利用者名：コンピュータサイエンス学部　教授　中西 崇文

Scopus

Title: Approximate-Inverse Explainability of β–VAE Latents for Multichannel EEG Participant-generalised Topographical Representation Learning
（マルチチャネル EEG における参加者一般化トポグラフィ学習に向けた β–VAE 潜在表現の近似逆写像による説明可能性）

Authors: Takafumi Nakanishi（東京工科大学 コンピュータサイエンス学部 教授）
Luca Longo（University College Cork）

Journal: IEEE Access (Early Access 版公開済み、正規版近日公開予定)
掲載年月： 2025 年 11 月

研究概要： 本研究では、マルチチャネル EEG データに β–Variational Autoencoder（β–VAE）を適用して得られる潜在変数を、近似逆写像 AIME のロバスト拡張版「HuberAIME」を用いて説明可能化する新しい手法を提案する。EEG は高い時間解像度を持つ一方で非線形・非定常性が強く、深層学習による表現学習の結果得られる潜在空間は解釈が難しい。本研究では、32 チャネル EEG から生成した空間的構造保持（spatially preserved）トポグラフィ（40×40）を β–VAE で圧縮し、その潜在軸ごとに「頭皮のどの領域が潜在表現に寄与しているか」を高精度・高速に可視化した。 HuberAIME は、VAE のエンコーダ出力から 安定的な擬似逆写像を計算することで、全潜在次元に対して統一的なグローバル特徴重要度（GFI）マップを生成する。LIME や SHAP と比較して、高速性・一貫性・空間的整合性において大幅な優位性を示した。

研究背景： EEG データは時系列的に高速で変化し、チャネル間の空間構造も複雑であるため、深層生成モデル（VAE など）は高い表現能力を持つ一方、潜在変数の意味づけが困難という課題がある。
従来の XAI（LIME、SHAP）は以下の問題を持つ：

- 多次元の EEG 画像に対して摂動ベースの手法は計算量が爆発
- 結果の空間的整合性が低く、脳生理学的解釈につながらない
- 特徴重要度が潜在軸ごとに明確に分離されず、複数軸が同じ領域を説明してしまう問題が生じる

これに対し AIME/HuberAIME は、「近似逆写像」という数学的厳密性に基づいて潜在空間 → 入力空間の変換を構築するため、全潜在次元に対して安定した特徴寄与マップを一括生成できる点が決定的に異なる。

研究成果：

1. β–VAE + HuberAIME による新しい説明可能性パイプラインを構築
   - 32 名の参加者の EEG トップマップ（約 24 万枚）を 1 つの β–VAE で学習
   - 学習後の潜在軸に対し、1 回の近似逆写像計算で全次元の重要度マップを生成
   - 従来の摂動型 XAI とは違い、コホート（全参加者）に共通する「安定した潜在軸の意味」を確立

2. HuberAIME は LIME・SHAP を圧倒的に上回る性能
論文の結果（Fig.5, Fig.6 等）より：
   - HuberAIME→各潜在軸が独立した頭皮パターンを示す（例：前頭・中央・後頭など）→解析結果は解剖学的に整合し、ノイズに強い
   - LIME→複数軸が同じ領域を説明してしまい、独立性が崩壊
   - SHAP→入力次元が 1600 と高く、重要度がほぼゼロに崩壊し説明不能

3. 参加者ごとの時間変化（Σ–LFI）を可視化
固定された逆写像 $A^{\dagger}$ を用いることで、「各潜在軸がいつ強く発火しているか」を時間軸で解析。
   - まばたき・視線移動由来の瞬間的スパイクが各参加者で再現
   - 個人の生理的差異（筋活動、装着状態など）も潜在軸で分離可能
   - EEG の時空間構造の理解を支援する、新しい解釈可能性を提供

