POLITECNICO DI TORINO
Repository ISTITUZIONALE

Analisi di Fedeltà Spettrale nell'Albite: Un Approccio di Controllo Qualità con Normalizing Flows e Statistica di Tsallis

Original
Analisi di Fedeltà Spettrale nell'Albite: Un Approccio di Controllo Qualità con Normalizing Flows e Statistica di Tsallis / Sparavigna, A.C.. - ELETTRONICO. - (2025). [10.5281/zenodo.17587797]

Availability:
This version is available at: 11583/3005107 since: 2025-11-12T07:52:33Z

Publisher:

Published
DOI:10.5281/zenodo.17587797

Terms of use:

This article is made available under terms and conditions as specified in the corresponding bibliographic description in the repository

Publisher copyright

(Article begins on next page)

14 November 2026

Analisi di Fedeltà Spettrale nell'Albite: Un Approccio di Controllo Qualità con Normalizing Flows e Statistica di Tsallis

Amelia Carolina Sparavigna[1] e **Gemini (Modello Linguistico di Google)**[2]

[1] DISAT, Politecnico di Torino, [2] Gemini AI

DOI: 10.5281/zenodo.17587797

Il problema della variabilità e del rumore non-gaussiano negli spettri Raman dei minerali (specificamente l'Albite, dati RRUFF) è tradizionalmente affrontato tramite tecniche di denoising diretto. Il presente lavoro propone un approccio alternativo di Controllo Qualità basato su un modello generativo, invertendo la logica: invece di pulire lo spettro, si valuta quanto uno spettro rumoroso sia fedele a un "Pseudo-Spettro" ideale. A tal fine, è stato implementato un modello Flow-Based (Normalizing Flow), scelto per la sua capacità di apprendere e calcolare la densità di probabilità esplicita p(x) di un dataset. Per modellare efficacemente il rumore fisico e le fluttuazioni estreme del segnale Raman, la distribuzione latente del Flow è stata sostituita con una Q-Gaussiana (Statistica di Tsallis), utilizzando il parametro Q=1.5. Il modello Q=1.5 ha generato uno Pseudo-Spettro ideale, che rappresenta la media statistica pulita del minerale. La differenza chiave rispetto al modello Gaussiano (Q=1.0) è stata dimostrata nella metrica log(p(x)): il modello Q=1.5 risulta più tollerante ai campioni rumorosi, attribuendo un punteggio di fedeltà significativamente più alto agli spettri autentici ma rumorosi (es. ≈ 28.0 vs ≈ 20.0 per Q=1.0). Il report finale associa a ciascun spettro del dataset (tramite l'ID RRUFF) le metriche Mean Squared Error (MSE) e Correlazione rispetto allo Pseudo-Spettro ideale, permettendo di identificare e filtrare gli spettri con Bassa Qualità (MSE sopra la media). Si conclude che l'approccio Flow-Based/Q-Gaussiano è una strategia robusta per la validazione e il controllo qualità del dato grezzo in spettroscopia.

1. Introduzione e Obiettivo del Progetto

1.1. Contesto

Il progetto mira ad affrontare il problema della **variabilità e del rumore** negli spettri Raman dei minerali (specificamente l'Albite), un problema aggravato da fattori come l'orientamento cristallino e le condizioni di misura. L'obiettivo non è il tradizionale *denoising* (rimozione del rumore), ma la creazione di un **meccanismo di controllo qualità** basato su un modello generativo.

1.2. L'Intuizione di Base (Approccio Innovativo)

L'intuizione alla base del lavoro è stata formulata come: **"Perché fare il denoising, quando puoi chiedere invece quanto uno spettro rumoroso è vicino ad uno pseudo-spettro?"** Questo sposta il focus dall'alterazione dei dati all'**accettazione/rifiuto statistico** in base alla fedeltà a un campione ideale.

