Statistics and Its Interface Volume 9 (2016) 445–452

Direct regression modelling of high-order moments in big data

Ruibin Xi and Nan Lin*

Big data problems present great challenges to statistical analyses, especially from the computational side. In this paper, we consider regression estimation of high-order moments in big data problems based on the U-statistic-based Functional Regression Model (U-FRM) model. The U-FRM model is a nonparametric method that allows direct estimation of higher-order moments without imposing parametric assumptions on the high order-moments. Despite this modeling advantage, its estimation relies on a U-statistics-based estimating equation whose computational complexity is generally too high for big data. In this paper, we propose using the "divide-and-conquer" strategy to construct a computationally more succinct surrogate estimating equation. Through both theoretical proof and simulations, we show that our method significantly reduces the computational time and meanwhile enjoys the same asymptotic behavior as the original estimation method. We then apply our method to a genomic problem to illustrate its performance on real data.

Keywords and phrases: Big data, Higher-order moment, U-statistics, Estimating equation, Divide-and-conquer, Aggregation, Consistency, Asymptotic normality, Data cube.

1. INTRODUCTION

The recent technology breakthroughs have made data collection very efficient and cost-effective in many different fields such as biology, astronomy and business. For example, in biology, the latest sequencing platform HiSeq X Ten can generate up to 1.8 Tb of sequencing data in a 3-day run. As data collection becomes easier, data analysis or computation is becoming the bottleneck for many researches and real applications. The very first question in big data analysis is how to computationally efficiently perform available statistical analyses. In big data analyses, an algorithm's ideal computational complexity is $O(n)$, where n stands for the size of the data. When an algorithm's computational complexity is more than $O(n^2)$, it becomes computationally very difficult or even infeasible for many big data analyses.

There are at least two basic strategies to address the computational problems in big data analysis. The first strategy is sub-sampling. The naive sub-sampling method is to uniformly sample from the entire data set and perform statistical analyses on the sub-sampled data. More advanced methods take advantage of the properties of the statistical analysis under consideration (e.g. linear regression) and develop sampling methods that can give estimates closer to the estimates based on the entire data sets than simple uniform sampling [3, 4, 17]. The other strategy is "divide-and-conquer", i.e. dividing the entire data set to many small subsets, compressing each subset to a small number of summary statistics and then performing the statistical analysis based on aggregating the summary statistics. This strategy was long ago employed by computer scientists for calculating simple statistics such as mean and sum [5]. Later, it was further developed for linear regression [2, 6], general multiple linear regression [1, 16], logistic regression analysis [24], predictive filters [1], generalized estimating equations [14] and Bayesian analysis [25]. Methods based on the divide-and-conquer strategy can be easily applied to parallel computing, distributed computing and fast query in the data cube setting [7]. For linear statistical analyses such as linear regression analysis, methods based on the divide-and-conquer strategy can exactly recover the estimates as the ones based on the entire data set.

In this paper, we consider in big data the estimation problem in a special semiparametric transformation regression model [10] that is particularly useful for modeling higher-order moments. This model was named as functional regression model (FRM) in its original reference. But to avoid confusion with the popular regression models for functional data, we rename it here the U-FRM as the transformation was formulated through U-statistics.

Traditional regression models such as the popular Generalized Estimating Equation (GEE) methods are mostly interested in the dependence of the mean of the response variable on explanatory variables. However, to understand the dependence structure in complex data, higher-order moments are often of direct interest. For example, Wang et al. [21] considered the epistatis effects of a pair of genetic markers, such as single nucleotide polymorphism (SNP), on the correlation pattern of a pair of genes. In this research, the authors focused on the dependence of correlation of gene pairs on genetic markers instead of the mean expression level of genes. Another example is the popular Gaussian graphical model [28, 26, 20], which also emphasizes on understanding

*Corresponding author.

the correlation structure instead of the mean of the response variables.

The U-FRM can be viewed as a generalization of the GEE, where the single subject-based response y_i is replaced with a function $f(y_{i_1}, \cdots, y_{i_k})$ of several responses $y_{i_1}, \cdots, y_{i_k}$ from multiple subjects $i_1, \cdots, i_k$. Examples of the U-FRM include parameter estimation in zero inflated regression models [27], in correlation analysis for assessing rater reliability and precision of diagnosis [19] and in spatial data analysis [12]. Similar to GEE, the U-FRM only needs to specify a model for the mean of $f(y_{i_1}, \cdots, y_{i_k})$ and this semi-parametric nature makes U-FRM more robust than parametric likelihood methods. For example, consider the random effect model $y_{ikt} = t^2 + r_{kt} + \epsilon_{ikt}$, where $k = 0, 1$, (r_{1t}, r_{2t}) follows a two dimensional distribution and ϵ_{ikt} are independent identically distributed normal random errors. This model can be viewed as a model for assessing the agreement of two measurements $k = 0, 1$ at different time points t and we are mostly interested in the correlation ρ_t between y_{i1t} and y_{i2t}. To estimate this correlation, we could use linear mixed effect (LME) model techniques. However, LME models usually assume that the distribution of (r_{1t}, r_{2t}) is Gaussian. If this assumption is not correct, statistical inference based on LME model would be inaccurate. In addition, if the LME model is mis-specified (e.g. t^2 is is mis-specified as t or the random effect r_{kt} is mis-specified as only depending on t), the LME model could give mis-leading estimates about the correlation ρ_t. The U-FRM instead provides an alternative way of estimating ρ_t that avoids the parametric specification of the dependence of y_{ikt} on t, k and avoids the normal assumption of the random effect and thus would be more robust than LME models. This U-FRM is given in Simulation 1. In general, parameter estimation of the U-FRM employs a U-statistic based generalized estimating equation (UGEE). While the U-FRM provides an appealing nonparametric framework for solving high-order moments regression problems, its computational complexity is likely to hinder its application to big data for that computing a U-statistic of degree $m \geq 2$ generally has complexity $O(n^m)$ (see definition in Section 2.1), and solving the estimating equation in U-FRM often involves many iterations each of which needs to compute the U-statistic.

In this paper, we propose a new efficient computational strategy for estimating the U-FRM that provides statistically equivalent estimates. Our method was motivated by the computationally more efficient surrogate of U-statistics for i.i.d. data, namely the aggregated U-statistics (AU-statistics) [13]. The AU-statistic significantly reduces the computational burden by utilizing the "divide-and-conquer" strategy and meanwhile maintains first-order asymptotic equivalence to the raw U-statistic. In our method, we replace the UGEE for a U-FRM by an AU-statistics-based generalized estimating equation (AUGEE). And we show that the estimator from the AUGEE is asymptotically equivalent to that from the UGEE. Simulation studies also show that the estimator obtained from the AUGEE is nearly as efficient as the estimator obtained from the UGEE but computationally much more efficient.

