les decisions preses respecte als valors dels paràmetres i el seu efecte posterior en el rendiment del model, comporta també un cost molt elevat en termes de temps i recursos humans.

Davant d'aquesta situació, l'objectiu d'aquest projecte s'articula al voltant de tres grans línies de millora respecte a l'estat actual del departament.

En primer lloc, es pretén desenvolupar un sistema capaç de **generar de manera automàtica i selectiva el flux de preprocessament adequat per a les diferents variables**, amb l'objectiu de poder executar aquest procés de forma massiva sobre grans volums de dades. Aquesta automatització ha de permetre reduir de manera significativa la latència associada al preprocessament i els recursos humans actualment necessaris per dur-lo a terme.

En segon lloc, es vol **definir una metodologia estandarditzada** que permeti automatitzar la cerca de paràmetres de preprocessament quasi òptims en aquells passos repetitius i no fortament dependents del context específic de cada projecte. Aquesta metodologia s'ha de basar en un marc teòric robust i ben fonamentat, que proporcioni un criteri objectiu i consensuable per justificar l'elecció dels valors concrets associats a cada pas de preprocessament automatitzat.

Finalment, el projecte té com a objectiu el desenvolupament d'una **interfície gràfica en forma d'aplicació d'escriptori** que permeti al científic de dades d'AN importar conjunts de dades i governar de manera integral tot el procés de preprocessament, alhora que disposa d'eines d'automatització. En particular, l'aplicació ha de permetre automatitzar determinats passos del flux, comprovar de manera gràfica i quantitativa els resultats obtinguts, modificar manualment paràmetres de preprocessament i observar en temps real el seu efecte sobre les distribucions resultants o intermèdies. Addicionalment, s'han d'oferir funcionalitats per visualitzar l'impacte de cada pas sobre l'estat anterior del preprocessament, consultar mètriques de qualitat o adequació de la distribució final, exportar de manera massiva els paràmetres de preprocessament de totes les variables i automatitzar el tractament en *batch* de grans grups de variables de manera simultània.

En conjunt, aquest projecte pretén empoderar el científic de dades amb una eina que redueixi significativament el temps invertit en el preprocessament, millori la qualitat i la consistència dels resultats obtinguts i garanteixi que el procés sigui replicable, estandarditzat, consensuat i teòricament fonamentat.

En definitiva, els objectius plantejats en aquest treball busquen transformar un procés de preprocessament actualment limitat, costós i fortament dependent del criteri individual, en un sistema formalitzat, automatitzable i justificable des d'un punt de vista teòric i pràctic.

2 ESTAT DE L'ART

El preprocessament de dades en moltes ocasions en entorns industrials ha estat relegat a un conjunt de bones pràctiques transmeses per experiència més que no pas a un camp d'estudi amb entitat pròpia. Encara que hi ha literatura relativament extensa, moltes pràctiques s'han consensuat per inèrcia i amb tècniques que s'han considerat evidents, no obstant no resulta tant trivial la seva elecció des d'un punt rigorós. Tanmateix, en el context del *deep learning*, aquesta etapa té un impacte directe tant en l'estabilitat del procés d'entrenament com en la capacitat de generalització dels models. A més, existeix un factor corporatiu rellevant, tant en l'eficiència, replicabilitat i la sistematització, com en un seguit de diverses mètriques econòmiques com els recursos humans, tecnològics i capitals emprats per a la generació d'aquests models, en els quals **un preprocessament de dades adequat pot tenir un impacte significatiu**.

Treballs fundacionals com *Efficient Backprop* (LeCun et al., 1998) ja posaven de manifest la importància de certes propietats de les dades d'entrada, especialment el centrament al voltant de zero, l'escalat homogeni de la variància i la reducció de correlacions entre variables. Aquests criteris estan estretament relacionats amb la dinàmica del descens de gradient, l'optimització de la convergència i la prevenció de fenòmens com la saturació de funcions d'activació o la propagació ineficient del gradient.

Com a conseqüència directa d'aquests resultats, s'ha consolidat l'ús de tècniques de normalització que projecten les dades cap a una distribució normal estàndard. Aquesta pràctica s'ha convertit gairebé en un estàndard *de facto*, especialment en *pipelines* a gran escala, tot i que, després d'una extensa recerca en la literatura, aquesta no ofereix una demostració formal de la seva optimalitat universal.

De fet, la creixent complexitat i dimensió dels models actuals porta a la intuïció que la relació entre la forma de la distribució d'entrada i el rendiment del model ha de ser més complexa. Determinades arquitectures neuronals, funcions d'activació o dominis d'aplicació poden beneficiar-se de distribucions amb propietats diferents de la normalitat

estricta. Això ha conduit a aquest projecte a **explorar propietats estadístiques de més baix nivell**, que es on es troben les característiques fonamentals i la informació latent de les distribucions. Agafant suport de la investigació en els articles existents i en proves empíriques, aquest projecte s'ha recolzat en els moments estadístics, mètriques que han resultat ser potencials criteris optimitzadors de les distribucions.

3 METODOLOGIA

3.1 Presentació del sistema de preprocessament *PTTN*

El sistema de preprocessament *PTTN* (Pertorbació, Truncament, Transformació i Normalització) es defineix com una **arquitectura modular orientada al tractament de variables numèriques** prèvies a l'entrenament de models de *deep learning (DL)*. El disseny del sistema defineix una separació explícita entre les fases d'intervenció humana i els processos d'inferència automatitzada, amb l'objectiu de combinar criteri expert i optimització estadística dins d'un mateix flux de treball.

L'arquitectura es divideix en dos mòduls principals, diferenciats segons la seva naturalesa operativa i el tipus de decisió que admeten:

- **Mòdul 1: Pertorbació** — configuració manual i interactiva.
- **Mòdul 2: Truncament, transformació i normalització** — inferència automatitzada de paràmetres i optimització estadística.

Aquesta divisió permet independitzar la fase de pertorbació, de caràcter inductiu i dependent del context, de la fase automàtica, gestionada per criteris quantitatius i mètriques formals definides en un marc teòric.

3.2 Mòdul 1: Pertorbació — configuració interactiva

El mòdul de pertorbació constitueix el primer pas del *pipeline* de preprocessament. Aquest pas no admet inferència automàtica dels seus paràmetres, ja que les operacions de pertorbació (augmentació de dades) responen a decisions de caràcter contextual, fortament vinculades als objectius del model, al domini d'aplicació i a la intenció del científic de dades.

Per aquest motiu, el sistema proporciona una **interfície interactiva que permet a l'usuari definir manualment els paràmetres de pertorbació** i visualitzar en temps real l'efecte d'aquestes operacions sobre la distribució de les dades. Aquesta visualització immediata facilita l'anàlisi qualitativa dels canvis introduïts i afavoreix la formació de decisions informades sobre l'impacte a les distribucions. Cal destacar que també es proporciona la configuració de perturbació nul·la (per defecte), on la sortida de dades d'aquest pas és exactament la d'entrada, permetent ometre aquest pas si el projecte ho requereix.

La sortida d'aquest mòdul es considera l'estat inicial de les dades sobre el qual s'aplica posteriorment el procés d'inferència automatitzada del segon mòdul en les variables numèriques. En conseqüència, la forma distribucional resultant després de la pertorbació condiciona directament les distribucions resultants de les etapes següents.

3.3 Mòdul 2: Truncament, transformació i normalització — inferència automatitzada i ajust dinàmic

El segon mòdul del sistema integra els passos de truncament, transformació i normalització, i constitueix el nucli automatitzat del preprocessament. En aquesta etapa, el sistema **analitza les propietats estadístiques de les dades per inferir paràmetres** que aproximin una distribució de sortida considerada quasi òptima segons els criteris definits al marc teòric i la mètrica de bondat associada al marc teòric.

