Globally Adaptive Longitudinal Quantile Regression with High Dimensional Compositional Covariates

Huijuan Ma,

Qi Zheng,

Zhumin Zhang,

Huichuan Lai,

Limin Peng

East China Normal University, University of Louisville, University of Wisconsin-Madison, University of Wisconsin-Madison, Emory University

Abstract

In this work, we propose a longitudinal quantile regression framework that enables a robust characterization of heterogeneous covariate-response associations in the presence of high-dimensional compositional covariates and repeated measurements of both response and covariates. We develop a globally adaptive penalization procedure, which can consistently identify covariate sparsity patterns across a continuum set of quantile levels. The proposed estimation procedure properly aggregates longitudinal observations over time, and ensures the satisfaction of the sum-zero coefficient constraint that is needed for proper interpretation of the effects of compositional covariates. We establish the oracle rate of uniform convergence and weak convergence of the resulting estimators, and further justify the proposed uniform selector of the tuning parameter in terms of achieving global model selection consistency. We derive an efficient algorithm by incorporating existing R packages to facilitate stable and fast computation. Our extensive simulation studies confirm the theoretical findings. We apply the proposed method to a longitudinal study of cystic fibrosis children where the association between gut microbiome and other diet-related biomarkers is of interest.

Key words and phrases:

Compositional covariates; Globally adaptive penalization; Longitudinal data; Quantile regression

---

Corresponding author: lpeng@emory.edu (Limin Peng).

The first two authors have equal contributions.

Huijuan Ma, KLATASDS-MOE, Academy of Statistics and Interdisciplinary Sciences, East China Normal University, Shanghai 200062, China

Qi Zheng, Department of Bioinformatics and Biostatistics, University of Louisville, Louisville, KY 40202, USA

Zhumin Zhang, Department of Nutritional Sciences, University of Wisconsin-Madison, Madison, WI 53706, USA

Huichuan Lai, Department of Nutritional Sciences, University of Wisconsin-Madison, Madison, WI 53706, USA

Limin Peng, Department of Biostatistics and Bioinformatics, Emory University, Atlanta, GA 30322, USA

1. Introduction

Compositional data have been frequently encountered in a variety of research fields. Examples include household expenditure compositions in economics, geochemical compositions of rocks in geology, and human microbiome compositions in medical studies. Compositional data consist of proportions bounded between zero and one and summed up to one, and are often high-dimensional. For instance, human microbiome data are usually captured by percentages (or relative abundance) of gene sequencing reads (Tyler et al., 2014) at a certain taxonomy level, and the number of operational taxonomy units (OTU) (e.g. phyla or genus) can range over hundreds, thousands, or even millions. With technology advancement, an increasing number of studies have collected such compositional data repeatedly over time. A common question of substantive interest is how these longitudinal compositional measurements are associated with other longitudinal biomarkers or clinical outcomes. This poses a regression problem subject to multi-fold complications, including a large number of covariates, positiveness and unit-sum constraints to the covariates, and within-subject dependence of longitudinal observations.

To deal with the high-dimensionality of covariates, a notable line of research has been established in the penalization framework (for example, Meinshausen and Buhlmann, 2006; Zhang and Huang, 2008; Kim et al., 2008; Lv and Fan, 2009; Fan and Lv, 2011). Extensions to longitudinal settings have been developed (for example, Wang et al., 2012; Zheng et al., 2018). When covariates are compositional, given the unit-sum constraint, an increase in one covariate must induce a decrease in another covariate. Applying the traditional penalization regression methods without accounting for the compositional nature of covariates may lead to results that are difficult to interpret. A common strategy for accommodating compositional covariates is to apply a sensible operation to the compositional proportions before incorporating them into a regression model, as in the linear log-contrast model and logistic normal multinomial regression model (Aitchison, 1982; Aitchison and Bacon-shone, 1984; Aitchison, 2003; Xia et al., 2013). Many efforts have also been devoted to deal with covariates that are both compositional and high-dimensional. For example, Lin et al. (2014) proposed Lasso-penalized method for the linear log-contrast regression model that properly accounts for the compositional nature of covariates. Shi et al. (2016) studied an extension of Lin et al. (2014)'s model with a set of linear constraints. Lu et al. (2019) further generalized the model to a generalized linear log-contrast model and proposed a l_1-penalized likelihood estimation procedure.

Methods for addressing high-dimensional compositional covariates in a longitudinal setting, however, has been scarce. Moreover, existing approaches are mostly based on mean-based linear regression, which typically confine covariate effects to be location-shifts and thus can be restrictive for real data. Quantile regression (Koenker and Bassett, 1978), as characterized by its flexibility to assess covariate effects across different quantile levels, has demonstrated promising utility to identify and depict dynamic covariate-response associations that often shed useful scientific insight. The modeling strategy of quantile regression has been incorporated to the analysis longitudinal data under various perspectives (Koenker, 2004; Wang and Fygenson, 2009; Ma, Peng, and Fu, Ma et al., for example). In the present of high-dimensional covariates, many authors (Li et al., 2007; Zou and Yuan, 2008; Wang et

al., 2012; Zheng et al., 2013; Fan et al., 2014, among others) have studied penalized quantile regression methods. These methods model a single or multiple pre-specified quantiles of the response; in other words, are locally concerned. The locally concerned methods are subject to inherent issues such as undesirable variability in variable selection results across neighboring quantile levels, and potential failure to detect some important variables due to an off-target selection of quantile levels. To address these limitations, Zheng et al. (2015) proposed the perspective of globally concerned quantile regression, which allows for simultaneously examining regression quantiles over a continuum set of quantile levels and thus reflects the underlying scientific interest in a more robust way. While demonstrating improved stability and "power" of variable selection compared to locally concerned quantile regression approaches, Zheng et al. (2015)'s method is not suitable for handling either longitudinal data or compositional covariates.

In this work, we develop a globally concerned longitudinal quantile regression framework, which is tailored to evaluate the effects of high-dimensional longitudinal compositional covariates on longitudinal responses. We consider a longitudinal linear log-contrast quantile regression model, where quantiles of the longitudinal response are linked to the log contrasts of the corresponding compositional covariates. To avoid the shortfall associated with selecting an irrelevant covariate as the reference in log-contrasts, we reformulate the model into a symmetric form with zero-sum constraint of coefficients, which ensures sensible interpretations of the effects of compositional covariates. We propose a regularization method, where a globally adaptive Lasso penalty is imposed to the longitudinal quantile loss function that appropriately aggregates the repeated measurements from the same subject. We further adapt the *rq.fit.fnc()* function in the existing R package *quantreg* to facilitate the estimation in the presence of the zero-sum constraint of coefficients.

We conduct theoretical studies for the proposed method in the ultra-high dimensional setting, where the number of covariates p may increase exponentially with sample size n (i.e. $\log p = o(n^b)$ for some $b > 0$) and the number of relevant covariates s also increases with n. We attain the uniform convergence rate of the proposed estimator as $O_p(\sqrt{s \log n/n})$, which is fastest possible. Because the longitudinal quantile loss function is not differentiable, to attain this result, we cannot adapt the existing work on linear regression based methods for high-dimensional compositional data, such as Lin et al. (2014), which penalizes a smooth least-squares loss function. Instead, we employ theoretical techniques, including chaining theory (Talagrand, 2005), contraction inequality (Ledoux and Talagrand, 1991), and empirical process (van der Vaart and Wellner, 1996), like in Zheng et al. (2015), which however did not address the longitudinal data structure and the compositional constraint for high dimensional covariates. In this work, we develop new lines of arguments to account for these special data features. A notable effort is that we properly formulate and establish a crucial Karush-Kuhn-Tucker (KKT) condition tailored to compositional data, which is new in literature. In addition, we thoroughly justify that penalizing the proposed longitudinal quantile loss function, which adopts the simple working independence assumption, is capable of accommodating longitudinal data with dependent repeated measures.

Our theoretical studies confer some useful results that were not discussed in existing work that handles high-dimensional compositional covariates based on log-contrast models, such

as Lin et al. (2014) and Shi et al. (2016). For example, our theoretical investigation reveals that the asymptotic behavior of the globally adaptive estimator based on a constrained linear log-contrast quantile regression model is asymptotically equivalent to its unconstrained counterpart as long as the reference variable for the latter is a truly relevant variable, which however would not be known in advance. In addition, we establish the weak convergence of any linear combination of the proposed estimator to a Gaussian process. We develop a GIC-type uniform tuning parameter selector. We show that the proposed estimation and tuning parameter procedures can correctly identify all globally relevant variables with probability tending to one (i.e. global model selection consistency).

The rest of this article is outlined as follows. In Section 2, we introduce a globally concerned framework built upon a longitudinal linear log-contrast quantile regression model. Then we propose a globally adaptive regularization procedure based on a symmetric model representation with the zero-sum coefficient constraint. In Section 3, we present the asymptotic studies for the proposed estimation procedure. In Section 4, we investigate the finite sample performance of proposed method through simulation studies. Finally, our methodology is illustrated by an application to a longitudinal observational study of cystic fibrosis children.

