CENTRAL CHINA NORMAL UNIVERSITY

Sparse Gaussianized Canonical Correlation Analysis with Applications to Portfolio Analysis

Presenter: Ben-Zheng Li

Central China Normal University

March 22, 2026

Reference: He Di, Hui Zou, Sparse Gaussianized Canonical Correlation Analysis with Applications to Portfolio Analysis, *Journal of the American Statistical Association*, 2026.

Motivation: Exploring Inter-variable Relationships

- **The Scientific Need**
  - Exploring linear relationships between two sets of variables is a core task in modern science.
  - **Biostatistics:** Interactions between groups of genes [4].
  - **Finance:** Correlation structures between different types of stocks (e.g., cyclical vs. non-cyclical) [2].
- **The Classical Solution: Canonical Correlation Analysis (CCA) [3]**
  - Introduced as a technique to find pairs of linear projections that maximize correlation.

Classical CCA

Optimization Framework [3]

For two random vectors $y \in \mathbb{R}^q$ and $x \in \mathbb{R}^p$, CCA seeks successive pairs of canonical vectors (α_k, β_k) to maximize the k-th canonical correlation ρ_k:

$$\rho_k = \max_{\alpha_k, \beta_k} \frac{\alpha_k^\top \Sigma_{yx} \beta_k}{\sqrt{\alpha_k^\top \Sigma_{yy} \alpha_k} \sqrt{\beta_k^\top \Sigma_{xx} \beta_k}}$$

subject to unit variance and orthogonality constraints:

$$\alpha_k^\top \Sigma_{yy} \alpha_k = 1, \beta_k^\top \Sigma_{xx} \beta_k = 1; \quad \alpha_k^\top \Sigma_{yy} \alpha_l = 0, \beta_k^\top \Sigma_{xx} \beta_l = 0 \ (\forall l < k)$$

- Σ_{yy}, Σ_{xx}: Within-set covariance matrices of y and x.
- Σ_{yx}: Between-set covariance matrix.
- Successive pairs (α_k, β_k) capture diminishing levels of association.

Challenges in Modern Data Analysis

Classical CCA faces significant hurdles when applied to high-dimensional and complex datasets.

- **High-Dimensionality ($p, q > n$)**
  - **Singularity:** Sample covariance matrices become singular, making the classical solution infeasible.
  - **Interpretability:** Linear combinations include all variables, making it hard to distinguish important variables from noise.
- **Non-normality and Heavy Tails**
  - **Model Violation:** Most sparse CCA methods rely on multivariate normal assumptions, which are often violated in practice.
  - **Financial Reality:** Asset returns are typically skewed and exhibit heavy tails (kurtosis > 3).
- **The Research Gap**
  - A robust method is needed to handle **high-dimensionality** and **non-normality** simultaneously.

Proposed Methodology

To bridge the gap between non-normality and CCA, the authors adopt a **Semiparametric Gaussian Copula Model**.

Model Assumption

$(y, x) = (Y_1, \ldots, Y_q, X_1, \ldots, X_p)$ follows a $(q+p)$-dimensional nonparametric normal distribution if there exist unknown monotone increasing functions (f, g) such that:

$$(f(y), g(x)) = (f_1(Y_1), \ldots, f_q(Y_q), g_1(X_1), \ldots, g_p(X_p)) \sim N_{q+p}(0, \Sigma)$$

where the covariance matrix is partitioned as $\Sigma = \begin{pmatrix} \Sigma_{ff} & \Sigma_{fg} \\ \Sigma_{gf} & \Sigma_{gg} \end{pmatrix}$.

Step 1: Coordinatewise Gaussianization

To estimate the unknown monotone functions (f, g), the authors utilize a **normal score estimator** to ensure the transformed variables follow a standard normal distribution.

