Universidad Nacional de Colombia, Sede Bogotá
Facultad de Ciencias, Carrera de Estadística
Bogotá, noviembre de 2015

Predictive modeling of anticancer drug sensitivity with the Cancer Cell Line Encyclopedia

Undergraduate Research Experience Purdue-Colombia 2015

Andrés Nicolás López[a]
Author

Hyonho Chun[b]
Advisor

Departamento de Estadística, Facultad de Ciencias, Universidad Nacional de Colombia, Bogotá, Colombia
Department of Statistics, College of Science, Purdue University, West Lafayette IN, United States of America

Abstract

In cancer research, genomic data generated by high throughput experiments have been applied in drug sensivity prediction. Drug sensitivity phenotype measurements and a set of genetic features are obtained from a tumor culture with the aim of identifying genetic predictors of drug response. In this work, we present a predictive modeling approach of drug sensitivity considering the dependence structure across the cell lines with a linear mixed-effects model while capturing the underlying functionality of the cell lines by introducing the notion of endogenous variables regressing gene expression on genetic variants. Informative genetic variants for gene expression profiles are obtained through shrinkage methods for linear models, gene expression and drug sensitivity association is also regularized. The parameters of our model are optimized by 10-fold cross-validation for each compound and expression feature. The method is applied to data from the Cancer Cell Line Encyclopedia.

***Key words*:** Cancer Cell Line Encyclopedia, genetic variants, instrumental variables, mixed effect model, pharmacogenomics, predictive modeling.

Resumen

En investigación del cáncer, datos genómicos generados por experimentos de alto rendimiento han sido aplicados en la predicción de sensibilidad a compuestos anti cáncer. Características fenotípicas de sensibilidad a los compuestos y mediciones genómicas son obtenidas de un cultivo del tumor con el objetivo de identificar predictores genéticos de la respuesta al compuesto. En este trabajo presentamos una metodología para la predicción de sensibilidad a los compuestos considerando la dependencia entre las líneas celulares con un modelo lineal de efectos mixtos y a su vez capturando la funcionalidad subyacente de las líneas celulares mediante el uso de variables endógenas, regresando expresión génica en variantes genómicas. Variantes genómicas informativas para los perfiles de expresión son obtenidas a través de métodos de regularización en modelos lineales, la asociación entre expresión génica y sensibilidad al compuesto también es regularizada. Los parámetros del modelo son optimizados mediante validación cruzada en 10 particiones para cada compuesto y expresión génica. El método es aplicado a los datos de Cancer Cell Line Encyclopedia.

***Palabras clave*:** Cancer Cell Line Encyclopedia, variaciones genómicas, variables instrumentales, modelo de efectos mixtos, farmacogenómica, modelamiento predictivo.

[a] Estudiante de Estadística. E-mail: anlopezl@unal.edu.co
[b] Assistant Professor of Statistics. E-mail: chunh@purdue.edu

1. Introduction

In high throughput technologies several products are measured across different experimental units allowing a large amount of measurements per unit of analysis. The application of high throughput technologies in genomics has lead to the study of the mRNA expression of thousands of genes at the same time in a single experiment. High throughput microarray tecnologies as genomic tools have permitted not just to translate the human genome sequence into gene function (Bubendorf 2001), but also it has been used to detect single nucleotide polymorphisms, alterations in gene copy number, among other genomic applications. Further to this, DNA microarray tecnologies has made feasible to relate physiological cell states to gene expression patterns (Trevino et al. 2007).

In cancer research, the use of genomic characterization has allowed the discovery of gene signatures highly predictive for clinical outcomes (Trevino et al. 2007, Lin et al. 2015) as well as the evaluation of the influence of genetic differences in pharmacological responses. Owing to the inclusion of genetic information from the tumor sample for a more complete analysis of complex diseases and prediction of therapy response, cancer pharmacogenomics is responsible for the prediction of drug response based on a patient's genetic profile (Watters & McLeod 2003).

Large-scale pharmacogenomic studies provide a complete genomic characterization for several cancer cell lines coupled with in vitro pharmacological profiling allowing the joint study of drug response and genomic data from different cancer types. The Cancer Cell Line Encyclopedia (CCLE) is an aggregation of genetic and pharmacological characterization from a large collection of cell lines with the objective of associating distinct pharmacologic responses to genomic patterns. The CCLE is a collaboration between the Broad Institute and the Novartis Institutes for Biomedical Research (Barretina et al. 2012).

