Efficient Canonical Correlation Analysis with Sparsity

Elena Tuzhilina

Department of Statistical Sciences, University of Toronto

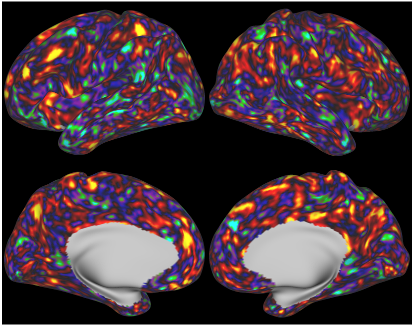
elena.tuzhilina@utoronto.ca

18 March 2026

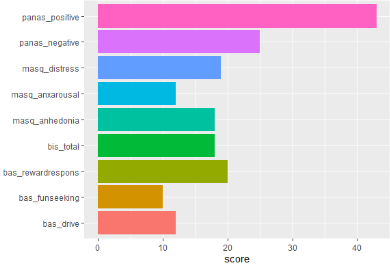
McGill Biostatistics Seminar

Motivating Example: Brain-Behavior Associations

Goal: Discover relationships between brain activity and behavioral performance during a cognitive task.



$X \in \mathbb{R}^{n \times p}$: brain activations during a gambling task



$Y \in \mathbb{R}^{n \times q}$: behavioral performance, mood, and reward-related scores

$n = 153$ participants; $p = 229$ brain regions; $q = 9$ behavioral scores
High-dimensional setting: $n \ll p$

Outline

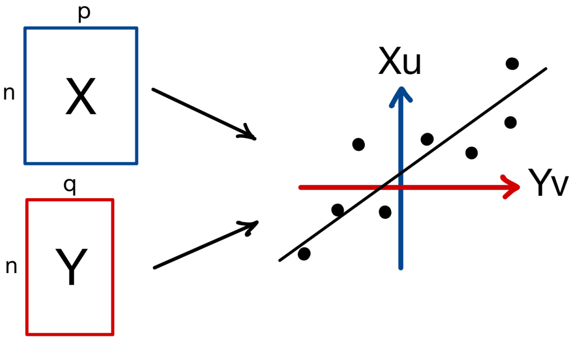
1. **Background**
    - Classical CCA and its limitations in high dimensions
    - Review of sparse CCA methods
2. **ECCAR Method**
    - CCA as reduced-rank regression
    - Algorithmic framework
3. **Experiments**
    - Simulation studies
    - Applications in genetics & neuroscience
    - Application to LLM interpretability
4. **Theoretical guarantees**
5. **Discussion**

Background

Canonical Correlation Analysis (CCA)

Goal: Given two datasets $X \in \mathbb{R}^{n \times p}$ and $Y \in \mathbb{R}^{n \times q}$ find directions $u \in \mathbb{R}^p$, $v \in \mathbb{R}^q$ that maximize the correlation

$$\max_{u,v} \quad \mathrm{cor}(Xu, Yv)$$



Canonical Correlation Analysis (CCA)

Goal: Given two datasets $X \in \mathbb{R}^{n \times p}$ and $Y \in \mathbb{R}^{n \times q}$ find directions $u \in \mathbb{R}^p, v \in \mathbb{R}^q$ that maximize the correlation

$$\max_{u,v} \quad \mathrm{cor}(Xu, Yv) = \frac{u^\top \Sigma_{XY} v}{\sqrt{u^\top \Sigma_X u}\sqrt{v^\top \Sigma_Y v}}$$

Canonical Correlation Analysis (CCA)

Goal: Given two datasets $X \in \mathbb{R}^{n \times p}$ and $Y \in \mathbb{R}^{n \times q}$ find directions $u \in \mathbb{R}^p, v \in \mathbb{R}^q$ that maximize the correlation

$$\max_{u,v} \quad \mathrm{cor}(Xu, Yv) = \frac{u^\top \Sigma_{XY} v}{\sqrt{u^\top \Sigma_X u}\sqrt{v^\top \Sigma_Y v}}$$

Constrained form: for the 1st pair of canincal direcitons (u_1, v_1)

$$\begin{aligned}(u_1, v_1) = \operatorname*{argmax}_{u,v} \quad & u^\top \Sigma_{XY} v \\ \text{s.t.} \quad & u^\top \Sigma_X u = 1, \quad v^\top \Sigma_Y v = 1\end{aligned}$$

After whitening:
Let $u_0 = \Sigma_X^{1/2} u_1, \quad v_0 = \Sigma_Y^{1/2} v_1$

$$\begin{aligned}(u_0, v_0) = \operatorname*{argmax}_{u,v} \quad & u^\top \left(\Sigma_X^{-1/2} \Sigma_{XY} \Sigma_Y^{-1/2}\right) v \\ \text{s.t.} \quad & \|u\| = \|v\| = 1\end{aligned}$$

Canonical Correlation Analysis (CCA)

Constrained form: for the 1^{st} pair of canincal direcitons (u_1, v_1)

$$(u_1, v_1) = \operatorname*{argmax}_{u,v} \quad u^\top \Sigma_{XY} v$$
$$\text{s.t.} \quad u^\top \Sigma_X u = 1, \quad v^\top \Sigma_Y v = 1$$

After whitening:

Let $u_0 = \Sigma_X^{1/2} u_1, \quad v_0 = \Sigma_Y^{1/2} v_1$

$$(u_0, v_0) = \operatorname*{argmax}_{u,v} \quad u^\top \left(\Sigma_X^{-1/2} \Sigma_{XY} \Sigma_Y^{-1/2}\right) v$$
$$\text{s.t.} \quad \|u\| = \|v\| = 1$$