- **Normal Score Transformation**:

$$\hat{f}_j = \Phi^{-1} \circ \left(\frac{n}{n+1} \hat{F}_j \right), \quad \hat{g}_l = \Phi^{-1} \circ \left(\frac{n}{n+1} \hat{G}_l \right)$$

- where $\hat{F}_j$ and $\hat{G}_l$ represent the empirical CDFs of variables Y_j and X_l, and the factor $\frac{n}{n+1}$ is an adjustment to avoid the boundary issue where $\Phi^{-1}(1) = \infty$.

Key Property: Invariance

The estimated canonical pairs are **invariant** against monotone transformations of any variable. This makes SGCCA highly robust to heavy-tailed distributions common in financial asset returns.

Step 2: Sparse Optimization Objective

Let $\hat{F}$ and $\hat{G}$ be the Gaussianized data matrices. SGCCA identifies the k-th pair of vectors $(\hat{\alpha}_k, \hat{\beta}_k)$ through the following penalized framework:

SGCCA Objective Function

$$(\hat{\alpha}_k, \hat{\beta}_k) = \arg\min_{\alpha_k, \beta_k} \left\{ \frac{1}{2n} \|\hat{F}\alpha_k - \hat{G}\beta_k\|_2^2 + \underbrace{\alpha_k^\top \left(\sum_{l<k} \hat{\rho}_l \hat{\Sigma}_{ff} \hat{\alpha}_l \hat{\beta}_l^\top \hat{\Sigma}_{gg} \right) \beta_k}_{\text{Orthogonal Correction}} + \lambda_{\alpha_k} \|\alpha_k\|_1 + \lambda_{\beta_k} \|\beta_k\|_1 \right\}$$

$$\text{subject to: } \alpha_k^\top \hat{\Sigma}_{ff} \alpha_k = 1 \text{ and } \beta_k^\top \hat{\Sigma}_{gg} \beta_k = 1.$$

- **Orthogonal Correction**: Incorporates info from the previous $k-1$ pairs $(\hat{\rho}_l, \hat{\alpha}_l, \hat{\beta}_l)$ to yield **nested solutions**.
- **Sparsity**: ℓ_1 penalties enable variable selection in high dimensions.

Methodology: Algorithm 1

Algorithm 1: An Iterative Algorithm for Solving SGCCA

1. Deflation and Initialization

For $k = 1$, set $\Omega_1 = I_n$. For $k \geq 2$, compute the deflation matrix: $\Omega_k = I_n - \hat{F} A_{k-1} R_{k-1} B_{k-1}^\top \hat{G}^\top / n$. Initialize $\{\hat{\alpha}_k^{(0)}, \hat{\beta}_k^{(0)}\}$.

2. Iterative Updates (Repeat for $m = 1, \ldots$ until convergence)

a) **Update** β_k: Set $\hat{Y}_k^{(m)} = \Omega_k^\top \hat{F} \hat{\alpha}_k^{(m-1)}$. Solve the Lasso-type problem:

$$\breve{\beta}_k^{(m)} = \arg\min_{\beta_k} \left\{ \frac{1}{2n} \|\hat{Y}_k^{(m)} - \hat{G}\beta_k\|_2^2 + \lambda_{\beta_k} \|\beta_k\|_1 \right\}$$

Set $\hat{\beta}_k^{(m)} = [\{\breve{\beta}_k^{(m)}\}^\top \hat{\Sigma}_{gg} \breve{\beta}_k^{(m)}]^{-1/2} \cdot \breve{\beta}_k^{(m)}$.

b) **Update** α_k: Set $\hat{X}_k^{(m)} = \Omega_k \hat{G} \hat{\beta}_k^{(m)}$. Solve:

$$\breve{\alpha}_k^{(m)} = \arg\min_{\alpha_k} \left\{ \frac{1}{2n} \|\hat{X}_k^{(m)} - \hat{F}\alpha_k\|_2^2 + \lambda_{\alpha_k} \|\alpha_k\|_1 \right\}$$

Set $\hat{\alpha}_k^{(m)} = [\{\breve{\alpha}_k^{(m)}\}^\top \hat{\Sigma}_{ff} \breve{\alpha}_k^{(m)}]^{-1/2} \cdot \breve{\alpha}_k^{(m)}$.