The CCLE has been a valuable source of information used to study specific cancer type genomic characteristics and drug sensitivity prediction either within a tumor lineage or across different cancer types. Nichols et al. (2014) studied head and neck squamous cell cancer (HNSCC) with the objective of clearly describing the mutational and copy number status of HNSCC cell lines and identifying candidate drugs with elevated efficacy in HNSCC. Sun & Liu (2015) performed a comparison of breast cancer cell lines with breast tumors based on the copy number variation and gene expression profiles. Palla et al. (2013) define a partition-valued process on an arbitrary covariate space using Gaussian processes. The process is used to construct two models, a multi task clustering model (MCM) which partitions data points in a similar way across multiple data sources and a network model. The MCM is applied to CCLE dataset to defining cancer subtypes. Berlow et al. (2014) proposed a novel approach for the drug sensitivity prediction model based on the genetic and functional data generated from the tumor culture. An integrative framework is based on prediction using genetic characterization and the possible targets of the drugs (a data extension that refers to the measure of inhibition of proteins that can be targeted by the drug). Furthermore, the CCLE study applied a regularization technique and a discrete classifier to identify predictive features of drug sensitivity. The mentioned studies do not acknowledge either the dependence structure across the cell lines nor the effect of external sources of variability that may affect the gene expression levels. [1]. The predictive performance of drug sensitivity might be enhanced by incorporating the use of genomic variants on gene expression estimation and the similarity structure across the cell lines.

The work of Lin et al. (2015) proposes a two-stage shrinkage method for high dimensional instrumental variable regression in order to assess the issue of confounding on a high dimensional scale. Motivated by the use of sparse instrumental variables in the study of complex traits, this study presents a drug sensitivity prediction model using genomic predictors as covariables, considering the dependence across the cell lines and the effect of external sources of variability in gene expression.

[1] It is noteworthy to mention that a recent publication (Zhang et al. 2015) has considered the cell line dependence for drug sensitivity prediction using the CCLE study as benchmark dataset. The implemented methodology does not consider the use of genomic variants for predicting drug response by introducing the notion of endogenous variables.

2. Materials

The CCLE project offers a compilation of genomic features from 947 human cancer cell lines coupled with pharmacological profiles for 24 anticancer drugs across 479 of the cell lines (Barretina et al. 2012). Figure 1 illustrates a simplified flow chart of the pharmacogenomic project. Multiple quality control steps were incorporated at each stage of cell culture and data production to ensure consistency of all datasets (Barretina et al. 2012) (Supplementary methods).

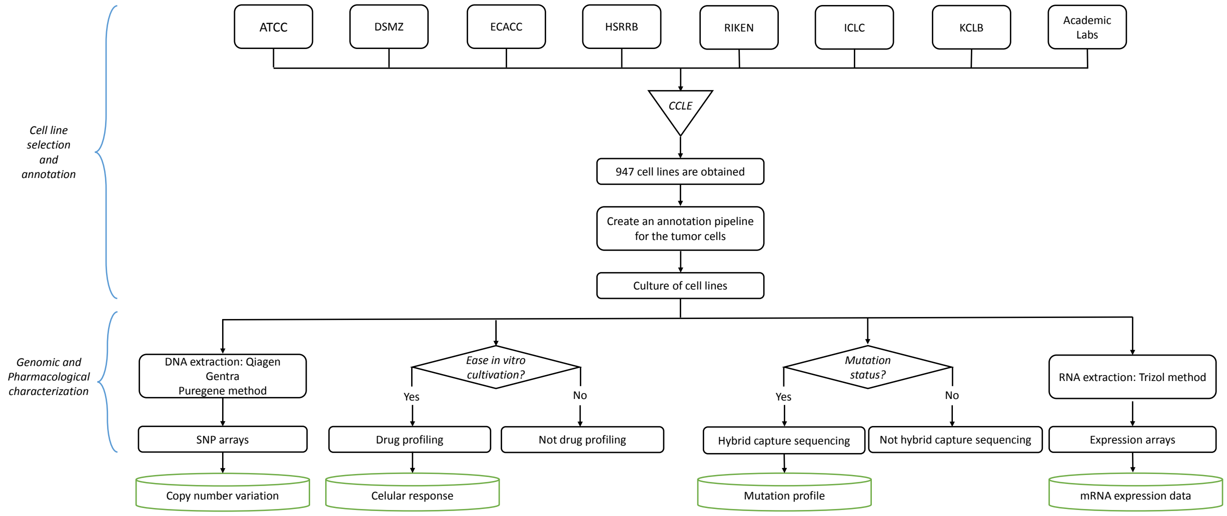
ATCC
DSMZ
ECACC
HSRRB
RIKEN
ICLC
KCLB
Academic Labs
CCLE
Cell line selection and annotation
947 cell lines are obtained
Create an annotation pipeline for the tumor cells
Culture of cell lines
Genomic and Pharmacological characterization
DNA extraction: Qiagen Gentra Puregene method
Ease in vitro cultivation?
Yes
No
Mutation status?
Yes
No
RNA extraction: Trizol method
SNP arrays
Drug profiling
Not drug profiling
Hybrid capture sequencing
Not hybrid capture sequencing
Expression arrays
Copy number variation
Celular response
Mutation profile
mRNA expression data

FIGURE 1: Simplified flow chart CCLE project. The CCLE offers a high throughput genetical genomic dataset coupled with drug sensitivity profiles. The datasets show at the bottom of the figure are used in the current analysis.

For the genomic characterization, gene expression profiles, copy number variation and mutation status for relevant human cell lines are provided. Copy number analysis was performed using genome-wide human Affymetrix SNP array 6.0. The expression data from CCLE was obtained using Affymetrix Human Genome U133 Plus 2.0 arrays. Mutation profiling using hybrid capture sequencing was obtained. For the pharmacological characterization in the CCLE, a total of up to 24 anti-cancer compounds for each cancer cell line were profiled. The cellular response of a target drug had between 6 to 8 point dose response, this data was summarized by a fitted model and quantitative measurements of this curves (drug sensitivities) were obtained.