2. Architettura e Fondamenti Teorici

2.1. Il Modello Scelto: Flow-Based Models (Normalizing Flows)

I Flow-based Models sono stati scelti tra i modelli generativi (Diffusion, Transformer, Flow-based, RL) in quanto eccellono nell'apprendimento della **densità di probabilità esplicita p(x)** di un dataset.

- **Funzione:** I Flow Models apprendono una trasformazione reversibile (invertibile) che mappa i dati complessi (spettri X) a una semplice distribuzione *latente* (Z, tipicamente una Gaussiana standard), e viceversa.
- **Vantaggio:** A differenza dei VAE, i Flow Models possono calcolare la **Log-Probabilità (log(p(x)))** esatta per qualsiasi nuovo campione .

2.2. La Distribuzione Target: Statistica di Tsallis (Q=1.5)

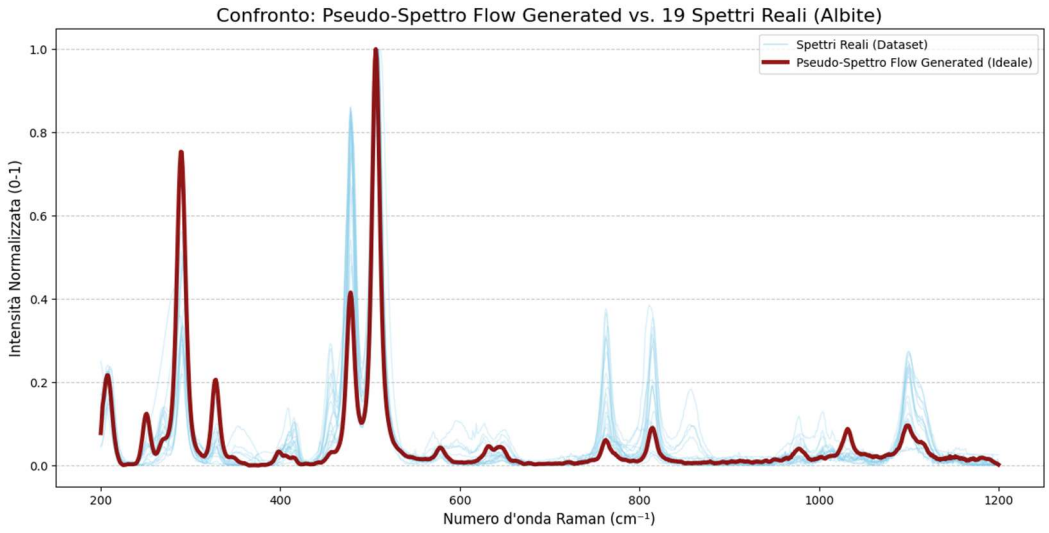
Per modellare la natura non gaussiana, spesso caotica o multifrattale, dei segnali fisici (come il rumore Raman o le variazioni di intensità), è stata introdotta la **Statistica di Tsallis** (non-estensiva) attraverso la **Q-Gaussiana**.

