PCA Structured lasSO (PCASSO)

Jeremy Rubin[1], Jarcy Zee[1,2]

[1]Department of Biostatistics, Epidemiology, and Informatics, University of Pennsylvania; [2]Children's Hospital of Philadelphia

INTRODUCTION

- Kidney disease diagnoses have conventionally been based on visual assessment of structural changes in the kidney biopsy tissue
- Digital pathology and computational image analysis methods provide an opportunity to extract additional information from biopsy images
- **Pathomic features**: Computer-generated quantitative measurements derived from segmented histologic objects
  - Quantify heterogeneity of histologic objects
  - E.g., **tubule-specific characteristics** of shape, texture, orientation[3]
- **Pathomic feature-based prediction[4] of clinical outcomes may be more reliable than…**
  - Using clinical data alone
  - Standard pathology descriptors only (i.e., from pathologist's manual visual assessments)
- **Previous use of pathomic features: aggregated to the patient level**: **loss of potentially useful information!**

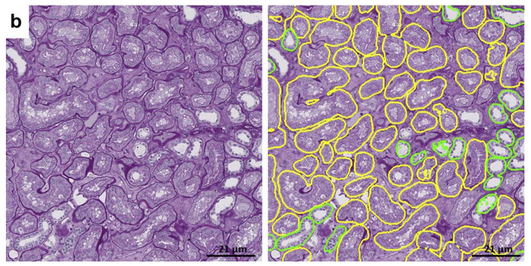

Figure 1. Segmentation of proximal (yellow) and distal (green) tubules. Adapted from *Development and evaluation of deep learning-based segmentation of histologic structures in the kidney cortex with multiple histologic stains*[7].

DATA STRUCTURE

- Scalar outcome y_i for each subject $i = 1, \dots, n$
- Matrix-valued predictors $X_i \in \mathbb{R}^{p_i \times q}$, with $q < p_i$, p_i large

X_1	Feature 1	Feature 2	…	Feature q
Tubule 1	x_{11}	x_{12}	…	x_{1q}
Tubule 2	x_{21}	x_{22}	…	x_{2q}
…	…	…	…	…
Tubule p_1	$x_{p_1 1}$	$x_{p_1 2}$	…	$x_{p_1 q}$

Histologic object on one dimension and features on the other

X_n	Feature 1	Feature 2	…	Feature q
Tubule 1	x_{11}	x_{12}	…	x_{1q}
Tubule 2	x_{21}	x_{22}	…	x_{2q}
…	…	…	…	…
Tubule p_n	$x_{p_n 1}$	$x_{p_n 2}$	…	$x_{p_n q}$

Figure 2. Structure of matrix-valued predictors of features per histologic object (e.g., tubule)

METHODS

- **PCASSO**: **Novel scalar-on-matrix regression technique using structured lasso with the first q PCA components of $X_i^T \in \mathbb{R}^{q \times p_i}$ as predictors, $q < p_i$**
  - Allows for unbalanced X_i
  - Preserves hierarchical structure
  - Estimable feature-level effects β
  - Enforces sparsity on row/column-level effects
  - Built-in dimensionality reduction
  - Collinearity ↓ in segmented object dimension
- **Structured Lasso**: Performs scalar-on-matrix regression for **balanced matrices** by solving the optimization problem[12]

$$\mathbf{argmin}_{(\alpha,\beta)\in\epsilon} \frac{1}{n}\sum_{i=1}^{n}\left(y_i - \alpha^T X_i \beta\right)^2 + \lambda_n \|\alpha\|_1 \|\beta\|_1$$

- **Principal Components Analysis (PCA)**: Finds linear transformation that maximizes variability in a dataset, and uses this transformation to get uncorrelated variables[4]
  - Decomposes matrix Z_i into orthogonal scores T_i and loadings L_i: $Z_i = T_i L_i^T$
  - Regression on first a_i columns of T_i: Large dimensionality reduction (for $a_i \ll$ minimum number of rows and columns)
  - Regression on principal components (PCs)[1,5,8-10] reduces collinearity of predictors → advantageous for lasso[2]
- X_i^T assumed to be rank $r \leq q$, as $q < p_i$
- Maximum number of independent PCs of X_i^T is q, and each PC is a q-dimensional vector
- Stack first q PCs of X_i^T row-wise to form X_i^*
- Each X_i^* has dimension $q \times q$ to enforce **balanced design:** dimensionality reduction that encapsulates the underlying structure of X_i^T

SIMULATION SETUP

- **Number of subjects**: $n = 200$
- **Number of features**: $q = 50$
- P_i: Population of objects which exist for subject i, from which p_i are observed
  - $P = P_1 = \cdots = P_n = 500$
  - Assumption: No segmentation errors (p_i objects correctly identified)
- $\widetilde{X}_i \in \mathbb{R}^{P \times q}$: **Matrix-valued predictors with full information** (features on all objects, not observed in real data) and i.i.d. $N(0,1)$ entries
- **Row coefficients** $\alpha^* \in \mathbb{R}^{P \times 1}$ and **column coefficients** $\beta^* \in \mathbb{R}^{q \times 1}$, each with i.i.d. $N(0,1)$ entries
- **Sparsity**
  - Let $s_\alpha = 90$ denote the sparsity coefficient index for α^* and $s_\beta = 96$ denote the sparsity coefficient index for β^*
  - Randomly sampled s_α and s_β percent of indices of α^* and β^*, respectively, and set α^* and β^* to be zero at these locations
- **Outcome**: $y_i = (\alpha^*)^T \widetilde{X}_i \beta^* + \epsilon_i, \epsilon_i \sim N(0,1)$
- $M = \{p_i : 1 \leq i \leq n\}, \sigma_M^2 = Var(M), \mu = E(M)$
  - $\mu = 250$
  - $\sigma_M^2 = 500$
- Sampled each p_i from a discrete uniform distribution $U(a,b)$ with midpoint μ and (a,b) such that $H = b - a + 1$ is the largest odd integer less than or equal to $\sqrt{12\sigma_M^2 + 1}$
- $X_i \in \mathbb{R}^{p_i \times q}$: **Matrix-valued predictors of observed objects** derived by randomly sampling p_i rows of $\widetilde{X}_i$

