Integrative analysis of multi-platform genomics data

Shisi He* Qingqi Yue † Ao Yuan‡

Abstract

In multi-platform genomic data analysis, multiple large sequences are observed from normal and abnormal subjects, and the sample sizes are relatively small. How to analyze such data jointly is challenging. Existing methods include the Lasso or graphical Lasso by selecting only few of the data for the analysis, resulting in information loss and possible biases. Here we propose two new methods: the integrative correlation to characterize the innate relationships within the sequences, and the empirical process/the smoothing method to combine information for prediction. These methods take the full data into the analysis and are very simple to use. Simulation studies are conducted to evaluate the performance of the methods, and a real multi-platform genomic data is analyzed to illustrate the application of the second method.

Key Words: Genomic data, prediction, gene copy number, methylation, protein, SNP.

1. Introduction

With the development of biotechnology, ultra-high dimensional genetic data are available. A typical character for such data is small n (sample size, often 10 - 100), large m (data dimension, often 10,000 or millions) and often with several different types of data with varied sample sizes and dependence relationships. How to analyze such data is a practical challenge, and new statistical methods are in demand for interpreting the wealth of data into biologically and clinically meaningful information. The integrative analysis is a recent research topic, which aims to deal with this problem.

Many existing methods analyze each variable separately. For example, individual datasets for genes, methylations or microRNAs were analyzed extensively for the disease diagnosis and prognosis study (Ramaswamy et al., 2001; Mikeska et al. 2012), little has been done on combining all

*Department of Biostatistics, Bioinformatics and Biomathematics, Georgetown University, Washington DC 20057
†NIH Clinical Center, Rehabilitation Medicine Department, Bethesda, MD 20892
‡Department of Biostatistics, Bioinformatics and Biomathematics, Georgetown University, Washington DC 20057

the information as an integrated dataset for the same study. DNA methylation has been thought as being an essential role of gene expression, usually negatively correlated with gene expression. However, recent studies on the relationship between DNA methylation, gene expression suggested a more complicated situation (Jones et al., 2013; Wagner et al., 2014). The methylation can be either active or passive related to gene expression, and there also exist situations that they are independently affected by small nucleotide polymorphism (SNP). Therefore, genomic data obtained from gene, transcript or methylation studies may have both common and unique information regarding the disease progress.

One may attempt to use traditional methods for these types of data, such as regression, but the large number of parameters is prohibitive for estimation. Existing methods are to use the LASSO (Tibshirani, 1996; 1997) to select only a few significant components, or to use the graphical model (For example, Edward, 2000; Anandkumar et al., 2012) on each variable separately. These methods will lead to information loss.

The integrative analysis is aimed at analyzing the mentioned types of data jointly, to gain more information and efficiency. Although there is no formal definition of integrative analysis, Li et al., (2009) used such method in the analysis of gene expression data. Zhang, Fang and Li (2015) used such method for gene expression and methylation analysis. Lin et al (2015) considered survival model for a similar problem. Wei (2015) gives a review of methods in this filed, and summarized the current methods as for several goals: dealing with batch effects for single data type, multiple data type, single data with survival data, multiple data with multiple study. In terms of methods used, these methods are summarized into linear/regression model, hierarchical model, Bayesian model, and survival model. In integrative analysis, we are to analyze the data jointly, instead of separately. We need to infer the relationships among the different types of data for this problem. Here we consider two methods for this problem under different assumptions of the data information, one attempts to capture the inner relationships among the data types, another aims at the basic structures of the data sequences. Simulation studies are conducted to evaluate the performance of the methods, and then one method is used to analyze a real data.

2. The proposed methods

We consider two methods for this problem based on different assumptions. The first method, the integrative correlation, assuming patterned structure of the data, attempts to capture the inner relationships among the data types. The second method, the smoothed empirical process method, assuming known ordering of the data sequence, aims at the basic structures of the data sequences. Below we describe them one by one.

2.1 the integrative correlation

With this method, we assume that the ordering of the data is unknown, so it cannot be treated as data processes. Also, the correspondence relationships among the measurements are unknown, and they have different lengths, so the relationships among them cannot be directly computed. With these assumption on the data, we introduce the integrative correlation.

We consider a typical data format: data collected from $n(= 150)$ normal individuals (control) and $n_1(= 100)$ tumor patients (case). For each individual, measurements of $m_1(= 20,000)$ genes (G, copy numbers), of $m_2(= 7000)$ proteins (P), of $m_3(= 9000)$ methylations (M), and of $m_4(= 10000)$ SNPs (S, 0,1,2 valued) are collected. The variables G, P, M are continuous. The goal is to predict patient/normal status based on an individual's measurements.