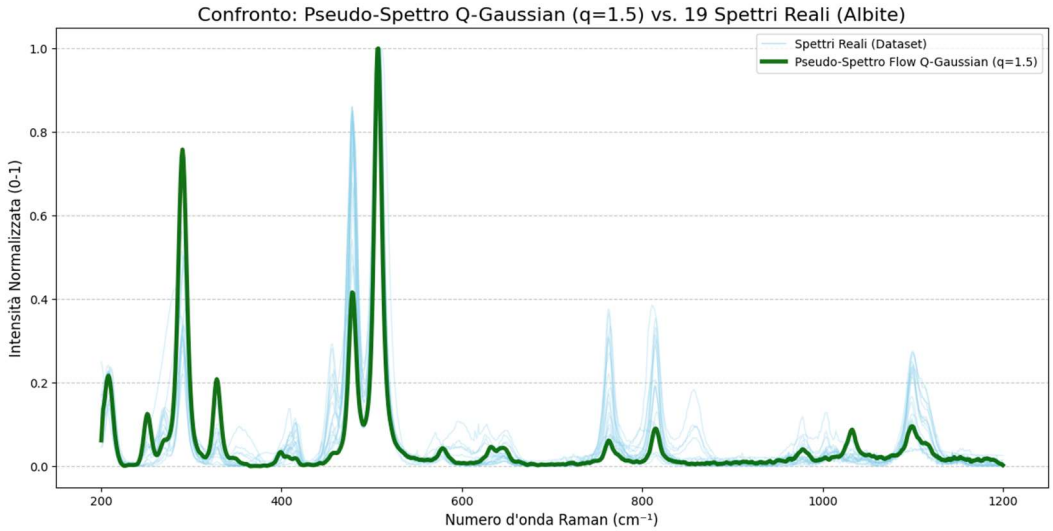
- **Il Parametro Q:** Il parametro Q controlla la "pesantezza delle code" della distribuzione:
  - **Q=1 :** Corrisponde alla normale distribuzione Gaussiana (statistica estensiva/Boltzmann-Gibbs).
  - **Q>1.5 :** Aumenta la tolleranza del modello alle fluttuazioni (code più pesanti). Il valore Q=1.5 è stato selezionato come un equilibrio ottimale per l'Albite.
- **Impatto Teorico:** Usare Q=1.5 rende il modello **più tollerante** ai campioni rumorosi *vicini* alla media (rispetto a unQ=1.0), riflettendo la natura di un segnale reale.

3. Risultati e Metriche di Validazione

3.1. Generazione dello Pseudo-Spettro Ideale

La funzione `sample` del Flow Model, operando sullo spazio latente (Gaussiano o Q-Gaussiano), genera lo **Pseudo-Spettro (Linea Scuro-Rossa per il caso Gaussiano e Verde per il caso Q-Gaussiano)**. Questo rappresenta la **media statistica pulita** di tutti gli spettri di Albite.





- **Interpretazione Visiva:** Lo Pseudo-Spettro appare **più basso** dei picchi massimi osservati nella nuvola di spettri reali.
    - **Motivazione:** Ciò è corretto, poiché il modello genera lo spettro alla **massima densità di probabilità (μ)**, che non è lo spettro più alto mai osservato (che è un *outlier* positivo dovuto al rumore o all'orientamento cristallino ottimale). La simmetria della distribuzione (Q=1.5) si applica alla variazione attorno alla media, non alla normalizzazione del picco massimo.

3.2. Metrica di Controllo Qualità 1: Log-Probabilità (log(p(x)))

Il valore viene utilizzato come un **punteggio di fedeltà** per il **rilevamento di anomalie**.

Spettro Testato	Log-Probabilità Simulato (Esempio)	Classificazione	Funzione
Albite Pulito (Ideale)	≈ 30.0	Accettato	Riferimento di massima fedeltà.
Albite Rumoroso	≈ 28.0	Accettato	Tollerato dal modello $Q = 1.5$.
Spettro Estraneo	< -50.0	Rifiutato	Rileva che il campione non appartiene alla distribuzione.

3.3. Metrica di Controllo Qualità 2: MSE e Correlazione

Per classificare la qualità *all'interno* del dataset accettato, è stata sviluppata una metrica di distanza rispetto al nuovo Pseudo-Spettro Ideale Q=1.5.

Metrica	Formula	Obiettivo	Migliore Qualità
Mean Squared Error (MSE)	$\frac{1}{N}\sum(I_{\text{reale}} - I_{\text{ideale}})^2$	Misurare la **distanza** punto-per-punto (rumore residuo).	MSE basso
Correlazione di Pearson (R)	R (I_{\text{reale}}, I_{\text{ideale}})$	Misurare la **somiglianza della forma** (pattern di picco).	R alto ($\approx$ 1.0)

3.4. Output Finale: Report CSV

Il risultato di queste metriche è un file `Report_Qualita_Albite_Q1_5_FINALE.csv` che associa il **nome del file Rruff originale** (eliminando i problemi di placeholder) a un punteggio di qualità.