Inicialment, el disseny del sistema contemplava la normalització com a part del procés d'automatització. No obstant, aquest pas ha estat exclòs de l'optimització automàtica, degut a que resulta indiferent respecte a la mètrica de bondat basada en moments estadístics normalitzats. Aquesta decisió queda justificada formalment a l'Apèndix (A.4). En conseqüència, **la normalització queda fora de l'espai d'optimització i es considera una operació final d'estandardització**. No obstant, aquest darrer pas permet a l'usuari canviar el mode de normalització entre dues configuracions: mitjana amb desviació estàndar o mediana amb rang interquartílic; així ho va sol·licitar explícitament AN.

El sistema permet tant l'execució completament automàtica del mòdul (ometent la normalització) com la intervenció manual de l'usuari sobre qualsevol dels paràmetres inferits. Qualsevol modificació manual activa un recàlcul dinàmic dels passos posteriors i la compilació de les seves gràfiques, per garantir la coherència del *pipeline*. Per exemple, una modificació en el truncament força l'actualització en temps real de la transformació i la normalització, mentre que no afecta els passos anteriors.

Pel que fa a l'automatització, el sistema admet dues estratègies d'inferència, seleccionables segons el context d'ús:

- Inferència conjunta de truncament i transformació, mitjançant una *grid search* sobre les transformacions suportades per l'arquitectura vigent (tals com logarítmiques, potencials, o de la família *Box-Cox*, entre d'altres) i un conjunt discret de valors de truncament predefinits pel departament d'AN.
- Inferència de la transformació únicament.

En ambdós casos, la mètrica de bondat del marc teòric actua com a criteri d'optimització per seleccionar la combinació de paràmetres que millor s'ajusta als objectius estadístics definits.

3.4 Plantejament i recerca per al marc teòric

3.4.1 Motivació per als criteris d'aproximació a l'optimalitat

En primera instància, és possible considerar que la mecànica del segon mòdul és anàloga a un **problema d'optimització**, ja que, es parteix d'una distribució de dades inicials i es busca obtenir-ne una versió aproximadament òptima com a entrada per a un model de *DL*. Conseqüentment,

desconeixem quina és l'entrada concreta al *PTTN* i els passos òptims intermedis, però és plausible definir una sortida aproximadament òptima, ja que la intenció es definir criteris ideals d'ingesta de dades per a un model genèric de *DL*. És a dir, *a priori* sembla possible definir una solució teòricament òptima, però no les solucions parcials òptimes ni el punt de partida concret. Per tant, s'ha d'abordar des del final (la sortida òptima objectiu) i construir retrospectivament les transformacions necessàries per aconseguir-la, aplicant *reverse engineering*.

Tot i que el preprocessament no és un problema de decisió dinàmica en el sentit clàssic, el disseny del mòdul *TTN* s'inspira en la lògica de la descomposició seqüencial. Seguint una línia de pensament anàloga al Principi d'Optimalitat de *Bellman*, s'assumeix que per assolir una configuració de sortida òptima segons una mètrica, cadascuna de les etapes del *pipeline* ha d'actuar com un optimitzador local sobre propietats específiques de la distribució, usant de guia un conjunt de criteris predefinits com a desitjables. (Bellman, 1957)

3.4.2 Criteris d'optimalitat

En aquest context, el terme optimalitat s'usa en el sentit operatiu d'un problema d'optimització: no pressuposa una definició axiomàtica d'òptim, sinó **el compliment i maximització d'un conjunt de propietats estadístiques desitjables**. Així, l'"optimalitat" fa referència al grau d'alineació amb aquestes mètriques, i no a l'existència d'un estat universalment òptim de la distribució.

Per formalitzar una sortida objectiu, s'hipotetitza un conjunt de **criteris estadístics d'optimalitat** que defineixen l'estat desitjat de la distribució resultant. Aquests criteris són (i) l'objectiu del procés d'optimització i (ii) la guia per definir mètriques d'avaluació dins del *pipeline*.

Tradicionalment, s'ha assumit la normal estàndard com a referència preferent per a la ingestió de dades en models genèrics de *DL*; tanmateix, en la revisió realitzada **no s'ha trobat cap evidència directa que n'estableixi l'optimalitat universal**. No obstant, sí que s'observa és que algunes propietats que afavoreixen l'entrenament (p. ex., mitjana propera a 0 i variància controlada) són característiques de la normal estàndard.

Això suggereix que la "hegemonia" de la normal pot ser massa restrictiva: altres distribucions (p. ex., *Laplace*, *t* de Student, *Poisson*, beta simètrica, etc.) podrien ser igualment adequades si preserven propietats estadístiques clau. Per aquest motiu, el plantejament consisteix a extreure i formalitzar les característiques que fan útil la normal, i usar-les com a **mètriques d'optimalitat**. D'aquesta manera, si la normal estàndard resulta ser realment l'òptim universal, el preprocessament hi convergiria; en cas contrari, emergirien distribucions diferents però coherents amb el marc.

En conseqüència, es parteix de la hipòtesi que **no hi ha evidencies suficients que justifiquin l'ús de la normal estàndard com a entrada òptima universal** per a models de *DL*.

A partir d'aquesta hipòtesi, s'han recollit **criteris candidats** inspirats en propietats essencials de la normal (Ross, 2014) i en criteris habituals per facilitar l'entrenament (p. ex., centrament i control de l'escala) (LeCun et al., 1998) (Apèndix A.1). En conjunt, aquests criteris (centrament, escala, continuïtat, variància controlada, curtosi moderada, etc.) defineixen una estructura mínima perquè la distribució d'entrada pugui considerar-se quasi-òptima i promogui un aprenentatge estable i equilibrat.

Finalment, per sistematitzar la selecció i l'avaluació d'aquests criteris, es proposa un marc teòric que no assumeix optimalitat absoluta, sinó que permet justificar candidats directament (via literatura), empíricament (experimentalment) o indirectament (via propietats emergents del mateix marc), sempre que aquest demostri ser prou robust.

3.4.3 Fonaments del marc teòric

Per sistematitzar els criteris, e**s proposa un marc basat en els moments estadístics de la distribució**, amb l'objectiu de capturar la interdependència entre propietats, ja que no són independents i els seus efectes es condicionen mútuament. En lloc d'avaluar cada criteri de manera aïllada, el marc planteja descriure la distribució mitjançant una representació jeràrquica.

Els moments permeten aquesta formalització amb múltiples ordres on cada un resumeix una dimensió complementària: la localització, la dispersió, l'asimetria, la curtosi, l'hiperasimetria, l'hipercurtosi, etc. (Apèndix A.2). D'aquesta manera, els moments no només resumeixen la distribució, sinó que també proporcionen una base quantificable per integrar i comparar els efectes combinats de les transformacions aplicades (Kendall & Stuart, 1946).

3.5 Mesura de bondat de la distribució

Per tal de convertir el marc teòric en una eina pràctica i mesurable, es proposa definir una **mesura de bondat** $G(X)$ que sintetitzi totes les propietats desitjables de la distribució X. Aquesta mesura serà la materialització real del nostre marc teòric i permetrà comparar i optimitzar distribucions d'entrada dins del *pipeline* del mòdul 2 (*TTN*). Aquesta mètrica té la intenció de quantificar el que en algun moment s'ha esmentat com la 'qualitat' de la distribució de les dades.

Sigui X una variable aleatòria amb distribució d'entrada al model. Recolzant-nos en els moments del 3r al 6è, es proposa una definició de mesura de bondat com:

$$G(X) = \sum_{i=3}^{6} w_i \cdot g_i(X) \tag{1}$$

on cada $g_i(X) \in [0, 1]$ representa una funció de compliment per a cada criteri:

- $g_3(X)$: mesura si la *skewness* és inferior al llindar establert amb la desigualtat de *Pearson* ($\sqrt{2}$) (Pearson, 1916).
- $g_4(X)$: mesura la proximitat del valor de la curtosi a 3 (*mesocúrtica* òptima) (Larasati et al., 2018), formalitzant concentració i forma de cues.
- $g_5(X)$ i $g_6(X)$: mesuren el control sobre valors extrems i acotament.

