Integrative Analysis Methods for Biological Problems Using Data Reduction Approaches

by

Ziheng Yang

A dissertation submitted in partial fulfillment
of the requirements for the degree of
Doctor of Philosophy
(Statistics)
in The University of Michigan
2017

Dissertation Committee:

Professor George Michailidis, Co-Chair
Professor Ji Zhu, Co-Chair
Professor Hui Jiang
Research Assistant Professor Alla Karnovsky
Professor Kerby Shedden

Ziheng Yang

yangzi@umich.edu

ORCID iD: 0000-0003-3351-7981

To my parents

ACKNOWLEDGEMENTS

I would like to thank the members of the dissertation committee for their helpful comments, in particular my advisor George Michailidis for being an essential source of guidance in ways not limited to academics. Even before beginning my journey in the graduate program, I had met with George to discuss future prospects and program information. I am confident to say that I would not be where I am now without his counsel and encouragement.

I also wish to thank all of the Statistics Department at the University of Michigan for providing an environment in which I was able to develop professionally as well as personally. Many thanks to Naisyin Wang, Moulinath Banerjee, Edward Ionides, and of course George Michailidis for their unique perspectives and teaching styles that helped shaped my current view of statistics and the academic world.

I would also like to acknowledge my friends, with whom I have shared many memorable experiences during these interesting five years, and most importantly my family, whose valuable support has made this work possible, among other accomplishments.

TABLE OF CONTENTS

LIST OF FIGURES

FIGURE

LIST OF TABLES

TABLE

LIST OF APPENDICES

APPENDIX

ABSTRACT

The “big data” revolution of the past decade has allowed researchers to procure or access biological data at an unprecedented scale, on the front of both volume (low-cost high-throughput technologies) and variety (multi-platform genomic profiling). This has fueled the development of new integrative methods, which combine and consolidate across multiple sources of data in order to gain generalizability, robustness, and a more comprehensive systems perspective. The key challenges faced by this new class of methods primarily relate to heterogeneity, whether it is across cohorts from independent studies or across the different levels of genomic regulation. While the different perspectives among data sources is invaluable in providing different snapshots of the global system, such diversity also brings forth many analytic difficulties as each source introduces a distinctive element of noise. In recent years, many styles of data integration have appeared to tackle this problem ranging from Bayesian frameworks to graphical models, a wide assortment as diverse as the biology they intend to explain. My focus in this work is dimensionality reduction-based methods of integration, which offer the advantages of efficiency in high-dimensions (an asset among genomic datasets) and simplicity in allowing for elegant mathematical extensions. In the course of these chapters I will describe the biological motivations, the methodological directions, and the applications of three canonical reductionist approaches for relating information across multiple data groups.

CHAPTER I

Introduction

Background and Motivation

The ability to measure, store, and process vast amounts of biological data has improved exponentially during the "big data" revolution of the past decade. Low-cost high-throughput technologies such as genomic microarrays (Ventimiglia and Petralia, 2013; Angenendt, 2005) as well as sequencing and mass spectrometry-based assay techniques (van Dijk et al., 2014; Yates et al., 2009) have led to the profiling of DNA, proteins, and small molecules at an unprecedented scale. Even without the state-of the art tools, researchers can access at their fingertips petabytes (10^{15} bytes) of information available in public repositories (Cook et al., 2016; Weinstein et al., 2013). Our strategies for computation are also transforming to keep pace with this deluge of data as more and more computing systems migrate to cloud-based (Dai et al., 2012) and heterogeneous environments (Schadt et al., 2010).

Nevertheless many obstacles still lie ahead in the endeavor of explaining biological systems quantitatively. As always there are technological considerations, such as balancing signal stability versus cost-effectiveness (Ventimiglia and Petralia, 2013) and conserving protein functionality during immobilization assays (Angenendt, 2005). Facilitating consistent curation and expedient exchange of data across publications and databases also remains an important task for preserving the value of data (Howe et al., 2008). Furthermore, there are data-related challenges that even improvements in analytic capabilities cannot fully address. By the famous "curse of dimensionality" (Somorjai et al., 2003), achieving sufficient statistical power becomes difficult when the number of samples is dwarfed by the number of molecular measurements taken. This bottleneck has led to, among other complications, highly unstable genomic signatures that change in

composition depending on the samples included (Ein-Dor et al., 2005). On another front, the "horizontally exhaustive" approach of extensively analyzing single data types has often proved insufficient for modeling the multi-dimensional nature of biology (Hoheisel et al., 2006). For this reason, researchers have begun to look across multiple "omics" data platforms (genomics, proteomics, metabolomics, etc.) to gain a broader view (Kitano, 2002). The advent of big data brought forth many promises, but our current ability to access information far outpaces our ability to comprehend it.

What is perhaps the most central problem for biological data is that it is inherently complex and heterogeneous (Marx, 2013). The molecular dynamics of the cell span numerous levels of regulatory networks (transcription, translation, epigenetic regulation, etc.) that comprise of countless members interacting both within and between levels. Moreover the molecular signatures (or any other biological attributes) are very noisy and tend to vary across individuals, which is especially problematic when sample sizes are small. It should come as no surprise then that analyses restricted among single types of data or single sample cohorts experience limitations in scope. To gain insight into more complex and consequential phenomena, evolving from a series of disconnected snapshots to a multi-faceted and holistic view of the biological system will be essential.

The advancements in our data infrastructure and tools ushered in many new analytic approaches, one of which is the integration of multiple sources of data as a new class of methodologies. Drawing upon multiple sources has two main advantages. The first is that combining data across studies (across observations) grants higher statistical power and improves generalizability. The second is that combining data across data types (across variables) provides a broader systems perspective from multiple vantage points. In the past, limitations in data availability generally reduced the need for serious consideration on how to integrate data. The handling of multiple datasets more often played a restricted role separate from the main analysis, for instance in the form of batch effect adjustment (Benito et al., 2004) or meta-analysis (Rhodes et al., 2002). Now, as collecting and accessing large quantities of data from multiple sources become more feasible, a new wave of full-fledged integration-based techniques has emerged.

Wei (2015) provides an extensive but not exhaustive review of recent integrative techniques. Unsurprisingly, the statistical approaches represented are as diverse as the

biology they aim to model. Perhaps the most abundant are Bayesian hierarchical models (Ruan and Yuan, 2011; Conlon et al., 2006; Wang et al., 2013; Lock and Dunson, 2013; Xing et al., 2011), whose flexible parameterizations offer a natural and interpretable way of combining different biological data types and relationships that connect them. Meanwhile, others adopt a more mathematical approach of using norm penalization to directly induce similarity in model coefficients across data groups (Ma et al., 2011; Fan and Li, 2002; Wang et al., 2009). In order to leverage the unique relational information of genomic regulatory pathways and molecular interaction networks, a number of graphical methods feature ways to incorporate existing knowledge of pathways into traditional frameworks (Khatri et al., 2012; Mitrea et al., 2013) or to consolidate multiple networks into a single model (Kolar et al., 2014; Wang et al., 2014). Finally, dimensionality reduction-based methods (Witten and Tibshirani, 2009; Zhang et al., 2012; Li et al., 2012; Lock et al., 2013) seek to alleviate the daunting complexity of biological data by decomposing or describing the observed variation with only a few key components or modules.

Description and Outline

In this thesis, I will introduce a series of novel dimensionality reduction-based techniques for analyzing multiple groups of data. As with other methods of this class, the goal is to reduce complex patterns into simple elements in the hopes of improving interpretability and efficiency. However, in the process of expanding this reductionist approach to multiple data sources, the intention is to explore the canonical ways in which data can be related across distinct groups. Therefore although the motivating questions at play here all originate from biology, there will be more focus on statistical issues rather than domain-specific ones so as to establish sound principles applicable to general integrative problems. Just as these methods are meant to decompose information across groups, so too do they represent strategies in decomposing the problem of data integration into its fundamental components.

Chapter 2 will focus on the integration of multiple nonnegative datasets under a matrix factorization framework. The problem arises in the cross-platform analysis of

omics expression data in which the goal is to find regulatory modules detectable across multiple data types (e.g. gene, protein, and miRNA expression). Nonnegative data are well-suited for studying biological data as they can be readily interpreted as signal strengths (e.g. of microarray probes), and nonnegative factor solutions provide the intuitive perspective of explaining observed patterns as a sum of parts. However, nonnegativity also brings analytic difficulties in the form of less uniquely defined solutions and unconventional algorithmic approaches. With this in mind, an alternative factorization structure is proposed, which distinguishes between common and distinct variation across datasets. Supplemented by an unconventional yet natural tuning procedure, this produces not only a flexible modeling of the heterogeneity among the data sources, but also an adaptive one.

Chapter 3 will cover an integrative method for classification based on partial least squares, which is designed to consolidate multiple cohorts of predictor and response datasets. This is relevant for developing gene expression-based prognostic signatures that are robust among heterogeneous data from different populations or studies. The methodological basis is partial least squares discriminant analysis, a well-established prediction tool for genomic data that affords great efficiency thanks to its dimensionality reduction basis and iterative regression procedure. The main challenge lies with accounting for additional data groups in a way that retains model simplicity. To this end, a slightly modified parameterization is used which represents the commonality among the data groups in terms of alignment between partial least squares weights. This adjustment proves to be highly compatible to the method's factorized regression approach as well as its sequential algorithmic procedure, achieving robustness through simplicity.

Finally, in Chapter 4 I will discuss an expansion of the principal component analysis framework to study principal variation across multiple data groups. The method will be applied to systematically navigate the multiple experimental conditions of a factorial design cell line study. Borrowing principles from analysis of variance, I establish an analogous multivariate decomposition of variation that preserves the original within- and between-group structure. Translating to the multivariate setting requires a few additional considerations, such as the use of principal angles to quantify the discrepancy between groups. The result is a novel view of multiple datasets in terms of a

reduced set of within and between group components, of which the complexity and commonality can be quantified and utilized for unique types of inferences.

CHAPTER II

A Non-negative Matrix Factorization Method for Detecting Modules in Heterogeneous Omics Multi-modal Data

2.1 Introduction

Technological advances allow biomedical researchers to collect a wide variety of omics data on a common set of samples. Data repositories such as The Cancer Genome Atlas (TCGA) provide multiple types of omics data, thus enabling in-depth investigation of molecular events at different stages of biology and for different tumor types. However, the latter task requires developing methods for data integration, a topic that has received increased attention in the literature.

In genomic studies the integration of multifaceted data is becoming increasingly viable and insightful (Gehlenborg et al., 2010; Jörnsten et al., 2011; Imielinski et al., 2012; Mo et al., 2013). Cellular signals and processes depend on the coordinated interaction and communication among a wide variety of biomolecules including genes, proteins, metabolites, and epigenetic regulators. There are multiple layers in which regulation takes place, and therefore multiple vantage points from which to observe biological activity. A joint analysis of data on the same set of samples from multiple omics sources has potential to achieve more perceptive results over separate analyses, as well as provide a more comprehensive global view of the biological system.

A key challenge for integration methods is dealing with heterogeneous data. Data from different sources are difficult to compare due to inherent discrepancies. Different genomic variables are measured and collected in different ways, and they are associated with different types of noise and confounding effects. Most importantly, they represent different aspects of the biological system. The discrepancy among data sources contributes to a useful multifaceted view of the system, but it also brings forth a new level of complexity that makes it hard to distinguish the coordinated signal.

There are many integration techniques that deal with the complexity of multiple sources by relying on prior knowledge of the relationships that connect them. Some procedures seek to map different experimental data types, such as gene expression, miRNA expression, and copy number variation to a common space of known biological pathways or sets (Khatri et al., 2012; Mitrea et al., 2013, and references therein). Others select features or assign weights to features based on prior knowledge, possibly using such information in a linear-based model (Jensen et al., 2007; Stingo et al., 2011; Jauhiainen et al., 2012) or in a framework for identifying modules (Li et al., 2012; Srihari and Ragan, 2013). All of these approaches require the consultation of an external resource, such as signaling pathways or gene interaction networks. While this supervised approach is convenient (and sensible in certain respects), it relies heavily on the external information being valid and representative, which is not always guaranteed, even in the modern era of data availability. In addition, relating variables based on previously established findings can introduce an element of bias and subjectivity that hinders the discovery of new associations.

In contrast to such supervised approaches, our objective is to develop an integration method that directly leverages the advantage of multiple data sources in order to deal with heterogeneity. In multiple data sets, the signal of interest is typically common among all sources (homogeneous), while extraneous effects tend to differ across sources (heterogeneous). The main principle of our approach is to separate the homogeneous and heterogeneous effects among the sources in order to extract the coordinated signal from extraneous noise. Many existing integration techniques similarly make the distinction between common and distinct effects across sources, such as those extending the Dirichlet mixture model (Lock and Dunson, 2013) and principal component analysis (Lock et al., 2013).

Our proposed method extends an integrative nonnegative matrix factorization framework (Zhang et al., 2012) via a partitioned factorization structure that captures homogeneous and heterogeneous effects. A novel tuning selection procedure allows the model to adapt to the level of heterogeneity among the data sets. We apply our approach to an integrated study of ovarian cancer involving three types of genomic variables, and

discover multi-dimensional modules exhibiting topological patterns of expression across known cancer-related pathways.

2.2 Methods

2.2.1 Nonnegative Matrix Factorization

Nonnegative matrix factorization (NMF) is a powerful tool for data reduction and exploration that has seen popular use in analyzing high-throughput genomic data (Brunet et al., 2004; Tamayo et al., 2007; Devarajan, 2008). The method is related to principal component analysis (PCA), except that it employs the constraint of nonnegativity in lieu of orthogonality. As a result, NMF solutions are less uniquely defined, but are more interpretable.

Given nonnegative data matrix $X_{N\times M}$, NMF finds a nonnegative factorization WH of rank D that best approximates X, typically in terms of the Frobenius norm (Lee and Seung, 1999):

$$\min_{W,H} \quad \| X - WH \|_F^2$$
$$\text{s.t.} \quad W \geq 0, H \geq 0.$$

While Euclidean distance assumes a Gaussian distribution of values, alternative formulations of NMF using Bregman divergences have been proposed (Sra and Dhillon, 2005). Bregman divergences, which bear a strong connection with exponential families (Banerjee et al., 2005), emcompass a wide range of distributional assumptions (e.g. Poisson, Exponential, and probabilistic distributions). Although we use Euclidean distance in the formulation of our method later, alternative loss functions may be accomodated via adjustments to the algorithm.

The factor $H_{D\times M}$ can be interpreted as the basic components of the data, while the elements of $W_{N\times D}$ can be thought of as latent factors associated with these components. Thus, each observation (row of X) is approximated by a linear combination of components (rows of H) with weights given by each row of W. The full data is explained by a sum of additive parts. In biological contexts, this is intuitive because

biological entities and mechanisms can be naturally described with a signal that is either present or absent.

Due to the constraint of nonnegativity of the approximation elements, solutions to NMF are only unique up to scalings and rotations. Specifically, scaling and rotating of the columns of W and rows of H appropriately will not alter the overall matrix product WH. For this reason, what is of interest in practice is not the values of the matrix elements, but their relative magnitudes with respect to each column of W or row of H.

At its core, NMF studies the data from a different vantage point (the origin) than orthogonality-based approaches (center of mass) such as PCA, partial least squares regression (PLS), and canonical correlation analysis (CCA). Besides being more intuitive, this also offers certain advantages such as the ability to capture context-dependent patterns (Devarajan, 2008). Meanwhile, for our purposes, the flexibility of the factorization is also convenient for dealing with heterogeneous data.

2.2.2 Joint NMF

Joint NMF (jNMF) was developed as an extension to NMF for integrating multiple data sets with a common set of observations (Zhang et al., 2012). For K data matrices $(X_1)_{N\times M_1}, \dots, (X_K)_{N\times M_K}$, the formal problem is:

$$\min_{W,H_1,\dots,H_K} \quad \sum_{k=1}^{K} \| X_k - WH_k \|_F^2$$

$$\text{s.t.} \quad W \geq 0, H_k \geq 0, k = 1, \dots, K,$$

with $W_{N\times D}, (H_k)_{D\times M_k}$ producing K rank D approximations. The method can be described as multiple NMF problems subject to a shared factor matrix. Other decomposition-based integration methods have been proposed, including multiple Canonical Correlation Analysis (Witten et al., 2009), multi-block Partial Least Squares (Li et al., 2012), and Joint and Individual Variation Explained (Lock et al., 2013). Such approaches use the orthogonality constraint, whereas jNMF and our proposed method employ nonnegativity.

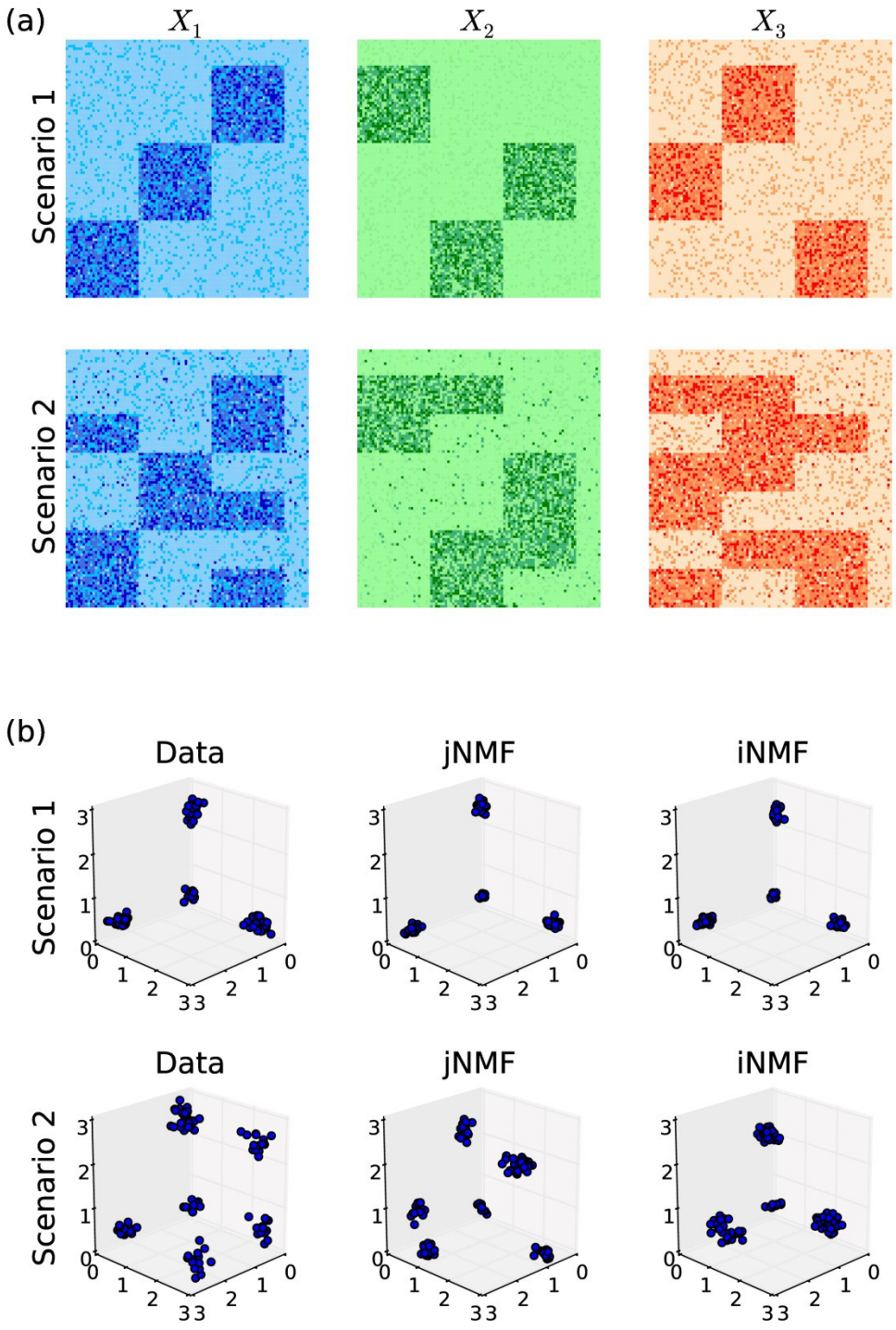

Figure 2.1: (a) An example of multi-dimensional modules across three different data sources. Three modules are distinguishable in Scenario 1 as strong associations between subsets of variables across sources and a common subset of observations. Scenario 2 contains the same data with added random noise and confounding effects. (b) Low-dimensional representations of the data (X_2), jNMF approximations (W), and iNMF approximations (W). The modules are clearly detected by both methods in Scenario 1, but only by iNMF in Scenario 2.

The method was shown to be able to detect coordinated activity across multiple genomic variables in the form of multi-dimensional modules. The exact definition of modules slightly differs across studies (Li et al., 2012; Roy et al., 2013; Jin and Lee, 2015), but their general purpose is to group variables based on common function or association. This serves as a useful preliminary step to reduce the dimensionality of the problem. Multidimensional modules capture common signals across multiple sources of data (see Figure 2.1a). In jNMF, as well as in our method, each module represents a biclustering of both observations and variables, which can be visualized as a block in the data matrix after appropriate rotation.

A limitation of jNMF is that it is not methodologically different from standard NMF. In fact it is easy to show that the problems are equivalent by setting $X = (X_1, \dots, X_K)$ and $H = (H_1, \dots, H_K)$. As a consequence, the optimization step of jNMF does not distinguish between different variable sources when integrating. This is a disadvantage when dealing with heterogeneous data.

The toy example in Figure 2.1 illustrates this. The heatmaps (Figure 2.1a) depict two scenarios of a three-source integration problem. In Scenario 1, three modules are easily distinguishable in all sources as blocks, which associate different subsets of variables with the common observation groups. Scenario 2 contains the same data, except with added noise (generated as discussed in Section 3.1). In particular, the additional block structures that are misaligned with the underlying modules represent confounding effects that vary from source to source.

Figure 2.1b plots (in low-dimensional space) the data and the corresponding solutions of jNMF and our proposed method (iNMF). Both methods clearly distinguish the signal when the signal is clean (Scenario 1), but jNMF is less robust to heterogeneous noise across the sources (Scenario 2). While jNMF is very effective for detecting homogeneous effects, its factorization structure WH_k leaves no room for heterogeneous approximations. As a result, jNMF is especially sensitive to random noise and confounding effects, because they typically have different structures across sources. We seek to remedy this via expanding the factorization structure.

2.2.3 Integrative NMF

Our proposed method, integrative NMF (iNMF), leverages the advantage of multiple data sources in order to gain robustness to heterogeneous perturbations. While jNMF considers homogeneous effects WH_k, iNMF additionally considers heterogeneous effects V_kH_k. Formally, for nonnegative observationally-linked data sets $X_1, \dots, X_K$ as defined previously, the optimization problem is the following:

$$\min_{\substack{W,H_1,\dots,H_K,\\ V_1,\dots,V_K}} \sum_{k=1}^{K} \| X_k - (W + V_k)H_k \|_F^2 + \lambda \sum_{k=1}^{K} \| V_kH_k \|_F^2$$

$$\text{s.t.} \quad W \geq 0, H_k \geq 0, V_k \geq 0, k = 1, \dots, K.$$

To retain identifiability, we penalize the Frobenius norm of the heterogeneous effects V_kH_k, as WH_k can always be expressed in terms of V_kH_k but not vice-versa. Rewriting $V_kH_k = (W + V_k)H_k - WH_k$, we see that the objective function is simply a partitioned version of the jNMF objective, which penalizes $X_k - WH_k$.

The idea of combining homogeneous and heterogeneous parts across sources is reminiscent of the one-way ANOVA model, in which the total variation is explained by joint and individual effects across groups: $y_i = \mu + \alpha_j + \epsilon_{ij}$. However, while the ANOVA common effect μ is estimated to be the sample mean, the iNMF homogeneous effect W is actually the element-wise minimum of the approximated latent factors $W + V_k$, since $V_k \geq 0$. For this reason, W, V_k cannot be directly used to infer the level of joint and individual effects among the sources, since W will be overestimated (and V_k underestimated) when parts of the individual effects are homogeneous. Thus, it is more appropriate to refer to W, V_k as approximations of the true joint and individual effects rather than their estimates.

Interestingly, restricting $W \geq 0, V_k \geq 0$ is methodologically equivalent to restricting $W + V_k \geq 0, V_k \leq 0$. In the latter, the approximated common factor W represents the element-wise maximum of $W + V_k$, rather than the element-wise minimum. Therefore, imposing nonnegativity on V_k does not lead to bias issues, but instead a particular perspective on the joint effects. It is also possible to allow for both positive and negative values for V_k if we set $W = \text{mean}(V_k)$, for instance.

The parameter λ can be viewed as the homogeneity parameter, since larger values induce smaller $V_k H_k$. The advantage of iNMF can be summarized as follows. When data sets from multiple sources contain homogeneous elements, performing separate analyses ($\lambda = 0$) sacrifices power; when they contain heterogeneous elements, a purely joint analysis ($\lambda = +\infty$) is sensitive to extraneous noise. In real applications, data consists of a mixture of both, and so iNMF functions as a mixture of jNMF and NMF.

2.2.4 Algorithm

The classical algorithm for NMF was introduced by Lee and Seung (2001), and consists of simple multiplicative updates derived from auxiliary functions. Over the years, new approaches based on gradient descent and alternating least squares have been proposed (Berry et al., 2007; Lin, 2007), which offer faster convergence and better convergence guarantees. However, these alternatives generally involve an explicit projection step to ensure nonnegativity of solutions, whereas with multiplicative updates nonnegativity is implicitly guaranteed. We base our algorithm for iNMF on the original method of Lee and Seung (2001), as it provides a more natural and flexible foundation from which to develop extensions, although other approaches are certainly viable.

Begining with random positive initializations, we perform the following element-wise updates at each iteration until convergence:

$$W_{ij} \leftarrow W_{ij} \frac{(\sum_k X_k H_k^T)_{ij}}{(\sum_k (W + V_k) H_k H_k^T)_{ij}}$$

$$(H_k)_{ij} \leftarrow (H_k)_{ij} \frac{((W + V_k)^T X_k)_{ij}}{((W + V_k)^T (W + V_k) H_k + \lambda V_k^T V_k H_k)_{ij}}$$

$$(V_k)_{ij} \leftarrow (V_k)_{ij} \frac{(X_k H_k^T)_{ij}}{((W + V_k) H_k H_k^T + \lambda V_k H_k H_k^T)_{ij}}.$$

Since the iNMF objective function is non-convex, one should perform many repetitions and choose the minimizer of the objective function as the final solution. The proof of monotonicity of the objective function under these updates is provided in Appendix A.1.

2.2.5 Sparse Formulation

Although NMF naturally gives rise to parsimonious solutions (Lee and Seung, 1999), sparsity can be further induced via penalization. We adopt a method similar to the one used in Mankad and Michailidis (2013), which applies the L1-norm to elements of H_k. This produces a slightly different objective function:

$$\sum_{k=1}^{K} \| X_k - (W + V_k)H_k \|_F^2 + \lambda \sum_{k=1}^{K} \| V_k H_k \|_F^2 + \lambda_s \sum_{k=1}^{K} \| H_k \|_1,$$

and algorithm:

$$W_{ij} \leftarrow W_{ij} \frac{(\sum_k X_k H_k^T)_{ij}}{(\sum_k (W + V_k)H_k H_k^T)_{ij}}$$

$$(H_k)_{ij} \leftarrow (H_k)_{ij} \frac{((W + V_k)^T X_k)_{ij}}{((W + V_k)^T (W + V_k)H_k + \lambda V_k^T V_k H_k)_{ij} + \lambda_s}$$

$$(V_k)_{ij} \leftarrow (V_k)_{ij} \frac{(X_k H_k^T)_{ij}}{((W + V_k)H_k H_k^T + \lambda V_k H_k H_k^T)_{ij}}.$$

A similar sparsity formulation involving the same penalization term can be derived for jNMF.