Computation via SVD:

- Let u_0 and v_0 be top singular vectors of $\Sigma_X^{-1/2} \Sigma_{XY} \Sigma_Y^{-1/2}$
- Map back $u_1 = \Sigma_X^{-1/2} u_0, \quad v_1 = \Sigma_Y^{-1/2} v_0$

CCA: Multiple Canonical Directions

General formulation: for the i^{th} pair of canincal directions (u_i, v_i)

$$(u_i, v_i) = \underset{u,v}{\operatorname{argmax}} \quad u^\top \Sigma_{XY} v$$
$$\text{s.t.} \quad u^\top \Sigma_X u = 1, \quad v^\top \Sigma_Y v = 1$$
$$u^\top \Sigma_X u_j = 0, \quad v^\top \Sigma_Y v_j = 0 \quad \forall j < i$$

Computation via SVD: to find all directions at once

$$U = [u_1 | \cdots | u_r], \quad V = [v_1 | \cdots | v_r], \quad r = \min(p, q)$$

- Whiten cross-covariance and take SVD

$$\Sigma_X^{-1/2} \Sigma_{XY} \Sigma_Y^{-1/2} = U_0 \Lambda V_0^\top$$

- Recover canonical directions

$$U = \Sigma_X^{-1/2} U_0, \quad V = \Sigma_Y^{-1/2} V_0$$

CCA Terminology

- Number of components: $r \le \min(p, q)$
- Canonical directions: (u_i, v_i) for $i = 1, \dots, r$
- Direction matrices:

$$U = [u_1 \mid \cdots \mid u_r] \in \mathbb{R}^{p \times r}, \quad V = [v_1 \mid \cdots \mid v_r] \in \mathbb{R}^{q \times r}$$

- Canonical variates: (Xu_i, Yv_i) for $i = 1, \dots, r$
- Canonical correlations:

$$\lambda_i = u_i^\top \Sigma_{XY} v_i = \mathrm{cor}(Xu_i, Yv_i), \quad i = 1, \dots, r$$

can be stored as a diagonal matrix

$$\Lambda = \mathrm{diag}(\lambda_1, \dots, \lambda_r)$$

CCA in High Dimensions

High-dimensional regime: $\min(p, q) \geq n$

- *Unstable* and *non-reproducible* canonical directions
- *Overfitting:* correlations on training data fail to generalize
- *Empirical evidence:* poor stability in practice (e.g., neuroscience applications)

CCA can produce *spurious and misleading associations* when $p, q \gg n$

Idea: Impose sparsity on canonical directions u_i and v_i to select only relevant variables.

Sparse CCA Methods: Heuristic-Based

Idea: Impose sparsity on canonical directions using ℓ_1-penalties; get subsequent directions via sequential deflation.

[Witten et al., 2009] Assumes $\Sigma_X = I$ and $\Sigma_Y = I$

$$\max_{u,v} \; u^\top \Sigma_{XY} v \quad \text{s.t.} \quad \|u\|_2 \leq 1, \; \|v\|_2 \leq 1, \; \|u\|_1 \leq t_u, \; \|v\|_1 \leq t_v$$

[Wilms et al., 2015] Solves ALS problems

$$\min_{u,v} \|Xu - Yv\|^2 + \rho_u \|u\|_1 + \rho_v \|v\|_1$$

Pros: Fast, scalable
Cons: Sensitive to initialization and diagonal covariance assumption

Sparse CCA Methods: Theory-Based

Idea: Respect full covariance structure and enforce sparsity by limiting the number of non-zero rows in U and V.

[Gao et al., 2015/17; Gao & Ma 2023]

- Bound the number of non-zero rows:

$$\Sigma_{XY} = \Sigma_X U \Lambda V^\top \Sigma_Y \quad \text{with} \quad U^\top \Sigma_X U = I, \;\; V^\top \Sigma_Y V = I$$
$$\text{supp}(U) \leq s_u, \;\; \text{supp}(V) \leq s_v$$

- Estimate rank-r subspace via Fantope relaxation:

$$\max_{F \in \mathcal{F}} \langle \Sigma, F \rangle - \rho \|F\|_{11} \quad \text{where } \mathcal{F} = \{F : 0 \preceq F \preceq I, \; \text{tr}(F) = r\}$$

Here $F = \begin{pmatrix} U \\ V \end{pmatrix} \begin{pmatrix} U^\top & V^\top \end{pmatrix}$ is a large $(p+q) \times (p+q)$ matrix.

Pros: Consistent, accurate
Cons: Computationally intensive; scales poorly for large datasets

Efficient Sparse CCA via Reduced Rank Regression (ECCAR)

- *Efficient Canonical Correlation Analysis with Sparsity* Z. Wu, C. Rousseau, E. Tuzhilina, C. Donnat (2025)
- *Canonical Correlation Analysis as Reduced Rank Regression in High Dimensions* C. Donnat, E. Tuzhilina (2024)
- package `ccar3` is available on CRAN

CCA via Reduced-Rank Regression

Idea: Recast CCA as a reduced-rank regression problem.

