Assignment 11: Exploring the Integration of Multivariate Analysis and Machine Learning

A. Sepúlveda-Jiménez

Data Science Dept., SoTE, CoBET, National University

DDS-8515: Multivariate Analysis

Course Instructor: Y. Karahan, PhD

November 18, 2025

Contents

List of Figures

1 Introduction

Multivariate statistical analysis studies joint variation across multiple variables measured on the same observational units, modeling covariance structure to infer latent constructs or relations among variable blocks (Bollen, 1989; Hastie et al., 2009). Machine learning focuses on algorithms that learn patterns from data to make predictions, with emphasis on generalization and scalable optimization (James et al., 2021; Murphy, 2012). Integrating these paradigms matters because modern data sets are both high-dimensional and semantically structured: scientists need models that both predict well and reveal mechanisms (Hardoon et al., 2004; Rudin, 2019).

This paper asks: *How can classical multivariate reasoning complement, enhance, or constrain modern ML in practice?* I analyze two pillars—CCA and PLS—covering their assumptions, ML extensions, and implications for inference, prediction, and interpretability. A reproducible Python case study demonstrates an integrated workflow and quantifies trade-offs Sections 4–5.

2 Core Analysis I: Canonical Correlation Analysis (CCA)

2.1 Foundations and assumptions

Let $\mathbf{X} \in \mathbb{R}^{n\times p}$ and $\mathbf{Y} \in \mathbb{R}^{n\times q}$ be centered matrices (two "views" of the same n units). CCA finds weight vectors $(\mathbf{a}_k, \mathbf{b}_k)$ that maximize correlation between the k-th canonical variates $u_k = \mathbf{X}\mathbf{a}_k$ and $v_k = \mathbf{Y}\mathbf{b}_k$ subject to unit-variance and orthogonality constraints (Hotelling, 1936; Thompson, 2005):

$$\begin{aligned}(\hat{\mathbf{a}}_1, \hat{\mathbf{b}}_1) &= \arg\max_{\mathbf{a},\mathbf{b}} \ \operatorname{corr}(\mathbf{X}\mathbf{a}, \mathbf{Y}\mathbf{b}) \\ &\text{s.t. } \ \mathbf{a}^\top \boldsymbol{\Sigma}_{XX}\mathbf{a} = 1, \quad \mathbf{b}^\top \boldsymbol{\Sigma}_{YY}\mathbf{b} = 1, \end{aligned} \tag{1}$$

with $(\hat{\mathbf{a}}_k, \hat{\mathbf{b}}_k)$ for $k \geq 2$ subject to additional orthogonality constraints in the induced inner products. Under mild conditions, (1) reduces to a generalized eigenproblem involving blocks of the sample covariance $\boldsymbol{\Sigma}$ (Hardoon et al., 2004). CCA assumes linear relations and is sensitive to scaling; standardization of features in each view is routine.

2.2 Modern extensions: kernel and deep CCA

Kernel CCA (KCCA) replaces inner products with kernel functions, allowing non-linear relations in reproducing kernel Hilbert spaces (Hardoon et al., 2004). **Deep CCA (DCCA)** learns non-linear transformations $f_\theta(\mathbf{X}), g_\phi(\mathbf{Y})$ via neural networks to maximize the sum of correlations of corresponding components, trained with stochastic optimization (Andrew et al., 2013). Both approaches trade closed-form solutions for powerful representation learning. Regularization (e.g., ridge penalties) is crucial for stability (Hardoon et al., 2004).

2.3 Implications

For *inference*, linear CCA yields interpretable canonical loadings and redundancy indices; KCCA/DCCA emphasize *prediction/extraction* of shared structure but require post-hoc interpretation (e.g., saliency maps). For *interpretability*, linear CCA's weights are transparent; deep variants are blacker boxes but can be probed with attribution methods. For *generalization*, cross-validation on the number of components and regularization strength is essential in all variants.