2.2.6 Tuning Selection

As with other sparse NMF formulations (Gao and Church, 2005; Kim and Park, 2007; Mankad and Michailidis, 2013), the sparsity parameter λ_s is best left to be chosen manually to adjust for interpretability, although it should be noted that too large of a choice leads to degenerate solutions. For selecting the number of modules D, a common method is to use a consensus-based approach (Brunet et al., 2004), which determines the credibility of each tuning choice based on the stability of the corresponding solutions. From basic intuition, given the most appropriate ranks $D_k, k = 1, \dots, K$ for individual data sets, the integrated rank should lie somewhere between $\max_k D_k$ and $\sum_k D_k$. However, it is sometimes preferable to choose a smaller rank for a simpler representation consisting of the top D modules.

Although a consensus-based strategy may also be used for the homogeneity parameter λ, the nature of the iNMF framework actually allows a simpler procedure. To separate the homogeneous and heterogeneous parts, we rely on measuring the level of heterogeneity across the sources. We do this by comparing the objective values of jNMF, which represent complete homogeneity, and separate NMFs (sNMF), which represent complete heterogeneity.

Given a decreasing sequence of λ, the procedure is as follows:

1. Perform jNMF and sNMF on the data sets, and store the unsquared residual quantities:

$$R_J = \sum_k \| X_k - W^{(J)} H_k^{(J)} \|_F , R_S = \sum_k \| X_k - W_k^{(S)} H_k^{(S)} \|_F .$$

2. For each λ in the decreasing sequence:
   a. Perform iNMF with homogeneity parameter λ and store:

$$R_I^{(\lambda)} = \sum_k \| X_k - W^{(I,\lambda)} H_k^{(I,\lambda)} \|_F .$$

   b. If $R_I^{(\lambda)} - R_J > 2(R_J - R_S)$, then stop and select the previous λ.

By selecting the smallest λ for which the threshold is not exceeded, we seek to attribute as much of the data as possible to heterogeneous effects ($V_k H_k$) before overfitting. Here, overfitting is detected when the difference between the iNMF and jNMF residuals, $R_I^{(\lambda)} - R_J$, becomes significantly large, as typically we would expect jNMF to detect some of the joint signal. More discussion on this procedure can be found in Appendix A.3.

2.3 Simulation Study

We compare jNMF and iNMF based on their abilities to identify the structure of the true modules, which amounts to identifying the correct biclusters of observations and variables. We generated data based on a joint block diagonal structure representing the modules (or joint effects) of interest. We then perturbed the data using three different methods, as follows. To simulate heterogeneous effects from extraneous factors, we

randomly add blocks with probability σ_h to the base structures. These blocks are aligned with the columns of the modules, but not their rows so as to be heterogeneous with respect to variable sources. To simulate random noise, we applied two types of error (scattered and uniform) independently to each data cell. Scattered error switches each entry value between zero and nonzero with probability σ_s, while uniform error adds a random Unif$(-\sigma_u, \sigma_u)$ variable to the entry and takes the absolute magnitude. Further details on the data generation process can be found in Appendix A.3. The final generated data matrices resemble those in the bottom row of Figure 2.1a.

The Frobenius norm error of the approximation is not useful here as a performance measure, since the goal is to identify the true modules rather than to approximate the data. Instead, we measure the level of signal detected relative to noise by considering the matrices WH_k, which represent the approximated homogeneous effects. For each data set X_k, the module detection score S is defined as:

$$S = (\mu_{\text{signal}} - \mu_{\text{noise}})_+ / \mu_{\text{signal}},$$

where $\mu_{\text{signal}}, \mu_{\text{noise}}$ are the averages of the values of WH_k that lie inside and outside of the true modules, respectively. This score is invariant to rotations and scalings of W, H_k, and it measures how well observations and variables are grouped according to the true modules. We take the average score S over all K data sources as the final module detection score.

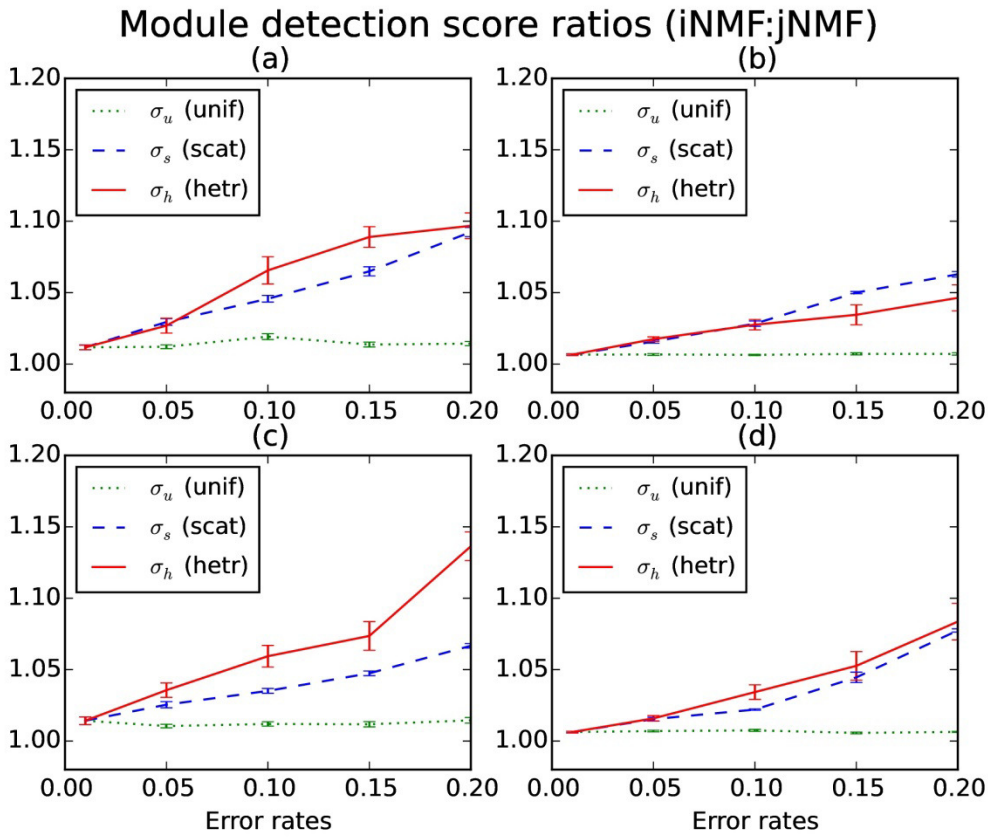

Figure 2.2: Average ratios (iNMF:jNMF) of detection performance (S) over 25 trials (with standard errors) under four data and module dimensions, with three types of perturbations (uniform, scattered, heterogeneous). The leftmost common point in each subplot represents the error scenario $\sigma_u = \sigma_s = \sigma_h = 0.01$, while each trajectory represents raising the level of a single type of error. (a): 2 sources of 40×40, 4 modules of 8×8; (b): 2 sources of 80×80, 8 modules of 8×8; (c): 2 sources of 72×72, 4 modules of 16×16; (d): 4 sources of 40×40, 4 modules of 8×8.

We compared the performance of jNMF and iNMF (200 repetitions used in each) under four different data scenarios: baseline (a), large number of modules (b), large size of modules (c), and larger number of data sets (d). Figure 2.2 plots the average ratios between the iNMF and jNMF detection scores. Under high levels of scattered and heterogeneous error, iNMF significantly outperforms jNMF in identifying the true modules. Higher levels of uniform error does not seem to lead to significant differences. The two methods are only comparable under homogeneous and noise-free settings, otherwise the advantage of iNMF is clear. This adaptivity of iNMF makes it robust to heterogeneity and noise that would meanwhile confound jNMF.

2.4 Application to Detecting Multi-modal Modules of Ovarian Cancer

2.4.1 Data Preparation and Preprocessing

For our application, we conduct a joint analysis of genetic and epigenetic variables to study biomarkers associated with ovarian cancer. The data was downloaded from The Cancer Genome Atlas (TCGA) on August 28, 2014 from the platforms Illumina 27K (DM), Agilent G4502A-07-2, Agilent G4502A-07-3 (GE), and Agilent H-miRNA 8x15K v2 (ME). All variables were Level 3 processed. The full data consists of 15661 DNA methylation (DM), 14821 gene expression (GE), and 799 miRNA expression (ME) variables from a common set of 592 ovarian cancer samples.

Variables with missing observations were omitted. Variance stabilization and nonnegativity transformations were applied as follows. GE data was randomly truncated at $-4+\epsilon$ and $4-\epsilon$ where $\epsilon \sim$ i.i.d. Unif$(0, 10^{-3})$, and then shifted +4 units. This is equivalent to applying the function $f(x) = \min\{\max\{x, -4+\epsilon\}, 4-\epsilon\} + 4$ to each entry. Random truncations serve to prevent data singularity issues. ME data was log2 transformed, truncated at $2+\epsilon$ and $6-\epsilon$ with the same method, and shifted -2 units. Each data set (DM, GE, ME) was then normalized according to its within-source standard deviation. Other normalization strategies are discussed in Appendix A.4. Next, we removed DM variables with means below the 15th percentile, and then DM and GE variables with variances below the 15th percentile, which produced the final data sets described above. This filtering procedure is similar to the one used in Zhang et al. (2011).

2.4.2 Module Discovery and Validation

We performed the sparse versions of jNMF and iNMF (200 repetitions each) on the post-processed TCGA data with $\lambda = 0.1$ (as chosen by our selection procedure) for a range of sparsity parameter choices $\lambda_s = 10^{-4}, 10^{-3}, 0.01, 0.1, 1$. We first evaluated the validity of the findings based on concordance with reference DM, GE, and ME variables clusters from relevant literature. These reference clusters consist of either two or four groups of variables each, and so we chose $D = 2,4$ to allow for more appropriate comparisons. Our own empirical variable clusters were computed from the factor

matrices H_k. We normalized each row of H_k by its mean, and assigned each variable to a cluster $1, \dots, D$ based on the maximum in each column.

Our first two reference clusters were derived from an integrative study of ovarian cancer by Bell et al. (2011) using DNA methylation, gene expression, miRNA expression, and DNA copy number variation data from TCGA. Consensus NMF clustering established four disease subtypes based on prominent gene markers in each cluster. These four groups of genes, and their associated DNA methylation variables (information provided by TCGA), comprised our reference GE and DM clusters. Another integrated analysis by Creighton et al. (2012) identified sets of miRNAs significantly associated with better or worse survival rates for ovarian cancer patients. We used these two groups of variables as our ME reference. A full list of these reference clusters is provided in Appendix A.5.

We assessed concordance between our empirical results and the reference using two metrics, the Gini impurity index (Hastie et al., 2009) and the cluster purity Kim and Park (2008). The Gini index for empirical cluster i is defined as:

$$I_i = \sum_{d=1}^{D} \hat{p}_{d,i}(1 - \hat{p}_{d,i}),$$

where $\hat{p}_{d,i}$ is the proportion of elements in empirical cluster i belonging to reference cluster d. For each data source, we compute this quantity for each empirical cluster $i = 1, \dots, D$, and take the average as the impurity score I. The cluster purity is defined as:

$$P = \frac{1}{n} \sum_{d=1}^{D} \max_{1 \le d \le D} n(d, i),$$

where n is the total number of members in all empirical clusters and $n(d, i)$ is the number of members of empirical cluster i belonging to reference cluster d. Whereas I measures the level of disagreement within each empirical cluster, P measures the level of agreement between the empirical and reference clusters.

Table 2.1: Impurity (I) and purity (P) scores (in percentages) of empirical clusters obtained from jNMF and iNMF with respect to three reference clusters. Shading indicates significantly ($\geq$2 sd) higher concordance compared to both the alternative method and the null distribution.

		I			P		
		DM	GE	ME	DM	GE	ME
Null clusters	mean	61	58	44	49	50	65
	st.dev.	4	7	2	5	8	1
$\lambda_s = 1$	jNMF	57	42	42	58	69	65
	iNMF	52	33	35	58	77	76
$\lambda_s = 0.1$	jNMF	64	12	44	58	92	65
	iNMF	46	22	41	67	85	68
$\lambda_s = 0.01$	jNMF	61	40	44	50	69	65
	iNMF	53	18	16	58	85	91
$\lambda_s = 10^{-3}$	jNMF	64	12	42	50	92	65
	iNMF	62	32	39	58	77	71
$\lambda_s = 10^{-4}$	jNMF	58	32	42	50	77	65
	iNMF	55	32	37	58	77	74

For each of these statistics, we simulated null distributions (1000 samples) by randomizing cluster assignments. Table 2.1 compares the impurity and purity scores with respect to all three reference clusters, applied to modules obtained by jNMF and iNMF (as well as from the null distribution) for a range of sparsity parameter choices. We see that the iNMF clusters are generally more concordant with established findings as well as more stable, as evidenced by the scores corresponding to the GE reference. This reflects iNMFs ability to more clearly distinguish the joint signals in the midst of heterogeneous confounders that are likely present among the DM, GE, and ME variables.

Table 2.2: Overlap in membership between observational clusters. Our results from iNMF are concordant with (a) csNMF clusters (498 samples), but not with (b) netNMF clusters (225 samples). Shading indicates maxima in both rows and columns.

(a)	csNMF			
	I	P	D	M
I	65	23	11	14
P	2	105	16	6
D	19	11	76	9
M	22	2	34	83

(b)	netNMF			
	1	2	3	4
I	12	23	0	14
P	15	47	0	9
D	4	34	1	5
M	39	18	1	3

The second step of our validation involves assessing the observational clusters generated by our modules (using the results for $\lambda = 0.1, D = 4$. Similar to before, we partitioned our 592 observations into four groups based on the maximum value within each row of the column-mean normalized W matrix. We compared these clusters with results from Bell et al. (2011) (results obtained from Verhaak et al. (2013)) and Hofree et al. (2013) who analyzed samples overlapping with ours. The first group used consensus NMF (csNMF) clustering, while the second applied a network-regularized NMF (netNMF) based on networks from public databases. Concordance tables are presented in Table 2.2.

Our empirical clusters largely coincide with those of csNMF, indicating that the underlying true signal among DM, GE, and ME variables is strong. However, there are some discrepancies, particularly among the modules (I) and (M). This suggests that the samples from these modules contain higher levels of heterogeneous noise. Because iNMF is able to adjust to this type of noise, its clusters are likely a more accurate reflection of the true clusters. Meanwhile, there is not as strong concordance between iNMF and netNMF clusters, which is likely due to the influence of external network information in the latter method. While the incorporation of such information brings in new perspectives, the reliability of the procedure is heavily dependent on the accuracy and relevance of the information. In addition, tuning selection is a delicate issue, as it is difficult to determine where exactly the underlying truth lies between what are suggested by observed patterns and prior input.

Although relying on external information can be useful in guiding the analysis, there are a few disadvantages. One is that such information may be unreliable. Although public databases are becoming increasingly extensive and well-curated, their results are

nevertheless aggregated from many studies with different designs and objectives, and are thus prone to accumulated errors and oversimplification. Incorporating additional information can be misleading if the information is messy or incongruous with the research question, as demonstrated in our validation step with observational clusters.

Furthermore, when the procedure is supervised, findings will naturally tend towards the reference. This is somewhat favorable, since results that largely deviate from well-established findings are less credible. However, for the purpose of discovery, there is limited utility in selecting new candidates based solely on to existing results. It is less subjective to withhold external information until after the analysis. We address both of these concerns by performing integration independently of enrichment, thereby allowing our module discovery step to be data-driven rather than input-driven.

2.4.3 Follow-up Analysis of Modules

Current methods of attaching biological relevance to discovered modules frequently involve enrichment according to either pathways gathered from various gene or interaction databases or experimental results (Zhang et al., 2012; Li et al., 2012; Roy et al., 2013; Jin and Lee, 2015). In such studies, the number of modules being considered is very high, which is suitable for associating with large collections of biological pathways and interactions. In contrast, our study deals with substantially fewer modules, which represent broader effects that are more appropriately associated with disease subtypes. Our analysis will span multiple cancer-related pathways extracted from BioCarta and relevant literature. Based on the distribution of module expression among these pathways, we will observe topological patterns of genomic expression and connect them with ovarian cancer subtypes.

For the rest of this section, we will focus on the modules discovered by iNMF at $\lambda_s = 0.01$, as they appear to be most concordant with the reference variable clusters, in particular the GE cluster that is associated with four subtypes of ovarian cancer: Immunoreactive, Proliferative, Differentiated, and Mesenchymal (Bell et al., 2011). These subtypes were defined based on high expression of gene markers associated with

responsiveness to antigens (I), proliferation (P), cell differentiation (D), stromal cell development (M).

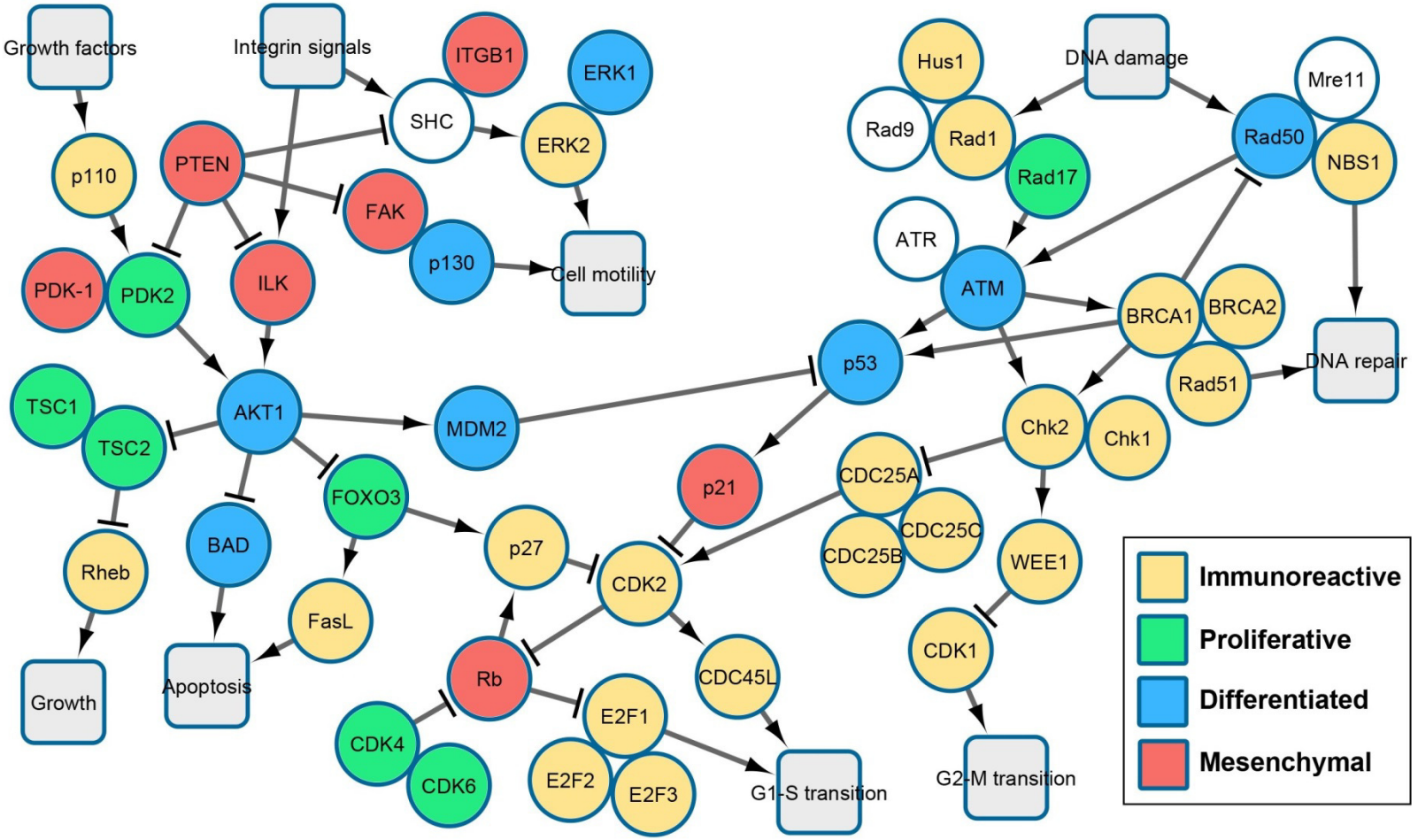

Figure 2.3: Module memberships of genes (from iNMF) arranged according to pathways derived from BioCarta and relevant literature, and include processes of DNA repair (top right), cell cycle regulation (bottom), cell survival and proliferation (left), and cell migration (top left).

As in our validation step, we assigned genes to modules (I/P/D/M) based on the maximum value within each column of the normalized H_k matrix. Thus, membership to a module means that a gene is most highly expressed in that module relative to other modules. Figure 2.3 shows the distribution of the modules across multiple cancer-related processes, which include DNA repair (top right), cell cycle regulation (bottom), cell survival and proliferation (left), and cell migration (top left). Visualization was performed with Cytoscape (Cline et al., 2007).

The DNA repair pathway begins with the Rad9/Hus1/Rad1 and Rad50/Mre11/NBS1 complexes, which sense DNA damage. The signal is transduced via the protein kinases ATM and ATR to checkpoint regulators p53, Chk1, and Chk2 that delay cell cycle progression, as well as to inducers of homologous repair BRCA1, BRCA2, and Rad51 (Yoshida and Miki, 2004; Houtgraaf et al., 2006). Cell cycle progression is managed by CDK2-activated CDC45 (initiates DNA replication), transcription factors E2F (activate S phase progression), and CDK1 (promotes G2-M transition). Also, Rb1 is a tumor suppressor involved in regulating many cellular processes, including G1-S transition, proliferation, and differentiation (Giacinti and Giordano, 2006).

In the PI3-Kinase pathway, growth factors activate PI3K, of which p110 is a catalytic subunit, and directly opposes PTEN in phosphorylating PIP2 into the lipid messenger PIP3. PIP3 recruits the kinase AKT, which begins a variety of signaling cascades that lead to growth, survival, and proliferation. AKT inhibits proapoptotic BAD and growth-inhibiting TSC, as well as activates MCM2, which degrades the cell cycle regulator p53. Phosphorylation of FOXO by AKT retains it in the cytoplasm and prevents its transcriptional activation of cell cycle regulation (via p21) and apoptosis (via FASLG), thus promoting proliferation and survival (Chalhoub and Baker, 2009). Lastly, transmembrane integrin signals activate FAK and SHC which initiate cell migratory pathways involved in directional migration and random motility, respectively. Both of these pathways are inhibited by PTEN via dephosphorylation.

By viewing the collection of pathways in light of the module memberships, we see several interesting patterns and connections. Members of module (I) are the most common, and are mainly distributed among the DNA repair and cell cycle regulation

pathways. This may represent a baseline biomarker signature that is persistent throughout a cell's life cycle. Members of module (P) are associated, appropriately, with proliferation and survival pathways. Genes in module (D) are more dispersed, and participate in a number of processes including checkpoint regulation, survival, and cell migration. Finally, genes in module (M) seem to be involved in upstream regulation of cell migration as well as tumor suppression, indicating late stages of tumor development.

It is important to note that our discovered modules do not necessarily equate to subtypes of observations or variables. Although the modules can certainly be used to characterize subtypes as we have shown, there is not necessarily a one-to-one correspondence between the two. For instance, in our above analysis (Figure 2.3) module (I) was most highly expressed among many variables, but the distribution of the other modules (P/D/M) may reveal alternative ways to subtype these variables. The modules discovered here describe genomic and observational patterns that additively construct the observed data most efficiently. In this sense, they represent the underlying latent mechanisms that give rise to both observation and variable subtypes, but not necessarily the subtypes themselves.

2.5 Discussion

As data collection technologies improve and data repositories expand, the quality and accessibility of data from multiple biological sources will continue to grow. As a result, the combined perspectives from internal signatures (e.g. genes, proteins, and metabolites) as well as external information (e.g. clinical status, patient history, and environmental factors) are contributing to an increasingly rich and complex model of the biological system. However, the abundance and diversity of data is accompanied by the problem of heterogeneity, both in the nature of data sources and in the data collection processes. It is important for strategies of data integration to evolve alongside these new challenges.

We have introduced a novel method of data integration based on a classical matrix decomposition technique. Our method was applied to an integrative study of ovarian cancer, in which we discovered multi-dimensional modules consistent with

previously established variable-based subtypes as observational clusters. These modules express notable topological patterns among cancer-related pathways, suggesting a connection with underlying biomarker signatures associated with disease subtypes.

The key merits of our approach are as follows. As with jNMF, iNMF is able to detect coordinated signals across multiple data sets. However, iNMF is also equipped to deal with issues arising from heterogeneous data. With its more flexible factorization structure, iNMF is able to adapt to the level of disparity between the data sets, in order to extract the joint signal of interest from heterogeneous confounders. To distinguish between common patterns spanning multiple sources and distinct patterns unique to individual sources is the first step for developing a proper integration procedure.

The basic framework of iNMF leaves room for further regularization beyond sparsity to be incorporated. One possibility is to consider relationships between individual variables from the same data source (gene-gene interactions) or from different sources (miRNA-gene or DNA methylation-gene regulations) (Li and Li, 2008; Zhang et al., 2011). Another approach is to induce adherence to known biological networks or observational relations by means of network statistics. The main challenges are adapting the penalties to the NMF framework and finding effective strategies for tuning selection.

Although our analysis examined several types of genomic variables, our results capture only a snapshot of cancer biology. For future investigations, it may be fruitful to explore more types of genomic data, such as DNA copy number variation and mutation status, or even clinical information. It may also be worthwhile to expand the analysis to multiple types of cancers. With the right tools, having a wider selection of data sources will only help in understanding complex disease mechanisms.

CHAPTER III

An Adaptive Partial Least Squares Classifier for Robust Prognostic Gene Signatures

3.1 Introduction

In recent years, studies on the genomic patterns of breast cancer have led to a number of molecular signatures associated with various disease attributes such as metastasis (Wang et al., 2005), p53 status (Miller et al., 2005), response to treatment (Ayers et al., 2004), and clinical outcome (Finak et al., 2008). These results have led to a better understanding of the complex genetic landscape underlying the disease (Sørlie et al., 2003; Hu et al., 2006; Prat et al., 2010), as well as the development of new microarray-based tools for guiding treatment strategies such as MammaPrint (van't Veer et al., 2002) and Oncotype (Paik et al., 2004). However the clinical utility of these molecular signatures, or their value in helping to distinguish between high-risk and low-risk cases, is limited in that they at best provide only incremental improvements from using standard clinicopathological parameters and are unable to replace them (Sotiriou and Lajos, 2009). Moreover, the prognostic power of these signatures is primarily based on proliferation-related and estrogen receptor (ER) signaling-related genes (Weigelt et al., 2012), which suggests that there is much potential variation of the disease not yet accounted for.

One key obstacle for developing effective genomic signatures is the heterogeneity among studies, as the differences in tumor samples and experimental procedures inevitably confound the observed data and ultimately the findings. Coupled with the complex correlation structure among genes, this leads to highly unstable and cohort-dependent signature contents (Ein-Dor et al., 2005). One answer is to combine multiple studies in order to dampen the cohort-specific noise (Hu et al., 2006; Teschendorff et al., 2006; Wirapati et al., 2008), which has proved useful in identifying similar groups of

patients with poor outcome (Weigelt and Reis-Filho, 2010). Since the molecular patterns of breast cancer are especially diverse (Sørlie et al., 2003; Hu et al., 2006), drawing upon multiple data sources may be essential to producing robust and generalizable results.

