Author Manuscript

Author Manuscript

Author Manuscript

Author Manuscript

HHS Public Access

Author manuscript

Biometrics. Author manuscript; available in PMC 2018 October 06.

Published in final edited form as:

Biometrics. 2018 June ; 74(2): 448–457. doi:10.1111/biom.12775.

A GLM-Based Latent Variable Ordination Method for Microbiome Samples

Michael B. Sohn and **Hongzhe Li***

Department of Biostatistics and Epidemiology, University of Pennsylvania Perelman School of Medicine, Philadelphia, PA 19104, U.S.A

SUMMARY:

Distance-based ordination methods, such as principal coordinates analysis (PCoA), are widely used in the analysis of microbiome data. However, these methods are prone to pose a potential risk of misinterpretation about the compositional difference in samples across different populations if there is a difference in dispersion effects. Accounting for high sparsity and overdispersion of microbiome data, we propose a GLM-based Ordination Method for Microbiome Samples (GOMMS) in this paper. This method uses a zero-inflated quasi-Poisson (ZIQP) latent factor model. An EM algorithm based on the quasi-likelihood is developed to estimate parameters. It performs comparatively to the distance-based approach when dispersion effects are negligible and consistently better when dispersion effects are strong, where the distance-based approach sometimes yields undesirable results. The estimated latent factors from GOMMS can be used to associate the microbiome community with covariates or outcomes using the standard multivariate tests, which can be investigated in future confirmatory experiments. We illustrate the method in simulations and an analysis of microbiome samples from nasopharynx and oropharynx.

Keywords

16S sequencing; Factor models; Microbiome; Zero-inflated models

1. Introduction

In microbiome studies, it is often of interest to visually inspect compositional differences among microbiome samples before conducting stringent statistical tests on characteristics of the samples. A common approach is to use a distance-based ordination method, such as the principal coordinate analysis (PCoA), which can map high-dimensional data onto low-dimensional displays (Legendre and Legendre, 1998). However, the distance-based ordination may be inappropriate for microbiome samples when there are large differences in dispersion between samples because it doesn't take the effect of dispersion into consideration. As a result, the distance-based ordination can pose a potential risk of misinterpretation about a compositional difference between samples.

* hongzhe@upenn.edu.

6. Supplementary Materials

R codes to implement the methods and the real data sets are available at https://cran.r-project.org/web/packages/gomms/

Author Manuscript Author Manuscript Author Manuscript Author Manuscript

It has been shown that the distance-based ordination methods are incapable of distinguishing between mean and dispersion effects (Warton et al., 2012). The former reflects the difference in relative abundance and the latter the difference in variability. That is, even though there is no mean effect, a distance-based ordination method can show significantly different centroids of two populations when there is a strong dispersion effect. The opposite is also possible: no difference in the centroids of two populations even if there is a mean effect. This undesirable result might happen in human microbiome studies since dispersion effects between different populations have been observed in numerous studies (Finegold et al., 2010; Zeller et al., 2014; Bäckhed et al., 2015).

A solution to this potential problem is to use a generalized linear model approach, which explicitly models the mean-variance relationship, thus incorporating dispersion effects. A family of Poisson factor models exists that can be used for ordination analysis (Shen and Huang, 2008; Lee et al., 2013). However, they are not suitable for microbiome data in which apparent overdispersion has been evidenced (Chen and Li, 2013; McMurdie and Holmes, 2014). In this paper, we propose a new GLM-based Ordination Method for Microbiome Samples (GOMMS) that uses a zero-inflated quasi-Poisson (ZIQP) factor model. This method accounts for characteristics of microbiome data (e.g., highly skewed non-negative counts with excessive zeros) while reducing dimensionality. In microbiome studies, a zero count is assigned to an absent taxon (e.g., species or gene) in a given sample if the taxon is detected in some other samples. Thus, a zero count can be either a result of true absence or undetected presence, suggesting that a mixture model is appropriate for modeling such zeros. GOMMS can also be used as a tool to associate the microbial communities to response variables or covariates that can be investigated in future confirmatory experiments.

In simulation studies, we demonstrate potential problems of distance-based ordination methods, particularly, PCoA and non-metric multidimensional scaling (NMDS). We show that GOMMS performs comparatively to the distance-based approaches when dispersion effects are negligible and consistently even when a distance-based ordination method exhibits the potential problems due to significant dispersion effects. We also show that the standard multivariate test such as the Hotelling's T^2 test can be applied using the estimated loading coefficients or coordinates for testing the differential community compositions. We perform an exploratory analysis on the upper respiratory tract microbial data (Charlson et al., 2010) that consist of the left and right of nasopharyngeal and oropharyngeal samples from smoking and nonsmoking healthy adults. GOMMS finds no difference in the centroid and variation of the samples of the left and right sides of the upper respiratory tract but significant difference in the samples of nasopharynx and oropharynx. Moreover, we find a significant correlation between a covariate, the elapsed time from the last smoke, and an estimated factor for the oropharyngeal samples from smokers, implying a possible effect of smoking on short-term changes in oropharyngeal microbial communities, which is an interesting hypothesis for a future confirmatory experiment.

2. Methods

In the factor analysis, we aim to find latent variables or factors $\boldsymbol{f} = (f_1,\ldots,f_p)^\top$, which represent unobserved constructs, such that factors can capture variability among correlated

Author Manuscript Author Manuscript Author Manuscript Author Manuscript

observed variables $\boldsymbol{x} = (x_1, \ldots, x_m)^\top$ with a smaller number of variables (i.e., $p < m$), where the superscript $\top$ represents the transpose operator. Each observed variable is modeled as a linear combination of factors:

$$x_j = \beta_{j1} f_1 + \cdots + \beta_{jp} f_p + \epsilon_j, \quad (1)$$

where ϵ_j represents the j^{th} unique component including a random error and $j = 1, 2, \ldots, m$. A graphical representation of this factor model is given in Figure (1).