Theoretical Properties of SGCCA

Theorem 3: Estimation Consistency

Under mild regularity conditions and $s\sqrt{\log(p \vee q)/n} \to 0$, the SGCCA estimator $\{\hat{\alpha}_k, \hat{\beta}_k\}$ satisfies:

$$\|\hat{\beta}_k - \beta_k\|_2 = O_p\left(\sqrt{\frac{s\log(p \vee q)}{n}}\right)$$

where s is the sparsity level. This guarantees that the estimation error vanishes even when p, q grow exponentially with n.

Theorem 4: Variable Selection Consistency

Suppose the minimum signal strength satisfies $\min_{j\in\mathcal{A}} |\beta_{k,j}| \gg \sqrt{\frac{\log(p\vee q)}{n}}$. With a properly chosen penalty λ, we have:

$$\mathbb{P}\left(\text{supp}(\hat{\beta}_k) = \text{supp}(\beta_k)\right) \to 1, \quad \text{as } n \to \infty$$

- **Summary:** These theorems provide a solid mathematical foundation for SGCCA in ultra-high dimensional and non-normal settings.

Simulation Settings and Competing Methods

The performance of SGCCA is evaluated through extensive simulations, repeated 200 times for each scenario, with dimensions p, q ranging from 25 to 1600.

Comprehensive List of Competing Methods

1. **convSCCA:** Conventional sparse CCA with SCAD penalty [5].
2. **PMDSCCA:** Penalized matrix decomposition sparse CCA [8].
3. **robustSCCA:** Robust sparse CCA designed for outliers [9].
4. **mixedSCCA:** Mixed sparse CCA for hybrid data types [10].
5. **SCCA:** A modern sparse CCA framework [11].

- **Dimensions:** High-dimensional settings where p, q grow up to 1600.
- **Robustness Test:** Scenarios include heavy-tailed distributions (t-distribution) and skewed data (Lognormal) to challenge the normality assumption of conventional methods.

Simulation Results

Table 1. The average numbers of correct (C) and incorrect (I) nonzero coefficients of first two pairs of estimated canonical vectors with their standard errors (in parentheses) over 200 replicates in Covariances 1 under different cases of (n, p, q).

			normal		lognormal		transformation		contaminated normal		$t(3)$	
Case		Method	C	I	C	I	C	I	C	I	C	I
$(n = 100, p = q = 25)$	$\widehat{\mathbf{A}}$	convSCCA	3.55	5.79	3.90	11.82	4.01	9.87	3.53	6.70	3.62	8.12
			(0.06)	(0.24)	(0.07)	(0.34)	(0.06)	(0.3)	(0.07)	(0.24)	(0.07)	(0.29)
		PMDSCCA	4.95	5.48	4.20	11.12	4.66	8.13	3.36	9.02	2.87	8.62
			(0.02)	(0.16)	(0.08)	(0.41)	(0.05)	(0.32)	(0.09)	(0.37)	(0.1)	(0.33)
		robustSCCA	5.00	19.89	4.97	19.81	4.99	19.88	4.98	19.91	4.96	19.74
			(0)	(0.03)	(0.01)	(0.03)	(0.01)	(0.03)	(0.01)	(0.02)	(0.02)	(0.04)
		mixedSCCA	4.98	0.67	4.99	0.59	4.99	0.68	4.97	0.90	4.87	1.28
			(0.01)	(0.07)	(0.02)	(0.06)	(0.01)	(0.07)	(0.02)	(0.08)	(0.04)	(0.1)
		SCCA	5.00	3.68	4.48	11.66	4.83	6.24	4.69	11.47	4.46	12.58
			(0)	(0.3)	(0.07)	(0.4)	(0.04)	(0.37)	(0.05)	(0.39)	(0.06)	(0.36)
		SGCCA	4.98	3.52	5.00	3.54	5.00	3.60	4.92	6.00	4.83	8.22
			(0.01)	(0.3)	(0)	(0.28)	(0)	(0.29)	(0.03)	(0.33)	(0.04)	(0.38)
	$\widehat{\mathbf{B}}$	convSCCA	3.63	5.76	3.97	12.96	4.09	10.26	3.54	6.36	3.66	7.94
			(0.06)	(0.25)	(0.07)	(0.34)	(0.06)	(0.33)	(0.06)	(0.26)	(0.07)	(0.26)
		PMDSCCA	4.96	5.72	4.22	11.34	4.67	8.31	3.35	9.25	2.81	8.87
			(0.02)	(0.18)	(0.08)	(0.39)	(0.05)	(0.31)	(0.09)	(0.38)	(0.1)	(0.34)
		robustSCCA	4.92	19.61	4.98	19.72	4.99	19.87	4.98	19.72	4.97	19.60
			(0.02)	(0.05)	(0.01)	(0.05)	(0.01)	(0.03)	(0.01)	(0.04)	(0.01)	(0.07)
		mixedSCCA	4.99	0.77	4.99	0.65	4.99	0.68	4.96	0.84	4.88	1.36
			(0.01)	(0.07)	(0.02)	(0.07)	(0.01)	(0.06)	(0.02)	(0.07)	(0.03)	(0.11)
		SCCA	5.00	3.66	4.43	11.95	4.86	6.43	4.72	11.55	4.48	12.66
			(0)	(0.31)	(0.07)	(0.41)	(0.03)	(0.36)	(0.05)	(0.39)	(0.06)	(0.34)
		SGCCA	4.98	3.61	5.00	3.30	5.00	3.42	4.93	5.68	4.87	8.21
			(0.01)	(0.29)	(0)	(0.27)	(0)	(0.26)	(0.02)	(0.33)	(0.03)	(0.37)