Figure 2 represents a sigmoidal fit for a target cell line (adapted from Barretina et al. (2012)). The area between the response curve and a fixed or variable reference (activity area) was estimated as the sum of differences between the measured activities at different concentrations and the reference level. A small activity area value denotes a low reactive compound for a given target. Further to this, Figure 2 shows the $IC50$ value of the fitted curve, it is defined as the concentration at which the drug response reached an absolute inhibition of 50%. A low $IC50$ represents a highly reactive compound for the target.

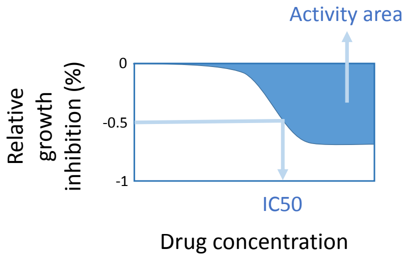

FIGURE 2: Sigmoid fit representation. Activity area and $IC50$.

As can be seen from Figure 2, the $IC50$ measurement only consider a single point in the fitted curve and activity area is measured over the complete curve. This implies that $IC50$ may be a noisier drug sensitivity measurement and also that activity area better captures the drug effects and cell responses (Zhang et al. 2015). Moreover, one of the recommended modeling guidelines for drug sensitivity prediction with gene expression data provided by Jang et al. (2014) highlight the use of the area over the fitted drug response curve as sensitivity measurement. For some of the compounds considered in the CCLE, activity area has shown to provide most accurate predictors in shrinkage regression methods (such as ridge and elastic net regression) (Jang et al. 2014).

Drug sensitivity data, gene expression profile, mutation profiling and copy number variation status were downloaded from the CCLE website `http://www.broadinstitute.org/ccle`. From the gene expression data type, the gctx file available at the CCLE website has the genomic and pharmacological profile from 486 common cancer cell lines. The gct file describing the gene expression profile across the cell lines and the Mutation Annotation File (MAF) corresponding to the mutation information against the reference genome were used to validate the gctx dataset. The MAF file was also used to obtain mutation features across the cell lines.

3. Methods

Drug sensitivity is predicted using gene expression as explanatory variable. Gene expression data is further estimated using mutation status and copy number variation profile across genes in order to control for external factors that may affect gene expression levels. This allows to jointly consider gene expression and genomic variants to predict drug profile. The current modeling approach makes use of gene expression profile across the cell lines, mutation status and copy number variation. Drug sensitivity measurements obtained from dose response curves are the response of interest.

3.1. Sample dependency

The association between the response variable and multiple regressors is assessed using a linear mixed-effects model which takes into account for the sample dependency with a random effect for each one of the cell lines. The response y_i on the ith cell line can be modeled as:

$$y_i = \beta_0 + \beta_1 x_{i1} + ... + \beta_p x_{ip} + \xi_i, \quad i = 1, ..., n \tag{1}$$

A standard regression model assuming independent errors is unreasonable given the dependence among the cell lines. The error term is then decomposed in error components:

$$\xi_i = b_i + \epsilon_i, \quad i = 1, ..., n \tag{2}$$

Thus, the error term is spitted in two parts: the dependent noise b_i that is given by the similarity of the cell lines and the independent noise ϵ_i. The two error terms ϵ_i and b_i are assumed to be independent of each other and b_i is considered to be a random intercept. For the random variables ϵ_i and b_i, $i = 1, ..., n$ a distribution is specified. Both ϵ_i (independent noise) and b_i (dependent noise) are assumed to be normally distributed, but the error terms ϵ_i are assumed to be independent whereas the random intercepts b_i are assumed to be correlated as the dependence between the cell lines (population structure) is introduced by this error.

In vectorized form, let $\boldsymbol{y} = (y_i, ..., y_n) \in \mathbb{R}^n$ be the response vector, $\boldsymbol{\beta} = (\beta_0, ..., \beta_p) \in \mathbb{R}^p$ the vector of fixed parameters, $\boldsymbol{Z} = \boldsymbol{I}_n$ an identity matrix of size n and $\boldsymbol{X} \in \mathbb{R}^{n \times p}$ the design matrix. Let $\boldsymbol{\epsilon} = (\epsilon_1, ..., \epsilon_n) \in \mathbb{R}^n$ and $\boldsymbol{b} = (b_1, ..., b_n) \in \mathbb{R}^n$ be the error terms. The model given in (1) can be rewritten as:

$$\boldsymbol{y} = \boldsymbol{X\beta} + \boldsymbol{Zb} + \boldsymbol{\epsilon} \tag{3}$$

As the dependent structure across the cell lines is modeled by $\boldsymbol{b}$, linear mixed-effect model given in (3) is a generalization of a multiple linear regression model.