As an assumption, there are patterned relationships among the measurements in normal individuals, while there is no regular relationships among the measurements in patients. So, we only need data from normal individuals to infer such relationships. Let g_{ij} be the measurement of j-th gene of the i-th individual, p_{ij} be that for the j-th protein, m_{ij} be that of the j-th methylation, and s_{ij} be that for the j-th SNP; $\mathbf{g}_i = (g_{i1}, ..., g_{i,m_1})$, $\mathbf{p}_i = (p_{i1}, ..., p_{i,m_2})$, $\mathbf{m}_i = (m_{i1}, ..., m_{i,m_3})$, and $\mathbf{s}_i = (s_{i1}, ..., s_{i,m_3})$, $(i = 1, ..., n_1)$. For two data sequences $\mathbf{x} = (x_1, ..., x_m)$ and $\mathbf{y} = (y_1, ..., y_m)$, there are $m(m-1)/2$ correlation coefficients among their components. If we use canonical correlation between $\mathbf{x}$ and $\mathbf{y}$, we still need to compute all the $m(m-1)/2$ correlations. When m is large and the sample size n is relatively small (like our case $m \approx 10,000$, $n \approx 100$), estimating these coefficients does not make sense. For a normal individual, often the components of genetic material have some innate connection. For example, the measurements $\mathbf{g}_i$ of genes from individual i may have the form

$$g_{ij} = g(t_j) + \epsilon_{ij}, \qquad (i = 1, ..., n; j = 1, ..., m)$$

where $g(\cdot)$ is some determinant function reflecting the inner relationship among the genes, t_j is the position of the j-th gene, and ϵ_{ij} is random error. While for abnormal individuals, often such a systematic relationship does not function, or only partially functions, so the measurments among the sequence are more random, or the relationship obeys different rules.

To capture the inner relationship among the components of sequence measurements, we introduce a single new measure, the *integrative correlation* between two sequences.

Let $\bar{x} = m^{-1}\sum_{i=1}^{m} x_i$, $\bar{y} = m^{-1}\sum_{i=1}^{m} y_i$, $Con(\mathbf{x},\mathbf{y}) = m^{-1}\sum_{i=1}^{m}(x_i-\bar{x})(y_i-\bar{y}) = m^{-1}\sum_{i=1}^{n} x_i y_i - \bar{x}\bar{y}$, $D^2(\mathbf{x}) = m^{-1}\sum_{i=1}^{m}(x_i-\bar{x})^2 = m^{-1}\sum_{i=1}^{n} x_i^2 - \bar{x}^2$,$D^2(\mathbf{y}) = m^{-1}\sum_{i=1}^{m}(y_i-\bar{y})^2 = m^{-1}\sum_{i=1}^{n} y_i^2 - \bar{y}^2$, and define the *integrative correlation* between two sequences $\mathbf{x}$ and $\mathbf{y}$ as

$$Icor(\mathbf{x},\mathbf{y}) = E_{(\mathbf{x},\mathbf{y})}[Cor(\mathbf{x},\mathbf{y})], \quad \text{with} \quad Cor(\mathbf{x},\mathbf{y}) = \frac{Con(\mathbf{x},\mathbf{y})}{D(\mathbf{x})D(\mathbf{y})}.$$

Note that the quantities $D^2(\mathbf{x})$, $Con(\mathbf{x},\mathbf{y})$ and $Icor(\mathbf{x},\mathbf{y})$ look like the empirical variance of some random variable x, the empirical covariance and the empirical correlation between two random variables (x,y), respectively, but in fact they are not. For independent observations $(x_1,...,x_n) \sim x$ and $(y_1,...,y_n) \sim y$, if x and y are independent, then the empirical correlation based on $(x_1,...,x_n)$ and $(y_1,...,y_n)$ is ≈ 0. But for two independent sequence data, $\mathbf{x}$ and $\mathbf{y}$, $Icor(\mathbf{x},\mathbf{y})$ can take any value in the interval $[-1,1]$. It measures the concordance among the components of $\mathbf{x}$ and those of $\mathbf{y}$. If $\mathbf{x}$ and $\mathbf{y}$ are in perfect positive concordance, i.e., $y_i \approx x_i$ $(i = 1,...,m)$, then $Icor(\mathbf{x},\mathbf{y}) \approx 1$, even if $\mathbf{x}$ and $\mathbf{y}$ are independent; if $\mathbf{x}$ and $\mathbf{y}$ are in perfect negative concordance, i.e., $y_i \approx -x_i$ $(i = 1,...,m)$, then $Icor(\mathbf{x},\mathbf{y}) \approx -1$, even if $\mathbf{x}$ and $\mathbf{y}$ are independent; if the components of $\mathbf{x}$ and $\mathbf{y}$ have no concordance, then $Icor(\mathbf{x},\mathbf{y}) \approx 0$. Also, the correlation between two iid copies of a random variable itself is always 1, but $Icor(\mathbf{x}_1,\mathbf{x}_2)$ between two iid copies of $\mathbf{x}$ is generally less than 1. *For the gene measurement example, if $g(\cdot) \equiv$ some constant, only then the integrative correlation is the classical correlation.*

We define the integrative correlation of $\mathbf{x}$ itself as $Icor(\mathbf{x}_1,\mathbf{x}_2)$, for any two iid copies of $\mathbf{x}$, and denote it as $Icor(\mathbf{x})$. Intuitively, generally we have $Icor(\mathbf{x},\mathbf{y}) < Icor(\mathbf{x})$, and a pair of sequences from the normal population will be more concordant than a pair with one from a normal individual and one from an abnormal individual. Based on this, we can construct statistics using the integrative correlations to diagnose an individual to be normal or not.