2.   Methodology

2.1   Longitudinal Linear Log-contrast Quantile Regression Model

Consider a longitudinal study with n subjects. Let $Y_i(t)$, $\mathbf{X}_i(t)$, and $\mathbf{W}_i(t)$, respectively, denote the longitudinal response, $r \times 1$ vector of regular covariates including 1 as the first component, and $p \times 1$ vector of compositional covariates at time t for subject i ($i = 1, \ldots, n$). A component of $\mathbf{X}_i(t)$ may flexibly represent the value of a time-dependent covariate measured at time t or a summary of the covariate history up to time t. We consider the setting where r is fixed and p increases with n satisfying $\log p = o(n^b)$ for some $b > 0$. At each time point t, the compositional covariates in $\mathbf{W}_i(t)$ are subject to the unit-sum constraint. That is, $\mathbf{W}_i(t)$ is belong to the $(p-1)$-dimensional positive simplex $\mathbb{S}^{p-1} = \{(w_1, \ldots, w_p): w_j > 0, j = 1, \ldots, p; \ \sum_{j=1}^{p} w_j = 1\}$. Suppose $Y_i(t)$, $\mathbf{X}_i(t)$, and $\mathbf{W}_i(t)$ are observed at m_i time points, denoted by $\{t_i^{(k)}, k = 1, \ldots, m_i\}$. Define a counting process for the observation time as $N_i(t) = \sum_{k=1}^{m_i} I(t_i^{(k)} \le t)$.

To gain a comprehensive and flexible view of how covariates influence the response, we adopt quantile regression modeling to formulate covariate effects on the τth conditional quantile of $Y(t)$ given $\mathbf{X}(t)$ and $\mathbf{W}(t)$, which is defined as $Q_{Y(t)}\{\tau | \mathbf{X}(t), \mathbf{W}(t)\} = \inf\{y : \Pr\{Y(t) \le y | \mathbf{X}(t), \mathbf{W}(t)\} \ge \tau\}$. However, directly plugging in $\mathbf{W}(t)$ into a regression model would be problematic because the components of $\mathbf{W}(t)$ can not change freely due to the unit-sum constraint, and thus interpreting the coefficients for $\mathbf{W}(t)$ would be difficult. To deal with the unit-sum constraint, we apply Aitchison and Bacon-shone (1984)'s log-contrast (or log-ratio) transformation that transforms the compositional $\mathbf{W}_i(t)$ from $\mathbb{S}^{p-1}$ to $\mathbf{Z}_i^p(t) \doteq \{\log\{W_{i1}(t)/W_{ip}(t)\}, \ldots, \log\{W_{i,p-1}(t)/W_{ip}(t)\}\}^\top$, where $W_{ij}(t)$ denotes the jth component of $\mathbf{W}_i(t)$. The transformation from $\mathbf{W}(t)$ to $\mathbf{Z}_i^p(t)$ is one-to-one and $\mathbf{Z}_i^p(t)$ is freely

ranged in R^{p-1} without any constraint. A log-contrast transformation requires selecting a reference covariate. For $\mathbf{Z}_i^p(t)$, the p-th component of $\mathbf{W}(t)$, $W_{ip}(t)$, serves as the reference.

We consider the following longitudinal linear log-contrast quantile regression model:

$$Q_{Y_i(t)}\{\tau \mid \mathbf{X}_i(t), \mathbf{W}_i(t)\} = \mathbf{X}_i(t)^\top \boldsymbol{\alpha}_0(\tau) + \mathbf{Z}_i^p(t)^\top \boldsymbol{\beta}_{0,\backslash p}(\tau) \quad \text{for} \quad \tau \in \Delta, \quad (2.1)$$

where $\boldsymbol{\alpha}_0(\tau)$ is a $r \times 1$ vector of regression coefficients for $\mathbf{X}_i(t)$, $\boldsymbol{\beta}_{0,\backslash p}(\tau) \doteq \{\beta_{0,1}(\tau), \ldots, \beta_{0,p-1}(\tau)\}^\top$ is a $(p-1) \times 1$ vector of regression coefficients for $\mathbf{Z}_i^p(t)$, and $\Delta \subset (0,1)$ is a set of quantile levels, pre-specified to align with the scientific problem of interest. For example, if the interest is to identify variables influencing the center of the response distribution, one may choose $\Delta = [0.4, 0.6]$. If the interest lies in the upper tail of the response distribution, one may choose $\Delta = [0.75, 0.9]$. One subtle drawback of model (2.1) is that any variable selection based on model (2.1) would automatically include $W_{ip}(t)$ as a relevant covariate, even when $W_{ip}(t)$ is not a relevant variable.

Following the strategy employed in the setting of linear regression with compositional covariates (Lin et al., 2014; Shi et al., 2016), we define $\beta_{0,p}(\tau) = -\sum_{j=1}^{p-1} \beta_{0,j}(\tau)$, and re-express model (2.1) as

$$Q_{Y_i(t)}\{\tau \mid \mathbf{X}_i(t), \mathbf{Z}_i(t)\} = \mathbf{X}_i(t)^\top \boldsymbol{\alpha}_0(\tau) + \mathbf{Z}_i(t)^\top \boldsymbol{\beta}_0(\tau), \text{ subject to } \sum_{j=1}^{p} \beta_{0,j}(\tau) = 0,$$
$$\text{for} \quad \tau \in \Delta. \quad (2.2)$$

Here $\mathbf{Z}_i(t) = \{\log\{W_{i1}(t)\}, \ldots \log\{W_{ip}(t)\}\}^\top$ and $\boldsymbol{\beta}_0(\tau) = \{\beta_{0,1}(\tau), \ldots, \beta_{0,p-1}(\tau), \beta_{0,p}(\tau)\}^\top$ with $\beta_{0,j}(\tau)$ denoting the jth component of $\boldsymbol{\beta}_0(\tau)$. Unlike model (2.1), model (2.2) takes a symmetric form and does not involve a choice of the reference covariate. The symmetric form of model (2.2) also enables estimation that possesses the desirable properties including scale invariance, permutation invariance and selection invariance (Aitchison, 1982; Lin et al., 2014).

Many longitudinal quantile regression models studied in literature (for example, Lipsitz et al., 1997; Wang and Fygenson, 2009; Sun et al., 2016; Cho et al., 2016; Gao and Liu, 2019) bear similar forms to model (2.1) or (2.2) but do not involve the zero-sum coefficient constraint. In addition, they were investigated under the locally concerned perspective.

We study a globally concerned framework based on the longitudinal quantile regression model (2.2), where a covariate is considered as relevant if it has nonzero effects on the conditional quantiles of $Y(t)$ at some, not necessarily all, quantile levels in Δ. That is, the set of relevant (or active) compositional covariates is defined as

$$S_\Delta = \{j \in \{1, \ldots, p\} : \exists \tau \in \Delta, |\beta_{0,j}(\tau)| > 0\}.$$

It is clear that $S_\tau \doteq S_{\{\tau\}} \subset S_\Delta$ when $\tau \in \Delta$. The globally concerned perspective warrants a global sparsity assumption, i.e. $s \doteq |S_\Delta| = o(n)$, for the model identifiability purpose, where $|\cdot|$ denotes the cardinality.

2.2 Globally adaptive L_1 penalized estimation

The observed longitudinal data can be generally formulated as $\{(Y_i(t)dN_i(t), \mathbf{X}_i(t)dN_i(t),$ $\mathbf{Z}_i(t)dN_i(t)), i = 1, \ldots, n\}$. When p is fixed, model (2.2) without the zero-sum coefficient constraint can be estimated by minimizing the longitudinal quantile loss function,

$$Q(\boldsymbol{\alpha}, \boldsymbol{\beta}; \tau) = \frac{1}{n} \sum_{i=1}^{n} \int_{0}^{\infty} \rho_{\tau}\Big\{Y_i(t) - \mathbf{X}_i(t)^{\top}\boldsymbol{\alpha} - \mathbf{Z}_i(t)^{\top}\boldsymbol{\beta}\Big\} dN_i(t),$$

where $\rho_{\tau}(t) = t(\tau - I\{t \le 0\})$ is the τth quantile loss function. By the definition, $Q(\boldsymbol{\alpha}, \boldsymbol{\beta};$ $\tau)$ takes equal weight summation of the quantile loss function assessed at all within-subject observation time points. This mimics the idea of constructing a generalized estimating equation (GEE) for longitudinal under the working independence assumption (Liang and Zeger, 1986). The same strategy has also been adopted by existing work on longitudinal quantile regression (Wang and Fygenson, 2009; Sun et al., 2016, for example). Estimation based on $Q(\boldsymbol{\alpha}, \boldsymbol{\beta}; \tau)$, like GEE approach, can properly accommodate longitudinal data with correlated repeated measures.

We propose to apply the adaptively weighted L_1 regularization to $Q(\boldsymbol{\alpha}, \boldsymbol{\beta}; \tau)$ to address the high-dimensionality of $\mathbf{Z}_i(t)$. This renders a regression coefficient estimator $\hat{\boldsymbol{\gamma}}(\tau)$ as a solution to the following constrained minimization problem,

$$\hat{\boldsymbol{\gamma}}(\tau) \doteq \left(\hat{\boldsymbol{\alpha}}(\tau)^{\top}, \hat{\boldsymbol{\beta}}(\tau)^{\top}\right)^{\top} = \operatorname{argmin}_{\boldsymbol{\alpha}, \boldsymbol{\beta}, \sum_{j=1}^{p}\beta_j = 0}\left(Q(\boldsymbol{\alpha}, \boldsymbol{\beta}; \tau) + \lambda \sum_{j=1}^{p} \omega_j(\tau)|\beta_j|\right). \qquad (2.3)$$

Aligning with the perspective of globally concerned quantile regression, λ is a tuning parameter which is a constant over τ and controls for the global sparsity over $\tau \in$, namely, S . Here $\omega_j(\tau)$ is a nonnegative adaptive weight function that gauges the impoFance of $Z_{ij}(t)$, the j-th component of $\mathbf{Z}_i(t)$, $j = 1, \ldots, p$. The adaptive weights may take the following forms:$(w1)$ $\omega_j(\tau) = 1/\left|\breve{\beta}_j(\tau)\right|$; $(w2)$ $\omega_j(\tau) = 1/\left(\sup_{\tau \in \Delta}\left|\breve{\beta}_j(\tau)\right|\right)$; $(w3)$ $\omega_j(\tau) = 1/\int_{\Delta}\left|\breve{\beta}_j(\tau)\right|d\tau$, where $\breve{\beta}(\tau)$ is a uniformly consistent estimator of $\boldsymbol{\beta}_0(\tau)$. As discussed in Zheng et al. (2015), $(w2)$ and $(w3)$ are two globally adaptive weights that capture the global impact of a covariate, and may be theoretically and empirically preferable. A uniformly consistent estimator $\breve{\beta}(\tau)$ can be obtained from directly adapting Belloni and Chernozhukov (2011)'s approach to high-dimensional longitudinal compositional data (i.e. solving the minimization problem (2.3) with the penalty term and tuning parameter selector presented Belloni and Chernozhukov (2011)). This can be justified by slightly modifying the proof of Theorem 1 (in Section 3), combined with the techniques of Belloni and Chernozhukov (2011).

