Joint and individual analysis of breast cancer histologic images and genomic covariates

Brigham and Women's Hospital computational pathology group
Harvard Medical School

Iain Carmichael

`https://idc9.github.io`

12/5/19

NSF Mathematical Sciences Postdoctoral Fellow
Department of Statistics
University of Washington

Overview of Carmichael et al. (2019)

- Develop methods to understand connections between complex and differing modalities of data
  - Histology and genetics
- Data integration using *angle-based joint and individual variation explained* (AJIVE)
  - Directly explore similarities and differences between these two modalities
- Image feature extraction with *convolutional neural networks* (CNNs)
- Methods for interpreting signals in the data captured by CNN features
- Results provide many interpretable connections and contrasts between histology and genetics

Overview of Carmichael et al. (2019)

- Develop methods to understand connections between complex and differing modalities of data
  - Histology and genetics
- Data integration using *angle-based joint and individual variation explained* (AJIVE)
  - Directly explore similarities and differences between these two modalities
- Image feature extraction with *convolutional neural networks* (CNNs)
- Methods for interpreting signals in the data captured by CNN features
- Results provide many interpretable connections and contrasts between histology and genetics

Overview of Carmichael et al. (2019)

- Develop methods to understand connections between complex and differing modalities of data
  - Histology and genetics
- Data integration using *angle-based joint and individual variation explained* (AJIVE)
  - Directly explore similarities and differences between these two modalities
- Image feature extraction with *convolutional neural networks* (CNNs)
- Methods for interpreting signals in the data captured by CNN features
- Results provide many interpretable connections and contrasts between histology and genetics

Overview of Carmichael et al. (2019)

- Develop methods to understand connections between complex and differing modalities of data
  - Histology and genetics
- Data integration using *angle-based joint and individual variation explained* (AJIVE)
  - Directly explore similarities and differences between these two modalities
- Image feature extraction with *convolutional neural networks* (CNNs)
- Methods for interpreting signals in the data captured by CNN features
- Results provide many interpretable connections and contrasts between histology and genetics

Overview of Carmichael et al. (2019)

- Develop methods to understand connections between complex and differing modalities of data
  - Histology and genetics
- Data integration using *angle-based joint and individual variation explained* (AJIVE)
  - Directly explore similarities and differences between these two modalities
- Image feature extraction with *convolutional neural networks* (CNNs)
- Methods for interpreting signals in the data captured by CNN features
- Results provide many interpretable connections and contrasts between histology and genetics

Collaborators

- **Statistics**: Jan Hannig, J.S. Marron
- **Pathology**: Benjamin Calhoun, Joseph Geradts
- **Cancer genetics**: Katherine Hoadley, Charles Perou
- **Epidemiology**: Linnea Olsson, Melissa Troester
- **Computer science**: Heather Couture, Marc Niethammer

Outline

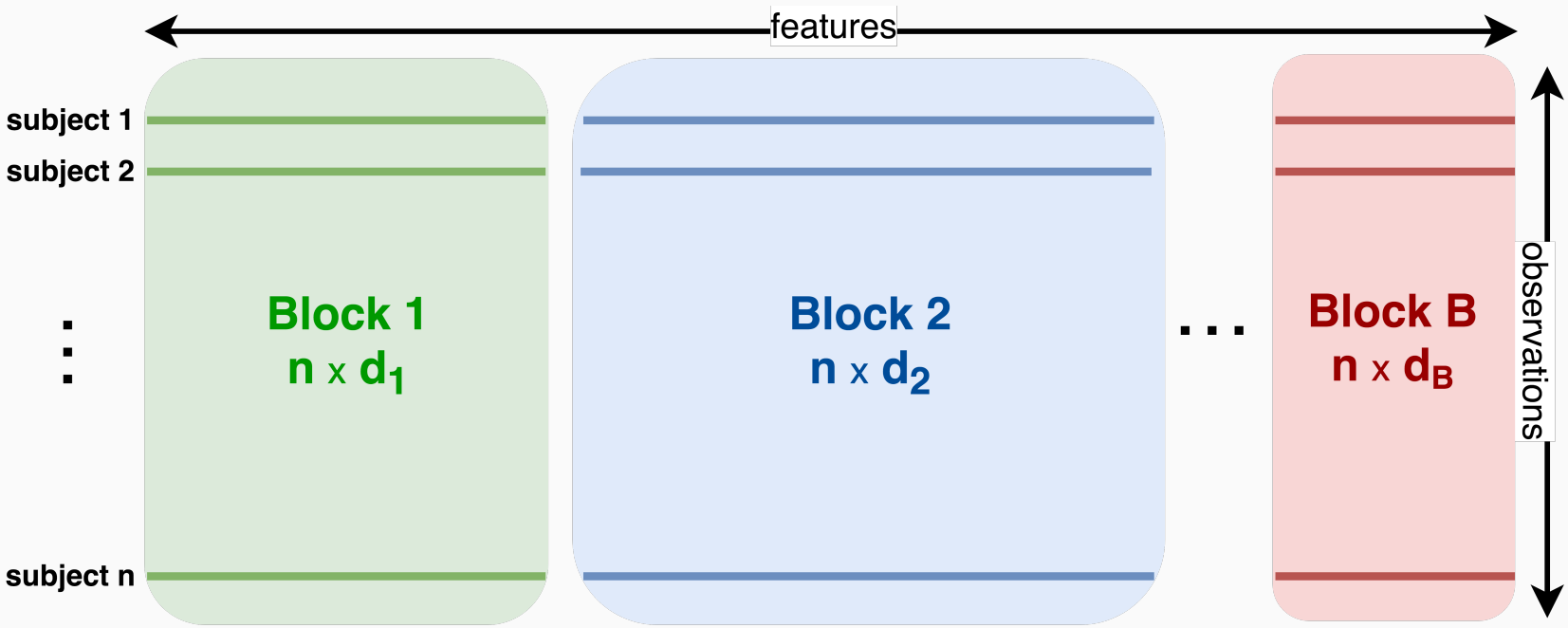
1. Angle-based joint and individual variation explained

2. Analysis of CBCS data

Angle-based joint and individual variation explained

Multi-block data setting

Fixed set of n observations, $d_1, \ldots, d_B$, sets of variables



*multi-block, muti-view, multi-omic...

Angle-based Joint and Individual Variation Explained

- Goal: find *joint* signals, if any exist, which are common to all data blocks as well as *individual* signals which are specific to each block
- Statistical inference for joint signal extraction
- Computational bottlenecks: low rank singular value decomposition and resampling procedures

(Lock et al., 2013; Feng et al., 2018)

Two block joint and individual factor model

Observed random vectors $\mathbf{x} \in \mathbb{R}^{d_x}, \mathbf{y} \in \mathbb{R}^{d_y}$

$$\mathbf{x} = A_x \mathbf{c} + B_x \mathbf{s}_x + \mathbf{e}_x$$

$$\mathbf{y} = A_y \mathbf{c} + B_y \mathbf{s}_y + \mathbf{e}_y$$

- Random latent (scores) vectors: $\mathbf{c}, \mathbf{s}_x, \mathbf{s}_y$
  - Joint signal $\mathbf{c} \in \mathbb{R}^{r_J}$, where $r_J :=$ joint rank
  - Individual signals $\mathbf{s}_x \in \mathbb{R}^{r_x}$, $\mathbf{s}_y \in \mathbb{R}^{r_y}$
- Random noise vectors $\mathbf{e}_x, \mathbf{e}_y$
- All random vectors are independent of each other
- Fixed loadings matrices A_x, A_y, B_x, B_y

AJIVE takes a different approach than maximum likelihood estimation

Two block joint and individual factor model

Observed random vectors $\mathbf{x} \in \mathbb{R}^{d_x}, \mathbf{y} \in \mathbb{R}^{d_y}$

$$\mathbf{x} = A_x \mathbf{c} + B_x \mathbf{s}_x + \mathbf{e}_x$$

$$\mathbf{y} = A_y \mathbf{c} + B_y \mathbf{s}_y + \mathbf{e}_y$$

- Random latent (scores) vectors: $\mathbf{c}, \mathbf{s}_x, \mathbf{s}_y$
  - Joint signal $\mathbf{c} \in \mathbb{R}^{r_J}$, where $r_J :=$ joint rank
  - Individual signals $\mathbf{s}_x \in \mathbb{R}^{r_x}$, $\mathbf{s}_y \in \mathbb{R}^{r_y}$
- Random noise vectors $\mathbf{e}_x, \mathbf{e}_y$
- All random vectors are independent of each other
- Fixed loadings matrices A_x, A_y, B_x, B_y

AJIVE takes a different approach than maximum likelihood estimation

AJIVE as matrix decomposition

Observed data blocks $X^1, \ldots, X^B$, $X^b \in \mathbb{R}^{n \times d_b}$

Decompose each data block into joint, individual and noise terms

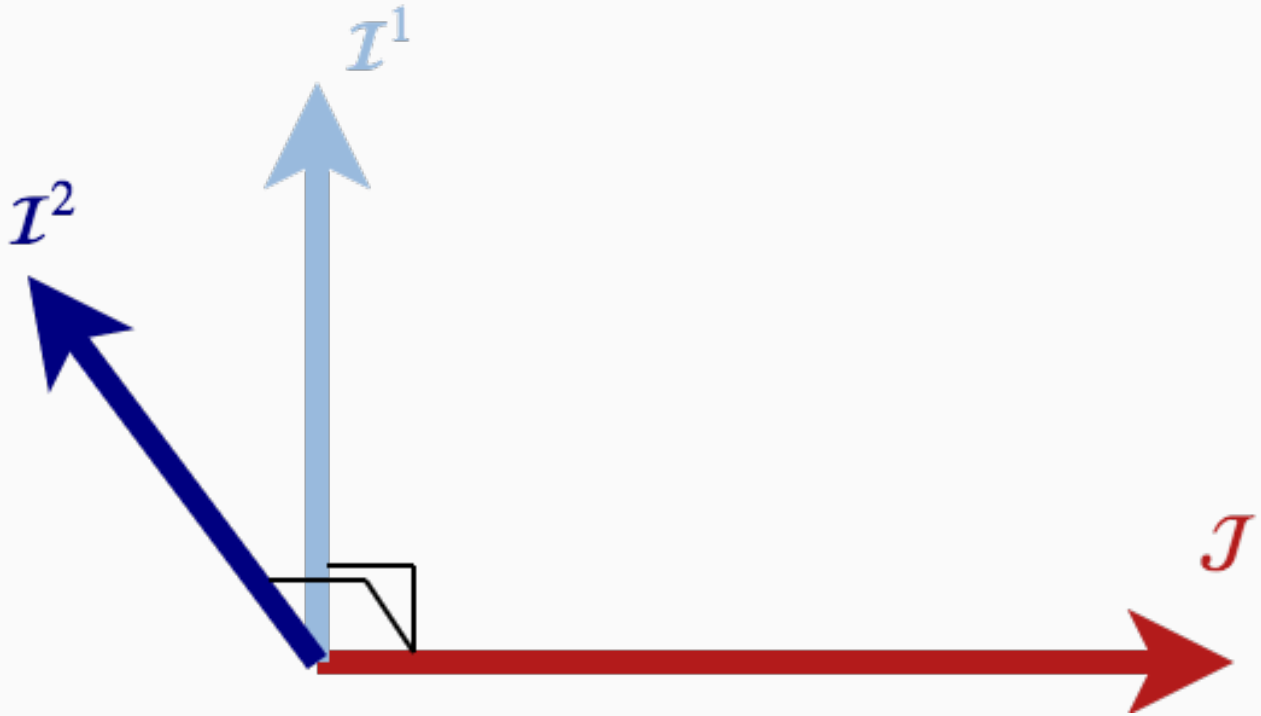
$$X^b = J^b + I^b + E^b \text{ for } b = 1, \ldots B$$

$\mathcal{J} = \text{col-span}(J^1) = \cdots = \text{col-span}(J^B)$ (1 subspace)

$\mathcal{I}^b = \text{col-span}(I^b)$ for $b = 1, \ldots, B$ (B subspaces)

$\mathcal{J} \perp \mathcal{I}^b$ for $b = 1, \ldots, B$

JIVE decomposition (n=3 samples)



Canonical correlation analysis (CCA)

Observed data blocks $X \in \mathbb{R}^{n \times d_x}$, $Y \in \mathbb{R}^{n \times d_y}$ (mean centered)

First find the most correlated directions

After finding first $k-1$ components,

$$\underset{w_x, w_y}{\text{maximize}} \ \text{corr}(Xw_x, Yw_y)$$

$$\text{s.t. } ||Xw_x||_2 = ||Yw_y||_2 = 1$$

$$Xw_{\mathbf{x}} \perp Xw_{\mathbf{x},1}, \ldots, Xw_{\mathbf{x},k-1}$$

$$Yw_{\mathbf{x}} \perp Yw_{\mathbf{x},1}, \ldots, Yw_{\mathbf{x},k-1}$$

- Loadings vector $w_{x,k} \in \mathbb{R}^{d_x}$ (similarly for y)
- Scores vector $u_{x,k} := Xw_{x,k} \in \mathbb{R}^n$ (similarly for y)
- Canonical correlation $\rho_k := \text{corr}(u_{x,k}, u_{x,k})$
- Common scores $c_k = u_{x,k} + u_{y,k}$

for $k = 1, \ldots, \min(d_x, d_y)$.

(Hotelling, 1936)

Canonical correlation analysis (CCA)

Observed data blocks $X \in \mathbb{R}^{n \times d_x}$, $Y \in \mathbb{R}^{n \times d_y}$ (mean centered)

First find the most correlated directions

After finding first $k-1$ components,

$$\underset{w_x, w_y}{\text{maximize}} \ \text{corr}(Xw_x, Yw_y)$$

$$\text{s.t. } ||Xw_x||_2 = ||Yw_y||_2 = 1$$

$$Xw_{\mathbf{x}} \perp Xw_{\mathbf{x},1}, \ldots, Xw_{\mathbf{x},k-1}$$

$$Yw_{\mathbf{x}} \perp Yw_{\mathbf{x},1}, \ldots, Yw_{\mathbf{x},k-1}$$

- Loadings vector $w_{x,k} \in \mathbb{R}^{d_x}$ (similarly for y)
- Scores vector $u_{x,k} := Xw_{x,k} \in \mathbb{R}^n$ (similarly for y)
- Canonical correlation $\rho_{\mathbf{k}} := \text{corr}(u_{x,\mathbf{k}}, u_{x,\mathbf{k}})$
- Common scores $c_{\mathbf{k}} = u_{x,\mathbf{k}} + u_{y,\mathbf{k}}$

for $k = 1, \ldots, \min(d_x, d_y)$.

(Hotelling, 1936)

Canonical correlation analysis (CCA)

Observed data blocks $X \in \mathbb{R}^{n \times d_x}$, $Y \in \mathbb{R}^{n \times d_y}$ (mean centered)

First find the most correlated directions

After finding first $k-1$ components,

$$\underset{w_x, w_y}{\text{maximize}} \ \text{corr}(Xw_x, Yw_y)$$

$$\text{s.t. } ||Xw_x||_2 = ||Yw_y||_2 = 1$$

$$Xw_{\mathbf{x}} \perp Xw_{\mathbf{x},1}, \ldots, Xw_{\mathbf{x},k-1}$$

$$Yw_{\mathbf{x}} \perp Yw_{\mathbf{x},1}, \ldots, Yw_{\mathbf{x},k-1}$$

- Loadings vector $w_{x,k} \in \mathbb{R}^{d_x}$ (similarly for y)
- Scores vector $u_{x,k} := Xw_{x,k} \in \mathbb{R}^n$ (similarly for y)
- Canonical correlation $\rho_{\mathbf{k}} := \text{corr}(u_{x,\mathbf{k}}, u_{x,\mathbf{k}})$
- Common scores $c_{\mathbf{k}} = u_{x,\mathbf{k}} + u_{y,\mathbf{k}}$

for $k = 1, \ldots, \min(d_x, d_y)$.