For an exploratory analysis, a factor model doesn't place any structure on the linear relationships between observed variables nor between observed variables and factors. It specifies only the number of factors, thus a factor model being a proper method for the ordination. In other words, we can use 2 or 3 factors to graphically display similarities of data as we usually do with the principal components of PCoA. With a number p of factors, the factor analysis tries to find p factors that represent the covariance matrix as well as possible (Jolliffe, 2002), which implies that the estimated two factors of a 2-factor model can represent the covariance matrix better than or as well as any pair of factors of a p-factor model can do, where $p > 2$. Note that the focus of GOMMS is to accurately display properties of data in a low dimensional space, not to identify the optimal number of latent factors that explain most of the variation in data, as in typical factor analysis.

2.1 ZIQP Factor Model

Typical microbiome data consist of highly skewed non-negative counts with an excess of zeros. The source of these zero counts can be either true absence of microbes or undetected presence of microbes. To fit these sparse data, we introduce a ZIQP factor model. Let x_{ij} denote a count assigned to taxon j in sample i. Applying a factor approach to the log of rate μ_{ij}, a ZIQP factor model can be given by

$$\begin{cases} X_{ij} \sim \begin{cases} 0 & \text{with probability } \eta_j \\ \mathrm{QP}(\mu_{ij}, \phi) & \text{with probability } 1 - \eta_j \end{cases} \\ \log(\mu_{ij}) = \beta_{i0} + \beta_{i1} f_{j1} + \cdots + \beta_{ip} f_{jp} \end{cases}, \quad (2)$$

where ϕ is an overdispersion parameter of the quasi-Poisson (QP) distribution, β_{i0} serves as a specific or unique factor loading for the i^{th} sample that depends on the sequencing depth, β_{ik} is the k^{th} factor loading for the i^{th} rate profile, and f_{jk} is the k^{th} factor score for the j^{th} feature, where $i = 1, \ldots, n$, $j = 1, \ldots, m$, and $k = 1, \ldots, p$. In this model, the logarithm is the canonical link function for a Poisson model in the generalized linear model (GLM) framework (McCullagh and Nelder, 1989). In this model, the mean and dispersion of the observed count is

$$\mathrm{E}(X_{ij}) = (1 - \eta_j)\mu_{ij}, \ \mathrm{Var}(X_{ij}) = (1 - \eta_j)\phi\mu_{ij},$$

Author Manuscript Author Manuscript Author Manuscript Author Manuscript

where ϕ is used to account for extra-variability. Since the main purpose of this model is for ordination analysis, p is usually chosen as 2 or 3, in which case one can plot the estimated factor loadings (β_{i1}, β_{i2}, β_{i3}) in a 2D or 3D plot.

2.2 Estimators of Parameters

The generalized log quasi-likelihood function is defined as

$$Q\left(\mu \mid x\right) = \int_{x}^{\mu} \frac{x - t}{\phi V(t)} dt, \quad (3)$$

where $\phi V(\cdot)$ is variance. Denote $\boldsymbol{B}_{n\times(p+1)}$ as a factor coefficient or loading matrix, $\boldsymbol{F}_{m\times(p+1)}$ as a factor matrix with a vector of 1's in the first column, and $\boldsymbol{B}_i$ and $\boldsymbol{F}_j$ as the ith row of $\boldsymbol{B}$ and the jth row of $\boldsymbol{F}$, respectively. Then, the log quasi-likelihood function for the ZIQP model (2) can be expressed as

$$\begin{aligned} \ell(\boldsymbol{\theta} \mid \mathrm{x}) = & \sum_{x_{ij}=0} \log\left[\eta_j - \left(1-\eta_j\right)\frac{\exp\left(\boldsymbol{B}_i\boldsymbol{F}_j^{\intercal}\right)}{\phi}\right] \\ & + \sum_{x_{ij}>0}\left[\log\left(1-\eta_j\right) + \int_{x_{ij}}^{\exp\left(\boldsymbol{B}_i\boldsymbol{F}_j^{\intercal}\right)} \frac{x_{ij}-t}{\phi t} dt\right], \end{aligned} \quad (4)$$

where $\boldsymbol{\theta} = (\boldsymbol{\eta}, \boldsymbol{B}, \boldsymbol{F}, \phi)$. The separation of zero and non-zero counts complicates the maximization of $\ell(\boldsymbol{\theta}|\mathbf{x})$. However, assuming we could observe $Z_{ij} = 1$ when X_{ij} comes from an unknown zero state and $Z_{ij} = 0$ when X_{ij} comes from QP, the complete log quasi-likelihood function (4) can be expressed as

$$\begin{aligned} \ell(\boldsymbol{\theta} \mid \mathrm{x}) = & \sum_{i=1}^{n}\sum_{j=1}^{m} z_{ij}\log\left(\eta_j\right) \\ & + \sum_{i=1}^{n}\sum_{j=1}^{m}\left(1-z_{ij}\right)\log\left(1-\eta_j\right) \\ & + \sum_{i=1}^{n}\sum_{j=1}^{m}\left(1-z_{ij}\right)\int_{x_{ij}}^{\exp\left(\boldsymbol{B}_i\boldsymbol{F}_j^{\intercal}\right)} \frac{x_{ij}-t}{\phi t} dt. \end{aligned}$$

The maximum likelihood estimation (MLE) for η_j is then the solution of

$$\nabla_{\eta_j}\ell\left(\boldsymbol{\theta} \mid \mathrm{x}\right) = \sum_{i=1}^{n} \frac{z_{ij} - \eta_j}{\eta_j\left(1-\eta_j\right)} = 0, \quad (5)$$

that is, $\hat{\eta}_j = \Sigma_{i=1}^{n} z_{ij} \Big/ n$, and MLEs for $\boldsymbol{B}$ and $\boldsymbol{F}$ are the solutions of