To solve the constrained minimization problem in (2.3), we first write the objective function as a classical quantile loss function. Let $\mathbf{e}_j$ be a p–dimensional vector with the jth component equal to 1 and all others equal to 0, $j = 1, \ldots, p$. Besides, for any integer $m \ge 2$, denote the m-vector of ones and zeros by $\mathbf{1}_m$ and $\mathbf{0}_m$, respectively. Since $\rho_{\tau}(u) + \rho_{\tau}(-u) = |u|$,

$$\lambda \sum_{j=1}^{p} \omega_j(\tau)|\beta_j| = \sum_{j=1}^{p} \left\{ \rho_\tau \left(Y_j^* - \mathbf{X}_j^{*\top} \boldsymbol{\alpha} - \mathbf{Z}_j^{*\top} \boldsymbol{\beta} \right) + \rho_\tau \left(Y_{p+j}^* - \mathbf{X}_{p+j}^{*\top} \boldsymbol{\alpha} - \mathbf{Z}_{p+j}^{*\top} \boldsymbol{\beta} \right) \right\},$$

where $\left(Y_j^*, \mathbf{X}_j^*, \mathbf{Z}_j^* \right) = (0, \mathbf{0}_r, \lambda \omega_j(\tau)\mathbf{e}_j)$ and $\left(Y_{p+j}^*, \mathbf{X}_{p+j}^*, \mathbf{Z}_{p+j}^* \right) = (0, \mathbf{0}_r, -\lambda \omega_j(\tau)\mathbf{e}_j)$. Letting $\boldsymbol{\gamma} = (\boldsymbol{\alpha}^\top, \boldsymbol{\beta}^\top)^\top$, we then formulate the equality constraint $\sum_{j=1}^{p} \beta_j = 0$ as two inequality constraints $\sum_{j=1}^{p} \beta_j \geq 0$ (or expressed as $(\mathbf{0}_r^\top, \mathbf{1}_p^\top)^\top \boldsymbol{\gamma} \geq 0$ in matrix form) and $-\sum_{j=1}^{p} \beta_j \geq 0$ (or expressed as $(\mathbf{0}_r^\top, -\mathbf{1}_p^\top)^\top \boldsymbol{\gamma} \geq 0$ in matrix form). Then the quantile regression problem in (2.3) with linear inequality constraints can be solved by the existing function *rq.fit.fnc()* in R package *quantreg*, using the augmented dataset $\left\{ Y_i(t_i^{(k)}), \mathbf{X}_i(t_i^{(k)}), \mathbf{Z}_i(t_i^{(k)}), k = 1, ..., m_i; i = 1, ..., n \right\}$, coupled with $\left\{ \left(Y_j^*, \mathbf{X}_j^*, \mathbf{Z}_j^* \right), \left(Y_{p+j}^*, \mathbf{X}_{p+j}^*, \mathbf{Z}_{p+j}^* \right), j = 1, ..., p \right\}$.

The set of relevant compositional covariates, S, is estimated by

$$\hat{S}_\Delta \doteq \left\{ j \in \{1, ..., p\} : \exists \tau \in \Delta, \left| \hat{\beta}_j(\tau) \right| > 0 \right\}.$$

2.3   Tuning parameter selection

Tuning parameter selection plays an important role in variable selection. In the globally concerned setting, a critical idea is to set λ as a common tuning parameter across all $\tau \in$ as a means to control the overall model complexity and avoid overall fitting. We adapt the generalized information criterion (GIC) (Nishii, 1984; Fan and Tang, 2013) to the problem setting of globally concerned longitudinal quantile regression with compositional covariates.

Specifically, we propose the following uniform selector of tuning parameter by minimizing

$$\text{GIC}(\lambda) = \int_\Delta \log \hat{\sigma}_\lambda(\tau) d\tau + \left(\left| \hat{S}_\lambda \right| - 1 \right) \phi_n,$$

where $\hat{S}_\lambda = \left\{ j \in \{1, ..., p\} : \sup_{\tau \in \Delta} \left| \hat{\beta}_{j,\lambda}(\tau) \right| \neq 0 \right\}$,

$$\hat{\sigma}_\lambda(\tau) = \frac{1}{n} \sum_{i=1}^{n} \int_0^\infty \rho_\tau \left\{ Y_i(t) - \mathbf{X}_i(t)^\top \hat{\boldsymbol{\alpha}}_\lambda(\tau) - \mathbf{Z}_i(t)^\top \hat{\boldsymbol{\beta}}_\lambda(\tau) \right\} dN_i(t),$$

and ϕ_n is a sequence converging to 0 with n. Here $\hat{\beta}_{j,\lambda}(\tau)$, $\hat{\boldsymbol{\alpha}}_\lambda(\tau)$, and $\hat{\boldsymbol{\beta}}_\lambda(\tau)$ stand for the proposed estimates for $\beta_j(\tau)$, $\boldsymbol{\alpha}(\tau)$, and $\boldsymbol{\beta}(\tau)$, respectively, with the tuning parameter fixed at λ. A popular choice of ϕ_n is $n^{-1} \log(p) \log(\log(n))$. Note that the model size pertaining to the compositional covariates is $\left| \hat{S}_\lambda \right| - 1$ due to the zero-sum constraint.

As shown in Theorem 3, with a properly chosen ϕ_n and a reasonable upper bound imposed to the model size, the proposed tuning parameter $\hat{\lambda}$, which is the minimizer of $\text{GIC}(\lambda)$ with respect to λ, can consistently identify the true model S. In other words, with probability tending to 1, $\hat{S}_{\hat{\lambda}} = S_\Delta$.

2.4 Grid-based Approximation

With finite sample sizes, minimizing (2.3) for all $\tau \in$ would yield estimates that are exactly piecewise constant functions of τ. While the exact breakpoints of these piecewise constant functions can be identified by adapting Koenker and d'Orey (1987) and Portnoy (1991)'s procedure, the computation expense can be overwhelming in the ultra-high dimensional cases. Therefore, we approximate $\widehat{\alpha}(\cdot)$ and $\widehat{\beta}(\cdot)$ by piecewise constant functions that only jump at the grid points of a pre-specified sufficiently fine τ-grid in to alleviate the computation burden. Let $\mathcal{S}_n$ denote the τ-grid in , $\tau_0 < \tau_1 < \ldots < \tau_{M(n)}$, and define the size of $\mathcal{S}_n$ as $\|\mathcal{S}_n\| = \max\{\tau_k - \tau_{k-1} : k = 1, \ldots, M(n)\}$. The grid-based approximations are given by $\widehat{\alpha}^{\mathcal{S}_n}(\cdot) = \sum_{k=1}^{M(n)} \widehat{\alpha}(\tau_k) I(\tau_{k-1} < \tau \le \tau_k)$, and $\widehat{\beta}^{\mathcal{S}_n}(\cdot) = \sum_{k=1}^{M(n)} \widehat{\beta}(\tau_k) I(\tau_{k-1} < \tau \le \tau_k)$. With certain smoothness assumption for $\boldsymbol{\alpha}_0(\cdot)$ and $\boldsymbol{\beta}_0(\cdot)$, we can show that the $\left(\widehat{\alpha}^{\mathcal{S}_n}(\cdot)^\top, \widehat{\beta}^{\mathcal{S}_n}(\cdot)^\top\right)^\top$ and $\left(\widehat{\alpha}(\cdot)^\top, \widehat{\beta}(\cdot)^\top\right)^\top$ have the same convergence rate and asymptotic distribution if $\|\mathcal{S}_n\|$ converges to 0 at the rate $o((ns)^{-1/2})$.

3. Theoretical Results

Without loss of generality, we assume that r, the number of usual covariates, is finite. Let $S = \{1, \ldots, s\}$ and we use $S_\Delta^c = \{s+1, \ldots, p\}$ to denote the collection of all irrelevant compositional variables. We allow the number of compositional covariates $p_n \doteq p$ and the true model size $s_n \doteq s$ to increase with the sample size n. To ease the presentation, we often omit the subscript n, when it is clear from the context.