Data integration methods in the genomics setting come in a large variety. Meta-analyses statistics such as the Cox regression score (Teschendorff et al., 2006) or the module score (Wirapati et al., 2008) take the most direct approach of summarizing the signal from each cohort with a convenient metric, but this only provides a snapshot of the information gathered from each study. Methods such as those based on statistical discrimination (Marron et al., 2007) and surrogate variable analysis (Leek and Storey, 2007) aim to eliminate systemic differences (batch effects) between datasets, but they are designed to adjust the data independently of the main analysis, and thus may be used concurrently with other integration techniques (such as this one). Adaptivity is the hallmark of a Bayesian paradigm, and there have been a number of Bayesian frameworks which provide flexible modeling with a hierarchical structure (Shen et al., 2004; Conlon et al., 2006; Scharpf et al., 2009). However, these approaches generally depend on a heavily parameterized model and such complexity may find it difficult to produce generalizable results for genomics applications. Notably, the aforementioned papers do not approach validation on independent datasets.

Not considered here, but of interest to the genomics setting, is the integration of data gathered from the same observations but across different platforms, such as gene expression, miRNA expression, DNA methylation, and clinical variables. There have been a handful of methods designed to model the complex relationship between genetic and epigenetic variables. These include approaches based on linear regulatory modules (Zhu et al., 2016), Bayesian hierarchical modeling (Wang et al., 2013), hypothesis testing (van Iterson et al., 2013), and partial least squares (Li et al., 2012). These methods address a different problem than the one considered in this paper, which is the analysis of multiple datasets of the same data type.

We introduce integrative partial least squares (iPLS), an new classification tool for consolidating multiple cohorts of data. The method is a natural extension of a classical dimensionality reduction-based regression method known for its efficiency in high dimensions. A simple but powerful modification results in an adaptive model that is

robust to heterogeneity between cohorts. Simulation results and a data example for predicting cancer recurrence using multiple gene expression data cohorts demonstrate the method's advantage over existing benchmarks.

3.2 Methods

3.2.1 Partial Least Squares

Partial least squares (PLS) regression has been widely used in genomics as a tool for classification (Nguyen and Rocke, 2002; Pérez-Enciso and Tenenhaus, 2003; Fort and Lambert-Lacroix et al., 2005), dimensionality reduction (Boulesteix, 2004; Gidskehaug et al., 2007), and more recently variable selection (Lê Cao et al., 2008; Chun and Keleş, 2010; Li et al., 2012). Fundamentally, PLS decomposes explanatory and response data into linear combinations of latent factors, which represent the underlying key mechanisms of the system. The inherent dimensionality reduction step makes the method ideal for genomic data.

The PLS weights are vectors that represent the most important patterns relating the predictors and the response, and are found by maximizing the covariance between the two. Given data matrices $X_{N\times p}, Y_{N\times q}$ (whose columns are centered and normalized) and choice of dimension D, PLS finds D iterations of weights $\{w_d, v_d\}_{d=1}^{D}$ by solving (Wold, 1985):

$$w_d, v_d \leftarrow \underset{w,v:\|w\|=\|v\|=1}{\operatorname{argmax}} \operatorname{cov}(Xw, Yv) = w^T X^T Y v. \tag{3.1}$$

We consider here only univariate classification ($q = 1$) in which case the condition $v_d = 1$ always holds and can be omitted.

After each iteration of weight computations, the matrices X, Y are adjusted (or deflated) to avoid overlapping solutions with the previous iteration. Many versions of PLS exist which differ on their methods of deflation (Wegelin, 2000; De Jong, 1993): PLS-mode A, PLS2, PLS-SVD, and SIMPLS. We focus here on the prediction-oriented PLS2, whose deflation step is:

$$X_{\text{new}} \leftarrow X - X w_d w_d^T X^T X / w_d^T X^T X w_d,$$
$$Y_{\text{new}} \leftarrow Y - X w_d w_d^T X^T Y / w_d^T X^T X w_d.$$

Typically the weights (w_d, v_d), scores (ξ_d, η_d), and loadings (γ_d, δ_d) for X, Y respectively:

$$\xi_d = Xw_d, \gamma_d = X^T Xw_d / w_d^T X^T Xw_d,$$
$$\eta_d = Yv_d, \delta_d = Y^T Xw_d / w_d^T X^T Xw_d,$$

are computed and stored as columns in their respective matrices:

$$w_{p\times D}, v_{q\times D}, \xi_{N\times D}, \eta_{N\times D}, \gamma_{p\times D}, \delta_{q\times D},$$

to facilitate calculations. For PLS2, the prediction scores given new data X^* is computed as $\hat{Y} = X^*\beta$, with the coefficient $\beta_{p\times q}$ given by:

$$\beta = (1/\sigma_X) \cdot w(\xi^T \xi)^+ \delta^T \cdot \sigma_Y^T, \tag{3.2}$$

where the column standard deviations $(\sigma_X)_{p\times 1}, (\sigma_Y)_{q\times 1}$ of X, Y serve as normalizing constants, and M^+ denotes the Moore-Penrose pseudoinverse of M (Pedregosa et al., 2011).

For binary classification problems these prediction scores are used to rank new samples, to be later used for classification. This method is known as partial least squares discriminant analysis (PLS-DA) (Pérez-Enciso and Tenenhaus, 2003), which we will focus on in the course of this paper. The PLS solution can also be used as input for other classification techniques such as linear or quadratic discriminant analysis (Nguyen and Rocke, 2002) and logistic regression (Fort and Lambert-Lacroix et al., 2005).

3.2.2 Integrative PLS

Given multiple cohorts of data, two standard approaches are to apply PLS separately on each dataset (and average the resulting prediction scores) or to apply PLS jointly to the combined datasets. We refer to these approaches as separate and joint PLS (sPLS and jPLS). Broadly speaking, the level of differences among the groups of data determines the relative performance of the predictions generated from these approaches. If the data are largely homogeneous, then the joint analysis achieves higher power. If there is substantial heterogeneity in the relationships mapping the predictors to the response, then the separate analysis is less affect by confounding. While nature of these cohort differences is typically unknown, we propose that gains in accuracy can still be achieved from an adaptive model.

The principle of integrative PLS (iPLS) is to consider a spectrum of intermediate models between sPLS and jPLS by evaluating the similarity in weights across groups. Given K cohorts of data pairs $(X_k)_{N_k \times p}, (Y_k)_{N_k \times 1}, k = 1, \ldots, K$ (whose columns are centered and normalized) and a choice of dimension D, we compute D iterations of K weights $\{\{w_d\}_{k=1}^K\}_{d=1}^D$ by solving:

$$\{w_{d,k}\}_{k=1}^K \leftarrow \underset{w_k : \|w_k\|=1}{\operatorname{argmax}} \frac{1}{p} \sum_k w_k^T X_k^T Y_k + \frac{\lambda_C}{K^2} \sum_{k,k'} w_k^T w_{k'}, \tag{3.3}$$

for chosen commonality parameter $\lambda_C \geq 0$. With the correlation computation in the second term, larger λ_C induces commonality across groups. This computation also makes the objective non-convex, and so multiple iterations of our coordinate descent algorithm (discussed later) may be needed to bypass local optima. As λ_C approaches 0 and $+\infty$, the iPLS solution tends towards the solutions of sPLS and jPLS respectively.

3.2.3 Algorithm

Apart from the calculation of the weights, the majority of the iPLS algorithm resembles that of PLS. Given K data pairs $(X_k)_{N_k \times p}, (Y_k)_{N_k \times 1}, k = 1, \ldots, K$ as before, the weights are found by repeating the following update across $k = 1, \ldots, K$ until convergence of the objective (Equation 3.3):

$$w_k \propto \frac{1}{p} X_k^T Y_k + \frac{2\lambda_C}{K^2} \sum_{k' \neq k} w_{k'},$$

where the updated quantity is normalized to unit length. This update maximizes the iPLS objective function with respect to w_k at each iteration. This can be easily seen by rewriting Equation 3.3 as:

$$\widetilde{\mathcal{F}} = \frac{1}{p} \sum_k w_k^T X_k^T Y_k + \frac{2\lambda_C}{K^2} \sum_{k' \neq k} w_k^T w_{k'},$$

which is equivalent due to the unit-length constraints on w_k.

We note that the algorithm appears to perform sufficiently well with only two repetitions, one with initialization at the sPLS solution and the other with initialization at the jPLS solution. As with PLS, the weights $(w_{d,k})$, scores $(\xi_{d,k}, \eta_{d,k})$, and loadings $(\gamma_{d,k}, \delta_{d,k})$ are stored as matrix columns and the coefficients β_k are computed (as shown

in Equation 3.2) for each group k. To make predictions on a new dataset X^*, we take the average of these coefficients and proceed as in standard PLS-DA with $\hat{Y} = X^* \sum_k \beta / K$.

Selection of D, λ_C is performed with 10-fold cross validation with the area under the curve (AUC) of the receiver operating characteristic, stratified with respect to the response variable and the data group membership. The following tuning ranges are suggested: $D = 1$ to 4, $\lambda_C \in \{10^{-2.5}, 10^{-2.25}, \dots, 10^{2.5}\}$. Since large λ_C induces a wide range of iPLS solutions (w_k only need to be close across k to produce large values in Equation 3.3), we improve stability by replacing the iPLS solution with that of jPLS if the largest selection of λ_C (i.e. $10^{2.5}$) is selected.

3.3 Simulation Study

For our simulations, we generated a system of K data pairs $(X_k)_{N_k \times p}, (Y_k)_{N_k \times 1}, k = 1, \dots, K$ according to a basic linear model with normal errors as follows:

$$(X_k)_{ij} \sim \text{i.i.d.} \mathcal{N}(0,1), (\epsilon_k)_{ij} \sim \text{i.i.d.} \mathcal{N}(0,0.1),$$

$$Y_k \sim 1\{X_k \beta_k + \boldsymbol{\epsilon}_k \geq 0\}.$$

The regression coefficients β_k represent the causal structure between predictors and response for each group, with non-zero entries (all of which take unit value) representing the causal signal. We specified that among these non-zero elements, N_C are common and $N_{D,k}$ are distinct across groups. For instance, $N_C = 4, N_{D,1:2} = \{3,1\}$ may produce the following slopes:

$$B_1 = (1 \quad 1 \quad 1 \quad 1 \quad 1 \quad 1 \quad 1 \quad 0 \quad 0 \quad \cdots),$$
$$B_2 = (1 \quad 1 \quad 1 \quad 1 \quad 0 \quad 0 \quad 0 \quad 1 \quad 0 \quad \cdots).$$

We evaluated iPLS against a collection of standard prediction methods consisting of PLS, logistic regression (LR), support vector machines (Vapnik, 1998) with the Gaussian kernel (gSVM), and random forest (Breiman, 2001) (RF), all implemented with the scikit-learn package (Pedregosa et al., 2011) in Python. With the exception of iPLS, all methods were performed with separate and joint approaches, meaning that they were applied to the cohorts independently and as a single merged dataset (denoted with the

prefixes "s-" and "j-" respectively). Parameters for PLS-type methods were chosen as per the recommendations discussed above. AUCs were computed from resulting predictions (or predicted probabilities) and averaged over 25 trials. For each of the separate approaches, the average of the K sets of predictions was taken to be the final prediction, similar to the strategy for iPLS.

To reflect the data example discussed in the next section, we simulated the case of training on three cohorts ($K = 3, N_{1:K} = 150, p = 200$) to make predictions on a fourth cohort ($N_{K+1} = 150, p = 200$). To study performance over all levels of similarity between groups, we varied the amount of overlapping signal N_C among $\{5,10, \dots ,50\}$. Note that the data was generated using $K = 4$ under the method described above so that the fourth dataset can be used for independent evaluation. The number of non-overlapping causal variables dictates the level of heterogeneity among cohorts

We first consider the specifications $N_{D,1:K} = \{150 - N_C, 150 - N_C, 150 - N_C, 0\}$ representing the setting in which the heterogeneity in the causal signal is distributed across training cohorts. As Table 3.1 shows, the top predictors include iPLS, jPLS, sPLS, jLR, and jSVM, with jLR performing slightly better under total commonality between groups ($N_C = 50$, i.e. total signal overlap) and iPLS and sPLS performing slightly better under minimal commonality ($N_C = 5$ or minimal signal overlap). In general, joint applications of methods tended to fare better than separate applications. Random forest performs poorly in this setting most likely because the broad dependency structure of our generated data is difficult to capture via simple trees. Importantly, iPLS maintains competitive accuracy across the range of signal overlap.

Table 3.1: Simulated predictive accuracy (average AUCs) evaluated on independent cohorts, under varying degrees of signal overlap (N_C). Dimensions: $(150 \times 200) * 3$; signal structure: $N_{D,1:K} = \{50 - N_C, 50 - N_C, 50 - N_C, 0\}$.

N_C	iPLS	jPLS	sPLS	jLR	sLR	jSVM	sSVM	jRF	sRF
10	0.69	0.68	0.68	0.66	0.61	0.67	0.6	0.54	0.52
15	0.73	0.72	0.73	0.7	0.66	0.72	0.64	0.55	0.52
20	0.77	0.77	0.77	0.75	0.69	0.77	0.68	0.56	0.53
25	0.8	0.8	0.79	0.78	0.71	0.79	0.69	0.57	0.55
30	0.82	0.83	0.82	0.81	0.73	0.82	0.71	0.58	0.55
35	0.84	0.84	0.83	0.83	0.74	0.83	0.72	0.58	0.55
40	0.86	0.86	0.84	0.86	0.75	0.85	0.73	0.57	0.55
45	0.89	0.89	0.86	0.89	0.78	0.87	0.76	0.59	0.56
50	0.9	0.9	0.87	0.92	0.79	0.88	0.77	0.6	0.58

Table 3.2 shows predictive performance under $N_{D,1:K} = \{150 - N_C, 0,0,0\}$ representing the setting in which there is an uneven distribution of heterogeneity in the causal signal due to one problematic cohort. Here, the advantage of iPLS over others is more pronounced under low levels of signal overlap. Since the standard joint and separate applications of predictors do not take into account the level of discrepancy between groups, their approach to handling the noise in the first cohort amounts to distributing it across the analysis. On the other hand, iPLS features a more adaptive approach in which the computation of its solution (w_k) is intimately tied with how the solution differs across groups ($\| \sum_k w_k \|^2$). By searching among a spectrum of intermediate models of varying levels of discrepancy among groups, the method seeks to identify the best performing model to maximize robustness. Importantly, iPLS achieves such adaptivity under a relatively simplistic framework and algorithm, which is precisely advantageous for the noisy and heterogeneous genomics data environments.

Table 3.2: Simulated predictive accuracy (average AUCs) evaluated on independent cohorts, under varying degrees of signal overlap (N_C). Dimensions: $(150 \times 200) * 3$; signal structure: $N_{D,1:K} = \{50 - N_C, 0,0,0\}$.

N_C	iPLS	jPLS	sPLS	jLR	sLR	jSVM	sSVM	jRF	sRF
10	0.85	0.83	0.82	0.81	0.79	0.82	0.77	0.63	0.58
15	0.86	0.85	0.83	0.84	0.79	0.83	0.77	0.61	0.58
20	0.86	0.86	0.84	0.86	0.79	0.85	0.76	0.6	0.56
25	0.88	0.88	0.85	0.87	0.79	0.86	0.77	0.61	0.58
30	0.88	0.88	0.86	0.88	0.8	0.86	0.77	0.6	0.58
35	0.89	0.89	0.87	0.89	0.8	0.87	0.77	0.6	0.56
40	0.89	0.89	0.87	0.9	0.79	0.88	0.77	0.6	0.57
45	0.9	0.9	0.87	0.91	0.79	0.88	0.77	0.58	0.57
50	0.9	0.9	0.87	0.92	0.79	0.88	0.77	0.6	0.58

3.4 Application to Constructing Robust Prognostic Genetic Signatures

3.4.1 Data Preparation and Preprocessing

We obtained gene expression and clinical data from a total of 814 breast cancer samples from four cohorts: Sotiriou et al. (2006), Wang et al. (2005), Ivshina et al. (2006), and Pawitan (2005). The datasets are available from their journal articles and the Gene Expression Omnibus database (http://www.ncbi.nlm.nih.gov/geo) with accession numbers GSE2990, GSE2034, GSE4922, and GSE1456. Samples were selected based on availability of clinical information (event of relapse and ER status). Only probe sets without missing values and present in all four cohorts were used. Transformations were applied to convert the probe intensities (all measured on the Affymetrix Human Genome U133A Array) to a log2 scale. Values from multiple probe sets mapping to the same gene were combined by taking the median.

We conducted separate analyses for ER positive and ER negative patients, as the molecular distinction between the two diseases has been well-established (Gruvberger et al., 2001; Sørlie et al., 2003; Hu et al., 2006). We considered in our analysis the top 200 genes with the highest absolute correlation with the response (event of relapse). While such a large quantity of genes may not necessarily optimize predictions, it allows for the inclusion of a broad range of genes to ameliorate the issue of volatile identifications of gene sets (Ein-Dor et al., 2005). The final dimensions of the datasets were $(147 + 209 + 211 + 62) \times 200$ for both ER positive and ER negative cancer samples.

3.4.2 Prediction on Independent Cohorts

We evaluated the same methods considered in our simulation study for predicting the event of relapse using gene expression profiles. Parameters for iPLS and PLS were chosen with cross validation as previously described. We note that the patients among these cohorts were heterogeneous in various aspects such as the type of treatment received (e.g. radiotherapy, systemic therapy) and duration of the follow-up. Nevertheless cancer recurrence serves as an important measure of disease severity which can guide treatment strategies. To reproduce the effect of applying learned genomic

signatures on new samples, we trained on three of the cohorts ($K = 3$) to make predictions on the remaining cohort. This was repeated for each combination of cohorts and each ER status for a total of eight evaluations.

Figures 3.1a and 3.1b shows the predictive accuracies for each method under each of the training-testing combinations. In general, the most accurate predictions occurred on the Pawitan dataset, and the least accurate on the Wang dataset. This is indicative of the heterogeneity between the cohorts involving the samples and methodology. For instance, samples from the Wang cohort originated from the Netherlands while samples of the other cohorts originated from primarily Sweden. Also, the Ivshina and Pawitan studies both employed the global mean method for normalizing raw expression data, while the Sotiriou and Wang studies each applied different normalization techniques. In general, predictions among the ER positive group were more favorable, which reflects both the limited availability and molecular diversity of ER negative breast cancer samples.

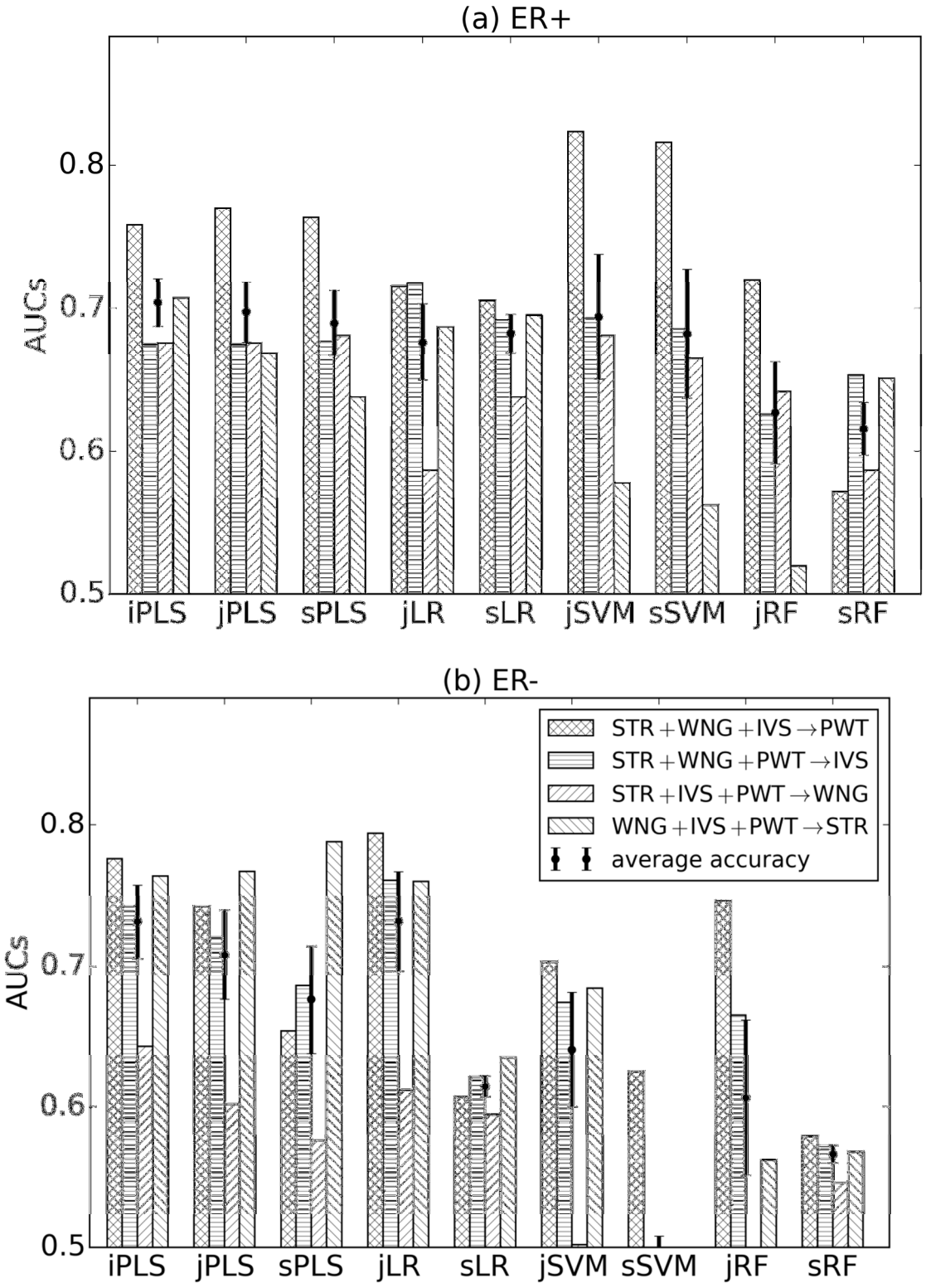

Figure 3.1: Predictive accuracy (AUCs) in predicting cancer recurrence from gene expression profiles among the ER positive (a) and ER negative (b) samples. Each dataset among the Sotiriou (STR), Wang (WNG), Ivshina (IVS), and Pawitan (PWT) cohorts was considered for independent testing. Error bars denote the standard error of the average AUCs across combinations.

On average, iPLS was the most accurate in predicting cancer relapse for both the ER statuses, as well as the most stable among the top predictors. Other methods (e.g. logistic regression and support vector machines) had at most competitive accuracy for only a fraction of cohort combinations that was not sustained for other combinations. Overall the instability in prediction can be attributed to the lack of adaptivity to cohort differences, a tendency to overfit, and difficulty in modeling the complex dependency structure of genes. Importantly, we find that iPLS consistently performs at least as well as both the joint and separate applications of PLS, and significantly better when the common signal among cohorts is overshadowed by cohort discrepancies, as it appears to be the case for the ER negative sample.

3.4.3 Follow-up Analysis of Signatures

Although iPLS was designed for making robust predictions, we may also investigate its output to identify the most important contributing variables. The average of the iPLS weights w_k (the initial layer before deflation) across groups can be used to rank variables based on absolute magnitude. The same ranking can be done for jPLS and sPLS, although due to an identifiability issue the sPLS predictor weights must be manually oriented to have positive pairwise correlation. This is not a concern for iPLS as its optimization objective naturally induces the proper orientation. For logistic regression and random forest, we ranked the variables based on absolute magnitude in the regression coefficient and variable importance respectively, and applied similar averaging for the separate approaches. We could not obtain such gene signatures from support vector machines due to its non-linear kernel.

The top 10 genes identified by each prediction method from the best performing cohort combination (testing on the Pawitan dataset) are shown in Tables B1 and B2 in Appendix B1. Even when considering all methods, only a fraction of these top genes overlap with existing signatures such as a recurrence-predicting signature by Chanrion et al. (2008) (TPX2, PRC1, RRM2, CDK1, ASPM, MMP1), a cell proliferation-based signature by Dai et al., (2005) (PRC1, SNRPA1), and a consensus signature derived from meta-analysis by Teschendorff et al. (2006) (RACGAP1, ZWINT, CDCA8). This

discrepancy reflects the highly variable nature of gene signature contents (Ein-Dor et al., 2005), although it should be noted that our total of 200 considered genes overlapped with only 40-50\% of each of these three signatures. There also appears to be better concordance with existing signatures among the ER positive group, which is expected due to the relative scarcity of ER negative breast cancers and the tendency of studies to base signatures on primarily or entirely ER positive samples.

From Table B1 of Appendix B1, we can see that the most concordant signatures are produced by methods that are generally better predictors (i.e. PLS-type methods). However, predictive performance often differs despite similar gene rankings, as seen among the PLS-type methods. In fact, as Table B3 of Appendix B1 shows, Methods with similarly strong prediction can have very similar regression coefficients (iPLS and jPLS) or very different ones (iPLS and sPLS). While predictive and inferential accuracies are somewhat correlated, it appears that a good prognosis signature is not restricted to a stable set of members, which is attributable in part to the high dependency among the genetic drivers of the disease.

From the gene ontology annotations, the signature developed from applying iPLS to the ER positive data (testing on the Pawitan cohort) primarily contain genes involved in cell division/mitosis (RACGAP1, PRC1, AURKA, NCAPG, KIF11, ASPM) and DNA replication (RRM2, GINS2, TOP2A). RACGAP1 and TOP2A are two proliferation markers linked to early recurrence in luminal breast cancers (Milde-Langosch et al., 2013). High expression of RACGAP1 is associated with poor survival (Pliarchopoulou et al., 2013), while TOP2A appears to be predictive of responsiveness to treatment (Villman et al., 2006). PRC1 and RRM2 are both potential therapeutic targets overexpressed in breast cancer and associated with tumor growth (Yun et al., 2008; Shimo et al., 2007; Putluri et al., 2014). AURKA has been linked via a Wnt signaling pathway to metastatic spread and recurrence (Eterno et al., 2016), and PBK has been shown to facilitate tumor growth via DNA damage response (Ayllon and O'Connor, 2007). GINS2 is a recently discovered independent prognostic marker for relapse-free and metastasis-free survivals that is implicated in cell growth and aneuploidy (Zheng et al., 2014; Rantala et al., 2010). The majority of genes identified in this group are associated with cancer cell proliferation.

For the ER negative group, the iPLS signature included genes with functions in lipid biosynthesis/metabolism (PTPLB, IDI1, CPT1A) and apoptosis regulation (MST4, HSPB1). In fact, ER negative breast cancer cells have been found to undergo structural reorganization of lipid rafts during transition to invasiveness (Ostapkowicz et al., 2006), and recently lipid metabolism-related proteins have been implicated in the risk of ER negativity (Kim et al., 2015). NUTF2 has been proposed as a strong prognostic factor that promotes cell proliferation and is repressed by estrogen (Oka et al., 2011). PTPLB is another potential independent prognostic factor which correlates with ER status, although its role in breast cancer is yet to be fully determined (Soysal et al., 2013). MST4 (along with MST3) is linked to tumor invasiveness via a STRIPAK complex-regulated promotion of cell migration and metastasis (Madsen et al., 2015). NDRG1 encodes a protein associated with breast cancer cell differentiation and metastasis suppression (Fotovati et al., 2011; Ellen et al., 2008). Recently, CPT1A has been identified as a potential anti-cancer therapeutic target, whose transcript variant induces anti-apoptosis and tumor invasiveness in place of its original function of fatty acid transportation (Pucci et al., 2016). Although this ER negative signature contains fewer overlapping members with existing signatures, many of the genes identified represent fairly recent discoveries.