Optimization problem: Let $B = U\Lambda V^\top \in \mathbb{R}^{p\times q}$ then the CCA problem is equivalent to

$$\min_B \left\| \frac{1}{n} XBY^\top - I_n \right\|_F^2 \quad \text{s.t.} \quad \text{rank}(B) = r$$

ECCAR Algorithm

Input: $X \in \mathbb{R}^{n \times p}$, $Y \in \mathbb{R}^{n \times q}$, r

1. **Regression:** $B = \text{argmin}_B \left\| \frac{1}{n} XBY^\top - I_n \right\|_F^2$
2. **Rank-r SVD:** $B \approx U_B \Lambda_B V_B^\top$
3. **Normalize:** $U = U_B (U_B^\top \Sigma_X U_B)^{-1/2}$, $V = V_B (V_B^\top \Sigma_Y V_B)^{-1/2}$
4. **Correlations:** $\Lambda = \mathrm{diag}(U^\top \Sigma_{XY} V)$

Output: (U, V, Λ)

- Solution via *regression + SVD*
- Learns a *low-rank coupling* between X and Y
- Avoids explicit inversion of Σ_X and Σ_Y
- Extendable to *high-dimensional settings*

ECCAR in High Dimensions

Input: $X \in \mathbb{R}^{n\times p}$, $Y \in \mathbb{R}^{n\times q}$, number of components r, penalty factor ρ

1. **Sparse regression:** $B = \text{argmin}_B \left\|\frac{1}{n}XBY^\top - I_n\right\|_F^2 + \rho \; pen(B)$
2. **Rank-r SVD:** $B \approx U_B \Lambda_B V_B^\top$
3. **Normalize:** $U = U_B(U_B^\top \Sigma_X U_B)^{-1/2}, \quad V = V_B(V_B^\top \Sigma_Y V_B)^{-1/2}$
4. **Correlations:** $\Lambda = \mathrm{diag}(U^\top \Sigma_{XY} V)$

Output: (U, V, Λ)

Computational complexity is $O\big(p^2 n + q^2 n + T((p \wedge q)n^2 + pqn)\big)$.

Flexible penalty

Sparsity in U and V can be controlled via flexible penalty on $B = U\Lambda V^\top$.

- Sparse

$$pen(B) = \|B\|_{11} = \sum_{i,j} |B_{ij}|$$

- Row-sparse

$$pen(B) = \|B\|_{21} = \sum_{i} \sqrt{\sum_{j} B_{ij}^2}$$

- Group-sparse

$$pen(B) = \sum_{G \in \mathcal{G}} \sqrt{N_G} \left\| B_G \right\|_F = \sum_{G \in \mathcal{G}} \sqrt{N_G} \sqrt{\sum_{(i,j) \in G} B_{ij}^2}$$

Parameter Tuning

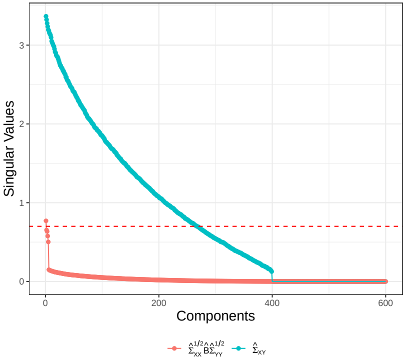
Penalty parameter ρ

- Cross-validation (or BIC)
- $\rho \asymp \sqrt{\frac{\log(p+q)}{n}}$

Number of components r

- Cross-validation
- Scree plot of $\Sigma_X^{1/2} B \Sigma_Y^{1/2}$

Clear eigengap $\Rightarrow$ reliable rank recovery



Singular values of $\widehat{\Sigma}_X^{1/2} \widehat{B} \widehat{\Sigma}_Y^{1/2}$ vs. $\widehat{\Sigma}_{XY}$ for $n = 400,\ p = q = 600,\ r = 5$.

Red dashed lines show true values.

Experiments

Simulation Study

Setup:

- Sparse canonical directions $U \in \mathbb{R}^{p \times r},\ V \in \mathbb{R}^{q \times r}$ with s non-zero rows.
- Covariances Σ_X, Σ_Y with $n_{\text{pca}} = 5$ sparse principal components.
- Signal strength $\Lambda = \lambda I$

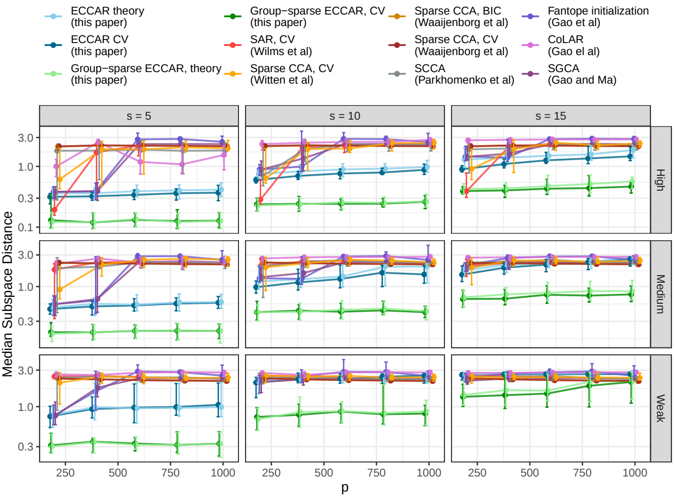
$$\lambda = 0.9 \text{ (high)}, \quad \lambda = 0.7 \text{ (medium)}, \quad \lambda = 0.5 \text{ (low)}$$

- Cross-covariance $\Sigma_{XY} = \Sigma_X U \Lambda V^\top \Sigma_Y$,

Data generation:

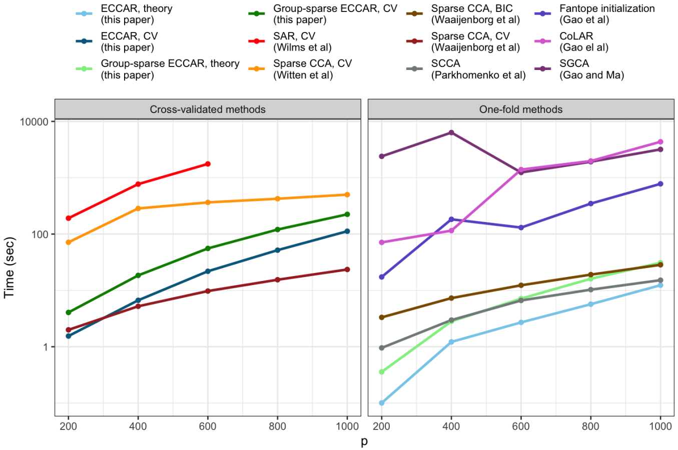
$$(X_i, Y_i)_{i=1}^n \sim \mathcal{N}_{p+q}\left(0, \begin{pmatrix} \Sigma_X & \Sigma_{XY} \\ \Sigma_{XY}^\top & \Sigma_Y \end{pmatrix}\right)$$

Evaluation: subspace distance (principal angles) between true and estimated canonical variates.



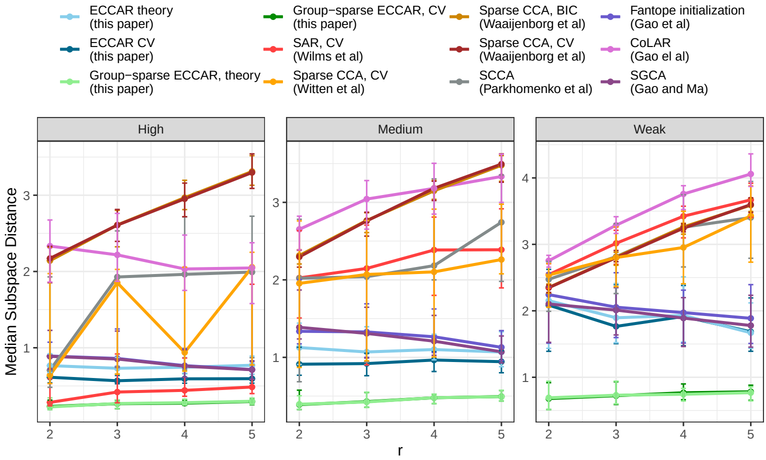
Median subspace distance averaged over 25 independent experiments for $n = 400$, $p = q$ and $r = 2$.

Simulation Study



Mean time averaged over 25 independent experiments for $n = 400$, $p = q$, strong signal strength and support $s = 5$.

Simulation Study



Median subspace distance averaged over 25 independent experiments for $n = 400$, $p = q = 200$.

Alcohol Use Disorder (AUD) Data

Goal: Jointly analyze gene expression and DNA methylation to identify AUD-associated patterns

Data

- $n = 46$ samples (23 AUD and 23 controls)
- X: gene expression ($p = 300$ genes)
- Y: DNA methylation ($q = 500$ CpG sites)

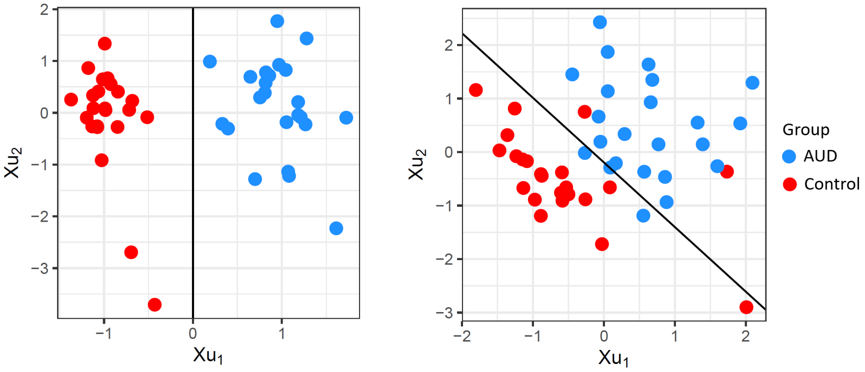
Model & Tuning:

- Number of components $r = 2$
- Regularization parameter selected via 8-fold CV (or BIC, or permutation)

Evaluation:

- MSE $\|XU - YV\|$
- Canonical correlations
- Classification accuracy of AUD labels (SVM on XU)

Alcohol Use Disorder (AUD) Data

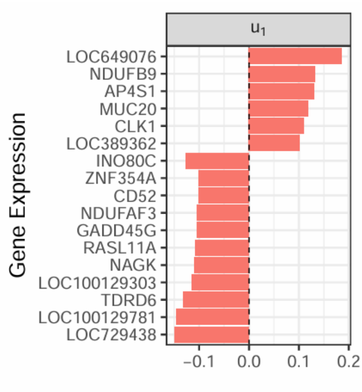


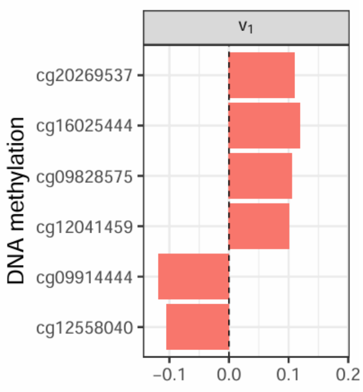
ECCAR (left) achieves clear separation between AUD and controls, while sparse CCA (right) by Witten et al. shows noticeable overlap