Els pesos $w_i \geq 0$, amb la restricció $\sum_i w_i = 1$, permeten ajustar la rellevància relativa de cada criteri segons la seva importància teòrica i experimental. La normalització de cada funció a l'interval $[0, 1]$ garanteix que cada criteri contribueixi proporcionalment, evitant dominàncies indesitjades.

Aquesta mesura de bondat proporciona:

1. Una **quantificació objectiva** de com de ben ajustada està la distribució d'entrada als criteris quasi-òptims definits pel marc teòric.
2. Una base per a **optimització automàtica** dins del segon mòdul (*TTN*), ja que cada pas pot ser ajustat per maximitzar $G(X)$.
3. La possibilitat de **comparar distribucions** i seleccionar aquella que millor compleixi els criteris teòrics i experimentals.

La definició completa de la mesura de bondat i els seus càlculs i metodologies específiques queden detallats a l'apèndix (A.3). En síntesi, aquesta mètrica **proporciona un criteri únic, objectiu i comparable per facilitar l'automatització.** A continuació es presenta la interfície d'escriptori que materialitza aquest criteri i permet governar el procés de manera visual i interactiva.

3.6 Aplicació d'escriptori

Com a part integrant de la metodologia, aquest projecte desenvolupa **una aplicació d'escriptori que serveix com a interfície gràfica** per al sistema *PTTN*. L'objectiu de la interfície és permetre als científics de dades explorar, ajustar i validar els paràmetres del preprocessament de manera visual i interactiva, sense necessitat d'interactuar directament amb codi. D'aquesta manera, es facilita un flux de treball eficient, intuïtiu i reproduïble.

3.6.1 Objectius i requeriments de la interfície

La interfície es dissenya seguint els requisits consensuats amb el departament d'AN, prioritzant:

- La càrrega de conjunts de dades i la classificació de variables segons el seu tipus.
- La representació gràfica en temps real de les distribucions de cada variable.
- La aplicació dels quatre passos de preprocessament de manera modular i ajustable per l'usuari.
- La coherència del flux de preprocessament: qualsevol modificació en un pas actualitza automàticament els passos posteriors i les representacions gràfiques.
- La automatització flexible dels passos de preprocessament. Es pot automatitzar un sol pas d'una variable, tots els passos d'una variable individual o tots els passos de totes les variables simultàniament.
- L'arxiu de les variables ja preprocessades, mantenint el control del procés.
- L'exportació de tots els paràmetres de preprocessat de totes les variables en un sol format.

3.6.2 Suport per a variables categòriques i binàries

A més de les variables numèriques, la interfície suporta de manera separada variables categòriques i binàries. **Aquestes variables poden ser sotmeses a una pertorbació opcional** que redefineix aleatòriament la seva categoria. El procés es basa en una distribució de probabilitat de reclassificació que segueix una funció normal sobre l'índex de les categories disponibles, introduint així un component de soroll controlat en les dades discretes. La pertorbació inclou també un component de probabilitat individual per cada dada, de manera que no totes les observacions necessàriament es veuen afectades, introduint així més complexitat a l'augmentació de dades.

Aquest mecanisme permet explorar la robustesa del model davant canvis aleatoris de categoria i simula escenaris de soroll en dades discretes, i **satisfà la necessitat d'AN** de poder gestionar tots els tipus de dades que els seus experiments suporten.

3.6.3 Implementació i funcionalitat

L'aplicació ha sigut desenvolupada en *Python* utilitzant **PyQt**, aprofitant les seves capacitats per a interfícies modulars, interactives i amb actualització en temps real. La interfície funciona com un entorn integrat que:

- Permet carregar les dades i els fitxers de configuració, amb validació d'errors i format.
- Visualitza la distribució de cada variable abans i després de cada pas de preprocessament.
- Integra els mòduls de *PTTN* de manera que el resultat de cada pas serveix com a entrada per al següent, assegurant la consistència i traçabilitat del procés.
- Permet la pertorbació de variables numèriques, categòriques i binàries, amb control sobre la magnitud i la probabilitat de modificació.
- Ofereix opcions d'automatització parcial o completa dels passos de truncament i transformació, utilitzant la mètrica de bondat com a criteri d'optimització.
- Resumeix els resultats finals amb mètriques estadístiques i el *score* de bondat, permetent l'arxiu de les variables processades.
- Permet exportar els paràmetres de preprocessat de cada variable en un fitxer de format *YAML*.

D'aquesta manera, la interfície facilita l'exploració i el control manual del procés, alhora que integra les optimitzacions estadístiques definides a la metodologia, assegurant un flux de treball coherent i reproduïble per a tot tipus de variables.

4 RESULTATS

4.1 Avaluació comparativa de la bondat de diferents distribucions estadístiques

Amb l'objectiu de validar empíricament els criteris teòrics de bondat s'ha dut a terme un experiment computacional en *Python* orientat a avaluar i analitzar diferents distribucions estadístiques habituals segons la seva adequació morfològica. Aquest estudi no pretén establir un criteri definitiu d'optimalitat, sinó **aportar evidència empírica i intuïtiva** que doni suport a les decisions conceptuals adoptades en el marc teòric.

L'experiment s'ha implementat mitjançant un únic *script* en *Python*. S'han generat mostres de diferents distribucions estadístiques típiques, entre les quals s'inclouen, entre d'altres: normal estàndard, beta simètrica, *Laplace*, logística, *Weibull*, *Rayleigh*, exponencial, *chi*-quadrat, log-normal, *Pareto*, *Cauchy*. etc.

Per a cada distribució s'han calculat les quatre mètriques estadístiques referents als moments que composen la mesura de Bondat: l'asimetria (*skewness*), la curtosi, l'*hiperskewness* i l'hipercurtosi.

Seguint la formalització del marc teòric, a partir d'aquestes mètriques s'ha definit el valor escalar de la **bondat** que **permet ordenar les distribucions segons la seva adequació global.** Aquesta mesura es construeix com una combinació ponderada de les quatre mètriques, assignant un pes de $0{,}35$ a la *skewness* i a la curtosi, i un pes de $0{,}15$ a la *hiperskewness* i a la hipercurtosi.

Un cop calculades les mètriques i la bondat associada, s'ha representat gràficament, per a cada distribució, tant la seva forma empírica com els valors de les quatre mètriques i el valor final de bondat. A la següent figura es mostren exemples representatius de les distribucions analitzades, juntament amb els valors de les mètriques.

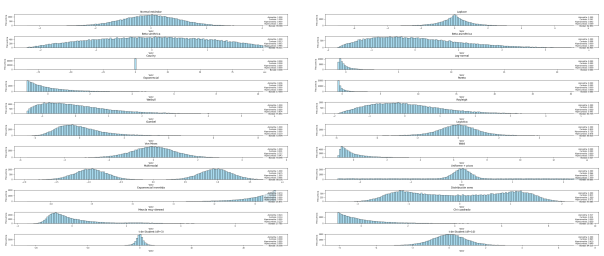

Fig. 1: Representació de diferents distribucions estadístiques generades amb `NumPy`, juntament amb els valors de les mètriques de *skewness*, curtosi, hiperskewness, hipercurtosi i la bondat agregada.

Finalment, totes les distribucions s'han ordenat de major a menor segons el seu valor de bondat, i s'ha representat conjuntament l'evolució de la bondat i de cadascuna de les quatre mètriques al llarg del rànquing. Aquesta visualització permet observar de manera clara com les distribucions amb formes més regulars i equilibrades tendeixen a ocupar les primeres posicions, mentre que distribucions amb cues pesades, asimetries extremes o moments no finits queden relegades a les últimes posicions.