Notably, the weights produced by jPLS and sPLS are precisely the covariances between the predictors (gene expression) and the response (cancer recurrence) for the datasets studied concurrently and separately respectively. Meanwhile, iPLS weights lie between these two extremes as an adaptive intermediate, which we have seen produces better overall predictions. In principle, we should expect iPLS to generate more concordant signatures, however there was little change in the rankings of prominent genes. Although our results produced from iPLS appear to align with existing signatures no worse than the other binary classifiers for our data, there is still much work to be done to arrive at a stable set of genetic markers for breast cancer severity.

3.5 Discussion

Since the number of reliable samples in genomics data is typically fewer than the number of measured variables, combining multiple cohorts is often considered as a tool

for improving the statistical strength and generalizability of predictive signatures. However, the common signal of interest that is amplified in doing so is inevitably accompanied by cohort discrepancies that are unknown in nature. We have presented here a predictive model for gauging and adapting to the extent of these discrepancies to arrive at more robust performance. By augmenting the foundation of the PLS dimensionality reduction technique with a simple adjustment and algorithm, we are able to choose among a continuum of models. With artificial and real data, we have demonstrated that such an adaptive approach can offer improvements over a purely joint or purely separate analysis.

Normalization is an important consideration for studying multiple sources of data. In iPLS, we center and normalize the datasets as a single block in order to apply a common adjustment to the data that avoids undue bias. This is one of the simplest approaches, and gives consideration to each data group proportional to its sample size. An alternative is to center and normalize within each dataset, but we felt that doing so in the setting of genomic expression data would produce less stable and more biased adjustments. Additional options are to adjust weights according to relative confidence in the information of each cohort or similarities between subsets of data groups. These strategies are easily implementable prior to the main iPLS algorithm to achieve optimal predictions.

Accounting for the degree of cohort differences allows iPLS to somewhat reduce the confounding effects of combining multiple studies, although it does not completely eliminate these effects. Careful consideration of batch effects and other confounders is still recommended, as the adaptations of iPLS handle a specific (albeit basic) type of heterogeneity among groups. As with other strategies of analysis, a more diverse selection of cohort provides additional benefits to the robustness of model performance in terms of both prediction and inference.

CHAPTER IV

An ANOVA-based Procedure for PCA, Decomposing Variation and Dimensionality

4.1 Introduction

In the modern age of information, the availability of biological data (particularly genomic data) far exceeds our ability to process it. The wealth of information poses an analytic problem in terms of both volume and variety, as biological expression is not only typically high-dimensional but also notoriously heterogeneous (Marx, 2013). While there exist techniques for operating under large p small n settings (Amini et al., 2008; Carvalho et al., 2008; Shao et al., 2011), the sparsity consideration alone does not address the fact that biological systems are complex in nature and that their observed patterns are both volatile across sample groups and diverse across variable types. For this reason, the integration of data from multiple sources (Wei, 2015) is perhaps essential in providing a global systems-level perspective and a basis for robust and reproducible results. In this work, we devise a method of data integration from the foundational perspective of dimensionality reduction, which aims to explore the consolidation of data patterns across groups in terms of its fundamental components.

Dimensionality reduction is commonly used as a first step in navigating the complex landscape of biological expression data (Yeung and Ruzzo, 2001; Dai et al., 2006; Teodoro et al., 2003; Das et al., 2006). Its central goal is to reduce data patterns into relatively few basic components or dimensions of highest importance. The definition of this "importance" will vary depending on the structure of the data and problem, ranging from the accuracy in reconstructing the data from nonnegative (Lee and Seung, 2001) or orthogonal (Wold et al., 1987) parts, to the ability to preserve local neighborhoods (Roweis and Saul, 2000) or pairwise distances (Kruskal, 1964) in low-dimensional manifolds. In any case, a simplified view is valuable (and oftentimes

necessary) for deriving meaningful insight from the observed patterns of many interconnected biological entities and processes. In the face of prevalent heterogeneity in biological data, we aim to translate this value to the analysis of multiple data groups.

Reducing dimensionality across multiple data groups presents the unique option of comparing these groups across their reduced components. Since these components represent the key elements among the datasets, they provide a novel and potentially more efficient means of relating the patterns among them, such as in terms of the similarity and complexity among groups. This has broad applicability in any study or experiment involving grouped or stratified data such as subtypes of diseases, subpopulations, or even different experimental conditions. Just as traditional dimensionality reduction reduces the view of single datasets, the integrative variant we propose reduces the comparison of multiple datasets.

There are currently a handful of approaches developed for adapting dimensionality reduction to multiple groups. Multi-block partial least squares (Li et al., 2012) applies weighted averaging on predictor variable scores to identify multi-dimensional regulatory modules most highly associated with the response. Multiple canonical correlation analysis (Witten and Tibshirani, 2009) combines simultaneous covariance calculations between all dataset pairs to extract highly correlated linear combinations of variables across groups. Integrative nonnegative matrix factorization (Yang and Michailidis, 2015) distinguishes between and relates the strengths of common and distinct effects to distill homogeneous modules from heterogeneous noise. In each case the defining objective of the original reduction technique (maximizing covariance, maximizing correlation, minimizing residual error, etc.) is preserved while some modification (e.g. a penalty or constraint) is incorporated to offset the additional model flexibility. We apply the same strategy for expanding the classical principal component analysis (PCA) to multiple datasets, a method known for its efficiency in explaining data patterns via variance-maximizing components.

Our approach consists of two stages, the first of which borrows principles from analysis of variance to construct an overarching framework that relates variation between and within data groups. This involves viewing PCA from the alternative perspective of residual minimization rather than variance maximization, and produces novel estimates

for the commonality and complexity shared among the data groups. The next stage uses this estimate to formulate hypothesis-based inference on the patterns common and distinct across datasets. We show how this inference can be applied to simply and shape the analysis procedure for a breast cancer cell line study with factorial design.

4.2 Methods

4.2.1 Principal Component Analysis

Principal components analysis (PCA) is a popular dimensionality reduction tool used in virtually every field that handles high-dimensional data including bioinformatics (Price et al., 2006; Zheng et al., 2012), economics (Vyas and Kumaranayake, 2016; Olawale and Garwe, 2010), and computer vision (De la Torre and Black, 2001; Kim et al., 2002). The method essentially deconstructs the observed data into of a few basic layers that are defined by explaining maximum information. Formally, given column-centered matrix $X_{N\times p}$, the goal is to find an orthogonal linear transformation of X in lower dimensions $D << p$ that retains maximal variance (Hastie et al., 2009):

$$\hat{W} = \underset{W^T W = I_D}{\operatorname{argmax}} \| XW \|_F^2,$$

where $\|\cdot\|_F$ denotes the Frobenius norm.

The loading matrix $\hat{W}_{p\times D}$ defines the mapping which transforms the data X to the lower dimensional space, whereas the score matrix $(\hat{XW})_{N\times D}$ is the mapped data whose variance is maximized. The model can be interpreted as reducing the observed patterns of the data into a set of observational latent factors (scores, i.e. rows of $\hat{XW}$) and key variable signatures (loadings, i.e. columns of $\hat{W}$). The solution is uniquely defined as is typically found via singular value decomposition (SVD):

$$X = U\Sigma V^T,$$

where the top D right singular vectors (in V) are taken as the loadings $\hat{W}$. Note that this guarantees that the solutions are nested, i.e. incrementing the dimension D leads to larger sets of loadings and scores which include the previous loadings and scores.

The product of the scores and loadings is a low-rank approximation of the data consisting of D rank-1 layers, which are often referred to as principal components:

$$\hat{X}\hat{W}\hat{W}^T = \sum_{k=1}^{K} \hat{X}\hat{W}_{\cdot d}\hat{W}_{\cdot d}^T.$$

We will be considering an alternative but equivalent formulation of PCA (Zou et al. 2006), which involves minimizing the sum-of-squares of the residuals of this PCA approximation:

$$\hat{W} = \underset{W^TW=I_D}{\operatorname{argmin}} \| X - XWW^T \|_F^2.$$

4.2.2 The ANOVA Decomposition for PCA

In classical one-way ANOVA (Cox, 2006), the variation of univariate data $Y \in \mathbb{R}^N$ decomposes into the variation within and between groups:

$$\sum_{k,i} (Y_{ki} - \overline{Y})^2 = \sum_{k,i} (Y_{ki} - \overline{Y}_k)^2 + \sum_k N_k (\overline{Y}_k - \overline{Y})^2,$$

where Y_{ki} represents each sample and $\overline{Y}_k, \overline{Y}$ denote the group and grand means. Note that the center in each sum-of-squares term is a sample mean of some form. Our approach to extending PCA involves constructing a similar decomposition for multivariate data $X \in \mathbb{R}^{N \times p}$ in which the centers for the sum-of-squares terms are PCA approximations $\hat{X}\hat{W}\hat{W}^T$.

Suppose that we have K observation groups of column-centered datasets $X_k \in \mathbb{R}^{N_k \times p}, k = 1, \ldots, K$ (letting $N = \sum_k N_k$) normalized as $X_k / \| X_k \|_F^2$. Consider performing PCA on the full data $X_f = (X_k)_{k=1}^K \in \mathbb{R}^{N \times p}$ jointly and each individual dataset X_k separately, which entails solving:

$$\hat{W}_f = \underset{W^TW=I_D}{\operatorname{argmin}} \| X_f - X_f WW^T \|_F^2, \hat{W}_k = \underset{W^TW=I_D}{\operatorname{argmin}} \| X_k - X_k WW^T \|_F^2.$$

At the uniquely defined optima, the PCA approximation $X_f \hat{W}_f \hat{W}_f^T$ becomes the rank-D center of the sum-of-squares term associated with X_f, and each PCA approximation

$\hat{X}_k\hat{W}_k\hat{W}_k^T$ becomes the rank-D center of the sum-of-squares term associated with its corresponding X_k. It is easy to notice that these sum-of-squares terms:

$$\text{TSS} = \| X_f - \hat{X}_f\hat{W}_f\hat{W}_f^T \|_F^2 \text{ ,WSS} = \sum_k \| X_k - \hat{X}_k\hat{W}_k\hat{W}_k^T \|_F^2. \quad (4.1)$$

can play an analogous role to the total and within group sum-of-squares of ANOVA.

For the between-group sum-of-squares, the classical ANOVA approach is to compute the sum-of-squares of the group means $\overline{Y}_k$ with the center being the grand mean $\overline{Y}$, computed from taking the mean of the group means. Similarly, we compute the sum-of-squares of the separate PCA approximations $\hat{X}_k\hat{W}_k\hat{W}_k^T$ with the center obtained from applying PCA to these PCA approximations. Thus we combine the separate approximations as a single matrix $\hat{X}_s = (\hat{X}_k\hat{W}_k\hat{W}_k^T)_{k=1}^K \in \mathbb{R}^{N\times p}$ and reapply PCA, which means solving:

$$\hat{W}_s = \underset{W^TW=I_D}{\operatorname{argmin}} \| \hat{X}_s - \hat{X}_s WW^T \|_F^2.$$

At the unique optimum, the PCA approximation $\hat{X}_s\hat{W}_s\hat{W}_s^T$ becomes the rank-D center of the sum-of-squares term associated with X_s:

$$\text{BSS} = \| \hat{X}_s - \hat{X}_s\hat{W}_s\hat{W}_s^T \|_F^2. \quad (4.2)$$

Combing the terms in Equations (4.1) and (4.2), we have the PCA analog to the ANOVA decomposition of sum-of-squares:

$$\| X_f - \hat{X}_f\hat{W}_f\hat{W}_f^T \|_F^2 \approx \sum_k \| X_k - \hat{X}_k\hat{W}_k\hat{W}_k^T \|_F^2 + \| \hat{X}_s - \hat{X}_s\hat{W}_s\hat{W}_s^T \|_F^2. \quad (4.3)$$

A schematic is given in Figure 4.1, which summarizes the entire procedure as sequential applications of PCA. The decomposition in Equation (4.3) can be viewed as the (approximate) equivalence of two paths for arriving at a rank-D approximation of the complete data. One path is to perform rank-D PCA on the joint data X_f. The other is to perform rank-D PCA separately on each X_k and reapply rank-D PCA on the combined approximations X_s. With general random matrices, Equation (4.3) appears to hold approximately, with small discrepancy due to the fact that data columns are centered

differently depending on whether PCA is applied jointly or separately. Consequently, $\hat{W}_f, \hat{W}_s$ are in general not identical.

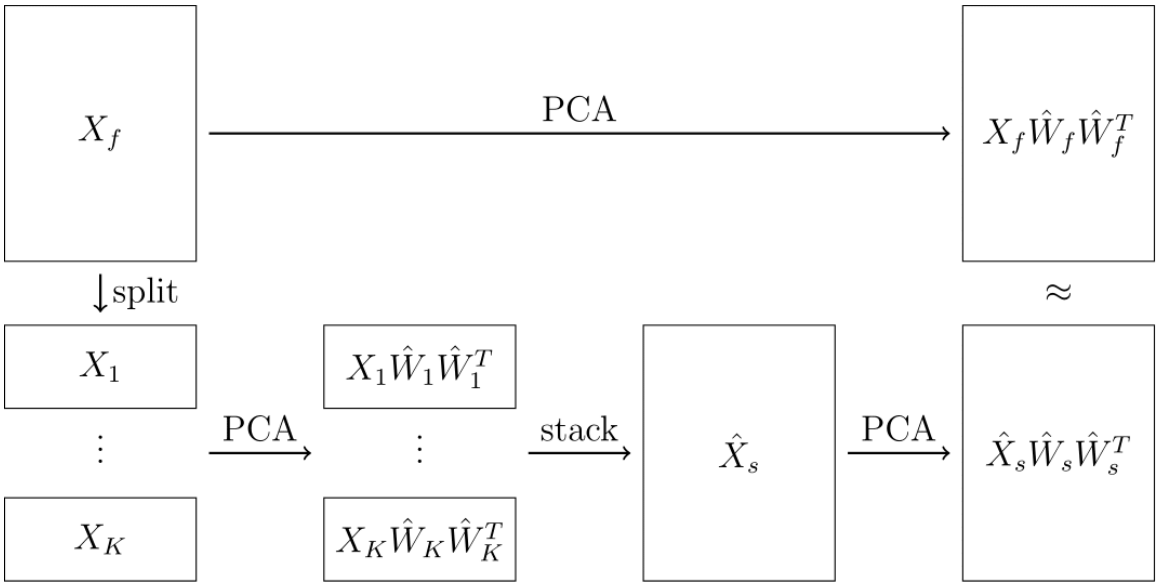

Figure 4.1: Schematic of the gPCA procedure: the multivariate extension of the ANOVA decomposition of sum-of-squares can be viewed as the (approximate) equivalence of two paths for arriving at a rank-D approximation of the complete data. One path is to perform rank-D PCA on the joint data X_f. The other is to perform rank-D PCA separately on each X_k and reapply rank-D PCA on the combined approximations X_s.

Using this ANOVA-based framework, we can arrive at estimates for the commonality ($\hat{\alpha}$), noise level ($\hat{\sigma}$), and common complexity ($\hat{D}$) associated with the collective data groups:

$$\hat{\alpha}_d = \frac{2}{K(K-1)} \sum_{k>k'} \sqrt{s_{\min}(\hat{W}_{k;\cdot d}^T \hat{W}_{k';\cdot d})},$$

$$\hat{\sigma}_d^2 = \frac{\mathrm{WSS}}{\mathrm{BSS}} \frac{d(K-1)(1-\hat{\alpha}_d^2)}{K(p-d)},$$

$$\hat{D} = \underset{d=1,\ldots,D_{\max}}{\operatorname{argmin}} \hat{\sigma}_d,$$

where $s_{\min}(\cdot)$ denotes the smallest singular value of a matrix and $\hat{W}_{k;\cdot,d}$ represents the empirical loading generated from the top d singular vectors. The criterion $\hat{\sigma}_d$ produces a rank estimate $\hat{D}$ which is used to evaluate $\hat{\sigma}_d, \hat{\alpha}_d$ to obtain $\hat{\sigma}, \hat{\alpha}$.

The full derivation is provided in Appendix C.2 and C.3, and involves applying properties of principal angles. As it is evident from this derivation, $\alpha \in [0,1]$ represents the proportion of common signal (in terms of loadings) present among the data groups, assuming that the data is generated from a mixture of mutually orthogonal common and distinct components. Meanwhile $\sigma > 0$ represents the noise-to-signal level associated within each dataset. Finally, D is the common rank of all PCA approximations used in the decomposition. To improve stability, we can apply non-parametric bootstrapping (100 bootstrap samples) and take the average values as our estimates $\hat{\alpha}_d$ and $\hat{\sigma}_d$.

4.2.3 Groupwise PCA

While $\hat{D}, \hat{\sigma}$ provide useful insight about the structure and strength of the signal across data groups, they primarily serve to set an appropriate stage for the computation of $\hat{\alpha}$. The entire procedure, which we refer to as groupwise PCA (gPCA), consists of estimating α and calculating significance via two complementary approaches. We let $p_1(a), p_2(a)$ denote the p-values associated with observing $\hat{\alpha}$ at least as large as α and at least as small as α respectively.

For $p_1(a) = P(\hat{\alpha} \geq a)$, we invoke linear algebra theory (discussed in Appendix C.4) to obtain a probability distribution for $\hat{\alpha}$. The probability of observing $\hat{\alpha}$ at least as large as α is:

$$p_1(a) = \frac{\Gamma(\frac{D+1}{2})\Gamma(\frac{p-D+1}{2})}{\Gamma(\frac{1}{2})\Gamma(\frac{p+1}{2})}(1-a^4)^{\frac{D(p-D)}{2}}F_{2,1}(\frac{p-D}{2},\frac{1}{2};\frac{p+1}{2};(1-a^4)I_D),$$

$$\text{for } a \in (0,1].$$

where $\Gamma(z) = \int_0^\infty t^{z-1}e^{-t}dt$ is the gamma function and $F_{2,1}$ is the Gaussian hypergeometric function of matrix argument. For large p, we can approximate $\Gamma(\frac{p-D+1}{2})/\Gamma(\frac{p+1}{2})$ with Sterling's formula:

$$\Gamma(\frac{p-D+1}{2})/\Gamma(\frac{p+1}{2}) \approx (\frac{2}{p-1})^{D/2}.$$

The null hypothesis associated with this p-value is that originating subspaces are independently and uniformly generated on Grassmann manifold of D-planes in $\mathbb{R}^p$ (with $D < \frac{p+1}{2}$), which is the space of all D-dimensional linear subspaces of $\mathbb{R}^p$ (Absil et al., 2006).

$$H_{0,p_1}: \mathcal{W}_1, \dots, \mathcal{W}_K \sim \text{i.i.d. Grass}(D, p).$$

This is not the same as assuming $\alpha = 0$ (i.e. that these subspaces are orthogonal).

For $p_2(a) = P(\hat{\alpha} \leq a)$, we use a different approach based on bootstrapping. For $k = 1, \dots, K$, we reproduce new data groups of the same dimensions by sampling (with replacement) within the k-th data group only (1000 repetitions for each k). This simulates data with homogeneous distribution across groups, from which we can apply gPCA to compute $\hat{\alpha}$ under the null hypothesis of $\alpha = 1$.

$$H_{0,p_2}: \alpha = 1.$$

The proportion of simulated $\hat{\alpha}$ which is at least as large as α is taken to be our p-value:

$$p_2(a) = \#\{\hat{\alpha}_{\text{boot}} \leq a\}/N_{\text{boot}}.$$

With our two measures of significance, we are equipped to apply both sides of hypothesis testing to our estimate of α.

4.2.4 Connections with JIVE

Recently, Lock et al. (2013) proposed Joint and Individual Variation Explained (JIVE), an integrative model which also adapts principles from PCA to decompose multiple datasets into joint ($J = (J_1^T \quad \cdots \quad J_K^T)^T$) and individual ($A_k$) structures:

$$X_1 = J_1 + A_1 + \epsilon_1,$$
$$\vdots$$
$$X_K = J_K + A_K + \epsilon_K.$$

These structures are restricted to be of certain ranks, and their rows are constrained to be orthogonal.

$$\text{rank}(J) = r, \text{rank}(A_k) = r_k, JA_k^T = \mathbf{0}_{N \times N_k}.$$

The solution is found via an iterative algorithm that alternates between estimating J and $A_1, \ldots, A_K$, the result of which can be factored into scores and loadings:

$$J_k = U_{c,k} W_c^T, A_k = U_{d,k} W_{d,k}^T, k = 1, \ldots, K.$$

The method relies on choosing the joint (r) and individual r_k ranks beforehand via a permutation testing procedure that generates null singular value distributions from disrupted joint or individual structures and compares singular values.

The difference between gPCA and JIVE is evident in their model structures. Both methods view the observed data X_k as the sum of common and distinct (JIVE refers to these as joint and individual) components with orthogonal rows, i.e. $W_c^T W_{d,k} = W_{d,k}^T W_{d,k'} = 0$ for all $k \neq k'$. However, gPCA scores remain the same across these components of the same group, whereas JIVE scores are free to vary.

$$X_k = U_{c,k} W_c^T + U_{c,k} W_{d,k}^T \text{ (gPCA)}, X_k = U_{c,k} W_c^T + U_{d,k} W_{d,k}^T \text{ (JIVE)}.$$

This means that gPCA has a stricter definition of group differences (perturbations of the loadings) than that of JIVE (perturbations of the data). This also means that the common rank D of gPCA should be interpreted differently than the joint rank r of JIVE. The former represents the estimated collective rank of the data groups adjusted for group-specific discrepancies in the variable signatures, while the latter represents the collective rank adjusted for group-specific orthogonal shifts in the data.

The JIVE framework is worth mentioning here because it is very similar in spirit and structure to that of gPCA, but ultimately it addresses the separate analytic question of explaining groups of data in terms of orthogonal joint and individual components. The estimation of the complexity of the joint component is accomplished with a permutation testing procedure comparing singular values, which is not based on the JIVE model. In constrast, gPCA focuses on estimating the common complexity and commonality, providing a basis for any subsequent assessment of components.

4.3 Simulation Study

We provide here numerical results for the estimation of D and α as well as associated p-values. Data was generated under the model provided in Appendix C.2. In

short, we assume datasets $X_k \in \mathbb{R}^{N_k \times p}$ (letting $N = \sum_k N_k$) to be generated under a basic factor model:

$$X_k = Z_k W_k^T + E_k, (Z_k)_{i\cdot} \sim \text{i.i.d.}\mathcal{N}(0, I_p), (E_k)_{i\cdot} = \text{i.i.d.}\mathcal{N}(0, \sigma^2 I_p),$$

with true loadings W_k composed of mixtures of common (W_c) and distinct ($W_{d,k}$) parts:

$$W_k = \alpha W_c + \sqrt{1-\alpha^2} W_{d,k},$$

$$W_c^T = (I_D \quad \mathbf{0}_{D\times(p-D)}), W_{d,k}^T = (\mathbf{0}_{D\times kD} \quad I_D \quad \mathbf{0}_{D\times(p-(k+1)D)}).$$

The data lie primarily on the D-dimensional subspaces $\mathcal{W}_k$ generated from W_k. The parameters α and σ denote the levels of commonality between groups and noise within groups respectively.

We first evaluated the rate of gPCA correctly selecting the true common rank D, and compared with conventional PCA rank selection methods. These include the Laplace method (Minka, 2000) and the Bayesian information criterion (Kass and Raftery, 1995), which focus on maximizing evidence under a Bayesian framework. The Kaiser-Guttman method refers to the classical approach of using the average of eigenvalues as stopping threshold (Jackson, 1993), which is comparable to gauging the "elbow" point of the eigenvalue scree plot. Kritchman and Nadler (2008) apply principles from random matrix theory to a sequential hypothesis testing procedure. All methods were implemented in Python, with support from the scikit-learn package (Pedregosa et al., 2011). JIVE was also considered using its Matlab implementation (Lock et al., 2013). The dimensions were chosen in part to resemble the data application in the next section as well as to highlight stable performance even under small samples. Results were averaged over D with 25 repetitions each for a total of 100 repetitions for each scenario.

Tables 4.1 and 4.2 show the rates of correctly identifying the common rank D using gPCA, JIVE, and conventional PCA rank selectors. The joint rank r of JIVE was evaluated for its accuracy in selecting the common rank D. As expected, gPCA drastically outperforms all others in selecting a quantity for which it was developed, simply because other model frameworks are incongruous to our heterogeneously generated datasets. Even JIVE which was designed for extending PCA to multiple datasets is largely inaccurate except in the homogeneous setting. This is due to the slight discrepancies in the model definitions and interpretations for the joint and common ranks

as previously discussed. For the purpose of supporting further inference regarding group commonality (α), the gPCA common rank selector appears to provide the most reliable estimate of common complexity.

Table 4.1: Accuracy rates for selecting D using various integrative and non-integrative methods across $\alpha \in \{0.0, 0.1, \ldots, 1.0\}$. Results are averaged over D with 25 repetitions each (100 in total). Methods: gPCA, JIVE, Bayesian information criterion (BIC), Laplace method (LP), Kaiser-Guttman method (KG), Kritchman and Nadler's method (KN). Specifications: $\{N_k, p, K, \sigma\} = \{13,16,3,0.1\}$ and $D \in \{1,2,3,4\}$, $D_{\max} = 4$.

	0.0	0.1	0.2	0.3	0.4	0.5	0.6	0.7	0.8	0.9	1.0
gPCA	0.97	0.97	0.97	0.98	1.0	1.0	1.0	1.0	1.0	1.0	0.99
JIVE	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.38	0.51	0.75
BIC	0.01	0.01	0.01	0.01	0.01	0.05	0.09	0.17	0.46	0.64	1.0
LP	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.13	0.36	0.75	1.0
KG	0.25	0.25	0.25	0.25	0.25	0.25	0.25	0.25	0.35	0.76	1.0
KN	0.0	0.0	0.01	0.02	0.04	0.09	0.17	0.23	0.25	0.26	0.97

Table 4.2: Accuracy rates for selecting D using various integrative and non-integrative methods across $\alpha \in \{0.0, 0.1, \ldots, 1.0\}$. Results are averaged over D with 25 repetitions each (100 in total). Methods: gPCA, JIVE, Bayesian information criterion (BIC), Laplace method (LP), Kaiser-Guttman method (KG), Kritchman and Nadler's method (KN). Specifications: $\{N_k, p, K, \sigma\} = \{39,16,2,0.1\}$ and $D \in \{1,2,3,4\}$, $D_{\max} = 4$.

	0.0	0.1	0.2	0.3	0.4	0.5	0.6	0.7	0.8	0.9	1.0
gPCA	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0
JIVE	0.01	0.01	0.02	0.03	0.04	0.04	0.04	0.04	0.26	0.45	0.66
BIC	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.07	0.71	1.0
LP	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.15	0.66	1.0
KG	0.25	0.25	0.25	0.25	0.25	0.25	0.25	0.25	0.26	0.74	1.0
KN	0.25	0.25	0.25	0.25	0.25	0.25	0.25	0.25	0.25	0.25	0.99

Tables 4.3 and 4.4 display the distributions of $\hat{\alpha}, p_1, p_2$ across a range of true commonality levels. The estimate $\hat{\alpha}$ appears to be accurate, except at low levels of α where the common signal is comparable in strength to the noise. Conveniently, the p-values p_1, p_2 appear to shrink towards zero as α approaches above and below roughly 0.8 respectively. This makes p_1, p_2 very appropriate for determining whether data groups can be characterized by total commonality. As we demonstrate in the next section, this is useful not only for deciding whether groups can be combined for analytic purposes, but also for conducting multivariate comparisons analogous to the univariate ANOVA comparisons.