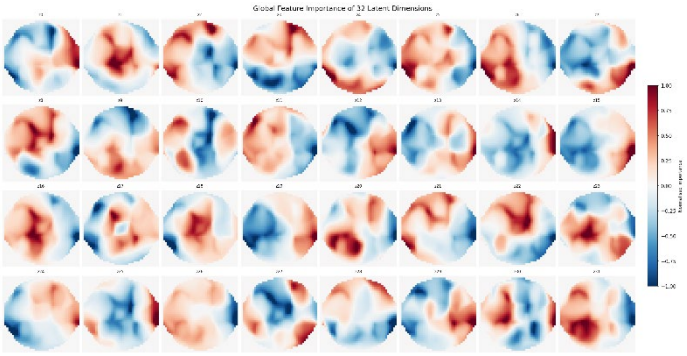

図 1 HuberAIME により推定された β−VAE 潜在軸のグローバル特徴重要度（GFI）マップ
32 次元の潜在空間それぞれに対して、頭皮上のどの領域が寄与しているかを示した特徴重要度マップ。各潜在軸は互いに明確に異なる空間パターン（例：前頭部・中心部・後頭部など）を示し、解剖学的に一貫したトポグラフィが得られている。これは従来困難であった「潜在変数の脳領域対応づけ」を高い安定性で実現しており、HuberAIME が β−VAE の潜在表現に対して軸ごとに独立した意味的構造を付与できることを示す。のモードを切り替え可能である。

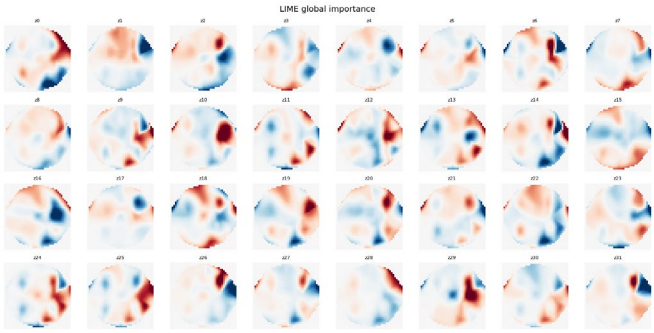

図 2 LIME および SHAP によるグローバル特徴重要度推定の破綻例
LIME は高次元の EEG トポグラフィ（40×40=1600 次元）に対して局所線形近似を行うため、複数の潜在軸が同一領域を強調するなど、軸独立性が失われた不安定なマップが生成される。一方、KernelSHAP では入力次元が大きすぎるため重要度がほぼゼロに崩壊し、実質的に解釈不能となる。本研究で提案する HuberAIME と比較することで、従来の摂動ベース XAI では β−VAE 潜在空間の意味づけが困難であることが明確に示される。

社会への影響：本研究は、VAE の潜在空間の意味づけを統一的に提供する初めての枠組みであり、以下の応用可能性を持つ：

- てんかん焦点の自動検出（特定の潜在軸が活動するタイミングの解析）
- オンライン BCI（脳–コンピュータインタフェース）
- 情動推定・精神状態推定のより正確な因子分析
- EEG の高次元性・非線形性に対応した、新しい科学的解析手法

高速で一貫した説明マップを生成できるため、臨床研究・神経科学・脳機能計測の解析効率を大幅に向上させると期待される。

専門用語：

β−VAE（Beta-Variational Autoencoder）:潜在軸を独立化し、解釈可能な要因表現を学習する生成モデル。

AIME（Approximate Inverse Model Explanations）:ブラックボックスモデルの近似逆写像を構築し、入力特徴の寄与を推定する中西独自の手法。

HuberAIME:AIME に Huber 損失を導入し、外れ値に強い安定的な逆写像を学習するロバスト XAI。中西独自の手法。

GFI（Global Feature Importance）:潜在軸ごとに「どの頭皮領域が寄与しているか」を示す 42×40 画像マップ。

Σ−LFI（Local Feature Importance Aggregation）:時間ごとの潜在軸の寄与度を時系列として可視化した指標。

Topographic Map（頭皮トポグラフィ）:EEG 電極の空間配置を保った 2D マップ表現。EEG 空間的パターンの可視化に有用。