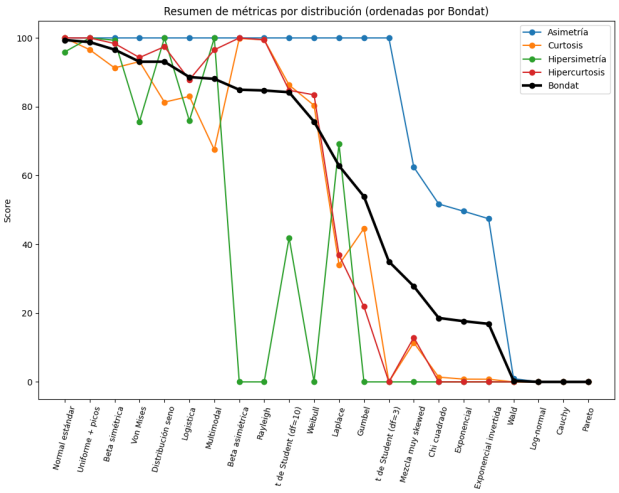

Fig. 2: Rànquing de distribucions ordenades segons la bondat agregada. Es mostren simultàniament els valors normalitzats de la bondat i de les quatre mètriques estadístiques utilitzades.

Aquest estudi experimental reforça la idea que la bondat d'una distribució no pot dependre d'una única mètrica, sinó que emergeix de l'equilibri entre simetria, concentració central i control dels extrems. Així mateix, justifica l'ús d'una mesura agregada i ponderada com a base conceptual per la automatització del procés de preprocessament (*PTTN*).

A més, s'observa que la mesura de bondat **tendeix a afavorir i fer convergir el rànquing cap a distribucions morfològicament semblants a la normal estàndard** (o, més generalment, cap a distribucions centrades, simètriques i amb cues moderades), sense imposar-la explícitament com a solució. Això constitueix, en part, una demostració empírica indirecta de l'adequació pràctica de la normal com a referència habitual en *DL*, però alhora és un indicador que el nostre marc teòric basat en moments és lògic, coherent i operatiu: és capaç de capturar i reforçar positivament distribucions equilibrades que compleixen els criteris desitjables recolzats per la literatura, i de penalitzar aquelles que presenten deficiències latents.

4.2 Avaluació qualitativa del rendiment

Per tal d'avaluar si el preprocessament proposat millora, empitjora o no afecta el rendiment dels models, s'ha dissenyat un experiment deliberadament senzill amb l'objectiu d'analitzar de manera qualitativa l'impacte d'aquest procés. Tot i que l'experiment mesura i compara diverses mètriques, la seva naturalesa limitada i no exhaustiva fa que els resultats s'interpretin principalment des d'una **perspectiva qualitativa**, inferint tendències generals sobre el rendiment

en comparació amb un entrenament basat en dades no preprocessades amb aquesta eina, més enllà d'una normalització estàndard.

Per aquest motiu, s'ha utilitzat un conjunt de dades genèric àmpliament emprat en experiments de validació, on l'objectiu no és obtenir bones mètriques predictives *per se*, sinó avaluar el comportament relatiu de les eines de preprocessament. El *dataset* seleccionat com a conjunt de control ha estat el conegut *Wine Quality*.

S'han construït dos conjunts de dades: el primer correspon al *dataset* original, emprat com a control, al qual només s'ha aplicat tractament de valors nuls i una normalització per mitjana i desviació estàndard; el segon conjunt també inclou el tractament de nuls, però ha estat extret al final de la *pipeline PTTN*, després d'aplicar una automatització global sobre totes les variables.

Sobre ambdós conjunts de dades **s'han entrenat diversos models d'aprenentatge automàtic**, com ara *XGBoost*, arbres de decisió i màquines de vectors de suport, amb l'objectiu de classificar els vins segons el seu tipus (blanc o negre). Els resultats obtinguts es mostren a les següents figures.

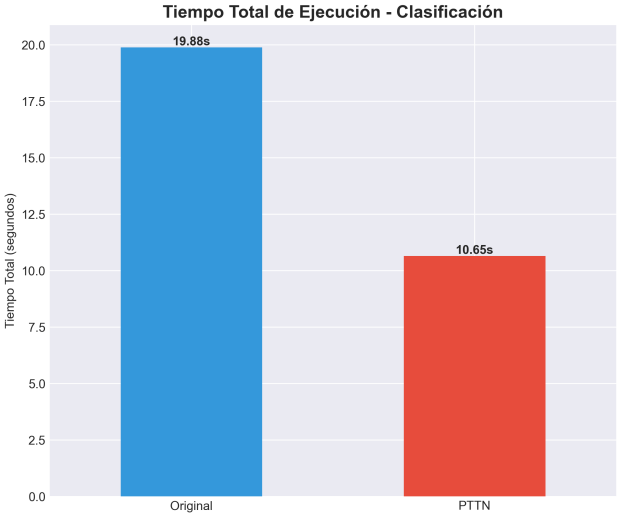

Fig. 3: Gràfica comparativa de temps d'execució (mitjà)

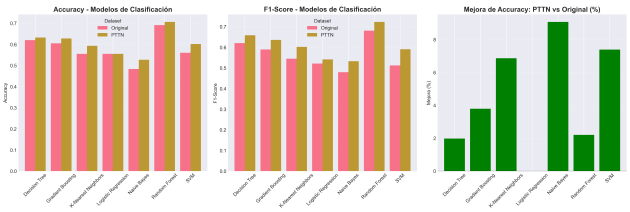

Fig. 4: Gràfiques comparatives del rendiment de diferents models per als dos *datasets*

La lleugera millora observada en el preprocessament *PTTN* respecte al conjunt de control probablement es deu al fet que el truncament redueix la influència dels *outliers*, mentre que les transformacions aplicades eliminen distribucions poc òptimes amb geometries irregulars, convertint-les en distribucions unimodals, centrades i escalades, amb un grau adequat de simetria i curtosi. Pel que fa al temps d'execució, és raonable inferir que aquests mateixos factors que afavoreixen el rendiment predictiu també contribueixen a una millor convergència dels models, i per tant redueix el nombre d'iteracions i, conseqüentment, el temps d'execució.

Aquest experiment no permet concloure que el preprocessament sigui clarament superior en termes de rendiment, ni que produeixi una millora significativa en la convergència, ja que es tracta d'un estudi limitat, amb recursos computacionals reduïts i una quantitat i diversitat de dades moderades. No obstant això, sí que confirma que, com a mínim, l'aplicació del preprocessament no empitjora els resultats, i reforça una intuïció generalment acceptada: **les distribucions truncades, lliures de valors extrems, i transformades cap a geometries més regulars tendeixen a afavorir tant la convergència com el rendiment dels models**.

Finalment, s'ha dut a terme una **validació preliminar utilitzant arquitectures de xarxes neuronals** (*MLP*) per observar el comportament del sistema en un entorn de *deep learning*. Cal subratllar que, a causa de la simplicitat de l'entorn experimental, aquest test s'ha de considerar com una aproximació superficial i de naturalesa principalment qualitativa. L'objectiu no és validar mètriques absolutes, sinó identificar tendències i comportaments que el *pipeline PTTN* introdueix en el procés d'aprenentatge. En aquest sentit, ja s'ha acordat amb el departament d'AN que l'avaluació definitiva es realitzarà en un marc experimental d'escala industrial i corporativa, utilitzant grans volums de dades i arquitectures més complexes.

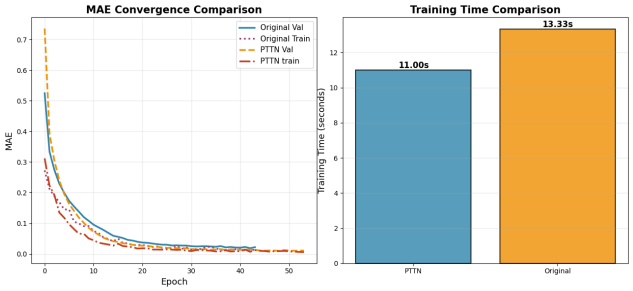

Fig. 5: Gràfiques comparatives de temps i convergència d'una arquitectura elemental *MLP*, per als dos *datasets*

Tot i el caràcter exploratori d'aquesta validació preliminar, els resultats obtinguts revelen tendències consistents que reforcen el valor del sistema *PTTN*, destacant especialment una convergència accelerada on el model redueix l'error (*MAE*/*MSE*) amb major celeritat tant en el conjunt d'entrenament com de validació.