The vector $\boldsymbol{\epsilon}$ of independent errors is assumed to have multivariate normal distribution with mean $\mathbf{0}$ and covariance matrix $\sigma_\epsilon^2 \boldsymbol{I}_n$. The vector $\boldsymbol{b}$ of dependent errors is assumed to be multivariate normally distributed with vector mean $\mathbf{0}$ and covariance matrix $\sigma_b^2 \boldsymbol{\Sigma}$ where $\boldsymbol{\Sigma} \in \mathbb{R}^{n \times n}$ represents the correlation structure of random intercepts, $\boldsymbol{\epsilon}$ and $\boldsymbol{b}$ are assumed to be independent.

Under this assumptions, the expected value of $\boldsymbol{y}$ given $\boldsymbol{X}$ is:

$$E(\boldsymbol{y}|\boldsymbol{X}) = \boldsymbol{X}\boldsymbol{\beta}$$

And the variance of $\boldsymbol{y}$ given $\boldsymbol{X}$ is:

$$V(\boldsymbol{y}|\boldsymbol{X}) = \sigma_b^2 \boldsymbol{Z}'\boldsymbol{\Sigma}\boldsymbol{Z} + \sigma_\epsilon^2 \boldsymbol{I}_n = \sigma_\epsilon^2 \left(\boldsymbol{I}_n + \rho\boldsymbol{\Sigma}\right) = \sigma_\epsilon^2 \boldsymbol{V}_\rho$$

With $\rho = \sigma_b^2/\sigma_\epsilon^2$ and $\boldsymbol{V}_\rho = (\boldsymbol{I}_n + \rho\boldsymbol{\Sigma})$. The tuning parameter ρ strikes a balance between dependent and independent noise:

- If ρ becomes larger, the random error is mostly explained by the dependent noise, it means that the error term $\boldsymbol{\xi} = (\xi_1, ..., \xi_n)$ can be explained by $\sigma_b^2 \boldsymbol{\Sigma}$ and we then can ignore the independent component $\boldsymbol{\epsilon}$.
- As the ρ value becomes smaller, the random error $\boldsymbol{\xi}$ turns into independent noise $\boldsymbol{\epsilon}$ and the model turns into a regular linear model.

We transform the model (3) multiplying by the inverse of the Cholesky decomposition $\boldsymbol{R}_\rho$ of the covariance matrix $\boldsymbol{V}_\rho$ at both sides of the equation so as to get independent noise for the error component:

$$\boldsymbol{R}_\rho^{-1}\boldsymbol{y} = \boldsymbol{R}_\rho^{-1}\boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{R}_\rho^{-1}\left(\boldsymbol{Z}\boldsymbol{b} + \boldsymbol{\epsilon}\right)$$

Let $\boldsymbol{y}_\rho = \boldsymbol{R}_\rho^{-1}\boldsymbol{y}$, $\boldsymbol{X}_\rho = \boldsymbol{R}_\rho^{-1}\boldsymbol{X}$ and $\boldsymbol{\delta}_\rho = \boldsymbol{R}_\rho^{-1}\left(\boldsymbol{Z}\boldsymbol{b} + \boldsymbol{\epsilon}\right) = \boldsymbol{R}_\rho^{-1}\left(\boldsymbol{b} + \boldsymbol{\epsilon}\right)$. Then we can rewrite model (3) as:

$$\boldsymbol{y}_\rho = \boldsymbol{X}_\rho\boldsymbol{\beta} + \boldsymbol{\delta}_\rho \tag{4}$$

The error term of the transformed model has expected value equal to zero:

$$E(\boldsymbol{\delta}_\rho) = E(\boldsymbol{R}_\rho^{-1}\left(\boldsymbol{b} + \boldsymbol{\epsilon}\right)) = \boldsymbol{R}_\rho^{-1}E\left(\boldsymbol{b} + \boldsymbol{\epsilon}\right) = \mathbf{0}$$

An a covariance structure that is homoscedastic and uncorrelated:

$$V(\boldsymbol{\delta}_\rho) = E\left[(\boldsymbol{\delta}_\rho - E(\boldsymbol{\delta}_\rho))(\boldsymbol{\delta}_\rho - E(\boldsymbol{\delta}_\rho))'\right] = E(\boldsymbol{\delta}_\rho\boldsymbol{\delta}_\rho') = \sigma_\epsilon \boldsymbol{I}_n$$

The first level of tuning in the model is given by the parameter ρ. Depending on which ρ value is used, we obtain a different parameter estimate $\boldsymbol{\beta}$. The second level of tuning is given by the penalization of the $\boldsymbol{\beta}$ parameter, it is required to obtain a sparse solution when $p >> n$.