As an example, $\mathbf{x} = (x_1,...,x_m)$, $x_i = sin(i) + \epsilon_i$, the $\epsilon_i's$ are iid $N(0,1)$. $\mathbf{x}_1 = (x_{11},...,x_{1m})$ and $\mathbf{x}_2 = (x_{21},...,x_{2m})$ are two iid copies of $\mathbf{x}$. Then by Cauchy-Schwarz inequality $-1 \leq$

$Cor(\mathbf{x}_1, \mathbf{x}_2) \leq 1$, with "=" if and only if $x_{1i} = cx_{2i}$ $(i = 1, .., m)$ for some non-zero and finite constant c. Since this condition is not satisfied, we have $-1 < Cor(\mathbf{x}_1, \mathbf{x}_2) < 1$ for any iid copies $\mathbf{x}_1$ and $\mathbf{x}_2$ of $\mathbf{x}$, and so $-1 < Icor(\mathbf{x}) < 1$.

Given n iid copies $\mathbf{x}_1, ..., \mathbf{x}_n \sim \mathbf{x}$, since the distribution of $\mathbf{x}$ is unknown, an estimate of $Icor(\mathbf{x})$ is the U-statistic (Hoeffding, 1948; 1961)

$$\widehat{Icor}_n(\mathbf{x}) = \frac{1}{n(n-1)} \sum_{i \neq j}^{n} Cor(\mathbf{x}_i, \mathbf{x}_j).$$

Now coming to our problem, we have four measures $Icor(\mathbf{g})$, $Icor(\mathbf{p})$, $Icor(\mathbf{m})$ and $Icor(\mathbf{s})$ for $\mathbf{g}, \mathbf{p}, \mathbf{m}$ and $\mathbf{s}$. Their estimates are

$$\widehat{Icor}_n(\mathbf{g}) = \frac{1}{n(n-1)} \sum_{i \neq j}^{n} Icor(\mathbf{g}_i, \mathbf{g}_j), \quad \widehat{Icor}_n(\mathbf{p}) = \frac{1}{n(n-1)} \sum_{i \neq j}^{n} Icor(\mathbf{p}_i, \mathbf{p}_j),$$

$$\widehat{Icor}_n(\mathbf{m}) = \frac{1}{n(n-1)} \sum_{i \neq j}^{n} Icor(\mathbf{m}_i, \mathbf{m}_j), \quad \widehat{Icor}_n(\mathbf{s}) = \frac{1}{n(n-1)} \sum_{i \neq j}^{n} Icor(\mathbf{s}_i, \mathbf{s}_j). \tag{1}$$

Let $\boldsymbol{\delta} = (Icor(\mathbf{g}), Icor(\mathbf{p}), Icor(\mathbf{m}), Icor(\mathbf{s}))'$, and

$$\hat{\boldsymbol{\delta}}_n = (\widehat{Icor}_n(\mathbf{g}), \widehat{Icor}_n(\mathbf{p}), \widehat{Icor}_n(\mathbf{m}), \widehat{Icor}_n(\mathbf{s}))' \tag{2}$$

Then from standard U-statistics theory, as $n \to \infty$, $\hat{\boldsymbol{\mu}}_n \overset{a.s.}{\to} \boldsymbol{\mu}$, and

$$\sqrt{n}\big(\hat{\boldsymbol{\delta}}_n - \boldsymbol{\delta}\big) \overset{D}{\to} N(\mathbf{0}, \Omega), \quad \text{or} \quad n\big(\hat{\boldsymbol{\delta}}_n - \boldsymbol{\delta}\big)\Omega^{-1}\big(\hat{\boldsymbol{\delta}}_n - \boldsymbol{\delta}\big)' \overset{D}{\to} \chi^2_4.$$

To compute Ω, we need the following notations. Let $\mathbf{y}_i = (\mathbf{g}_i, \mathbf{p}_i, \mathbf{m}_i, \mathbf{s}_i)$,

$$\mathbf{h}(\mathbf{y}_i) = \big(E[Cor(\mathbf{g}_i, \mathbf{g}_j)\big|\mathbf{g}_i], E[Cor(\mathbf{p}_i, \mathbf{p}_j)\big|\mathbf{p}_i], E[Cor(\mathbf{m}_i, \mathbf{m}_j)\big|\mathbf{m}_i], E[Cor(\mathbf{s}_i, \mathbf{s}_j)\big|\mathbf{s}_i]\big)' - \boldsymbol{\delta},$$

then (see, for example, Serfling, 1980)

$$\Omega = 4E[\mathbf{h}(\mathbf{y})\mathbf{h}'(\mathbf{y})].$$

Let

$$\mathbf{h}_n(\mathbf{y}_i) = \frac{1}{n} \sum_{j \neq i}^{n} \big(Cor(\mathbf{g}_i, \mathbf{g}_j), Cor(\mathbf{p}_i, \mathbf{p}_j), Cor(\mathbf{m}_i, \mathbf{m}_j), Cor(\mathbf{s}_i, \mathbf{s}_j)\big)' - \hat{\boldsymbol{\delta}}_n.$$

Then Ω is estimated by

$$\hat{\Omega}_n = \frac{4}{n} \sum_{i=1}^{n} \mathbf{h}_n(\mathbf{y}_i)\mathbf{h}'_n(\mathbf{y}_i). \tag{3}$$

Thus for given nominal level α, the $(1-\alpha)$-th joint confidence region $\mathbf{R}_n(\alpha)$ of $\boldsymbol{\delta}$ can be obtained as

$$\mathbf{R}_n(\alpha) = \{\ \boldsymbol{\mu}:\ \ n\big(\boldsymbol{\delta}-\hat{\boldsymbol{\delta}}_n\big)'\Omega^{-1}\big(\boldsymbol{\delta}-\hat{\boldsymbol{\delta}}_n\big) \le \chi^2_4(1-\alpha)\ \ \},$$

where $\chi^2_4(1-\alpha)$ is the $(1-\alpha)$-th upper quantile of χ^2_4.