Aquesta millora, especialment visible en les primeres èpoques, **suggereix que el preprocessament optimitza la geometria de l'espai d'entrada i actua com un inductive bias** que facilita l'estabilitat inicial de l'aprenentatge. Així mateix, s'ha observat una millora en la generalització gràcies a un marge reduït entre les corbes de train i validation, la qual cosa apunta a una regularització implícita que genera representacions internes més robustes i mitiga el sobreajust.

En l'àmbit de l'eficiència computacional, els tests han registrat una reducció del temps d'entrenament amb *early stopping* d'entre el 17% i el 20%, un factor crític en entorns industrials que permet reduir el nombre d'èpoques necessàries per convergir. Aquests valors són orientatius i s'han de prendre amb caràcter qualitatiu ja que no hi ha garanti-

es que aquests valors de millora es mantinguin exactament igual en experiments més complexos.

Finalment, s'ha constatat l'estabilitat del rendiment final, on el sistema assoleix una precisió lleugerament superior sense introduir les penalitzacions habituals de les tècniques de preprocessament mal ajustades.

Aquests resultats, tot i ser de naturalesa qualitativa i basats en un test superficial, serveixen com a **evidència empírica de les tendències de millora del sistema i constitueixen la base per als experiments industrials i rigorosos ja acordats amb el departament d'AN.**

5 Conclusions

Aquest treball ha abordat la necessitat d'automatitzar, governar i estandarditzar el preprocessament de dades en entorns corporatius, iterant el sistema *PTTN* replantejant-lo com a flux modular de pertorbació, truncament, transformació i normalització, orientat a preparar variables d'entrada per a models de *deep learning*. El resultat principal és una **arquitectura que combina eines per al control expert amb inferència automàtica de paràmetres**, i que garanteix la coherència del *pipeline* mitjançant actualitzacions dinàmiques dels passos posteriors.

Des del punt de vista teòric, s'ha definit un **marc d'avaluació basat en moments** que evita imposar la normal estàndard com a solució *a priori*, i formalitza el concepte de "qualitat" d'una distribució com l'equilibri entre simetria, concentració central i control d'extrems. A partir d'aquest marc, s'ha proposat una **mesura de bondat** $G(X)$ que permet comparar distribucions i governar l'optimització automàtica del mòdul *TTN* amb un criteri únic i mesurable.

En la validació empírica, l'experiment de rànquing de distribucions ha mostrat que la mesura de bondat tendeix a afavorir distribucions centrades, simètriques i amb cues moderades, i que el rànquing convergeix de manera natural cap a famílies morfològicament properes a la normal sense imposar-la explícitament. Addicionalment, l'avaluació qualitativa en eines de *ML* amb el *dataset Wine Quality* suggereix que el preprocessament *PTTN*, com a mínim, **no empitjora el rendiment** i pot aportar millores lleugeres associades al control d'*outliers* i a la regularització de les distribucions.

Aquests resultats s'han vist reforçats per **una validació preliminar amb xarxes neuronals** (*MLP*), on s'ha observat que el preprocessament *PTTN* actua com un *inductive bias* que, *a priori*, accelera la convergència, millora la generalització i redueix el temps d'entrenament. Tot i que aquests resultats són de caràcter exploratori, estableixen una base sòlida per a les línies futures ja acordades amb el departament d'AN, que inclouen: (i) la validació a escala industrial amb models neuronals de major complexitat, (ii) ampliar l'espai de transformacions i estratègies de cerca, i (iii) estudiar la sensibilitat dels pesos w_i i dels llindars de les funcions de compliment.

En conjunt, el projecte transforma un procés manual i poc consensuat en **un sistema replicable, justificable i automatitzable**, aportant una base teòrica i una eina pràctica per reduir la latència del preprocessament i millorar la consistència de les decisions en entorns de producció.

Ús de la IA generativa

Els *LLM* de *IA generativa* en aquest projecte es limiten exclusivament a dues funcions diferenciades. D'una banda, part de la fase d'investigació es recolzarà en l'ús d'aquestes tecnologies per oferir suport auxiliar en consultes sobre l'*estat de l'art*, la viabilitat de determinades eines, informació del camp de les matemàtiques, enginyeria i *ML* de caràcter general o elemental i en tasques de *troubleshooting*. A més, totes les contribucions han estat contrastades amb literatura i documentació fiable i rellevant. D'altra banda, pel que fa a la redacció del document, la *IA generativa* es farà servir únicament com a eina de suport per a la correcció ortogràfica, de cohesió, síntesi i coherència del contingut, el qual ha estat elaborat i redactat manualment per l'autor.

Agraïments

Vull expressar el meu agraïment a David Morán Pomés, responsable del projecte a GCO, pel seu suport, orientació tècnica i seguiment durant el desenvolupament d'aquest projecte.

Agraeixo també al meu tutor del TFG, Jordi Serra Ruiz, pel seu acompanyament i indicacions proporcionades al llarg del projecte.

Igualment, vull donar les gràcies a Jordi Pons Aróztegui, coordinador del grau, per la seva disponibilitat i rapidesa en resoldre consultes relacionades amb el projecte.

Finalment, agraeixo a la meva família, parella i amics el suport i la paciència mostrada durant tot el procés.

Referències

Bellman, R. (1957). *Dynamic Programming*. Princeton University Press.

Box, G. E. P., & Cox, D. R. (1964). An Analysis of Transformations. *Journal of the Royal Statistical Society: Series B (Methodological)*, *26*(2), 211 - 252.

Goodfellow, I., Bengio, Y., & Courville, A. (2016). *Deep Learning*. MIT Press.

Ioffe, S., & Szegedy, C. (2015). Batch Normalization: Accelerating Deep Network Training by Reducing Internal Covariate Shift. *Proceedings of the 32nd International Conference on Machine Learning (ICML)*, 448 - 456.

Kendall, M. G., & Stuart, A. (1946). *Advanced Theory of Statistics, Volume 1: Distribution Theory*. Charles Griffin & Company.

Larasati, A., Dwiastutik, A., Ramadhanti, D., & Mahardika, A. (2018). The effect of Kurtosis on the accuracy of artificial neural network predictive model. *MATEC Web of Conferences*, *204*, 02018. https://doi.org/10.1051/matecconf/201820402018

LeCun, Y., Bottou, L., Orr, G. B., & Müller, K.-R. (1998). Efficient BackProp. *Neural Information Processing Systems*, 9 - 50. https://dl.acm.org/citation.cfm?id=668382

Pearson, K. (1905). The Problem of the Random Distribution of Correlations. *Biometrika*, *4*(1/2), 13 - 20.

Pearson, K. (1916). Mathematical contributions to the theory of evolution. — XIX. Second supplement to a memoir on skew variation. *Philosophical Transactions of the Royal Society of London. Series A*, *216*, 429 - 457. https://doi.org/10.1098/rsta.1916.0009

Ross, S. (2014). *Introduction to Probability and Statistics*. Academic Press.

Shaham, U., Zaidman, I., & Svirsky, J. (2020). Deep Ordinal Regression using Optimal Transport Loss and Unimodal Output Probabilities. *arXiv preprint arXiv:2011.07607*. https://arxiv.org/abs/2011.07607

APÈNDIX

A.1 Recerca de Criteris

A continuació es presenten un conjunt de criteris candidats derivats de característiques essencials de la normal estàndard

- Acotament: La distribució ha d'estar limitada dins d'un rang finit per evitar valors extrems i *outliers*, i millorar l'estabilitat del model.
- Simetria: La densitat de probabilitat ha de ser aproximadament simètrica respecte al seu centre.
- Centrament: La mitjana (o mediana) ha de situar-se a prop de 0 per facilitar la convergència del model.
- Unimodalitat: La distribució ha de presentar només un pic principal.
- Continuïtat: La variable transformada ha de ser contínua, evitant discontinuïtat, concentracions discretes i distribucions degenerades.
- No constància: La distribució ha de ser variable i generar un mínim d'entropia.
- Homocedasticitat: La variància ha d'estar controlada i ser consistent a través del domini, per maximitzar la resolució del gradient.
- Escala: Les cotes i els rangs han d'estar dins d'uns valors estàndards que afavoreixin un gradient estable, com $[-1, 1]$ o $[0, 1]$.
- Equilibri entròpic: El nivell d'entropia de la distribució no pot arribar a un nivell de soroll, redundància i *sparsity*.
- Curtosi moderada: La distribució ha de contenir una concentració de valors al voltant de la mitjana; no pot ser ni excessivament concentrada en el centre ni excessivament plana.