Let $\mathbf{V}_i(t) = (\mathbf{X}_i(t)^\top, \mathbf{Z}_i(t)^\top)^\top$ and $\boldsymbol{\gamma}(\tau) = (\boldsymbol{\alpha}(\tau)^\top, \boldsymbol{\beta}(\tau)^\top)^\top$ satisfying $\sum_{j=1}^p \beta_j(\tau) = 0$. Thus, $\boldsymbol{\gamma}_0(\tau) = (\boldsymbol{\alpha}_0(\tau)^\top, \boldsymbol{\beta}_0(\tau)^\top)^\top$. We decompose $\mathbf{Z}_i(t)$ into $(\mathbf{Z}_{ia}(t)^\top, \mathbf{Z}_{ib}(t)^\top)^\top$ and $\mathbf{V}_i(t)$ into $(\mathbf{V}_{ia}(t)^\top, \mathbf{V}_{ib}(t)^\top)^\top$, where $\mathbf{Z}_{ia}(t) = (Z_{i,1}(t), \ldots, Z_{i,s}(t))^\top$, $\mathbf{V}_{ia}(t), = (\mathbf{X}_i(t)^\top, \mathbf{Z}_{ia}(t)^\top)^\top$ and $\mathbf{V}_{ib}(t) = \mathbf{Z}_{ib}(t) = (Z_{i,s+1}(t), \ldots, Z_{i,p}(t))^\top$. Likewise, $\boldsymbol{\beta}(\tau) = (\boldsymbol{\beta}_a(\tau)^\top, \boldsymbol{\beta}_b(\tau)^\top)^\top$ and $\boldsymbol{\gamma}(\tau) = (\boldsymbol{\gamma}_a(\tau)^\top, \boldsymbol{\gamma}_b(\tau)^\top)^\top$, where $\boldsymbol{\beta}_a(\tau) = (\beta_1(\tau), \ldots, \beta_s(\tau))^\top$, $\boldsymbol{\gamma}_a(\tau) = \boldsymbol{\alpha}(\tau)^\top, \boldsymbol{\beta}_a(\tau)^\top)^\top$, and $\boldsymbol{\gamma}_b(\tau) = \boldsymbol{\beta}_b(\tau) = (\beta_{s+1}(\tau), \ldots, \beta_p(\tau))^\top$. Regularity conditions (C1)–(C5) are stated in Section S1 of the Supplementary Materials.

In Theorem 1, we show that the proposed estimator is uniformly consistent over , with the convergence rate, $O_p(\sqrt{(r+s)\log n/n})$, which is fastest possible and is as good as that of Zheng et al. (2015)'s globally adaptive estimator. Concerning a single τ or a finite number of τ's, we establish a faster convergence rate, $O_p(\sqrt{(r+s)/n})$, as stated in Corollary 1.

Theorem 1. *Suppose conditions (C1)–(C5) (stated in the* Supplementary Materials) *hold. Furthermore, we assume that $n/((r+s)^3 \log^2 \max\{n, r+p\}) \to \infty$ and*

$$\sup_{j > r+s, \boldsymbol{\delta} \in R_{r+s-1}} \frac{E\left[\int_0^\infty \left\{V_{ij}(t)\mathbf{V}_i(t)^\top \boldsymbol{\delta}\right\}^2 dN_i(t)\right]}{\|\boldsymbol{\delta}\|^2} = o\left(\frac{\log\max\{n, r+p\}}{(r+s)\log n}\right).$$

If $r + s = o(n^{1/3})$, $\sup_{j \in S_\Delta, \tau \in \Delta} \lambda w_j(\tau) = O_p(\sqrt{n \log n})$, $\lambda/(\sqrt{r+s}\log\max\{n, r+p\}) \to \infty$ and $(\inf_{j > r+s, \tau \in \Delta} w_j(\tau))^{-1} \sqrt{n}/\sqrt{(r+s)\log\max\{n, r+p\}} = O_p(1)$ then the proposed estimator satisfies

$$\sup_{\tau \in \Delta} \| \hat{\gamma}(\tau) - \gamma_0(\tau) \| = O_p(\sqrt{(r+s)\log n/n}).$$

Corollary 1. *Suppose the conditions in Theorem 1 hold. Then the proposed estimator satisfies*

$$\| \hat{\gamma}(\tau_0) - \gamma_0(\tau_0) \| = O_p(\sqrt{(r+s)/n}).$$

In Theorem 2, we establish the weak convergence of the proposed estimator.

Theorem 2. *Suppose the conditions in Theorem 1 hold. If $(r+s)^3 \log^4 n = o(n)$, for any given $\xi \in R_{r+s-1}$ and $\|\xi\| = 1$, we have the following results:*

a.   *If $\sqrt{n/\{(r+s)\log n\}}\, \inf_{1 \le j \le s, \tau \in \Delta} |\beta_{0j}(\tau)| \to \infty$, then*

$$n^{1/2}\xi^{\top}\left[\mathbf{H}_{\tau}\{\hat{\gamma}(\tau) - \gamma_0(\tau)\} + \frac{\lambda}{n}\boldsymbol{\varpi}(\tau)\right]$$

converges weakly to a mean zero Gaussian process with covariance

$$\boldsymbol{\Sigma}(\tau, \tau') = E\{h_{n,\xi,\tau}(\mathbf{V}(t), Y)h_{n,\xi,\tau'}(\mathbf{V}(t), Y)\} - E\{h_{n,\xi,\tau}(\mathbf{V}(t), Y)\}E\{h_{n,\xi,\tau'}(\mathbf{V}(t), Y)\},$$

where $h_{n,\xi,\tau}(\mathbf{V}(t), Y) = \int_0^\infty \xi^{\top}\mathbf{V}(t)\psi_{\tau}\{Y(t) - \mathbf{V}(t)^{\top}\gamma_0(\tau)\}dN(t)$, $\psi_{\tau}(u) = \tau - I(u < 0)$,

$$\mathbf{H}_{\tau} = \begin{pmatrix} E\left[\int_0^\infty f_{t,\tau}\{0 \mid \mathbf{V}_i(t)\}\mathbf{V}_{ia}(t)\mathbf{V}_{ia}(t)^{\top}dN_i(t)\right] & \mathbf{0} \\ \mathbf{0} & \mathbf{0} \end{pmatrix},$$

$\boldsymbol{\varpi}(\tau) = \left(\mathbf{0}_r^{\top}, (\boldsymbol{\omega}(\tau) \circ sign(\boldsymbol{\beta}_0(\tau)))^{\top}, \mathbf{0}_{p-s}^{\top}\right)^{\top}$, $\bigcirc$ denotes the Hadamard product, and $\boldsymbol{\omega}(\tau) = (\omega_1(\tau), \ldots, \omega_p(\tau))^{\top}$;

b.   *If $\sup_{\tau \in \Delta} n^{-1/2}\{\sum_{j \in S_\tau}\lambda w_j^2(\tau)\}^{1/2} = o_p(1)$, then $n^{1/2}\xi^{\top}\mathbf{H}_{\tau}\{\hat{\gamma}(\tau) - \gamma_0(\tau)\}$ converges weakly to a mean zero Gaussian process with covariance $\boldsymbol{\Sigma}(\tau, \tau')$.*

To establish the asymptotic properties of the GIC tuning parameter selector, we assume the following condition (C5+), which is an enhanced version of (C5) presented in the Supplementary Materials:

(C5+) (a)

$$0 < \Lambda_{\min} := \inf_{\delta \in R_\ell, \ \ell \le r+\kappa, \delta \ne 0} \frac{\delta^{\top}E\left[\int_0^\infty \mathbf{V}_i(t)\mathbf{V}_i(t)^{\top}dN_i(t)\right]\delta}{\|\delta\|^2}$$
$$\le \sup_{\delta \in R_\ell, \ \ell \le r+\kappa, \delta \ne 0} \frac{\delta^{\top}E\left[\int_0^\infty \mathbf{V}_i(t)\mathbf{V}_i(t)^{\top}dN_i(t)\right]\delta}{\|\delta\|^2} := \Lambda_{\max} < \infty.$$

(b)

$$q' := \sup_{\delta \in R_\ell, \ \ell \le r + \kappa, \delta \ne 0} \frac{E\left[\int_0^\infty \left|\mathbf{V}_i(t)^\top \delta\right|^2 dN_i(t)\right]^{3/2}}{E\left[\int_0^\infty \left|\mathbf{V}_i(t)^\top \delta\right|^3 dN_i(t)\right]} > 0,$$

where $R_\ell = \left\{\delta = (\delta_x^\top, \delta_z^\top)^\top : \delta_x \in \mathbb{R}^r, \sum_{j=1}^p \delta_{zj} = 0, \|\delta_z\|_0 \le \ell - r\right\}$ with $\|\cdot\|_0$ denoting the L_0 norm.

In addition, we set a model size upper bound, denoted by κ, with $s < \kappa < p$.

$$\xi_n = \min\left\{\min_{1 \le j \le r} \int_\Delta |\alpha_{0j}(\tau)| d\tau, \ \min_{1 \le j \le s} \int_\Delta |\beta_{0j}(\tau)| d\tau\right\},$$

which measures the minimal overall effect of usual and compositional relevant variables upon the conditional distribution. Theorem 3 and Corollary 2 present results on the consistency of tuning parameter selection based on GIC.

Theorem 3. *Suppose the conditions in Theorem 1 and (C5+) hold. Further,* $\log(r + p)/n = o(\phi_n)$, $\phi_n = o(\xi_n^{5/2})$, *and* $\kappa n^{-1} \log \max\{n, r + p\} = o(\xi_n^3)$, *then*

$$P\left(\inf_{S \ne S_\Delta, |S| \le \kappa} GIC(S) > GIC(S_\Delta)\right) \to 1.$$

Corollary 2. *Under the same conditions as in Theorem 3, if*

$$\left\{\inf_{j > r + s, \tau \in \Delta} w_j(\tau)\right\}^{-1} \sqrt{n}/\sqrt{(r + s)\log \max\{n, r + p\}} = O_p(1)$$

and $\sup_{\tau \in \Delta, j \in S_\tau} w_j(\tau) = O_p(\sqrt{n}/(\sqrt{r + s}\log \max\{n, r + p\}))$, *then* $P\left(\hat{S}_{\hat{\lambda}} = S_\Delta\right) \to 1$.