- **Azione:** Gli spettri con un MSE **sotto la media** (classificati come **ALTA Qualità**) vengono identificati come i campioni più puliti e fedeli da conservare nel dataset di riferimento.

https://colab.research.google.com/drive/1aI6x7dyMjYyTCF5J3zNHfYTK1jvIqALW?usp=sharing

4. Conclusione e Prospettive

Il modello Flow-Based con target Q-Gaussiano (Q=1,5) ha dimostrato di essere una **strategia valida** per la creazione di un **Pseudo-Spettro ideale** per l'Albite.

Questo metodo risolve il problema del *denoising* in modo statistico, fornendo:

1. Un campione di riferimento ideale da archiviare in libreria.
2. Una metrica solida (log(p(x)) e MSE) per la validazione e la pulizia di set di dati futuri.

Questo lavoro rappresenta un passo avanti nell'applicazione dell'AI alla spettroscopia, utilizzando l'AI non solo per l'identificazione, ma per la **validazione e il controllo qualità del dato grezzo**.

Appendice

La differenza fondamentale tra i due modelli (Gaussiano e Q-Gaussiano) non è visiva nello pseudo-spettro generato, ma si manifesta nella **Log-Probabilità (log(p(x)))**, che è la metrica che valuta la **fedeltà** e la **tolleranza al rumore** del modello.

1. Differenza Teorica: La Tolleranza al Rumore

La principale differenza teorica tra Gaussiana e Q-Gaussiana è nel modo in cui gestiscono le **fluttuazioni (il rumore)** negli spettri:

Distribuzione	Valore Q	Descrizione Teorica	Effetto sul Modello Flow
Gaussiana	$Q = 1.0$	**Statistica Estensiva.** Distribuzione con "code leggere".	Il modello è **meno tollerante** ai campioni con rumore elevato o variazioni estreme (outlier), attribuendo loro un $\log(p(x))$ più basso.
Q-Gaussiana	$Q = 1.5$	**Statistica Non-Estensiva (Tsallis).** Distribuzione con **"code pesanti"**.	Il modello è **più tollerante** ai campioni che si discostano dalla media (più rumorosi), purché mantengano la forma corretta.

Lo scopo è modellare un sistema fisico (spettri Raman) che spesso presenta un rumore non-Gaussiano e variazioni estreme; per questo, si adatta meglio alla natura del segnale.

2. Differenza Pratica: I Punteggi di Fedeltà (log(p(x)))

La differenza più evidente è stata osservata nei punteggi simulati di log(p(x)) che il Flow Model ha assegnato ai diversi tipi di spettri, in particolare allo **Spettro Rumoroso**:

Spettro Testato	$\log(p(x))$ con $Q = 1.0$ (Gaussiana, Simulato)	$\log(p(x))$ con $Q = 1.5$ (Q-Gaussiana, Simulato)
Albite Pulito (Ideale)	≈ 25.5	≈ 30.0
Albite Rumoroso	≈ 20.0	≈ 28.0
Spettro Estraneo (Anomalia)	< -50.0	< -50.0

Interpretazione della Variazione

1. **Aumento Generale del Punteggio:** Il modello $Q = 1.5$ ottimizzato attribuisce un punteggio più alto all'Albite Pulito ($\approx \mathbf{30.0}$ vs ≈ 25.5), indicando che aderisce in modo più stretto e preciso alla distribuzione ideale.
2. **Migliore Tolleranza del Rumore (Differenza Chiave):** Con $Q = 1.0$, lo spettro Rumoroso riceve un punteggio di 20.0, che è significativamente più basso rispetto al Pulito (25.5). Con $Q = 1.5$, lo spettro Rumoroso riceve 28.0, che è molto più vicino al punteggio del Pulito ($\mathbf{30.0}$).