Author Manuscript Author Manuscript Author Manuscript Author Manuscript

$$\nabla_{\boldsymbol{B}}\ell\left(\boldsymbol{\theta}+\mathbf{X}\right)=\left\{\left(\mathbf{1}-\boldsymbol{Z}\right)\circ\left[\boldsymbol{X}-\exp\left(\boldsymbol{B}\boldsymbol{F}^{\intercal}\right)\right]\right\}\boldsymbol{F}=\mathbf{0}\text{ subject to}\boldsymbol{B}_{0}\geq\mathbf{0},\quad(6)$$

$$\nabla_{\boldsymbol{F}}\ell\left(\boldsymbol{\theta}\mid\mathbf{X}\right)+\left\{\left(\mathbf{1}-\boldsymbol{Z}\right)\circ\left[\boldsymbol{X}-\exp\left(\boldsymbol{B}\boldsymbol{F}^{\intercal}\right)\right]\right\}^{\intercal}\boldsymbol{B}_{-0}=\mathbf{0},\quad(7)$$

where $\boldsymbol{X}$ and $\boldsymbol{Z}$ are the matrices of $\{x_{ij}$ and $\{z_{ij}\}$, respectively, the subscript 0 indicates the first column of a matrix, −0 indicates the exclusion of the first column, and ∘ is the Hadamard product operator. The score functions (6) and (7) are identical to those of Poisson regression and Poisson regression with an offset $\boldsymbol{B}_0$. However, the solution of these parameters is not identifiable since both $\boldsymbol{B}$ and $\boldsymbol{F}$ are unknown. To achieve identifiability, we use as a constraint

$$\boldsymbol{F}_{k}^{\intercal}\boldsymbol{F}_{k'}=\delta_{kk'},\quad(8)$$

where $\delta_{kk'}$ is the Kronecker delta, and adopt an alternating maximum likelihood method. More details about the constraint and the alternating maximum likelihood method can be found in Shen and Huang (2008).

For the estimation of $\boldsymbol{\phi}$, Equation (3) doesn't behave like a log likelihood. The conventional approach is a moment estimator based on the residuals (McCullagh and Nelder, 1989), that is,

$$\hat{\phi}_{i}=\frac{1}{m-p}\sum_{j=1}^{m}\frac{\left(x_{ij}-\hat{\mu}_{ij}\right)^{2}}{\hat{\mu}_{ij}},\quad(9)$$

where p is the number of factors. We use a global overdispersion parameter $\hat{\phi}=\text{median}\left(\hat{\phi}_{i}\right)$. Note that it is more realistic to use an overdispersion parameter for each feature. This will, however, often lead some MLEs for $\boldsymbol{B}$ and $\boldsymbol{F}$ to be divergent during alternating maximum likelihoods in an expectation-maximization (EM) framework.

2.3 EM algorithm for a ZIQP Factor Model

The constraint and estimators (5) - (9) are used in an EM framework to obtain unique estimations of the parameters, $\boldsymbol{\theta}=(\boldsymbol{\eta},\boldsymbol{B},\boldsymbol{F},\boldsymbol{\phi})$.

2.3.1 Initialization Step.—Singular value decomposition (SVD) allows us to express a matrix $\boldsymbol{A}_{n\times m}$ as

$$\boldsymbol{A}=\boldsymbol{U}\boldsymbol{\Sigma}\boldsymbol{V}^{\intercal}\Leftrightarrow\boldsymbol{A}\boldsymbol{V}=\boldsymbol{U}\boldsymbol{\Sigma},\quad(10)$$

Author Manuscript Author Manuscript Author Manuscript Author Manuscript

where $\boldsymbol{U}$ is an $n \times r$ orthogonal matrix, Σ is an $r \times r$ non-negative diagonal matrix, $\boldsymbol{V}$ is an $m \times r$ orthogonal matrix, and $r \leqslant \min(m, n)$. It can be viewed as a transformation of the unit sphere in $\mathbb{R}^n$ into an ellipsoid in $\mathbb{R}^r$ that exposes the substructure of $\boldsymbol{A}$ more clearly, ordering it based on the amount of variation. Therefore, $\boldsymbol{V}$ is a good candidate for $\boldsymbol{F}$ since it satisfies the constraint (8) while SVD allows us to approximate $\boldsymbol{A}$ by selecting the first p values of Σ. The SVD approach has been used in various genomic studies (Holter et al., 2000; Lee et al., 2013).

Since assuming a factor model on the logarithm of rate parameter, we apply SVD to the log of the standardized $\boldsymbol{X}$ after adding 0.5 to avoid singularity to obtain $\boldsymbol{V}_{m\times p}$ and set $\boldsymbol{F}^{(0)} = \boldsymbol{V}$. For $z_{ij}^{(0)}$, the proportion of zeros in taxon j is used when $x_{ij} = 0$. To estimate $\boldsymbol{B}^{(0)}$, we fit n quasi-Poisson regressions with weights $\mathbf{1}_i - \boldsymbol{Z}_i^{(0)}$ using $\boldsymbol{X}_\mathrm{i}$ as a response and $\boldsymbol{F}^{(0)}$ as covariates, where $\mathbf{1}$ is an $m \times n$ matrix with 1's; $\mathbf{1}_i$ and $\boldsymbol{Z}_i$ are the i^{th} rows of $\mathbf{1}$ and $\boldsymbol{Z}$, respectively. With $\mu_{ij}^{(0)} = \exp\left(\boldsymbol{B}_i^{(0)}\boldsymbol{F}_j^{(0)T}\right)$, we estimate $\hat{\phi}_i^{(0)}$ by (9) and then $\hat{\phi}^{(0)} = \mathrm{median}\left(\hat{\phi}_i^{(0)}\right)$.