For any $1 \le l \le s$, we use $\mathbf{Z}_i^l(t)$ to denote the log-ratio transformed $\mathbf{W}_i(t)$ when the reference is the lth component, i.e. $\mathbf{Z}_i^l(t)$ is the vector $\mathbf{Z}_i(t) - Z_{i,l}(t)\mathbf{1}_p$ with the lth component removed. We also define $\mathbf{V}_i^l(t) = \left(\mathbf{X}_i(t)^\top, \mathbf{Z}_i^l(t)^\top\right)^\top$. Let $\boldsymbol{\gamma}_{\backslash l}(\tau) = (\boldsymbol{\alpha}(\tau)^\top, \boldsymbol{\beta}_{\backslash l}(\tau)^\top)^\top$ where $\boldsymbol{\beta}_{\backslash l}(\tau) = (\beta_1(\tau), \dots, \beta_{l-1}(\tau), \beta_{l+1}(\tau), \dots, \beta_p(\tau))^\top$, Let $\hat{\boldsymbol{\gamma}}_{\backslash l}(\tau)$ be the solution of the following unconstrained minimization problem

$$\frac{1}{n}\sum_{i=1}^n \int_0^\infty \rho_\tau\left\{Y_i(t) - \mathbf{V}_i^l(t)^\top \boldsymbol{\gamma}_{\backslash l}\right\} dN_i(t) + \lambda \sum_{j=1, j \ne l}^p \omega_j(\tau)|\beta_j|, \tag{3.4}$$

where $\boldsymbol{\gamma}_{\backslash l} = (\alpha_1, \cdots, \alpha_r, \beta_1, \cdots, \beta_{l-1}, \beta_{l+1}, \cdots, \beta_p)^\top$. Then the globally adaptive unconstrained estimator $\hat{\boldsymbol{\gamma}}_l^u(\tau)$ with the lth component as the reference is

$(\hat\gamma_{1,\vee}(\tau), \cdots, \hat\gamma_{r,\vee}(\tau), \hat\gamma_{r+1,\vee}(\tau), \cdots, \hat\gamma_{r+l-1,\vee}(\tau), -\sum_{k=1,k\neq l}^{p}\hat\gamma_{r+k,\vee}(\tau), \hat\gamma_{r+l+1,\vee}(\tau), \cdots, \hat\gamma_{r+p,\vee}(\tau))^{\top}$. We state the asymptotic properties of $\hat\gamma_l^u(\tau)$ in the following theorem:

Theorem 4. *Under the same conditions as in Theorem 2, if* $(r+s)^3 \log^4 n = o(n)$ *and* $\sup_{\tau\in\Delta, j\in S_\tau} n^{-1/2}\lambda w_j(\tau) = o_p(1)$, *then for any given* $\boldsymbol{\xi}\in R_{r+s-1}$, $\|\boldsymbol{\xi}\| = 1$, *and* $1 \le l \le s$, *we have*

a. $n^{1/2}\boldsymbol{\xi}^{\top}\mathbf{H}_r\{\hat\gamma_l^u(\tau) - \gamma_0(\tau)\}$ *converges weakly to a mean zero Gaussian process with covariance* $\Sigma(\tau, \tau')$ *and* $P(\sup_{\tau\in\Delta}\|\hat\gamma_{b,l}^u(\tau)\|_\infty = 0) \to 1$;

b. $n^{1/2}\boldsymbol{\xi}^{\top}\{\hat\gamma_l^u(\tau) - \gamma_0(\tau)\}$ *and* $n^{1/2}\boldsymbol{\xi}^{\top}\{\hat\gamma(\tau) - \gamma_0(\tau)\}$ *are asymptotically equivalent.*

Theorem 4 indicates that the proposed constrained estimator is asymptotically equivalent to an unconstrained estimator that uses a relevant variable as the reference. The latter approach however requires preliminary knowledge about truly relevant variables, which may not be available in real applications.

By our theorems, the technical constraints for s include $(r+s)^3 \log^2 \max\{n, r+p\} = o(n)$ and $(r+s)^3 \log^4 n = o(n)$. When $p = O(n^a)$ $(a > 0)$, we can allow s to be close to but smaller than $o(n^{1/3})$, which is the fastest model size growth rate derived in Welsh (1989) and He and Shao (2000) for an unpenalized quantile regression estimator to achieve asymptotic normality. Theorem proofs are provided in the Supplementary Materials (Section S4).

4. Simulation Studies

In this section, we carry out simulation studies to evaluate the finite sample performance of the proposed method. We consider the sample size $n = 100$ and generate $Y(t)$ based on the assumed quantile regression model with $r = 4$ and $p = 400$. Specifically, we generate the longitudinal observation times $t_i^{(k)}$, $k = 1, \ldots, m_i$, from a standard Poisson process, where m_i is the integer part of $2 + U_i$ with $U_i \sim Uniform(0, 2)$. With $r = 4$, we generate X_{i1} from $Uniform(0, 1)$ and X_{i2} from $Bernoulli(0.5)$. For each observed time point $t = t_i^{(k)}$, we first generate a p-dimensional vector $\widetilde{\mathbf{Z}}_i(t) = (\widetilde{Z}_{i1}(t), \ldots, \widetilde{Z}_{ip}(t))^{\top}$ from a multivariate normal distribution $N_p(\mathbf{0}, \Sigma)$, where $\Sigma = (\rho^{|i-j|})$ with $\rho = 0.5$. Next, we set $\breve{Z}_{ij}(t) = \Phi(\widetilde{Z}_{ij}(t))$ for $j \le 7$ and $\breve{Z}_{i7}(t) = -\Phi(\widetilde{Z}_{i7}(t))$, and then standardize $\breve{Z}_{ij}(t)$ so that its second moment equals 1, where $\Phi(\cdot)$ is the standard normal distribution function and $j = 1, \ldots, p$. The standardized $\breve{Z}_{ij}(t)$'s $(j = 1, \ldots, p)$ form the covariate vector $\mathbf{Z}_i(t)$.

To generate the longitudinal responses, we consider the following four setups: Setup (I): Data are generated from a longitudinal linear model with independent homogeneous errors,

$$Y_i(t) = -X_{i1} + X_{i2} - t + \mathbf{Z}_i(t)^{\top}\mathbf{b} + \epsilon_i(t),$$

where $\mathbf{b} = (1, 0.8, 0.9, 1, 2, -1.5, -4.2, 0, \ldots, 0)^{\top}$, $\epsilon_i(t) \sim N(0, 1)$ for any $t > 0$, and $\epsilon_i(t)$ and $\epsilon_i(t')$ are independent for $t > 0$, $t' > 0$ and $t \neq t'$.

Setup (II): Data are generated from a longitudinal linear model with dependent homogeneous errors,

$$Y_i(t) = -X_{i1} + X_{i2} - t + \mathbf{Z}_i(t)^\top \mathbf{b} + a_i + \epsilon_i(t),$$

where $\mathbf{b} = (1, 0.8, 0.9, 1, 2, -1.5, -4.2, 0, \ldots, 0)^\top$, $a_i \sim N(0, 1/2)$ $\epsilon_i(t) \sim N(0, 1/2)$ for $t > 0$, $\epsilon_i(t)$ and $\epsilon_i(t')$ are independent for $t > 0$, $t' > 0$ and $t \neq t'$, and a_i and $\epsilon_i(t)$ are independent for $t > 0$.

Setup (III): Data are generated from a longitudinal linear model with independent heterogeneous errors,

$$Y_i(t) = -X_{i1} + X_{i2} - t + \mathbf{Z}_i(t)^\top \mathbf{b}_1 + \left(X_{i1} + \mathbf{Z}_i(t)^\top \mathbf{b}_2 \right) \epsilon_i(t),$$

where $\mathbf{b}_1 = \mathbf{b} = (1, 0.9, 0.75, 0.5, 0.8, 1, -4.95, 0, \ldots, 0)^\top$, $\mathbf{b}_2 = (0, 0.25, 0, 1, 0, 0, -1.25, 0, \ldots, 0)^\top$, $\epsilon_i(t) \sim N(0, 1)$ for any $t > 0$, and $\epsilon_i(t)$ and $\epsilon_i(t')$ are independent for $t > 0$, $t' > 0$ and $t \neq t'$.

Setup (IV): Data are generated from a longitudinal linear model with dependent heterogeneous errors,

$$Y_i(t) = -X_{i1} + X_{i2} - t + \mathbf{Z}_i(t)^\top \mathbf{b}_1 + \left(X_{i1} + \mathbf{Z}_i(t)^\top \mathbf{b}_2 \right)(a_i + \epsilon_i(t)),$$

where $\mathbf{b}_1 = \mathbf{b} = (1, 0.8, 0.9, 1, 2, -1.5, -4.2, 0, \ldots, 0)^\top$, $\mathbf{b}_2 = (0, 0.2, 0, 0.1, 0, 0, -0.3, 0, \ldots, 0)^\top$, $a_i \sim N(0, 1/2)$ and $\epsilon_i(t) \sim N(0, 1/2)$ for $t > 0$, $\epsilon_i(t)$ and $\epsilon_i(t')$ are independent for $t > 0$, $t' > 0$ and $t \neq t'$, and a_i and $\epsilon_i(t)$ are independent for $t > 0$.