2. THE U-FRM

2.1 U-statistics and AU-statistics

Let $X_1, \cdots, X_N$ be N i.i.d. random variables from an unknown distribution P in a nonparametric family $\mathcal{P}$. Suppose that $h(x_1, \cdots, x_m)$ is a measurable function defined on $\mathbb{R}^m$ that is symmetric in its arguments and satisfies $\vartheta = E[h(X_1, \cdots, X_m)] < \infty$. Then an unbiased estimator of ϑ is given by

$$U_N = \binom{N}{m}^{-1} \sum_{1 \leq i_1 < \cdots < i_m \leq N} h(X_{i_1}, \cdots, X_{i_m}), \tag{1}$$

where the summation is over the set of all $\binom{N}{m}$ combinations of m integers, $i_1 < i_2 < \ldots < i_m$ chosen from $\{1, 2, \ldots, N\}$. Here, U_N is called a U-statistic with kernel h and degree m. Many commonly used nonparametric statistics can be viewed as U-statistics, such as the Mann-Whitney-Wilcoxon test statistic [18, 22] and Kendall's τ rank correlation [9]. The time complexity of computing the U-statistics in (1) is generally $O(N^m)$, which increases very rapidly as the sample size increases for $m \geq 2$.

To reduce the computational burden of U-statistics, Lin and Xi (2010) [13] introduced AU-statistics defined as follows. First, partition a random sample $\{X_1, \cdots, X_N\}$ into K subsets with observations in the kth subset denoted by $\{X_{k1}, \cdots, X_{kn_k}\}$ and the U-statistic based on them as U_{kn_k}. It is obvious that $\sum_{k=1}^{K} n_k = N$. Then, the AU-statistic is given by the following weighted average,

$$\tilde{U}_N = \frac{1}{N} \sum_{k=1}^{K} n_k U_{kn_k}. \tag{2}$$

Since the AU-statistics only depends on m-tuples within each subset not across, its computational complexity is much lower then the original U-statistics. Lin and Xi (2010) [13] also showed that under some mild regularity conditions (allowing K tending to ∞), the asymptotic distribution of AU-statistics is the same as that of the U-statistics and thus they are statistically equivalent.

2.2 The UGEE

Consider a regression setup based on independent observations $Z_1 = (Y_1, X_1), \cdots, Z_N = (Y_N, X_N)$. Let $\mathbf{f}$ and $\mathbf{g}$ be two known measurable q-dimensional vector-valued functions satisfying the following equation,

$$\begin{aligned} & E[\mathbf{f}(Y_{i_1}, \cdots, Y_{i_m}) | X_{i_1}, \cdots, X_{i_m}] \\ = \; & \mathbf{g}(X_{i_1}, \cdots, X_{i_m}; \boldsymbol{\theta}_0), \end{aligned} \tag{3}$$

where $\boldsymbol{\theta}_0$ is a p-dimensional unknown parameter. We then call (3) the U-FRM [10]. Without loss of generality, we may assume the functions $\mathbf{f}$ and $\mathbf{g}$ are symmetric about their arguments. Otherwise, they can be easily symmetrized.

Suppose that $H(x_1, \cdots, x_m)$ is a measurable $p \times q$ dimensional matrix-valued function and is symmetric about its arguments and

$$\begin{aligned}(4)\qquad &\mathbf{h}(z_1, \cdots, z_m; \boldsymbol{\theta}) \\ &= H(x_1, \cdots, x_m)[\mathbf{f}(y_1, \cdots, y_m) - \mathbf{g}(x_1, \cdots, x_m; \boldsymbol{\theta})],\end{aligned}$$

where $z_i = (y_i, x_i)$. The following UGEE is used to estimate $\boldsymbol{\theta}_0$ in the U-FRM,

$$\begin{aligned}(5)\qquad \mathbf{U}_N(\boldsymbol{\theta}) &= \binom{N}{m}^{-1} \sum_{1 \le i_1 < \cdots < i_m \le N} \mathbf{h}(Z_{i_1}, \cdots, Z_{i_m}; \boldsymbol{\theta}) \\ &= 0.\end{aligned}$$

However, solving equation (5) is computationally expensive for $m \ge 2$ and large N. Interestingly, Liang et al. (2013) [12] recently proposed to use a resampling method to estimate parameters in the UGEE (5). In the next section, we will use the AU-statistics [13] to reduce the computational complexity for solving (5).

2.3 The AUGEE

Now, we can introduce the alternative AUGEE for the U-FRM and show that the estimator obtained from the AUGEE is asymptotically equivalent to the estimator from the original UGEE.

As in the case of AU-statistics, we first partition the data set $\{Z_1, \cdots, Z_N\}$ into K subsets $\{Z_{k1}, \cdots, Z_{kn_k}\}$, $k = 1, \ldots, K$. Let $\mathbf{U}_k(\boldsymbol{\theta})$ be the U-statistic based function in (5) based on the kth subset. Then, we can solve the following alternative AUGEE to get an estimate of $\boldsymbol{\theta}_0$,

$$(6)\qquad \tilde{\mathbf{U}}_N(\boldsymbol{\theta}) = \frac{1}{N} \sum_{k=1}^{K} n_k \mathbf{U}_{kn_k}(\boldsymbol{\theta}) = \mathbf{0}.$$

Let $\tilde{\boldsymbol{\theta}}_{K,N}$ be the solution to the estimating equation (6). Note that $\tilde{\boldsymbol{\theta}}_{1,N}$ is just the solution to the UGEE (5). Because the estimating equation (6) uses m-tuples less than the estimating equation (5), the computational complexity of solving (6) would be much lower. If we use the Newton-Raphson algorithm and choose n_k to be the same for all k, the computational complexity of solving the AUGEE (6) would be at the order of $O(N^m/K^{m-1})$ in each iteration, but the computational complexity of solving the UGEE (5) is at the order of $O(N^m)$ in each iteration. Therefore, the AUGEE tremendously reduces the computational burden of estimating FRMs when $m \ge 2$.

Let $z_1, \cdots, z_m$ be fixed numbers. Define

$$\boldsymbol{\vartheta} = E[\mathbf{h}(Z_1, \cdots, Z_m)],$$

$$\begin{aligned}\mathbf{h}_k(z_1, \ldots, z_k) &= E[\mathbf{h}(z_1, \ldots, z_k, Z_{k+1}, \ldots, Z_m)], \\ \boldsymbol{\zeta}_k(\mathbf{h}) &= \mathrm{Var}(\mathbf{h}_k(Z_1, \cdots, Z_k)).\end{aligned}$$

Before presenting the asymptotic property of the estimator $\tilde{\boldsymbol{\theta}}_{K,N}$, we give the following conditions.