2.3.2 E Step.—Estimate $z_{ij}^{(k)}$ given current estimates of $\boldsymbol{\theta}$ by

$$= \begin{cases} \dfrac{\hat{\eta}_j^{(\ell-1)}}{\hat{\eta}_j^{(\ell-1)} + \left(1-\hat{\eta}_j^{(\ell-1)}\right)\exp\left(-\hat{\mu}_{ij}^{(\ell-1)} \Big/ \phi^{(\ell-1)}\right)} & \text{if } x_{ij} = 0, \\ 0 & \text{otherwise}. \end{cases}$$

2.3.3 M Step.—Given the current estimate of $\boldsymbol{Z}^{(\ell)}$, the following steps are performed in order:

(1) Maximize $\boldsymbol{B}_i^{(\ell)}$ with weights $(\mathbf{1} - \boldsymbol{Z}^{(\ell)})_i$ by quasi-Poisson regressing $\boldsymbol{X}_i$ onto $\boldsymbol{F}^{(\ell-1)}$ for all i.

(2) Estimate $\boldsymbol{\phi}^{(\ell)}$ with $\boldsymbol{B}_i^{(\ell)}$ and $\boldsymbol{F}^{(\ell-1)}$.

(3) Maximize $\boldsymbol{F}_j^{(\ell)}$ with weights $\left(1 - \boldsymbol{Z}^{(\ell)}\right)_j^\intercal$ by quasi-Poisson regressing $\boldsymbol{X}_j^\intercal$ onto $\boldsymbol{B}^{(\ell)}$ for all j.

(4) Apply SVD to the standardized $\boldsymbol{B}^{(\ell)}\,\mathbf{F}^{(\ell)T}$ to obtain $\boldsymbol{V}^{(\ell)}$ and reset $\boldsymbol{F}^{(\ell)} = \boldsymbol{V}^{(\ell)}$.

Repeat the EM steps until $\delta\boldsymbol{F} = \max(\boldsymbol{F}^{(\ell)} - \boldsymbol{F}^{(\ell-1)}) < \delta_0$, where δ_0 is a small number, such as $\delta_0 = 10^{-6}$.

3. Simulation Results

3.1 Data generated from negative binomial distributions

To illustrate potential problems of the distance-based ordination method and compare the performance of GOMMS with that of PCoA and NMDS, we simulated count data for two populations, each containing n samples from negative binomial (NB) models with the estimated proportion p_j of taxon j from real datasets including the Human Microbiome

Author Manuscript Author Manuscript Author Manuscript Author Manuscript

Project (The Human Microbiome Project Consortium, 2012), where NB(μ, ϕ) represents the probability mass function of a negative binomial with the parameters of mean μ and overdispersion ϕ,

$$P_{NB}\left(X = x \mid \phi\right) = \frac{\Gamma\left(x + \phi^{-1}\right)}{\Gamma\left(\phi^{-1}\right)\Gamma\left(x+1\right)}\left(\frac{1}{1+\phi\mu}\right)^{\phi^{-1}}\left(\frac{\mu}{\phi^{-1}+\mu}\right)^{x}.$$

The dispersion Var(X) = $\mu(1 + \phi\mu)$. We set $\mu_{ij} = N_{0pj}$ for two populations and $\phi_{G2} = 10_{\phi G1}$, where N_0 is an arbitrary constant, and ϕ_{G1} and ϕ_{G2} are global overdispersions for the first and second population, respectively. In other words, there is no mean difference in taxa between the two populations but a strong difference in dispersions of the two populations. We used a global overdispersion parameter for each population to demonstrate potential problems of the distance-based ordination method more clearly. For the following simulations, 80 samples (40 samples per population) and 100 taxa were simulated.

As for comparisons, we used the Bray-Curtis distance (Bray and Curtis, 1957) for PCoA and NMDS as their distance matrix, which is commonly used in microbiome data analysis. Figure (2) (a) shows 2-D ordination results for the three methods: GOMMS, PCoA, and NMDS. GOMMS shows a similar centroid but clear difference in dispersions. NMDS also shows a similar centroid but very different dispersions. However, PCoA shows differences in both centroids and dispersions, which is misleading.

To view the mean difference under strong dispersion effects, we swapped the mean counts for the first 10 taxa with the following 10 taxa in the second population. The results for this case are shown in Figure (2) (b). GOMMS and PCoA display both mean and dispersion differences while NMDS doesn't display a mean effect. As shown in Figure (2), under a strong dispersion effect, PCoA tends to show a mean difference; however, NMDS tends not to display any mean difference whether or not there is a mean difference, clearly showing the potential problems of the distance-based ordination methods.

3.2 Data generated from zero-inflated negative binomial distributions with taxon-specific dispersion parameters

Secondly, we used a Zero-Inflated Negative Binomial (ZINB) model with randomly generated mean counts for taxa, namely

$$x_{ij} = \eta_j \mathrm{I}\left(x_{ij} = 0\right) + \left(1 - \eta_j\right)\mathrm{NB}\left(\mu_{ij}, \phi_j\right)\mathrm{I}\left(x_{ij} \geq 0\right),$$

where η_j is the proportion of taxon j from a zero state and I is the indicator function. For ZINB models, we used taxon-specific overdispersion parameters, that is, a value of overdispersion for each taxon ϕ_j was randomly generated. For the range of the rate parameter, we used the products of randomly selected values of the three parameters: an expected common mean $u_j \sim$ int(1,10) for taxon j, a difference factor $w_j \sim$ int(2, 5) for taxon j between two populations, and a sample scale $s_i \sim$ int(1,10) for sample i to mimic different sequencing depths, where int(a,b) denotes any randomly selected integer between a and b.

Author Manuscript Author Manuscript Author Manuscript Author Manuscript

For the mixture parameter η_j for sample j, we used $\eta_j = \exp(-4.6/\mu_j)$ to make the probability of true zero depending on the means. For the cases of significant dispersion differences between two populations, the overdispersion parameter for each taxon ϕ_j in one population was randomly generated from Unif(0.5,1) and that in the other population from Unif(3,10). For the cases of little or no significant dispersion effect between two populations, a value of overdispersion for each taxon ϕ_j for each population was randomly generated from Unif(0.5,10).