(Hotelling, 1936)

Principal angle analysis (PAA)

Subspaces $\mathcal{X}, \mathcal{Y} \subseteq \mathbb{R}^n$ of dimensions d_x, d_y

First find the closes pair of directions

After finding first $k-1$ components,

$$\underset{u_x, u_y \neq 0}{\text{minimize}} \ \text{angle}(u_x, u_y)$$

$$u_x \in \mathcal{X}, u_y \in \mathcal{Y}$$

$$u_{\mathbf{x}} \perp u_{\mathbf{x},1}, \ldots, u_{\mathbf{x},k-1}$$

$$u_{\mathbf{y}} \perp u_{\mathbf{y},1}, \ldots, u_{\mathbf{y},k-1}$$

- Principal vector $u_{x,k} \in \mathbb{R}^n$ (similarly for y)
  - Think of $u_{x,k}$ as either a 1-dimensional subspace or a unit vector with a fixed orientation
- Principal angle $\theta_{\mathbf{k}} := \text{angle}(u_{x,\mathbf{k}}, u_{y,\mathbf{k}})$

for $k = 1, \ldots, \min(d_x, d_y)$.

(Bjorck and Golub, 1973; Edelman et al., 1998)

Principal angle analysis (PAA)

Subspaces $\mathcal{X}, \mathcal{Y} \subseteq \mathbb{R}^n$ of dimensions d_x, d_y

First find the closes pair of directions

After finding first $k-1$ components,

$$\underset{u_x, u_y \neq 0}{\text{minimize}} \ \text{angle}(u_x, u_y)$$

$$u_x \in \mathcal{X}, u_y \in \mathcal{Y}$$

$$u_{\mathbf{x}} \perp u_{\mathbf{x},1}, \ldots, u_{\mathbf{x},k-1}$$

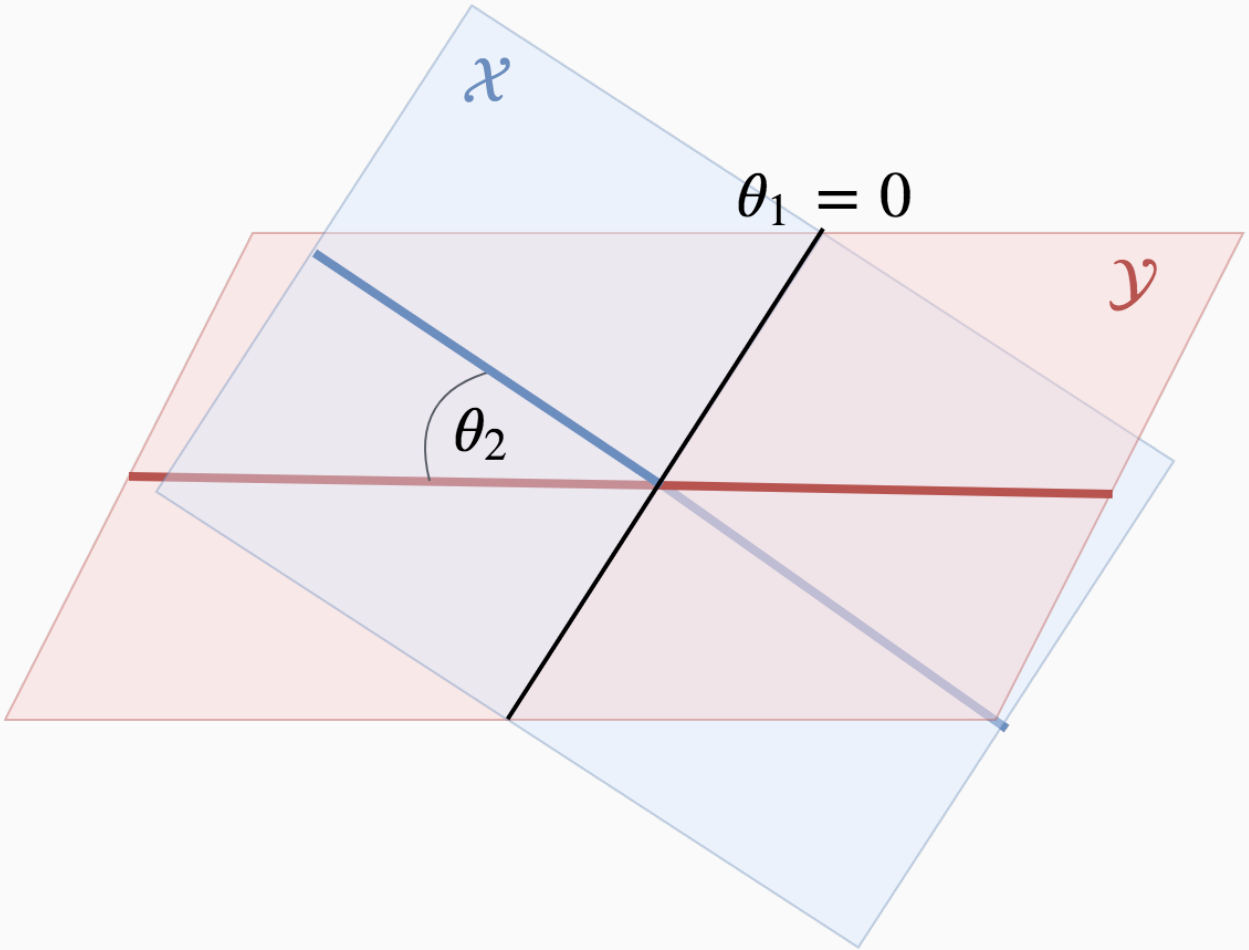
$$u_{\mathbf{y}} \perp u_{\mathbf{y},1}, \ldots, u_{\mathbf{y},k-1}$$

- Principal vector $u_{x,k} \in \mathbb{R}^n$ (similarly for y)
  - Think of $u_{x,k}$ as either a 1-dimensional subspace or a unit vector with a fixed orientation
- Principal angle $\theta_{\mathbf{k}} := \text{angle}(u_{x,\mathbf{k}}, u_{y,\mathbf{k}})$

for $k = 1, \ldots, \min(d_x, d_y)$.

(Bjorck and Golub, 1973; Edelman et al., 1998)

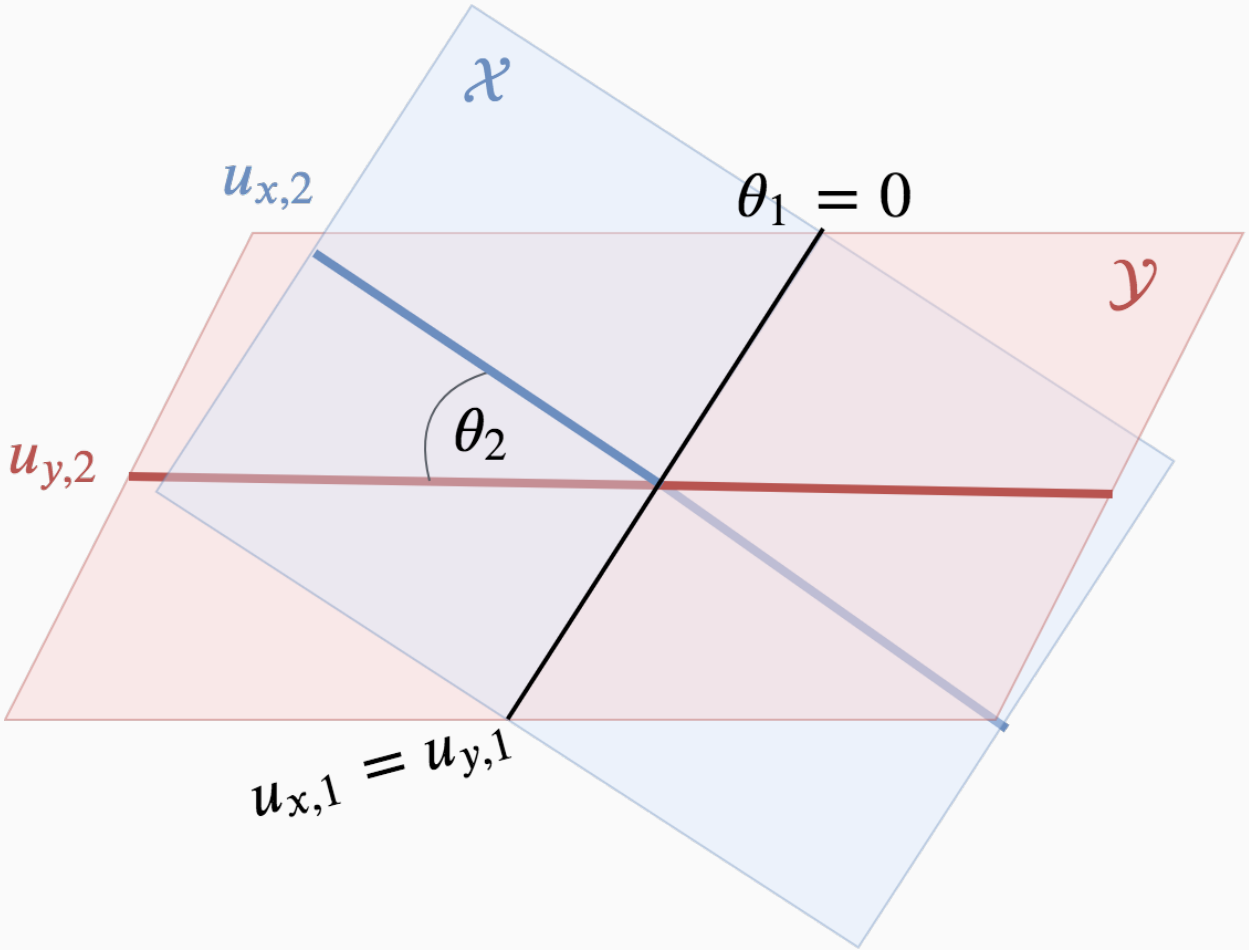
Principal Angle Analysis



CCA = principal angle analysis of col-span(X), col-span(Y)

$\rho_k = \cos(\theta_k)$, principal vectors = CCA scores

Principal Angle Analysis



CCA = principal angle analysis of col-span(X), col-span(Y)

$\rho_k = \cos(\theta_k)$, principal vectors = CCA scores

Angle between two random vectors

- Isotropic random direction $v \in \mathbb{R}^n$
  - span(v) where $v \sim N(0, I_n)$
  - Uniform distribution over directions i.e. Grassmanian(1, n)
- Random angle distribution

$$\theta = \text{angle}(v_1, v_2)$$

where v_1, v_2 are independent, isotropic random directions

Angle between two random vectors

- Isotropic random direction $v \in \mathbb{R}^n$
  - span(v) where $v \sim N(0, I_n)$
  - Uniform distribution over directions i.e. Grassmanian(1, n)
- Random angle distribution

$$\theta = \text{angle}(v_1, v_2)$$

where v_1, v_2 are independent, isotropic random directions

Random principal angle distribution

- $\mathcal{A} \subseteq \mathbb{R}^n$, a random, isotropic d dimensional subspace
  - Sample $A \in \mathbb{R}^{n \times d}$ with iid $N(0, 1)$ entries
  - $\mathcal{A} =$ col-span(A)
  - Uniform distribution on d dimensional subspaces in $\mathbb{R}^n$ i.e. Grassmanian(d, n)

$\theta \sim RPA(p, q)$ random principal angle

1. Sample isotropic d dimensional subspace $\mathcal{A}$
2. Sample isotropic p dimensional subspace $\mathcal{B}$
3. $\theta =$ smallest principal angle between $\mathcal{A}$ and $\mathcal{B}$

Random principal angle distribution

- $\mathcal{A} \subseteq \mathbb{R}^n$, a random, isotropic d dimensional subspace
  - Sample $A \in \mathbb{R}^{n \times d}$ with iid $N(0, 1)$ entries
  - $\mathcal{A} =$ col-span(A)
  - Uniform distribution on d dimensional subspaces in $\mathbb{R}^n$ i.e. Grassmanian(d, n)

$\theta \sim RPA(p, q)$ random principal angle

1. Sample isotropic d dimensional subspace $\mathcal{A}$
2. Sample isotropic p dimensional subspace $\mathcal{B}$
3. $\theta =$ smallest principal angle between $\mathcal{A}$ and $\mathcal{B}$

Joint rank selection with random direction bound (two blocks)

Observed data blocks $X \in \mathbb{R}^{n \times d_x}$, $Y \in \mathbb{R}^{n \times d_y}$ (mean centered)

Idea: retain principal vectors/CCA components which are "closer together than random"

- Compute observed principal angles between col-span(X), col-span(Y)

$$\theta \in \mathbb{R}^{\min(d_x, d_y)}$$

- Let $\theta_{threshold} = 5$th percentile of RPA(d_x, d_y) and let $\widehat{r_j}$ be the smallest j s.t. $\theta_j < \theta_{threshold}$
- Classical CCA rank selection method known as Roy's latent root test (Johnstone, 2008)

Joint rank selection with random direction bound (two blocks)

Observed data blocks $X \in \mathbb{R}^{n \times d_x}$, $Y \in \mathbb{R}^{n \times d_y}$ (mean centered)

Idea: retain principal vectors/CCA components which are "closer together than random"

- Compute observed principal angles between col-span(X), col-span(Y)

$$\theta \in \mathbb{R}^{\min(d_x, d_y)}$$

- Let $\theta_{threshold} = 5$th percentile of RPA(d_x, d_y) and let $\widehat{r}_j$ be the smallest j s.t. $\theta_j < \theta_{threshold}$
- Classical CCA rank selection method known as Roy's latent root test (Johnstone, 2008)

Joint rank selection with random direction bound (two blocks)

Observed data blocks $X \in \mathbb{R}^{n \times d_x}$, $Y \in \mathbb{R}^{n \times d_y}$ (mean centered)

Idea: retain principal vectors/CCA components which are "closer together than random"

- Compute observed principal angles between col-span(X), col-span(Y)

$$\theta \in \mathbb{R}^{\min(d_x, d_y)}$$

- Let $\theta_{threshold} = 5$th percentile of RPA(d_x, d_y) and let $\widehat{r_J}$ be the smallest j s.t. $\theta_j < \theta_{threshold}$
- Classical CCA rank selection method known as Roy's latent root test (Johnstone, 2008)

Joint rank selection with random direction bound (two blocks)

Observed data blocks $X \in \mathbb{R}^{n \times d_x}$, $Y \in \mathbb{R}^{n \times d_y}$ (mean centered)

Idea: retain principal vectors/CCA components which are "closer together than random"

- Compute observed principal angles between col-span(X), col-span(Y)

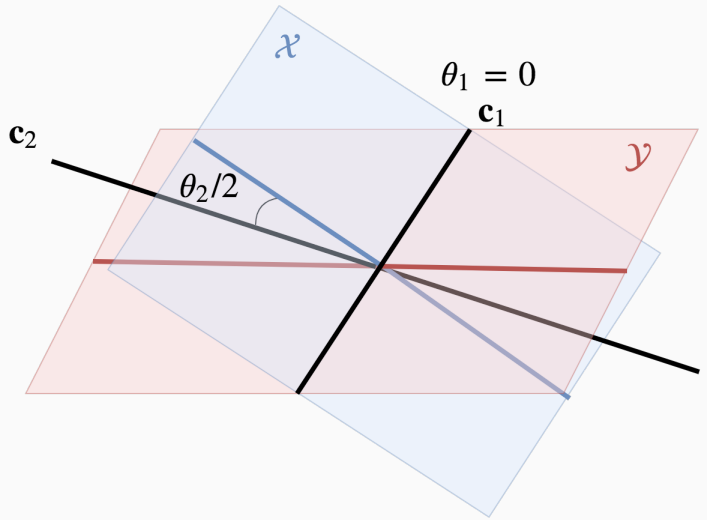
$$\theta \in \mathbb{R}^{\min(d_x, d_y)}$$