For a new individual j with measurements $\mathbf{y}_j = (\mathbf{g}_j, \mathbf{p}_j, \mathbf{m}_j, \mathbf{s}_j)$, compute the statistic

$$\mathbf{t}(\mathbf{y}_j) = \frac{1}{n}\sum_{i=1}^{n}\big(Cor(\mathbf{g}_i,\mathbf{g}_j), Cor(\mathbf{p}_i,\mathbf{p}_j), Cor(\mathbf{m}_i,\mathbf{m}_j), Cor(\mathbf{s}_i,\mathbf{s}_j)\big)'. \tag{4}$$

If $\mathbf{t}(\mathbf{y}_j) \in \mathbf{R}_n(\alpha)$ or

$$n\big(\mathbf{t}(\mathbf{y}_j)-\hat{\boldsymbol{\delta}}_n\big)'\Omega^{-1}\big(\mathbf{t}(\mathbf{y}_j)-\hat{\boldsymbol{\delta}}_n\big) \le \chi^2_4(1-\alpha) \tag{5}$$

we diagnose individual j as normal, otherwise abnormal.

2.2 Smoothed Empirical process

The method of integrative correlation assumes that the normal and abnormal populations have different innate relation structures. In practice, many data may not satisfy this assumption. In this case we consider the smoothed empirical process method, it requires the ordering of the sequences to be known. Numerous other methods can be used for this problem. We are more interested in methods using all the data information, thus excluding variable selection methods such as LASSO and principal components analysis. A natural candidate is the empirical process method. The genes, SNPs, and proteins are well studied, their ordering and genetic distances among them, and the correspondence between each gene and each protein can all be find from internee. But methylation has no ordering and location information. Thus this method only applies to the $(\mathbf{g}_i(\cdot), \mathbf{p}_i(\cdot), \mathbf{s}_i(\cdot))$'s $(i = 1, ..., n)$, viewed as observed processes. Here the $(\mathbf{g}_i(\cdot), \mathbf{p}_i(\cdot), \mathbf{s}_i(\cdot))$'s are the re-arranged versions of the original data, in orderings according to their genetic orderings and correspondences as searched from the Internet.

Let $\mathbf{x}_i(\cdot) = (\mathbf{g}_i(\cdot), \mathbf{p}_i(\cdot), \mathbf{s}_i(\cdot))'$, $\bar{\mathbf{x}}(\cdot) = n^{-1}\sum_{i=1}^{n}\mathbf{x}_i(\cdot)$, and $\boldsymbol{\mu}(\cdot) = E[\mathbf{x}_i(\cdot)]$. Then under suitable conditions,

$$\sqrt{n}\big(\bar{\mathbf{x}}(\cdot)-\boldsymbol{\mu}(\cdot)\big) \overset{D}{\Rightarrow} \mathbb{W}(\cdot),$$

where $\mathbb{W}(\cdot)$ is a mean zero Gaussian process, with a covariance function $\sigma(s,t) = E[W(s)W(t)]$

which can be determined. In particular, at each fixed t, the variance at t is estimated by

$$\sigma^2(t) = \frac{1}{n-1}\sum_{i=1}^{n}\big(\mathbf{x}_i(t) - \bar{\mathbf{x}}(t)\big)\big(\mathbf{x}_i(t) - \bar{\mathbf{x}}(t)\big)'$$

and for given level α, we can construct level $(1-\alpha)$ confidence band for $\boldsymbol{\mu}(\cdot)$.

Let $\boldsymbol{\mu}_0(\cdot)$ and $\boldsymbol{\mu}_1(\cdot)$ be the empirical mean for the normal and abnormal populations, $\sigma_0^2(\cdot)$ and $\sigma_1^2(\cdot)$ be the corresponding variance functions, and n_0 and n_1 be the corresponding sample sizes. Then the level $1-\alpha$ confidence bands for normal and abnormal populations can be obtained as

$$[\boldsymbol{\mu}_0(\cdot) \pm n_0^{-1/2}1.96\sigma_0(\cdot)], \quad [\boldsymbol{\mu}_1(\cdot) \pm n_1^{-1/2}1.96\sigma_1(\cdot)]$$

For a new data $\mathbf{x}_j(\cdot)$, we classify the underlying individual as normal or abnormal by comparing its relationships to the two bands.