Simulation Results

Table 2. The medians and standard errors (in parentheses) of $\mathbf{Err}(\widehat{\mathbf{A}}) = \|\mathbf{P}_{\widehat{\mathbf{A}}} - \mathbf{P}_{\mathbf{A}^*}\|_F$ and $\mathbf{Err}(\widehat{\mathbf{B}}) = \|\mathbf{P}_{\widehat{\mathbf{B}}} - \mathbf{P}_{\mathbf{B}^*}\|_F$ over 200 replicates in Covariances 1 under different cases of (n, p, q).

Case		Method	normal	lognormal	transformation	contaminated normal	$t(3)$
$(n = 100, p = q = 25)$	$\hat{\mathbf{A}}$	convSCCA	1.418	1.545	1.426	1.419	1.422
			(0.02)	(0.013)	(0.017)	(0.02)	(0.019)
		PMDSCCA	0.424	1.507	0.972	1.626	1.807
			(0.018)	(0.023)	(0.029)	(0.019)	(0.018)
		robustSCCA	1.741	1.744	1.848	1.746	1.674
			(0.012)	(0.01)	(0.008)	(0.011)	(0.015)
		mixedSCCA	0.308	0.291	0.300	0.334	0.418
			(0.013)	(0.013)	(0.013)	(0.018)	(0.023)
		SCCA	0.255	1.331	0.603	0.624	1.358
			(0.009)	(0.026)	(0.026)	(0.031)	(0.033)
		SGCCA	0.275	0.272	0.272	0.349	0.478
			(0.013)	(0.011)	(0.008)	(0.02)	(0.026)
	$\hat{\mathbf{B}}$	convSCCA	1.417	1.520	1.424	1.417	1.427
			(0.02)	(0.014)	(0.019)	(0.019)	(0.02)
		PMDSCCA	0.410	1.474	0.963	1.618	1.826
			(0.018)	(0.023)	(0.029)	(0.019)	(0.017)
		robustSCCA	1.774	1.732	1.850	1.765	1.682
			(0.012)	(0.011)	(0.007)	(0.011)	(0.014)
		mixedSCCA	0.290	0.297	0.289	0.335	0.436
			(0.012)	(0.012)	(0.012)	(0.018)	(0.023)
		SCCA	0.255	1.352	0.569	0.625	1.335
			(0.008)	(0.026)	(0.027)	(0.032)	(0.033)
		SGCCA	0.268	0.257	0.261	0.356	0.488

Application to Portfolio Analysis

Table 9. Numbers of nonzero and positive coordinates for the kth canonical vectors of cyclical and non-cyclical stocks on the full dataset.