- Let $\theta_{threshold} = 5$th percentile of RPA(d_x, d_y) and let $\widehat{r_J}$ be the smallest j s.t. $\theta_j < \theta_{threshold}$
- Classical CCA rank selection method known as Roy's latent root test (Johnstone, 2008)

$B \geq 2$ blocks generalization

- SUMCORR-AVGVAR CCA (Kettenring, 1971; Nielsen, 2002; Asendorf, 2015)
- Subspace flag mean (Draper et al., 2014)
- Random direction bound generalizations
  - Any partially shared structure (Feng et al., 2018)
  - Fully shared structure only (in progress, unpublished)

Flag mean



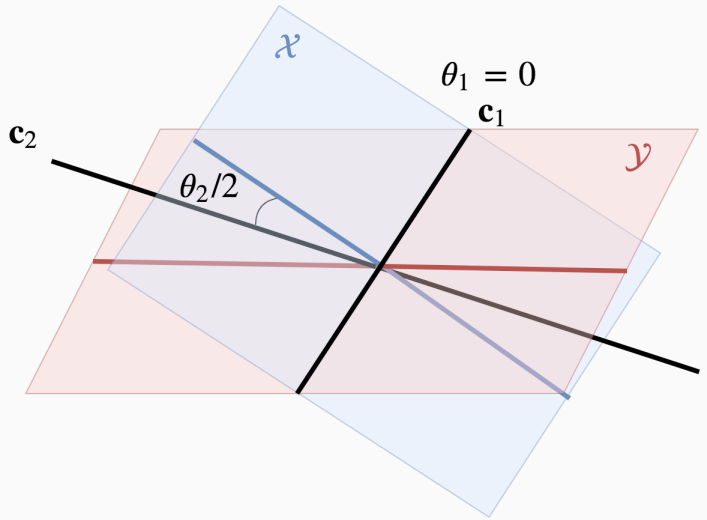
Subspace flag mean: finds the "most central" subspace of a collection of $B \geq 2$ subspaces (Draper et al., 2014)

- For $B = 2$, the flag mean is the average of the CCA scores

$$c = u_x + u_y$$

- For $B \geq 2$, a similar relation holds for SUMCORR-AVGVAR CCA

Flag mean



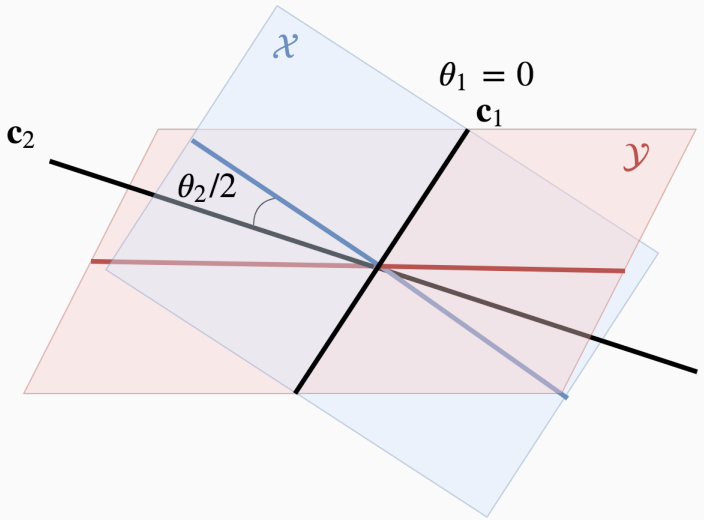
Subspace flag mean: finds the "most central" subspace of a collection of $B \geq 2$ subspaces (Draper et al., 2014)

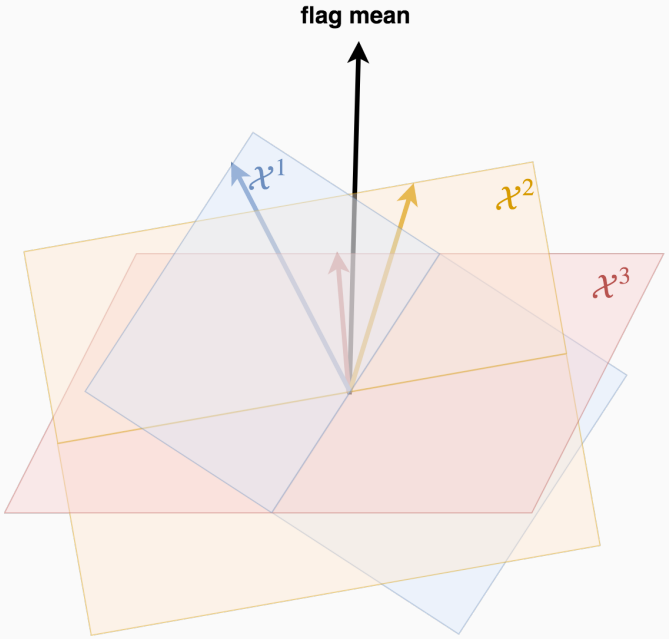
- For $B = 2$, the flag mean is the average of the CCA scores

$$c = u_x + u_y$$

- For $B \geq 2$, a similar relation holds for SUMCORR-AVGVAR CCA

Flag mean





Subspace flag mean: finds the "most central" subspace of a collection of $B \geq 2$ subspaces (Draper et al., 2014)

- For $B = 2$, the flag mean is the average of the CCA scores

$$c = u_x + u_y$$

- For $B \geq 2$, a similar relation holds for SUMCORR-AVGVAR CCA

Key components of AJIVE

- Joint rank estimation, $\widehat{r_J}$
  - Random direction bound
  - Wedin bound
- Common scores/flag mean summarize joint information
  - $C \in \mathbb{R}^{n \times \widehat{r_J}}$
- Joint and individual spaces
  - Estimated joint subspace spanned by common scores

$$\widehat{\mathcal{J}} = \text{col-span}(C)$$

  - Estimated individual information is orthogonal to joint space

$$\widehat{\mathcal{I}^x}, \widehat{\mathcal{I}^y} \subseteq \text{col-span}(C)^{\perp}$$

Key components of AJIVE

- Joint rank estimation, $\widehat{r_J}$
  - Random direction bound
  - Wedin bound
- Common scores/flag mean summarize joint information
  - $C \in \mathbb{R}^{n \times \widehat{r_J}}$
- Joint and individual spaces
  - Estimated joint subspace spanned by common scores

$$\widehat{\mathcal{J}} = \text{col-span}(C)$$

  - Estimated individual information is orthogonal to joint space

$$\widehat{\mathcal{I}^x}, \widehat{\mathcal{I}^y} \subseteq \text{col-span}(C)^{\perp}$$

Key components of AJIVE

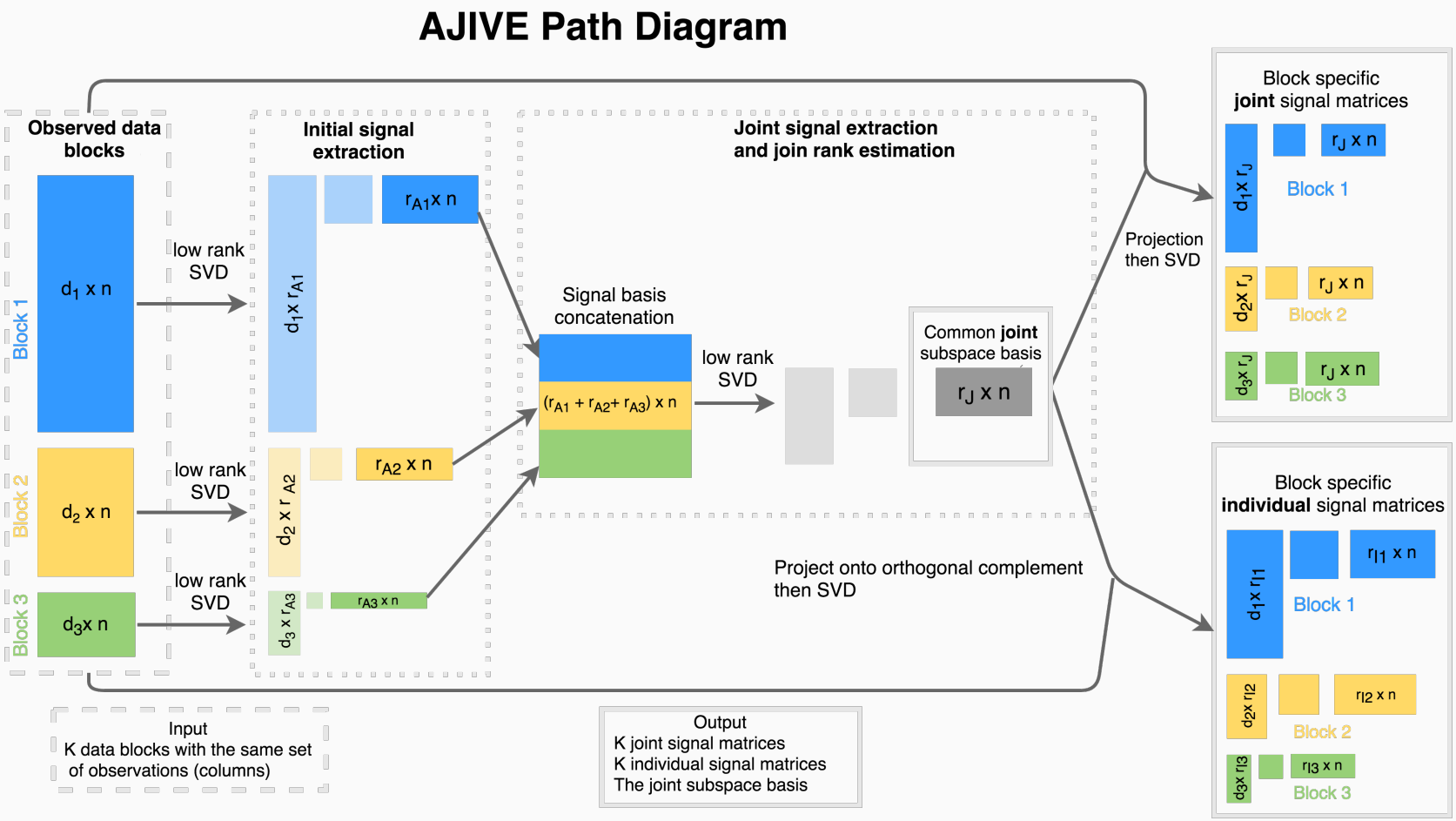
- Joint rank estimation, $\widehat{r_J}$
  - Random direction bound
  - Wedin bound
- Common scores/flag mean summarize joint information
  - $C \in \mathbb{R}^{n \times \widehat{r_J}}$
- Joint and individual spaces
  - Estimated joint subspace spanned by common scores

$$\widehat{\mathcal{J}} = \text{col-span}(C)$$

  - Estimated individual information is orthogonal to joint space

$$\widehat{\mathcal{I}^x}, \widehat{\mathcal{I}^y} \subseteq \text{col-span}(C)^{\perp}$$

AJIVE three step procedure



Details in Feng et al. (2018)

Analysis of CBCS data

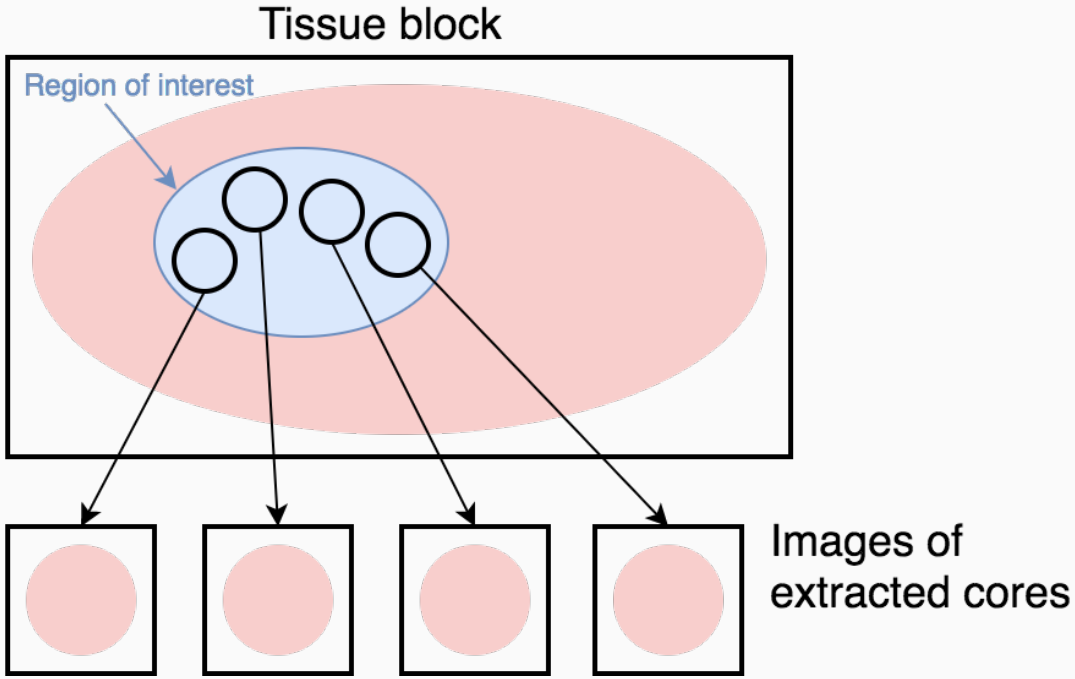
Carolina breast cancer study, phase 3 (CBCS)

- Population-based study of black and white women with invasive breast cancer diagnosed between 2008-2013 in North Carolina
  - 1191 subjects available in phase 3
- Histology
  - Pathologist selected regions of interest
  - Average of four 1mm cores per patient
- PAM50 expressions
- Other clinical variables
  - ER status, clinical Her2 status, histology type, proliferation score, PAM50 subtype, age, race, etc

(Troester et al., 2017; Allott et al., 2018)

Carolina breast cancer study, phase 3 (CBCS)

- Population-based study of black and white women with invasive breast cancer diagnosed between 2008-2013 in North Carolina
  - 1191 subjects available in phase 3
- Histology
  - Pathologist selected regions of interest
  - Average of four 1mm cores per patient



- PAM50 expressions
- Other clinical variables
  - ER status, clinical Her2 status, histology type, proliferation score

Carolina breast cancer study, phase 3 (CBCS)

- Population-based study of black and white women with invasive breast cancer diagnosed between 2008-2013 in North Carolina
  - 1191 subjects available in phase 3
- Histology
  - Pathologist selected regions of interest
  - Average of four 1mm cores per patient
- PAM50 expressions
- Other clinical variables
  - ER status, clinical Her2 status, histology type, proliferation score, PAM50 subtype, age, race, etc

