Submitted to the Annals of Applied Statistics
arXiv: arXiv:1912.00434

SUPPLEMENTARY MATERIAL FOR: JOINT AND INDIVIDUAL ANALYSIS OF BREAST CANCER HISTOLOGIC IMAGES AND GENOMIC COVARIATES

By Iain Carmichael[*], Benjamin C. Calhoun[†], Katherine A. Hoadley[†], Melissa A. Troester[†], Joseph Geradts[‡], Heather D. Couture[§], Linnea Olsson[†], Charles M. Perou[†], Marc Niethammer[†], Jan Hannig[†], and J.S. Marron[†]

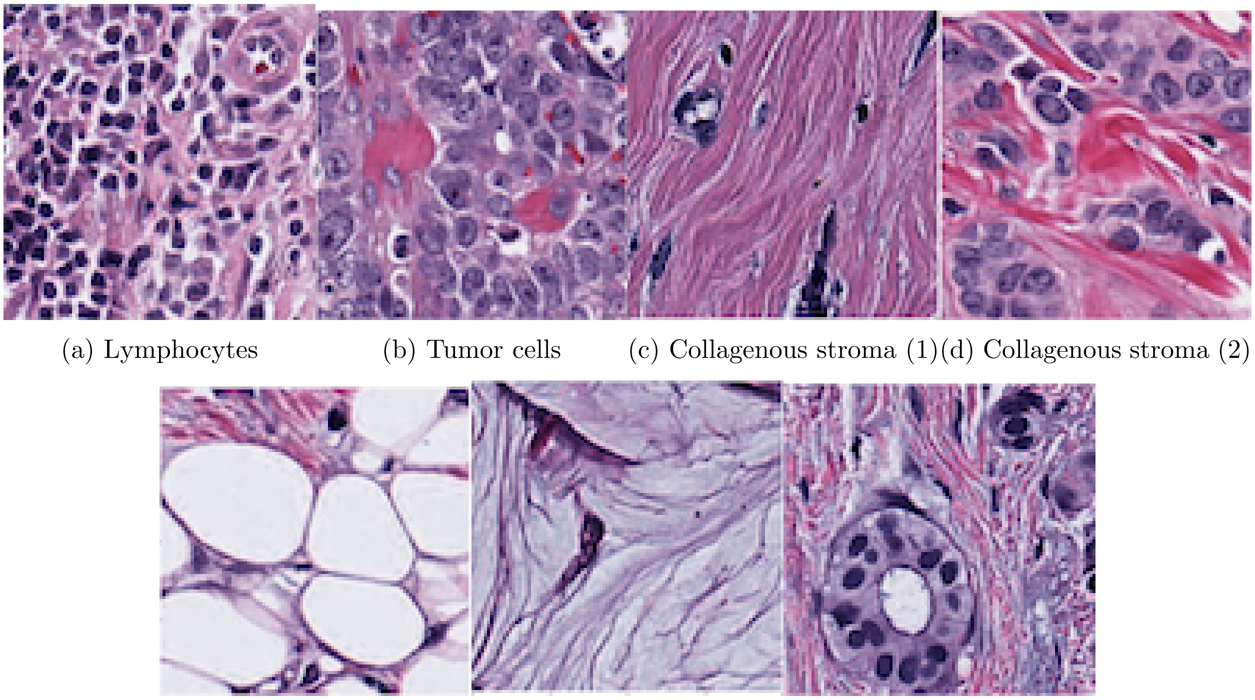
University of Washington[*], *University of North Carolina at Chapel Hill*[†], *City of Hope National Medical Center* [‡], *and Pixel Scientia Labs*[§]

Section A discusses some of the common tissue structures that play important roles in the results section of the paper. Section B provides the AJIVE diagnostic plot for the CBCS data. Section C discusses the statistical methods used to compare the AJIVE scores with the clinical variables (including multiple testing control). Additional visualizations can be found at https://marronwebfiles.sites.oasis.unc.edu/AJIVE-Hist-Gene/

APPENDIX A: COMMON TISSUE STRUCTURES

Below we give examples of some of the tissue structures which are relevant to this paper. Histopathology images are quite complex and pathologists are trained for years to interpret them[1]. For a more in-depth discussion of breast cancer pathology see Rosen (2001); Schnitt and Collins (2009).