However, in practice often the data are very noise, directly using the empirical process method does not work well. In this case, we consider smoothing method to reduce the noises. For this, let $k(\cdot)$ be a density function (typically, we can take $k(\cdot) = \phi(\cdot)$, the density function of N(0,1)). We use the Nadaraya-Watson smoother (Nadaraya, 1964; Watson, 1964). For simple of exposition, we first deal with each measurement separately. Let $\mu_0(t)$and $\mu_1(t)$ as given before. For simple of exposition, we conly consider gene only. The smoothed version is

$$\bar{\mu}_0(t) = \frac{\sum_{j=1}^{m}\mu_0(t_j)k(\frac{t_j-t}{h})}{\sum_{j=1}^{m}k(\frac{t_j-t}{h})}, \qquad t \in T.$$

where $h > 0$ is the *bandwidth*. It is known that the method is not sensitive to the choices of the kernel, but is very sentsitive to the choice of the bandwidth. For convenience, we choose $k(\cdot)$ to be the density of N(0,1). There are various choices of the badwidth h, vary from simple to complicated. For simplicity we choose $h = O(n^{-5/12})$. The bandwidth can be used to control the amount for smoothness. The larger h is, the more smooth the curve.

The variance of $\bar{\mu}_0(t)$ is approximated as

$$\bar{\sigma}_0^2(t) \approx \sigma_0^2(t)\frac{\int t^2k(t)dt}{nh}.$$

Define $\bar{\mu}_1(t)$ similarly and obtain $\bar{\sigma}_1^2(t)$ similarly. The $1-\alpha$ confidence bands for normal and abnormal populations as, for α =5%,

$$[\bar{\boldsymbol{\mu}}_0(\cdot) \pm n_0^{-1/2}1.96\bar{\sigma}_0(\cdot)], \quad [\bar{\boldsymbol{\mu}}_1(\cdot) \pm n_1^{-1/2}1.96\bar{\sigma}_1(\cdot)].$$

In R, the function *ksmooth* in *stat* package is used to compute $\bar{\boldsymbol{\mu}}(\cdot)$.

For a coming subject with observation $\{g(t) : t \in T\}$. We first smooth it by the method above, to get $\bar{g}(\cdot)$. Then compute its projections onto the two profiles $\bar{\mu}_0(\cdot)$ and $\bar{\mu}_1(\cdot)$,

$$\langle \bar{g}, \bar{\mu}_0 \rangle = \frac{\int_T \bar{g}(t)\bar{\mu}_0(t)dt}{||\bar{g}||\,||\bar{\mu}_0||}, \qquad \langle \bar{g}, \bar{\mu}_1 \rangle = \frac{\int_T \bar{g}(t)\bar{\mu}_1(t)dt}{||\bar{g}||\,||\bar{\mu}_1||},$$

where, for a function g, $||g||^2 = \int_T g^2(t)dt$. Geometrically, $\langle \bar{g}, \bar{\mu}_0 \rangle$ is the cosine value of the angle between the $\bar{g}$ and $\bar{\mu}_0$, and large value of it means closeness of the two curves. Thus we classify this subject is normal if $\langle \bar{g}, \bar{\mu}_0 \rangle > \langle \bar{g}, \bar{\mu}_1 \rangle$, otherwise abnormal.

3. Simulation study and application

3.1 Simulation study for integrative correlation method

We simulate $n = 150$ normal individuals and $n_1 = 100$ abnormal individuals, each with measurements on log of gene copy numbers $\mathbf{g}$ on $m_1 = 20,000$ genes, on proteins $\mathbf{p}$ with $m_2 = 12,000$ measurements, on methylations $\mathbf{m}$ with $m_3 = 16,000$ components, and on SNPs $\mathbf{s}$ with $m_4 = 12,000$ locus sites. The measurements are correlated within the same individual, to reflect this, we use successive conditioning sampling. We distinguish normal and abnormal individuals.

Data for normal individuals are generated as below.

For each $i = 1, ..., n$, to sample $\mathbf{g}_i = (g_{i,1}, ..., g_{i,m_1})$,

$$g_{ij} \sim \sin(2j) + \frac{1}{100}\epsilon_{ij}, \quad \epsilon_{ij} \sim N(0, 10^2), \quad (j = 1, ..., m_1).$$

To sample $\mathbf{p}_i = (p_{i,1}, ..., p_{i,m_2})$,

$$p_{ij} \sim \sin(2j) + \frac{1}{100}\epsilon_{ij}, \quad \epsilon_{ij} \sim N(0, 12^2), \quad (j = 1, ..., m_2).$$

To sample $\mathbf{m}_i = (m_{i,1}, ..., m_{i,m_3})$,

$$m_{ij} \sim \sin(4j) + \frac{1}{100}\epsilon_{ij}, \quad \epsilon_{ij} \sim N(0, 11^2), \quad (j = 1, ..., m_3).$$

To sample $\mathbf{s}_i = (s_{i,1}, ..., s_{i,m_4})$, $s_{i,1} \sim multinomial(1, p_0)$, with $p_0 = (0.3, 0.35, 0.35)'$. We first sample a continuous $\mathbf{x}_i = (x_{i,1}, ..., x_{i,m_4})$'s, $x_{i,1} \sim |\sin(20)| + \epsilon_{i1} + 2$, with $\epsilon_{i1} \sim N(\mu_x, \sigma_x^2)$, $\mu_x = 0$, $\sigma_x = 0.55$. For $j = 2, ..., s_1$, sample $x_{i,j}|x_{i,j-1} \sim |\sin(20j)| + \epsilon_{ij} + 2$, with $\epsilon_{ij} \sim$