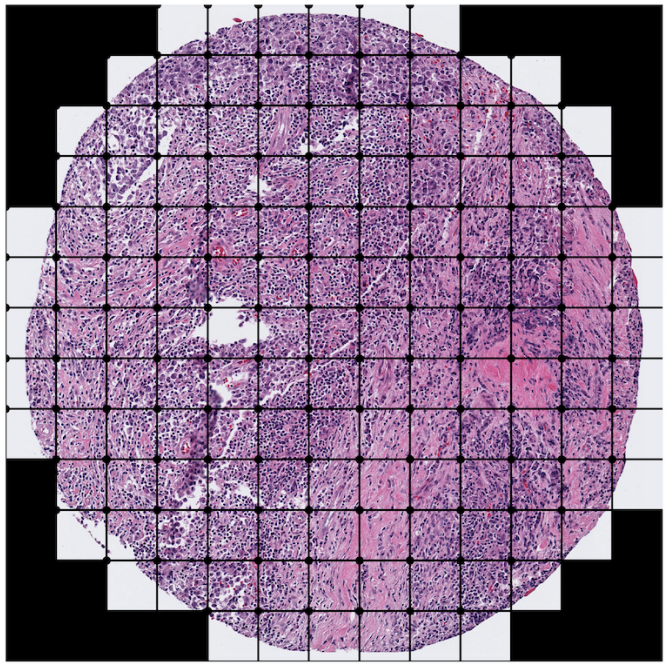
(Troester et al., 2017; Allott et al., 2018)

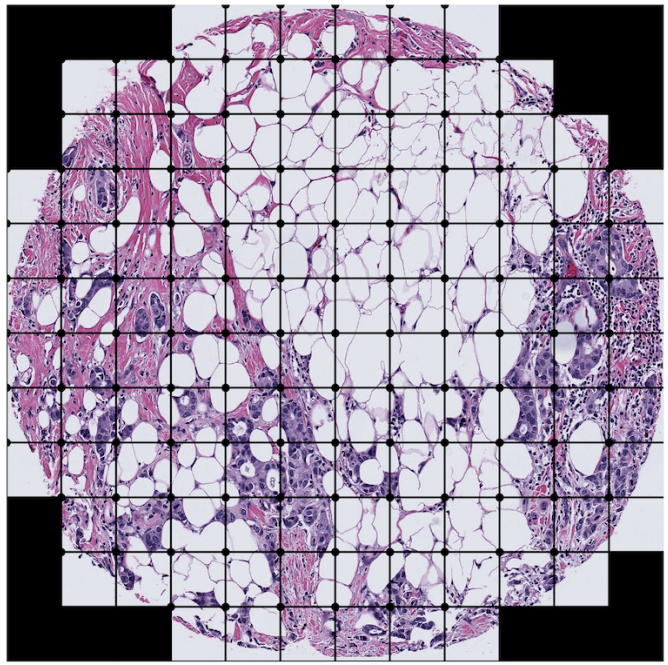
Carolina breast cancer study, phase 3 (CBCS)

- Population-based study of black and white women with invasive breast cancer diagnosed between 2008-2013 in North Carolina
  - 1191 subjects available in phase 3
- Histology
  - Pathologist selected regions of interest
  - Average of four 1mm cores per patient
- PAM50 expressions
- Other clinical variables
  - ER status, clinical Her2 status, histology type, proliferation score, PAM50 subtype, age, race, etc

(Troester et al., 2017; Allott et al., 2018)

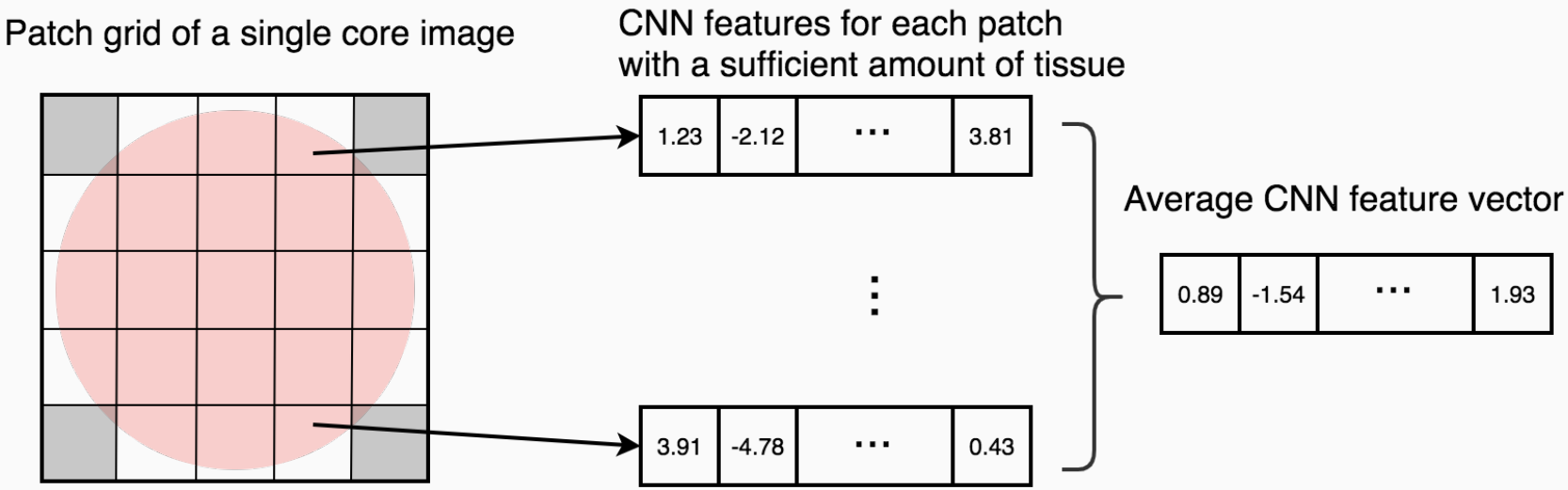
Image patch representation





- Preliminary stain normalization Macenko et al. (2009)
- Background mask estimated with weighted combination of Otsu's method and the Triangle method (Zack et al., 1977; Otsu, 1979)
- Patches with more than 90% background are ignored

Image feature extraction



- CNN features extracted from each patch
  - Transfer learning
  - Mean pool of last convolutional layer of VGG16
- Cores are represented by an average of their patch features
- Subjects are represented by an average of their cores

AJIVE output

- $X^{\text{histology}} \in \mathbb{R}^{1,191 \times 512}$, $X^{\text{genetic}} \in \mathbb{R}^{1,191 \times 50}$
- Estimated AJIVE ranks
  - Joint: 7
  - Genetic individual: 25
  - Image individual: 76
- Quantities of interest (conceptually 3 PCAs)
  - Common normalized scores and common loadings
  - Image individual scores and loadings
  - Genetic individual scores and loadings

First joint component, neural features loadings vector

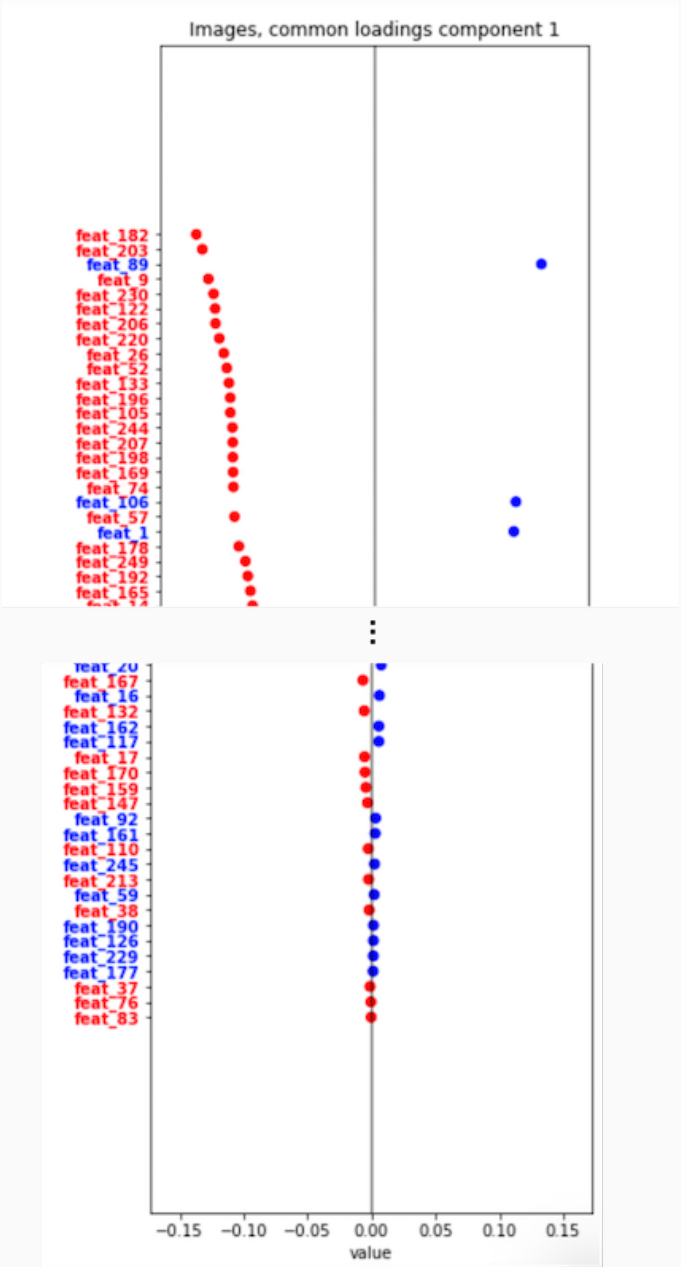


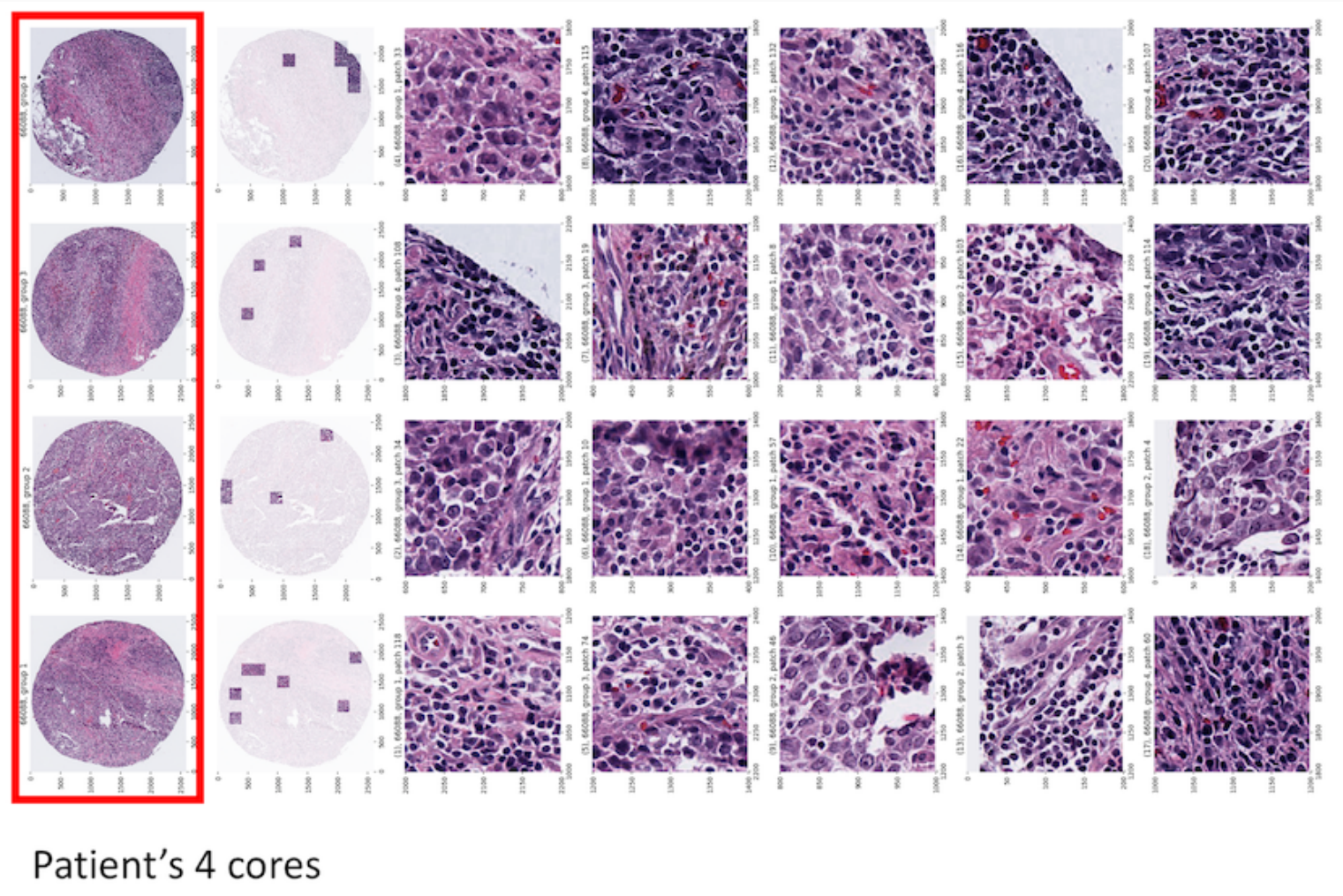
Image signal interpretation

Goal: understand what visual signals are associated with a given loadings vector (mode of variation)

- Multi-scaled approach
- For each end (positive/negative) of each AJIVE component we study
    - Patient level similarities: cores of top 15 patients (i.e. patients with most negative/positive scores)
    - Important features within each patient: representative patch views of top 15 patients

Representative patch view

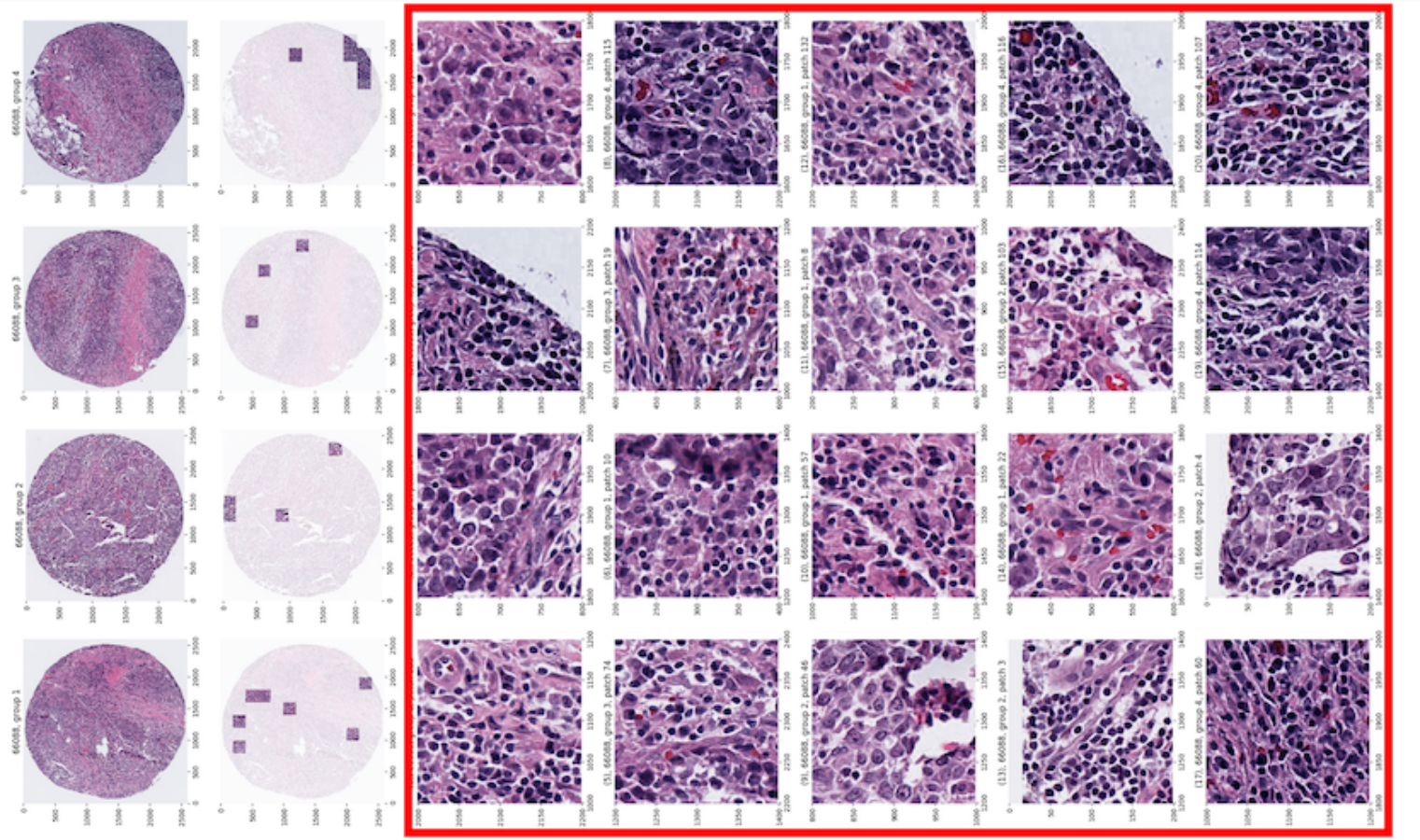
Representative patches are selected by projecting each patch's features onto a loadings vector then picking the top patches



