December 11, 2009
CS229 Final Project

Andrew Beck
Winnie Cheng
Franklin Wu

Integrated Analysis of Breast Cancer Gene Expression, Morphology, and Patient Survival

Acknowledgements: We received support on this work from Professor Daphne Koller, Stephen Gould, and Daniela Witten. We thank Drs Rob West, Matt van de Rijn, and Marc van de Vijver for providing access to the NKI breast cancer images.

I. Introduction

Since the mid-19th century, scientists have recognized that the process of carcinogenesis produces characteristic morphological changes in cancer cells. To this day, careful morphological analysis of microscopic images remains the single most important diagnostic tool used to classify cancer in routine clinical practice. In the past decade, the molecular understanding of cancer has increased tremendously, driven in large part by gene expression profiling experiments, which quantify the expression of virtually all known transcripts in the human genome[2]. Despite widespread acceptance of the clinically and biologically useful information contained in both morphological features and gene expression profiles, there has to date been no effort to systematically integrate gene expression changes in cancer with a computational assessment of morphologic changes.

In this paper, we developed tools to permit an integrative analysis of gene expression and morphology, thus allowing the identification of gene sets that are associated with morphologic phenotypes. To accomplish this, we first developed image analysis tools to permit the measurement of tissue and cellular structures in breast cancer microscopic images. We then applied sparse canonical correlation analysis (CCA) to identify relatively small sets of genes that show maximal correlation with relatively small sets of image features. Lastly, we assessed the association of a score integrating data from image and gene feature with patient survival. This analysis permits the first steps towards a molecular understanding of the morphologic variability in breast cancer and suggest pathways through which molecular changes drive tumor morphologic changes, which ultimately impact patient outcome.

II. Materials and Methods

IIa. Breast Cancer Dataset

We are using the Netherlands Cancer Institute (NKI) dataset, which contains gene expression profiling data (~24K probes per sample) and survival data from 295 breast cancer patients [13]. This data is publically available (http://microarray-pubs.stanford.edu/wound_NKI/explore.html). In addition to the gene expression microarray and clinical data, we have obtained 1282 (~5 images per patient) microscopic images of H&E stained breast cancer histologic sections from 249 of the NKI patients.

IIb. Epithelial/Stromal and Nuclear/Cytoplasmic Segmentation

Epithelial cancers are composed of two tissue compartments: the malignant epithelium (which contains the breast cancer cells) and the cancer stroma, which contains a milieu of "supportive" cell types (including fibroblasts, endothelial cells, and smooth muscle cells)[1]. The epithelium itself is comprised of individual epithelial cells, each of which contains a nucleus (with the cell's genetic material) as well as the cytoplasm. All tissue and cellular compartments of cancer undergo changes, but these changes are often most profound in the epithelial nuclei, as the nuclei contain DNA, and cancer is primarily a disease of genetic abnormalities (acquired and/or inherited). Therefore, an important first stage of our project was to develop a classifier to separate epithelium from stroma, and then to develop a second classifier to work within epithelial cells to separate nuclei from cytoplasm.

To develop a classifier to distinguish tissue regions (epithelium and stroma) and cellular regions (nucleus and cytoplasm), we adapted a standard pixelwise CRF model, as described as the baseline method in Gould et al. [5]. Given an image $\mathcal{I}$ composed of pixels p, region class labels S, and pixel appearance vector α_p, this model took the form of a unified energy function:
$E(S|\mathcal{I}) = \sum_p \psi_p(S_p, S_q;\ \alpha_p, \alpha_q) + \ \theta \sum_{pq} \psi_{pq}(S_p, S_q; \alpha_p \alpha_q)$, where ψ_p is a multi-class logistic over boosted appearance features, ψ_{pq} is a boundary penalty to encourage adjacent pixels to take the same value, and θ are model parameters. The model was implemented using functions provided in the STAIR vision library [6]. For each pixel in the training set, the local pixel appearance vector α_p was constructed with both raw image features of the pixel and of surrounding pixels. The raw features measured for the pixel and surrounding pixels are the 17-dimensional color and texture features determined in Shotton et al.[11]. In training the classifier, the raw image features are augmented by region predictions made by a one-vs-all boosted classifier trained on the raw features of a given pixel's neighboring pixels. Inference of region label is performed with a two-sage hill climbing approach to minimize the energy function.