Table 4.3: Distributions of $\hat{\alpha}, p_1, p_2$ from gPCA averaged over D with 25 repetitions each (100 in total) across $\alpha \in \{0.0, 0.1, \ldots, 1.0\}$. Specifications: $\{N_k, p, K, \sigma\} = \{13, 16, 3, 0.1\}$ and $D \in \{1,2,3,4\}$, $D_{\max} = 4$.

	0.0	0.1	0.2	0.3	0.4	0.5	0.6	0.7	0.8	0.9	1.0
$\hat{\alpha}$	0.17	0.17	0.18	0.22	0.3	0.41	0.52	0.63	0.73	0.84	0.95
p_1	0.87	0.87	0.85	0.78	0.61	0.35	0.12	0.02	0.0	0.0	0.0
p_2	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.01	0.02	0.17

Table 4.4: Distributions of $\hat{\alpha}, p_1, p_2$ from gPCA averaged over D with 25 repetitions each (100 in total) across $\alpha \in \{0.0, 0.1, \ldots, 1.0\}$. Specifications: $\{N_k, p, K, \sigma\} = \{39, 16, 2, 0.1\}$ and $D \in \{1,2,3,4\}$, $D_{\max} = 4$.

	0.0	0.1	0.2	0.3	0.4	0.5	0.6	0.7	0.8	0.9	1.0
$\hat{\alpha}$	0.12	0.12	0.14	0.23	0.35	0.46	0.57	0.67	0.78	0.88	0.99
p_1	0.93	0.93	0.91	0.76	0.48	0.22	0.07	0.02	0.0	0.0	0.0
p_2	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.09

4.4 Application to Studying Growth Factor Responsiveness across Breast Cancer Subtypes

4.4.1 Background and Data Processing

Breast cancer is a very diverse disease both in terms of its clinical attributes (Anderson and Matsuno, 2006) and molecular signature (TCGAN, 2013). Clinically, there are primarily four subtypes determined based on the presence of immunohistochemical markers estrogen receptor (ER), progesterone receptor (PR), and human epidermal growth factor receptor 2 (HER2) in the tissue (Parise et al., 2009), which are associated with different rates of survival (Onitilo et al., 2008) and treatment strategies (Weigel and Dowsett, 2010). These subtypes have been shown to correlate with molecular subtypes derived from gene expression (Sorlie, 2003) and microRNA expression (Blenkiron et al., 2007) profiles. Such molecular heterogeneity presents a significant challenge and opportunity not only for understanding breast cancer at the genomic level but also for developing more refined and personalized therapies that target clinical heterogeneity.

In the following data application we demonstrate how the gPCA model can be used for group-wise inferences in a factorial design study investigating growth factor responsiveness in breast cancer cell lines. The target study (Niepel et al., 2014) accounts for a variety of experimental conditions with replicated measurements across different

breast cancer subtypes, ligand (growth factor) types, concentrations of ligand, exposures times, and signaling pathways. Using gPCA, we derive estimates for the complexity and commonality (along with significance calculations) to provide an ANOVA-based decomposition of variation along the levels of each of these factors. This allows us to answer the main question of assessing ligand responsiveness across disease subtypes, while also accounting for each experimental factor in an efficient and comprehensive fashion.

We obtained pre- and post-treatment phosphorylation levels of ERK and AKT kinases to quantify activity in their respective growth-related MAPK and PI3K/AKT pathways. The log10 of fold-change (post- to pre-treatment ratio) was assessed for these two kinases following exposure to 15 different growth factors at two doses (1 and 100 ng/ml) and three time points (10, 30, and 90 minutes). Measurements were repeated across a total of 39 breast cancer cell lines spanning three clinical subtypes: 18 triple negative (TN, ER-/PR-/HER2-), 11 HER2-overexpressing (H2, HER2+), and 10 hormone receptor positive (HR, ER+ or PR+). The data can be found from http://lincs.hms.harvard.edu/niepel-bmcbiol-2014/.

4.4.2 Multivariate Comparisons across Experimental Factors

Given the relatively large number of ligand types, we elected to study how the distribution of responsiveness towards these ligands compares across each of the factors: ligand concentration, kinase type, time of measurement, and disease subtype. For each of these factors, gPCA was applied globally (among all factor levels) and pairwise (between each pair of factor levels) with $D_{\max} = 4$. In addition, we repeated the analysis over all combinations of the remaining factors, except for disease subtype as it was associated with the 39 cell line samples. Tables C3-C6 in Appendix C.5 summarize our findings, which include estimates of complexity and commonality, the p-values for commonality, and the proportion of within group variation attributed ($\tilde{\text{WSS}} = \text{WSS}/(\text{WSS} + \text{BSS})$). Data dimensions are provided for each scenario, along with the number of effective independent tests m, which we use for Bonferroni correction (i.e. significance level lowered to $0.05/m$).

Across the first three experimental factors (Tables C3-C5 in Appendix C5), the distribution in growth factor responsiveness appears to be generally homogeneous, as evidenced by unanimous significance in p_1. In other words, ligand concentration, kinase type, and time of measurement seem to have little effect on the overall observed patterns of growth factor-induced kinase activity among the different cell line and ligand combinations. This suggests that adjusting these experimental factors should not significantly alter the overall findings. Had any one of these factors showed signs of heterogeneity (i.e. insignificant p_1 or significant p_2) however, it would support conducting further analysis into potential discrepancies among their levels. We note that performing traditional tests of homogeneity (e.g. with the one-way ANOVA framework) for each ligand across subtypes not only would require many more tests, but also does not consider the collective distribution of ligand responses as a whole.

It is interesting that these findings of homogeneity contrast with those of the original study (Niepel et al., 2014) which found that responses among ligands were in fact significantly different when compared across measurement time, dosage, and kinase type. We note that their comparisons generally focus on individual ligands or cell lines, whereas our comparisons are across many cell lines and ligands. It is possible for individual combinations of cell lines and ligands to show marked differences across the factors while the collective data groups exhibit homogeneity as a whole. One advantage of gPCA is that it allows for a more global comparison strategy over individual checks under each experimental factor.

In addition to measuring the homogeneity among groups, gPCA provides additional insight in terms of the common complexity $\hat{D}$ and proportion of variation attributable to within groups $\widetilde{\text{WSS}}$. The observed patterns across ligand concentration, kinase type, and time of measurement were generally simplistic, reflecting the fact that a one-dimensional model was usually enough to adequately approximate the homogeneous data patterns across these factors. Meanwhile, within group variation seems to correlate with commonality, which is expected since more variation attributed within groups implies less variation attributed to group differences. This correlation is not perfect

because the similarity (α) between each group's data signature is not tied to the relative noise level (σ) within each group.

When disease subtype was considered as the factor variable (Table C4 in Appendix C5), we observed some degree of heterogeneity in growth factor responsiveness, which was largely restricted to measurements of pAKT. To investigate whether a subgroup of ligand types were primarily responsible for the variation in growth factor sensitivity, we separated the ligands into ErbB-associated (EGF, EPR, BTC, HRG) and non-ErbB-associated (VEGF165, INS, IGF-1/2, PDGF-BB, HGF, SCF, FGF-1/2, NGF-beta, EFNA1) groups. Figures 4.2 and 4.3 show heatmaps of pAKT and pERK responsiveness to ligands in all cell lines, measured 30 minutes after treatment of 100 or 1 ng/ml of ligand respectively. Moreover the heterogeneity and homogeneity of patterns across subtypes in these respective settings can be readily observed.

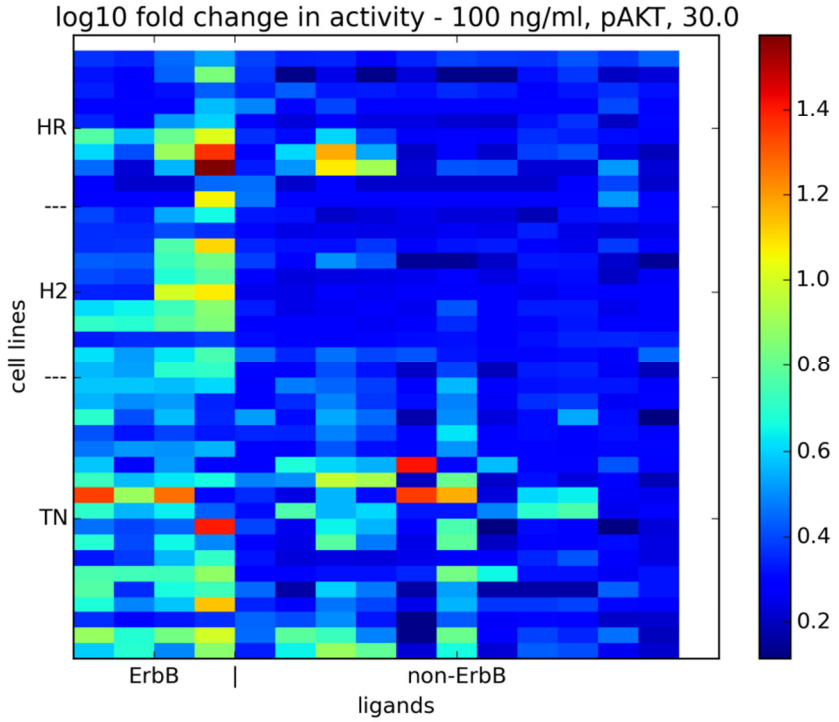

Figure 4.2: Heatmap of log10 fold change of pAKT measured 30 minutes in response to 100 ng/ml treatment of ligand, among all ligand types and all cell lines. Breast cancer subtype and ligand subgroup memberships are indicated. Growth factor sensitivity of AKT is heterogeneous in distribution but similar in level across subtypes.

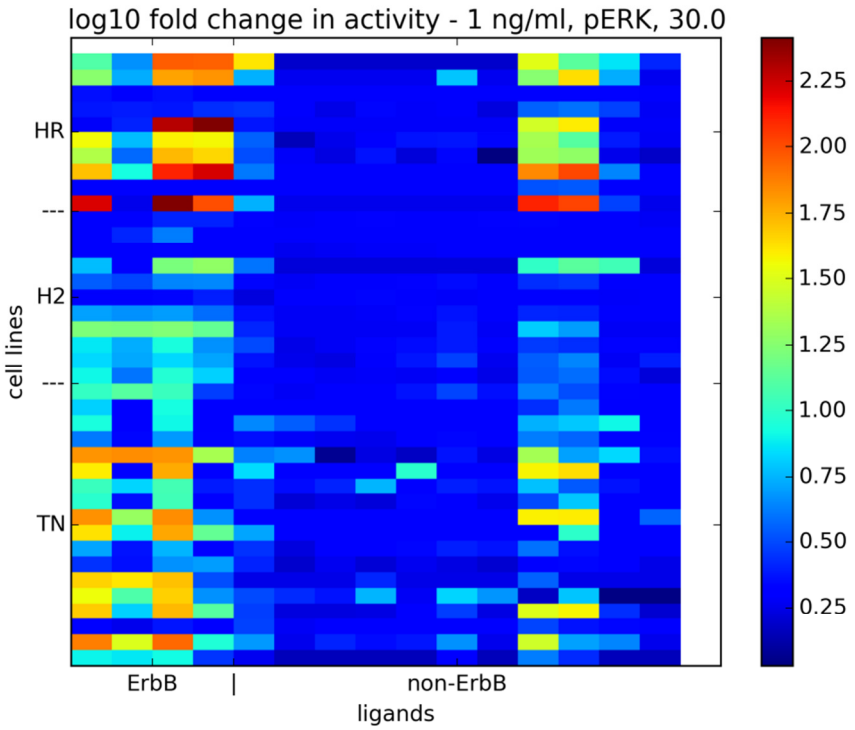

Figure 4.3: Heatmap of log10 fold change of pERK measured 30 minutes in response to 1 ng/ml treatment of ligand, among all ligand types and all cell lines. Breast cancer subtype and ligand subgroup memberships are indicated. Growth factor sensitivity of ERK is homogeneous in distribution but different in level across subtypes.

Using the gPCA procedure, we have reduced the vast range of combinations of experimental conditions to identify the main source of variation as the relationship between disease subtype and ligand type. From this point, more targeted analysis may proceed. In the original analysis, Niepel et al. (2014) report that cells of the TN subtype were the most responsive, while H2 and HR were particularly sensitive to certain classes of growth factors (ErbB2-related and FGF1/FGF2/HRG respectively). These findings are evident in the heatmaps and can be confirmed statistically via univariate comparison techniques such as the F-test and t-test. For instance, under the specifications of Figure 4.2, the mean ErbB ligand response was significantly different ($t = 10.4$, p-value $= 1.3e^{-19}$) from that of non-ErbB ligands in the H2 subtype. Under the specifications of Figure 4.3, the mean response of FGF1, FGF2, and HRG was significantly different ($t = 4.28$, p-value $= 3.3e^{-5}$) than that of the rest in the HR subtype.

Under these same two specifications, we used gPCA to compare ligand responsiveness among ligand types across disease subtypes, for ErbB and non-ErbB ligand subgroups (Tables 4.5 and 4.6). For AKT, ligand response patterns were very

diverse across subtypes, particularly between TN and HR subtypes. Meanwhile, non-ErbB-induced responses of ERK were generally more homogeneous across subtypes than ErbB-induced responses. It is interesting to compare these results with those of F-test comparisons from ANOVA (Tables 4.7 and 4.8). For AKT, mean ErbB-induced activity was largely similar across subtypes, as was mean non-ErbB-induced activity to a lesser extent. In contrast, we see dramatic differences in mean ligand-induced ERK response across almost all subtypes for both ligand subgroups. Thus we see that growth factor sensitivity of AKT is heterogeneous in distribution but similar in level across subtypes, and that growth factor sensitivity of ERK is homogeneous in distribution but different in level across subtypes. Although breast cancer subtype is a major determinant for growth factor-induced responsiveness for both kinases, it appears that different signal transduction pathways across subtypes are involved in activating AKT while similar pathways lead to activation of ERK at different magnitudes across subtypes.

Table 4.5: Summary of gPCA findings (AKT only) for ligand responsiveness among ligand types across disease subtypes. Data dimensions: (18 + 11 + 10 cell lines × 4 or 11 ligands). Significance (denoted by shading) was assessed at level $0.05/m$ for $m = 8$ independent tests in each scenario).

Specifications	Factor levels	$\hat{D}$	$\hat{\alpha}$	p_1	p_2	$\tilde{\text{WSS}}$
100 ng/ml; AKT; 30 min (ErbB ligands)	TN, H2, HR	1	0.796	2.51E-01	0.001	0.679
	TN, H2	1	0.735	3.47E-01	0.065	0.741
	TN, HR	1	0.732	3.51E-01	0.002	0.697
	H2, HR	1	0.948	3.78E-02	0.175	0.868
100 ng/ml; AKT; 30 min (non-ErbB ligands)	TN, H2, HR	1	0.596	2.57E-01	0.002	0.643
	TN, H2	1	0.597	2.56E-01	0.386	0.866
	TN, HR	1	0.452	5.25E-01	0.004	0.661
	H2, HR	1	0.756	5.20E-02	0.092	0.788

Table 4.6: Summary of gPCA findings (ERK only) for ligand responsiveness among ligand types across disease subtypes. Data dimensions: (18 + 11 + 10 cell lines × 4 or 11 ligands). Significance (denoted by shading) was assessed at level $0.05/m$ for $m = 8$ independent tests in each scenario).

Specifications	Factor levels	$\hat{D}$	$\hat{\alpha}$	p_1	p_2	$\tilde{\text{WSS}}$
1 ng/ml; ERK; 30 min (ErbB ligands)	TN, H2, HR	1	0.931	5.75E-02	0.055	0.699
	TN, H2	1	0.962	2.43E-02	0.104	0.846
	TN, HR	1	0.9	9.60E-02	0.030	0.657
	H2, HR	1	0.951	3.49E-02	0.077	0.799
1 ng/ml; ERK; 30 min (non-ErbB ligands)	TN, H2, HR	1	0.929	2.93E-04	0.087	0.819
	TN, H2	1	0.912	7.95E-04	0.305	0.844
	TN, HR	1	0.943	1.12E-04	0.082	0.916
	H2, HR	1	0.927	3.37E-04	0.115	0.777

Table 4.7: Summary of ANOVA findings (AKT only) for growth factor responsiveness among ligand types across disease subtypes. Data dimensions: (18 + 11 + 10 cell lines × 4 or 11 ligands). Significance (denoted by shading) was assessed at level $0.05/m$ for $m = 8$ independent tests in each scenario).

Specifications	Factor levels	F	p-value	Factor level	Mean
	TN, H2, HR	0.116	8.90E-01		
100 ng/ml; AKT; 30 min	TN, H2	0.346	5.58E-01	TN	3.8
(ErbB ligands)	TN, HR	0.003	9.56E-01	H2	3.39
	H2, HR	0.109	7.42E-01	HR	3.74
	TN, H2, HR	8.28	2.98E-04		
100 ng/ml; AKT; 30 min	TN, H2	14.49	1.69E-04	TN	1.92
(non-ErbB ligands)	TN, HR	3.59	5.90E-02	H2	0.99
	H2, HR	6.15	1.38E-02	HR	1.39

Table 4.8: Summary of ANOVA findings (ERK only) for growth factor responsiveness among ligand types across disease subtypes. Data dimensions: (18 + 11 + 10 cell lines × 4 or 11 ligands). Significance (denoted by shading) was assessed at level $0.05/m$ for $m = 8$ independent tests in each scenario).

Specifications	Factor levels	F	p-value	Factor level	Mean
	TN, H2, HR	13.4	4.32E-06		
1 ng/ml; ERK; 30 min	TN, H2	14.1	2.80E-04	TN	18.28
(ErbB ligands)	TN, HR	10.8	1.38E-03	H2	4.84
	H2, HR	16.4	1.18E-04	HR	48.27
	TN, H2, HR	8.72	1.94E-04		
1 ng/ml; ERK; 30 min	TN, H2	5.84	1.62E-02	TN	2.92
(non-ErbB ligands)	TN, HR	7.93	5.19E-03	H2	1.43
	H2, HR	10.23	1.58E-03	HR	7.38

Just as these standard statistical methods are used to compare univariate data across groups, gPCA provides the framework for comparing multivariate patterns across groups. The former should be used to quantify differences in magnitude between data groups, while the latter should be used to quantify differences in variational signatures of the data. This distinction is most obvious as we consider that information on the averages is forgoed in the column-centering process of PCA. Altogether, the two methods offer complementary insights towards the characterization of data patterns across groups.

4.5 Discussion

The availability of biological data is expanding in terms of not only volume but also variety. This has led to growing interest in developing new analytic methods which are able to draw upon multiple data sources and consolidate the information within. In this paper we have approached this challenge from the perspective of dimensionality reduction, a widely used tool for reducing complex biological patterns to fundamental components. At the level of these core components, we construct an integrative framework that provides novel insight into the variational patterns within and between multiple data groups.

Our approach combines the variation decomposition structure of ANOVA with the variation reduction procedure of PCA. We expand the study of within and between variation to the multivariate setting, which deals with comparisons of subspaces of the groups signatures via principal angles. This key element provides a novel estimation scheme for the complexity and commonality that characterizes multiple groups of data. Applying properties of principal angles in turn allows for theoretical significance calculations in addition to the conventional permutation-based calculations.

In our data application, we have demonstrated the utility of this multivariate comparison framework to a factorial design study of growth factor response signatures. Not surprisingly, the method provides an ANOVA-type layout of conclusions composed of global assessments (of homogeneity or heterogeneity) across all data groups followed by pairwise assessments between groups. Owing to the ability of PCA-type decompositions to reduce multivariate patterns, we have a way of performing such assessments over matrices of data values rather than vectors. This allows a novel analytic approach for studying data organized into groups on a broader scale.

Despite originating from fairly classical statistical methods, the style of inference provided by gPCA represents a unique approach to multivariate and multi-group analysis. However, there are many areas for refinement, such as in the assumptions used in deriving the gPCA estimators for commonality and common complexity. In particular, the loadings of each data group play an important role to these estimators yet are assumed to be perfectly estimated in our model. Random matrix theory may reveal further avenues for development. Moreover, there may be other unexplored ways of deriving

insight from the gPCA model. The perspective of data groups in terms of within and between complexity structures may give rise to alternative views of data patterns than the one presented in this work.

CHAPTER V

Conclusion

The main contribution of this work is the generalization of several canonical dimensionality reduction methods for the setting of multiple data groups, in ways that further each method's underlying reduction approach. A recurring theme among these integrative extensions is to develop an overarching framework which includes the standard joint and separate approaches as special cases. Once this is accomplished, the remaining task (and the key innovative element) is to develop a strategy for selecting the most appropriate intermediate among a full spectrum. Effectively, this amounts to selecting a level of heterogeneity relative to homogeneity for the resulting model that is most representative of the data or most conducive to positive performance. In order to better identify or utilize the common signal among different data sources, it is merely principled to first tease out the distinctive noise.

In Chapter 2, I proposed a partitioned factorization model intended for distinguishing between homogeneous and heterogeneous patterns among multiple sources of nonnegative data. In this setting, such patterns are nonnegative signal components whose additive combination approximates the observed data. This sum-of-parts perspective (though interpretable and biologically applicable) leads to non-uniqueness of the solution, particularly in the heterogeneous parts. To deal with this, I enacted penalization on precisely these parts to counterbalance their unrestricted nature in the factorization. However, the new problem then becomes how to achieve this balance, which is complicated by the unwieldly nonnegativity constraint on the algorithm, also a consequence of the sum-of-parts perspective. Using a novel empirical approach that gauges the magnitude of approximation residuals, this balance is detected as the point at which the relationship among these residuals begins to deteriorate due to overfitting. The

resulting integrative factorization demonstrates marked advantages in the detection of common regulatory modules in synthetic and real genomic data.

In Chapter 3, I discussed a variant of the partial least squares classifier for adapting to heterogeneity across multiple cohort studies. Here, the units of integration are not the patterns within each dataset, but the associations that connect them as captured by the regression coefficient. To preserve the efficient iterative computation of these coefficients but also gain model flexibility, we introduced a simple modification that reduces the comparison across cohorts to a single correlation-based term. Naturally, this translates to efficiency in the algorithm, mainly due to the ease with which this term is absorbed into the existing computations. These adjustments grant improved robustness and accuracy over standard classifiers, as shown in simulations and real data applications on the development of molecular prognostic signatures.

Lastly, in Chapter 4, I combined principles of variance decomposition and variance reduction to conduct multivariate comparisons across the many experimental conditions of a cell line study. While our hybrid framework preserved many aspects ANOVA and PCA, the nature of these methods required additional considerations, the most pivotal of which was the incorporation of principal angles. As the centers of the sum-of-squares elements are transformed from scalars to subspaces, the largest principal angle served as the crucial metric of distance for quantifying heterogeneity between data groups. This led to new statistics for quantifying complexity and commonality, as well as significance calculations for the latter, which altogether provide novel insight and strategies for grouped data analysis.

As noted before, the main intention of this work is to explore different styles of dimensionality reduction as they are adapted to handle multiple datasets. In both iNMF and iPLS, the increased model flexibility produced by accommodating multiple groups was addressed via tuning selection. For the latter, this was rather straightforward, as response data was available for performing cross validation. For the former, this called for the alternative approach of gauging the residuals of the joint components of the dimensionality reduction, a strategy which takes advantage of the perspective of the data as nonnegative parts. Due to the simple and direct nature of PCA solutions, the unique optimality of principal components was not as conducive to a tuned intermediate model

as was in the other methods. This invited the application of principles from ANOVA that provided a novel view on principal variation as within and between groups, which ultimately led to not an intermediate model but rather an intermediate level of commonality. Since accounting for the potential differences among multiple data groups naturally leads to higher degrees of freedom, any relaxation of a model's framework to allow such differences must be coupled with restrictive measures. As we have seen here, how each classical method approaches these measures reflects their respective data and problem structures.

APPENDICES

APPENDIX A

Supplementary Material for "A Non-negative Matrix Factorization Method for Detecting Modules in Heterogeneous Omics Multi-modal Data"

A.1 Derivation of the iNMF Algorithm

We show here that the multiplicative updates used to solve iNMF ensure that the objective function $\mathcal{F}(W, H, V)$ is monotonically decreasing:

$$\mathcal{F}(W, H, V) = \sum_k \| X_k - (W + V_k)H_k \|_F^2 + \lambda \sum_k \| V_k H_k \|_F^2$$

All quantities X_k, W, H_k, V_k are as defined in the main article. For convenience, we use H and V to denote $\{H_1, \dots, H_K\}$ and $\{V_1, \dots, V_K\}$, respectively.

1. The bulk of the proof involves auxiliary functions and some algebraic manipulation, but an application of duality theory reveals some useful relations. The corresponding dual problem of iNMF is:

$$\max_{\Theta} \inf_{W,H,V} \mathcal{F}(W, H, V) + \text{tr}(\Phi W^T) + \sum_k \text{tr}(\Psi_k H_k^T) + \sum_k \text{tr}(\Xi_k V_k^T) \quad \text{(A1)}$$

$$\text{subject to: } \Phi \geq 0, \Psi_k \geq 0, \Xi_k \geq 0, k = 1, \dots, K,$$

where $\Theta = \{\Phi, \Psi_1, \dots, \Psi_K, \Xi_1, \dots, \Xi_K\}$ are matrices whose elements are the Lagrangian multipliers for the elements of $\{W, H_1, \dots, H_K, V_1, \dots, V_K\}$, respectively. By definition, we have $\Phi \in \mathbb{R}^{N \times D}, \Psi_k \in \mathbb{R}^{D \times M_k}$ and $\Xi_k \in \mathbb{R}^{N \times D}$ for all $k = 1, \dots, K$.

From the first order conditions of the Lagrangian function in Equation A1, we may solve for the Lagrangian multipliers:

$$\Phi = 2 \sum_k (X_k H_k^T - (W + V_k) H_k H_k^T)$$

$$\Psi_k = 2((W + V_k)^T X_k - (W + V_k)^T (W + V_k) H_k - \lambda V_k^T V_k H_k), k = 1, \dots, K$$

$$\Xi_k = 2(X_k H_k^T - (W + V_k)H_k H_k^T - \lambda V_k H_k H_k^T), k = 1, \dots, K$$

By the complementary slackness property, we have the following relations at the optimal solution for all indices (i, j):

$$W_{ij} \sum_k (X_k H_k^T - (W + V_k)H_k H_k^T)_{ij} = 0$$

$$(H_k)_{ij}((W + V_k)^T X_k - (W + V_k)^T (W + V_k)H_k - \lambda V^T V H_k)_{ij} = 0, k = 1, \dots, K$$

$$(V_k)_{ij}(X_k H_k^T - (W + V_k)H_k H_k^T - \lambda V_k H_k H_k^T)_{ij} = 0, k = 1, \dots, K.$$

These relations lead to our multiplicative updates after some algebraic manipulation.

2. The last portion of the proof involves auxiliary functions, defined below:

Definition. *$G(h, h')$ is an auxiliary function for F(h) if the following are satisfied:*

$$G(h, h') \geq F(h) \forall h$$

$$G(h, h) = F(h).$$

Auxiliary functions have the following property:

Lemma A1. *If G is an auxiliary function for F, and $h^{(t+1)} = arg\min_h G(h, h^{(t)})$, then*

$$F(h^{(t+1)}) \leq F(h^{(t)}).$$

Proof. $F(h^{(t+1)}) \leq G(h^{(t+1)}, h^{(t)}) \leq G(h^{(t)}, h^{(t)}) = F(h^{(t)})$. □

If G is easier to minimize than F, then we may take repeated iterations of $h^{(t+1)} = \arg\min_h G(h, h^{(t)})$ instead of directly dealing with F.