Patient's 4 cores

Representative patch view

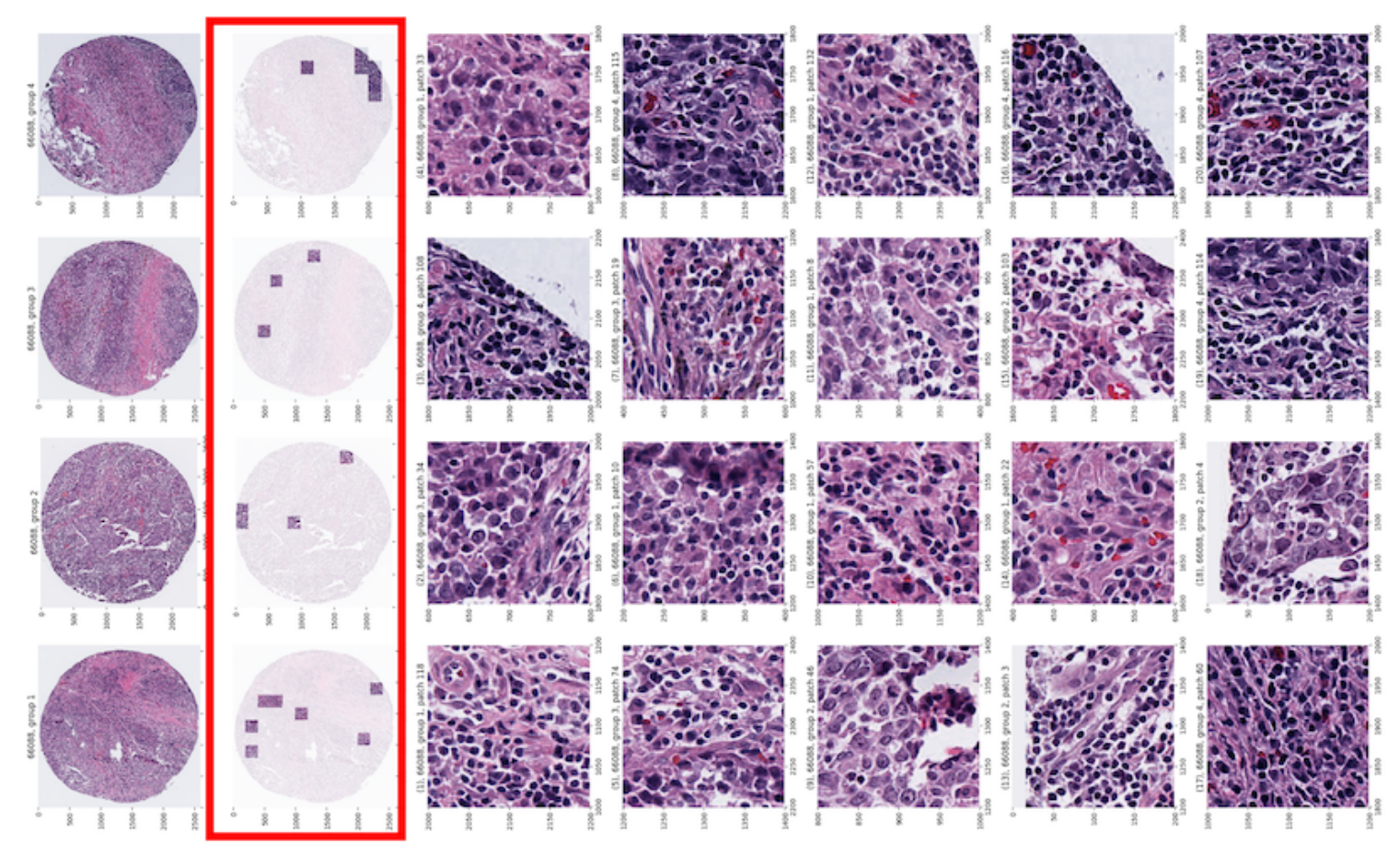
Representative patches are selected by projecting each patch's features onto a loadings vector then picking the top patches



20 representative patches for given mode of variation
e.g. negative end of first joint component

Representative patch view

Representative patches are selected by projecting each patch's features onto a loadings vector then picking the top patches

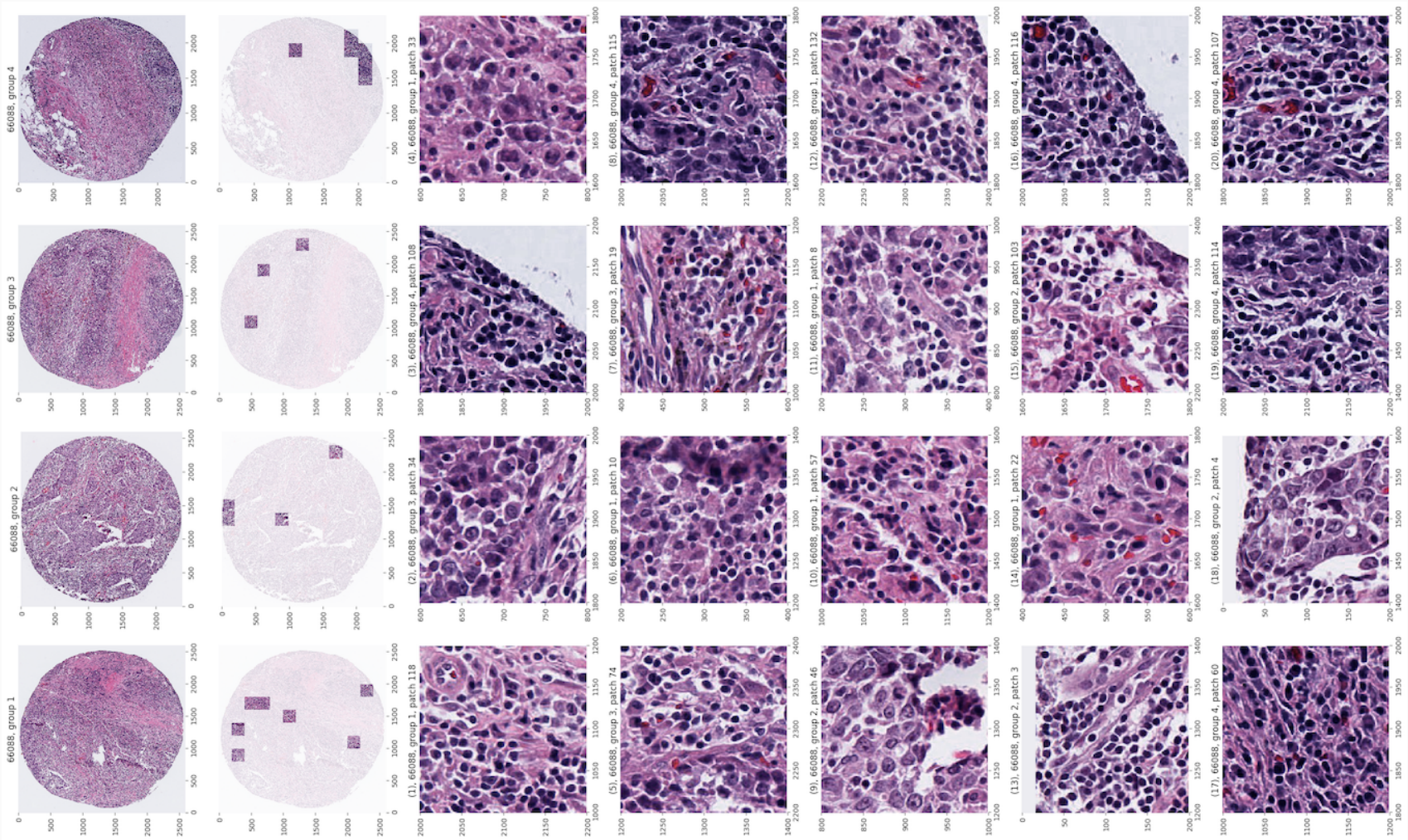


Locations of representative patches

Pathology review

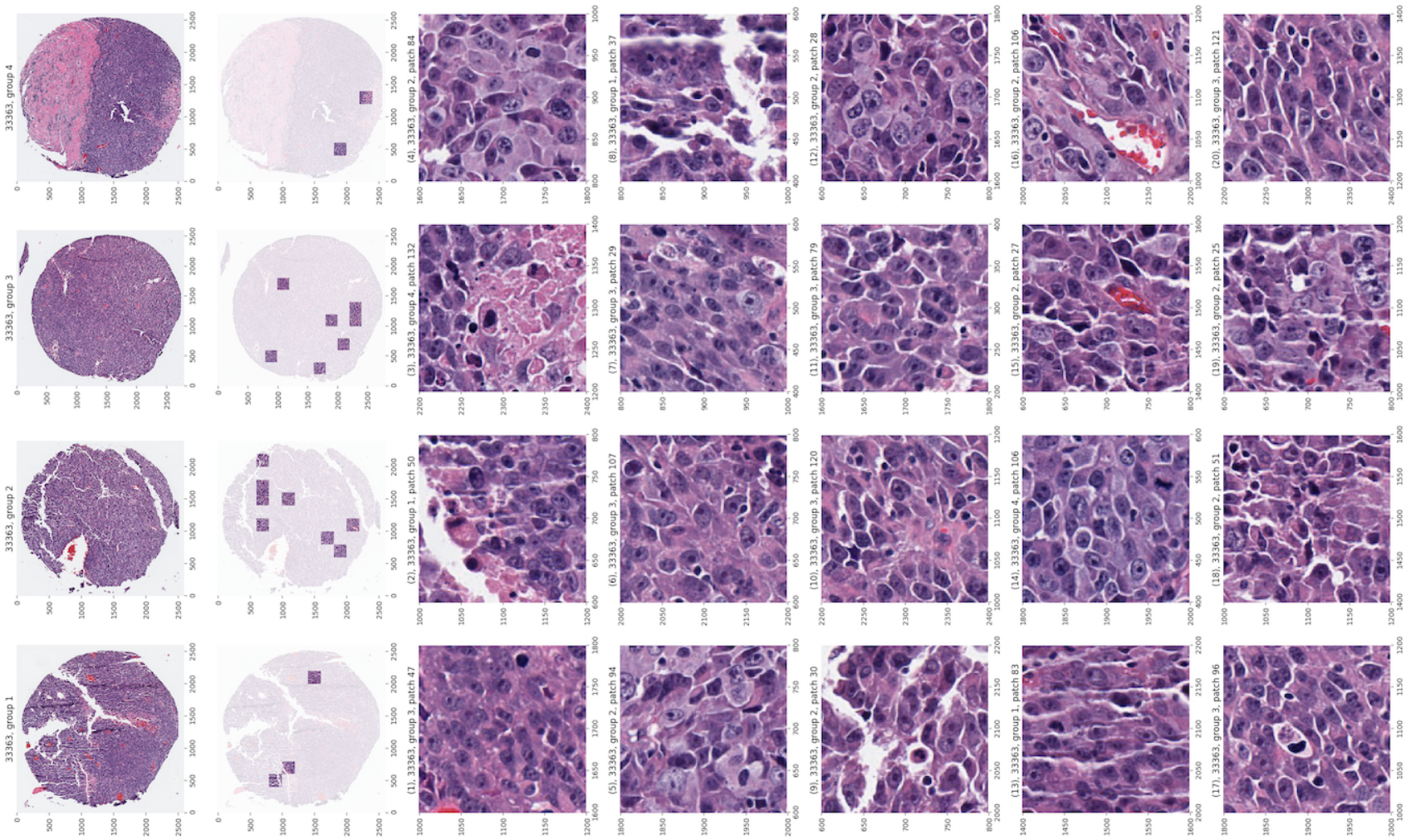
component	end	homogeneous	tumor cellularity	tubule formation	nuclear grade	adipocytic stroma	collagenous stroma	lymphocytes	necrosis
1	positive	no	low	yes	1, 2	yes	yes	no	no
	negative	yes	high	no	3	no	limited	yes	yes
2	positive	no	variable	yes	3	focal	yes	few	no
	negative	yes	moderate/high	yes	2	focal	yes	no	no
3	positive	no	variable	yes	3	yes	limited	yes	no
	negative	yes	moderate/high	no	3	no	yes	no	no