(C1) $E[\mathbf{h}(Z_1, \cdots, Z_m; \boldsymbol{\theta}_0)^T \mathbf{h}(Z_1, \cdots, Z_m; \boldsymbol{\theta}_0)] < \infty$ and $\boldsymbol{\zeta}_1(\mathbf{h}_{\boldsymbol{\theta}_0})$ is positive definite, where $\mathbf{h}_{\boldsymbol{\theta}_0}(z_1, \cdots, z_m) = \mathbf{h}(z_1, \cdots, z_m; \boldsymbol{\theta}_0)$.

(C2) $\mathbf{h}(z_1, \cdots, z_m; \boldsymbol{\theta})$ is twice differentiable in a neighborhood of $\boldsymbol{\theta}_0$ and

$$B = E\left[\frac{\partial \mathbf{h}}{\partial \boldsymbol{\theta}}(Z_1, \cdots, Z_m; \boldsymbol{\theta}_0)\right]$$

is an invertible matrix.

(C3) Suppose that $h^s(z_1, \cdots, z_m; \boldsymbol{\theta})$ $(s = 1, \cdots, p)$ is the sth entry of the vector function $\mathbf{h}(z_1, \cdots, z_m; \boldsymbol{\theta})$ and $b(z_1, \cdots, z_m)$ is a measurable function which is symmetric about its argument and $E\big[b(Z_1, \cdots, Z_m)\big]^2 < \infty$. We have for all $s, i, j = 1, \cdots, p$

$$E\left[\frac{\partial h^s}{\partial \theta_j}(Z_1, \cdots, Z_m; \boldsymbol{\theta}_0)\right]^2 < \infty$$

and

$$\left|\frac{\partial^2 h^s}{\partial \theta_i \partial \theta_j}(z_1, \cdots, z_m, \boldsymbol{\theta})\right| \le b(z_1, \cdots, z_m)$$

in a neighborhood of $\boldsymbol{\theta}_0$.

Theorem 1. *Suppose that m is fixed and that $\tilde{\boldsymbol{\theta}}_{K,N}$ is the solution to the AUGEE (6). If Conditions (C1), (C2) and (C3) are satisfied and $K = o(N)$, the estimator $\tilde{\boldsymbol{\theta}}_{K,N}$ is a consistent estimator of $\boldsymbol{\theta}_0$ and*

$$(7)\qquad \sqrt{N}(\tilde{\boldsymbol{\theta}}_{K,N} - \boldsymbol{\theta}_0) \xrightarrow{d} \mathcal{N}(\mathbf{0}, m^2 G \boldsymbol{\zeta}_1(\mathbf{h}_{\boldsymbol{\theta}_0}) G^T),$$

where $G = B^{-1}$.

Note that Theorem 1 applies to the case of $K = 1$, so it also establishes the asymptotic normality of the estimator from the original UGEE, which was missing in the original reference of the U-FRM [10]. A quick corollary of Theorem 1 is that the estimators $\tilde{\boldsymbol{\theta}}_{K,N}$ and $\tilde{\boldsymbol{\theta}}_{1,N}$ are asymptotically equivalent when $K = o(N)$. Therefore, the aggregation method reduces the computational complexity while maintaining the asymptotic efficiency of the estimator $\tilde{\boldsymbol{\theta}}_{1,N}$. The proof of Theorem 1 is given in Appendix.

When $m \ge 2$, the AUGEE is computationally much more efficient than the UGEE. The most time consuming step in obtaining the UGEE is to calculate the U-statistic of (5) in each iteration of a root-finding algorithm. In each iteration, the computational complexity is $O(N^m)$ for calculating the U-statistic. If we take $K = \sqrt{N}$, the computational complexity of calculating the AU-statistic in (6) is only $O(K(N/K)^m)$. For instance, if $m = 2$ and $K = \sqrt{N}$,

the time complexity for the AUGEE is $O(N^{3/2})$ while that for UGEE is $O(N^2)$. When N is moderately large (e.g. $N = 10{,}000$), this can be several magnitudes of computational time saving. In big data applications, the number of observations N is usually in tens of thousands or even in millions. We recommend to use a large K (e.g. $K = \sqrt{N}$) since in such cases the asymptotic results will be very accurate and the UGEE and AUGEE will be very close to each other. If the data size is not large (a few hundreds or less), we recommend to use a small K since a too large K would give less accurate estimates and the computational burden of calculating the UGEE in such applications is not very demanding.

3. SIMULATION STUDIES

In this section, we will show by simulation that the estimator obtained from the AUGEE is statistically equivalent to the estimator obtained from the UGEE, while the former is computationally more efficient.

SIMULATION 1. Suppose that y_{1it}, y_{2it} are two measurements on subject i at time t ($i = 1, \cdots, n; t = 1, \cdots, T$). Assume that subjects are independent. Let σ^2_{kt} be the variance of y_{kit} ($k = 1, 2$) and ρ_t be the correlation between the two measurements y_{1it} and y_{2it} at time t. By the independence assumption, it follows that

$$\begin{aligned} E[(y_{1it} - y_{1jt})^2/2] &= \sigma^2_{1t} \\ E[(y_{2it} - y_{2jt})^2/2] &= \sigma^2_{2t} \\ E[(y_{1it} - y_{1jt})(y_{2it} - y_{2jt})/2] &= \rho_t \sqrt{\sigma^2_{1t}}\sqrt{\sigma^2_{2t}} \end{aligned}$$

Let $\mathbf{y}_{it} = (y_{1it}, y_{2it})$ and $\mathbf{y}_i = (\mathbf{y}_{i1}, \cdots, \mathbf{y}_{iT})$. Denote $f_{kt}(\mathbf{y}_i, \mathbf{y}_j) = (y_{kit} - y_{kjt})^2/2$, $h_{kt} = \sigma^2_{kt}$ ($k = 1, 2$), $f_{3t}(\mathbf{y}_i, \mathbf{y}_j) = (y_{1it} - y_{1jt})(y_{2it} - y_{2jt})/2$ and $h_{3t} = \rho_t \sqrt{\sigma^2_{1t}}\sqrt{\sigma^2_{2t}}$. Then the U-FRM model is

$$E[f_{kt}(\mathbf{y}_i, \mathbf{y}_j)] = h_{kt} \quad k = 1, 2, 3 \;\; t = 1, \cdots, T.$$