Under Setups (I) and (II), we can show that

$$Q_{Y_i(t)}\{\tau \mid \mathbf{X}_i(t), \mathbf{Z}_i(t)\} = Q_e(\tau) - X_{i1} + X_{i2} - t + \mathbf{Z}_i(t)^\top \mathbf{b},$$

where $Q_e(\tau)$ is the τ–th quantile of standard normal distribution. Under Setups (III) and (IV), we can show that

$$Q_{Y_i(t)}\{\tau \mid \mathbf{X}_i(t), \mathbf{Z}_i(t)\} = \{-1 + Q_e(\tau)\}X_{i1} + X_{i2} - t + \mathbf{Z}_i(t)^\top \{\mathbf{b}_1 + \mathbf{b}_2 Q_e(\tau)\}.$$

In all setups, the true regression coefficients for $\mathbf{Z}_i(t)$ satisfy the zero-sum constraint at each τ.

We evaluate the finite-sample performance of the proposed globally adaptive Lasso estimators with weights ($w2$) and ($w3$), denoted by AW_2 and AW_3 respectively. We set $= [0.1, 0.9]$ and the τ-grid $\mathcal{S}_n$ as $\{0.1 < 0.125 < \ldots < 0.9\}$. We select the tuning parameter λ by a GIC criterion with $\phi_n = \log(\log n) \log p/n$, except for that in the initial estimator. The candidate values for λ include $N/4$ equally-spaced grid points between $N/150$ and $N/15$, where $N = \sum_{i=1}^n m_i$ is the total number of longitudinal observations. We adapt Belloni and

Chernozhukov (2011)'s method over $\quad$ to get the estimator $\breve{\beta}(\tau)$ for calculating the adaptive weight functions.

We compare AW_2 and AW_3 to the locally concerned adaptive Lasso estimator at a single predetermined quantile level $\tau = 0.2, 0.5$, or 0.8, denoted by $SS(\tau)$, as well as the pointwise approach, which simply combines the estimates from $SS(\tau)$ over $\tau \in \quad$, denoted by PS. We also consider four other benchmark estimation procedures, ALasso (i), ALasso (ii), ALasso (iii), and ALasso (iv). The ALasso (i) estimators are the unconstrained estimator obtained by minimizing (3.4) with the reference, the lth component, randomly chosen. The ALasso (ii) estimators are the globally adaptive estimators derived from model (2.2) without considering the zero-sum constraint. That is, the ALasso (ii) estimators, $\left(\widehat{\alpha}(\tau)^{(ii)}, \widehat{\beta}(\tau)^{(ii)}\right)$, are obtained as

$$\text{argmin}_{\alpha, \beta} \left\{ \frac{1}{n} \sum_{i=1}^{n} \int_0^\infty \rho_\tau \left\{ Y_i(t) - \mathbf{X}_i(t)^\top \boldsymbol{\alpha} - \mathbf{Z}_i(t)^\top \boldsymbol{\beta} \right\} dN_i(t) + \lambda \sum_{j=1}^{p} \omega_j(\tau) |\beta_j| \right\}.$$

ALasso (iii) estimator is obtained from fitting the log-contrast model based on the relevant variables selected by ALasso (ii) approach. ALasso (iv) estimator is obtained by solving the minimization problem (2.3) without including the zero-sum constraint, using the selected relevant variables to fit a log-contrast model, and then selecting the tuning parameter based on the GIC criterion and determining the final estimator.

We assess the variable selection performance of the different methods described above in terms of mean number of correctly identified relevant variables (NCN), mean number of incorrectly selected variables (NIN), percentage of under-fitted models (PUUF), percentage of correctly fitted models (PCF), and percentage of over-fitted models (POF). To evaluate the global estimation accuracy over $\tau \in \quad$, we consider three different types of average estimation errors, AEE_{ℓ_1}, AEE_{ℓ_2} and AEE_{ℓ_∞}, where

$$AEE_{\ell_q} \doteq \frac{1}{|\Delta|} \int_\Delta \| \widehat{\beta}(\tau) - \beta^*(\tau) \|_q d\tau.$$

For $SS(\tau)$, we calculate the average estimation errors by extrapolating the coefficient estimate as the constant value of the whole coefficient function over $\tau \in \quad$. To assess how well the estimated coefficients satisfy the zero-sum constraint, we adopt the criterion, SUM, which is defined as $SUM = \sum_{j=1}^{p} \beta_j(\tau_*)$, where $\beta_j(\cdot)$ stands for the estimated coefficient function and $\tau_* = \text{argmax}_{\tau \in \Delta} |\sum_{j=1}^{p} \beta_j(\tau)|$. Better performance would be indicated by NCN closer to 7, the true number of relevant covariates, PCF closer to 100%, NIN, PUF and POF closer to 0, smaller AEE_{ℓ_1}, AEE_{ℓ_2} and AEE_{ℓ_∞}, and SUM closer to or equal to 0.

The simulation results for setups (I)–(IV) are presented in Table S1, Table S2, Table S3, and Table 1 respectively, where Tables S1–S3 are provided in the Supplementary Materials. Simulation results are summarized based on 300 replicates. As seen from these tables, the proposed estimators with the globally adaptive weights, AW_2 and AW_3, perform well in all setups where the error terms can be homogeneous or heterogeneous, and can be independent or dependent across different time points. In all setups, the PCFs based on

these estimators are around or above 85% and the zero-sum constraint is always met by the estimated coefficient functions. As shown by additional simulations reported in the Supplementary Materials (see Table S4–S5), the PCFs can further increase as the variance of the longitudinal error deceases. In setups (I) and (II) where the effects of $\mathbf{Z}(t)$ are constant over τ, the estimation accuracy is comparable between the proposed globally adaptive estimators and the locally concerned estimators, SS(τ), τ = 0.2, 0.5, 0.8. However, variable selection based on SS(τ) is more likely to miss relevant variables, as reflected by the higher PUFs, particularly when τ = 0.2 or 0.8. In setups (III) and (IV) where the effects of $\mathbf{Z}(t)$ are not constant over τ, SS(τ) has much worse performance of variable selection than AW$_2$ and AW$_3$. This may lead to the deterioration in the average estimation errors for SS(τ) observed in setup (III) and (IV). In all setups, PS method produces average estimation errors similar to those of AW$_2$ and AW$_3$. However, PS method tends to overfit with POF equal to 31.7% in setup (I), 26.3% in setup (II), and 23% in setup (III) and setup (IV).

When examining the results from the globally adaptive estimators under ALasso (i), we note a common overfitting problem associated with adopting the unconstrained log-contrast model. This is because ALasso (i) procedure automatically includes the reference compositional covariate, which may not be a truly relevant covariate. The results under ALasso (ii) suggest that the underlying zero-sum constraint of coefficients would not be satisfied if it is not carefully accounted for in the estimation procedure. In such a situation, interpreting the resulting coefficient estimates as the effects of compositional covariates would be problematic. ALasso (iii) approach renders satisfactory rates of correct fitting but yields larger estimation errors compared to the proposed method. ALasso (iv) method tends to overfit with the percentages of overfitting above 25%. The enlarged estimation errors or the overfitting behavior reflect the disadvantage of handling the zero-sum constraint separately from model estimation and variable selection. In summary, the simulation results suggest the importance of the proposed globally adaptive estimators as well as their satisfactory empirical performance.

5. A real data example

We applied the proposed method to a longitudinal dataset from the *Feeding Infants Right… from the STart* (FIRST) study. The FIRST study is an ongoing perspective observational study, which has enrolled and followed up children with cystic fibrosis (CF) from the neonatal period. In this study, various diet-related biomarkers have been collected repeatedly at pre-specified CF care visits. For example, fecal specimens were collected approximately at 2, 4, 6, 8, and 12 months of age for each child. The gut microbiome composition data were extract from fecal specimen by 16S rRNA gene pyrosequencing and comprise of relative abundances of 364 unique genera subject to the unit-sum constraint. The levels of calprotectin, a biomarker for the inflammation in the gastrointestinal (GI) tract, were also tracked over time and recorded in the unit of microgram per gram of stool. In our analysis of the FIRST dataset, the specific question of interest is how the gut microbiome composition is associated with the calprotectin level over time. Identifying the subcompositional bacterial taxa that are linked to the variations in calprotectin can shed useful insight about the early CF disease mechnisam.

The final dataset includes 135 subjects and a total of 328 longitudinal records after the exclusion of 7 children with low birth weight. Table S6 in the Supplementary Materials presents basic summary statistics for gender, number of longitudinal records, and calprotectin levels. It is shown that 56% of subjects are boys, and about 50% of subjects have 3 or 4 longitudinal records. It is also noted that the calprotectin levels present a skewed distribution with median (= 64.5) considerably smaller than mean (= 111.2). In this case, adopting longitudinal quantile regression modeling can deliver a more comprehensive and robust view about how the gut microbiome composition influences calprotectin levels.

In our analysis, we implement the proposed globally adaptive methods with the adaptive weights ($w2$) and ($w3$) and   = (0.2, 0.8] (denoted by AW_2 and AW_3 respectively), the locally concerned adapive-Lasso method $SS(\tau)$ with $\tau = 0.2, 0.3, \ldots, 0.8$, and the pointwise method (denoted by PS), which is a union set for $SS(\tau)$ with $\tau = 0.2, 0.225, \ldots, 0.8$. We include gender as a regular covariate. The compositional covariates are the relative abundances of 364 genera measured from the gut microbiome samples. We exclude six genera that have relative abundance below the detection limit in all samples. In addition, we replace all non-detectable relative abundance by an extremely small constant 10^{-20}, which is much smaller than the minimum nonzero relative abundance captured in our dataset, 4.418×10^{-6}. For the tuning parameter selection, the candidate values of λ include $N/4$ equally-spaced grid points between $N/150$ and $N/15$, where $N = 328$. To avoid selecting boundary λ's, ϕ_n in GIC criteria is chosen as $\log(\log n) \log p/(20n)$ for globally concerned quantile regression and locally concerned quantile regression. The estimates below 10^{-4} are shrunk to zero.