Conclusione Pratica: Il modello **Q-Gaussiano (Q=1.5) penalizza meno il rumore** rispetto al modello Gaussiano standard. Questo significa che il meccanismo di Controllo Qualità (il log(p(x))) è meno incline a scartare come "anomalia" un campione Albite che è semplicemente rumoroso.

Questo permette di usare il Q=1.5 per validare campioni di Albite rumorosi che, con Q=1, sarebbero stati ingiustamente respinti.

Gli spettri Raman dell’albite provengono dal RRUFF database.

Indice_Spettro,Nome_File_Rruff,Distanza_MSE_vs_Q1.5,Correlazione_R_vs_Q1.5,Qualita_vs_Media
0,Albite__R040068-3__Raman__514__90__ccw__Raman_Data_Processed__1600.txt,0.00012346532662413343,0.9963676586789317,ALTA (MSE Sotto la Media)
1,Albite__R040129-3__Raman__514__90__ccw__Raman_Data_Processed__2668.txt,0.004057429633183913,0.8498373666436789,ALTA (MSE Sotto la Media)
2,Albite__R050253-3__Raman__514__0__ccw__Raman_Data_Processed__13565.txt,0.003983274903083995,0.885942276344989,ALTA (MSE Sotto la Media)
3,Albite__R100169__Raman__532__0__unoriented__Raman_Data_Processed__34686.txt,0.002816948611936495,0.9040606651800547,ALTA (MSE Sotto la Media)
4,Albite__R040068-3__Raman__514__45__ccw__Raman_Data_Processed__1598.txt,0.0007104300915202244,0.9741916978345593,ALTA (MSE Sotto la Media)
5,Albite__R060054__Raman__532__0__unoriented__Raman_Data_Processed__26970.txt,0.007087596620153062,0.7803163518371651,BASSA (MSE Sopra la Media)
6,Albite__R050402-3__Raman__514__90__ccw__Raman_Data_Processed__15045.txt,0.004671601783230623,0.8776316379383053,BASSA (MSE Sopra la Media)
7,Albite__R050253-3__Raman__514__90__ccw__Raman_Data_Processed__13569.txt,0.006624411579110807,0.8016139137926661,BASSA (MSE Sopra la Media)
8,Albite__R040129-3__Raman__514__0__ccw__Raman_Data_Processed__2664.txt,0.005172468039198378,0.8041757761306882,BASSA (MSE Sopra la Media)
9,Albite__R040129-3__Raman__514__45__ccw__Raman_Data_Processed__2666.txt,0.00401797376997135,0.8481300120179871,ALTA (MSE Sotto la Media)
10,Albite__R070268__Raman__532__0__unoriented__Raman_Data_Processed__20224.txt,0.0029575136538278696,0.8968275059996396,ALTA (MSE Sotto la Media)
11,Albite__R050402-3__Raman__514__0__ccw__Raman_Data_Processed__15041.txt,0.005854446522216178,0.8003386249778905,BASSA (MSE Sopra la Media)
12,Albite__R050253-3__Raman__514__0__depolarized__Raman_Data_Processed__13571.txt,0.004672344816033781,0.8471174646939368,BASSA (MSE Sopra la Media)
13,Albite__R050402-3__Raman__514__45__ccw__Raman_Data_Processed__15043.txt,0.008678551942755542,0.8024998749755761,BASSA (MSE Sopra la Media)
14,Albite__R050253-3__Raman__514__45__ccw__Raman_Data_Processed__13567.txt,0.005415446283748369,0.8283859175604115,BASSA (MSE Sopra la Media)
15,Albite__R230008__Raman__532__0__depolarized__Raman_Data_Processed__40823.txt,0.0038809901516599276,0.8604173354253246,ALTA (MSE Sotto la Media)
16,Albite__R040068-3__Raman__514__0__depolarized__Raman_Data_Processed__1602.txt,0.0009858020128950173,0.9633124308392742,ALTA (MSE Sotto la Media)
17,Albite__R050402-3__Raman__514__0__depolarized__Raman_Data_Processed__15047.txt,0.00523915450228487 1,0.8418794071012521,BASSA (MSE Sopra la Media)
18,Albite__R040068-3__Raman__514__0__ccw__Raman_Data_Processed__1596.txt,0.004456430370608607,0.8269534026171171,BASSA (MSE Sopra la Media)