Let $\mathbf{f}_t = (f_{1t}, f_{2t}, f_{3t})$, $\mathbf{h}_t = (h_{1t}, h_{2t}, h_{3t})$, $\mathbf{f} = (\mathbf{f}_1, \cdots, \mathbf{f}_T)$ and $\mathbf{h} = (\mathbf{h}_1, \cdots, \mathbf{h}_T)$. Then the U-FRM becomes $E[\mathbf{f}(\mathbf{y}_i, \mathbf{y}_j)] = \mathbf{h}$. Given observations $\mathbf{y}_1, \cdots, \mathbf{y}_N$, the following UGEE is used to estimate the parameters ρ_t, σ^2_{1t} and σ^2_{2t},

$$\mathbf{U}_N(\boldsymbol{\theta}) = \binom{N}{2}^{-1} \sum_{1 \le i < j \le N} \big(\mathbf{f}(\mathbf{y}_i, \mathbf{y}_j) - \mathbf{h}\big).$$

The U-FRM under consideration is based on a posttraumatic stress disorder (PTSD) study with a total of 95 women victims of sexual and non-sexual assault at the University of Pennsylvania Medical Center [10]. The two measurements are PTSD Symptom Scale and Beck Depression Inventory at 5 time points. The goal is to longitudinally examine the correlations between the two measurements.

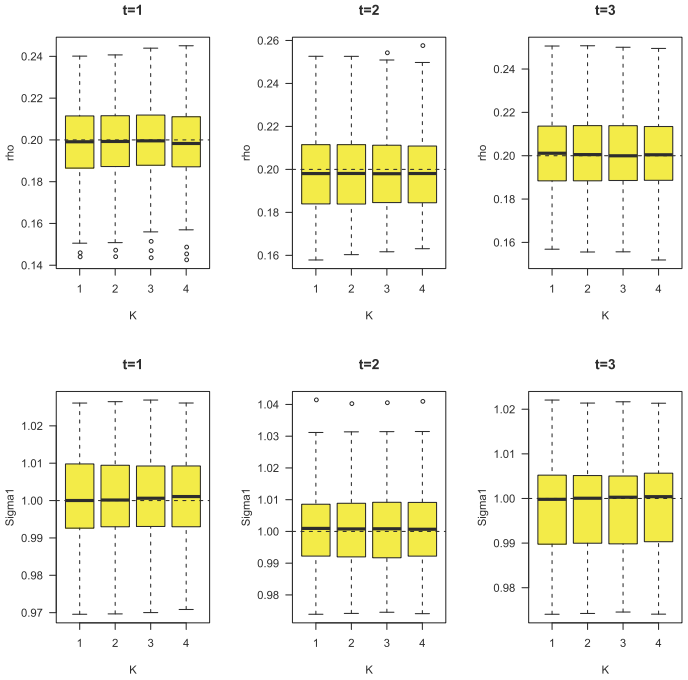

Figure 1. Box plots of correlation estimates from different estimating equations in Simulation 1. 1: UGEE; 2: AUGEE ($K = 5$); 3: AUGEE ($K = 10$) and 4: AUGEE ($K = 20$). Top row is for the correlations ρ_t and the bottom row is for the standard deviations σ_{1t}.

In this simulation, the number of time points is set as 3, i.e. $T = 3$. We generate 100 data sets each of which has 2,000 observations. In every data set, (y_{1it}, y_{2it}) are drawn from a mean 0 bivariate normal distribution. The parameters are set as $\sigma^2_{1t} = \sigma^2_{2t} = 1$ and $\rho_t = 0.2$ for all $t = 1, 2, 3$. We then compare their estimates from the UGEE and from the AUGEE with partition number $K = 5$, 10 and 20 using a program written in R. Since the estimates for σ_{2t} are very similar to σ_{1t}, here we only report the results for ρ_t and σ_{1t}. Figure 1 shows the box plots of the 100 estimates of the correlation ρ_t from the UGEE and three AUGEEs. And box plots for other parameters showed similar comparisons. Table 1 compares the sample means and sample variances of the 100 estimates and average computation time using the four different estimating equations. At all t, the four sample means are comparable and the sample variances remain at the same level for different K. However, the AUGEE saves a considerable amount of computation time compared with the UGEE. In all, the simulation clearly shows that AUGEEs provide estimators nearly as good as estimators obtained from UGEEs, while the computational burden of solving AUGEEs is much lower.

SIMULATION 2. In this simulation, we consider the estimation of the over-dispersion parameter with the UGEE. Let y_i and $\mathbf{x}_i$ denote some count response and vector of inde-

Table 1. Comparison of estimates from different estimating equations in Simulation 1

	K	Mean	Variance($\times 10^{-4}$)	Time (seconds)
	1	0.19813	3.71	71667.9
ρ_1 $(t=1)$	5	0.19813	3.69	13050.8
	10	0.19823	3.74	6441.2
	20	0.19820	3.87	3223.2
	1	0.19734	4.48	71667.9
ρ_2 $(t=2)$	5	0.19740	4.47	13050.8
	10	0.19729	4.47	6441.2
	20	0.19730	4.47	3223.2
	1	0.20021	3.30	71667.9
ρ_3 $(t=3)$	5	0.20017	3.32	13050.8
	10	0.20010	3.33	6441.2
	20	0.19997	3.39	3223.2
	1	1.00054	1.45	71667.9
σ_1 $(t=1)$	5	1.00063	1.45	13050.8
	10	1.00068	1.44	6441.2
	20	1.00073	1.46	3223.2
	1	1.00034	1.50	71667.9
σ_1 $(t=2)$	5	1.00030	1.50	13050.8
	10	1.00031	1.51	6441.2
	20	1.00042	1.47	3223.2
	1	0.999137	1.29	71667.9
σ_1 $(t=3)$	5	0.999203	1.28	13050.8
	10	0.999214	1.29	6441.2
	20	0.999304	1.29	3223.2

pendent variables from the ith subject. The classic quasi-Poisson log-linear model is given by

$$E(y_i|\mathbf{x}_i) = \mu_i = \exp(\mathbf{x}^T\boldsymbol{\beta}).$$

If y_i comes from a Poisson distribution, its mean and variance are equal. In case of over-dispersion, we have

$$E((y_i-\mu_i)^2) = \lambda\mu_i \quad (\lambda > 0).$$

The over-dispersion parameter may be estimated by the GEE method, but here we will use UGEE and AUGEE. Following Kowalski and Tu (2008) [10], define

$$\begin{aligned} f_1(y_i, y_j) &= y_i + y_j \\ f_2(y_i, y_j) &= (y_i - y_j)^2 \\ h_1(\mathbf{x}_i, \mathbf{x}_j; \boldsymbol{\beta}, \lambda) &= \mu_i + \mu_j \\ h_2(\mathbf{x}_i, \mathbf{x}_j; \boldsymbol{\beta}, \lambda) &= \lambda(\mu_i + \mu_j) + (\mu_i - \mu_j)^2 \\ \mathbf{f} &= (f_1, f_2) \\ \mathbf{h} &= (h_1, h_2). \end{aligned} \tag{8}$$

Then, we have $E(\mathbf{f}(y_i, y_j)|\mathbf{x}_i, \mathbf{x}_j) = \mathbf{h}(\mathbf{x}_i, \mathbf{x}_j; \boldsymbol{\beta}, \lambda)$ and hence we can construct the estimating equation as (6) to estimate the parameters $\boldsymbol{\beta}$ and λ.