First joint component, negative end



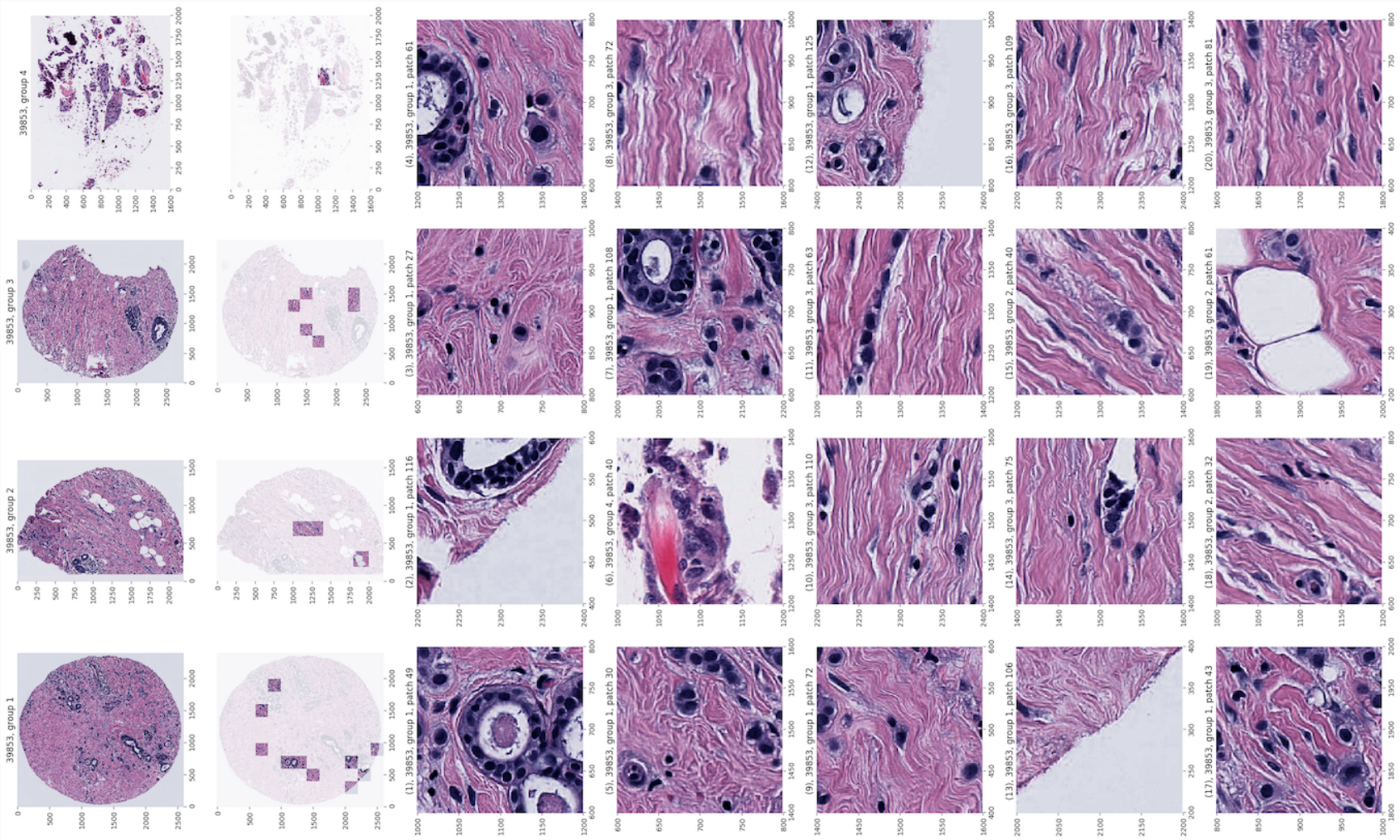
Tumor infiltrating lymphocytes

First joint component, negative end



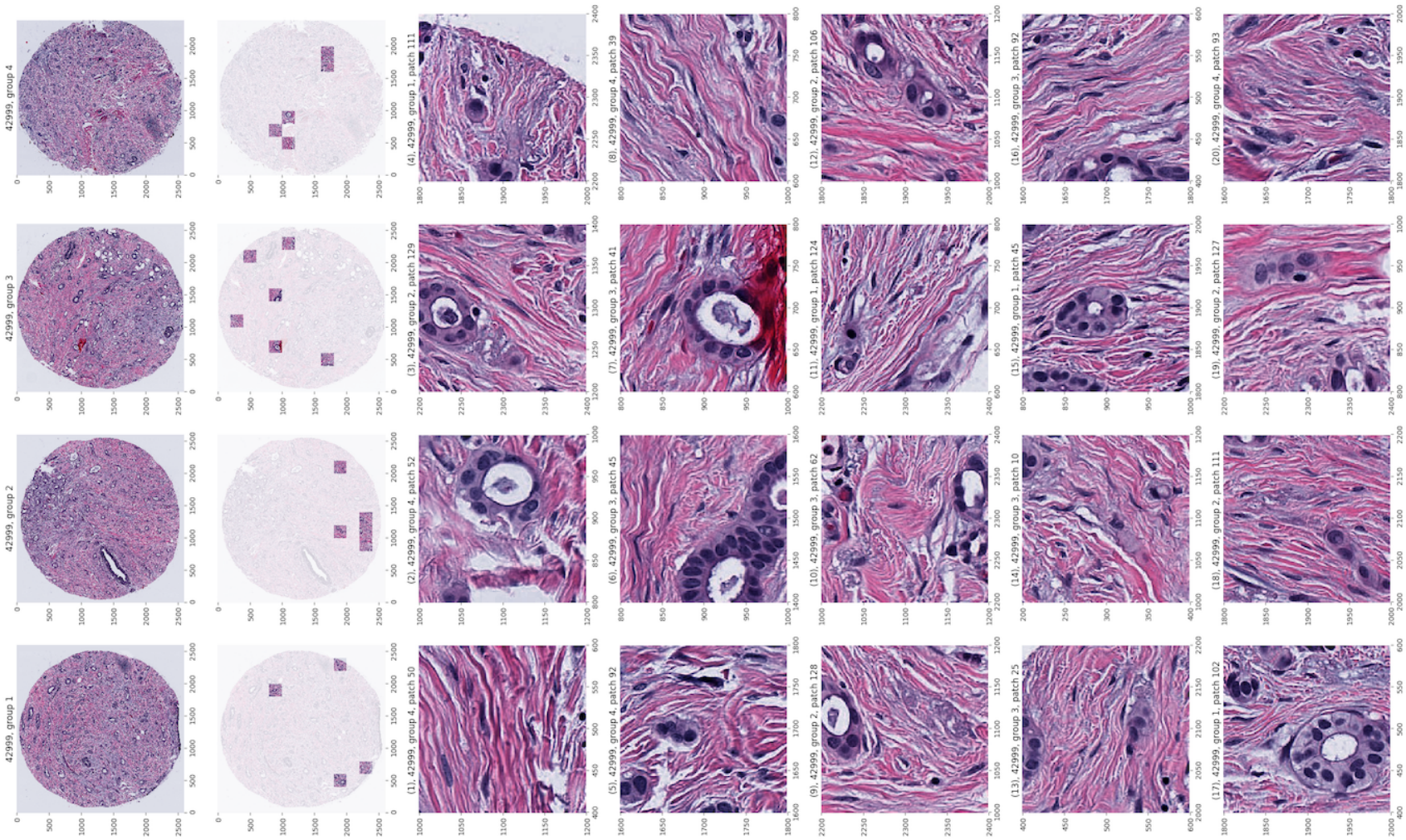
High nuclear grade tumor cells

First joint component, positive end



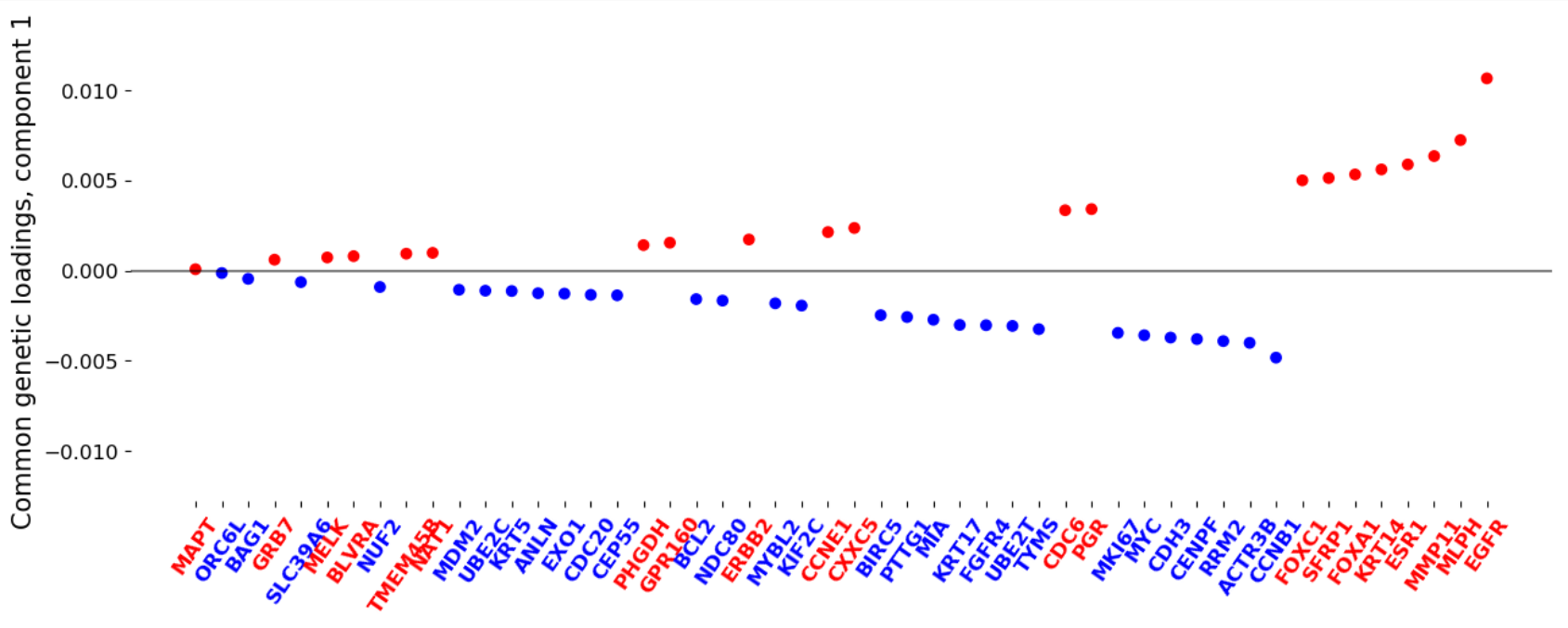
Mostly normal breast structure e.g. ducts, collagenous stroma

First joint component, positive end



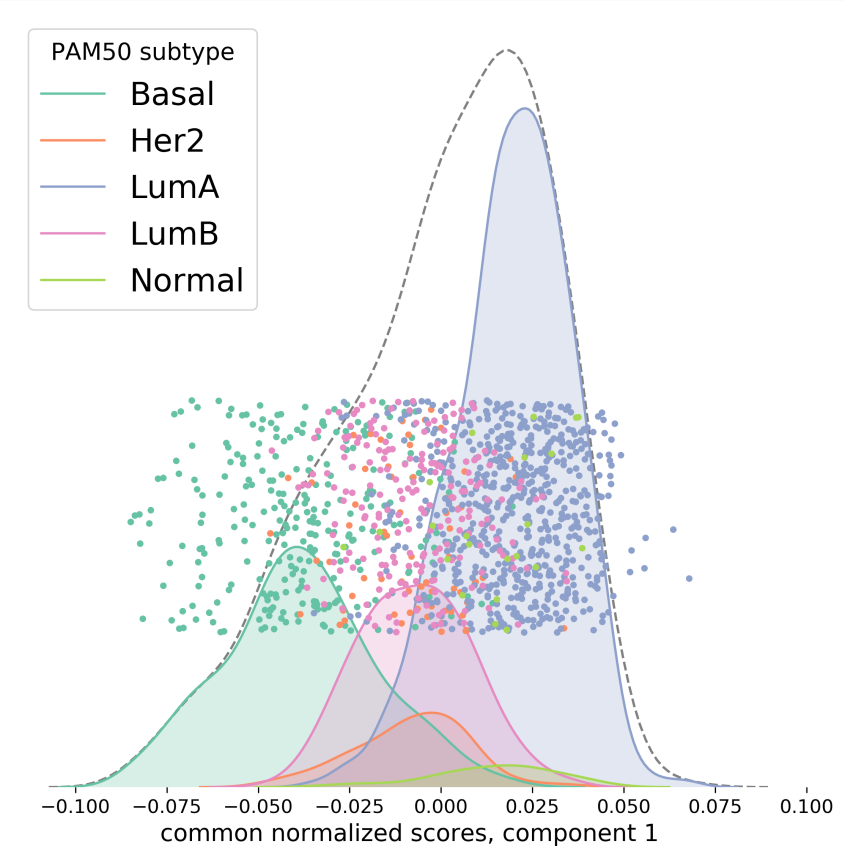
Mostly normal breast structure e.g. ducts, collagenous stroma

First joint component genetics loading vector



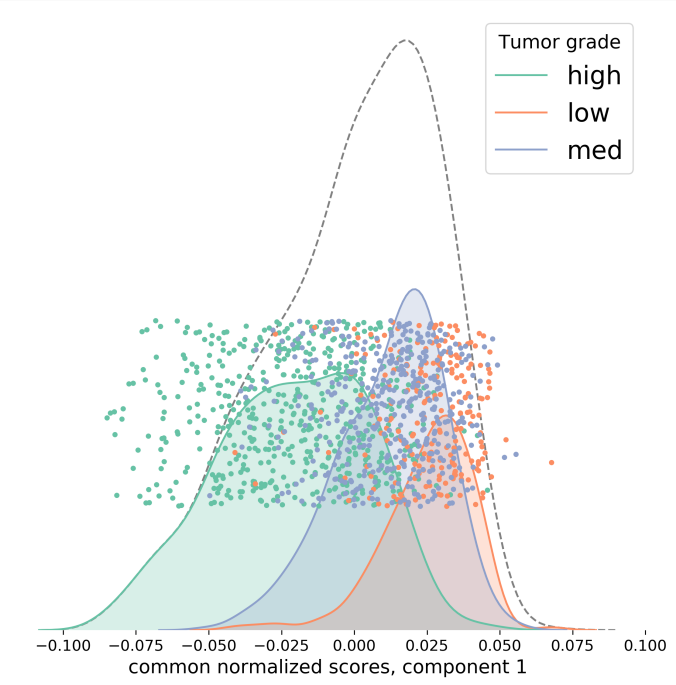
- Top negative genes associated with high tumor cell proliferation and tend to have low expression levels in normal breast tissue
- Top positive genes tend to have high expression levels in normal breast tissue

First joint component and PAM50 subtypes



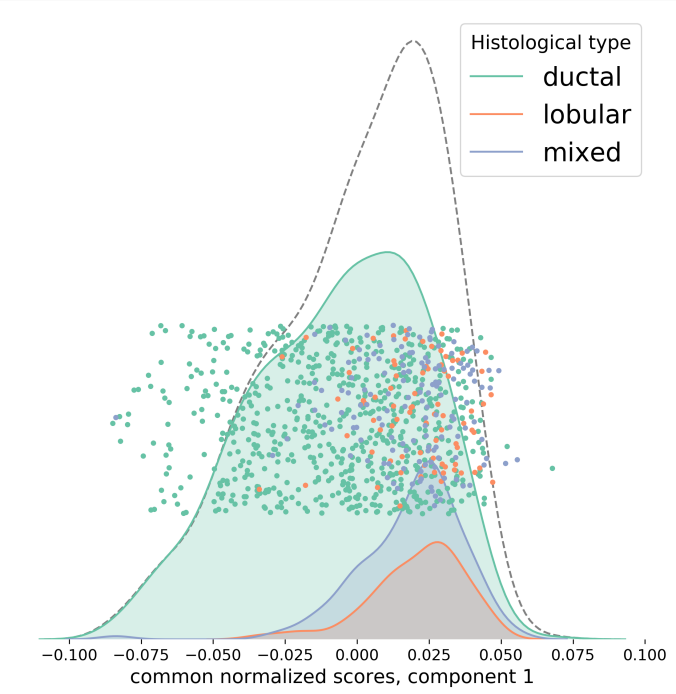
- Basal-like clusters on the negative end
  - Separated from other subtypes: Her2 (AUC = 0.87), Luminal A (0.98), Luminal B (0.89), Normal-like (0.97)
- Luminal B and Her2 (in the middle) are separated from Normal-like (0.82/0.88) and Luminal A (0.87/0.90)