Alcohol Use Disorder (AUD) Data

Method	Test MSE	Test Correlation	Accuracy (SVM)
ECCAR, CV (this paper)	**0.604**	0.400	**0.958**
SAR, CV (Wilms and Croux [49])	0.857	0.274	0.869
SAR, BIC (Wilms and Croux [49])	0.905	0.311	0.879
Sparse CCA, CV (Witten et al. [51])	1.070	0.258	0.823
Sparse CCA, permuted (Witten et al. [51])	0.731	**0.483**	0.933
Sparse CCA, CV (Waaijenborg and Zwinderman [48])	1.825	-0.351	0.955
Sparse CCA, BIC (Waaijenborg and Zwinderman [48])	1.865	-0.383	0.955
Fantope Initialization (Gao et al. [13])	40.081	0.106	0.570
SGCA (Gao and Ma [14])	1.617	0.193	0.736

Alcohol Use Disorder (AUD) Data





- AUD-associated CpG sites: cg20269537, cg11562309
- Differentially expressed genes: ZNF354A, RASL11A
- Alcohol-related pathways: NDUFAF3, NDUFB9
- Linked to alcohol-related disease: GADD45G

Autism Brain Imaging Data Exchange (ABIDE)

Goal: Link brain connectivity with behavior in Autism Spectrum Disorder (ASD)

Data:

- $n = 106$ subjects (65 ASD, 41 controls)
- X: functional connectivity between 110 brain regions (Harvard-Oxford atlas, $p = 5,995$)
- Y: Vineland Adaptive Behavior scores ($q = 14$)

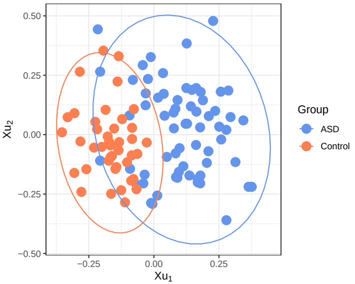
Grouping:

- 110 brain regions are grouped into 8 functional networks
- 5,995 features in X are grouped into 36 network interaction sets

Autism Brain Imaging Data Exchange (ABIDE)

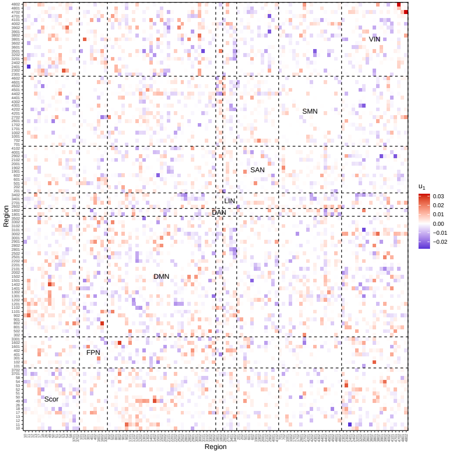
Method	Test MSE
ECCAR, CV (this paper)	2.19 (1.84; 2.63)
Group ECCAR, CV (this paper)	1.96 (1.63; 2.19)
Row-sparse ECCAR, CV (this paper)	**1.76** (1.39; 2.16)
SAR, CV (Wilms and Croux [50])	3.17 (2.16; 3.43)
SAR, BIC (Wilms and Croux [50])	3.75 (3.49; 3.77)
Sparse CCA, CV (Witten et al. [51])	2.48 (2.23; 2.92)
Sparse CCA, permuted (Witten et al. [51])	2.81 (2.22; 2.97)
SCCA, CV (Parkhomenko et al. [37])	2.33 (1.81; 2.62)

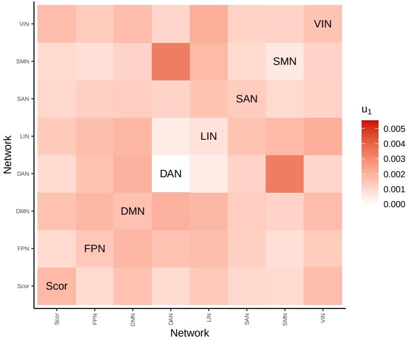
Performance on 10-fold cross-validation



First two canonical variates (row-sparse ECCAR)

Autism Brain Imaging Data Exchange (ABIDE)

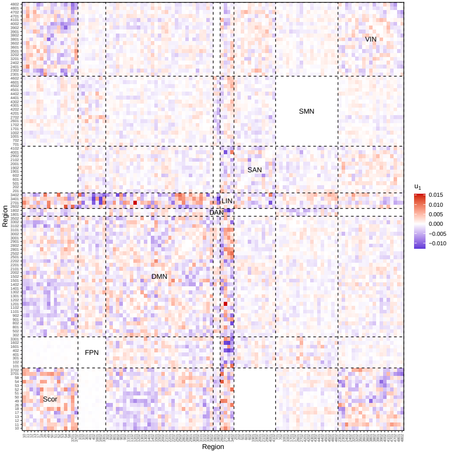


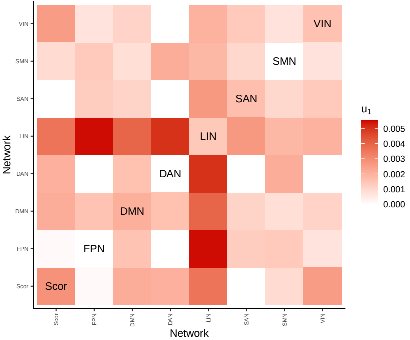


Direction u_1 for row-sparse ECCAR

Strong interaction between the dorsal attention network (DAN) and the somatomotor network (SMN)

Autism Brain Imaging Data Exchange (ABIDE)