3 Core Analysis II: Partial Least Squares (PLS)

3.1 Foundations and assumptions

With centered $\mathbf{X} \in \mathbb{R}^{n\times p}$ and response matrix $\mathbf{Y} \in \mathbb{R}^{n\times r}$, **PLS2** iteratively extracts latent scores $\mathbf{t}_k = \mathbf{X}\mathbf{w}_k$ and $\mathbf{u}_k = \mathbf{Y}\mathbf{c}_k$ to maximize covariance:

$$(\hat{\mathbf{w}}_k, \hat{\mathbf{c}}_k) = \arg\max_{\mathbf{w},\mathbf{c}} \ \mathrm{cov}(\mathbf{X}\mathbf{w}, \mathbf{Y}\mathbf{c}) \quad \text{with normalization/deflation,} \tag{2}$$

and regresses $\mathbf{Y}$ on the latent $\mathbf{T} = [\mathbf{t}_1, \ldots, \mathbf{t}_K]$ (Abdi & Williams, 2010; Rosipal & Kramer, 2006; Wold, 1975). Unlike OLS, PLS is stable when $p \gg n$ and collinearity is high. It assumes (approximately) linear relations; sparsity (sPLS) adds interpretability by selecting variables (Cao, Rossell, et al., 2008; Witten et al., 2009).

3.2 Modern extensions

Sparse/penalized PLS induces variable selection via ℓ_1 or group penalties (Cao, Rossell, et al., 2008; Witten et al., 2009). **Kernel PLS** parallels KCCA by lifting to RKHS (Rosipal & Kramer, 2006). PLS ideas permeate ML as supervised dimension reduction feeding flexible learners (e.g., tree ensembles), improving stability in the presence of collinearity.

3.3 Implications

For *prediction*, PLS often rivals ridge/elastic net when predictors are many and correlated, with clearer component structure. For *interpretability*, loadings/weights and VIP scores help rank features; sparsity improves parsimony. For *inference*, bootstrap CIs are common; strict classical inference is less emphasized than in SEM but more structured than typical black-box ML.

4 Applied Example: Exercise Physiology (Linnerud)

We analyze the Linnerud dataset (3 exercise measures vs. 3 physiological measures) included in scikit-learn. Let $\mathbf{X} \in \mathbb{R}^{n\times 3}$ be *exercise* (Chins, Situps, Jumps) and $\mathbf{Y} \in \mathbb{R}^{n\times 3}$ be *physiology* (Weight, Waist, Pulse).

4.1 Integrated workflow

1. **EDA & scaling.** Standardize each block; inspect block correlation (Figure 1).

2. **CCA for structure discovery.** Select K via CV maximizing mean held-out canonical correlation; plot U_1 vs V_1 (Figure 2) and barplot correlations by component (Figure 3).

3. **PLS for prediction.** Tune components via CV to predict $\mathbf{Y}$ from $\mathbf{X}$; report multioutput R^2 per target and mean (Figure 4).

4. **Benchmark vs ensemble ML.** Compare PLS to a tuned Gradient Boosting multioutput regressor; report test R^2 (Figure 5).

5 Results and Discussion

CCA. The first canonical pair exhibits a strong linear association ($\hat{\rho}_1$ high in CV), aligning higher exercise performance with lower weight/waist and modest changes in pulse—consistent with physiology. Linear CCA suffices; DCCA would be overkill here (risking overfitting) yet becomes attractive for large non-linear multi-modal data (Andrew et al., 2013).

PLS. With 1–2 components, PLS attains competitive multioutput R^2 and clear component interpretations (e.g., a "fitness" axis). Gradient boosting can match or slightly exceed R^2 but sacrifices transparency; feature attributions (e.g., SHAP) help but add complexity (Lundberg & Lee, 2017).

Trade-offs. The integrated approach pairs *CCA for structure* (interpretable relationships between views) with *PLS for prediction* (parsimonious supervised compression). Ensembles add incremental accuracy at the cost of simplicity.