Table 1: Type I error of Simulation Data for Integration Correlation Method

measure	gene	protein	methylation	SNP	Type I error
simulation 1	$N(0, 10^2)$	$N(0, 12^2)$	$N(0, 11^2)$	$N(0, 0.55^2)$	
				$\rho = 0.95$	0.040
simulation 2	$N(0, 11^2)$	$N(0, 12^2)$	$N(0, 9^2)$	$N(0, 0.5^2)$	
				$\rho = 0.961$	0.04667
simulation 3	$N(0, 8^2)$	$N(0, 14^2)$	$N(0, 12^2)$	$N(0, 0.5^2)$	
				$\rho = 0.96$	0.0533

$N(\mu_x + \rho_x(x_{i,j-1} - \mu_x), (1 - \rho_x^2)\sigma_x^2)$, with $\rho_x = 0.95$.let x_i standardize normal to y_i. Let q_0 be the p_0-th quantile of y_i, q_1 be the $(p_0 + p_1)$-th quantile, and $q_2 = \infty$. Then define $s_{i,j} = k$, if $q_k < y_{i,j} \leq q_{k+1}$, $(k = 0, 1, 2)$.

Data for normal individuals are generated below.

For each $i = 1, ..., n_1$, to sample $\mathbf{g}_i = (g_{i,1}, ..., g_{i,m_1})$,

$$g_{ij} \sim \sin(2j) + \frac{1}{100}\epsilon_{ij}, \quad \epsilon_{ij} \sim N(0, 10.3^2), \quad (j = 1, ..., m_1).$$

To sample $\mathbf{p}_i = (p_{i,1}, ..., p_{i,m_2})$,

$$p_{ij} \sim \sin(2j) + \frac{1}{100}\epsilon_{ij} + 2, \quad \epsilon_{ij} \sim N(0, 12.2^2), \quad (j = 1, ..., m_2).$$

To sample $\mathbf{m}_i = (m_{i,1}, ..., m_{i,m_3})$,

$$m_{ij} \sim \sin(4j) + \frac{1}{100}\epsilon_{ij} + 1, \quad \epsilon_{ij} \sim N(0, 11.2^2), \quad (j = 1, ..., m_3).$$

for the abnormal individuals the parameters change(for example gene expression with $\mu = 0, \sigma = 0.58$), detailed parameters are displayed in Table1 and Table 2.

After the data generated, for the normal individuals, compute $\hat{\boldsymbol{\mu}}_n$ as in (1) and (2), and $\hat{\Omega}_n$ as in (3). Then for each incoming individual j with data $\mathbf{y}_j = (\mathbf{g}_j, \mathbf{p}_j, \mathbf{m}_j, \mathbf{s}_j)$, compute $\mathbf{t}(\mathbf{y}_j)$ as in (4), and check if (5) is satisfied or not, to classify this individual as normal or abnormal. Type I error and Power are calculated to evaluate the performance of simulations.

Table 2: Power of Simulation Data for Integration Correlation Method

measure	gene	protein	methylation	SNP	power
simulation 1	$N(0, 10.3^2)$	$N(0, 12.2^2)$	$N(0, 11.2^2)$	$N(0, 0.58^2)$	
				$\rho = 0.95$	
				$p_0 = (0.3, 0.35, 0.35)'$	0.85
simulation 2	$N(0, 11.2^2)$	$N(0, 12.2^2)$	$N(0, 9.3^2)$	$N(0, 0.5^2)$	
				$\rho = 0.961$	
				$p_0 = (0.3, 0.364, 0.336)'$	0.87
simulation 3	$N(0, 8.3^2)$	$N(0, 14.2^2)$	$N(0, 12.2^2)$	$N(0, 0.5^2)$	
				$\rho = 0.9615$	
				$p_0 = (0.3, 0.35, 0.35)'$	0.94

Table 3: Simulation Data for Smoothed Emprical Process Method

measure	Normal	Abnormal	prediction accuracy for normal	prediction accuracy for abnormal
simulation 1	$N(0, 0.8^2)$	$N(0, 2.5^2)$	0.807	0.89
simulation 2	$N(0, 0.5^2)$	$N(0, 1.5^2)$	0.80	0.90

3.2 Simulation study for smoothed emprical process method

We simulate gene expression information for $n = 150$ normal individuals and $n_1 = 100$ abnormal individuals. Data for normal individuals are generated below. For each $i = 1, ..., n$, to sample $\mathbf{g}_i = (g_{i,1}, ..., g_{i,m_1})$,

$$g_{ij} \sim \sin(j) + \epsilon_{ij}, \quad \epsilon_{ij} \sim N(0, 0.8^2), \quad (j = 1, ..., m_1).$$

Data for abnormal individuals are generated below. For each $i = 1, ..., n_1$, to sample $\mathbf{g}_i = (g_{i,1}, ..., g_{i,m_1})$,

$$g_{ij} \sim \sin(j) + \epsilon_{ij}, \quad \epsilon_{ij} \sim N(0, 2.5^2), \quad (j = 1, ..., m_1).$$

Detailed simulation parameters and the proportion of correct classification for both normal and abnormal patients are reported in Table 3.