3.1.1. Likelihood and tuning parameter

For the first level of tuning, we consider that the response variable $\boldsymbol{y}$ given the matrix $\boldsymbol{X}$ of input variables is normally distributed with mean $\boldsymbol{X}\boldsymbol{\beta}$ and covariance matrix $\sigma_\epsilon^2\left(\boldsymbol{I}_n + \rho\boldsymbol{\Sigma}\right)$. In explicit form:

$$f(\boldsymbol{y}|\boldsymbol{\beta}, \rho, \sigma_\epsilon^2) = \frac{1}{(2\pi\sigma_\epsilon^2)^{n/2}\left|(\boldsymbol{I} + \rho\boldsymbol{\Sigma})\right|^{1/2}} \exp\left\{-\frac{(\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta})^\top\left(\boldsymbol{I}_n + \rho\boldsymbol{\Sigma}\right)(\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta})}{2\sigma_\epsilon^2}\right\}$$

For a given ρ, the conditional estimates for $\boldsymbol{\beta}$ and σ_ϵ^2 that maximize the likelihood are given by:

$$\hat{\beta}(\rho) = \quad \left[\boldsymbol{X}^\top\left(\boldsymbol{I}_n + \rho\boldsymbol{\Sigma}\right)^{-1}\boldsymbol{X}\right]^{-1}\left[\boldsymbol{X}^\top\left(\boldsymbol{I}_n + \rho\boldsymbol{\Sigma}\right)^{-1}\boldsymbol{y}\right]$$

And

$$\hat{\sigma}_\epsilon^2(\rho) = \quad \left(\boldsymbol{y} - \boldsymbol{X}\hat{\beta}(\rho)\right)^\top\left(\boldsymbol{I}_n + \rho\boldsymbol{\Sigma}\right)^{-1}\left(\boldsymbol{y} - \boldsymbol{X}\hat{\beta}(\rho)\right)/n$$

The obtained expressions allows to derive the profiled likelihood as function of the tuning parameter ρ. The ρ value estimation obtained from the profiled likelihood is used to retrieve a regression model with independent error term. When required, once the variables are transformed, the regularization is performed.

3.1.2. Regularization

Constrained optimization methods in regression such as ridge regression, LASSO regression (Tibshirani 1996) and elastic net regularization (Zou & Hastie 2005) has been applied to a variety of prediction tasks using genomic data (Waldron et al. 2011). For the second level of tuning, we consider a penalization of the parameter vector $\boldsymbol{\beta}$ in order to perform variable selection when required. In a regular linear model, a method for grouped variable selection is performed with elastic net regularization.

Elastic net regression analysis performs better in several aspects compared with other regression techniques. In broad terms, elastic net algorithm overcomes ordinary least squares (OLS) in both prediction and interpretation, what is more, in a high dimensional setting (when the number of features p is larger than the number of observations n) OLS approach is not appropriate due to overfitting of the data (James et al. 2013). Unlike ridge regression, elastic net allows to obtain a parsimonious model making some of the coefficient estimates to be equal to zero. Even though LASSO technique perform variable selection, the possible grouping effect of the covariates is not taken into account by this methodology and as a result the LASSO tends to select randomly only one of the correlated variables, what is more, ridge regression has been shown to outperform the LASSO for prediction in low dimensional data (Zou & Hastie 2005).

The elastic net regression overcomes this difficulties with a shrinkage of the parameter vector which is a generalization of both the LASSO (l_1 penalization) and the ridge (l_2 penalization) regularization. It is worth pointing out that neither of the aforementioned methodologies consider the dependence across the samples.

The elastic net regression analysis requires a method for selecting the tuning parameters. For the CCLE dataset experimental units were spitted into ten non-overlapping groups. The optimal setting of the tuning parameters is chosen to minimize the mean square error using 10-fold cross validation.

3.1.3. Prediction of random effects

For the prediction of random effects in the linear mixed model given in (3) note that:

$$\begin{pmatrix} \boldsymbol{y} \\ \boldsymbol{b} \end{pmatrix} = \begin{pmatrix} \boldsymbol{X\beta} \\ \boldsymbol{0} \end{pmatrix} + \begin{pmatrix} \boldsymbol{Z} & \boldsymbol{I} \\ \boldsymbol{I} & \boldsymbol{0} \end{pmatrix} \begin{pmatrix} \boldsymbol{b} \\ \boldsymbol{\epsilon} \end{pmatrix} \tag{5}$$

Where

$$\begin{pmatrix} \boldsymbol{b} \\ \boldsymbol{\epsilon} \end{pmatrix} \sim N\left(\begin{pmatrix} \boldsymbol{0} \\ \boldsymbol{0} \end{pmatrix}, \begin{pmatrix} \sigma_b^2\boldsymbol{\Sigma} & \boldsymbol{0} \\ \boldsymbol{0} & \sigma_\epsilon^2\boldsymbol{I}_n \end{pmatrix} \right)$$

Under the model assumptions we have:

$$\begin{pmatrix} \boldsymbol{y} \\ \boldsymbol{b} \end{pmatrix} \sim N\left(\begin{pmatrix} \boldsymbol{X\beta} \\ \boldsymbol{0} \end{pmatrix}, \begin{pmatrix} \sigma_b^2\boldsymbol{Z\Sigma Z}' + \sigma_\epsilon^2\boldsymbol{I}_n & \sigma_b^2\boldsymbol{Z\Sigma} \\ \sigma_b^2\boldsymbol{\Sigma Z}' & \sigma_b^2\boldsymbol{\Sigma} \end{pmatrix} \right) \tag{6}$$