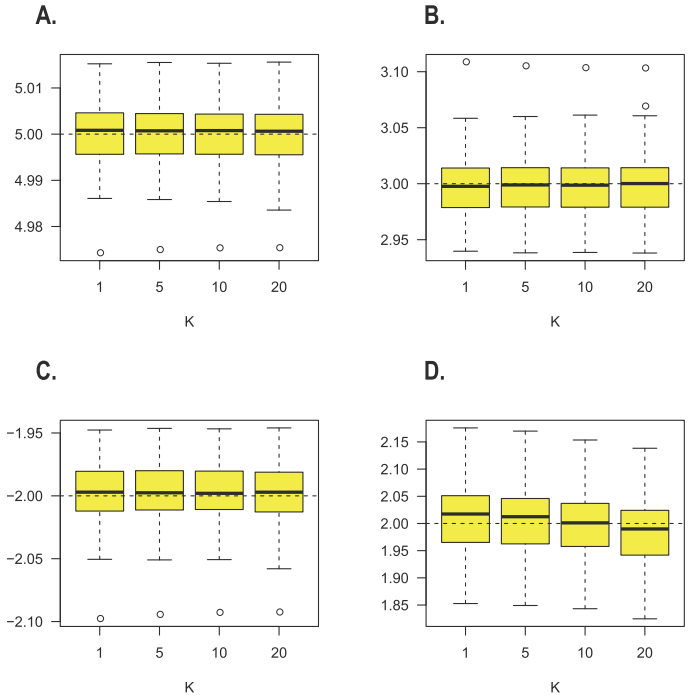

Figure 2. Box plots of parameter estimates (A. β_0, B. β_1, C. β_2, D. λ) for Simulation 2 from different types of estimating equations ($K = 1, 5, 10, 20$).

In this simulation, we set $\boldsymbol{\beta} = c(5, 3, -2)$, $\lambda = 2$. The covariates $\mathbf{x}_i = (1, x_{1i}, x_{1i}^2)$, where $x_{1i} = \exp(z_i)/(1+\exp(z_i))$ with z_i sampled from $N(0, 0.5^2)$. The responses are generated from negative binomial distribution with mean $\mu_i = \mathbf{x}_i^t\boldsymbol{\beta}$ and variance $\lambda\mu_i$. This model is motivated by the dependence of short read coverage on the GC content (i.e. proportion of G and C in a genomic region) in the high-throughput sequencing data (see more detailed description in Section 4). We generated 100 data sets and each data set contains 5,000 observations. We compared the performance of UGEE and AUGEE for $K = 5, 10, 20$. Note that UGEE can also be viewed as AUGEE with $K = 1$. Figure 2 shows the boxplots of the estimates of $\boldsymbol{\beta}$ and λ in the 100 simulations. Clearly, the estimates of $\boldsymbol{\beta}$ and λ with different choices of K are very close. Table 2 shows the mean computation time for different types of estimating equations as well as the mean and variances of the 100 estimates in each scenario. Clearly, as K increases the computation time decreases significantly, and the variances of the estimates largely remains at the same level for different K. For example, for $K = 5$, all estimates' mean and variance are all very close to the case of $K = 1$ (i.e. UGEE), but its computation time is only 1/5 of the UGEE.

4. REAL DATA ANALYSIS

In this section, we apply the AUGEE model to estimate the GC-dependence and the over-dispersion parameter in the high-throughput sequencing (HTS) genomic data. In recent years, the breakthrough of the HTS technology has

Table 2. Comparison of estimates from different estimating equations

	K	Mean	Variance($\times 10^{-4}$)	Time (seconds)
	1	5.000061	0.51	20647.9
β_0	5	5.000005	0.51	4194.0
	10	4.999919	0.51	2091.3
	20	4.999733	0.51	1039.3
	1	2.998614	7.58	20647.9
β_1	5	2.998798	7.52	4194.0
	10	2.998992	7.50	2091.3
	20	2.999377	7.57	1039.3
	1	−1.998244	6.15	20647.9
β_2	5	−1.998412	6.12	4194.0
	10	−1.998500	6.10	2091.3
	20	−1.998635	6.19	1039.3
	1	2.010236	43.5	20647.9
λ	5	2.005958	43.6	4194.0
	10	1.999551	42.4	2091.3
	20	1.985545	42.9	1039.3

revolutionized the research in many biological fields. The HTS technology has been applied in various biological assays such as SNP detection, copy number variation detection, gene expression analysis and epigenetic studies. It is well-known that the short read coverage of the HTS data can be influenced by many biological and technical factors [23]. If the technical factors are not properly accounted for, the biological analysis based on HTS data would be misleading. Here, we use the U-FRM to study one of the most important factors, the GC-content on the short read coverage in HTS data. The GC-content refers to the proportion of G and C in a genomic region.

We consider the HTS data HG00103 sequenced from the 1000 Genome Project [15]. After aligning the short reads to the human reference genome hg18 using BWA [11], we extracted the mapping positions of the short reads and binned the data to 50 Kb (K-basepair) bins. In each bin, we counted the total number of short reads and calculated the GC-percentage. The original HTS data is around 10 Gb (in fastq format) and we got around 238,000 bins after binning the data to 50 Kb bins. Figure 3 shows the dependence of the read count on the GC-content. We used the same model as in Simulation 2 in Section 3, where x_{1i} is the GC-proportion and y_i is the read count in the ith bin.