First joint component associations with other covariates



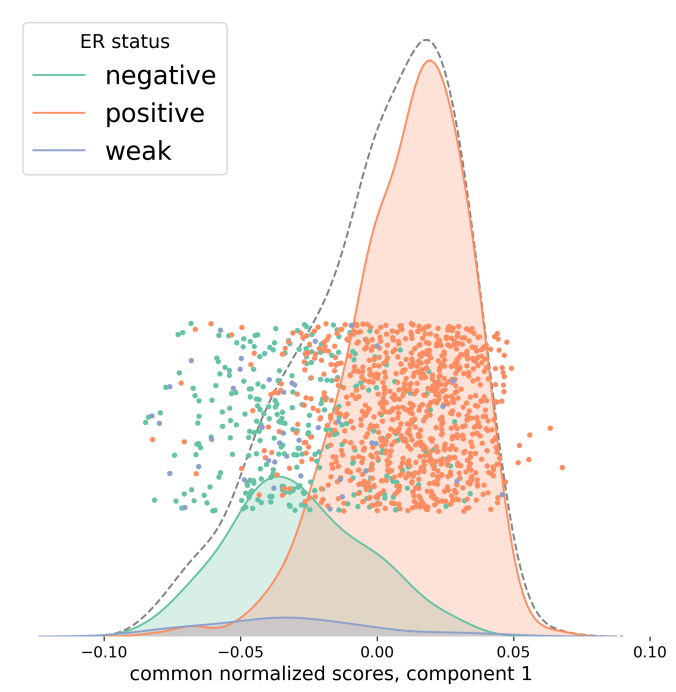
- Tumor grade
  - High-grade on negative end
  - High vs. low AUC = 0.94
- Histological type
  - Ductal on negative end
  - Ductal vs. lobular AUC = 0.79
- Estrogen receptor status
  - ER negative on negative end
  - Postive vs. negative AUC = 0.883
- Strong, negative correlation with proliferation score
- Risk of reoccurrence PT
  - ROR-PT high on negative end
  - High vs. low AUC = 0.999

First joint component associations with other covariates



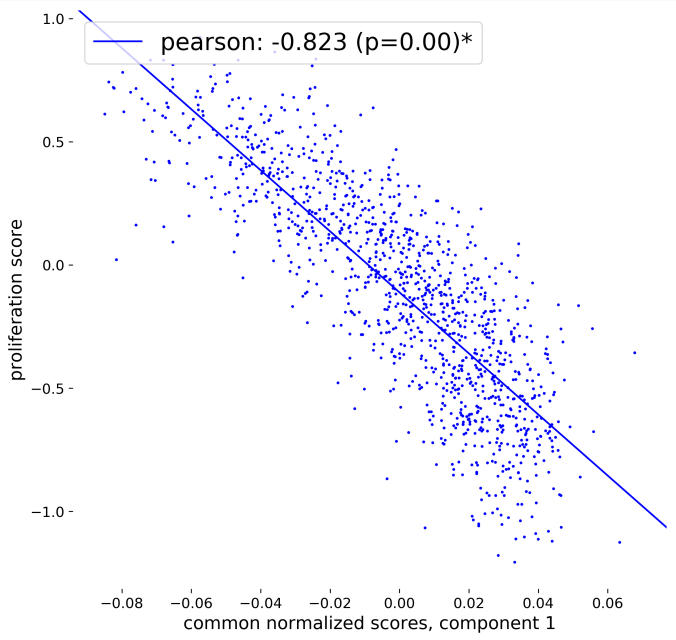
- Tumor grade
  - High-grade on negative end
  - High vs. low AUC = 0.94
- Histological type
  - Ductal on negative end
  - Ductal vs. lobular AUC = 0.79
- Estrogen receptor status
  - ER negative on negative end
  - Postive vs. negative AUC = 0.883
- Strong, negative correlation with proliferation score
- Risk of reoccurrence PT
  - ROR-PT high on negative end
  - High vs. low AUC = 0.999

First joint component associations with other covariates



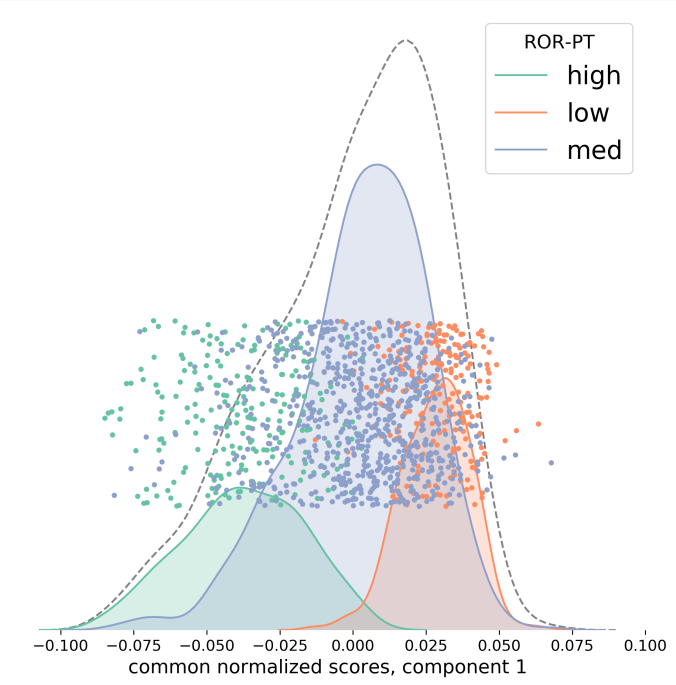
- Tumor grade
  - High-grade on negative end
  - High vs. low AUC = 0.94
- Histological type
  - Ductal on negative end
  - Ductal vs. lobular AUC = 0.79
- Estrogen receptor status
  - ER negative on negative end
  - Postive vs. negative AUC = 0.883
- Strong, negative correlation with proliferation score
- Risk of reoccurrence PT
  - ROR-PT high on negative end
  - High vs. low AUC = 0.999

First joint component associations with other covariates



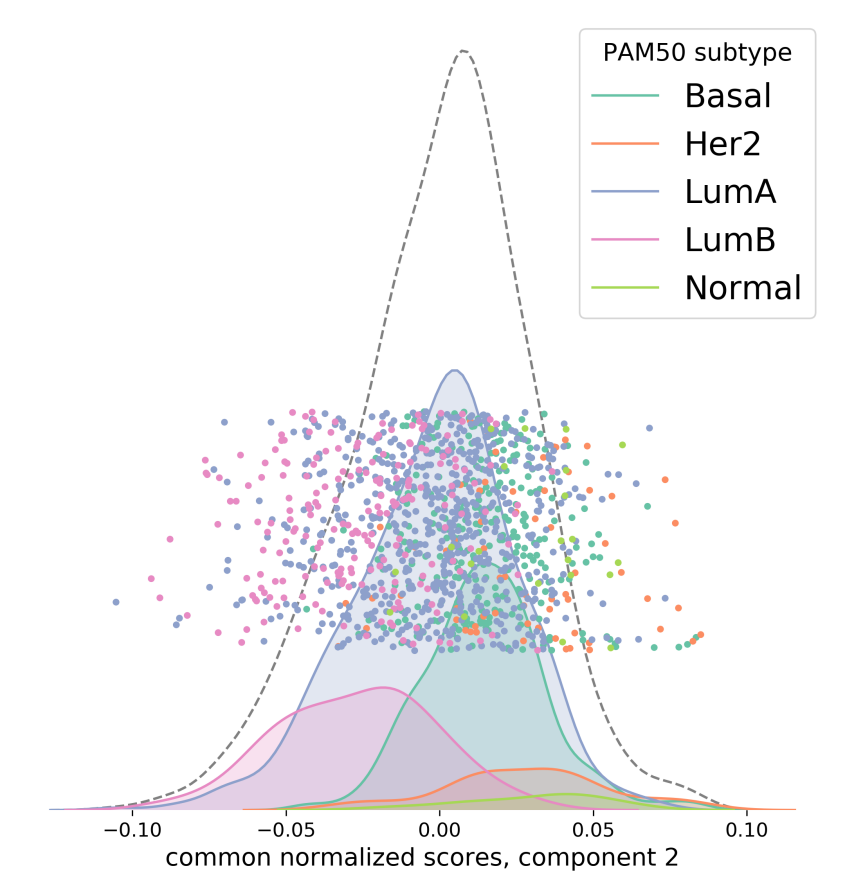
- Tumor grade
  - High-grade on negative end
  - High vs. low AUC = 0.94
- Histological type
  - Ductal on negative end
  - Ductal vs. lobular AUC = 0.79
- Estrogen receptor status
  - ER negative on negative end
  - Postive vs. negative AUC = 0.883
- Strong, negative correlation with proliferation score
- Risk of reoccurrence PT
  - ROR-PT high on negative end
  - High vs. low AUC = 0.999

First joint component associations with other covariates



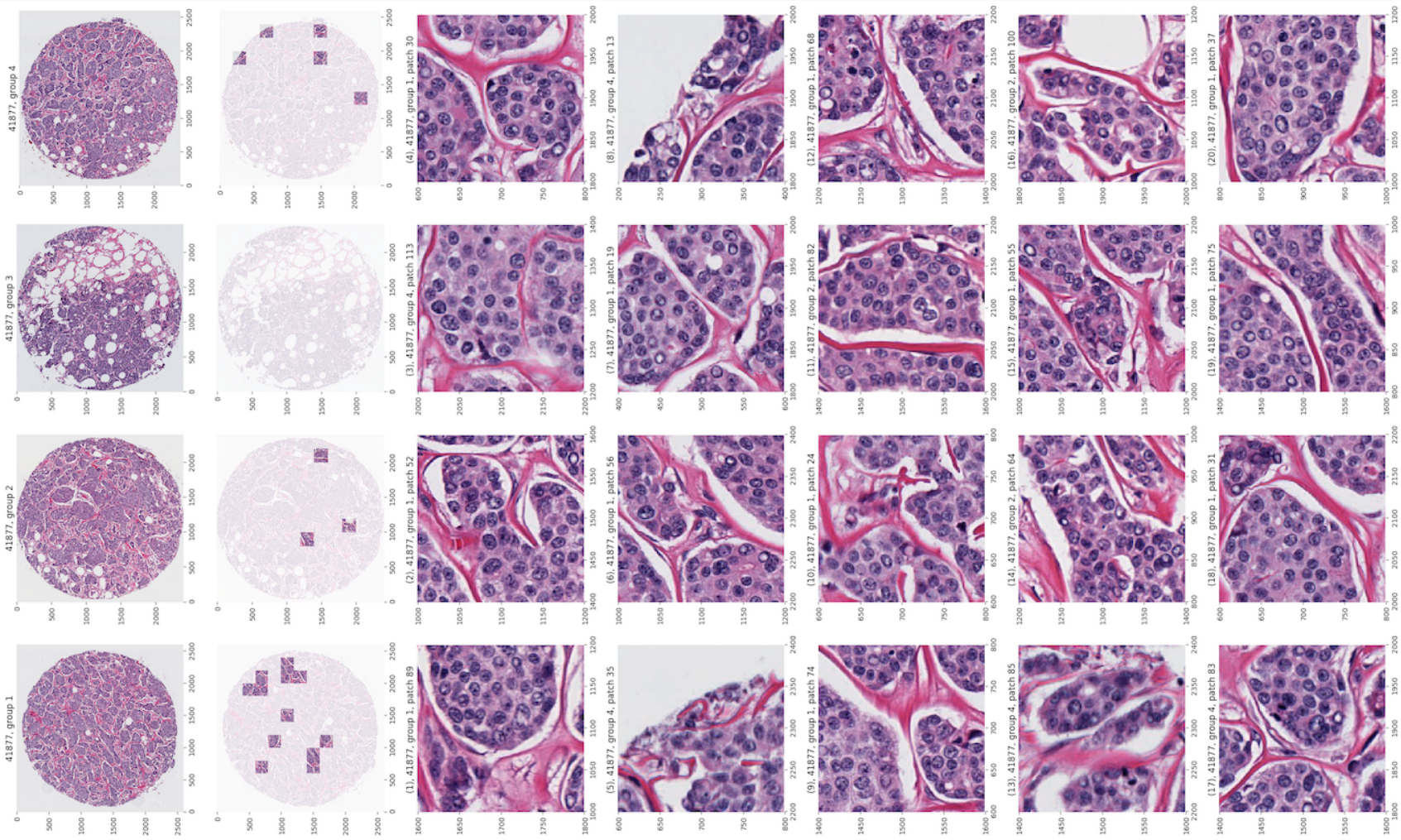
- Tumor grade
  - High-grade on negative end
  - High vs. low AUC = 0.94
- Histological type
  - Ductal on negative end
  - Ductal vs. lobular AUC = 0.79
- Estrogen receptor status
  - ER negative on negative end
  - Postive vs. negative AUC = 0.883
- Strong, negative correlation with proliferation score
- Risk of reoccurrence PT
  - ROR-PT high on negative end
  - High vs. low AUC = 0.999

Negative end of second joint component picks out Luminal B



Luminal B vs: Basal (AUC = 0.905), HER2 (0.933), Luminal A (0.760), Normal (0.950)

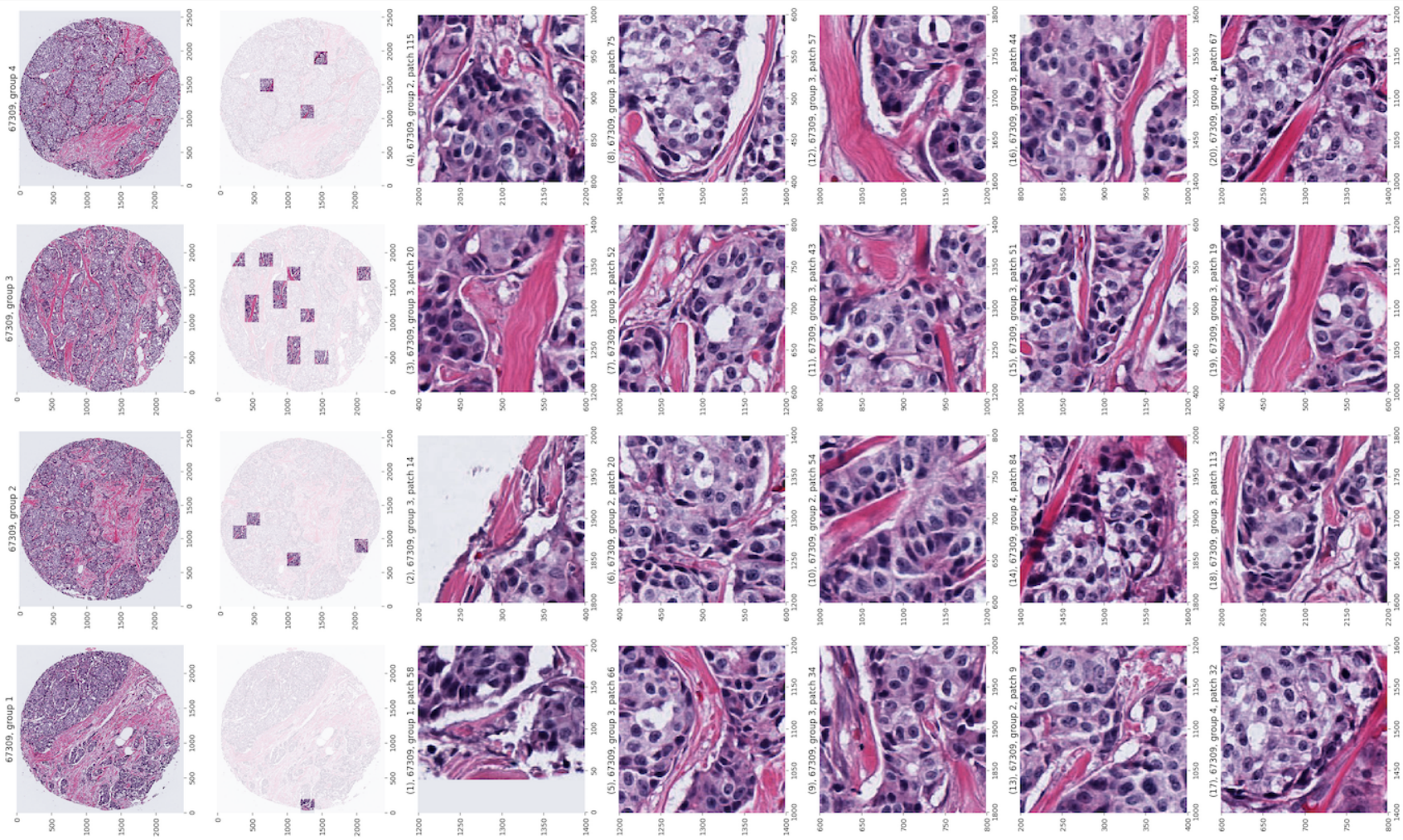
Second joint component, negative end morphology