[1] And pathologists don't always agree with each other about their interpretations Elmore et al. (2015).



(a) Lymphocytes (b) Tumor cells (c) Collagenous stroma (1) (d) Collagenous stroma (2)

(e) Fat cells (adipocytic stroma) (f) Mucin (g) Duct

Fig 1: Examples of some of the common histology structures discussed in this paper.

- Figure 1a shows tumor infiltrating lymphocytes (TILs), the nuclei of which are basophilic (dark blue or purple) in conventional hematoxylin and eosin-stained (H&E) tissue sections. Hematoxylin quantitatively stains nucleic acids (DNA and RNA). Stromal TILs are more common in certain subtypes of breast cancer and may be associated with prognosis and response to treatment.
- Figure 1b shows mostly high nuclear grade tumor cells (basophilic, dark blue or purple nuclei on H&E) as well as some stroma (eosinophilic or pink on H&E). Nuclear grade describes how abnormal the tumor cells look: “low grade” means the nuclei resemble those of normal cells and “high grade” means the nuclei are enlarged, hyperchromatic (more basophilic, darker blue/purple staining than normal nuclei), irregularly shaped and may contain multiple nucleoli. Nucleoli are small intranuclear organelles that contain DNA, RNA and protein and are responsible for the synthesis of ribosomes (ribosomes are the organelles that synthesize proteins).
- Figure 1c shows collagenous stroma which is synthesized by fibroblasts and myofibroblasts and appears as eosinophilic (pink) fibrillar extracellular material on H&E. Collagenous stroma is the connective tissue that provides the scaffolding and support for epithelial structures. The nuclei (basophilic, dark blue or purple on H&E) of a few fibroblasts are visible in a collagenous stroma.
- Figure 1d shows clusters of tumor cells separated by areas or eosinophilic (pink) collagenous stroma. On H&E-stained tissue sections, the tumor cell nuclei and their contents are basophilic (blue/purple) and the tumor cell cytoplasm varies from pale to eosinophilic (pink). This is a common histologic appearance of breast cancer as it invades into the stroma as aggregates or sheets of tumor cells.

- The large white spaces in Figure 1e are the cytoplasm of adipocytes, cells that synthesize lipids (i.e, fat). The lipid-filled cytoplasm of adipocytes appears as these optically clear areas because the solvents used in the routine preparation of H&E tissue sections dissolve the lipids, leaving blank spaces where the lipids in the cytoplasm were. The ratio of fatty stoma to collagenous or fibrous stroma varies with age. Older patients will have more fatty stroma than younger patients. This age-related decrease in breast stromal density accounts for the increased accuracy of mammography in older patients.
- Figure 1f shows extracellular mucin, a glycoprotein produced by epithelial cells which can be present in both normal and tumor tissue. Mucin appears almost clear or pale pink or blue in H&E-stained tissue sections. Invasive breast cancers with pure mucinous histology are often low-grade and are thought to have a better prognosis than invasive ductal carcinoma of no special type.
- Figure 1g shows a normal duct in the lower left and low cellularity invasive carcinoma in the upper right part of the image. The benign cells contain nuclei that lack the enlargement, irregular shape and multiple nucleoli often seen in tumor cell nuclei. The benign cells have ample pale eosinophilic (pink) cytoplasm. The cells rest on a thin basement membrane which appears as a circumferential eosinophilic (pink) band around the periphery of the duct. The optically clear space in the middle of the duct is the lumen.

REMARK A.1. *There are a couple of terms important to this paper which are similar but have different meanings. Clinical HER2 and molecular HER2 are two separate classifications used in breast cancer; the former is a immunohistochemical classification used in the clinic to determine clinical decision making while the latter is a genetic subtype. High nuclear grade refers to individual cancer cells; high tumor grade is based on a composite index including nuclear grade, tubule formation and mitotic activity. Collagenous stroma refers to (the pink) connective tissue; adipocytic stroma refers to (the white) fat cells.*

APPENDIX B: AJIVE DIAGNOSTICS

This section presents the AJIVE diagnostic plot based on the singular values of the concatenated basis matrix. This plot is described in Section 2 of (Feng et al., 2018).