To obtain ground truth labels, a pathologist (AB) manually labeled a subset of images (65 images were labeled for the epithelial/stromal classifier). After epithelial/stromal segmentation and prior to nuclear segmentation, the epithelial images were further processed to remove inflammatory cells, by implementing the strel and imopen functions in Matlab, which morphologically open the image based on a structure element. A threshold is then applied to this opened image and the pixel locations of the thresholded image are mapped to the original image of epithelium, with most inflammatory cells (which tend to be smaller and less cohesive than the epithelial cells) filtered out. Following this step, ground truth labels were obtained for nuclei and cytoplasm from 7 images and a nuclear/cytoplasmic classifier was trained and applied to the segmented epithelial regions. Following identification of nuclear regions (as described above), we performed segmentation of individual nuclei using an adaptive Otsu approach [10].

IIc. Image feature extraction

To measure epithelial and stromal image features, we have developed an epithelial nuclear feature measurement pipeline and a stromal feature measurement pipeline using the open source CellProfiler software (http://www.cellprofiler.org/), which implements routines in Matlab. We measured a variety of descriptors of image texture and heterogeneity as well as nuclear size, shape, intensity, texture, and crowding. Following development of the image feature extractions pipelines, we used the analytic pipelines to measure 120 epithelial features and 120 stromal features from our set of 1282 images obtained from 249 patients. For features measured from individual nuclei, we summarized the measurements per image by taking the mean. This process produced a 249 patient by 240 image feature data matrix.

IId. Identification of sets of genes that show maximal correlation with sparse sets of morphologic features

A primary goal of this project was to identify gene expression patterns that show maximal correlation with morphological phenotypes. A natural unsupervised learning algorithm for achieving this goal is canonical correlation analysis (CCA). The CCA procedure was initially proposed by Hotelling in 1936 [8] and can be applied to this project as follows: Let X_1 denote the n x p_1 data matrix of image features with n samples and p_1 image features, and let X_2 denote the n x p_2 data matrix of gene expression features with n samples and p_2 gene expression features. We first standardize the features in each dataset to have mean zero and standard deviation of 1. The CCA procedure will produce a p_1 dimensional weight vector w_1 and a p_2 dimensional weight vector w_2 that maximize the CCA criterion:

$$maximize_{w_1 w_2} w_1^T X_1^T X_2 w_2 \ subject\ to\ w_1^T X_1^T X_1 w_1 = \ w_2^T X_2^T X_2 w_2 = 1$$

Standard CCA is difficult to implement with high dimensional data, where the numbers of features measured (p) far exceeds the number of samples (n). Given the large number of image and gene features measured in our study, we chose to implement a penalized form of CCA to produce sparse linear combinations of genes highly correlated with sparse linear combinations of image features. We implemented penalized CCA using the technique of Witten et al.[14, 15] with the package "PMD" in the R language for statistical computing (http://cran.r-project.org/). The sparse CCA criterion is as follows:

$$maximize_{w_1 w_2} w_1^T X_1^T X_2 w_2 \ subject\ to\ \left||w_1|\right|_1 \ \le \ c_1\sqrt{p_1}, \left||w_2|\right|_1 \le \ c_2\sqrt{p_2}, \left||w_1|\right|^2 \le \ 1, \left||w_2|\right|^2 \le \ 1$$

The L_1 (Lasso [12]) penalty applied to the weight vectors (w_1 and w_2) ($\left||w_1|\right|_1 \ \le \ c_1\sqrt{p_1}, \left||w_2|\right|_1 \ \le \ c_2\sqrt{p_2}$) has the effect of driving most weights to zero for small values of the tuning parameter c, resulting in sparse and more easily interpretable sets of image and gene features with non-zero weights. It is noted that the sparse CCA criterion is biconvex, and with w_1 fixed is convex in w_2, and with w_2 fixed is convex in w_1, and thus can be solved by an iterative algorithm, in which w_2 is initialized to have an L_2 norm 1 and the following 2 steps are repeated until convergence [15]:

$$1)\ w_1 \coloneqq \arg max_{w_1}\ w_1^T X_1^T X_2 w_2 \ subject\ to\ \left||w_1|\right|_1 \ \le \ c_1\sqrt{p_1}, \left||w_1|\right|^2 \ \le \ 1$$

$$2)\ w_2 \coloneqq \arg max_{w_2}\ w_1^T X_1^T X_2 w_2 \ subject\ to\ \left||w_2|\right|_1 \ \le \ c_2\sqrt{p_2}, \left||w_2|\right|^2 \ \le \ 1$$