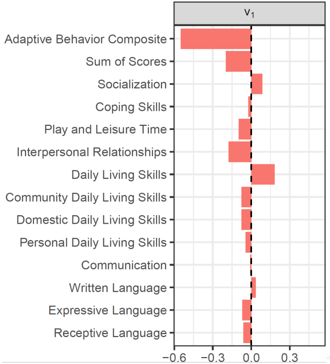


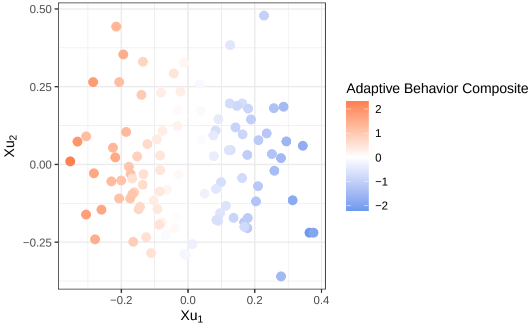


Direction u_1 for group-sparse ECCAR

Limbic network (LIN) reveals strong connections with dorsal attention network (DAN) and the frontoparietal network (FPN)

Autism Brain Imaging Data Exchange (ABIDE)



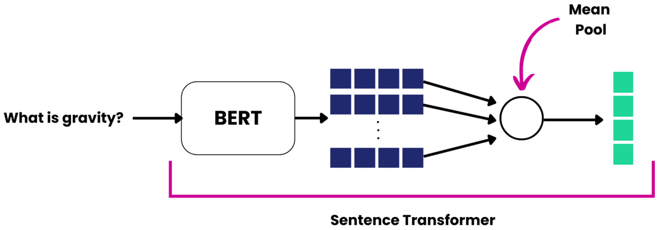


Adaptive Behavior Composite (ABC) exhibits the primary contribution to canonical correlation.

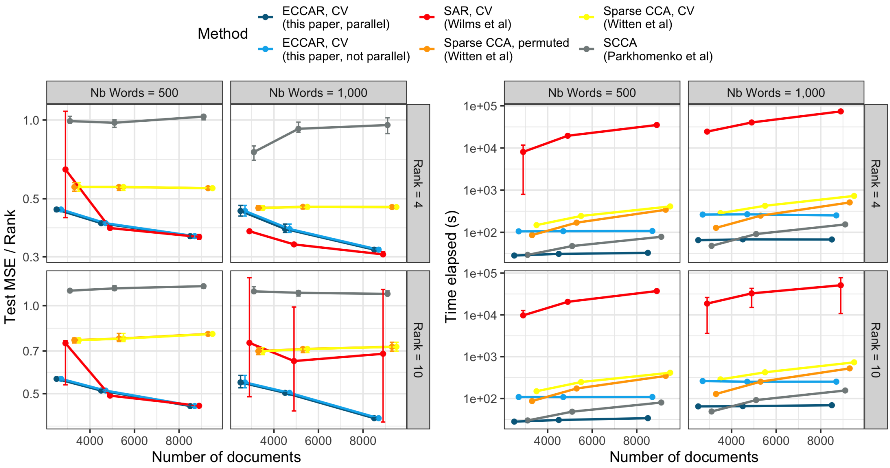
Goal: Relate high-dimensional, opaque LLM embeddings to interpretable word-level features

Data:

- n texts from 20 `Newsgroups` (20 topics, coarsened into 7 categories)
- X: `all-mpnet-base-v2` embeddings, $p = 750$
- Y: TF-IDF vectors of top q words by frequency

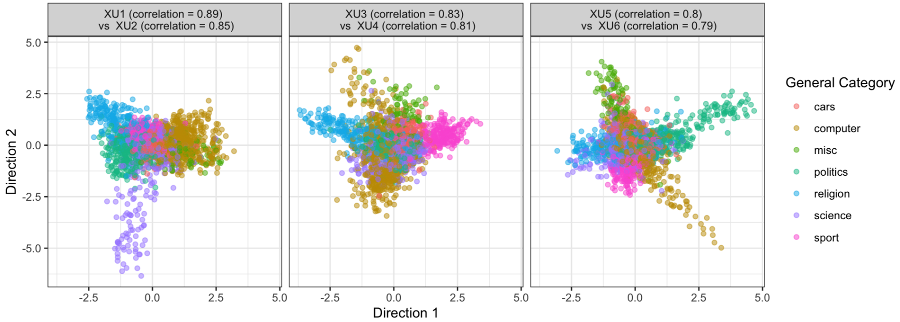


LLM interpretability



MSE (left) and compute time (right) as a function of n for sparse CCA models with $r = 4, 10$ and $q = 500, 1000$.

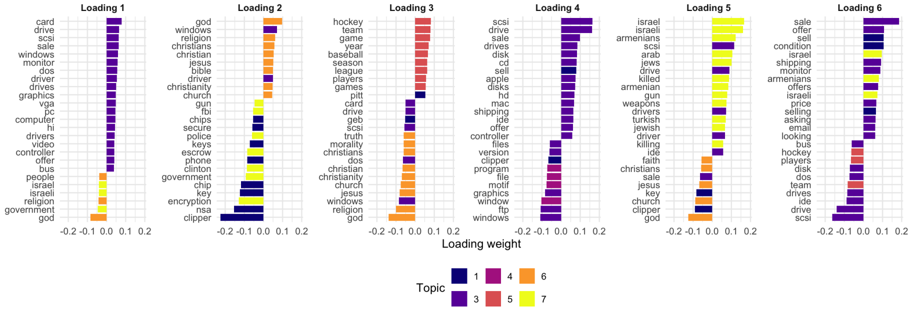
LLM interpretability



Canonical variates $(Xu_i)_{i=1}^{6}$ obtained by ECCAR with $n = 3000$ and $q = 1000$.