Intratumoral channels of stromal cells that are surrounded by cancer cells

Currently being validated as morphological feature of Luminal B cancers

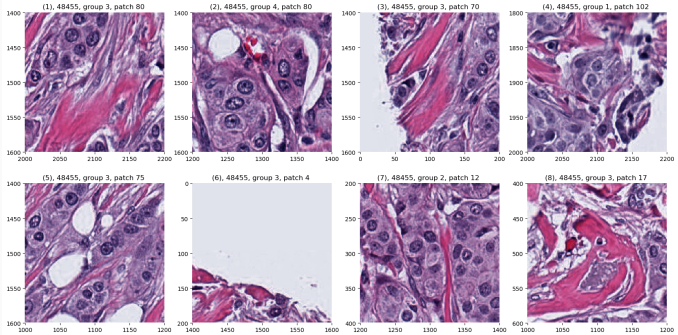
Second joint component, negative end morphology

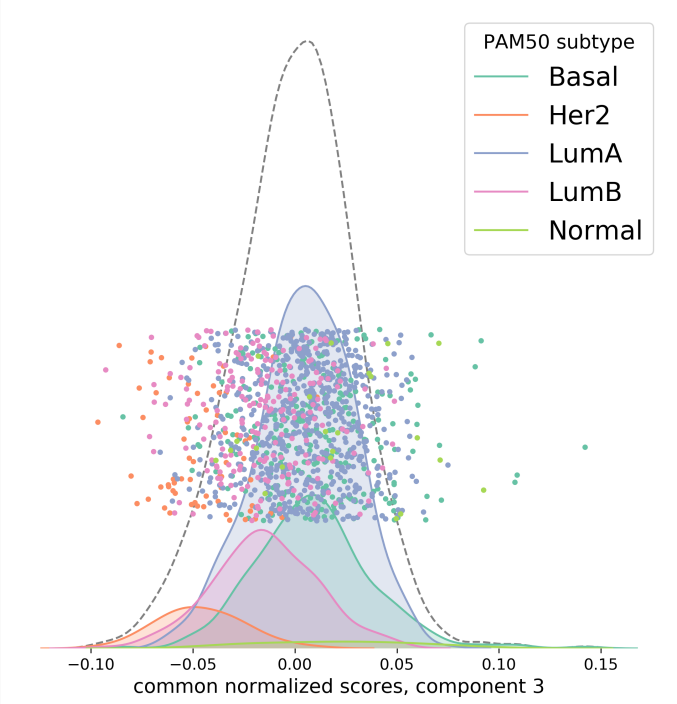


Intratumoral channels of stromal cells that are surrounded by cancer cells

Currently being validated as morphological feature of Luminal B cancers

Negative end of third joint component picks out molecular Her2



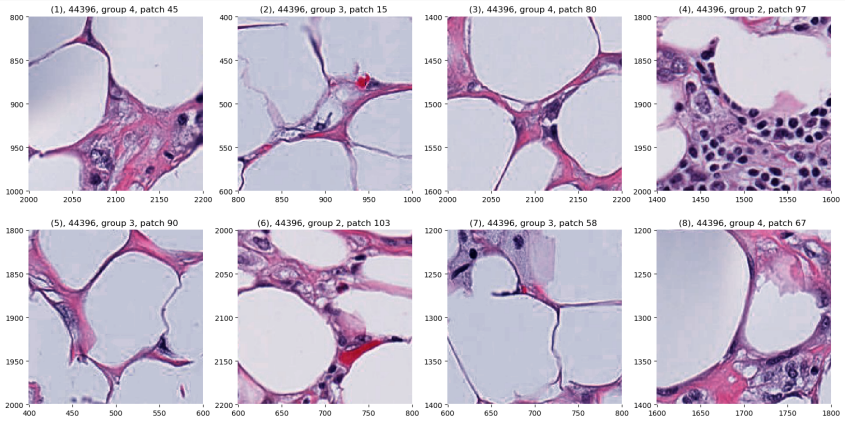


Molecular Her2 vs: Basal (AUC= 0.947), Luminal A (0.940), Luminal B (0.833), Normal (0.950)

Histology individual components

Variation related to tumor microenvironment*

- High fat content (component 1)
- Mucinious carcinoma (component 2)
- Degraded samples (component 3)

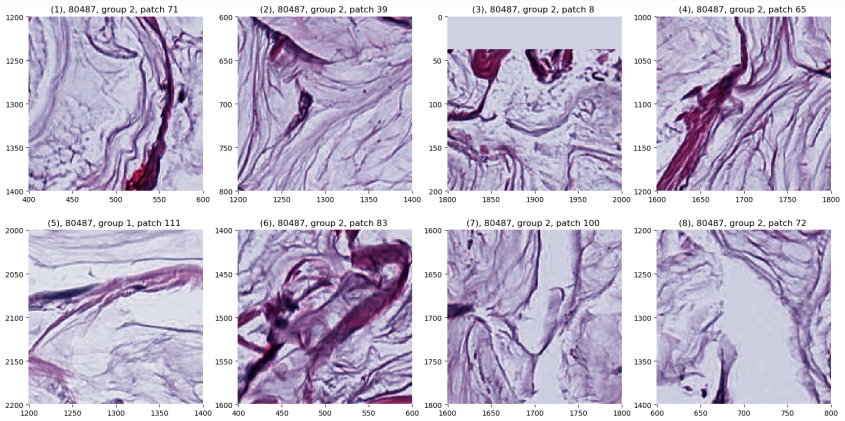


*PAM50 genes do not describe the tumor microenvironment (Perou et al., 2000)

Histology individual components

Variation related to tumor microenvironment*

- High fat content (component 1)
- Mucinious carcinoma (component 2)
- Degraded samples (component 3)

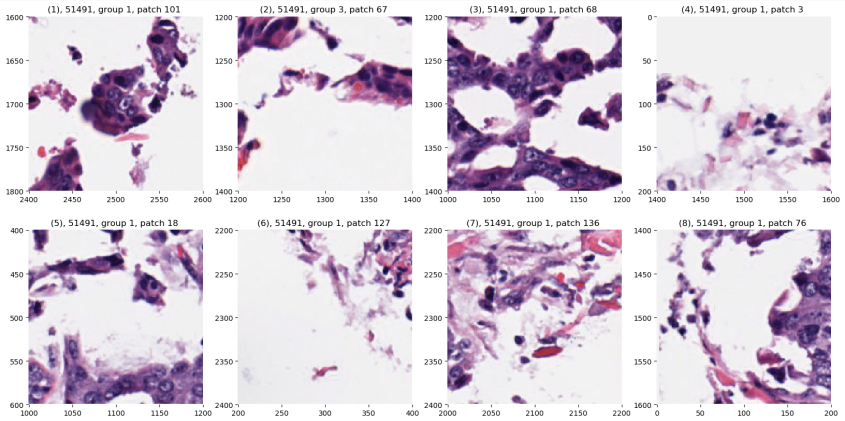


*PAM50 genes do not describe the tumor microenvironment (Perou et al., 2000)

Histology individual components

Variation related to tumor microenvironment*

- High fat content (component 1)
- Mucinious carcinoma (component 2)
- Degraded samples (component 3)

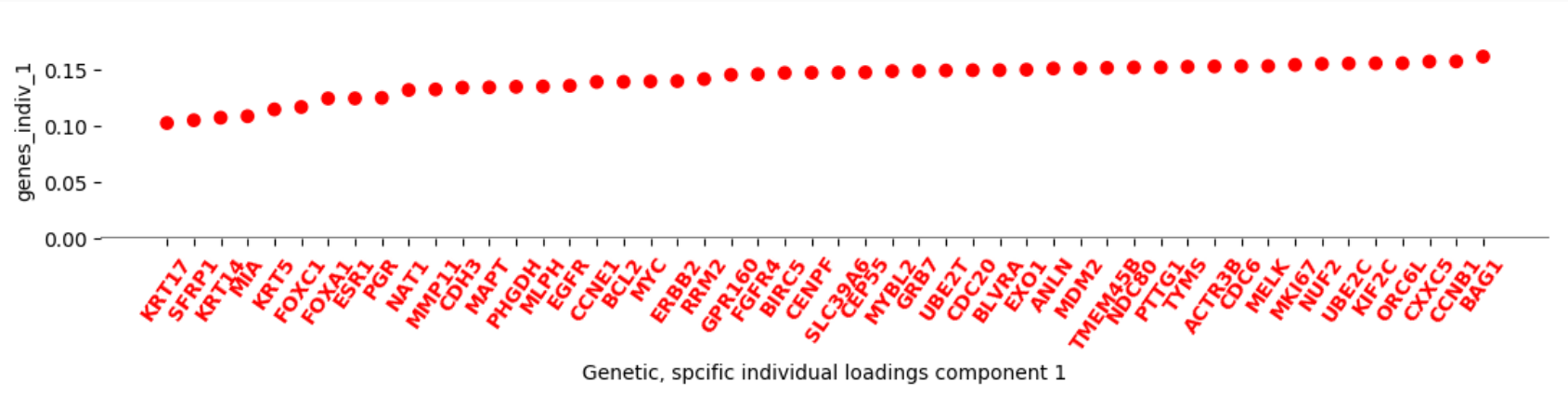


*PAM50 genes do not describe the tumor microenvironment (Perou et al., 2000)

Genetic individual components

Technical variational as well as additional PAM50 subtype information

- Overall expression level (component 1)
- Luminal A vs. Normal (component 2)
  - Top left: estrogen signaling pathway
  - Middle: proliferation
  - Bottom right: normal myoepithelium and Basal-like



Genetic individual components

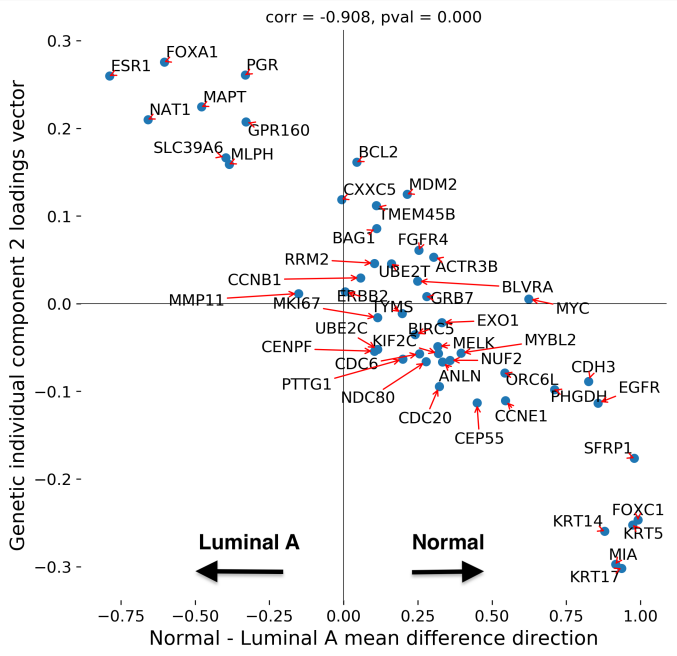
Technical variational as well as additional PAM50 subtype information

- Overall expression level (component 1)
- Luminal A vs. Normal (component 2)
  - Top left: estrogen signaling pathway
  - Middle: proliferation
  - Bottom right: normal myoepithelium and Basal-like

Genetic individual components

Technical variational as well as additional PAM50 subtype information

- Overall expression level (component 1)
- Luminal A vs. Normal (component 2)
  - Top left: estrogen signaling pathway
  - Middle: proliferation
  - Bottom right: normal myoepithelium and Basal-like



Paper available on arxiv

Carmichael et al. (2019): `https://arxiv.org/pdf/1912.00434.pdf`

Future directions in statistical data integration

- Multi-view clustering
  - Each view has a set of clusters
  - Model connections between clusters in different views
- Estimation of partially shared structures for $B > 2$ blocks
- Modeling multiple, complex data objects

Future directions in integrative cancer research

- Validate Luminal B morphology
  - Survival
  - Independent replication
- AJIVE/cluster analysis with other modalities (protein, copy number, multiplex immunofluorescence, etc)

Questions?

References

Allott, E. H., Geradts, J., Cohen, S. M., Khoury, T., Zirpoli, G. R., Bshara, W., Davis, W., Omilian, A., Nair, P., Ondracek, R. P., et al. (2018). Frequency of breast cancer subtypes among african american women in the amber consortium. *Breast Cancer Research*, 20(1):12.

Asendorf, N. A. (2015). *Informative data fusion: Beyond canonical correlation analysis*. PhD thesis, The University of Michigan.

Bjorck, A. and Golub, G. H. (1973). Numerical methods for computing angles between linear subspaces. *Mathematics of computation*, 27(123):579–594.

References II

Carmichael, I., Calhoun, B. C., Hoadley, K. A., Troester, M., Geradts, J., Couture, H. D., Olsson, L., Perou, C. M., Niethammer, M., Hannig, J., and Marron, J. (2019). Joint and individual analysis of breast cancer histologic images and genomic covariates. *arXiv preprint arXiv:1912.00434*.

Draper, B., Kirby, M., Marks, J., Marrinan, T., and Peterson, C. (2014). A flag representation for finite collections of subspaces of mixed dimensions. *Linear Algebra and its Applications*, 451:15–32.

Edelman, A., Arias, T. A., and Smith, S. T. (1998). The geometry of algorithms with orthogonality constraints. *SIAM journal on Matrix Analysis and Applications*, 20(2):303–353.

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

References III

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

Johnstone, I. M. (2008). Multivariate analysis and jacobi ensembles: Largest eigenvalue, tracy–widom limits and rates of convergence. *Annals of statistics*, 36(6):2638.

Kettenring, J. R. (1971). Canonical analysis of several sets of variables. *Biometrika*, 58(3):433–451.

Lock, E. F., Hoadley, K. A., Marron, J. S., and Nobel, A. B. (2013). Joint and individual variation explained (jive) for integrated analysis of multiple data types. *The annals of applied statistics*, 7(1):523.

Macenko, M., Niethammer, M., Marron, J. S., Borland, D., Woosley, J. T., Guan, X., Schmitt, C., and Thomas, N. E. (2009). A method for normalizing histology slides for quantitative analysis. In *2009 IEEE International Symposium on Biomedical Imaging: From Nano to Macro*, pages 1107–1110. IEEE.

References IV

Nielsen, A. A. (2002). Multiset canonical correlations analysis and multispectral, truly multitemporal remote sensing data. *IEEE transactions on image processing*, 11(3):293–305.

Otsu, N. (1979). A threshold selection method from gray-level histograms. *IEEE transactions on systems, man, and cybernetics*, 9(1):62–66.

Perou, C. M., Sørlie, T., Eisen, M. B., Van De Rijn, M., Jeffrey, S. S., Rees, C. A., Pollack, J. R., Ross, D. T., Johnsen, H., Akslen, L. A., et al. (2000). Molecular portraits of human breast tumours. *nature*, 406(6797):747.

Troester, M. A., Sun, X., Allott, E. H., Geradts, J., Cohen, S. M., Tse, C.-K., Kirk, E. L., Thorne, L. B., Mathews, M., Li, Y., et al. (2017). Racial differences in pam50 subtypes in the carolina breast cancer study. *JNCI: Journal of the National Cancer Institute*, 110(2):176–182.

References V

Zack, G., Rogers, W., and Latt, S. (1977). Automatic measurement of sister chromatid exchange frequency. *Journal of Histochemistry & Cytochemistry*, 25(7):741–753.