To select the penalty terms c_1 and c_2, we evaluated ten sets of c_1 and c_2 values evenly spaced within a range from 1 to $\sqrt{p_1}$ and $\sqrt{p_2}$, respectively. During each of these ten trials, sparse CCA is performed with a given set of tuning parameters c_1 and c_2 to generate a pair of weight vectors (w_1,w_2) and the value of $Cor(X_1 w_1, X_2 w_2)$ is recorded. The significance of the $Cor(X_1 w_1, X_2 w_2)$ for a given set of c_1 and c_2 is estimated by permutations. We performed 200 permutations. During each, the samples in X_1 and X_2 are randomly permuted to obtain matrices $X_1^* \ and\ X_2^*$. Sparse CCA was then performed on the permuted data to generate $w_1^* \ and\ w_2^*$ and $Cor(X_1^* w_1^*, X_2^* w_2^*)$. The sparse CCA algorithm then implements Fisher's transformation to convert the real correlation and set of 200 permuted correlations into random variables that were approximately normally distributed, which we denote as fc and fc*. A p value is then computed for each pair of w_1,w_2 resulting from a given value of the tuning parameter c, as the fraction of the 200 fc* that exceed fc. We ultimately selected tuning parameters c_1 and c_2 that gave a high $Cor(X_1 w_1, X_2 w_2)$, low p value, and an easily interpretable number of non-zero weights. To compute a second approximately orthogonal set of feature weights $(\mathrm{w}_1, \mathrm{w}_2)^{(2)}$, it is noted that the the sparse CCA criterion can be re-formulated as maximizing: $w_1^T Z_1 w_2 \ subject\ to\ \left||w_1|\right|_1 \ \le \ c_1\sqrt{p_1}, \left||w_2|\right|_1 \le \ c_2\sqrt{p_2}, \left||w_1|\right|^2 \le \ 1, \left||w_2|\right|^2 \le \ 1$, where $Z_1 = \ X_1^T X_2$. Z_2 is then computed as $Z_1 - \ w_1 Z_1 w_2 w_1 w_2^T$. The optimization problem stated above is then repeated with Z_2 replacing Z_1. We chose to compute only two factors for this project; however, the above procedure may be repeated to obtain as many sets of weight vectors as desired[7].

IId. Analysis of Functional Gene Set Enrichment

Following identification of genes with non-zero values in the gene weight vector for factors 1 and 2, we identified biological pathways significantly enriched in the sets of genes positively and negatively associated with the linear combination of image features by uploading our gene sets to the DAVID set of bioinformatics resources[9] and identifying clusters of annotation terms with highly significant enrichment in our gene sets.

IIe. Survival analysis

To assess the association of the image gene feature combinations with survival, we computed a single score for each patient that we called the Image-Gene Score (IGS), which was simply the sum of the linear combinations of image features and gene features for the two factors. For a given sample (i), where $X_{(i)}$ denotes the i^th row of matrix X and $w^{(k)}$ denotes the weight vector for the k^th factor :

$$IGS^{(i)} = \ X_{1(i)} w_1^{(1)} + \ X_{2(i)} w_2^{(1)} + \ X_{1(i)} w_1^{(2)} + \ X_{2(i)} w_2^{(2)}$$

To assess, the association of the IGS score with survival, we discretized the scores into three quantiles, plotted Kaplan Meier survival curves, and performed the log-rank test to assess statistical significance. To assess the additive effect of the IGS for predicting survival in a multivariate model containing histologic grade, we fitted a Cox proportional hazards model[3] and computed p values to assess the probability that the IGS and grade were each contributing to the predictive accuracy of the model.

III. Results

IIIa. Development of Epithelial/Stromal Classifier and Nuclear/Cytoplasmic Classifier

For the purposes of quantitatively assessing accuracy, we trained the classifier on 2/3 of our labeled images and tested it on the remaining 1/3. The epithelial/stromal classifier achieved an overall accuracy of 84% with recall/precision of 0.93/0.95 on background, 0.80/0.79 on epithelium, and 0.82/0.81 on stroma. An example test image is provided in Figure 1. We have not yet labeled enough nuclear/cytoplasmic regions on images to permit a quantitative evaluation of the nuclear/cytoplasmic classifier on new images, but it achieved an overall accuracy of ~92% on the training images and qualitatively performed well on the new images in the dataset. We also have not yet quantitatively assessed the accuracy of the nuclear region segmentation or nuclear measurements; however, they appear to work well in most cases.