Les propietats esmentades són només considerades candidates i seran posteriorment confirmades o descartades segons el que es pugui demostrar de forma robusta.

A.1.1 Criteris quasi-òptims segons la literatura vigent

En primera instància, per justificar si existeixen fonaments rigorosos per considerar un criteri determinat com a desitjable, s'ha dut a terme una cerca exhaustiva de la literatura existent per identificar referències explícites i contrastades que evidenciïn la necessitat d'incloure certes propietats com a criteri de disseny. Els resultats són sorprenentment ambigus i poc documentats per a certes propietats fonamentals; en canvi, per a altres, la literatura no deixa lloc a dubtes sobre la seva rellevància.

L'article *Efficient Backprop* (LeCun et al., 1998) ha establert les bases de molts aspectes del *deep learning* tal com es coneix avui dia. Es demostren determinats aspectes clau relacionats amb la distribució d'entrada del model. Textualment, pel que fa al centrament, s'indica:

- «La convergència sol ser més ràpida si la mitjana de cada variable d'entrada sobre el conjunt d'entrenament és propera a zero.»
- «En general, qualsevol desplaçament de l'entrada mitjana allunyant-se de zero esbiaixarà les actualitzacions en una direcció particular i, per tant, alentirà l'aprenentatge.»

Respecte a l'escala i la homocedasticitat (variància controlada):

- «La convergència és més ràpida no només si les entrades es desplacen com s'ha descrit anteriorment, sinó també si s'escalen de manera que totes tinguin aproximadament la mateixa covariància.»
- «L'escalat accelera l'aprenentatge perquè ajuda a equilibrar la velocitat a la qual aprenen els pesos connectats als nodes d'entrada.»

Pel que fa a la continuïtat:

- «Les entrades que són linealment dependents poden produir degeneracions que alentiran l'aprenentatge.»

Finalment, en relació amb la no correlació:

- «Les variables d'entrada haurien d'estar descorrelacionades si és possible.»

Referent a la curtosi, existeix evidència que una curtosi moderada, equivalent a la curtosi de la distribució normal estàndard (*mesocúrtica*), és desitjable en les dades d'entrada del model (Larasati et al., 2018).

La curtosi és una mesura estadística que caracteritza la concentració de valors al voltant de la mitjana d'una distribució, proporcionant informació sobre l'amplitud i forma de les cues. Les distribucions amb curtosi elevada (*leptocúrtica*) presenten pics més afilats i cues més llargues que la normal, mentre que les distribucions amb curtosi baixa (*platocúrtica*) són més planes i amb cues més curtes. Una distribució *mesocúrtica*, com la normal estàndard, representa un equilibri entre concentració i dispersió, la qual cosa optimitza la capacitat del model per aprendre de manera estable i convergeix més ràpidament.

Els resultats de l'estudi citat indiquen que el nivell de curtosi impacta de manera significativa en la precisió de la xarxa neural. Concretament, les dades *platocúrtiques* i *leptocúrtiques* presenten taxes d'error de classificació substancialment més altes que les dades mesocúrtiques, la qual cosa evidencia que un nivell moderat de curtosi contribueix a millorar tant l'eficiència com la fiabilitat de la predicció. Així, mantenir una distribució d'entrada amb curtosi propera a la normal estàndard esdevé un criteri fonamental per a la qualitat i estabilitat del model.

Per a la resta de propietats llistades no es disposa de referències explícites i directes que validin formalment que siguin desitjables. Això no implica que no ho siguin, poden ser-ho com a conseqüència d'altres característiques que sí que tenen evidència empírica o teòrica, o simplement mai s'ha formalitzat de manera explícita, ja que podrien ser considerades intuïtives o no afecten de manera crítica la convergència del model. No obstant, no existeixen referencies explícites on es deixi entendre que són desitjables en les distribucions d'entrada als models.

A.1.2 Criteris quasi-òptims indirectes

Justificació de l'unimodalitat

La unimodalitat és una propietat que intuïtivament sembla desitjable per a la geometria d'una distribució d'entrada en un model de *deep learning*. No obstant això, demostrar formalment la seva rellevància no és trivial i la literatura no proporciona referències explícites que estableixin la unimodalitat com a criteri general, més enllà de casos concrets com les dades ordinals (Shaham et al., 2020). Tot i això, s'ha pogut evidenciar de manera indirecta que és una característica desitjable a partir d'un experiment empíric relacionat amb la curtosi.

S'ha hipotetitzat que, donada la distribució normal estàndard amb curtosi $\gamma_2 = 3$ (nivell òptim), qualsevol distribució multimodal que s'assembli a la normal en termes de moments (per exemple, una distribució bimodal centrada en zero amb desviació estàndard 1 i pics situats a -1 i 1) presentaria valors de curtosi significativament diferents de 3. Això suggereix que la multimodalitat empitjora almenys un criteri quasi-òptim, com és la curtosi, i per tant pot reduir la qualitat de la distribució.

L'experiment s'ha desenvolupat de la manera següent: es genera una distribució normal estàndard i una bimodal amb els paràmetres indicats, i es calcula la seva curtosi.

Els resultats de l'experiment, mostren clarament que la distribució normal estàndard generada presenta una curtosi pròxima a 3, mentre que la distribució bimodal, tot i ser estandarditzada i semblant a la normal en termes de moments centrals, té una curtosi subòptima de 2.5. Aquesta diferència de 0.5 és significativa i evidencia que la multimodalitat compromet almenys un criteri quasi-òptim, confirmant de manera indirecta que la unimodalitat és desitjable per a una bona adequació de la distribució d'entrada.

Per tant, es pot concloure que, no només la unimodalitat es desitjable, sinó que el control d'aquesta queda contingut en la cerca d'una curtosi adequada, el qual correspon al quart moment estadístic.

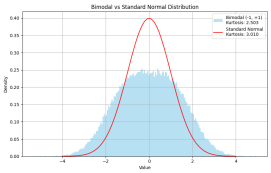

Fig. 6: Es mostren les dues distribucions (unimodal i bimodal) amb els resultats de curtosi. Es pot apreciar que els resultats de curtosi de la bimodal són subòptims.

Justificació de l'acotament

Considerem una variable aleatòria X amb mitjana μ i desviació típica σ. La curtosi està definida com:

$$\kappa = \frac{\mathbb{E}[(X-\mu)^4]}{\sigma^4}.$$

1. Necessitat d'acotament: Si X té suport no acotat, els valors extrems poden fer que la curtosi $\mathbb{E}[(X-\mu)^4]$ sigui molt gran i, per tant, subòptima. Això fa que la curtosi deixi de ser una mesura útil de concentració central. Per aquest motiu, és desitjable establir un interval d'acotament $[a, b]$ tal que $a, b \in \mathbb{R}$ i $a < b$, amb:

$$a \leq X \leq b \quad \Rightarrow \quad |X-\mu| \leq \max\left(|a-\mu|,\, |b-\mu|\right).$$

2. Efecte sobre la curtosi: A partir de l'acotament $a \leq X \leq b$, definim

$$M := \max(|a-\mu|,\, |b-\mu|).$$

Aleshores, per a qualsevol realització de X es compleix

$$|X-\mu| \leq M \quad \Rightarrow \quad (X-\mu)^4 \leq M^4.$$

Aplicant l'operador esperança a banda i banda s'obté

$$\mathbb{E}[(X-\mu)^4] \leq \mathbb{E}[M^4] = M^4,$$

ja que M és una constant.