Some direction separate text categories (e.g. direction 1 separates computer, direction 3 separates sport and religion)

LLM interpretability



Top 25 loadings for $(v_i)_{i=1}^{6}$ obtained by ECCAR ($n = 3000$, $q = 1000$). Colors show each word's most frequent LDA topic.

The loadings are dominated by few main categories (e.g. sport and religion in loading 3, and computers in loading 4)

Theoretical guarantees

Canonical pair model

The probability space is

$$\left\{(X_i, Y_i) \overset{\text{i.i.d.}}{\sim} \mathcal{N}_{p+q}(0, \Sigma) \mid \Sigma \in \mathcal{F}(s_u, s_v, p, q, r; M)\right\}.$$

- Covariance $\Sigma = \begin{pmatrix} \Sigma_X & \Sigma_{XY} \\ \Sigma_{XY}^\top & \Sigma_Y \end{pmatrix}$ admits decomposition

$$\Sigma_{XY} = \Sigma_X U \Lambda V^\top \Sigma_Y$$

  where $U = [u_1 \mid \ldots \mid u_r] \in \mathbb{R}^{p \times r}$, $V = [v_1 \mid \ldots \mid v_r] \in \mathbb{R}^{q \times r}$, and $\Lambda = \mathrm{diag}(\lambda_1, \ldots, \lambda_r)$, with $\lambda_1 \geq \cdots \geq \lambda_r$.
- Leading canonical directions are row-wise sparse:

$$|\mathrm{supp}(U)| \leq s_u, \quad |\mathrm{supp}(V)| \leq s_v.$$

- Covariances are well-conditioned:

$$\sigma_{\min}(\Sigma_X) \wedge \sigma_{\min}(\Sigma_Y) \geq \frac{1}{M}, \quad \sigma_{\max}(\Sigma_X) \vee \sigma_{\max}(\Sigma_Y) \leq M.$$

Theoretical guarantees

Theorem

Consider the parameter space $\mathcal{F}(s_u, s_v, p, q, r; M)$ *for the covariance matrix* Σ*, and let* $\Delta = \widehat{B} - B$*, where* $\widehat{B}$ *is the estimate obtained in step 1 of the sparse ECCAR algorithm and* B *is the underlying population quantity* $B = U\Lambda V^\top$*. Assume* $n \geq c s_u s_v \log(p+q)$ *for some sufficiently large constant* c*. There exist constants* a, b, C *depending on* M *and* c *such that if* $\rho \geq a\sqrt{\log(p+q)/n}$*, then with probability at least* $1 - \exp(-b s_u \log(ep/s_u)) - \exp(-b s_v \log(eq/s_v)) - (p+q)^{-b}$*, we have:*

$$\|\Delta\|_F \leq C\rho\sqrt{s_u s_v}.$$

In particular, if ρ *is of order of* $\sqrt{\log(p+q)/n}$*, we have:*

$$\|\Delta\|_F \lesssim \sqrt{\frac{s_u s_v \log(p+q)}{n}},$$

and $\widehat{B}$ *has sparse entries:* $\|\widehat{B}\|_0 \lesssim s_u s_v$*.*

Theoretical guarantees

Theorem

Consider the parameter space $\mathcal{F}(s_u, s_v, p, q, r; M)$ for the covariance matrix Σ. Assume $n \geq c s_u s_v \log(p+q)/\lambda_r^2$ for some sufficiently large constant c. There exist constants a_1, a_2, b, C depending on M and c such that if $\rho \in \left[a_1\sqrt{\frac{\log(p+q)}{n}}, a_2\sqrt{\frac{\log(p+q)}{n}}\right]$, then with probability at least $1 - \exp(-b(s_u + \log(ep/s_u))) - \exp(-b(s_v + \log(eq/s_v))) - (p+q)^{-b}$, we have

$$\max\left\{\min_{O\in\mathcal{O}_r}\|\widehat{U} - UO\|_F, \min_{O\in\mathcal{O}_r}\|\widehat{V} - VO\|_F\right\} \leq C\frac{1}{\lambda_r^2}\sqrt{\frac{s_u s_v \log(p+q)}{n}}.$$

Comparison to other frameworks

Feature	Our method (ECCAR)	Gao et al. (2017)	Gao & Ma (2023)
Statistical Rate ($\min_{O \in \mathcal{O}_r} \|\widehat{U} - UO\|_F$)	$O\left(\frac{1}{\lambda_r^\star}\sqrt{\frac{s_u s_v \log(p+q)}{n}}\right)$	$O\left(\frac{1}{\lambda_r^\star}\sqrt{\frac{s_u r \log(p)}{n}}\right)$	$O\left(\frac{1}{\lambda_r^\star}\sqrt{\frac{(s_u + s_v) r \log(p+q)}{n}}\right)$
Computational Cost	$O(T(pn^2 + pqn))$	$O(Tpq \min(p,q))$ Prohibitive due to Fantope and multi-parameter tuning	$O(T(p+q)^3)$ Prohibitive due to Fantope and multi-parameter tuning
Key Algorithmic Steps	ADMM on a penalized regression problem (step 1) SVD (step 2)	Fantope Projection (step 1) Regression-based Refinement (step 2) Normalization (step 3)	Fantope Projection (step 1) Thresholded Gradient Descent (step 2)
Sample Splitting?	No	Yes (3-fold)	No

Discussion

Discussion

Motivation: Canonical Correlation Analysis (CCA) should be

- *scalable* for high-dimensional data,
- *interpretable* for scientific insight,
- *sparsistent* with provable guarantees.