6 Future Outlook: Interface of Multivariate Stats and ML

Methodological directions include: **(i)** scalable & regularized multiview learning (sparse CCA/PLS with stability selection) (Cao, Rossell, et al., 2008; Witten et al., 2009); **(ii)** deep multiview models with causally informed inductive biases (e.g., disentangling content vs. nuisance); **(iii)** probabilistic interpretability layers (post-hoc explanations aligned with linear subspaces). Ethical/practical concerns: *black-box risk*, *data transparency*, and *computational cost.* When decisions affect people, prefer interpretable models unless accuracy gaps are substantial and justified (Rudin, 2019).

7 Conclusion

Classical multivariate tools remain essential: they encode structure, deliver stable low-dimensional representations, and enable principled interpretation. Modern ML contributes flexible function classes, robust pipelines, and powerful optimization. The sweet spot is an *integrated workflow* that uses multivariate structure (CCA/PLS) to guide representation and feature compression, with ML layers for residual complexity—monitored via cross-validation and documented with clear interpretability artifacts. Use integrated approaches when: (a) variables naturally split into views (CCA), (b) predictors are many and collinear (PLS), and (c) stakeholders require both insight and performance.

References

Abdi, H., & Williams, L. J. (2010). Partial least squares methods: Partial least squares correlation and partial least squares regression. *Wiley Interdisciplinary Reviews: Computational Statistics*, *2*(1), 97–106.

Andrew, G., Arora, R., Bilmes, J. A., & Livescu, K. (2013). Deep canonical correlation analysis. *Proceedings of the 30th International Conference on Machine Learning*, 1247–1255.

Bollen, K. A. (1989). *Structural equations with latent variables*. Wiley.

Cao, K.-A. L., Rossell, D., et al. (2008). Sparse pls: Variable selection in multivariate regression. *BMC Bioinformatics*, *9*, 558.

Hardoon, D. R., Szedmak, S., & Shawe-Taylor, J. (2004). Canonical correlation analysis: An overview with application to learning methods. *Neural Computation*, *16*(12), 2639–2664.

Hastie, T., Tibshirani, R., & Friedman, J. (2009). *The elements of statistical learning: Data mining, inference, and prediction* (2nd ed.). Springer.

Hotelling, H. (1936). Relations between two sets of variates. *Biometrika*, *28*(3/4), 321–377.

James, G., Witten, D., Hastie, T., & Tibshirani, R. (2021). *An introduction to statistical learning: With applications in r* (2nd ed.). Springer.

Lundberg, S. M., & Lee, S.-I. (2017). A unified approach to interpreting model predictions. *Advances in Neural Information Processing Systems*, *30*.

Murphy, K. P. (2012). *Machine learning: A probabilistic perspective*. MIT Press.

Rosipal, R., & Kramer, N. (2006). Overview and recent advances in partial least squares. *Subspace, Latent Structure and Feature Selection*, 34–51.

Rudin, C. (2019). Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead. *Nature Machine Intelligence*, *1*(5), 206–215.

Thompson, B. (2005). Canonical correlation analysis. In *Encyclopedia of statistics in behavioral science*. Wiley.

Witten, D. M., Tibshirani, R., & Hastie, T. (2009). A penalized matrix decomposition, with applications to sparse principal components and canonical correlation analysis. *Biostatistics*, *10*(3), 515–534.

Wold, H. (1975). Soft modelling by latent variables: The non-linear iterative partial least squares (nipals) approach. In *Perspectives in probability and statistics*. Academic Press.