3. Elecció del "bon"acotament: L'acotament no ha de ser ni massa estricte ni massa ample:

- Acotament massa petit: pot tallar massa dades i reduir la curtosi artificialment, deformant la representació de la distribució.
- Acotament massa gran: deixa passar valors extrems que inflen la curtosi i la fan inestable.

Per tant, cal triar un acotament raonable que preservi la massa central i controli els extrems, fent que la curtosi sigui una mesura robusta i útil per a les demostracions.

En conclusió, l'acotament adequat garanteix que la curtosi sigui finita, estable i significativa. Això permet considerar-ho com a una propietat desitjable per a una distribució.

A.2 Jerarquia de moments com a marc d'aproximació a l'optimalitat

1. Mitjana (primer moment): La mitjana estableix el centre de la distribució:

$$\mu = \mathbb{E}[X] = \int_{-\infty}^{\infty} x\, f(x)\, dx$$

Aquest primer moment emergeix directament dels criteri de centrament demostrat a la literatura. Definir la mitjana permet formalitzar matemàticament el centre de la distribució i serveix com a punt de referència necessari per a qualsevol mesura posterior de dispersió o simetria.

2. Variància (segon moment): La variància mesura la dispersió respecte a la mitjana:

$$\sigma^2 = \mathbb{E}[(X-\mu)^2] = \int_{-\infty}^{\infty} (x-\mu)^2 f(x) dx$$

Aquest moment deriva de manera natural del primer, ja que sense un centre ben definit no es pot mesurar la dispersió. La variància formalitza el criteri de escala: per assolir una escala òptima, la variància ha de ser controlada dins d'un rang moderat i coherent amb les restriccions establertes per la literatura. Per tant, la variància controlada (homocedasticitat) també queda formalitzada dintre del segon moment.

3. Asimetria (tercer moment): La *skewness* descriu la simetria de la distribució:

$$\gamma_1 = \frac{\mathbb{E}[(X-\mu)^3]}{\sigma^3} = \frac{\int_{-\infty}^{\infty} (x-\mu)^3 f(x) dx}{\sigma^3}$$

Aquest moment emergeix, en part, dels dos primers moments i la seva definició permet incorporar la simetria com a propietat a considerar dins del sistema. Condicionar la simetria respecte als moments anteriors assegura que els criteris demostrats (centrament i escala) es mantenen coherents.

A més, la simetria condiciona de manera directa el quart moment, la curtosi, que ja hem justificat com a desitjable segons la literatura. Concretament, l'asimetria imposa un límit inferior sobre la curtosi, establert per la *desigualtat de Pearson* (Pearson, 1916):

$$\gamma_2 \geq \gamma_1^2 + 1,$$

on γ_1 és la *skewness* i γ_2 la curtosi.

A partir d'aquesta relació podem definir un llindar crític de *skewness*

$$|\gamma_1|_{\text{crit}} = \sqrt{\gamma_2^{\text{òptim}} - 1},$$

per al qual la curtosi mínima ja arriba a 3 (el valor òptim de curtosi segons la literatura):

$$\gamma_2^{\text{òptim}} = 3, |\gamma_1|_{\text{crit}} = \sqrt{2} \approx 1.414$$

Això implica que, per a valors de *skewness* superiors a aquest llindar, la curtosi no pot assolir la *mesocúrtica* òptima. Per tant, la *skewness* com a tal no és una mètrica directament d'optimalitat, però existeix un llindar a partir del qual l'asimetria pot induir distribucions subòptimes respecte al criteri de curtosi.

4. Curtosi (quart moment): La curtosi mesura la concentració de valors i la forma de les cues:

$$\gamma_2 = \frac{\mathbb{E}[(X-\mu)^4]}{\sigma^4} = \frac{\int_{-\infty}^{\infty} (x-\mu)^4 f(x) dx}{\sigma^4}$$

La interpretació de la curtosi depèn dels tres moments anteriors i formalitza matemàticament el criteri de la curtosi observat a la literatura. La seva inclusió sorgeix de manera natural a partir de la jerarquia: un cop establerts centrament, escala i simetria, només llavors té sentit avaluar la concentració de valors extrems i la forma general de la distribució.

5. Cinquè moment: El cinquè moment captura la direccionalitat i l'asimetria de les cues més allunyades respecte al centre:

$$\mu_5 = \mathbb{E}[(X-\mu)^5] = \int_{-\infty}^{\infty} (x-\mu)^5 f(x) dx$$

Aquest moment depèn dels moments anteriors (mitjana, variància, *skewness* i curtosi) i permet detectar desequilibris subtils en les regions extremes de la distribució. La seva inclusió formalitza el control de valors atípics desequilibrats i permet un control de l'acotament de la distribució.

6. Sisè moment: El sisè moment mesura la concentració i intensitat dels valors extremadament allunyats de la mitjana:

$$\mu_6 = \mathbb{E}[(X-\mu)^6] = \int_{-\infty}^{\infty} (x-\mu)^6 f(x) dx$$

Aquest moment refina la informació proporcionada pel quart moment (curtosi), aportant una visió més detallada sobre la probabilitat de valors molt llunyans del centre. Permet ampliar el control de l'acotament, assegurant que la distribució mantingui les cues dins de límits raonables.

Així, la lògica subjacent és la següent: moltes de les propietats que la literatura ha demostrat ser desitjables per a la ingesta de dades en models de *DL* també fonamenten els moments com a mesura formal de la distribució.

D'aquesta manera, el marc jeràrquic de moments funciona com un model de metrització de la qualitat de la distribució: encapsula de manera estructurada i mesurable les característiques quasi-òptimes identificades, integrant tant els criteris directament demostrats com aquells derivats de manera lògica.

A.3 Formalització de la mètrica de bondat

Amb l'objectiu de convertir la mètrica de bondat $G(X)$ en una eina operativa i consistent, es formalitzen a continuació les funcions individuals $g_i(X)$ que la composen. Cada funció s'ha definit de manera que retorni un valor dins de l'interval $[0, 1]$, on 1 indica el compliment total del criteri i 0 la seva desviació màxima admissible. Els valors de referència s'han seleccionat segons l'evidència teòrica i experimental disponible.

A.3.1 Compliment per al centrament

Degut a que l'últim pas del *PTTN* és la normalització on s'estandarditzen els valors (mitjana aproximadament a 0) el centrament queda garantit intrínsecament. Per tant, el primer moment no s'ha de mesurar, ja que queda garantit.

A.3.2 Compliment per la variància

Per el mateix motiu que la propietat anterior, al estandarditzar els valors, també podem garantir que la desviació típica i, conseqüentment, la variància tindran un valor d'aproximadament 1. Per tant, el segon moment estadístic no s'inclou a la mètrica de la bondat ja que el control de la variància queda intrínsec al nostre model.

A.3.3 Funció de compliment a la skewness: $g_{\gamma_3}(X)$

Per mesurar quantitativament el criteri de simetria —és a dir, assegurar que la skewness no comprometi la curtosi òptima $\gamma_2^* = 3$— es defineix la funció de compliment que usa al tercer moment estadístic:

$$g_{\gamma_3}(X) = \begin{cases} 1, & |\gamma_1(X)| \leq \sqrt{2} \\ \exp\Big(- \dfrac{(|\gamma_1(X)| - \sqrt{2})^2}{2\sigma_\gamma^2} \Big), & |\gamma_1(X)| > \sqrt{2} \end{cases}$$

$$\sigma_\gamma > 0$$

Càlcul de la skewness empírica La skewness empírica $\gamma_1(X)$ es calcula a partir dels moments centrals de la distribució:

$$\gamma_1(X) = \frac{\frac{1}{n}\sum_{i=1}^{n}(x_i - \mu_X)^3}{\left(\frac{1}{n}\sum_{i=1}^{n}(x_i - \mu_X)^2\right)^{\frac{3}{2}}} = \frac{m_3}{m_2^{\frac{3}{2}}},$$

on m_2 i m_3 són el segon i tercer moment central, respectivament, i μ_X la mitjana empírica.