SIMULATION RESULTS

- **Naïve approaches using structured lasso**
  - **Max p_i balancing**: Resample[6,11] rows of X_i until number of rows $= \max\{p_i : 1 \leq i \leq n\}$
  - **Min p_i balancing**: Keep first $\min\{p_i : 1 \leq i \leq n\}$ rows of X_i
- **Naïve aggregated approach using standard lasso**: For each subject, average feature $j = 1, \dots, q$ across tubules
- **Model fitting**
  - 80/20 training/testing split
  - λ_n (lasso shrinkage parameter) chosen per method with 5-fold cross validation on training data

	PCASSO	Max p_i Balancing	Min p_i Balancing	Aggregated
% β correctly identified as nonzero	80%	3%	5%	4%
% β entries with correct positive sign	41%	2%	2%	0%
% β entries with correct negative sign	42%	1%	1%	1%

Table 1. Performance metrics comparing PCASSO to naïve scalar-on-matrix regression and aggregation methods under consideration from simulation study.

CONCLUSIONS

- PCASSO **most consistently identifies true nonzero feature effects** → better identification of which pathomic features are most informative of clinical outcome
- PCASSO **more consistently identifies correct sign of nonzero feature effects** → better identification of the directions of associations of pathomic features and clinical outcome
- Simulation results are preliminary; further simulations and real data analysis are needed to confirm performance of PCASSO

REFERENCES

1. Bair, E., Hastie, T., Paul, D., & Tibshirani, R. (2006). Prediction by Supervised Principal Components. Journal of the American Statistical Society, 101, 119–137. https://doi.org/10.1198/016214505000000628
2. Bühlmann, P., Rütimann, P., Geer, S. van de , & Zhang, C.-H. (2013). Correlated variables in regression: Clustering and sparse estimation. Journal of Statistical Planning and Inference , 143(11), 1835–1858. https://doi.org/10.1016/j.jspi.2013.05.019
3. Eloyan, A., Yue, M. S., & Khachatryan, D. (2020). Tumor Heterogeneity Estimation for Radiomics in Cancer. Statistics in Medicine, 39(30), 4704–4723. https://doi.org/10.1002/sim.8749
4. Ginsburg, S. B., Viswanath, S. E., Bloch, B. N., Rofsky, N. M., Genega, E. M., Lenkinski, R. E., & Madabhushi, A. (2015). Novel PCA-VIP Scheme for Ranking MRI Protocols and Identifying Computer-Extracted MRI Measurements Associated with Central Gland and Peripheral Zone Prostate Tumors. *Journal of Magnetic Resonance Imaging*, *41*(5), 1383–1393. https://doi.org/10.1002/jmri.24676
5. Hill, R. C., Fomby, T. B., & Johnson, S. R. (1977). Component selection norms for principal components regression. Communications in Statistics - Theory and Methods , 6(4), 309–334. https://doi.org/10.1080/03610927708827494
6. Hu, M., Crainiceanu, C., Schindler, M. K., Dewey, B., Reich, D. S., Shinohara, R. T., & Eloyan, A. (2020). Matrix decomposition for modeling lesion development processes in multiple sclerosis. Biostatistics. https://doi.org/10.1093/biostatistics/kxaa016
7. Jayapandian, C. P., Chen, Y., Janowczyk, A. R., Palmer, M. B., Cassol, C. A., Sekulic, M., Hodgin, J. B., Zee, J., Hewitt, S. M., O'Toole, J., Toro, P., Sedor, J. R., Barisoni, L., & Madabhushi, A. (2021). Development and evaluation of deep learning-based segmentation of histologic structures in the kidney cortex with multiple histologic stains . *Kidney International*, *99*(1), 86–101. https://doi.org/10.1016/j.kint.2020.07.044
8. Jolliffe, I. (1982). A Note on the Use of Principal Components in Regression. Journal of the Royal Statistical Society. Series C (Applied Statistics), 31(3), 300–303. https://www.jstor.org/stable/2348005
9. Mansfield, E. R., Webster, J. T., & Gunst, R. F. (1977). An Analytic Variable Selection Technique for Principal Component Regression. Journal of the Royal Statistical Society. Series C (Applied Statistics), 26(1), 34–40. https://www.jstor.org/stable/2346865
10. Mevik, B.-H., & Wehrens, R. (2007). The pls Package: Principal Component and Partial Least Squares in R. *Journal of Statistical Software*, *18*(2), 1–24. https://doi.org/10.18637/jss.v018.i02
11. Ren, S., Lai, H., Tong, W., Aminzadeh, M., Hou, X., & Lai, S. (2010). Nonparametric bootstrapping for hierarchical data. Journal of Applied Statistics, 37(9), 1487–1498. https://doi.org/10.1080/02664760903046102
12. Zhao, J., & Leng, C. (2014). Structured Lasso for Regression with Matrix Covariates. Statistica Sinica , 24(2), 799–814. https://doi.org/10.5705/SS.2012.033