Table 3 shows the parameter estimates and computational time of AUGEE with different partition number K. It is clear that all estimates using different K are very similar. In terms of the computational time, larger K would require much less computational time. For example, when $K = 20$, the computational time is 10,624.66 seconds, only 5% of the computational time for $K = 1$. Figure 3 shows the estimated function $\exp(\beta_0 + \beta_1 x + \beta_2 x^2)$ with different choices of K. Note that since these estimates are so close

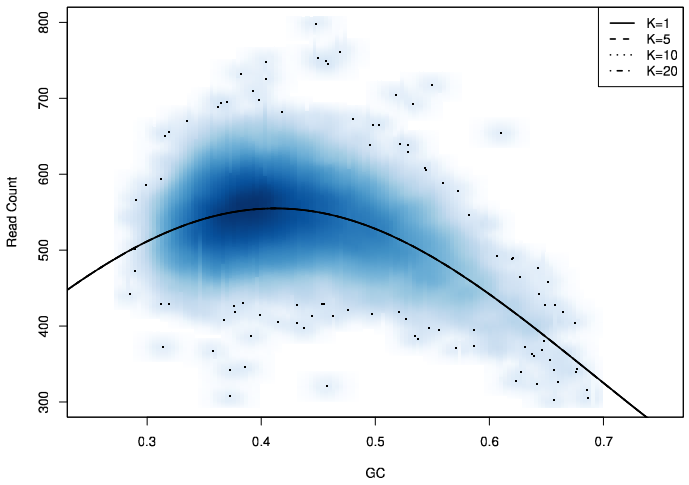

Figure 3. The GC dependence of Read Count in 50 KB for the 1000 Genome Individual HG00103; The curves are estimated using the model same as Simulation 2 with different partition number K.

that the fitted curves are almost the same and they visually seem the same.

5. CONCLUSION AND DISCUSSION

In this paper, we consider direct estimation of high-order moments in a regression setting and focus on the nonparametric U-FRM model that imposes no restrictions on the mean structure. Estimation of the U-FRM model was originally proposed to be estimated using a U-statistics-based estimating equation and its high computational complexity makes it difficult to be applied to big data problems. In this paper, we proposed an efficient computational strategy by constructing a surrogate estimating equation using the "divide-and-conquer" method. Our new approach significantly reduces the computational complexity and is proved to meanwhile maintain asymptotic equivalence. Its merit is further illustrated through an application in genomics. In addition, the original estimator of the U-FRM model was presented in [10] without its asymptotic behavior, and our theory also fills this blank (see Theorem 1). Further, by ideas similar in [24], our computational strategy also enables a storage scheme to support the online analytical processing of the U-FRM model in some big data environments such as data cubes and data streams.

Semiparametric and nonparametric models are often appealing for big data problems for that parametric models may not offer enough flexibility in accounting for the complex structure of such data. However, the high computational burden can be a serious bottleneck in promoting them. The method presented in this paper, along with our earlier work [13], provides a general computational strategy for implementing U-statistics-based methods in large scale problems. And we hope they may shed light on how to revive many traditional nonparametric methods in big data.

Table 3. Real Data Analysis

	K	Mean	Time (seconds)
β_0	1	5.221022	214399.57
	5	5.220018	42683.05
	10	5.219381	21320.32
	20	5.220067	10624.66
β_1	1	5.326345	214399.57
	5	5.330730	42683.05
	10	5.332967	21320.32
	20	5.330220	10624.66
β_2	1	−6.459848	214399.57
	5	−6.464496	42683.05
	10	−6.466170	21320.32
	20	−6.463578	10624.66
λ	1	2.764525	214399.57
	5	2.762958	42683.05
	10	2.761451	21320.32
	20	2.758854	10624.66

APPENDIX

In this section, we give the proof of Theorem 1. We first give a theorem about the asymptotic normality for the vector-valued AU-statistics, which itself is also of interest.

Theorem 2. *Suppose that* $\mathbf{h} = (h^1, \cdots, h^p)^T$ *is a* p*-dimensional vector-valued measurable functions which is symmetric about its arguments. Let* $\tilde{\mathbf{U}}_N$ *be the vector-valued AU-statistic with kernel* $\mathbf{h}$*. Suppose* $E[h^i(X_1, \cdots, X_m)]^2 < \infty$ *for all* $i = 1, \cdots, p$ *and* $\boldsymbol{\zeta}_1(\mathbf{h})$ *is positive definite. Then, if* $K = o(N)$*, one has*

$$\sqrt{N}[\tilde{\mathbf{U}}_N - \boldsymbol{\vartheta}] \xrightarrow{d} \mathcal{N}(0, m^2\boldsymbol{\zeta}_1(\mathbf{h})) \quad as \quad N \to \infty,$$

where $\boldsymbol{\vartheta} = E[\mathbf{h}(X_1, \cdots, X_m)]$.

Proof. It is sufficient to prove that for any nonzero vector $\mathbf{c} = (c_1, \cdots, c_p)^T \in \mathbb{R}^p$, we have

(9)
$$\sqrt{N}[\mathbf{c}^T\tilde{\mathbf{U}}_N - \mathbf{c}^T\boldsymbol{\vartheta}] \xrightarrow{d} \mathcal{N}(0, m^2\mathbf{c}^T\boldsymbol{\zeta}_1(\mathbf{h})\mathbf{c}) \quad \text{as} \quad N \to \infty.$$

It is easy to see that $\mathbf{c}^T\tilde{\mathbf{U}}_N$ is an AU-statistic with kernel $g = \mathbf{c}^T\mathbf{h}$ and $E[g(X_1, \cdots, X_n)] = \mathbf{c}^T\boldsymbol{\vartheta}$. Since $E[h^i(X_1, \cdots, X_m)]^2 < \infty$ for all $i = 1, \cdots, p$, we have

$$\begin{aligned} & E[g(X_1, \cdots, X_m)]^2 \\ = & \sum_{i,j=1}^{p} E[c_i c_j h^i(X_1, \cdots, X_m) h^j(X_1, \cdots, X_m)] < \infty. \end{aligned}$$

At last, since $\zeta_1(g) = \mathbf{c}^T\boldsymbol{\zeta}_1(\mathbf{h})\mathbf{c} > 0$ and $K = o(N)$, we get the asymptotic normality (9) from Theorem 2 in Lin and Xi (2010) [13]. □

Kantorovitch's theorem, whose proof can be found in [8], is needed in proving the consistency and asymptotic normality of the estimator $\tilde{\boldsymbol{\theta}}_{K,N}$. For ease of reference, we list Kantorovitch's theorem as the following lemma.