3. Each of the iNMF updates may be derived with an appropriate auxiliary function. We outline here only the derivation for V_k update, but the updates for W, H_k are similarly derived. Since the updates are performed element-wise, it is enough to show that the update $(V_k)_{ij}^{(t+1)}$ satisfies:

$$\mathcal{F}((V_k)_{ij}^{(t+1)}) \leq \mathcal{F}((V_k)_{ij}^{(t)}). \quad \text{(A2)}$$

The first two derivatives of $\mathcal{F}$ with respect to $(V_k)_{ij}$ are:

$$\mathcal{F}'_{ij} = \mathcal{F}'((V_k)_{ij}) = (-2X_k H_k^T + 2(W + V_k)H_k H_k^T + 2\lambda V_k H_k H_k^T)_{ij}$$

$$\mathcal{F}''_{ij} = \mathcal{F}''\left((V_k)_{ij}\right) = 2(1+\lambda)(H_k H_k^T)_{jj}.$$

Lemma A2. *The function:*

$$\mathcal{G}(h,(V_k)_{ij}) = \mathcal{F}((V_k)_{ij}) + \mathcal{F}'((V_k)_{ij})(h - \left(V_k\right)_{ij}) + \frac{\left((W + V_k + \lambda V_k)H_k H_k^T\right)_{ij}}{(V_k)_{ij}}(h - (V_k)_{ij})^2,$$

is an auxiliary function for $\mathcal{F}$.

Proof. $\mathcal{G}((V_k)_{ij},(V_k)_{ij}) = \mathcal{F}((V_k)_{ij})$ is easy to see. To show that $G(h,(V_k)_{ij}) \geq \mathcal{F}(h)$, we write out the Taylor expansion of $\mathcal{F}$ at $(V_k)_{ij}$:

$$\mathcal{F}(h) = \mathcal{F}((V_k)_{ij}) + \mathcal{F}'((V_k)_{ij})(h - \left(V_k\right)_{ij}) + (1+\lambda)(H_k H_k^T)_{jj}(h - (V_k)_{ij})^2.$$

Thus, it is sufficient to show that:

$$\frac{((W + V_k + \lambda V_k)H_k H_k^T)_{ij}}{(V_k)_{ij}} \geq (1+\lambda)(H_k H_k^T)_{jj}.$$

By nonnegativity of the matrix factors, we have:

$$\begin{aligned}\frac{\left((W + V_k + \lambda V_k)H_k H_k^T\right)_{ij}}{(V_k)_{ij}} &\geq (1+\lambda)\frac{(V_k H_k H_k^T)_{ij}}{(V_k)_{ij}} \\ &= (1+\lambda)\frac{\sum_l (V_k)_{il}\,(H_k H_k^T)_{lj}}{(V_k)_{ij}} \\ &= (1+\lambda)(H_k H_k^T)_{jj}.\ \square\end{aligned}$$

Combining Lemmas A1 & A2, we have that the update:

$$(V_k)_{ij}^{(t+1)} = \underset{h}{\operatorname{argmin}} G(h, \left(V_k\right)_{ij}^{(t)}),$$

guarantees Equation A2. But this minimizer can be expressed as:

$$\begin{aligned}\arg\min_h G(h,(V_k)_{ij}^{(t)}) &= (V_k)_{ij} - (V_k)_{ij}\frac{\mathcal{F}'((V_k)_{ij})}{2\left((W + V_k + \lambda V_k)H_k H_k^T\right)_{ij}} \\ &= (V_k)_{ij} - (V_k)_{ij}\frac{(-2X_k H_k^T + 2(W+V_k)H_k H_k^T + 2\lambda V_k H_k H_k^T)_{ij}}{2((W+V_k+\lambda V_k)H_k H_k^T)_{ij}} \\ &= (V_k)_{ij}\frac{(X_k H_k^T)_{ij}}{((W + V_k + \lambda V_k)H_k H_k^T)_{ij}},\end{aligned}$$

which is exactly our iNMF update for $(V_k)_{ij}$.

A.2 Intuition for the Tuning Selection Procedure

We discuss here the intuition behind the stopping threshold $R_I^{(\lambda)} - R_J > 2(R_J - R_S)$ from the tuning selection procedure. Let $X_k, k = 1, \ldots, K$ be observationally linked data sets, and let X_k^S, X_k^J, X_k^I be the approximating solutions of sNMF, jNMF, and iNMF:

$$X_k^S = W_k^S H_k^S, X_k^J = W^J H_k^J, X_k^I = (W^I + V_k^I) H_k^I.$$

Suppose that we adjust these solutions entry-wise with respect to the jNMF solution:

$$\tilde{x}_{k,ij}^S = x_{k,ij}^S - x_{k,ij}^J, \tilde{x}_{k,ij}^J = x_{k,ij}^J - x_{k,ij}^J, \tilde{x}_{k,ij}^{I,h} = (W^I H_k^I)_{k,ij} - x_{k,ij}^J.$$

Note that for iNMF we consider only the homogeneous portion. We will omit subscripts for the sake of brevity.

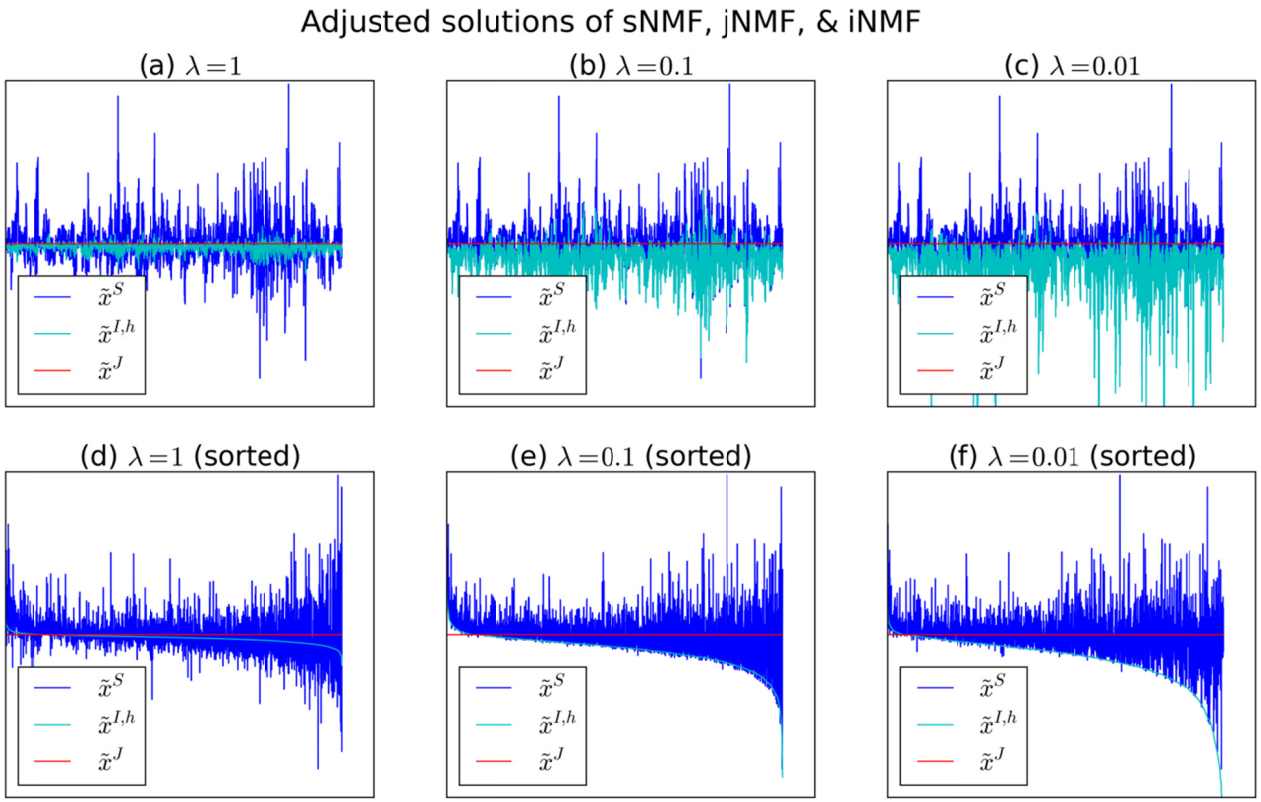

Figure A1: Adjusted sNMF, jNMF, and iNMF solutions with different λ choices for iNMF, computed from generated data ($\sigma_u, \sigma_s, \sigma_h = (0.01, 0.2, 0.01)$). First row: unsorted; second row: sorted with respect to the adjusted iNMF solution.

Figure A1 plots the entries of the adjusted solutions $\tilde{x}^S, \tilde{x}^{I,h}, \tilde{x}^J$ computed from simulated data (see Appendix A.3) over different choices of λ for iNMF. Naturally, $\tilde{x}^S$ (sNMF) are centered around $\tilde{x}^J$ (jNMF). Also, $\tilde{x}^{I,h}$ (iNMF, homogeneous) generally lie below $\tilde{x}^J$ (jNMF), since the other heterogeneous portion of iNMF is nonnegative. As the choice of λ shrinks, the iNMF solution becomes less homogeneous and $\tilde{x}^{I,h}$ becomes less resembling of $\tilde{x}^J$. When λ is small enough, iNMF begins to over fit the data. Our tuning selection procedure selects = 0:1 for this particular example, which in fact leads to optimal performance.

When we sort the adjusted solutions by $\tilde{x}^{I,h}$, we see some interesting relations. At the optimal λ (Figure A1e), the iNMF homogeneous solutions $\tilde{x}^{I,h}$ lie slightly above the minimum of the distribution of the sNMF solutions $\tilde{x}^S$. If the level of $\tilde{x}^{I,h}$ had been higher (Figure A1d), then the full iNMF solution would deviate from the sNMF solution, and hence yield a poor approximation of the data. If the level of $\tilde{x}^{I,h}$ had been lower (Figure A1f), then the approximation accuracy of iNMF will be slightly improved at the expense of losing detection of the joint signal. In principle, the optimal iNMF solution must (1) achieve good fit on the data and (2) maximize the homogeneous portion used to achieve that fit.

Now consider the distributions of the unsorted adjusted solutions (Figure A1a-c). What is notable about the iNMF solutions $\tilde{x}^{I,h}$ under optimal $\lambda = 0.1$ (Figure A1b) is that its distribution appears to match the lower half of the distribution of $\tilde{x}^S$. Similar to before, this is a distinguishing feature of an optimal iNMF solution. Another way of describing this is to say that the deviation between the iNMF (under optimal λ) and jNMF solutions is roughly twice the deviation between the jNMF and sNMF solutions. In fact, our stopping threshold $R_I^{(\lambda)} - R_J > 2(R_J - R_S)$ takes advantage of precisely this relation. As the selection procedure iteratively evaluates choices of λ, it effectively tunes the relative magnitude of the iNMF homogeneous solution to a level that matches that of the optimal solution.

In summary, selecting the optimal λ is akin to finding the iNMF solution with the most appropriate level of deviation from the jNMF and sNMF solutions. What remains is

to decide how to quantify this deviation for each data source. We use the (unsquared) Frobenius norm of the residuals for this task, summed across sources:

$$R_{S,k} = \| X_k - X_k^S \|_F, R_{J,k} = \| X_k - X_k^J \|_F, R_{I,k} = \| X_k - W^I H_k^I \|_F$$

Note the unconventional definition of the iNMF residual with respect to the homogeneous part only. Since the sNMF and jNMF are minimizers of their respective objective functions, we have $R_{S,k} \leq R_{J,k} \leq R_{I,k}$.

Our primary reasons for using the unsquared residuals are that (1) they are on approximately the same scale as the solutions and (2) they are more robust than comparing the solutions directly, particularly due to the relative non-identifiability of NMF-type solutions with respect to scale and rotation. Also, our tuning selection procedure suggests searching across a decreasing list of λ until the threshold is exceeded. This is slightly more conservative (to avoid overfitting) than finding the λ such that $R_I^{(\lambda)} - R_J$ is closest to $2(R_J - R_S)$, although the latter is an option.

A.3 Data Generation for the Simulation Study

We outline here our method of generating data sets containing multi-dimensional modules with various types of perturbations.

1. Generate a joint block diagonal support:
    a. Set $W_{N\times D}$ and $(H_k)_{D\times M_k}, k = 1, \ldots, K$ to be binary and block diagonal (D blocks) so that their products WH_k align with the desired data and module dimensions.
    b. Independently assign each nonzero entry in W and H_k a random value according to $\text{Beta}(2,2) * 2$ (this is arbitrary).
    c. Multiply to obtain the matrices $WH_k, k = 1, \ldots, K$.
2. Introduce heterogeneous perturbations:
    a. Set $(V_k)_{N\times D}$ to be zero matrices, and consider the D^2 regions whose rows and columns align with the D blocks (modules) in W.

b. In each of the D^2 regions, introduce a heterogeneous perturbation (with independent probability σ_h) by assigning either the top or lower half (with equal probability) to be ones.
c. Independently assign to each nonzero entry of $(V_k)_{N\times D}$ a random value according to $\text{Beta}(2,2) * 2$.
d. Add the products $V_k H_k$ to the previous results to obtain $X_k = (W + V_k)H_k$ (the data sets should resemble the ones in Scenario 2 of Figure A1a).

3. Introduce scattered and uniform error:
   a. For each entry in X_k, with independent probability σ_s, either replace a positive value with zero, or replace a zero with a randomly generated $(\text{Beta}(2,2) * 2)^2$ value.
   b. For each entry in X_k, with independent probability σ_u, add a random $\text{Unif}(-\sigma_u, \sigma_u)$ value, and take the absolute magnitude as the new entry.

A.4 Normalization for iNMF

In dealing with multiple data sources, integrative methods must find a way to represent the information from each source in a balanced way. In iNMF, we may consider attaching weights c_k to each data matrix to control the level of influence of each source over the analysis:

$$\mathcal{F}(W,H,V) = \sum_k \| c_k X_k - (W + V_k)H_k \|_F^2 + \lambda \sum_k \| V_k H_k \|_F^2$$

Of course, this is equivalent to scaling each data set X_k by a factor of c_k. Here, we explore how one should approach choosing these normalization coefficients.

Table A1: Impurity (I) and purity (P) scores (in percentages) of empirical clusters obtained from jNMF and iNMF with respect to three reference clusters. Shading indicates significantly (≥2 sd) higher concordance compared to both the alternative method and the null distribution.

		I			P		
		DM	GE	ME	DM	GE	ME
Null clusters	mean	61	58	44	49	50	65
	st.dev.	4	7	2	5	8	1
$\lambda_s = 1$	jNMF	58	23	36	50	77	74
	iNMF	56	49	24	58	62	85
$\lambda_s = 0.1$	jNMF	58	26	30	50	77	79
	iNMF	55	43	27	58	69	82
$\lambda_s = 0.01$	jNMF	62	30	26	50	77	82
	iNMF	46	31	27	67	69	82
$\lambda_s = 10^{-3}$	jNMF	45	56	30	67	54	79
	iNMF	48	48	27	67	62	82
$\lambda_s = 10^{-4}$	jNMF	56	31	41	58	69	68
	iNMF	54	62	27	58	54	82

In our application, we normalized with respect to the within-source variance of each data set (i.e. $c_k = 1/\text{std}(X_k)$). This accounts for the inherent levels of variation within the sources, but not the numbers of variables (about a 19:20:1 ratio). Therefore, we also consider here normalizing with respect to the sum-of-squares of each data set (i.e. $c_k = 1/\sqrt{\text{SS}(X_k)}$). Table A1 shows the validation results from repeating the analysis under this alternative normalization. Compared with those of the previous normalization, the GE clusters are less concordant with the reference while the ME clusters are more concordant. The scores for the DM clusters remain roughly consistent, likely due to these clusters having poor concordance to begin with.

In principle, the normalization weights should be chosen to address discrepancies in the variability of data and the number of variables in each source. However, the integrative value of a data source may depend on many other factors such as the reliability of each source, the relevance of each source to the research purpose, and the clarity of each source's signal. Therefore, dimensionality and variability should not completely dictate the normalization. As we have seen, applying the sum-of-squares normalization does not necessarily produce a more concordant joint approximation of modules, possibly due to differences in signal strength and fidelity between the sources.

As a general rule, dimensionality and data variability should guide the choice of normalization, but the nature of the sources themselves should also be taken into account. In our application, our follow-up analysis takes place in the space of genes, so it was natural to use the standard deviation normalization which produced more concordant GE results.

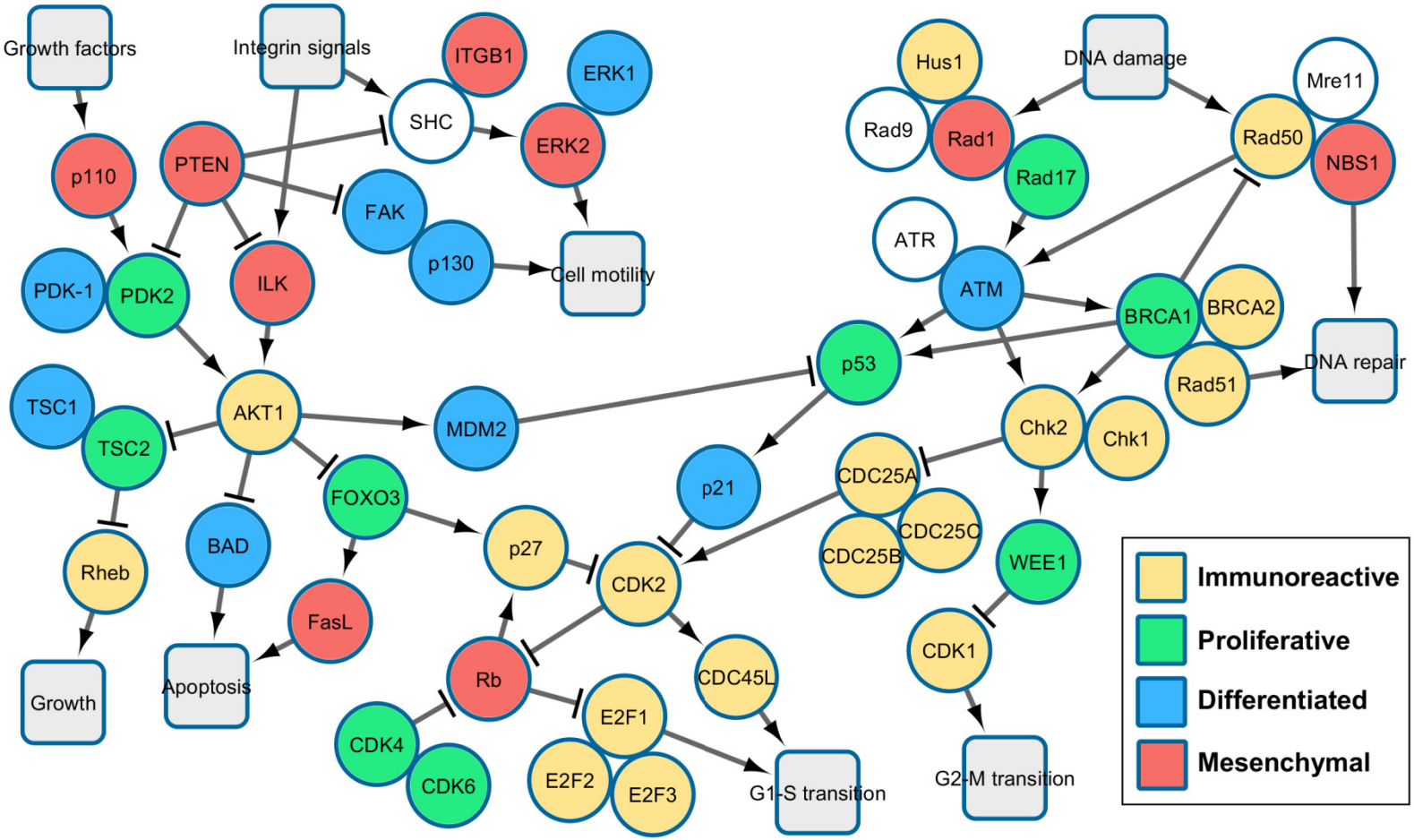

Figure A2: Module memberships of genes (from iNMF with alternative sum-of-squares normalization) arranged according to pathways derived from BioCarta and relevant literature.

In any case, it is recommended to check the robustness of the findings under different normalizations. Under the sum-of-squares normalization, iNMF produced the most concordant results under $\lambda = 0.01$ (Table A1). Using this result, we applied the same procedure as used before to obtain the visualization in Figure A2. Apart from minor discrepancies, all four modules (I/P/D/M) are distributed in roughly the same topological regions as before. The empirical memberships of the genes in these pathways appear stable between the two normalizations.

A.5 Reference Variable Clusters

GE reference cluster:

1. CXCL11, CXCL10, CXCR3
2. HMGA2, SOX11, MCM2, PCNA
3. MUC16, MUC1, SLPI
4. FAP, ANGPTL2, ANGPTL1

DM reference cluster:

1. cg08046471, cg01288089, cg08843314
2. cg03251079, cg20088964, cg08432727, cg20008332, cg10691006, cg15057726, cg02689825, cg04562739, cg25984124
3. cg06420088, cg07399355, cg17257175, cg24512973, cg12966875, cg23889010
4. cg08826839, cg09427311, cg11213150, cg07044282

ME reference cluster:

1. hsa-miR-19a, hsa-miR-19b, hsa-miR-136, hsa-miR-376c, hsa-miR-483-5p, hsa-miR-572, hsa-miR-575, hsa-miR-638, hsa-miR-671-5p, hsa-miR-769-5p, hsa-miR-923, hsamiR-1225-5p
2. hsa-miR-15b, hsa-miR-98, hsa-miR-135b, hsa-miR-146a, hsa-miR-148a, hsa-miR-148b, hsa-miR-150, hsa-miR-221*, hsa-miR-342-5p, hsa-miR-361-3p, hsa-miR-362-3p, hsamiR-374a, hsa-miR-374b, hsa-miR-450a, hsa-miR-454, hsa-miR-502-5p, hsa-miR-505, hsa-miR-532-3p, hsa-miR-582-5p, hsa-miR-625, hsa-miR-652, hsa-miR-660

APPENDIX B

Supplementary Material for "An Adaptive Partial Least Squares Classifier for Robust Prognostic Gene Signatures"

B.1 Supplementary Tables for the Multi-Cohort Prognostic Signature Study

Table B1: Top 10 genes of signatures identified from predictive methods (testing on Pawitan cohort, both ER statuses).

ER+

iPLS	jPLS	sPLS	jLR	sLR	jRF	sRF
RACGAP1	RACGAP1	RACGAP1	TPX2	LY6G6E	VTCN1	MAB21L1
PRC1	PRC1	RRM2	MRPL11	MMP12	MEIS3P1	UQCRC1
AURKA	AURKA	GINS2	SNRPE	ULK4	LOC100294145	MST4
RRM2	RRM2	PRC1	CDCA3	KIF18A	KIF11	FBXO5
NCAPG	NCAPG	CCNB1	RRM2	EYA4	C15orf63	C3orf14
PBK	PBK	CDK1	COG8	HMGB3	LTF	CHMP1A
GINS2	GINS2	AURKA	C12orf35	UQCRC1	DTL	SLC9A2
KIF11	KIF11	KIF11	C12orf44	CXCL2	TMEM106C	GTPBP4
ASPM	ASPM	ZWINT	SNRPA1	NQO1	RACGAP1	KERA
TOP2A	TOP2A	DTL	SHFM1	LOC157562	SEMA3G	CH25H

ER−

iPLS	jPLS	sPLS	jLR	sLR	jRF	sRF
PTPLB	PTPLB	PTPLB	MST4	S100P	AURKA	CPT1A
IDI1	IDI1	IDI1	CPT1A	MSMB	ESRP2	SHCBP1
MST4	NUTF2	ESRP2	SPAG16	ERCC6L	C14orf139	SNRPG
NDRG1	NDRG1	MST4	GINS1	MB	ECT2	SEC61G
NUTF2	MST4	CPT1A	CKS1B	UBE2S	MSMB	GIPC2
MLF1IP	MLF1IP	S100P	MRPL13	SHFM1	RAN	TPX2
C14orf156	C14orf156	C14orf156	CCNA2	CDC20	SHMT2	ENY2
CPT1A	ESRP2	MMP1	PGP	CNTNAP2	CYP4B1	CDCA8
HSPB1	MB	NUTF2	PSMA7	CDCA8	SURF2	EIF4EBP1
ESRP2	HSPB1	MLF1IP	C7	EIF4EBP1	DTL	C14orf156

Table B2: Pairwise correlations between regression coefficients generated from predictive methods (all cohort combinations and ER statuses). Values in bold represent comparisons between methods with best performance (within 0.03 of highest AUC among these methods) in predicting cancer relapse.

Testing on Pawitan

ER+	jPLS	sPLS	jLR	sLR
iPLS	0.98	0.55	0.61	0.5
jPLS	-	0.53	0.56	0.48
sPLS	-	-	0.39	0.04
jLR	-	-	-	0.56
ER-	jPLS	sPLS	jLR	sLR
iPLS	0.98	0.75	0.72	0.44
jPLS	-	0.74	0.69	0.42
sPLS	-	-	0.64	0.55
jLR	-	-	-	0.4

Testing on Ivshina

ER+	jPLS	sPLS	jLR	sLR
iPLS	0.78	0.32	0.32	0.66
jPLS	-	0.33	0.21	0.26
sPLS	-	-	0.5	0.09
jLR	-	-	-	0.28
ER-	jPLS	sPLS	jLR	sLR
iPLS	0.99	0.7	0.74	0.56
jPLS	-	0.72	0.75	0.56
sPLS	-	-	0.59	0.25
jLR	-	-	-	0.49

Testing on Wang

ER+	jPLS	sPLS	jLR	sLR
iPLS	1.0	0.4	0.59	0.4
jPLS	-	0.4	0.59	0.4
sPLS	-	-	0.19	0.32
jLR	-	-	-	0.26
ER-	jPLS	sPLS	jLR	sLR
iPLS	0.99	0.76	0.68	0.56
jPLS	-	0.73	0.67	0.57
sPLS	-	-	0.61	0.39
jLR	-	-	-	0.58

Testing on Sotiriou

ER+	jPLS	sPLS	jLR	sLR
iPLS	0.53	0.42	0.28	0.33
jPLS	-	0.73	0.62	0.31
sPLS	-	-	0.45	0.09
jLR	-	-	-	0.22
ER-	jPLS	sPLS	jLR	sLR
iPLS	1.0	0.88	0.77	0.54
jPLS	-	0.88	0.77	0.54
sPLS	-	-	0.69	0.45
jLR	-	-	-	0.46

APPENDIX C

Supplementary Material for "An ANOVA-based Procedure for PCA, Decomposing Variation and Dimensionality"

C.1 Principal Angles

Principal angles provide a notion of distance between subspaces, which is useful for studying the gPCA decomposition.

Definition. *Given D-dimensional linear subspaces $\mathcal{Y}, \mathcal{Z} \subseteq \mathbb{R}^p$, the principal angles $\theta_1 \leq \cdots \leq \theta_D \in [0, \pi/2]$ between $\mathcal{Y}, \mathcal{Z}$ are defined recursively as follows:*

$$\cos\theta_d = \max_{u \in \mathcal{Y}, v \in \mathcal{Z}} u^T v = u_d^T v_d, \| u \|_2 = \| v \|_2 = 1, u^T u_i = v^T v_i = 0, i = 1, \ldots, d-1.$$

Each principal angle is defined as the smallest possible angle between two vectors chosen from $\mathcal{Y}, \mathcal{Z}$, such that these chosen vectors are orthogonal to all of its predecessors. The principal vectors u_d, v_d are not uniquely defined, whereas the principal angles θ_d are.