Results of simulations are presented in Figure 3 for four different scenarios with various differences in means and/or dispersions. Similar conclusions can be drawn as the data simulated under the NB models. When there are differences in dispersions between two populations, NMDS fails to display differences in the means (Figures 3 (c) and (d)). On the other hand, PCoA tends to present a false difference in the means (Figure 3 (d)) in the presence of differences in dispersions. GOMMS seems to capture these differences clearly even there are taxon-specific dispersions.

3.3. Power/type 1 error comparisons

For a quantitative comparison of testing the difference in the overall means between two populations, we applied the Hotelling's T^2 test with unequal covariance matrices using the coordinates estimated from GOMMS, PCoA, and NMDS. By generating data as in the previous section using the ZINB models, we evaluated the effects of sample size n, the number of features, the number of differentially abundant features and the sparsity of the count data on the power and type 1 error of the Hotelling's T^2 test. Specifically, to evaluate the effect of sample size, we randomly generated datasets with various numbers of samples (n = 20, 50, 100) and a fixed number of taxa (p = 100). To evaluate the effect of the number of taxa, we used 50 samples with various numbers of taxa (p = 50, 100, 200) but a fixed number of differentially abundant taxa (20%). To evaluate the effect of the number of differentially abundant taxa, with 50 samples and 100 taxa, they were increased by two-fold from 10%. Finally, to evaluate the effect of the sparsity, we used a fixed sample size of 50 but varied the rate parameter as well as the number of taxa. Overall, we considered a total of 400 simulations for each case, including 100 simulations where there is no mean or dispersion differences, 100 simulations where there is a mean difference but no dispersion difference, 100 simulations where there is a dispersion difference but no mean difference, and 100 simulations where there are both mean and dispersion differences. The empirical power and type 1 error for α = 0.01 are calculated based on these simulations. Results for α level of 0.05 and 0.001 are similar and are omitted.

Table 1 summarizes the simulation results. Where there is no difference in dispersions of two populations, overall, we observe that the tests from all the three ordination methods have essentially the same power with the type 1 errors well controlled around α = 0.01. However, when there is a difference in dispersions between two populations, tests based on PCoA clearly yield extremely inflated type 1 errors. In contrast, tests based on GOMMS and NMDS control the type 1 errors within the specified 0.01 level although tests based on NMDS are slightly conservative. However, tests based on GOMMS have better power for detecting differences in the means between two populations.

Author Manuscript Author Manuscript Author Manuscript Author Manuscript

4. Exploratory Analysis of Real Data: Microbial Communities in the Upper Respiratory Tract

The original data sets consist of 16S rRNA gene sequences of 291 swap samples from the left and right sides of nasopharynx and oropharynx of each 29 smoking and 33 nonsmoking healthy adults (Charlson et al., 2010). We used 269 samples after removing duplicate samples from the same adult. The sequences were analyzed for a taxonomic assignment using the Qiime pipeline (Caporaso et al., 2010) with a default parameter setting. We first analyzed the samples of nonsmoking healthy adults to see a difference in microbial communities between different sites of the upper respiratory tract. With the taxonomic assignments at the genus level, we performed ordination with GOMMS. As shown in Figure (4) (left panel), there are little or no mean or dispersion difference in the samples of the left and right sides of the upper respiratory tract (p=0.98 for nasopharynx, and p = 0.98 for oropharynx, Hotelling's T^2 tests). However, there are mean and dispersion differences in the samples of nasopharynx and oropharynx (p=0.00, Hotelling's T^2 test, and p = 0.00, Box's M test). The samples of oropharynx seem to have a higher dispersion than those of nasopharynx.

Since there is no difference in the samples of the left and right sides, the samples for the two sides are treated as independent replicates and used to see the impact of smoking on airway bacterial communities. The ordination result with the taxonomic assignments at the genus level is shown in the right panel of Figure (4). We observed some effect of smoking on nasopharyngeal bacterial communities but not on oropharyngeal bacterial communities (p=0.02 and 0.24, respectively, Hotelling's T^2 tests). However, among the smokers, a wider variation in oropharyngeal samples is observed as compared to the nasopharyngeal samples (p=0.062, Box's M test).

To explain the variation in oropharyngeal bacterial communities among the smokers, we calculated the correlations between the three factors obtained by GOMMS and two quantitative variables provided with the upper respiratory tract dataset, including elapsed-time-since-last-meal and elapsed-time-since-last-smoke. We observe a significant correlation between elapsed-time-since-last-smoke and one of the factors (Kendall's rank correlation coefficient $\tau = -0.241$ and its p-value = 0.009, Figure (5)), which may imply potential short-term changes in oropharyngeal bacterial communities caused by smoking.

As for comparisons, results of PCoA and NMDS are also shown in Figure (4) (middle and bottom panels), both also indicating mean and dispersion differences in the samples of nasopharynx and oropharynx and higher dispersion in oropharyngeal samples than those of nasopharynx. As expected, no significant difference was observed between left and right sites using either of the loading scores. However, Hotelling's T^2 test with PCoA loading scores indicate a significant mean difference between smokers and non-smokers in oropharyngeal microbiomes (p=0.004), but not in the nasopharyngeal bacterial communities (p=0.22).

Author Manuscript Author Manuscript Author Manuscript Author Manuscript

5. Discussion

It is well known that microbiome data are overdispersed; the degree of dispersion can vary across different populations, such as diseased and healthy populations. Therefore, it is logical to extract the information about dispersion directly from data. However, the distance-based approach, which makes an implicit assumption about dispersion, is used for ordination in the vast majority of published articles in the microbiome study. This approach provides reliable results when two conditions are satisfied: 1) a distance metric used is appropriate for given data and 2) dispersion effects are small across different populations. However, if either of the two conditions is not satisfied, it might provide undesirable results as depicted in the Warton *et al.* (2012) article and this article, where the distance-based methods can confound location and dispersion effects.