Lemma 1 (Kantorovitch's theorem)**.** *Let* $\mathbf{a}_0$ *be a point in* $\mathbb{R}^p$*,* U *an open neighborhood of* $\mathbf{a}_0$ *and* $\mathbf{f} : U \mapsto \mathbb{R}^p$ *a differential mapping, with its derivative* $D\mathbf{f}(\mathbf{a}_0)$ *invertible. Define*

$$\mathbf{r}_0 = -D\mathbf{f}(\mathbf{a}_0)^{-1}\mathbf{f}(\mathbf{a}_0), \ \mathbf{a}_1 = \mathbf{a}_0 + \mathbf{r}_0,$$
$$U_0 = \{\mathbf{x} | \ \|\mathbf{x} - \mathbf{a}_1\| \le \|\mathbf{r}_0\|\}$$

If the derivative $D\mathbf{f}(\mathbf{a}_0)$ *satisfies the Lipschitz condition*

$$\|D\mathbf{f}(\mathbf{x}_1) - D\mathbf{f}(\mathbf{x}_2)\| \le M\|\mathbf{x}_1 - \mathbf{x}_2\|,$$

for all points $\mathbf{x}_1, \mathbf{x}_2 \in U_0$ *and if the inequality*

$$\|\mathbf{f}(\mathbf{a}_0)\| \cdot \|D\mathbf{f}(\mathbf{a}_0)^{-1}\|^2 M \le \frac{1}{2}$$

is satisfied, the equation $\mathbf{f}(\mathbf{x}) = \mathbf{0}$ *has a unique solution in* U_0*.*

Now we are ready to give the proof of Theorem 1.

Proof of Theorem 1. A. Consistency. Since $\mathbf{h}_{\boldsymbol{\theta}_0}$ satisfies Condition (C1), by Theorem 2 we have $\tilde{\mathbf{U}}_N(\boldsymbol{\theta}_0) = o_p(1)$. From Condition (C3), we get

$$\frac{\partial\tilde{\mathbf{U}}_N}{\partial\boldsymbol{\theta}}(\boldsymbol{\theta}_0) = E\left[\frac{\partial\mathbf{h}}{\partial\boldsymbol{\theta}}(Z_1, \cdots, Z_m; \boldsymbol{\theta}_0)\right] + o_p(1).$$

Then, $\frac{\partial\tilde{\mathbf{U}}_N}{\partial\boldsymbol{\theta}}(\boldsymbol{\theta}_0)$ is invertible in probability and $\mathbf{r}_N = -(\frac{\partial\tilde{\mathbf{U}}_N}{\partial\boldsymbol{\theta}}(\boldsymbol{\theta}_0))^{-1}\tilde{\mathbf{U}}_N(\boldsymbol{\theta}_0)$ tends to zero in probability. By Condition (C3), it is straightforward to show that there exists a neighborhood U of $\boldsymbol{\theta}_0$ and a constant M such that in probability

$$\left\|\frac{\partial\tilde{\mathbf{U}}_N}{\partial\boldsymbol{\theta}}(\boldsymbol{\theta}_1) - \frac{\partial\tilde{\mathbf{U}}_N}{\partial\boldsymbol{\theta}}(\boldsymbol{\theta}_2)\right\| \le M\|\boldsymbol{\theta}_1 - \boldsymbol{\theta}_2\|$$

for all $\boldsymbol{\theta}_1, \boldsymbol{\theta}_2 \in U$. Again, since $\tilde{\mathbf{U}}_N(\boldsymbol{\theta}_0) = o_p(1)$ and $(\frac{\partial\tilde{\mathbf{U}}_N}{\partial\boldsymbol{\theta}}(\boldsymbol{\theta}_0))^{-1}$ is bounded in probability, we have $\|\tilde{\mathbf{U}}_N(\boldsymbol{\theta}_0)\|\|(\frac{\partial\tilde{\mathbf{U}}_N}{\partial\boldsymbol{\theta}}(\boldsymbol{\theta}_0))^{-1}\|^2 M \le 1/2$ in probability. Then, by Kantorovitch's theorem, there exists a unique solution $\tilde{\boldsymbol{\theta}}_N$ in the neighborhood $U_N = \{\boldsymbol{\theta} | \ \|\boldsymbol{\theta} - \boldsymbol{\theta}_N\| \le \mathbf{r}_N\}$ in probability, where $\boldsymbol{\theta}_N = \boldsymbol{\theta}_0 + \mathbf{r}_N$. Then, we have $\|\tilde{\boldsymbol{\theta}}_N - \boldsymbol{\theta}_0\| \le 2\|\mathbf{r}_N\| = o_p(1)$ and the estimator $\tilde{\boldsymbol{\theta}}_{K,N}$ is consistent.

B. Normality. Since $\tilde{\boldsymbol{\theta}}_N$ is the solution to Equation (6), we have $\tilde{\mathbf{U}}_N(\tilde{\boldsymbol{\theta}}_N) = \mathbf{0}$. By the Taylor expansion of the vector-valued function $\tilde{\mathbf{U}}_N(\boldsymbol{\theta})$ at $\boldsymbol{\theta}_0$, we have

$$\mathbf{0} = \tilde{\mathbf{U}}_N(\tilde{\boldsymbol{\theta}}_N) = \tilde{\mathbf{U}}_N(\boldsymbol{\theta}_0) + \frac{\partial\tilde{\mathbf{U}}_N}{\partial\boldsymbol{\theta}}(\boldsymbol{\theta}_0)(\tilde{\boldsymbol{\theta}}_N - \boldsymbol{\theta}_0) + \mathbf{R}_2,$$

where $\mathbf{R}_2$ is the second order residual in the Taylor expansion. Therefore, we have the following representation

$$\begin{aligned} & \sqrt{N}(\tilde{\boldsymbol{\theta}}_N - \boldsymbol{\theta}_0) \\ = & -\left(\frac{\partial\tilde{\mathbf{U}}_N}{\partial\boldsymbol{\theta}}(\boldsymbol{\theta}_0)\right)^{-1}\sqrt{N}\left(\tilde{\mathbf{U}}_N(\boldsymbol{\theta}_0) - \mathbf{R}_2\right). \end{aligned}$$

By Conditions (C2) and (C3), we have $\frac{\partial \tilde{\mathbf{U}}_N}{\partial \boldsymbol{\theta}}(\boldsymbol{\theta}_0) \to B$ in probability. Let V_k be the U-statistic with kernel $b(\cdot)$ based on the observations $\{Z_{k1}, \cdots, Z_{kn_k}\}$ and $\tilde{V}_N = \sum_{k=1}^{K} n_k V_k / N$ be the corresponding AU-statistic. Since $\tilde{\boldsymbol{\theta}}_N$ is a consistent estimator of $\boldsymbol{\theta}_0$, we have $\|\mathbf{R}_2\| \le C\tilde{V}_N \|\tilde{\boldsymbol{\theta}}_N - \boldsymbol{\theta}_0\|^2$ in probability for some constant C. From the proof of Part A, we know that $\sqrt{N}\|\tilde{\boldsymbol{\theta}}_N - \boldsymbol{\theta}_0\|^2 \le \sqrt{N}\|\mathbf{r}_N\|^2 = o_p(1)$. Furthermore, $\tilde{V}_N$ goes to $\mathrm{E}[b(Z_1, \cdots, Z_m)]$ in probability. Hence, $\|\sqrt{N}\mathbf{R}_2\| = o_p(1)$ and Theorem 1 can then be proved using the delta method. □

ACKNOWLEDGEMENT

This work was partially supported by the grant NSF-DMS0906023 to N. L. and the National Natural Science Foundation of China (11471022) and the National Key Basic Research Program of China (2015CB856000) to R. X.