The initial signal ranks are 81 (image features) and 30 (genes). There were chosen by inspection of the the difference of the log-singular values and airing on the side of picking too high a rank.

Note the Wedin bound does not provide any value for these data. This is likely due to known conservativeness of the Wedin bound for non-square matrices. The random direction bound – which can be seen to be equivalent to the classical Roy's latent root test for CCA rank selection – estimates the joint rank to be 7. This estimated joint rank was fairly robust to moderate changes in the initial signal ranks. The image individual rank is estimated to be 76 and genetic individual rank is estimated to be 25. These are likely overestimates, however, we focus only on the first few individual components.

APPENDIX C: CLINICAL DATA INTERPRETATION METHODS

In addition to the H&E images and gene expression data, we have a variety of clinical variables which can be used to interpret the different AJIVE components (e.g. PAM50 subtype). We compare each clinical variable of interest with the AJIVE scores for each component (i.e. the common normalized scores, image individual scores and genetic individual scores).

For continuous variables (e.g. proliferation score) we create a scatter plot, report the Pearson correlation and use the standard t-test test to determine if the association is statistically significant.

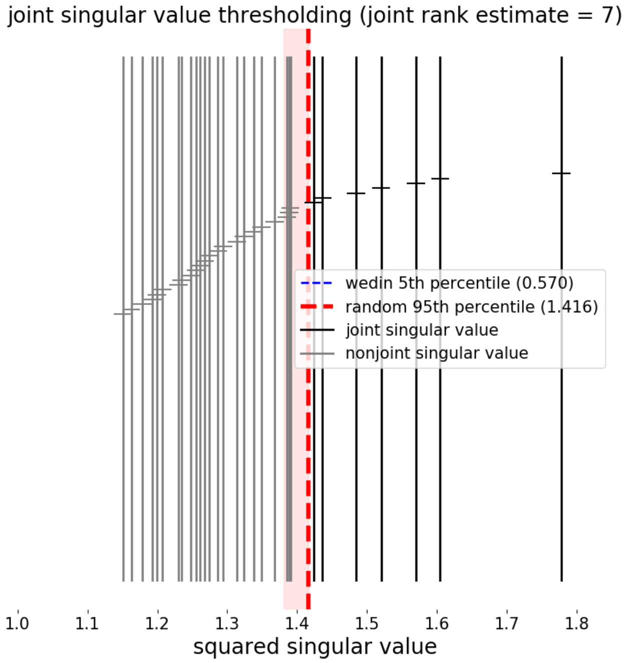


Fig 2: The AJIVE diagnostic plot shows that 7 possible joint directions are closer together than random. The vertical black/gray line segments show the principal angles between the X and Y signal subspaces (on the squared singular value scale). The vertical, red, shaded bar shows the 5-95th percentile of the distribution of leading principal angles between two random subspaces with the same dimensions as X and Y. The dashed, vertical, red line shows the random direction cutoff. The Wedin bound is less than one (off the figure) and is vacuous for this dataset. The gray line segments correspond to angles larger than the cutoff while the black line segments correspond to the angles smaller than the cutoff.

For example, Figure 7c shows the first joint component (x-axis) compared to the proliferation scores (y-axis). The text in the top left reports the Pearson correlation and is bolded if the correlation is statistically significant (after correction for multiple testing).

For categorical variables we show a conditional histogram of the scores and report difference in distribution tests for each possible class comparison using the Mann Whitney U test. This test test is used because it looks for location differences and because its test statistic is equivalent to the AUC statistic which gives an interpretable measure of how well separated two classes are. For categorical variables with more than 2 classes (e.g. there are five PAM50 subtypes) we do all one-vs-one comparisons.