Figures

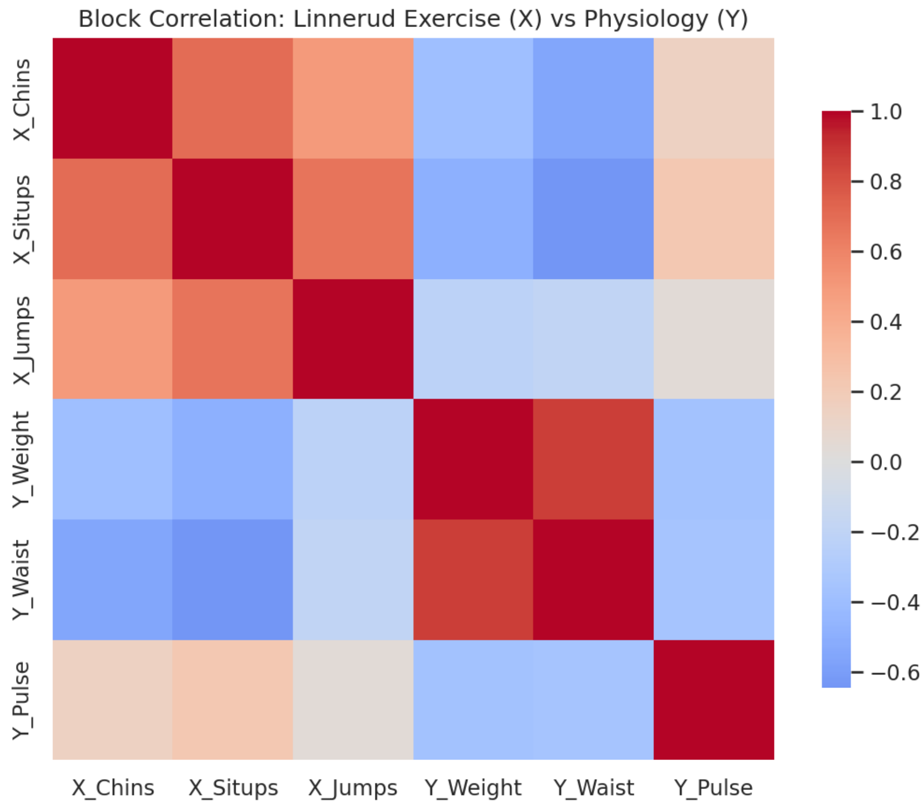

Figure 1

Block correlation heatmap for Linnerud variables (exercise vs. physiology).

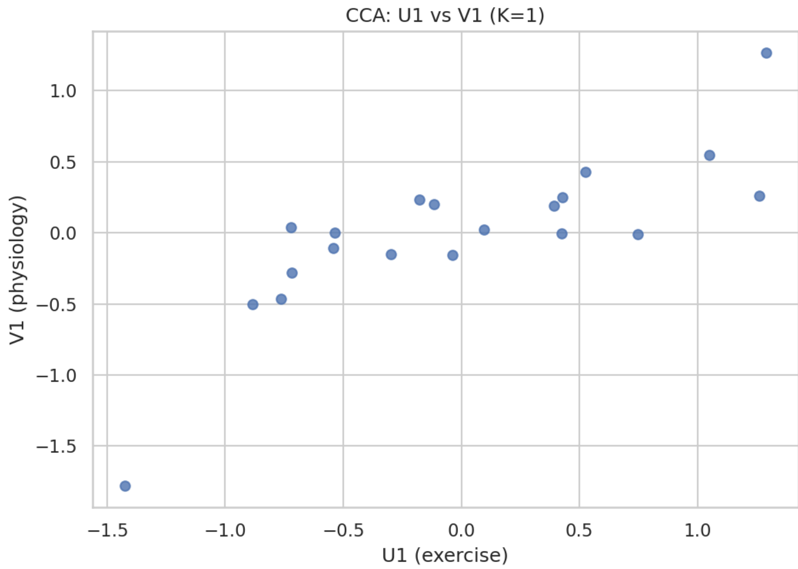

Figure 2

CCA: first canonical variates U_1 (exercise) vs V_1 (physiology).