		# Nonzero coordinates				# Positive coordinates			
	Method	$k=1$	$k=2$	$k=3$	$k=4$	$k=1$	$k=2$	$k=3$	$k=4$
Cyclical stocks	CCA	80	80	80	80	55	42	38	33
	SCCA	39	46	49	1	39	22	25	1
	SGCCA	41	32	48	1	41	17	27	1
Non-cyclical stocks	CCA	80	80	80	80	56	39	38	41
	SCCA	42	46	50	1	42	26	21	1
	SGCCA	44	41	45	1	43	20	22	1

NOTE: Strictly speaking, the signs of $(\widehat{\boldsymbol{\alpha}}_k, \widehat{\boldsymbol{\beta}}_k)$ are not unique because $(-\widehat{\boldsymbol{\alpha}}_k, -\widehat{\boldsymbol{\beta}}_k)$ can also be viewed as a pair of canonical vectors. For comparability, we adjust the signs so that the first nonzero element of the canonical vector of cyclical stocks is positive.

Table 10. The average of the first canonical correlations on the training and testing datasets from cyclical versus non-cyclical stocks, and the average numbers of selected variables on the training datasets.

Method	Correlation on the training sets	Correlation on the testing sets	# Selected cyclical stocks	# Selected non-cyclical stocks
CCA	0.965 (0.001)	0.495 (0.008)	80.0 (0.00)	80.0 (0.00)
SCCA	0.946 (0.001)	0.763 (0.007)	43.0 (0.41)	43.2 (0.42)
SGCCA	0.945 (0.001)	0.776 (0.007)	40.5 (0.25)	43.5 (0.37)

NOTE: Standard errors are shown in the parentheses.

References I

- Darolles, S., Florens, J.-P., and Gouriéroux, C. (2004). Kernel-based nonlinear canonical analysis and time reversibility. *Journal of Econometrics*, 119(2), 323-353.
- Firoozye, N., Tan, V., and Zohren, S. (2023). Canonical portfolios: Optimal asset and signal combination. *Journal of Banking & Finance*, 154, 106952.
- Hotelling, H. (1936). Relations between two sets of variates. *Biometrika*, 28(3/4), 321-377.
- Wang, Y. X. R., Jiang, K., Feldman, L. J., Bickel, P. J., and Huang, H. (2015). Inferring gene-gene interactions and functional modules using sparse canonical correlation analysis. *The Annals of Applied Statistics*, 9(1), 300-323.

References II

- Parkhomenko, E., Tritchler, D., and Beyene, J. (2009). Sparse canonical correlation analysis with application to genomic data. *Statistical Applications in Genetics and Molecular Biology*, 8(1), 1-34.
- Chalise, P., and Fridley, B. L. (2012). Comparison of penalty functions for sparse canonical correlation analysis. *Computational Statistics & Data Analysis*, 56(1), 245-254.
- Rodosthenous, T., Shahrezaei, V., and Evangelou, M. (2020). Integrating multi-omics data through sparse canonical correlation analysis. *Computational and Structural Biotechnology Journal*, 18, 1149-1157.
- Witten, D. M., Tibshirani, R., and Hastie, T. (2009). A penalized matrix decomposition, with applications to sparse principal components and canonical correlation analysis. *Biostatistics*, 10(3), 515-534.

References III

- Wilms, I., and Croux, C. (2016). Robust sparse canonical correlation analysis. *Journal of Multivariate Analysis*, 146, 156-167.
- Yoon, G., Carroll, R. J., and Gaynanova, I. (2020). Sparse semiparametric canonical correlation analysis for data of mixed types. *Biometrika*, 107(3), 609-625.
- Mai, Q., and Zhang, X. (2019). An iterative algorithm for sparse canonical correlation analysis. *Journal of Multivariate Analysis*, 171, 360-376.