Let matrices Y, Z denote the orthogonal projections onto subspaces $\mathcal{Y}, \mathcal{Z}$ respectively, and let $Z_\perp$ be the orthogonal projection onto the orthogonal complement of $\mathcal{Z}$. The following properties relate the principal angles to these projections.

Lemma C1. *The singular values of $Y^T Z$ are the cosines of the principal angles between $\mathcal{Y}, \mathcal{Z}$ (Björch and Golub, 1973):*

$$s(Y^T Z) = \{\cos\theta_1, \ldots, \cos\theta_D\}.$$

Lemma C2. *The singular values of $Y^T Z_\perp$ are the sines of the principal angles between $\mathcal{Y}, \mathcal{Z}$ (Qiu et al., 2005).*

$$s(Y^T Z_\perp) = \{\sin\theta_D, \ldots, \sin\theta_1\}.$$

The Grassmann manifold of D-planes in $\mathbb{R}^p$, denoted $\text{Grass}(D, p)$, is the space of all D-dimensional linear subspaces of $\mathbb{R}^p$. The following lemma provides the probability distribution for the largest principal angle (Absil et al., 2006).

Lemma C3. *Let subspaces $\mathcal{Y}, \mathcal{Z}$ be independently generated from the uniform distribution on* $\mathrm{Grass}(D, p)$ *with* $D < \frac{p+1}{2}$. *Then the probability distribution function of the largest principal angle θ_D between $\mathcal{Y}, \mathcal{Z}$ is given by:*

$$P(\theta_D < \theta) = \frac{\Gamma\left(\frac{D+1}{2}\right)\Gamma\left(\frac{p-D+1}{2}\right)}{\Gamma\left(\frac{1}{2}\right)\Gamma\left(\frac{p+1}{2}\right)}(\sin\theta)^{D(p-D)}$$

$$F_{2,1}(\frac{p-D}{2}, \frac{1}{2}; \frac{p+1}{2}; (\sin^2\theta)I_D), \theta \in [0, \frac{\pi}{2}),$$

where $\Gamma(z) = \int_0^\infty t^{z-1}e^{-t}dt$ *is the gamma function and* $F_{2,1}$ *is the Gaussian hypergeometric function of matrix argument.*

C.2 Approximation of WSS/BSS

The ratio between the within and between sum-of-squares is $\mathrm{WSS}/\mathrm{BSS} = \sum_k \mathrm{WSS}_k / \sum_k \mathrm{BSS}_k$ where:

$$\mathrm{WSS}_k = \| X_k - X_k\hat{W}_k\hat{W}_k^T \|_F^2, \mathrm{BSS}_k = \| X_k\hat{W}_k\hat{W}_k^T - X_k\hat{W}_k\hat{W}_k^T\hat{W}_s\hat{W}_s^T \|_F^2.$$

In this section we derive an approximation for this quantity. Script letters denote the subspaces generated from the span of the columns of the specified loading matrix (e.g. $\mathcal{W}_k = \mathrm{span}(W_k), \mathcal{W}_s = \mathrm{span}(W_s)$). Principal angles (in ascending magnitude) between subspaces $\mathcal{W}_k, \mathcal{W}_{k'}$ and $\mathcal{W}_k, \mathcal{W}_s$ are denoted as $\theta_{k,k';d}, \theta_{k,s;d}, d = 1, \ldots, D$ respectively.

1. We begin by adopting some distributional assumptions. Suppose that we have K datasets (letting $X_k \in \mathbb{R}^{N_k \times p}$) generated under a basic factor model:

   $$X_k = Z_k W_k^T + E_k, (Z_k)_{i\cdot} \sim \text{i.i.d.}\mathcal{N}(0, I_p), (E_k)_{i\cdot} = \text{i.i.d.}\mathcal{N}(0, \sigma^2 I_p). \quad \text{(C1)}$$

   Each X_k is generated from a product of true scores ($Z_k \in \mathbb{R}^{N \times D}$) and loadings ($W_k \in \mathbb{R}^{p \times D}$) with normally distributed noise. As a result, the data lie primarily on the D-dimensional subspaces $\mathcal{W}_k$ generated from the true loadings W_k of each group.

   Furthermore, we allow these loadings to be mixtures of common (W_c) and distinct ($W_{d,k}$) parts:

$$W_k = \alpha W_c + \sqrt{1-\alpha^2} W_{d,k},$$

$$W_c^T = (I_D \quad \mathbf{0}_{D\times(p-D)}), W_{d,k}^T = (\mathbf{0}_{D\times kD} \quad I_D \quad \mathbf{0}_{D\times(p-(k+1)D)}).$$

In other words, we have:

$$W_k^T = (\alpha I_D \quad \mathbf{0}_{D\times(k-1)D} \quad \alpha_c I_D \quad \mathbf{0}_{D\times(p-(k+1)D)}). \tag{C2}$$

The columns of $W_c, W_{d,k}$ are orthonormal in the sense that $W_c^T W_c = W_{d,k}^T W_{d,k} = I_D$ and $W_c^T W_{d,k} = \mathbf{0}_{D\times D}$, so that the columns of each W_k form an orthonormal basis ($W_k^T W_k = I_D$). The parameters α and σ denote the levels of commonality between groups and noise within groups respectively.

For simplicity, we assume that the X_k are already column-centered and normalized:

$$X_k^T \mathbf{1}_{N_k} = \mathbf{0}_p, \| X_k \|_F^2 = \| X_{k'} \|_F^2 = 1, \forall k, k'. \tag{C3}$$

We additionally assume that the empirical separate loadings coincide with the truth:

$$\hat{W}_k = W_k \forall k. \tag{C4}$$

Although this never holds perfectly except in an asymptotic setting, our final model tends to be robust to imperfect estimation of loadings as it relies more so on the variability produced by these loadings than their exact orientation.

2. Under these assumptions, we write:

$$\mathrm{WSS}_k = \| X_k(I_p - W_k W_k^T) \|_F^2, \mathrm{BSS}_k = \| X_k W_k W_k^T (I_p - W_s W_s^T) \|_F^2.$$

Here, $W_s = \underset{W^T W = I_D}{\operatorname{argmin}} \| X_s - X_s W W^T \|_F^2$ where $X_s = (X_k W_k W_k^T)_{k=1}^K \in \mathbb{R}^{N\times p}$.

Lemma C4. *Given Assumptions C1-C4,* WSS_k *follows a scaled chi-squared distribution:*

$$\mathrm{WSS}_k/\sigma^2 \sim \chi^2_{N_k(p-D)}.$$

Proof. Under Assumptions C1-C4, the values which contribute to WSS_k are portions of the noise E_k which are orthogonal to $\mathcal{W}_k$:

$$\mathrm{WSS}_k = \| E_k(I_p - W_k W_k^T) \|_F^2.$$

Since the distribution of E_k is rotationally invariant and $W_k W_k^T$ is a rank-D orthogonal projection, the sum-of-squares of each row of $E_k(I_p - W_k W_k^T)$ is distributed as χ^2_{p-D} scaled by σ. □

3. The subspace $\mathcal{W}_s$ is defined as capturing the maximal variation among the combined PCA approximations $(X_k W_k W_k^T)_{k=1}^K$, which suggests that $\mathcal{W}_s$ is oriented around the "center" of the subspaces of each group $\mathcal{W}_k$. We approximate this center by normalizing the columns of $\sum_k W_k$, which we denote by $W_s \propto \sum_k W_k$. Thus given Assumption C2, we are assuming:

$$W_s^T = (\alpha I_D \quad \tfrac{\alpha_c}{K} I_D \quad \cdots \quad \tfrac{\alpha_c}{K} I_D \quad \mathbf{0}_{D\times(p-(K+1)D)})/\sqrt{\alpha^2 + \alpha_c^2/K}, \qquad \text{(C5)}$$

where $\alpha_c = \sqrt{1-\alpha^2}$.

As with our construction of W_k from a mixture of common (W_c) and distinct (W_d) elements, this approximation of W_s is a mixture of the loadings (W_k) from each group. This approach is convenient for computations, but does not account for mixtures between columns of different rankings (e.g. $(W_k)_{\cdot 1}$ with $(W_{k'})_{\cdot 2}$) which may very well lead to better optimization of W_s. Moreover, weighting each group equally does not necessarily lead to the optimal W_s. These are important details to note, since imperfect estimation of W_s leads to overestimation of BSS. However, this does not appear to undermine overall performance, since our final estimator involves minimizing across a range of quantities so that the relative difference in WSS/BSS across rank choices plays a larger role than its precise magnitude.

The gPCA decomposition of sum-of-squares (Equation 4.3) is expected to hold under assumptions C1-C5. While Assumptions C1 and C2 are related to the structure of the model, Assumptions C3-C5 are suggestive of a noise-free environment. This decomposition is remarkably robust in numerical examples (Table C1), even as the noise level overwhelms the signal which muddles the estimation of the PCA loadings W_k. Where the gPCA method breaks down rather lies with the estimation of α.

Table C1: Distribution (mean and standard deviation) of the (WSS + BSS)/TSS ratio of gPCA across $\alpha \in \{0.0, 0.1, \ldots, 1.0\}$. Results are averaged over 25 repetitions. Specifications: $\{N_k, p, K\} = \{50, 100, 2\}$ and assuming correct selection of rank ($D = 2$).

σ	0.0	0.1	0.2	0.3	0.4	0.5	0.6	0.7	0.8	0.9	1.0
0.1	1.006	1.006	1.006	1.005	1.005	1.005	1.004	1.004	1.003	1.002	1.001
	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
0.5	1.013	1.013	1.013	1.013	1.013	1.014	1.014	1.014	1.015	1.016	1.017
	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
0.9	1.014	1.014	1.014	1.014	1.014	1.014	1.014	1.014	1.015	1.015	1.015
	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
1.3	1.014	1.014	1.014	1.014	1.014	1.014	1.014	1.014	1.014	1.015	1.015
	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0

The aspect of the gPCA method which is affected by high noise rather appears to be the estimation of α, as shown in Table C2. Estimation of the within- to between- sum-of-squares ratio is only stable if the common rank D is correctly estimated, but this relies on correct estimation of α. On the other hand, estimation of D is more forgiving of inaccurate estimation of WSS and BSS as only their ratio contributes to the entire procedure.

Table C2: Distribution (mean) of $\hat{\alpha}$ of gPCA across $\alpha \in \{0.0, 0.1, \ldots, 1.0\}$. Results are averaged over 25 repetitions. Specifications: $\{N_k, p, K\} = \{50, 100, 2\}$ and assuming correct selection of rank ($D = 2$).

σ	0.0	0.1	0.2	0.3	0.4	0.5	0.6	0.7	0.8	0.9	1.0
0.1	0.12	0.12	0.14	0.24	0.35	0.46	0.57	0.67	0.78	0.88	0.99
0.2	0.17	0.17	0.18	0.21	0.3	0.42	0.53	0.64	0.75	0.86	0.97
0.3	0.22	0.22	0.23	0.25	0.29	0.38	0.49	0.6	0.7	0.77	0.8
0.4	0.28	0.28	0.28	0.29	0.32	0.37	0.44	0.52	0.6	0.68	0.76

4. We can arrive at a similar distributional statement for the non-error portion of BSS_k, defined as:

$$\tilde{\text{BSS}}_k = \| Z_k W_k^T (I_p - W_s W_s^T) \|_F^2.$$

Regarding the projected errors $E_k W_k W_k^T (I_p - W_s W_s^T)$ as negligible, we use $\tilde{\text{BSS}}_k$ to approximate BSS_k.

Lemma C5. *Given Assumptions C1, C2, and C5,* $\tilde{\text{BSS}}_k$ *follows a scaled chi-squared distribution:*

$$\frac{\tilde{\text{BSS}}_k K}{(1-\alpha^2)(K-1)} \sim \chi^2_{N_k D}.$$

Proof. Let $W_{s,\perp}$ be a matrix whose columns span the orthogonal complement of $\mathcal{W}_s$. From Lemma C2, the singular values of $W_k^T W_{s,\perp}$ are the sines of the principal angles between $\mathcal{W}_k, \mathcal{W}_s$:

$$s(W_k^T W_{s,\perp}) = \{\sin\theta_{k,s;D}, \dots, \sin\theta_{k,s;1}\}.$$

This implies that the eigenvalues of $W_k^T(I_p - W_s W_s^T)W_k = W_k^T W_{s,\perp}(W_k^T W_{s,\perp})^T$ are the squared sines of these angles:

$$\lambda(W_k^T(I_p - W_s W_s^T)W_k) = \{\sin^2\theta_{k,s;D}, \dots, \sin^2\theta_{k,s;1}\}. \quad \text{(C6)}$$

From Assumptions *C1, C2, and C5*, the principal angles between subspaces $\mathcal{W}_k, \mathcal{W}_s$ are given by:

$$\cos\theta_{k,s;d} = \sqrt{\alpha^2 + \alpha_c^2/K}, d = 1, \dots, D,$$

which implies that:

$$\sin^2\theta_{k,s;d} = 1 - \cos^2\theta_{k,s;d} = 1 - \alpha^2 - \alpha_c/K = (1-\alpha^2)(K-1)/K. \quad \text{(C7)}$$

With Equations C6 and C7, we have:

$$\begin{aligned} \| Z_k W_k^T(I_p - W_s W_s^T) \|_F^2 &= \mathrm{tr}(Z_k W_k^T(I_p - W_s W_s^T)W_k U_k^T) \\ &= \mathrm{tr}\big(Z_k Q \sin^2\theta_{k,s;D} I_D Q^T Z_k^T\big) \text{ (spectral decomposition)} \\ &= \| Z_k \|_F^2 \ (1-\alpha^2)(K-1)/K, \end{aligned}$$

which follows the scaled chi-squared distribution of the theorem. □

The variation between the separate PCA approximations in BSS is thus explained by an angular relationship between the primary subspaces on which each data group lies.

5. Under Assumption C4 and considering the data X_k as stochastic, WSS and BSS are independent since $X_k(I_p - W_k W_k^T)$ and $X_k W_k W_k^T$ are independent. Therefore, combining Lemmas C4 and C5 (and assuming $\widetilde{\mathrm{BSS}}_k \approx \mathrm{BSS}_k$) gives our approximation for our sum-of-squares ratio:

$$\mathrm{WSS}/\mathrm{BSS} \approx \sigma^2(p-D)/\frac{D(1-\alpha^2)(K-1)}{K}. \quad \text{(C8)}$$

The numerator resembles that of the F-statistic from one-way ANOVA, with $p - D$ degrees of freedom associated with the within sum-of-squares.

However the denominator has a somewhat different form, which reflects the primary difference between decomposing variance in the univariate and multivariate settings. In ANOVA, the sum-of-squares centered around a point ($\overline{Y}$), whereas in gPCA the sum-of-squares is centered around effectively a subspace (XWW^T). Unlike comparisons in Euclidean space, such comparisons between subspaces involve principal angles and their associated geometric properties.

C.3 Estimation of Commonality and Complexity

From Equation C8, we obtain the estimates for complexity $\hat{D}$ and commonality $\hat{\alpha}$ for the gPCA procedure via an estimate of noise level $\hat{\sigma}$.

1. Applying Lemma C1 to Assumption C2, we have that the principal angles between subspaces $\mathcal{W}_k, \mathcal{W}_{k'}$ of any two groups are given by:

$$s(W_k^T W_{k'}) = \{\cos\theta_{k,k';1}, \dots, \cos\theta_{k,k';D}\} = \{\alpha^2\}_{d=1}^{D}. \quad \text{(C9)}$$

This leads to a series of natural estimates for α from the singular values of the covariance between group loadings, averaged over all group pairs:

$$\frac{2}{K(K-1)} \sum_{k>k'} \sqrt{s_d(\hat{W}_k^T \hat{W}_{k'})}, d = 1, \dots, D,$$

where $s_d(\cdot)$ denotes the d-th largest singular value.

We consider only the estimate generated from the largest principal angle $\hat{\theta}_{k,k';D}$ (i.e. smallest singular value s_1), as it appears to produce the most accurate estimates for α:

$$\hat{\alpha}_D = \frac{2}{K(K-1)} \sum_{k>k'} \sqrt{s_{\min}(\hat{W}_k^T \hat{W}_{k'})}. \quad \text{(C10)}$$

The reason for this is not entirely understood, however only the largest principal angle has been established as a metric on subspaces (Zhang, 2005), whereas the identifiability condition fails to hold for the remaining angles. Moreover the sine

of the largest principal angle is equivalent to the projection 2-norm which is widely used in engineering applications (Absil, 2006).

2. The estimate in Equation C10 would never be used in practice as it relies on knowledge of the true common complexity D, which determines the dimensions of each $\hat{W}_k$. Thus, for a given rank d, our estimate of α is the following:

$$\hat{\alpha}_d = \frac{2}{K(K-1)} \sum_{k>k'} \sqrt{s_{\min}(\hat{W}_{k;\cdot d}^T \hat{W}_{k';\cdot d})}.$$

Applying this estimate to Equation C8 and rearranging terms gives an estimate for σ based on an input estimate for d:

$$\hat{\sigma}_d^2 = \frac{\text{WSS}}{\text{BSS}} \frac{d(K-1)(1-\hat{\alpha}_d^2)}{K(p-d)}.$$

3. To complete the estimates of α, σ, what remains is to reliably estimate D, the rank of all PCA approximations used in the gPCA decomposition. Interestingly, the minimizer of $\hat{\sigma}$ across a range of rank choices $d = 1, \ldots, D_{\max}$ tends to accurately select the true D.

$$\hat{D} = \underset{d=1,\ldots,D_{\max}}{\text{argmin}} \hat{\sigma}_d.$$

In classical PCA, the estimated level of noise always decreases for as the choice of rank increases (Tipping and Bishop, 1999). This is natural since each new principal component adds a new layer of complexity to the nested solution structure, allowing more data variation to be explained. Thus it is intriguing that our criterion $\hat{\sigma}_d$ is no longer monotonically related to the choice of rank, simply due to applying an ANOVA-like framework for decomposing variation. Whereas the classical PCA model seeks to increase the variation explained within a single dataset (smaller residual sum-of-squares), the gPCA model seeks a balance between increasing the variation explained within data groups (smaller WSS) and decreasing the variation attributable to group differences (larger BSS).

C.4 Significance Calculation for Commonality

We discuss here the computation of $p_\alpha(a) = P(\hat{\alpha} > a)$. Lemma C3 provides the probability distribution function for the largest principal angle between subspaces generated independently on Grass(D, p) with $D < \frac{p+1}{2}$. From Equation C9, we have $\theta_{k,k'} = \arccos \alpha^2$, which relates the principal angles between the gPCA loading subspaces and the true commonality level. This can be used to rewrite the distribution function as:

$$P(\hat{\alpha} > a) = \frac{\Gamma\left(\frac{D+1}{2}\right)\Gamma\left(\frac{p-D+1}{2}\right)}{\Gamma\left(\frac{1}{2}\right)\Gamma\left(\frac{p+1}{2}\right)}(1-a^4)^{\frac{D(p-D)}{2}}$$

$$F_{2,1}\left(\frac{p-D}{2}, \frac{1}{2}; \frac{p+1}{2}; (1-a^4)I_D\right), a \in (0,1].$$

This is the p-value for observing $\hat{\alpha}$ at least as large as a under the null hypothesis that the originating subspaces are independently and uniformly generated on the space of all D-dimensional linear subspaces of $\mathbb{R}^p$ (with $D < \frac{p+1}{2}$). Note that this is not the same as assuming $\alpha = 0$, which is equivalent to assuming that these subspaces are orthogonal.

For large p, we can approximate $\Gamma(\frac{p-D+1}{2})/\Gamma(\frac{p+1}{2})$ with Sterling's formula:

$$\begin{aligned}
&\Gamma\left(\frac{p-D+1}{2}\right)/\Gamma\left(\frac{p+1}{2}\right) \\
&\quad \approx \sqrt{\pi(p-D-1)}\left(\frac{p-D-1}{2e}\right)^{(p-D-1)/2} / \sqrt{\pi(p-1)}\left(\frac{p-1}{2e}\right)^{(p-1)/2} \\
&\quad = \sqrt{(p-D-1)/(p-1)}\left(\frac{p-D-1}{p-1}\right)^{(p-D-1)/2} / \left(\frac{p-1}{2e}\right)^{D/2} \\
&\quad = \left(1 - \frac{D/2}{(p-1)/2}\right)^{(p-D)/2} / \left(\frac{p-1}{2e}\right)^{D/2} \approx \left(\frac{2}{p-1}\right)^{D/2}.
\end{aligned}$$

C.5 Supplementary Tables for the Growth Factor Responsiveness Study

Tables C3-C6 summarize the output obtained from applying gPCA to compare ligand response patterns among cell lines across ligand concentration, kinase type, time of measurement, and breast cancer subtype.

Table C3: Summary of gPCA findings for ligand responsiveness among ligand types across ligand concentrations. Data dimensions: (39 cell lines × 15 ligands) × 2 concentrations. Significance (denoted by shading) was assessed at level $0.05/m$ for $m = 6$ independent tests in each scenario).

Specifications	Factor levels	$\hat{D}$	$\hat{\alpha}$	p_1	p_2	$\tilde{\text{WSS}}$
AKT; 10 min	1, 100 ng/ml	1	0.949	1.71E-17	0.186	0.939
AKT; 30 min	1, 100 ng/ml	1	0.934	2.88E-15	0.221	0.929
AKT; 90 min	1, 100 ng/ml	1	0.924	4.91E-14	0.280	0.923
ERK; 10 min	1, 100 ng/ml	1	0.916	3.56E-13	0.334	0.916
ERK; 30 min	1, 100 ng/ml	1	0.945	7.54E-17	0.217	0.933
ERK; 90 min	1, 100 ng/ml	1	0.939	5.04E-16	0.215	0.931

Table C4: Summary of gPCA findings for ligand responsiveness among ligand types across kinase types. Data dimensions: (39 cell lines × 15 ligands) × 2 kinases. Significance (denoted by shading) was assessed at level $0.05/m$ for $m = 6$ independent tests in each scenario).

Specifications	Factor levels	$\hat{D}$	$\hat{\alpha}$	p_1	p_2	$\tilde{\text{WSS}}$
1 ng/ml; 10 min	AKT, ERK	1	0.864	2.47E-09	0.270	0.862
1 ng/ml; 30 min	AKT, ERK	1	0.835	7.50E-08	0.059	0.841
1 ng/ml; 90 min	AKT, ERK	1	0.835	7.26E-08	0.069	0.845
100 ng/ml; 10 min	AKT, ERK	1	0.871	9.85E-10	0.231	0.866
100 ng/ml; 30 min	AKT, ERK	1	0.869	1.32E-09	0.062	0.863
100 ng/ml; 90 min	AKT, ERK	1	0.836	6.89E-08	0.043	0.844

Table C5: Summary of gPCA findings for ligand responsiveness among ligand types across times of measurement. Data dimensions: (39 cell lines × 15 ligands) × 3 time points. Significance (denoted by shading) was assessed at level $0.05/m$ for $m = 12$ independent tests in each scenario).

Specifications	Factor levels	$\hat{D}$	$\hat{\alpha}$	p_1	p_2	$\tilde{\text{WSS}}$
1 ng/ml; AKT	10, 30, 90 min	2	0.992	2.47E-66	0.221	0.927
	10, 30 min	1	0.938	7.23E-16	0.332	0.935
	10, 90 min	1	0.939	5.60E-16	0.342	0.935
	30, 90 min	1	0.936	1.60E-15	0.329	0.935
1 ng/ml; ERK	10, 30, 90 min	2	0.931	2.98E-28	0.003	0.665
	10, 30 min	1	0.856	7.42E-09	0.089	0.852
	10, 90 min	1	0.862	3.52E-09	0.085	0.854
	30, 90 min	1	0.95	9.21E-18	0.283	0.945
100 ng/ml; AKT	10, 30, 90 min	2	0.988	2.08E-59	0.193	0.931
	10, 30 min	1	0.945	6.65E-17	0.359	0.938
	10, 90 min	1	0.96	1.18E-19	0.349	0.952
	30, 90 min	1	0.939	5.38E-16	0.243	0.938
100 ng/ml; ERK	10, 30, 90 min	2	0.943	2.08E-31	0.003	0.697
	10, 30 min	1	0.879	3.12E-10	0.096	0.869
	10, 90 min	1	0.885	1.33E-10	0.087	0.871
	30, 90 min	1	0.958	3.62E-19	0.279	0.948

Table C6: Summary of gPCA findings for ligand responsiveness among ligand types across disease subtypes. Data dimensions: 18 + 11 + 10 cell lines × 15 ligands. Significance (denoted by shading) was assessed at level $0.05/m$ for $m = 48$ independent tests in each scenario). See second portion on the next page.

Specifications	Factor levels	$\hat{D}$	$\hat{\alpha}$	p_1	p_2	$\tilde{WSS}$
1 ng/ml; AKT; 10 min	TN, H2, HR	1	0.729	3.43E-02	0.022	0.766
	TN, H2	1	0.678	7.31E-02	0.132	0.857
	TN, HR	1	0.679	7.25E-02	0.028	0.771
	H2, HR	1	0.851	1.53E-03	0.146	0.865
1 ng/ml; AKT; 30 min	TN, H2, HR	1	0.804	6.86E-03	0.052	0.822
	TN, H2	1	0.745	2.59E-02	0.016	0.869
	TN, HR	1	0.695	5.84E-02	0.028	0.815
	H2, HR	1	0.931	1.46E-05	0.197	0.904
1 ng/ml; AKT; 90 min .	TN, H2, HR	1	0.783	1.15E-02	0.061	0.818
	TN, H2	1	0.715	4.30E-02	0.169	0.854
	TN, HR	1	0.76	1.91E-02	0.055	0.849
	H2, HR	1	0.856	1.23E-03	0.192	0.873
1 ng/ml; ERK; 10 min	TN, H2, HR	1	0.995	3.48E-13	0.121	0.912
	TN, H2	1	0.994	1.14E-12	0.121	0.924
	TN, HR	1	0.993	2.85E-12	0.070	0.936
	H2, HR	1	0.997	1.70E-14	0.244	0.934
1 ng/ml; ERK; 30 min	TN, H2, HR	1	0.925	2.28E-05	0.043	0.814
	TN, H2	1	0.927	1.92E-05	0.184	0.865
	TN, HR	1	0.917	4.49E-05	0.026	0.837
	H2, HR	1	0.944	3.54E-06	0.057	0.820
1 ng/ml; ERK; 90 min	TN, H2, HR	1	0.877	4.96E-04	0.046	0.758
	TN, H2	1	0.896	1.78E-04	0.266	0.841
	TN, HR	1	0.855	1.28E-03	0.030	0.788
	H2, HR	1	0.887	3.04E-04	0.074	0.779

See legend and first portion on the previous page.