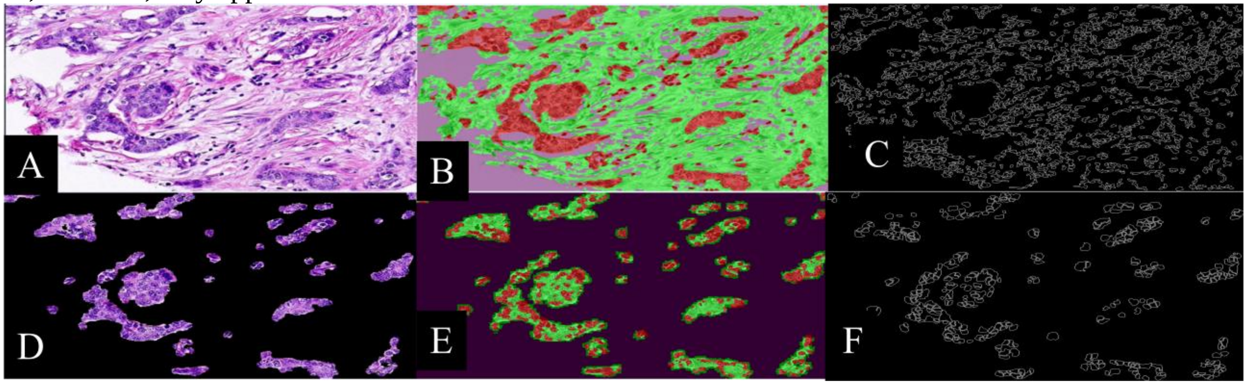


Figure 1. Segmented breast cancer image from test set. A. H&E stained microscopic image of breast cancer. B. Image with superimposed labeling of epithelium (red) and stroma (green). C. Outlines of segmented stromal nuclei objects. D. H&E stained segmented epithelial region. E. Epithelial region is further subdivided into nuclear (red) and cytoplasmic (green) regions. F. Outlines of segmented epithelial nuclei objects.

IIIb. Identification of highly correlated gene expression and image features As described in materials and methods, we performed sparse canonical correlation analysis with 2 factors and selected tuning parameters c1 and c2 based on the resulting $Cor(X_1 w_1, X_2 w_2)$, significance of the correlation, and desired number of non-zero weights for images and genes. A plot of the observed correlation for the set of ten penalty terms tested is shown in Figure 2. We selected c1 = 27.8 and c2 = 3.8 , which gave $Cor(X_1 w_1, X_2 w_2) = 0.95, p = 0.03$ and resulted in ~1000 non-zero gene weights and ~10 non-zero image feature weights.

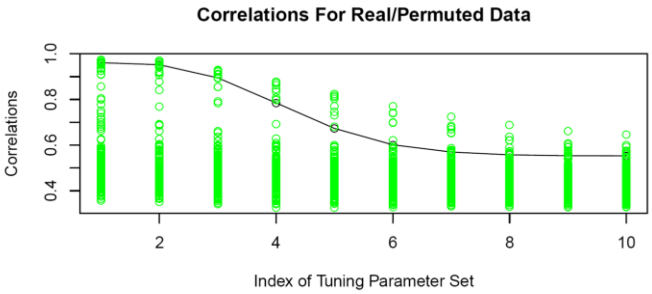


Figure 2. The x axis represents the index of the ten trials of sets of tuning parameters (c_1,c_2), which were evenly spaced from 1 to √c. The y axis represents the correlation between X_1w_1 and X_2w_2. The green dots represent the correlation obtained from permuted data and the black line represents correlation obtained on real data. In our analysis, we selected the terms c_1,c_2 that correspond to tuning parameter index set 2.

IIIc. Biological analysis of sets of highly correlated gene expression and image features