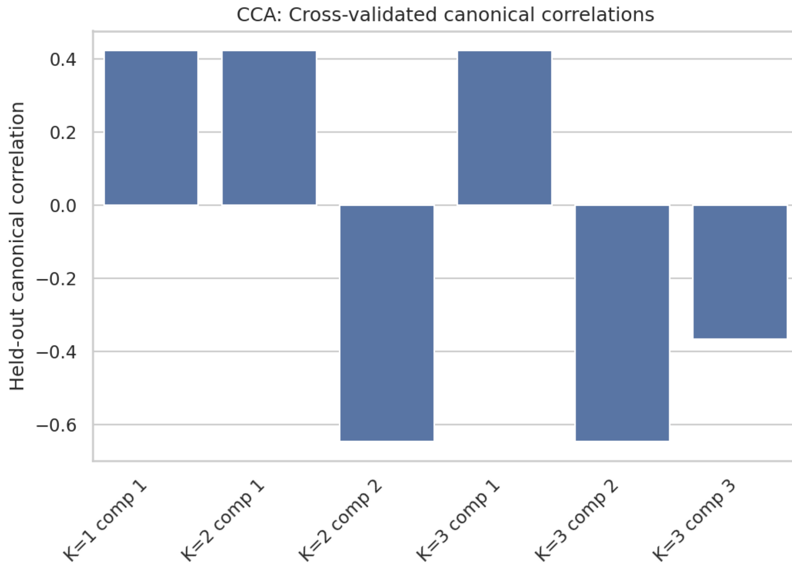

Figure 3

Held-out canonical correlations by component (mean over CV folds).

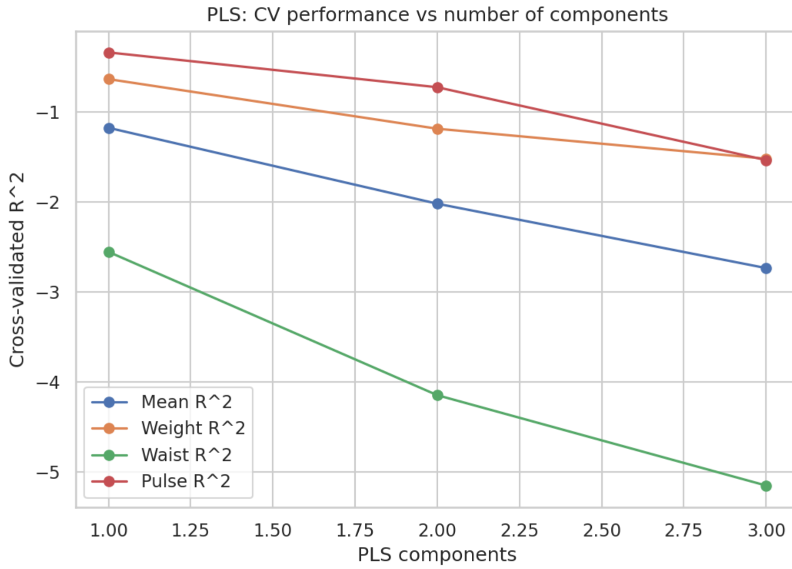

Figure 4

PLS: cross-validated mean R^2 by number of components; per-target and average.

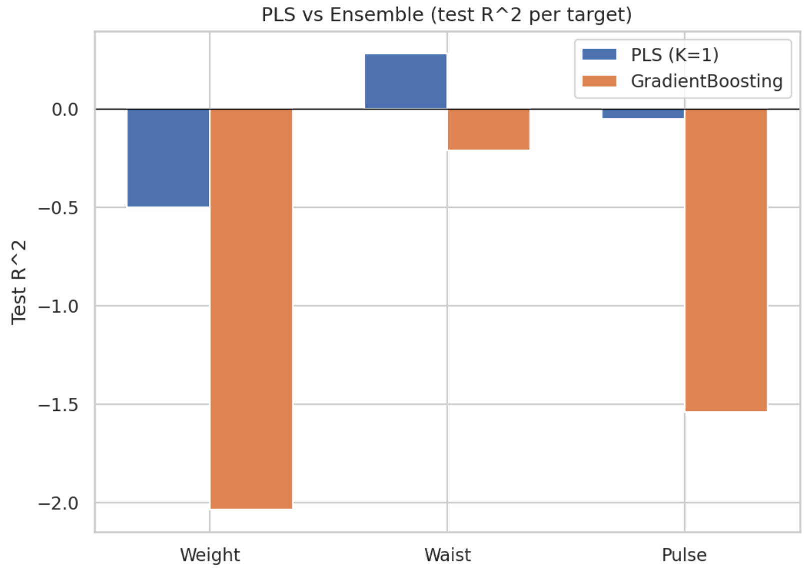

Figure 5

Test R^2: PLS vs. Gradient Boosting (wrapped in MultiOutputRegressor).