Specifications	Factor levels	$\hat{D}$	$\hat{\alpha}$	p_1	p_2	$\tilde{WSS}$
100 ng/ml; AKT; 10 min	TN, H2, HR	1	0.689	6.35E-02	0.014	0.728
	TN, H2	1	0.641	1.14E-01	0.069	0.830
	TN, HR	1	0.625	1.34E-01	0.015	0.748
	H2, HR	1	0.838	2.46E-03	0.115	0.827
100 ng/ml; AKT; 30 min	TN, H2, HR	1	0.605	1.64E-01	0.009	0.687
	TN, H2	1	0.537	2.79E-01	0.044	0.794
	TN, HR	1	0.525	3.01E-01	0.011	0.709
	H2, HR	1	0.822	4.05E-03	0.074	0.797
100 ng/ml; AKT; 90 min	TN, H2, HR	1	0.673	7.84E-02	0.008	0.675
	TN, H2	1	0.61	1.56E-01	0.008	0.748
	TN, HR	1	0.615	1.49E-01	0.006	0.756
	H2, HR	1	0.77	1.53E-02	0.035	0.666
100 ng/ml; ERK; 10 min	TN, H2, HR	1	0.989	6.18E-11	0.125	0.881
	TN, H2	1	0.987	1.60E-10	0.044	0.905
	TN, HR	1	0.988	1.35E-10	0.098	0.929
	H2, HR	1	0.99	2.26E-11	0.180	0.885
100 ng/ml; ERK; 30 min	TN, H2, HR	1	0.918	4.11E-05	0.077	0.819
	TN, H2	1	0.915	5.40E-05	0.112	0.838
	TN, HR	1	0.923	2.87E-05	0.053	0.877
	H2, HR	1	0.928	1.77E-05	0.102	0.836
100 ng/ml; ERK; 90 min	TN, H2, HR	1	0.91	7.20E-05	0.072	0.826
	TN, H2	1	0.92	3.70E-05	0.152	0.884
	TN, HR	1	0.907	8.98E-05	0.047	0.866
	H2, HR	1	0.909	8.09E-05	0.084	0.820

BIBLIOGRAPHY

Absil,P.A. et al. (2006) On the largest principal angle between random subspaces. *Linear Algebra Appl.,* **414,** 288-294.

Amini,A.A. and Wainwright,M.J. (2009) High-dimensional analysis of semidefinite relaxations for sparse principal components. *Ann. Stat.,* **37,** 2877-2921.

Anderson,W.F. and Matsuno,R. (2006) Breast cancer heterogeneity: a mixture of at least two main types?. *J Natl. Cancer I.,* **98,** 948-951.

Angenendt,P. (2005) Progress in protein and antibody microarray technology. *Drug Discov. Today,* **10,** 503-511.

Ayers,M. et al. (2004) Gene expression profiles predict complete pathologic response to neoadjuvant paclitaxel and fluorouracil, doxorubicin, and cyclophosphamide chemotherapy in breast cancer. *J. Clin. Oncol.,* **22,** 2284-2293.

Ayllon,V. and O'Connor,R. (2007) PBK/TOPK promotes tumour cell proliferation through p38 MAPK activity and regulation of the DNA damage response. *Oncogene,* **26,** 3451-3461.

Banerjee,A. et al. (2005) Clustering with Bregman divergences. *J. Mach. Learn. Res.,* **6,** 1705-1749.

Bell,D. et al. (2011) Integrated genomic analyses of ovarian carcinoma. *Nature,* **474,** 609-615.

Benito,M. et al. (2004) Adjustment of systematic microarray data biases. *Bioinformatics,* **20,** 105-114.

Berry,M.W. et al. (2007) Algorithms and applications for approximate nonnegative matrix factorization. *Comput. Stat. Data Anal.,* **52,** 155-173.

Besse,P. and de Falguerolles,A. (1993) Application of resampling methods to the choice of dimension in principal component analysis. *Computer Intensive Methods in Statistics,* Physica-Verlag HD, 167-176.

Björck,A. and Golub,G.H. (1973) Numerical methods for computing angles between linear subspaces. *Math. Comput.,* **27,** 579-594.

Blenkiron,C. et al. (2007) MicroRNA expression profiling of human breast cancer identifies new markers of tumor subtype. *Genome Biol.,* **8,** R214.

Boulesteix,A.-L. (2004) PLS dimension reduction for classification with microarray data. *Stat. Appl. Genet. Mol. Biol.,* **3,** 1-30.

Breiman,L. (2001) Random forests. *Machine Learning,* **45,** 5-32.

Bro,R. et al. (2008) Cross-validation of component models: a critical look at current methods. *Anal. Bioanal. Chem.,* **390,** 1241-1251.

Brunet,J.P. et al. (2004) Metagenes and molecular pattern discovery using matrix factorization. *Proc. Natl. Acad. Sci.,* **101,** 4164-4169.

Carvalho,C.M. et al. (2008) High-dimensional sparse factor modeling: applications in gene expression genomics. *J. Am. Stat. Assoc.,* **103,** 1438-1456.

Chalhoub,N. and Baker,S.J. (2009) PTEN and the PI3-kinase pathway in cancer. *Annu. Rev. Pathol.,* **4,** 127-150.

Chanrion,M. et al. (2008) A gene expression signature that can predict the recurrence of tamoxifen-treated primary breast cancer. *Clin. Cancer Res.,* **14,** 1744-1752.

Chun,H. and Keleş,S. (2010) Sparse partial least squares regression for simultaneous dimension reduction and variable selection. *J. Roy. Stat. Soc. B (Stat. Methodol.),* **72,** 3-25.

Cline,M.S. et al. (2007) Integration of biological networks and gene expression data using Cytoscape. *Nat. Protoc.,* **2,** 2366-2382.

Conlon,E.M. et al. (2006) Bayesian models for pooling microarray studies with multiple sources of replications. *BMC Bioinform.,* **7,** 247.

Cook,C.E. et al. (2016) The European Bioinformatics Institute in 2016: data growth and integration. *Nucleic Acids Res.,* **44,** D20-D26.

Cox,D.R. (2006) *Principles of statistical inference,* Cambridge University Press.

Creighton,C.J. et al. (2012) Integrated analyses of microRNAs demonstrate their widespread influence on gene expression in high-grade serous ovarian carcinoma. *PLOS ONE,* **7,** e34546.

Dai,H. et al. (2005) A cell proliferation signature is a marker of extremely poor outcome in a subpopulation of breast cancer patients. *Cancer Res.,* **65,** 4059-4066.

Dai,J.J. et al. (2006) Dimension reduction for classification with gene expression microarray data. *Stat. Appl. Genet. Molec. Biol.,* **5,** 1544-6115.

Dai,L. et al. (2012) Bioinformatics clouds for big data manipulation. *Biol. Direct,* **7,** 43.

Das,P. et al. (2006) Low-dimensional, free-energy landscapes of protein-folding reactions by nonlinear dimensionality reduction. *P. Natl. Acad. Sci. USA,* **103,** 9885-9890.

De Jong, S. (1993) SIMPLS: an alternative approach to partial least squares regression. *Chemometr. Intell. Lab. Syst.,* **18,** 251-263.

De la Torre,F. and Black,M.J. (2001) Robust principal component analysis for computer vision. *Eighth International Conference on Computer Vision (ICCV01),* **1,** 362-369.

Devarajan,K. (2008) Nonnegative matrix factorization: an analytical and interpretive tool in computational biology. *PLoS Comput. Biol.,* **4,** e1000029.

Ein-Dor,L. et al. (2005) Outcome signature genes in breast cancer: is there a unique set?. *Bioinformatics,* **21,** 171-178.

Ellen,T.P. et al. (2008) NDRG1, a growth and cancer related gene: regulation of gene expression and function in normal and disease states. *Carcinogenesis,* **29,** 2-8.

Eterno,V. et al. (2016) AurkA controls self-renewal of breast cancer-initiating cells promoting wnt3a stabilization through suppression of miR-128. *Sci. Rep.,* **6,** 28436.

Fan,J. and Li,R. (2002) Variable selection for Cox's proportional hazards model and frailty model. *Ann. Stat.,* **30,** 74-99.

Finak,G. et al. (2008) Stromal gene expression predicts clinical outcome in breast cancer. *Nat. Med.,* **14,** 518-527.

Fort,G. and Lambert-Lacroix,S. et al. (2005) Classification using partial least squares with penalized logistic regression. *Bioinformatics,* **21,** 1104-1111.

Fotovati,A. et al. (2011) N-myc downstream-regulated gene 1 (NDRG1) a differentiation marker of human breast cancer. *Pathol. Oncol. Res.,* **17,** 525-533.

Gao,Y. and Church,G. (2005) Improving molecular cancer class discovery through sparse non-negative matrix factorization. *Bioinformatics,* **21,** 3970-3975.

Gehlenborg,N. et al. (2010) Visualization of omics data for systems biology. *Nature,* **7,** S56-S68.

Giacinti,C. and Giordano,A. (2006) RB and cell cycle progression. *Oncogene,* **25,** 5220-5227.

Gidskehaug,L. et al. (2007) A framework for significance analysis of gene expression data using dimension reduction methods. *BMC Bioinformatics,* **8,** 346.

Gruvberger,S. et al. (2001) Estrogen receptor status in breast cancer is associated with remarkably distinct gene expression patterns. *Cancer Res.,* **61,** 5979-5984.

Hastie,T. et al. (2009) *The Elements of Statistical Learning. 2nd edn.,* Springer, New York.

Hofree,M. et al. (2013) Network-based stratification of tumor mutations. *Nat. Methods,* **10,** 1108-1115.

Hoheisel,J.D. (2006) Microarray technology: beyond transcript profiling and genotype analysis. *Nat. Rev. Genet.,* **7,** 200-210.

Houtgraaf,J.H. et al. (2006) A concise review of DNA damage checkpoints and repair in mammalian cells. *Cardiovasc. Revasc. Med.,* **7,** 165-172.

Howe,D. et al. (2008) Big data: the future of biocuration. *Nature,* **455,** 47-50.

Hu,Z. et al. (2006) The molecular portraits of breast tumors are conserved across microarray platforms. *BMC Genomics,* **7,** 96.

Imielinski,M. et al. (2012) Integrated proteomic, transcriptomic, and biological network analysis of breast carcinoma reveals molecular features of tumorigenesis and clinical relapse. *Mol. Cell. Proteomics,* **11,** M111.014910.

Ivshina,A.V. et al. (2006) Genetic reclassification of histologic grade delineates new clinical subtypes of breast cancer. *Cancer Res.,* **66,** 10292-10301.

Jackson,D.A. (1993) Stopping rules in principal components analysis: a comparison of heuristical and statistical approaches. *Ecology,* **74,** 2204-2214.

Jauhiainen,A. et al. (2012) Transcriptional and metabolic data integration and modeling for identification of active pathways. *Biostatistics,* **13,** 748-761.

Jensen,S.T. et al. (2007) Bayesian variable selection and data integration for biological regulatory networks. *Ann. Appl. Stat.,* **1,** 612-633.

Jin,D. and Lee,H. (2015) A computational approach to identifying gene-microRNA modules in cancer. *PLoS Comput. Biol.,* **11,** e1004042.

Jörnsten,R. et al. (2011) Network modeling of the transcriptional effects of copy number aberrations in glioblastoma. *Mol. Syst. Biol.,* **7,**

Kass,R.E. and Raftery,A.E. (1995) Bayes factors. *J. Am. Stat. Assoc.,* **90,** 773-795.

Khatri,P. et al. (2012) Ten years of pathway analysis: current approaches and outstanding challenges. *PLoS Comput. Biol.,* **8,** e1002375.

Kim,H. and Park,H. (2007) Sparse non-negative matrix factorizations via alternating non-negativity-constrained least squares for microarray data analysis. *Bioinformatics,* **23,** 1495-1502.

Kim,J. and Park,H. (2008) Sparse nonnegative matrix factorization for clustering. *Technical Report, Georgia Institute of Technology,* GT-CSE-08-01.

Kim,K.I. et al. (2002) Face recognition using kernel principal component analysis. *IEEE Signal Proc. Let.,* **9,** 40-42.

Kim,S. et al. (2015) Differential expression of lipid metabolism-related proteins in different breast cancer subtypes. *PLoS One,* **10,** e0119473.

Kitano,H. (2002) Computational systems biology. *Nature,* **420,** 206-210.

Kolar,M. et al. (2014) Graph estimation from multi-attribute data. *J. Mach. Learn. Res.,* **15,** 1713-1750.

Kritchman,S. and Nadler,B. (2008) Determining the number of components in a factor model from limited noisy data. *Chemometr. Intell. Lab.,* **94,** 19-32.

Kruskal,J.B. (1964) Multidimensional scaling by optimizing goodness of fit to a nonmetric hypothesis. *Psychometrika,* **29,** 1-27.

Lê Cao,K.-A. et al. (2008) A sparse PLS for variable selection when integrating omics data. *Stat. Appl. Genet. Mol. Biol.,* **7,** 1-32.

Lee,D.D. and Seung,H.S. (1999) Learning the parts of objects by non-negative matrix factorization. *Nature,* **401,** 788-791.

Lee,D.D. and Seung,H.S. (2001) Algorithms for non-negative matrix factorization. *Adv. Neural Inform. Proc. Syst.,* **13,** 556-562.

Leek,J.T. and Storey,J.D. (2007) Capturing heterogeneity in gene expression studies by surrogate variable analysis. *PLoS Genet.,* **3,** 1724-1735.

Li,C. and Li,H. (2008) Network-constrained regularization and variable selection for analysis of genomic data. *Bioinformatics,* **24,** 1175-1182.

Li,W. et al. (2012) Identifying multi-layer gene regulatory modules from multi-dimensional genomic data. *Bioinformatics,* **28,** 2458-2466.

Lin,C.J. (2007) On the convergence of multiplicative update algorithms for nonnegative matrix factorization. *IEEE Trans. Neural Netw.,* **18,** 1589-1596.

Lock,E.F. and Dunson,D.B. (2013) Bayesian consensus clustering. *Bioinformatics,* **29,** 2610-2616.

Lock,E.F. et al. (2013) Joint and individual variation explained (JIVE) for integrated analysis of multiple data types. *Ann. Appl. Stat.,* **7,** 523–542.

Ma,S. et al. (2011) Integrative analysis and variable selection with multiple high-dimensional data sets. *Biostatistics,* **12,** 763-775.

Madsen,C.D. et al. (2015) STRIPAK components determine mode of cancer cell migration and metastasis. *Nat. Cell Biol.,* **17,** 68-80.

Mankad,S. and Michailidis,G. (2013) Structural and functional discovery in dynamic networks with non-negative matrix factorization. *Phys. Rev. E,* **88,** 042812.

Marron,J.S. et al. (2007) Distance-weighted discrimination. *J. Am. Stat. Assoc.,* **102,** 1267-1271.

Marx,V. (2013) Biology: the big challenges of big data. *Nature,* **498,** 255-260.

Milde-Langosch,K. et al. (2013) Validity of the proliferation markers Ki67, TOP2A, and RacGAP1 in molecular subgroups of breast cancer. *Breast Cancer Res. Treat.,* **137,** 57-67.

Miller,L.D. et al. (2005) An expression signature for p53 status in human breast cancer predicts mutation status, transcriptional effects, and patient survival. *Proc. Natl. Acad. Sci. USA,* **102,** 13550-13555.

Minka,T.P. (2000) Automatic choice of dimensionality for PCA. *Technical Report 514, Massachusetts Institute of Technology*.

Mitrea,C. et al. (2013) Methods and approaches in the topology-based analysis of biological pathways. *Front. Psychol.,* **4,** 278.

Mo,Q. et al. (2013) Pattern discovery and cancer gene identification in integrated cancer genomic data. *Proc. Natl. Acad. Sci.,* **110,** 4245-4250.

Nguyen,D.V. and Rocke,D.M. (2002) Tumor classification by partial least squares using microarray gene expression data. *Bioinformatics,* **18,** 39-50.

Niepel,M. et al. (2014) Analysis of growth factor signaling in genetically diverse breast cancer lines. *BMC Biol.,* **12,** 20.

Oka,K. et al. (2011) Nudix-type motif 2 in human breast carcinoma: A potent prognostic factor associated with cell proliferation. *Int. J. Cancer,* **128,** 1770-1782.

Olawale,F. and Garwe,D. (2010) Obstacles to the growth of new SMEs in South Africa: A principal component analysis approach. *Afr. J. Bus. Manage.,* **4,** 729-738.

Onitilo,A.A. et al. (2009) Breast cancer subtypes based on ER/PR and Her2 expression: comparison of clinicopathologic features and survival. *Clin. Med. Res.,* **7,** 4-13.

Ostapkowicz,A. et al. (2006) Lipid rafts remodeling in estrogen receptor–negative breast cancer is reversed by histone deacetylase inhibitor. *Mol. Cancer Ther.,* **5,** 238-245.

Paik,S. et al. (2004) A multigene assay to predict recurrence of tamoxifen-treated, node-negative breast cancer. *N. Engl. J. Med.,* **351,** 2817-2826.

Parise,C.A. et al. (2009) Breast cancer subtypes as defined by the estrogen receptor (ER), progesterone receptor (PR), and the human epidermal growth factor receptor 2 (HER2) among women with invasive breast cancer in California, 1999–2004. *Breast J.,* **15,** 593-602.

Pawitan,Y. et al. (2005) Gene expression profiling spares early breast cancer patients from adjuvant therapy: derived and validated in two population-based cohorts. *Breast Cancer Res.,* **7,** R953.

Pedregosa,F. et al. (2011) Scikit-learn: Machine learning in Python. *J. Mach. Learn. Res.,* **12,** 2825-2830.

Pérez-Enciso,M. and Tenenhaus,M. (2003) Prediction of clinical outcome with microarray data: a partial least squares discriminant analysis (PLS-DA) approach. *Hum. Genet.,* **112,** 581-592.

Pliarchopoulou,K. et al. (2013) Prognostic significance of RACGAP1 mRNA expression in high-risk early breast cancer: a study in primary tumors of breast cancer patients participating in a randomized Hellenic Cooperative Oncology Group trial. *Cancer Chemoth. Pharm.,* **71,** 45-255.

Prat,A. et al. (2010) Phenotypic and molecular characterization of the claudin-low intrinsic subtype of breast cancer. *Breast Cancer Res.,* **12,** R68.

Price,A.L. et al. (2006) Principal components analysis corrects for stratification in genome-wide association studies. *Nat. Genet.,* **38,** 904-909.

Pucci,S. et al. (2016) Carnitine palmitoyl transferase-1A (CPT1A): a new tumor specific target in human breast cancer. *Oncotarget,* **7,** 19982-19996.

Putluri,N. et al. (2014) Pathway-centric integrative analysis identifies RRM2 as a prognostic marker in breast cancer associated with poor survival and tamoxifen resistance. *Neoplasia,* **16,** 390-402.

Qiu,L. et al. (2005) Unitarily invariant metrics on the Grassmann space. *SIAM J. Matrix Anal. A.,* **27,** 507-531.

Rantala,J.K. et al. (2010) Integrative functional genomics analysis of sustained polyploidy phenotypes in breast cancer cells identifies an oncogenic profile for GINS2. *Neoplasia,* **12,** 877-888.

Rhodes,D.R. et al. (2002) Meta-analysis of microarrays. *Cancer Res.,* **62,** 4427-4433.

Roweis,S.T. and Saul,L.K. (2000) Nonlinear dimensionality reduction by locally linear embedding. *Science,* **290,** 2323-2326.

Roy,S. et al. (2013) Integrated module and gene-specific regulatory inference implicates upstream signaling networks. *PLoS Comput. Biol.,* **9,** e1003252.

Ruan,L. and Yuan,M. (2011) An empirical bayes' approach to joint analysis of multiple microarray gene expression studies. *Biometrics,* **67,** 1617-1626.

Schadt,E.E. et al. (2010) Computational solutions to large-scale data management and analysis. *Nat. Rev. Genet.,* **11,** 647-657.

Scharpf,R.B. et al. (2009) A Bayesian model for cross-study differential gene expression. *J. Am. Stat. Assoc.,* **104,** 1295-1310.

Shao,J. et al. (2011) Sparse linear discriminant analysis by thresholding for high dimensional data. *Ann. Stat.,* **39,** 1241-1265.

Shen,R. et al. (2004) Prognostic meta-signature of breast cancer developed by two-stage mixture modeling of microarray data. *BMC Genomics,* **5,** 94.

Shimo,A. et al. (2007) Elevated expression of protein regulator of cytokinesis 1, involved in the growth of breast cancer cells. *Cancer Sci.,* **98,** 174-181.

Somorjai,R.L. et al. (2003) Class prediction and discovery using gene microarray and proteomics mass spectroscopy data: curses, caveats, cautions. *Bioinformatics,* **19,** 1484-1491.

Sørlie,T. et al. (2003) Repeated observation of breast tumor subtypes in independent gene expression data sets. *Proc. Natl. Acad. Sci. USA,* **100,** 8418-8423.

Sotiriou,C. and Lajos,P. (2009) Gene-expression signatures in breast cancer. *N. Engl. J. Med.,* **360,** 790-800.

Sotiriou,C. et al. (2006) Gene expression profiling in breast cancer: understanding the molecular basis of histologic grade to improve prognosis. *J. Natl. Cancer Inst.,* **98,** 262-272.

Soysal,S. et al. (2013) PTP1B expression is an independent positive prognostic factor in human breast cancer. *Breast Cancer Res. Tr.,* **137,** 637-644.

Sra,S. and Dhillon,I.S. (2005) Generalized nonnegative matrix approximations with Bregman divergences. *J. Mach. Learn. Res.,* **18,** 283-290.

Srihari,S. and Ragan,M.A. (2013) Systematic tracking of dysregulated modules identifies novel genes in cancer. *Bioinformatics,* **29,** 1553-1561.

Stingo,F.C. et al. (2011) Incorporating biological information into linear models: A Bayesian approach to the selection of pathways and genes. *Ann. Appl. Stat.,* **5,**

Tamayo,P. et al. (2007) Metagene projection for cross-platform, cross-species characterization of global transcriptional states. *Proc. Natl. Acad. Sci.,* **104,** 5959-5964.

Teodoro,M.L. et al. (2003) Understanding protein flexibility through dimensionality reduction. *J. Comput. Biol.,* **10,** 617-634.

Teschendorff,A.E. et al. (2006) A consensus prognostic gene expression classifier for ER positive breast cancer. *Genome Biol.,* **7,** R101.

The Cancer Genome Atlas Network,. (2012) Comprehensive molecular portraits of human breast tumors. *Nature,* **490,** 61-70.

Ulfarsson,M.O. and Solo,V. (2008) Rank selection in noist PCA with sure and random matrix theory. *IEEE T. Acoust. Speech,* 3317-3320.

van Dijk,E.L. et al. (2014) Ten years of next-generation sequencing technology. *Trends Genet.,* **30,** 418-426.

van Iterson, et al. (2013) Integrated analysis of microRNA and mRNA expression: adding biological significance to microRNA target predictions. *Nucleic Acids Res.,* **41,** e146.

van't Veer,L.J. et al. (2002) Gene expression profiling predicts clinical outcome of breast cancer. *Nature,* **415,** 530-536.

Vapnik,V. (1998) *Statistical learning theory,* Wiley.

Ventimiglia,G. and Petralia,S. (2013) Recent advances in DNA microarray technology: an overview on production strategies and detection methods. *Bionanoscience,* **3,** 428-450.

Verhaak,R.G.W. et al. (2013) Prognostically relevant gene signatures of high-grade serous ovarian carcinoma. *J. Clin. Invest.,* **123,** 517-525.

Villman,K. et al. (2006) TOP2A and HER2 gene amplification as predictors of response to anthracycline treatment in breast cancer. *Acta Oncol.,* **45,** 590-596.

Vyas,S. and Kumaranayake,L. (2006) Constructing socio-economic status indices: how to use principal components analysis. *Health Policy Plann.,* **21,** 459-468.

Wang,B. et al. (2014) Similarity network fusion for aggregating data types on a genomic scale. *Nat. Methods,* **11,** 333-337.

Wang,S. et al. (2009) Hierarchically penalized Cox regression with grouped variables. *Biometrika,* **96,** 307-322.

Wang,W. et al. (2013) iBAG: integrative Bayesian analysis of high-dimensional multiplatform genomics data. *Bioinformatics,* **29,** 149-159.

Wang,Y. et al. (2005) Gene-expression profiles to predict distant metastasis of lymph-node-negative primary breast cancer. *Lancet,* **365,** 671-679.

Wegelin,J.A. (2000) A survey of partial least squares (PLS) methods, with emphasis on the two-block case. *Technical Report, University of Washington.*

Wei,Y. (2015) Integrative analyses of cancer data: a review from a statistical perspective. *Cancer Inform.,* **14,** 173.

Weigel,M.T. and Dowsett,M. (2010) Current and emerging biomarkers in breast cancer: prognosis and prediction. *Endocr.-Relat. Cancer,* **17,** R245-R262.

Weigelt,B. and Reis-Filho,J.S. (2010) Molecular profiling currently offers no more than tumour morphology and basic immunohistochemistry. *Breast Cancer Res.,* **12,** S5.

Weigelt,B. et al. (2012) Challenges translating breast cancer gene signatures into the clinic. *Nat. Rev. Clin. Oncol.,* **9,** 58-64.

Weinstein,J.N. et al. (2013) The cancer genome atlas pan-cancer analysis project. *Nat. Genet.,* **45,** 1113-1120.

Wirapati,P. et al. (2008) Meta-analysis of gene expression profiles in breast cancer: toward a unified understanding of breast cancer subtyping and prognosis signatures. *Breast Cancer Res.,* **10,** R65.

Witten,D.M. and Tibshirani,R.J.. (2009) Extensions of sparse canonical correlation analysis with applications to genomic data. *Stat. Appl. Genet. Molec. Biol.,* **8,** 1-27.

Witten,D.M. et al. (2009) Extensions of sparse canonical correlation analysis with applications to genomic data. *Stat. Appl. Genet. Molec. Biol.,* **8,** 1-27.

Wold,H. (1985) Partial least squares. *Encyclopedia of the Statistical Sciences,* **6,** 581-591.

Wold,S. et al. (1987) Principal component analysis. *Chemometr. Intell. Lab.,* **2,** 37-52.

Xing,C. and Dunson,D.B. (2011) Bayesian inference for genomic data integration reduces misclassification rate in predicting protein-protein interactions. *PLoS Comput. Biol.,* **7,** e1002110.

Yamada,K.M. and Araki,M. (2001) Tumor suppressor PTEN: modulator of cell signaling, growth, migration and apoptosis. *J. Cell Sci.,* **114,** 2375-2382.

Yang,Z. and Michailidis,G. (2015) A non-negative matrix factorization method for detecting modules in heterogeneous omics multi-modal data. *Bioinformatics,* **32,** 1-8.

Yates,J.R. et al. (2009) Proteomics by mass spectrometry: approaches, advances, and applications. *Annu. Rev. Biomed. Eng.,* **11,** 49-79.

Yeung,K.Y. and Ruzzo,W.L. (2001) Principal component analysis for clustering gene expression data. *Bioinformatics,* **17,** 763-774.

Yoshida,K. and Miki,Y. (2004) Role of BRCA1 and BRCA2 as regulators of DNA repair, transcription, and cell cycle in response to DNA damage. *Cancer Sci.,* **95,** 866-871.

Yun,H.J. et al. (2008) Transcriptional targeting of gene expression in breast cancer by the promoters of protein regulator of cytokinesis 1 and ribonuclease reductase 2. *Exp. Mol. Med.,* **40,** 345-353.

Zhang,S. et al. (2011) A novel computational framework for simultaneous integration of multiple types of genomic data to identify microRNA-gene regulatory modules. *Bioinformatics,* **27,** i401-i409.

Zhang,S. et al. (2012) Discovery of multi-dimensional modules by integrative analysis of cancer genomic data. *Nucleic Acids Res.,* **40,** 9379-9391.

Zheng,M. et al. (2014) High GINS2 transcript level predicts poor prognosis and correlates with high histological grade and endocrine therapy resistance through

mammary cancer stem cells in breast cancer patients. *Breast Cancer Res. Tr.,* **148,** 423-436.

Zheng,X. et al. (2012) A high-performance computing toolset for relatedness and principal component analysis of SNP data. *Bioinformatics,* **28,** 3326-3328.

Zhu,R. et al. (2016) Integrating multidimensional omics data for cancer outcome. *Biostatistics,* **17,** 1-14.

Zou,H. et al. (2006) Sparse principal component analysis. *J Comput. Graph. Stat.,* **15,** 265-286.