The proposed model includes only one global overdispersion parameter in order to achieve more computationally stable and fast results. However, as shown in Section 3.2, this simple model can indeed capture the main features of the data in term of overall means and dispersions even when the data are generated with taxon-specific overdispersions. Similarly, for the purpose of ordination analysis, we suggest fitting the model with a very small number of latent factors. We observed in our simulations that even the true models include a large number of latent factors, fitting the proposed model with two to three factors can reveal the differences in means or dispersions between two populations.

GOMMS resolves the problems of the distance-based approach and also provides a way to link characteristics of a microbial community with covariates of interest through estimated latent factors. However, it has a disadvantage in computation time. The distance-based approach, particularly PCoA, is extremely fast and irrelevant to the problem of divergence. However, since GOMMS uses an alternating regression in the EM framework to estimate parameters, it is slower compared to the distance-based approach and parameter estimates. However, for practical sample sizes and number of taxa in microbiome studies, the average runtime is in several minutes. For example, for a sample of 50, the average runtime of GOMMS on a typical desktop PC is 22.46, 40.37, 66.35 seconds for 50, 100 and 200 taxa, compared to 0.12, 0.12 and 0.12 seconds for PCoA and 9.30, 14.96, 18.33 seconds for NMDS, respectively. Finally, in practice, to minimize the issue of convergence of the EM algorithm, GOMMS includes only the taxa whose numbers of nonzero counts are greater than the larger of two numbers: the number of factors and 5% of the sample size. When there are more than p distinct nonzero values for each taxon, non-convergence rarely occurs. The R package for GOMMS provides the result of convergence.

While the methods are developed for count data from 16S rRNA sequencing in microbiome studies, the ideas can be extended to the situations where only the relative abundances or the composition of the taxa are available, in which case a zero-inflated Beta distribution with latent factors can be considered. For the shotgun metagenomic data, one can align the sequencing reads to clade-specific marker genes (Segata et al., 2012) or a set of universal marker genes (Sunagawa et al., 2013) and obtain a set of read counts over these marker genes. The proposed quasi-Poisson model can be modified to include multivariate counts. Finally, it is also interesting to extend the method to incorporate covariates.

Author Manuscript Author Manuscript Author Manuscript Author Manuscript

R codes to implement the methods and the real data sets are available at https://cran.r-project.org/web/packages/gomms/

Acknowledgments

This research was supported by NIH grants CA127334 and GM097505. We thank the AE and two reviewers for their very detailed and helpful comments.

References

Bäckhed F, et al. (2015). Dynamics and Stabilization of the Human Gut Microbiome during the First Year of Life. Cell Host and Microbe, 17, 690–703. [PubMed: 25974306]

Bray JR and Curtis JT (1957) An ordination of the upland forest communities of southern Wisconsin. Ecological Monographs, 27, 325–349.

Caporaso J, et al. (2010). Qiime allows analysis of high-throughput community sequencing data. Nat. Methods, 7(5), 335–336. [PubMed: 20383131]

Charlson E, et al. (2010). Disordered Microbial Communities in the Upper Respiratory Tract of Cigarette Smokers. PLoS One, 5(12), e15216. [PubMed: 21188149]

Chen J and Li H (2013) Variable Selection for Sparse Dirichlet-Multinomial Regression with an Application to Microbiome Data Analysis. Ann Appl Stat, 7(1) 418–442.

Finegold SM, et al. (2010) Pyrosequencing study of fecal microflora of autistic and control children. Anaerobe, 16(4), 444–453. [PubMed: 20603222]

Holter N, Mitra M, Maritan A, Cieplak M, Banavar J, and Fedoroff N (2000) Fundamental patterns underlying gene expression profiles: simplicity from complexity. PNAS, 97(15), 8409–8414. [PubMed: 10890920]

Jolliffe IT (2002) Principal Component Analysis. 2nd ed. Springer-Verlag, New York.

Lee S, et al. (2013) Poisson Factor Models with Applications to Non-normalized MicroRNA Profiling. Bioinformatics, 29(9), 1105–1111. [PubMed: 23428639]

Legendre P and Legendre L (1998) Numerical Ecology: Developments in Environmental Modelling. Elsevier Science, Amsterdam.

McCullagh P and Nelder J (1989) Generalized Linear Models. 2nd ed. Chapman and Hall, London.

McMurdie PJ and Holmes S (2014) Waste Not, Want Not: Why Rarefying Microbiome Data Is Inadmissible. PLoS Comput Biol, 10(4), e1003531. [PubMed: 24699258]

Segata N , et al. (2012) Metagenomic microbial community profiling using unique clade-specific marker genes. Nature Methods, 8, 811814.

Shen H and Huang J (2008) Forecasting Time Series of Inhomogeneous Poisson Processes with Application to Call Center Workforcet a;e Management. The Annals of Applied Statistics, 2(2), 601–623.

Sunagawa S, et al. (2013) Metagenomic species profiling using universal phylogenetic marker genes. Nature Methods, 10, 11961199.

The Human Microbiome Project Consortium (2012) Structure, function and diversity of the healthy human microbiome. Nature, 486, 207–214. [PubMed: 22699609]

Warton DI, et al. (2012) Distance-based multivariate analyses confound location and dispersion effects. Methods in Ecology and Evolution, 3, 89–101.