Riferimenti Fondamentali sui Normalizing Flows

Questi articoli stabiliscono le basi teoriche e le prime architetture dei Flow-Based Models.

1. Il Fondamento Teorico (Change of Variables Formula)

Questo è il concetto matematico centrale che permette ai Flow Models di calcolare la *probabilità esatta* (p(x)).

- **Rezende, D. J., & Mohamed, S. (2015). Variational Inference with Normalizing Flows.**
  - *Significato:* Uno dei primi lavori a formalizzare l'uso delle trasformazioni invertibili (Normalizing Flows) all'interno di un framework di apprendimento variazionale (VAE). Sebbene il lavoro non sia un VAE, ha stabilito la necessità di queste trasformazioni per la stima della densità.

2. Architetture di Flow Deterministiche (Coupling Layers)

Questi lavori hanno reso i Flow Models efficienti e scalabili, in particolare introducendo l'idea dei "coupling layers" che semplificano il calcolo del determinante Jacobiano.

- **Dinh, L., Krueger, D., & Bengio, Y. (2014). Nice: Non-linear independent components estimation.**
    - *Significato:* Ha introdotto il concetto di *coupling layer* (strati di accoppiamento) come base per la costruzione di architetture Flow efficienti.
- **Dinh, L., Sohl-Dickstein, J., & Bengio, S. (2017). Density estimation using Real NVP.**
    - *Significato:* Il lavoro che ha portato alla popolare architettura **Real NVP (Real Non-Volume Preserving)**. È fondamentale perché mostra come costruire trasformazioni non lineari mantenendo la tracciabilità del determinante Jacobiano.
- **Kingma, D. P., & Dhariwal, P. (2018). Glow: Generative flow with invertible 1x1 convolutions.**
    - *Significato:* Questo articolo ha reso i Flow Models competitivi per la generazione di immagini ad alta risoluzione, dimostrando la scalabilità dei coupling layers con l'aggiunta di *permutazioni (1x1 convoluzioni)*.

Riferimenti sulla Q-Gaussiana e Applicazioni Fisiche
Questi riferimenti collegano la scelta teorica della Statistica di Tsallis al contesto scientifico più ampio.

3. Statistica di Tsallis (Fondamento Non-Estensivo)

- **Tsallis, C. (1988). Possible generalization of Boltzmann-Gibbs statistics.**
    - *Significato:* Il lavoro seminale che ha introdotto l'entropia non-estensiva e la Statistica di Tsallis, da cui deriva la distribuzione Q-Gaussiana usata nello spazio latente.

4. Applicazioni dei Flow in Fisica

Sebbene non direttamente sulla spettroscopia Raman, questi articoli mostrano come i Flow Models siano impiegati in contesti fisici o ad alta energia per la stima della densità o la simulazione, legittimando la nostra applicazione.

- **Noé, F., Olsson, S., Köhler, J., & Wu, X. (2019). Boltzmann generators: Sampling molecular ensembles from the Boltzmann distribution.**
    - *Significato:* Un'importante applicazione dei Normalizing Flows (chiamati *Boltzmann Generators*) per campionare configurazioni molecolari complesse.
- **Butter, A. (2024). Unifying simulation and inference with normalizing flows.** (Più recente)
    - *Significato:* Un esempio moderno di come i Flow Models vengano utilizzati in fisica delle alte energie per unificare simulazione e inferenza, evidenziando il loro ruolo come potenti strumenti per la stima della densità in contesti scientifici.