Interpretació i propietats

- La funció és 1 mentre $|\gamma_1(X)|$ estigui dins del rang acceptable ($\leq \sqrt{2}$), sense penalitzar (Pearson, 1916).
- Comença a penalitzar només quan la *skewness* supera el llindar de $\sqrt{2}$, de manera contínua i monòtona.
- σ_γ controla la rapidesa amb què decreix la funció quan la *skewness* és alta.
- Manté coherència amb la forma funcional de g_{μ_1} i g_{σ_2} i permet integrar-se fàcilment dins de la mètrica global $G(X)$.

A.3.4 Funció de compliment per a la curtosi: $g_{\gamma_4}(X)$

Per mesurar quantitativament el criteri de *mesocurtosi* —és a dir, la proximitat de la curtosi empírica $\gamma_2(X)$ al valor òptim $\gamma_2^* = 3$ (Larasati et al., 2018), que correspon a una distribució amb concentració i forma de cues equilibrada— es defineix la funció de compliment:

$$g_{\gamma_4}(X) = \exp\left(- \frac{(\gamma_2(X) - \gamma_2^*)^2}{2\sigma_\gamma^2} \right), \quad \sigma_\gamma > 0, \quad \gamma_2^* = 3$$

Càlcul de la curtosi empírica La curtosi empírica $\gamma_2(X)$ es calcula a partir dels moments centrals de la distribució:

$$\gamma_2(X) = \frac{\frac{1}{n}\sum_{i=1}^{n}(x_i - \mu_X)^4}{\left(\frac{1}{n}\sum_{i=1}^{n}(x_i - \mu_X)^2\right)^2} = \frac{m_4}{m_2^2},$$

on m_2 i m_4 són el segon i quart moment central, respectivament, i μ_X la mitjana empírica.

Interpretació i propietats

- $g_{\gamma_4}(X) = 1$ quan $\gamma_2(X) = \gamma_2^* = 3$, indicant compliment total de *mesocurtosi*.
- Penalitza tant distribucions amb curtosi massa baixa (*platicúrtica*) com massa alta (*leptocúrtica*).
- La funció és simètrica al voltant de l'òptim i sempre dins de l'interval $(0, 1]$.
- σ_γ controla la tolerància: valors petits fan que la penalització sigui més severa, valors grans fan que sigui més tolerant.
- La forma funcional és coherent amb les altres funcions de compliment, permetent la integració directa dins de la mètrica global $G(X)$.

A.3.5 Funció de compliment per al cinquè moment $g_{\mu_5}(X)$

Per mesurar quantitativament la contribució d'*asimetria en cues* —és a dir, l'efecte d'extrems que no queda reflectit completament en la *skewness* ni la curtosi— es defineix la funció de compliment basada en el cinquè moment estandarditzat (*hiperskewness*):

$$g_{\mu_5}(X) = \begin{cases} 1, & |\mu_5(X)| \leq t_5 \\ \exp\Big(- \dfrac{(|\mu_5(X)| - t_5)^2}{2\sigma_5^2} \Big), & |\mu_5(X)| > t_5 \end{cases}$$

$$t_5 \geq 0,\ \sigma_5 > 0$$

Càlcul empíric del cinquè moment El cinquè moment estandarditzat es calcula a partir dels moments centrals:

$$\mu_5(X) = \frac{\frac{1}{n}\sum_{i=1}^{n}(x_i - \mu_X)^5}{\left(\frac{1}{n}\sum_{i=1}^{n}(x_i - \mu_X)^2\right)^{5/2}} = \frac{m_5}{m_2^{5/2}},$$

on m_2 i m_5 són el segon i cinquè moment central, respectivament, i μ_X la mitjana empírica.

A.3.6 Funció de compliment per al sisè moment: $g_{\mu_6}(X)$

Per capturar la *intensitat i forma extrema de cues* més enllà de la curtosi, es fa servir el sisè moment estandarditzat, el qual pren com a referència el valor $\mu_6^* = 15$ propi d'una distribució normal.

$$g_{\mu_6}(X) = \begin{cases} 1, & |\mu_6(X) - \mu_6^*| \leq t_6, \\ \exp\left(-\dfrac{(\mu_6(X) - \mu_6^*)^2}{2\sigma_6^2}\right), & |\mu_6(X) - \mu_6^*| > t_6 \end{cases}$$

$$\mu_6^* = 15, \quad t_6 \geq 0, \quad \sigma_6 > 0.$$

Càlcul empíric del sisè moment L'estimació empírica de l'hipercurtosi es fa mitjançant:

$$\mu_6(X) = \frac{\frac{1}{n}\sum_{i=1}^{n}(x_i - \mu_X)^6}{\left(\frac{1}{n}\sum_{i=1}^{n}(x_i - \mu_X)^2\right)^3} = \frac{m_6}{m_2^3},$$

on m_2 i m_6 són el segon i sisè moment central, respectivament.

A.3.7 Ponderació de les mètriques

Un aspecte clau que a definir ha sigut els pesos w_i de la mesura de bondat $G(X)$, que s'han establert tenint en compte la naturalesa i la jerarquia dels moments centrals i superiors de la distribució. Els moments senars (3r i 5è) capturen l'asimetria, mentre que els moments parells (4t i 6è) capturen la curtosi i la intensitat de cues. Dins dels parells ((3r, 4t), (5è, 6è)), assignem el mateix pes, ja que compleixen funcions complementàries. A més, els moments centrals (3r i 4t) contenen gran part de la informació dels moments superiors, fet que justifica que s'assumeixi per conveniència una aproximació del doble d'influència en la mesura de bondat que els moments superiors (5è i 6è). Per coherència i comoditat, els pesos s'han normalitzat i aproximat:

$$w_3 = w_4 \approx 0.35, \quad w_5 = w_6 \approx 0.15,$$

d'aquesta manera mostra la prioritat dels moments centrals sobre els superiors, mantenint la suma total $\sum_{i=3}^{6} w_i = 1$.

A.4 Motius per l'omissió de la normalització en l'automatització

Inicialment, el disseny del projecte contemplava la inclusió de la normalització com a part del procés d'automatització del preprocessament. No obstant això, una anàlisi teòrica més detallada ha posat de manifest que, dins del marc proposat, aquesta operació resulta redundant i, en certs casos, conceptualment inconsistent amb la mètrica de bondat definida.

La mètrica utilitzada en aquest treball es basa en moments definits de manera invariant davant transformacions lineals d'escala. Sigui una variable aleatòria X amb mitjana μ i desviació estàndard σ, i considerem una transformació lineal de la forma

$$Y = aX + b, \quad a \neq 0.$$

Els moments centrals normalitzats d'ordre k es defineixen com

$$\gamma_k = \frac{\mathbb{E}[(X - \mu)^k]}{\sigma^k}.$$

És possible comprovar que aquests moments són invariants sota la transformació anterior.

$$\frac{\mathbb{E}[(Y - \mathbb{E}[Y])^k]}{\mathrm{Var}(Y)^{k/2}} = \frac{a^k \mathbb{E}[(X - \mu)^k]}{(|a|\sigma)^k} = \gamma_k.$$

Es pot apreciar que, per exemple, la asimetria ($k = 3$) i la curtosi ($k = 4$) no es veuen alterades per canvis d'escala ni per desplaçaments de la variable. Aquest fet implica que qualsevol operació de normalització lineal, com ara l'estandardització a mitjana zero i variància unitària, no aporta informació addicional ni modifica el valor de la mètrica de bondat utilitzada. De la mateixa forma, tampoc influeix el mètode de normalització, ja que serà indiferent fer-ho amb un mètode alternatiu com la mediana i el rang interquartílic, ja que només modifiquen al distribució a nivell d'escala i desplaçament.

Des d'aquesta perspectiva, la normalització actua com una transformació neutra respecte a l'objectiu del sistema i automatitzar-la dins del procés d'optimització resulta innecessari.

A més, incloure la normalització com a paràmetre optimitzable podria generar redundància dins de l'espai de cerca, augmentant la complexitat del procés innecessariament. Per aquest motiu, s'ha optat per excloure explícitament la normalització de l'automatització, assumint-la únicament com una operació fixa d'estandardització final.