Zeller G, et al. (2014). Potential of fecal microbiota for early-stage detection of colorectal cancer. Mol Syst Biol, 10: 766. [PubMed: 25432777]

Author Manuscript Author Manuscript Author Manuscript Author Manuscript

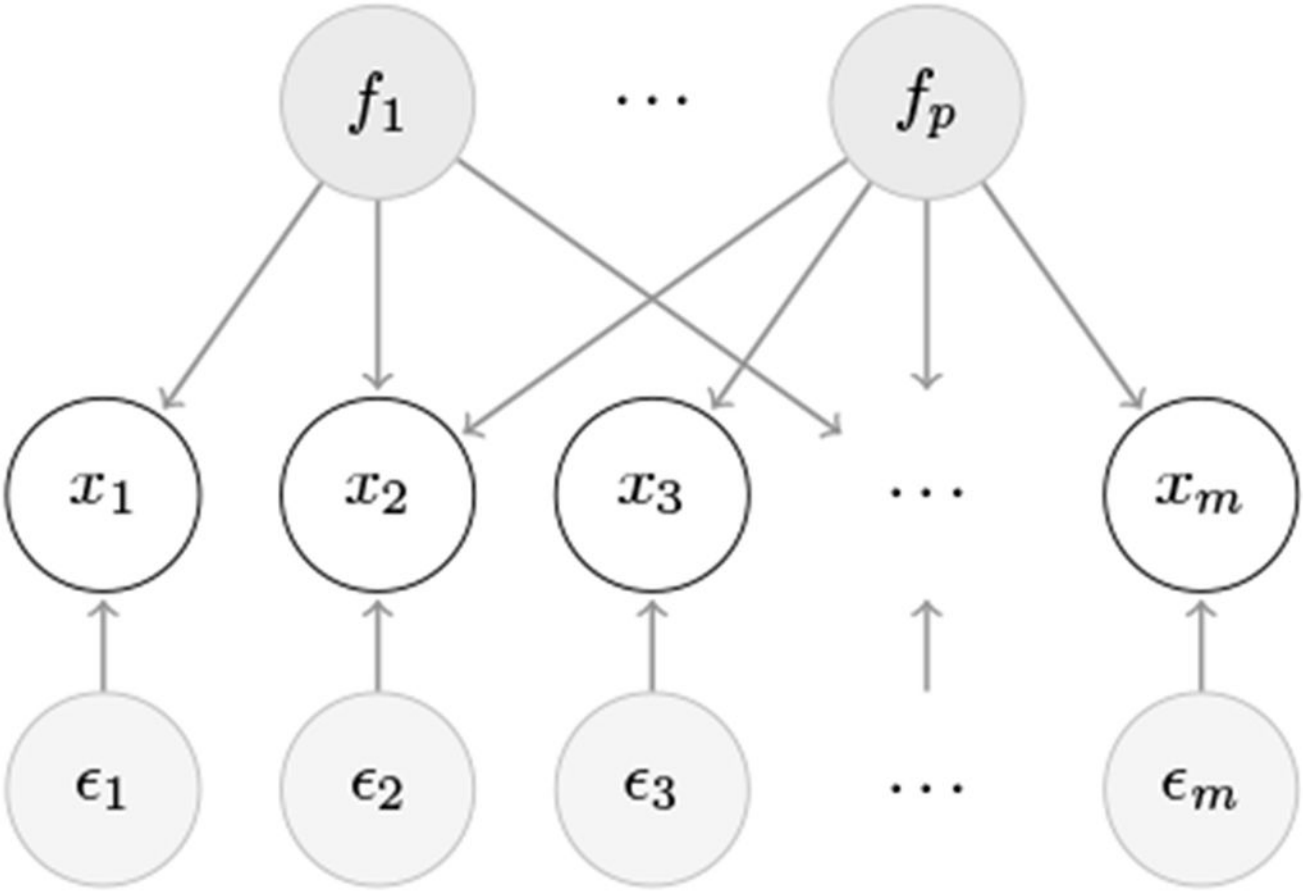

Figure 1.
A graphical representation of a factor model for an example involving p orthogonal factors and m observed variables.