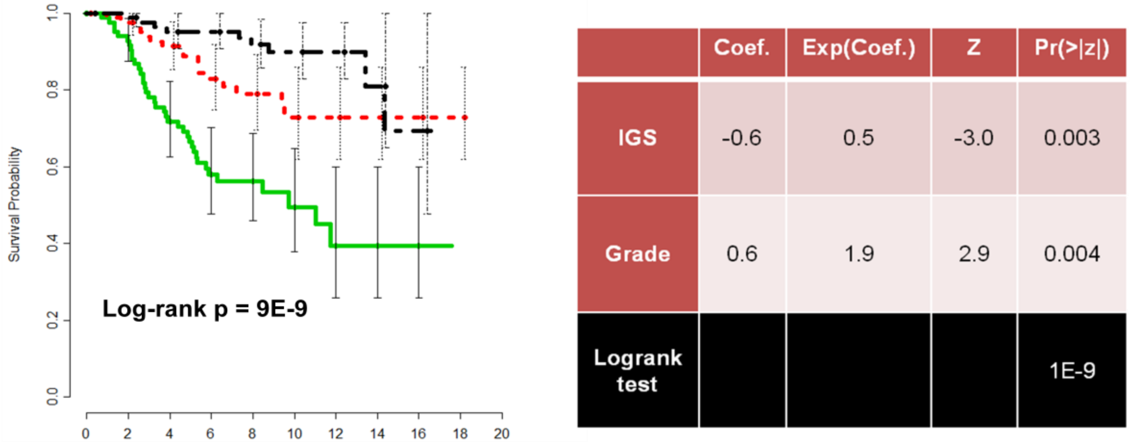
The 9 image features with non-zero weights in Factor 1 were all measurements of stromal density. Functional gene set enrichment analysis of the 1318 genes with non-zero weights in Factor 1 shows that the set of genes with non-zero weights (both positive and negative) is most enriched for membrane glycoproteins (Bonferroni p = 2E-7). The 10 image features with non-zero weights in Factor 2 were descriptors of epithelial nuclear size, shape, and texture. Functional gene set enrichment analysis of the 1297 genes with non-zero weights in Factor 2 shows that the set of genes with positive weights is highly significantly enriched for proteins operating in the stroma/extracellular matrix (Bonferroni p = 1.4E-20) and the set of genes with negative weights is highly significantly enriched for proteins that operate in cell cycle regulation (Bonferroni p = 1.9E-14). These findings represent the first step to a detailed understanding of the gene expression patterns driving stromal and epithelial morphologic variability in breast cancer. The findings from Factor 1 suggest that the expression of membrane glycoproteins may be a regulator of stromal density. Analysis of the Factor 2 gene set suggests that the expression of a coherent set of extracellular matrix proteins is negatively correlated with the expression of a set of proteins regulating DNA replication. These findings suggest a network of epithelial-stromal interactions and provide a candidate list of (likely) epithelial (including: TNFSF13B, PTTG2, CCNB2) and stromal proteins (including: TGFB3, COL3A1, SPON1, SPARC) whose interaction is tightly correlated with nuclear morphologic changes.

IIId. Analysis of Patient Survival

Although we did not supervise our sparse CCA analysis with survival data, it seemed plausible that genes with non-zero weights may be mechanistically important drivers of patient survival, as they were chosen due to their high correlation with tumor morphologic changes. To assess this hypothesis, as described in the Materials and Methods, we computed an IGS score for each patient to summarize the gene and image linear combinations generated by sparse CCA. We stratified patients into three equal-sized groups based on their IGS score and plotted Kaplan Meier curves, which demonstrated significantly different survival rates in the three groups (Log-rank test p = 9E-9) (Figure 3). We fit a Cox proportional-hazards survival model with grade and the discretized IGS score and found that both features contributed prognostic information to the model (both p < 0.01).

A. Kaplan Meier Survival Curves

B. Cox Multivariate Survival Analysis



	Coef.	Exp(Coef.)	Z	Pr(>|z|)
IGS	-0.6	0.5	-3.0	0.003
Grade	0.6	1.9	2.9	0.004
Logrank test				1E-9

Figure 3. Survival analysis. A. Kaplan Meier survival curve of breast cancer cases stratified by IGFS into 3 quantiles. The y axis is probability of survival and the x axis is time in years. The Log-rank test compares the equality of the three survival functions. The p value indicates that the probability of equality of the three survival functions is 9E-9. B. A multivariate Cox proportional hazards survival model was fitted with discretized IGFS and grade, which were both independent predictors of survival in the model. The Cox model takes the form $h_i(t) = h_0(t)\exp(\beta_{IGS}x_{IGS}^{(i)} + \beta_{Grade}x_{Grade}^{(i)})$, where $h_i(t)$ represents patient *i*'s risk of death at time *t*, $h_0(t)$ is the baseline hazard function, and the $\beta's$ are the coefficients for the covariates in the model[4]. Exp(Coef) can be interpreted as multiplicative effects of the covariates on the hazard. Holding grade constant, an IGS score of 3 reduces the yearly risk of death by 50% compared to an IGS score of 2. Column Z displays the ratio of each regression coefficient to its standard error, which is a Wald statistic. Both IGS and Grade have significant coefficients and their Z's are

associated with low p values. The logrank test p value on the final row indicates the probability of the truth of the null hypothesis that all of the β's are zero. This hypothesis is rejected.