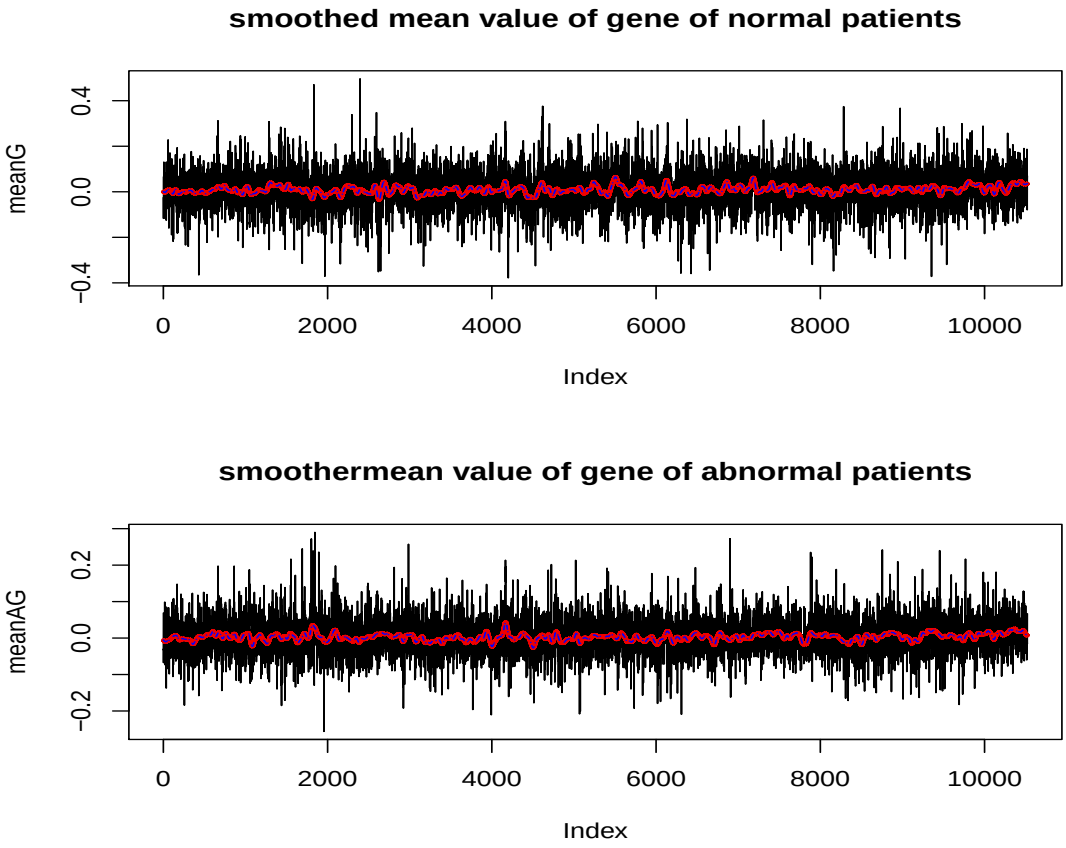
3.3 Application of the smoothed emprical process method to real data

In this section, we will use data from The Cancer Genome Atlas (TCGA) project at NIH for the study of the Ovarian Carcinoma disease from 2001-2010. There are 330 individuals in this data, with 92 disease free and 238 recurred/ progressed. For each patient, he or she contains information of 10521 gene and 172 protein.

For this data, the normal/abnormal status are known, so simulation study is not necessary and we analyze the data directly. The data is from two platforms: the gene expression data and the protein data. For this data, the ordering of the sequences are known, so both methods can be used. We computed integrative correlations for the normal and abnormal populations, and found no significance difference in the correlation structures. So we used the smoothed empirical process method to analyze the data. The results are plotted in Figures 1 and 2.

From the Figures, we see that smoothed empirical mean curves can characterize the stable structure of both gene and protein information. After smoothing these means, the projections of a new patient between normal and abnormal smoothed means are compared. For the gene expression data, 62.0% normal patients and 92.9% abnormal patients are correctly predicted. For the protein data, 64.1% normal patients and 67.6% abnormal patients are correctly predicted.