For example, Figure 7b shows the first joint component (x-axis) conditioned on PAM50 subtype. The legend lists the classes, number of subjects in each class, and the the type of test. For a given class, the legend lists the other classes which were statistically significantly separated (after multiple testing adjustment) and reports the test statistics (AUC score) in parentheses. The class name is bold if at least one other class is statistically significantly separated. For example, in Figure 7b molecular Her2 is statistically significantly separated from Basal (AUC = 0.876), Luminal A (AUC = 0.897) and Normal (AUC = 0.827), but not Luminal B.

We compare each of the joint and individual AJIVE components to 33 variables. Additionally, for each multi-class categorical variable we do all of the one vs. one tests (e.g. for 5 classes we do $\binom{5}{2}$ tests). Therefore adjustment for multiple testing is necessary to avoid spurious results. For each of the AJIVE joint, image individual and genetic individual components we use the Benjamini-Hochberg procedure Benjamini and Hochberg (1995) which is implemented in Statsmodels Seabold and Perktold (2010).

Some of the clinical variables (e.g. proliferation scores, PAM50 molecular subtype) compared to the AJIVE scores are derived directly from the PAM50 gene expression which were used in the AJIVE analysis. This raises issues related to *post selection inference* when we compute p-values for these comparisons. Because the focus of this paper is on exploratory analysis we leave these issues for follow up work.

All joint, image individual and genetic individual clinical data comparisons are shown provided in Supplement B.

REFERENCES

Benjamini, Y. and Hochberg, Y. (1995). Controlling the false discovery rate: a practical and powerful approach to multiple testing. *Journal of the Royal statistical society: series B (Methodological)* **57** 289–300.

Elmore, J. G., Longton, G. M., Carney, P. A., Geller, B. M., Onega, T., Tosteson, A. N., Nelson, H. D., Pepe, M. S., Allison, K. H., Schnitt, S. J. et al. (2015). Diagnostic concordance among pathologists interpreting breast biopsy specimens. *Jama* **313** 1122–1132.

Feng, Q., Jiang, M., Hannig, J. and Marron, J. (2018). Angle-based joint and individual variation explained. *Journal of multivariate analysis* **166** 241–265.

Rosen, P. P. (2001). *Rosen's breast pathology.* Lippincott Williams & Wilkins.

Schnitt, S. J. and Collins, L. C. (2009). *Biopsy interpretation of the breast.* Lippincott Williams & Wilkins.

Seabold, S. and Perktold, J. (2010). Statsmodels: Econometric and statistical modeling with python. In *Proceedings of the 9th Python in Science Conference* **57** 61. Scipy.

I. Carmichael
Department of Statistics
University of Washington
Seattle, WA, 98195
idc9@uw.edu

B.C. Calhoun
Department of Pathology and Laboratory Medicine
University of North Carolina at Chapel Hill
Chapel Hill, NC, 27599
ben.calhoun@unchealth.unc.edu

K.A. Hoadley
Department of Genetics
Lineberger Comprehensive Cancer Center
Computational Medicine Program
University of North Carolina at Chapel Hill
Chapel Hill, NC, 27599
hoadley@med.unc.edu

M.A. Troester
Department of Epidemiology
Department of Pathology and Laboratory Medicine
University of North Carolina at Chapel Hill
Chapel Hill, NC, 27599
troester@unc.edu

J. Geradts
Department of Population Sciences
City of Hope National Medical Center
Duarte, CA 91010
jgeradts@coh.org

H.D. Couture
Pixel Scientia Labs
Raleigh, NC, 27615
heather@pixelscientia.com

L. Olsson
Department of Epidemiology
University of North Carolina at Chapel Hill
Chapel Hill, NC, 27599
lolsson@live.unc.edu

C.M. Perou
Department of Genetics
Department of Pathology and Laboratory Medicine
University of North Carolina at Chapel Hill
Chapel Hill, NC, 27599
cperou@med.unc.edu

M. Niethammer
Department of Computer Science
University of North Carolina at Chapel Hill
Chapel Hill, NC, 27599
mn@cs.unc.edu

J. Hannig
Department of Statistics
University of North Carolina at Chapel Hill
Chapel Hill, NC, 27599
jan.hannig@unc.edu

J.S. Marron
Department of Statistics
University of North Carolina at Chapel Hill
Chapel Hill, NC, 27599
marron@unc.edu