To evaluate each method, we compute prediction errors as follows. We first randomly split the 135 subjects into a training set of size 120 and a testing set of size 15. We apply the method to the training data set and obtain the estimator of $(\boldsymbol{\alpha}_0(\tau)^\top, \boldsymbol{\beta}_0(\tau)^\top)^\top$, denoted by $\left(\widehat{\boldsymbol{\alpha}}^{\text{train}}(\tau)^\top, \widehat{\boldsymbol{\beta}}^{\text{train}}(\tau)^\top\right)^\top$. Then we calculate the prediction error in the testing set as

$$\text{PE}(\varDelta) = \frac{\sum_{i \in \mathscr{T}} \int_{\varDelta} \int_0^\infty \rho_\tau \left\{ Y_i(t) - \mathbf{X}_i(t)^\top \widehat{\boldsymbol{\alpha}}^{\text{train}}(\tau) - \mathbf{Z}_i(t)^\top \widehat{\boldsymbol{\beta}}^{\text{train}}(\tau) \right\} dN_i(t) d\tau}{\sum_{i=1}^n 1\{i \in \mathscr{T}\}},$$

where $\mathscr{T}$ denotes the test set. For $SS(\tau)$, we calculate PE( ) by treating the coefficient estimate as a constant valued function over $\tau \in$.

Table 2 lists the genus sets selected by different methods. The averages prediction errors (PE) along with the corresponding standard deviations (within parentheses) are also presented. The calculations of PE are based on 200 random splitting of training and test sets. From Table 2, we observe that the selected genus sets vary considerably across the locally concerned methods, $SS(\tau)$ with different choices of τ. These observations suggest that some genera may have varying influences on different quantiles of calprotectin level, and in part, may also reflect the variable selection instability associated with $SS(\tau)$ (Zheng et al., 2015). For example, the genus "g115" may only affect median calprotectin but not the other lower or upper quantiles of calprotectin. In contrast, the proposed globally concerned methods give

robust and yet parsimonious selection of genus sets. For example, the selected genus sets are almost identical between AW_2 and AW_3. The selected genera are mostly also selected by one of $SS(\tau)$'s. Naively pooling the results from $SS(\tau)$'s, as shown by PS method, lead to select an excessive number of genera (i.e. 26 genera). Some genera selected by $SS(\tau)$ but not by AW_2 or AW_3 are possibly "false positive" as suggested by the apparent overfitting behavior of PS method demonstrated in the simulation studies. Moreover, it is noted that the proposed method AW_3 yields the smallest prediction error. The prediction error of AW_2 is close to the second smallest one. The locally concerned $SS(\tau)$ methods produce larger prediction errors because they would neglect important genera that do not show effects at the τ-th quantile but are relevant to other quantiles. In summary, the proposed globally adaptive methods strikes the best balance between parsimonious variable selections and accurate predictions, while retaining sensible interpretations via the satisfaction of the zero-sum constraint of coefficients.

6. Discussions

In this work, we develop a globally concerned longitudinal quantile regression framework which accommodates high-dimensional compositional covariates. The proposed method can achieve the oracle convergence rate as well as the global model selection consistency, while enjoying interpretative advantages.

The longitudinal quantile regression model adopted in this work assumes all covariate effects do not change over time. To accommodate temporal covariate effects, model (2.1) or (2.2) can be extended with regression coefficients formulated as bivariate functions of τ and t. An intuitive approach to tackle this extension is to combine the proposed method with the strategy of Park and He (2017). That is, the longitudinal loss function may be modified by incorporating spline approximations to the regression coefficient functions with the penalty term adjusted accordingly. Nevertheless, this approach may be computationally prohibitive provided the additional high-dimensional layer induced by spline approximations. More specifically, suppose there are L spline basis functions, and $L = O(n^{1/5})$. Based on the proposed estimation for model (2.1), the computational complexity is about $O(n^2 \cdot p \cdot M(n))$, according to Klee and Minty (1972)'s result for simplex algorithm. When considering the spline based estimation for the extended model with time-varying coefficient coefficients, we expect that the computational intensity would be roughly equivalent to that of fitting a quantile regression model for a dataset with sample size $nM(n)$ and covariate dimension pL, which is about $O(n^2 M(n)^2 pL)$. Given $M(n) = O(n)$, as suggested by Zheng et al. (2015), tackling the more flexible model with time-varying coefficients would require $O(n^{6/5})$ times the computational effort needed for the proposed model (2.1), which can be computationally prohibitive for high-dimensional applications. How to address such an obstacle merits future research.

After applying the proposed method to a real dataset, assessing the adequacy of model (2.1) with the pre-specified quantile index set and the selected relevant variables may be of practical interest. To this end, we can adapt the model checking strategy of Peng and Huang (2008), and consider the stochastic process

$$K_n(\tau) = n^{-1/2} \sum_{i=1}^{n} \int_0^\infty W(\mathbf{V}_i(t)) \psi_\tau \left\{ Y_i(t) - \mathbf{X}_i(t)^\top \hat{\boldsymbol{\alpha}}(\tau) - \mathbf{Z}_i(t)^\top \hat{\boldsymbol{\beta}}(\tau) \right\} dN_i(t),$$

as an analogue of the martingale based diagnostic process employed by Peng and Huang (2008), where $\psi_\tau(u) = \tau - I(u < 0)$. Here $W(\cdot)$ is a known bounded function and $\mathbf{V}_i(t) = (\mathbf{X}_i(t)^\top, \mathbf{Z}_i(t)^\top)^\top$. A lack-of-fit test statistic may be constructed based on $\sup_{\tau \in} |K_n(\tau)|$. Following the lines of Peng and Huang (2008), the corresponding p value can be obtained by using a properly designed resampling scheme to approximate the distribution of $K_n(\cdot)$ under model assumption (2.1).

Following the idea of weighted GEE (Liang and Zeger, 1986) and the quasi-likelihood approach for median regression (Jung, 1996), we may incorporate within-subject correlation of repeated measures to further improve estimation efficiency of the proposed method. Specifically, we may consider a weighted penalized estimating equation,

$$n^{-1/2} \sum_{i=1}^{n} \mathbf{V}_i^\top \mathbf{Q}_i(\tau; \boldsymbol{\alpha}, \boldsymbol{\beta})^{-1} S_i(\tau; \boldsymbol{\alpha}, \boldsymbol{\beta}) + \lambda \sum_{j=1}^{p} \omega_j(\tau) \text{sign}(\beta_j) = 0,$$

subject to constraint $\sum_{j=1}^{p} \beta_j = 0$, where $\mathbf{V}_i = \left(\mathbf{V}_i(t_i^{(1)}) \ldots, \mathbf{V}_i(t_i^{(m_i)}) \right)^\top$, $S_i(\tau; \boldsymbol{\alpha}, \boldsymbol{\beta}) = (S_{i1}(\tau; \boldsymbol{\alpha}, \boldsymbol{\beta}), \ldots, S_{i,m_i}(\tau; \boldsymbol{\alpha}, \boldsymbol{\beta}))^\top$ with $S_{ik}(\tau; \boldsymbol{\alpha}, \boldsymbol{\beta}) = I\left(Y_i(t_i^{(k)}) - \mathbf{X}_i(t_i^{(k)})^\top \boldsymbol{\alpha} - \mathbf{Z}_i(t_i^{(k)})^\top \boldsymbol{\beta} \leq 0 \right) - \tau$, and $\mathbf{Q}_i(\tau, \boldsymbol{\alpha}, \boldsymbol{\beta})$ is a working covariance matrix that approximates the covariance of $\mathbf{S}_i(\tau, \boldsymbol{\alpha}, \boldsymbol{\beta})$. When $\mathbf{Q}_i(\tau, \boldsymbol{\alpha}, \boldsymbol{\beta})$ is an identity matrix $\mathbf{I}_{m_i}$, solving this estimating equation is equivalent to minimizing (2.3) which adopts the working independence assumption. However, we note that the weighted estimating equation loses the nice monotonicity property possessed by the unweighted version. In addition, the covariance of $\mathbf{Q}_i(\tau, \boldsymbol{\alpha}, \boldsymbol{\beta})$ is often unknown in practice and its empirical estimate may not be stable when sample size is not large, like in the FIRST dataset. As suggested by a referee, one possible solution to alleviate the computational issue is to adopt an iterative algorithm where one first solves the weighted estimating equation with the parameters $\boldsymbol{\alpha}$ and $\boldsymbol{\beta}$ in the weight function $\mathbf{Q}_i(\tau, \boldsymbol{\alpha}, \boldsymbol{\beta})^{-1}$ fixed and then updates the weight function with the resulting parameter estimates. In this case, the estimating equation involved in each iteration is still monotone. Applying this strategy may lead to an approach that improves estimation efficiency while being computationally viable. It is of future research interests to investigate this weighted method in more details.

Supplementary Material

Refer to Web version on PubMed Central for supplementary material.

Acknowledgements

References

Aitchison J (1982). The statistical analysis of compositional data. Journal of the Royal Statistical Society. Series B (Methodological) 44, 139–177.

Aitchison J (2003). The Statistical Analysis of Compositional Data. Caldwell, New Jersey: Blackburn Press.

Aitchison J and Bacon-shone J (1984). Log contrast models for experiments with mixtures. Biometrika 71, 323–330.