The conditional distribution of $\boldsymbol{b}$ given $\boldsymbol{y}$ is normal with vector mean:

$$\begin{aligned} \boldsymbol{\mu}_{\boldsymbol{b}|\boldsymbol{y}} =& \sigma_b^2\boldsymbol{\Sigma Z}'\left(\sigma_b^2\boldsymbol{Z\Sigma Z}' + \sigma_\epsilon^2\boldsymbol{I}_n\right)^{-1}(\boldsymbol{y} - \boldsymbol{X\beta}) \\ =& \sigma_b^2\boldsymbol{\Sigma}\left(\sigma_b^2\boldsymbol{\Sigma} + \sigma_\epsilon^2\boldsymbol{I}_n\right)^{-1}(\boldsymbol{y} - \boldsymbol{X\beta}) \\ =& \boldsymbol{\Sigma}\left(\boldsymbol{\Sigma} + \rho^{-1}\boldsymbol{I}_n\right)^{-1}(\boldsymbol{y} - \boldsymbol{X\beta}) \end{aligned}$$

And covariance matrix $\boldsymbol{\Sigma}_{\boldsymbol{b}|\boldsymbol{y}} = \sigma_b^2\boldsymbol{\Sigma}$. Given estimates of $\boldsymbol{\Sigma}$, $\boldsymbol{\beta}$ and ρ we obtain a prediction $\hat{\boldsymbol{b}}$ of the random effect $\boldsymbol{b}$ as $\hat{\boldsymbol{\mu}}_{\boldsymbol{b}|\boldsymbol{y}}$.

4. Results

In order to select the features that characterize the genomic dataset, the genes whose expression data varies the most across the different cell lines are selected. The gene expression profiles for 5000 genes with largest interquartile range are used as observed explanatory variables.

Not only the genomic characterization is used to predict drug efficiency, but also to assess similarity across the cell lines. The distance between cell lines is measured with their corresponding genomic features, which helps to create a similarity matrix. Figure 3 shows the first factorial plane for the gene expression data, this approximation by two dimensions is revealing the dependence existence across the cell lines by tissue. In addition, the graph shows that gene expression is quite varying even within the same tissue and also that different tissues can be similar to each other.

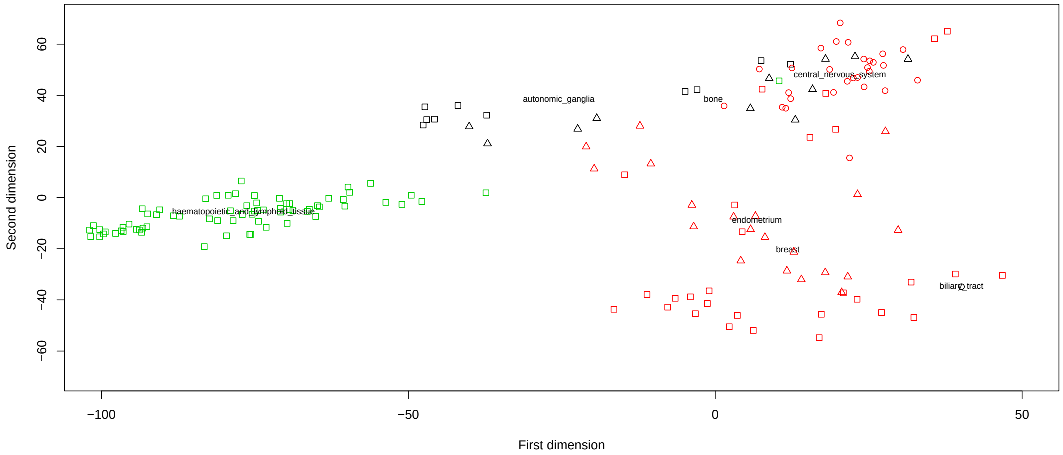

FIGURE 3: Singular Value Decomposition visualization based on the Euclidean distance. 5000 most varying gene expression are used. Cells from seven tissues are plotted.

It is known that gene expression data may differ by tissue and part of the observed dependence across cell lines may be driven by the functional similarity of the tissue and other external sources (Lin et al. 2015). A basal gene expression is estimated using instrumental variables methods in order to distinguish the underlying functionality of the cell using genetic variants as instrumental variables. This basal gene expression is outcome of interaction with structure of variation (genetic variants) and allows to distinguish between the signal and the noise from the expression data. Copy number variation and mutation data are used as instrument for the model. The impact on gene expression of mutations is analysed within cis-regulatory regions.

For each gene expression an association test is performed across each of the copy number profile considering the sample dependence. The copy number measurements were summarized by calling them into discrete values of *loss*, *normal* and *gain* for each sample through K-Means clustering.

4.1. Univariate feature selection

In order to reduce the number of noisy features across the cell lines a location-wise association test is performed. For each gene expression a marginal p value profile is obtained, each profile provides a list of differentially expressed features which will be used to pinpoint the structure of variation associated with gene expression. The union of significant structure features across the complete expression data are considered in the following analysis.

Figure 4 shows the image plot for the complete p value profiles by chromosome. Each p value plot is presented twice, first the p value profiles by chromosomes are plotted and then the gene expression (rows) are ordered with hierarchical clustering analysis to highlight their relatedness. The dissimilarity structure for the algorithm is based on the expression data.