Contribution: We introduced `ECCAR`, the sparse CCA estimator that achieves all three. The package `ccar3` is available from CRAN.

Impact:

- Matches state-of-the-art statistical performance while reducing computation.
- Applications in genetics, neuroscience, and LLM embeddings.
- Highlights how classical methods, when regularized, remain powerful in modern data science.

Future Work: Extension to multi-view datasets (e.g., multi-omics).

Acknowledgment

Claire Donnat
University of Chicago

NSERC, Discovery Grant
University of Toronto McLaughlin Center, Accelerator Grant
University of Toronto Data Science Institute, Catalyst Grant

References

C. Gao, Z. Ma, Z. Ren, and H. H. Zhou, *Minimax estimation in sparse canonical correlation analysis*, The Annals of Statistics, 43 (2015), pp. 2168–2197.

C. Gao, Z. Ma, and H. H. Zhou, *Sparse CCA: Adaptive estimation and computational barriers*, The Annals of Statistics, 45 (2017), pp. 2074–2101.

S. Gao and Z. Ma, *Sparse gca and thresholded gradient descent*, Journal of Machine Learning Research, 24 (2023), pp. 1–61.

H. Hotelling, *Relations between two sets of variables*, Biometrika, 28 (1936), pp. 321—-377.

P. G. Martin, H. Guillou, F. Lasserre, S. Déjean, A. Lan, J.-M. Pascussi, M. SanCristobal, P. Legrand, P. Besse, and T. Pineau, *Novel aspects of ppar α-mediated regulation of lipid and xenobiotic metabolism revealed through a nutrigenomic study*, Hepatology, 45 (2007), pp. 767–777.

E. Parkhomenko, D. Tritchler, and J. Beyene, *Sparse canonical correlation analysis with application to genomic data integration*, Statistical applications in genetics and molecular biology, 8 (2009).

L. Tozzi, E. Tuzhilina, M. F. Glasser, T. J. Hastie, and L. M. Williams, *Relating whole-brain functional connectivity to self-reported negative emotion in a large sample of young adults using group regularized canonical correlation analysis*, Neuroimage, 237 (2021), p. 118137.

S. Waaijenborg and A. H. Zwinderman, *Sparse canonical correlation analysis for identifying, connecting and completing gene-expression networks*, BMC Bioinformatics, 10 (2009).

I. Wilms and C. Croux, *Robust sparse canonical correlation analysis*, BMC Systems Biology, 10 (2016).

D. M. Witten and R. J. Tibshirani, *Extensions of sparse canonical correlation analysis with applications to genomic data*, Statistical applications in genetics and molecular biology, 8 (2009).

Thank you for your attention!

SAR via Alternating Lasso Regressions (Wilms et al.)

Input: data matrices $X \in \mathbb{R}^{n\times p}$, $Y \in \mathbb{R}^{n\times q}$ (centered); sparsity parameters λ_x, λ_y.

Algorithm: alternate

1. Solve lasso problem

$$b := \arg\min_b \|Xa - Yb\|_2^2 + \lambda_y \|b\|_1$$

   Normalize b.

2. Solve lasso problem

$$a := \arg\min_a \|Yb - Xa\|_2^2 + \lambda_x \|a\|_1$$

   Normalize a.

Output: (a, b).

Higher-order components: deflate X, Y

$$X := X - \frac{(Xa)(Xa)^\top}{\|Xa\|^2}, \quad Y := Y - \frac{(Yb)(Yb)^\top}{\|Yb\|^2}$$

Then repeat algorithm on the deflated data.

Fantope initialization

CCA formulation

$$\text{maximize}_A \quad \text{tr}(A^\top \Sigma A) \quad \text{subject to} \quad A^\top \Sigma_0 A = I$$

Here $A = \binom{U}{V}$ and $\Sigma_0 = \text{diag}(\Sigma_X, \Sigma_Y)$.

Rank-One Formulation:

$$\text{maximize}_F \quad \langle \Sigma, F \rangle \quad \text{subject to} \quad \Sigma_0^{\frac{1}{2}} F \Sigma_0^{\frac{1}{2}} \in \mathcal{P}$$

- $F = AA^\top$ is a $p + q \times p + q$ rank-r matrix;
- $\mathcal{P}$ encodes set of rank r projection matrices;
- we also need sparsity constraint for F.

Solution: Replace $\mathcal{P}$ with convex relaxation.

Fantope initialization

Convex Sparse Optimization:

$$\text{maximize}_{F \in \mathcal{F}} \ \langle \Sigma, F \rangle - \rho \|F\|_1$$

Fantope Relaxation:

$$\mathcal{F} = \{F : 0 \preceq F \preceq I, \ \ \text{tr}(F) = r\}$$

- ℓ_1 penalty promotes sparsity in U and V.
- F is no longer exactly rank-r, but concentrates mass on candidate supports.

Output: Extract supports (S_x, S_y) for Stage 2 refinement (CCA on reduced data).

TF-IDF

Convert text documents into numerical vectors for analysis.

Term Frequency (TF): Measures how often a word appears in a document.
Inverse Document Frequency (IDF): Downweights common words across documents.

$$\text{TF-IDF}(t, d) = \text{TF}(t, d) \times \log \frac{N}{\text{DF}(t)}$$

t: term, d: document, N: total documents, DF(t): number of documents containing t

Intuition: High TF-IDF $\Rightarrow$ word is frequent in this document but rare overall $\Rightarrow$ important/unique feature