IV. Conclusions

In this paper, we developed methods for computational analysis of tumor morphology and integration of these measurements with gene expression data and patient survival data. Using sparse CCA, we identified two factors, each comprised of a set of genes highly correlated with a set of image features. The image features in the first factor were entirely stromal, and the image features in the second factor were entirely epithelial nuclear. Functional gene set analysis of genes with non-zero weights in Factor 1 suggests that membrane glycoproteins may be important regulators of stromal morphology. Analysis of the Factor 2 gene set demonstrated that the genes with positive weights were highly significantly enriched for proteins operating in the stroma/extracellular matrix, while genes with negative weights were highly significantly enriched for proteins that regulate DNA replication/cell cycle. These findings suggest that this set of stromal and cell cycle proteins likely interact in a pathway that directly impacts epithelial nuclear morphologic changes. Since the sparse CCA analysis was designed to identify genes and image features that were mutually correlated, it seemed likely that these sets of morphologic and molecular changes might be mechanistically important and associated with patient survival. To assess the association of the factors identified by sparse CCA with outcome, we simply added the linear combinations of the gene and image features from the two factors into a unified score (IGS). We found that the IGS was highly associated with patient survival and provided additional prognostic information to histological grade. Taken together, our findings suggest that the integration of molecular and morphologic measurements represents a promising new strategy for studying breast cancer biology and for identifying sets of genes and image features associated with patient outcome.

References:

[1] A. K. Abbas, R. S. Cotran, N. Fausto, V. Kumar, J. A. Perkins and S. L. Robbins, *Robbins and Cotran Pathologic basis of disease [print/digital]*, Elsevier/Saunders, Philadelphia, 2005.

[2] P. O. Brown and D. Botstein, *Exploring the new world of the genome with DNA microarrays*, Nat Genet, 21 (1999), pp. 33-7.

[3] D. R. Cox, *Regression models and life-tables*, Journal of the Royal Statistical Society. Series B (Methodological) (1972), pp. 187-220.

[4] J. Fox, *Cox proportional-hazards regression for survival data*, An r and splus companion to applied regression (2002).

[5] S. Gould, R. Fulton and D. Koller, *Decomposing a Scene into Geometric and Semantically Consistent Regions*, *ICCV*, 2009.

[6] S. Gould, A. Y. Ng and D. Koller, *The STAIR Vision Library*, 2008.

[7] S. Gross, *Sparse canonical correlation analysis for the integrative analysis of genomic data User guide and technical document*, Department of Statistics, Stanford University.

[8] H. Hotelling, *Relations between two sets of variates*, Biometrika, 28 (1936), pp. 321-377.

[9] W. Huang da, B. T. Sherman and R. A. Lempicki, *Systematic and integrative analysis of large gene lists using DAVID bioinformatics resources*, Nat Protoc, 4 (2009), pp. 44-57.

[10] N. Otsu, *A threshold selection method from gray-level histograms*, Automatica, 11 (1975), pp. 285-296.

[11] J. Shotton, J. Winn, C. Rother and A. Criminisi, *Textonboost: Joint appearance, shape and context modeling for multi-class object recognition and segmentation*, Lecture Notes in Computer Science, 3951 (2006), pp. 1.

[12] R. Tibshirani, *Regression shrinkage and selection via the lasso*, Journal of the Royal Statistical Society. Series B (Methodological), 58 (1996), pp. 267-288.

[13] M. J. van de Vijver, Y. D. He, L. J. van't Veer, H. Dai, A. A. Hart, D. W. Voskuil, G. J. Schreiber, J. L. Peterse, C. Roberts, M. J. Marton, M. Parrish, D. Atsma, A. Witteveen, A. Glas, L. Delahaye, T. van der Velde, H. Bartelink, S. Rodenhuis, E. T. Rutgers, S. H. Friend and R. Bernards, *A gene-expression signature as a predictor of survival in breast cancer*, N Engl J Med, 347 (2002), pp. 1999-2009.

[14] D. M. Witten, R. Tibshirani and T. Hastie, *A penalized matrix decomposition, with applications to sparse principal components and canonical correlation analysis*, Biostatistics, 10 (2009), pp. 515-34.

[15] D. M. Witten and R. J. Tibshirani, *Extensions of Sparse Canonical Correlation Analysis with Applications to Genomic Data*, Statistical Applications in Genetics and Molecular Biology, 8 (2009).