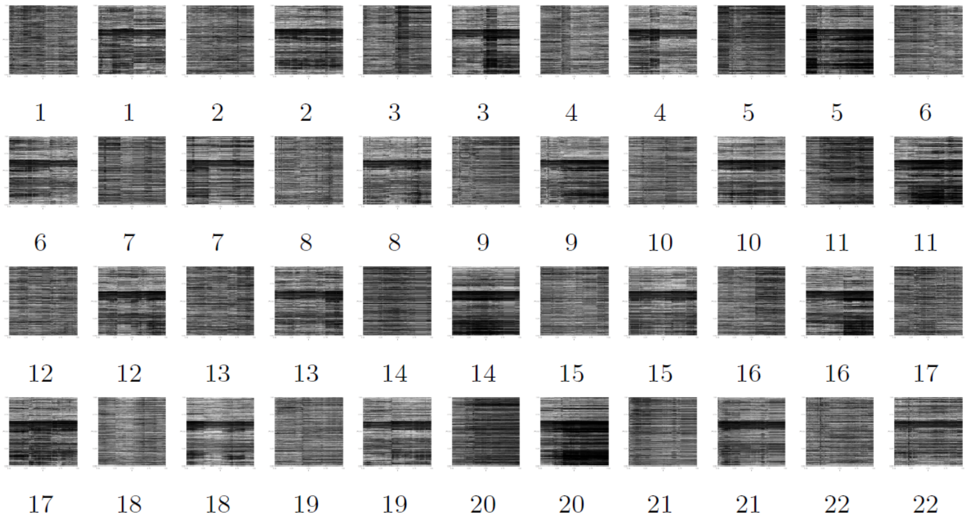

Figure 4: Image plot of p values profiles by chromosome (autosomes). Lower p values are represented in black.

4.2. Estimated gene expression

For the linear mixed effects model given in (3), the covariance matrix of the random effects is inferred with the genomic profile of the cell lines. Additionally, the conditional distribution of $\boldsymbol{b}$ is used to obtain a prediction of the random effects on the test data. Genomic variants are used as instruments of gene expression data.

A penalized multivariate regression approach is applied to determine optimal instruments from the obtained p value profiles. The differentially expressed genes by copy number calling is determined and shrinkage penalty is then applied to this feature subset in order to identify and estimate the nonzero fixed effects of the instruments to estimate gene expression data.

4.2.1. Population structure

Firstly, the population structure considered by the mixed linear effect model given in (3) is used to estimate gene expression without considering genomic variants. The importance of the random effect term is assessed through the tuning parameter ρ on training data. The tuning parameter ρ is used to retrieve an independent error model for gene expression in order to obtain an estimate for the fixed intercept of the model. Prediction of the random effects on test data is obtained with the conditional distribution of $\boldsymbol{b}$. Ten non-overlapping groups of cell lines are used where each one is used as test data so as to predict the random effects while the remaining cell lines are used to learn the fixed part of the model.

4.2.2. Population structure and genomic variants

Secondly, expression data is estimated from genomic variants considering the dependence structure across the cell lines. For the fixed part of the model, elastic net regression is performed where the optimal setting of the tuning parameters on training phase is estimated through 10-fold cross validation. The random part of the model is predicted similarly than in the previous scheme. Each fold in turn is used as test data and the remaining groups of cell lines are used to learn the fixed part of the model. Regularization is performed on the first stage of the procedure in order to obtain informative genomic variants.

Pre-selected genomic copy number features and mutation status are used to estimate gene expression. Elastic net regression is performed to identify relevant genomic variants in gene expression estimation. This informative genomic feature set is used to obtain a basal gene expression which later will be used to predict drug sensitivity.

5. Conclusions

Given the current availability of high throughput technologies for the study of complex traits and the large amount of information generated by large-scale pharmacogenomic studies, specialized statistical methodologies are required for the correct analysis of the vast amount of measurements obtained from the experimental units. With the purpose of understanding cell responses to different anti cancer compounds, the correct modeling of drug sensitivity allows an accurate detection of genomic predictors of drug response and the possibility of diminishing the error of drug response prediction.

The public available data from the Cancer Cell Line Encyclopedia provides access to high throughput data for genetic variants and gene expression profiles for hundreds of human cancer cell lines. When coupled with drug sensitivity profiles, the CCLE has allowed to simultaneously study the effects of different genomic features on drug response. Motivated by the work of Lin et al. (2015), we have made use of estimated gene expression for a given set of genomic variants in a high dimensional setting for predicting drug response. we have also considered the dependence structure across the cell lines with a linear mixed-effects model for the prediction of this complex trait. This modeling approach allows to jointly analyze gene expression, genetic variants and complex traits considering the dependence structure across samples.

Gene expression data from the CCLE showed the dependence existence across the cell lines, motivating the incorporation of genetic similarity of the cell lines for the prediction of drug sensitivity. Moreover, the available genomic variants of the profiled cell lines and the potential impact of external sources on gene expression levels is considered with the implementation of genomic variants as instruments of gene expression. The use of variants as optimal instruments provides a more effective use of this information when modeling drug response. The use of estimated gene expression obtained by using genomic variants allows to distinguish the underlying functionality of the cell lines by introducing the notion of endogenous variables.