Received 30 October 2014

REFERENCES

[1] Chen, Y., Dong, G., Han, J., Pei, J., Wah, B., and Wang, J. (2006). Regression cubes with lossless compression and aggregation. *IEEE Transactions on Knowledge and Data Engineering*, 18, 1585–1599.

[2] Chen, Y., Dong, G., Han, J., Wah, B. W., and Wang, J. (2002). Multi-dimensional regression analysis of time-series data streams. *Proceedings of the International Conference on Very Large Data Bases*, 323–334.

[3] Clarkson, R., Drineas, P., Magdon-Ismail, M., Mahoney, M., Meng, X., and Woodruff, D. (2013). The fast cauchy transform and faster robust linear regression. *Proceedings of the Twenty-Fourth Annual ACM-SIAM Symposium on Discrete Algorithms*, 466–477. MR3186768

[4] Dhillon, P., Lu, Y., Foster, D., and Ungar, L. (2013). New subsampling algorithms for fast least squares regression. *Advances in Neural Information Processing Systems*, 360–368.

[5] Gray, J., Chaudhuri, S., Bosworth, A., Layman, A., Reichart, D., Venkatrao, M., Pellow, F., and Pirahesh, H. (1997). Data cube: A relational aggregation operator generalizing group-by, cross-tab and sub-totals. *Data Mining and Knowledge Discovery*, 1, 29–54.

[6] Han, J., Chen, Y., Dong, G., Pei, J., Wah, B. W., Wang, J., and Cai, Y. (2005). Stream cube: An architecture for multi-dimensional analysis of data streams. *Distributed and Parallel Databases*, 18(2):173–197.

[7] Harinarayan, V., Rajaraman, A., and Ullman, J. (1996). Implementing data cubes efficiently. *Proceedings of the ACM SIGMOD*, 205–216.

[8] Hubbard, J. and Hubbard, B. (1999). *Vector Calculus, Linear Algebra, and Differential Forms*. Prentice-Hall, New Jersey. MR1657732

[9] Kendall, M. (1938). A new measure of rank correlation. *Biomerika*, 30:81–89.

[10] Kowalski, J. and Tu, X. (2008). *Modern Applied U-Statistics*. John Wiley & Sons, Hoboken, New Jersey. MR2368050

[11] Li, H. and Durbin, R. (2009). Fast and accurate short read alignment with burrows–wheeler transform. *Bioinformatics*, 25(14):1754–1760.

[12] Liang, F., Cheng, Y., Song, Q., Park, J., and Yang, P. (2013) A resampling-based stochastic approximation method for analysis of large geostatistical data. *Journal of the American Statistical Association*, 108(501):325–339. MR3174623

[13] Lin, N. and Xi, R. (2010). Fast surrogates of u-statistics. *Computational Statistics & Data Analysis*, 54(1):16–24. MR2558454

[14] Lin, N. and Xi, R. (2011). Aggregated estimating equation estimation. *Statistics and Its Interface*, 4:73–83. MR2775250

[15] Mills, R. et al. (2011). Mapping copy number variation by population-scale genome sequencing. *Nature*, 470(7332):59–65

[16] Liu, C., Zhang, M., Zheng, M., and Chen, Y. (2003). Step-by-step regression: A more efficient alternative for polynomial multiple linear regression in stream cube. *Proceedings of the Seventh Pacific-Asia Conference on Knowledge Discovery and Data Mining*, 437–448.

[17] Ma, P., Mahoney, M., and Yu, B. (2014). A statistical perspective on algorithmic leveraging. *Proceedings of The 31st International Conference on Machine Learning*, 91–99. MR3361306

[18] Mann, H. and Whitney, D. (1947). On a test of whether one of two random variables is stochastically larger than the other. *Annals of Mathematical Statistics*, 18:50–60. MR0022058

[19] Tu, X. M., Feng, C., Kowalski, J., Tang, W., Wang, H., Wan, C., and Ma, Y. (2007). Correlation analysis for longitudinal data: applications to hiv and psychosocial research. *Statistics in medicine*, 26(22):4116–4138. MR2405796

[20] Uhler, C. (2012). Geometry of maximum likelihood estimation in gaussian graphical models. *The Annals of Statistics*, 40(1):238–261. MR3014306

[21] Wang, L., Zheng, W., Zhao, H., and Deng, M. (2013). Statistical analysis reveals co-expression patterns of many pairs of genes in yeast are jointly regulated by interacting loci. *PLoS genetics*, 9(3):e1003414.

[22] Wilcoxon, F. (1945). Individual comparisons by ranking methods. *Biometrics Bulletin*, 1:80–83.

[23] Xi, R., Hadjipanayis, A., Luquette, L., Kim, T., Lee, E., Zhang, J., Johnson, M., Muzny, D., Wheeler, D., Gibbs R., Kucherlapati, R., and Park, P. (2011). Copy number variation detection in whole-genome sequencing data using the bayesian information criterion. *Proceedings of the National Academy of Sciences*, 108(46):e1128–1136

[24] Xi, R., Lin, N., and Chen, Y. (2009). Compression and aggregation for logistic regression analysis in data cubes. *IEEE Transactions on Knowledge and Data Engineering*, 21(4):479–492.

[25] Xi, R., Lin, N., Chen, Y., and Kim, Y. (2012). Compression and aggregation of bayesian estimates for data intensive computing. *Knowledge and information systems*, 33(1):191–212.

[26] Yin, Y. and Li, H. (2011). A sparse conditional gaussian graphical model for analysis of genetical genomics data. *The Annals of Applied Statistics*, 5(4):2630. MR2907129

[27] Yu, Q., Chen, R., Tang, W., He, H., Gallop, R.,Christoph, P. C., Hu. J., and Tu, X. M. (2013). Distribution-free models for longitudinal count responses with overdispersion and structural zeros. *Statistics in Medicine*, 32(14):2390–2405. MR3067391

[28] Yuan, M. and Lin, Y. (2007). Model selection and estimation in the gaussian graphical model. *Biometrika*, 94(1):19–35. MR2367824

Ruibin Xi
School of Mathematical Sciences
and Center of Statistical Science
Peking University
China
E-mail address: ruibinxi@math.pku.edu.cn

Nan Lin
Department of Mathematics
Washington University in St. Louis
USA
E-mail address: nlin@wustl.edu