Belloni A and Chernozhukov V (2011). ℓ_1-penalized quantile regression in high-dimensional sparse models. The Annals of Statistics 39, 82–130.

Cho H, Hong HG, and Kim M-O (2016). Efficient quantile marginal regression for longitudinal data with dropouts. Biostatistics 17(3), 561–575. [PubMed: 26951723]

Fan J, Fan Y, and Barut E (2014). Adaptive robust variable selection. The Annals of Statistics 42, 324–351. [PubMed: 25580039]

Fan J and Lv J (2011). Non-concave penalized likelihood with np-dimensionality. IEEE Transactions on Information Theory 57, 5467–5484. [PubMed: 22287795]

Fan Y and Tang CY (2013). Tuning parameter selection in high dimensional penalized likelihood. Journal of the Royal Statistical Society: Series B (Statistical Methodology) 75, 531–552.

Gao X and Liu Q (2019). Sparsity identification in ultra-high dimensional quantile regression models with longitudinal data. Communications in Statistics-Theory and Methods, 1–25.

He X and Shao Q-M (2000). On parameters of increasing dimensions. Journal of Multivariate Analysis 73(1), 120–135.

Jung S-H (1996). Quasi-likelihood for median regression models. Journal of the American Statistical Association 91, 251–257.

Kim Y, Choi H, and Oh H-S (2008). Smoothly clipped absolute deviation on high dimensions. Journal of the American Statistical Association 103, 1665–1673.

Klee V and Minty G (1972). How good is the simplex algorithm, inequalities-iii, edited by 0. shisha.

Koenker R (2004). Quantile regression for longitudinal data. Journal of Multivariate Analysis 91(1), 74–89.

Koenker R and Bassett G (1978). Regression quantiles. Econometrica 46, 33–50.

Koenker R and d'Orey V (1987). Algorithm as 229: Computing regression quantiles. Journal of the Royal Statistical Society, Series C (Applied Statistics) 36, 383–393.

Ledoux M and Talagrand M (1991). Probability in Banach Space: Isoperimetry and Processes. Ergebnisse der Mathematik und ihrer Grenzgebiete, Volume 23. Springer, Berlin.

Li Y, Liu Y, and Zhu J (2007). Quantile regression in reproducing kernel hilbert spaces. Journal of American Statistical Association 102, 255–268.

Liang K-Y and Zeger SL (1986). Longitudinal data analysis using generalized linear models. Biometrika 73(1), 13–22.

Lin W, Shi P, Feng R, and Li H (2014). Variable selection in regression with compositional covariates. Biometrika 101, 785–797.

Lipsitz SR, Fitzmaurice GM, Molenberghs G, and Zhao LP (1997). Quantile regression methods for longitudinal data with drop-outs: application to cd4 cell counts of patients infected with the human immunodeficiency virus. Journal of the Royal Statistical Society: Series C (Applied Statistics) 46(4), 463–476.

Lu J, Shi P, and Li H (2019). Generalized linear models with linear constraints for microbiome compositional data. Biometrics 75, 235–244. [PubMed: 30039859]

Lv J and Fan Y (2009). A unified approach to model selection and sparse recovery using regularized least squares. The Annals of Statistics 37, 3498–3528.

Ma H, Peng L, and Fu H. Quantile regression modeling of latent trajectory features with longitudinal data. Journal of Applied Statistics 46, 2884–2904.

Meinshausen N and Buhlmann P (2006). High-dimensional graphs and variable selection with the lasso. The Annals of Statistics 34, 1436–1462.

Nishii R (1984). Asymptotic properties of criteria for selection of variables in multiple regression. The Annals of Statistics 12, 758–765.

Park S and He X (2017). Hypothesis testing for regional quantiles. Journal of Statistical Planning and Inference 191, 13–24.

Peng L and Huang Y (2008). Survival analysis with quantile regression models. Journal of the American Statistical Association 103(482), 637–649.

Portnoy S (1991). Asymptotic behavior of the number of regression quantile breakpoints. SIAM Journal on Scientific and Statistical Computing 12, 867–883.

Shi P, Zhang A, and Li H (2016). Regression analysis for microbiome compositional data. The Annals of Applied Statistics 10, 1019–1040.

Sun X, Peng L, Manatunga A, and Marcus M (2016). Quantile regression analysis of censored longitudinal data with irregular outcome-dependent follow-up. Biometrics 72(1), 64–73. [PubMed: 26237289]

Talagrand M (2005). The Generic Chaining. Springer, Berlin.

Tyler AD, Smith MI, and Silverberg MS (2014). Analyzing the human microbiome: a "how to" guide for physicians. The American Journal of Gastroenterology 109, 983–993. [PubMed: 24751579]

van der Vaart A and Wellner J (1996). Weak Convergence and Empirical Processes: With Applications to Statistics. New York: Springer.

Wang H and Fygenson M (2009). Inference for censored quantile regression models in longitudinal studies. The Annals of Statistics 37, 756–781.

Wang L, Wu Y, and Li R (2012). Quantile regression for analyzing heterogeneity in ultrahigh dimension. Journal of the American Statistical Association 101, 1418–1429.

Wang L, Zhou J, and Qu A (2012). Penalized generalized estimating equations for high-dimensional longitudinal data analysis. Biometrics 68(2), 353–360. [PubMed: 21955051]

Welsh A (1989). On m-processes and m-estimation. The Annals of Statistics 17(1), 337–361.

Xia F, Chen J, Fung W, and Li H (2013). A logistic normal multinomial regression model for microbiome compositional data analysis. Biometrics 69, 1053–1063. [PubMed: 24128059]

Zhang C-H and Huang J (2008). The sparsity and bias of the Lasso selection in high-dimensional linear regression. The Annals of Statistics 36(4), 1567–1594.

Zheng Q, Gallagher C, and Kulasekera KB (2013). Adaptive penalized quantile regression for high dimensional data. Journal of Statistical Planning and Inference 143, 1029–1038.

Zheng Q, Peng L, and He X (2015). Globally adaptive quantile regression with ultra-high dimensional data. The Annals of Statistics 43, 2225–2258. [PubMed: 26604424]

Zheng X, Fu B, Zhang J, and Qin G (2018). Variable selection for longitudinal data with high-dimensional covariates and dropouts. Journal of Statistical Computation and Simulation 88(4), 712–725.

Zou H and Yuan M (2008). Composite quantile regression and the oracle model selection theory. The Annals of Statistics 36, 1108–1126.

Table 1:

Simulation results under Setup (IV) with dependent heterogeneous errors

	AEE_{ℓ_1}	AEE_{ℓ_2}	AEE_{ℓ_∞}	NCN	NIN	PUF (%)	PCF (%)	POF (%)	SUM
				Proposed					
AW_2	2.261	1.024	0.694	6.923	0.040	7.7	88.3	4.0	0.000
AW_3	2.297	1.041	0.707	6.877	0.017	12.0	86.7	1.3	0.000
SS(0.2)	2.828	1.275	0.841	6.110	0.007	57.3	42.0	0.7	0.000
SS(0.5)	1.908	0.863	0.577	6.670	0.017	29.0	69.3	1.7	0.000
SS(0.8)	2.623	1.206	0.829	6.273	0.023	43.3	54.7	2.0	0.000
PS	2.277	1.034	0.702	6.970	0.257	3.0	74.0	23.0	0.000
				ALasso (i)					
AW_2	2.456	1.067	0.706	6.917	1.030	8.0	0.7	91.3	0.000
AW_3	2.491	1.084	0.718	6.860	1.010	13.3	0.7	86.0	0.000
				ALasso (ii)					
AW_2	2.326	1.064	0.725	6.913	0.033	8.7	88.3	3.0	1.916
AW_3	2.352	1.076	0.734	6.857	0.017	14.0	84.3	1.7	−2.315
				ALasso (iii)					
AW_2	2.549	1.146	0.770	6.913	0.033	8.7	88.3	3.0	0.000
AW_3	2.668	1.199	0.811	6.857	0.017	14.0	84.3	1.7	0.000
				ALasso (iv)					
AW_2	2.294	1.030	0.685	6.963	0.380	3.7	63.7	32.7	0.000
AW_3	2.305	1.035	0.690	6.957	0.347	4.3	65.0	30.7	0.000

Table 2:

Analysis of the FIRST dataset

τ	Method	Selected Genus Sets	PE
[0.2, 0.8]	AW$_2$	g50 g93 g115 g137 g147 g152 g162 g178 g184 g197 g204 g210 g213 g219 g297 g319 g370	0.5279 (0.0837)
	AW$_3$	g50 g93 g115 g147 g152 g162 g178 g197 g204 g210 g213 g219 g297 g319 g370	0.5271 (0.0837)
	PS	g14 g32 g50 g64 g93 g115 g119 g137 g147 g152 g153 g162 g178 g183 g184 g188 g193 g197 g199 g204 g210 g213 g219 g297 g319 g370	0.5278 (0.0826)
0.2	SS	g147 g153 g213	0.7893 (0.1420)
0.3	SS	None	0.6833 (0.1206)
0.4	SS	g50 g93 g119 g147 g162 g183 g197 g199 g204 g213 g297	0.6135 (0.1038)
0.5	SS	g14 g115 g137 g147 g193 g197 g204 g213 g219 g297 g319	0.5855 (0.0976)
0.6	SS	None	0.5960 (0.0981)
0.7	SS	g147 g152 g178 g184 g197 g204 g213 g297 g319 g370	0.6621 (0.1025)
0.8	SS	g32 g147 g152 g162 g178 g197 g204 g213	0.7926 (0.1235)