Acknowledgements

Special thanks go to Professor Hyonho Chun for her guidance, patience and encouragement. To Professor Leonardo Trujillo and the Department of Statistics of Universidad Nacional de Colombia. To the Undergraduate Research Experience Purdue Colombia 2015. To the Department of Statistics of Purdue University.

References

Barretina, J., Caponigro, G., Stransky, N., Venkatesan, K., Margolin, A. A., Kim, S., Wilson, C. J., Lehár, J., Kryukov, G. V., Sonkin, D., Reddy, A., Liu, M., Murray, L., Berger, M. F., Monahan, J. E., Morais, P., Meltzer, J., Korejwa, A., Jané-Valbuena, J., Mapa, F. A., Thibault, J., Bric-Furlong, E., Raman, P., Shipway, A., Engels, I. H., Cheng, J., Yu, G. K., Yu, J., Aspesi, P., de Silva, M., Jagtap, K., Jones, M. D., Wang, L., Hatton, C., Palescandolo, E., Gupta, S., Mahan, S., Sougnez, C., Onofrio, R. C., Liefeld, T., MacConaill, L., Winckler, W., Reich, M., Li, N., Mesirov, J. P., Gabriel, S. B., Getz, G., Ardlie, K., Chan, V., Myer, V. E., Weber, B. L., Porter, J., Warmuth, M., Finan, P., Harris, J. L., Meyerson, M., Golub, T. R., Morrissey, M. P., Sellers, W. R., Schlegel, R. & Garraway, L. A. (2012), 'The Cancer Cell Line Encyclopedia enables predictive modelling of anticancer drug sensitivity'.

Berlow, N., Haider, S., Wan, Q., Geltzeiler, M., Davis, L. E., Keller, C. & Pal, R. (2014), 'An integrated approach to anti-cancer drug sensitivity prediction', *Computational Biology and Bioinformatics* **11**(6), 995–1008.

Bubendorf, L. (2001), 'High-throughput microarray technologies: from genomics to clinics.', *European urology* **40**(2), 231–8.
*http://www.ncbi.nlm.nih.gov

James, G., Witten, D., Hastie, T. & Tibshirani, R. (2013), *Springer Texts in Statistics An Introduction to Statistical Learning - with Applications in R.*

Jang, I. S., Neto, E. C., Guinney, J., Friend, S. H. & Margolin, A. A. (2014), 'Systematic assessment of analytical methods for drug sensitivity prediction from cancer cell line data.', *Pacific Symposium on Biocomputing. Pacific Symposium on Biocomputing* pp. 63–74.
*http://www.pubmedcentral.nih.gov

Lin, W., Feng, R. & Li, H. (2015), 'Regularization Methods for High-Dimensional Instrumental Variables Regression With an Application to Genetical Genomics', *Journal of the American Statistical Association* **110**(509), 270–288.

Nichols, A., Black, M., Yoo, J., Pinto, N., Fernandes, A., Haibe-Kains, B., , Paul C Boutros, B. & Barrett, J. (2014), 'Exploiting high-throughput cell line drug screening studies to identify candidate therapeutic agents in head and neck cancer', *Pharmacology and Toxicology* **15**, 1–10.

Palla, K., Knowles, D. & Ghahramani, Z. (2013), 'A dependent partition-valued process for multitask clustering and time evolving network modelling', p. 9.

Sun, Y. & Liu, Q. (2015), 'Deciphering the Correlation between Breast Tumor Samples and Cell Lines by Integrating Copy Number Changes and Gene Expression Profiles', *BioMed Research International* **2015**, 1–11.

Tibshirani, R. (1996), 'Regression and shrinkage via the Lasso', *J R Stat Soc, Ser B* **58**(1), 267–288.

Trevino, V., Falciani, F. & Barrera-Saldaña, H. A. (2007), 'Dna microarrays: a powerful genomic tool for biomedical and clinical research', *Molecular Medicine* **13**(9-10), 527.

Waldron, L., Pintilie, M., Tsao, M.-S., Shepherd, F. A., Huttenhower, C. & Jurisica, I. (2011), 'Optimized application of penalized regression methods to diverse genomic data.', *Bioinformatics (Oxford, England)* **27**(24), 3399–406.
*http://www.pubmedcentral.nih.gov

Watters, J. W. & McLeod, H. L. (2003), 'Cancer pharmacogenomics: Current and future applications', *Biochimica et Biophysica Acta - Reviews on Cancer* **1603**(2), 99–111.

Zhang, N., Wang, H., Fang, Y., Wang, J., Zheng, X. & Liu, X. S. (2015), 'Predicting Anticancer Drug Responses Using a Dual-Layer Integrated Cell Line-Drug Network Model', *PLoS Comput Biol* **11**.

Zou, H. & Hastie, T. (2005), 'Regularization and variable selection via the elastic net', *Journal of the Royal Statistical Society. Series B: Statistical Methodology* **67**(2), 301–320.