Figure 1: Gene expression for normal and abnormal populations



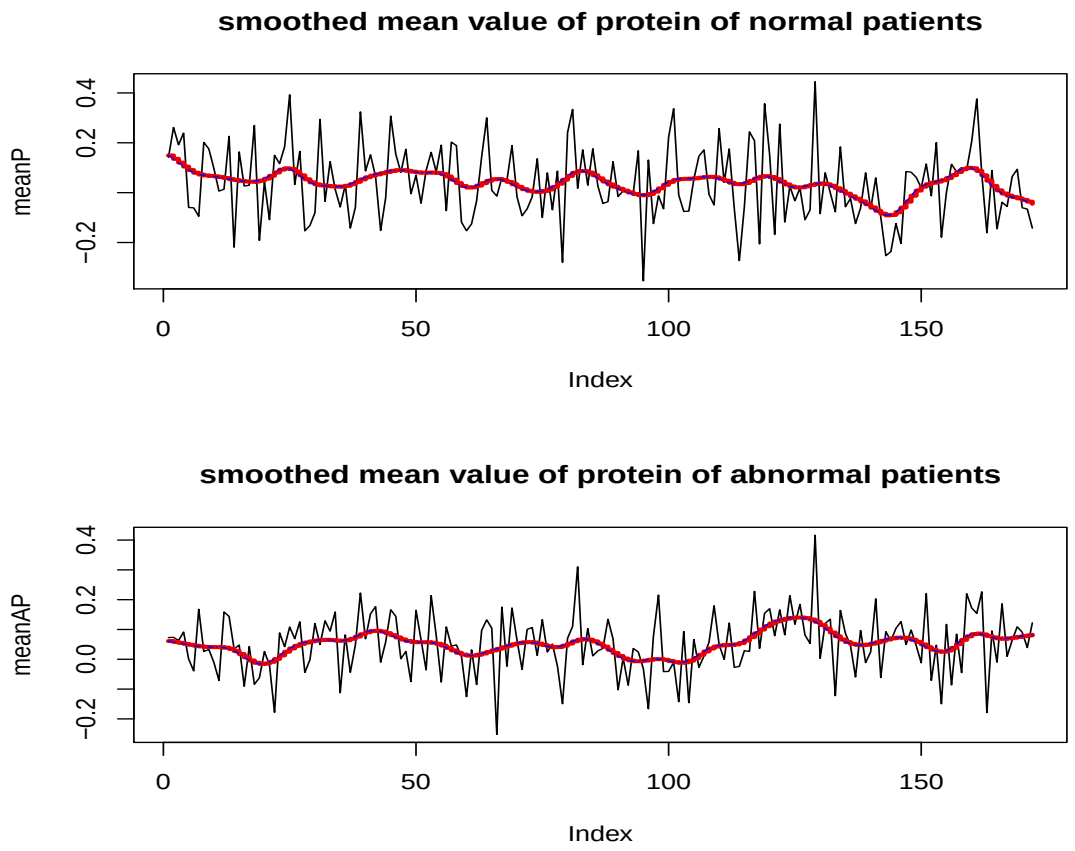


Figure 2: protein for normal and abnormal populations

REFERENCES

Anandkumar, A., Tan, V.Y.F., Huang, F., Willsky, A. (2012). High-dimensional Gaussian graphical model selection: walk summability and local separation criterion, *Journal of Machine Learning Research*, **13**, 2293-2337.

Akaike, H. (1973), "Information Theory and an Extension of the Maximum Likelihood Principle," in *Second International Symposium on Information Theory*, eds. B. N. Petrov and F. Càski, Budapest: Akademiai Kiàdo, pp. 267-281.

Cancer Genome Atlas Research Network. "Integrated genomic analyses of ovarian carcinoma." Nature 474.7353 (2011): 609-615..

Edwards, D. (2000). *An Introduction to Graphical Modelling*. Second Edition, Springer Verlag.

Friedman, J., Hastie, T., Tibshirani, R. (2008). Sparse inverse covariance estimation with the graphical lasso, *Biostatistics*, **9** (3), 432441.

Hoeffding, W. (1948). A class of statistics with asymptotically normal distribution. *Annals of Mathematical Statistics*,

19, 293-325.

Hoeffding, W. (1961). The strong law of large numbers for U-statistics. *Inst. Statist. Mimeo. Ser.*, No. 302, 1-10.

Jones, M. J., Fejes, A. P., Kobor, M. S. (2013). DNA methylation, genotype and gene expression: who is driving and who is along for the ride? *Genome Biol.*, **14**(7): 126.

Li, M., Balch, C., Montgomery, J.S., Jeong, M., Chung, J.H., Yan, P., Huang, T., Kim, S., Nephew, K.P. (2009). Integrated analysis of DNA methylation and gene expression reveals specific signaling pathways associated with platinum resistance in ovarian cancer, *BMC Medical Genomics*, **2**, 34. doi:10.1186/1755-8794-2-34.

Lin, D., Auman, T., Innocenti, F., Kosorock, M., Liu, Yufeng, Rathmell, W.K., Sun, W., Yeh, J.J., Zeng, D. (2015). Manuscript.

Mikeska T., Bock C., Do H. , and Dobrovic A., (2012). DNA methylation biomarkers in cancer: progress towards clinical implementation. *Expert Rev. Mol. Diagn.*, **12** (5), 473-487.

Nadaraya, E.A. (1964). On estimating regression,*Theory of probability and its applications*, **9**(1), 141-142.

Ramaswamy, S., Tamayo, P., Rifkin, R., Mukherjee, S., Yeang, et al. (2001). Multiclass cancer diagnosis using tumor gene expression signatures. *Proc. Natl. Acad. Sci. U S A*, **98**, 15149 -15154.

Serfling, R. (1980). *Approximation Theorems of Mathematical Statistics*, Jhon Wiley & Sons.

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

Tibshirani, R. (1997). The lasso method for variable selection in the Cox model. *Statistics in Medicine*, **16**, 385-395.

Wagner, J. R., Busche, S., Ge, B., Kwan, T., Pastinen, T., Blanchette, M. (2014). The relationship between DNA methylation, genetic and expression inter-individual variation in untransformed human fibroblasts. *Genome Biol.*, **15**(2): R37.

Watson, G.S. (1964). Smooth regression analysis, *Sankhya: The Indian Journal of Statistics* (Series A), **26**(4), 359-372.

Wei, Y. (2015). Integrative analyses of cancer data: a review from a statistical perspective, *Cancer Informatics*, **14** (S2), 173-181.

Yuan, M., Lin, Y. (2007). Model selection and estimation in the Gaussian graphical model, *Biometrika*, **94**(1), 19-35.

Zhang, X., Fang, H., Li, G. (2015). Integrated Analyses on Gene Expression and Methylation Microarray Data and Application to Biomarker Development of Ovarian Cancer, Manuscript.