Author Manuscript Author Manuscript Author Manuscript Author Manuscript

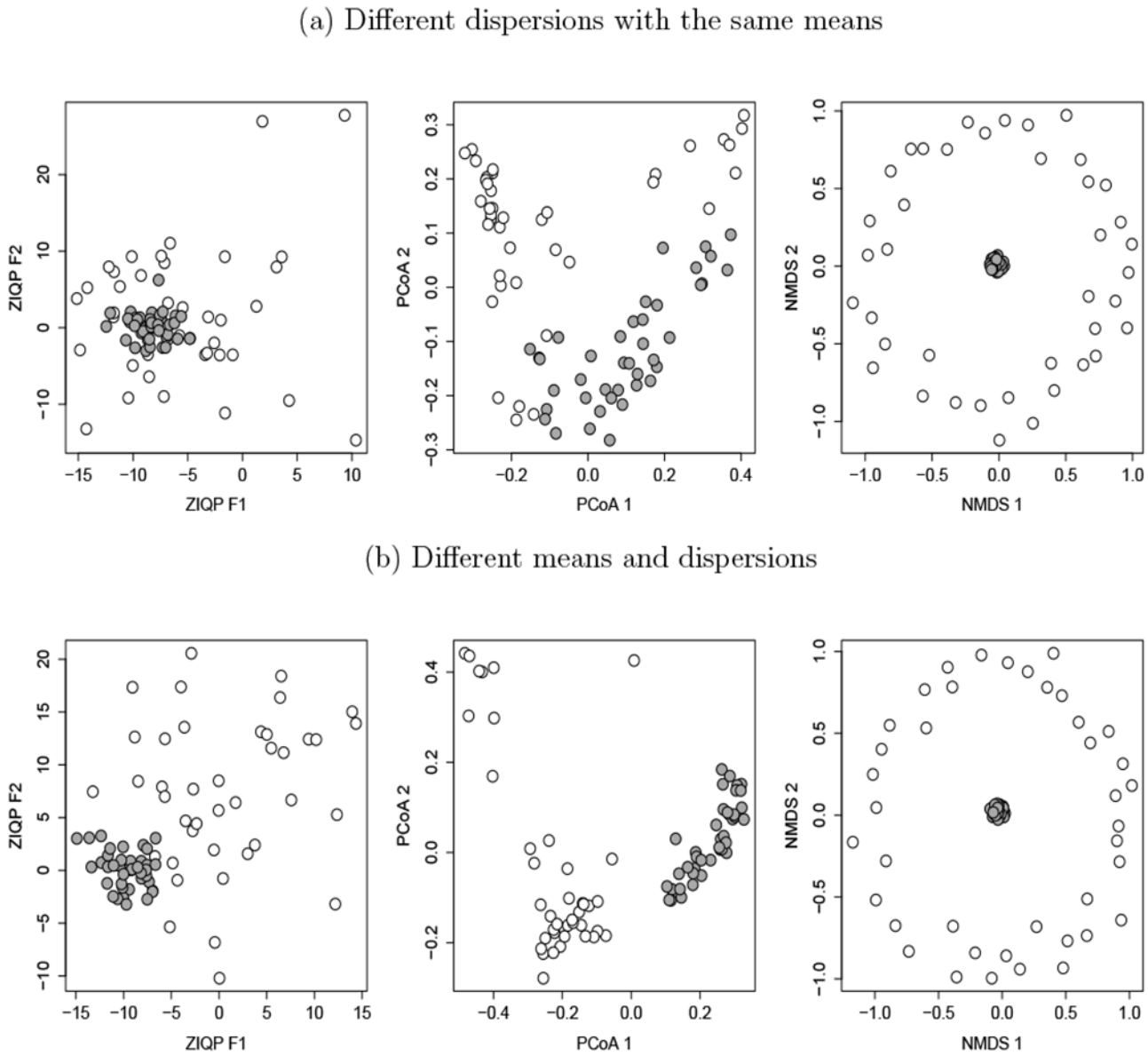

Figure 2.

Simulation results where data are generated from NB distributions. (a) There is a difference in dispersions but not in means between the two populations. (b) There is a difference in both means and dispersions between the two populations. The difference in dispersions between two populations is 10 fold. White circles represent samples in a population with higher dispersion.

Author Manuscript Author Manuscript Author Manuscript Author Manuscript

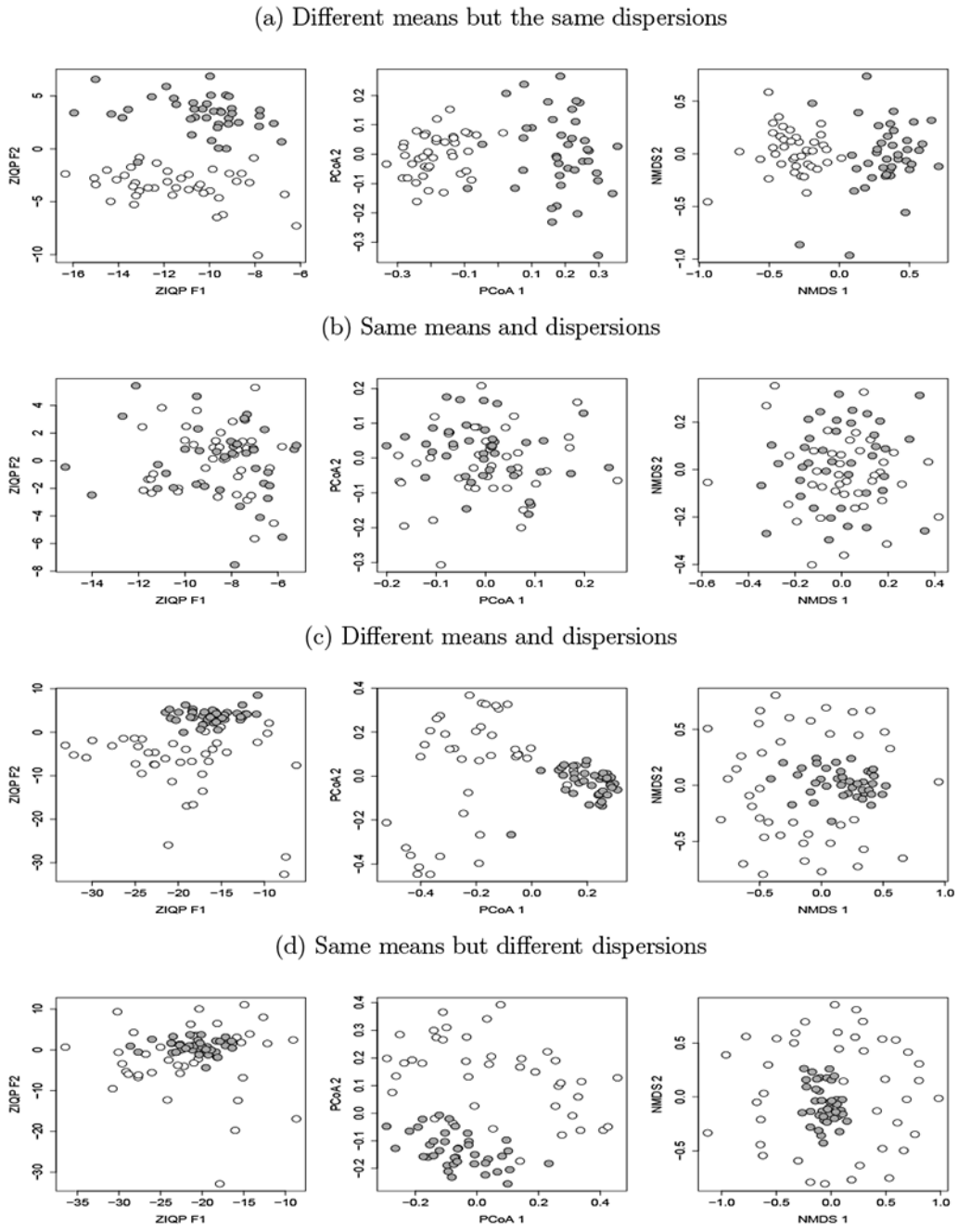

Figure 3.

Simulation results where data are generated from ZINB distributions with taxon-specific overdispersion parameters. Four scenarios are presented, where each population is represented by a different color and white circles represents samples in a population with higher dispersion.

Author Manuscript Author Manuscript Author Manuscript Author Manuscript

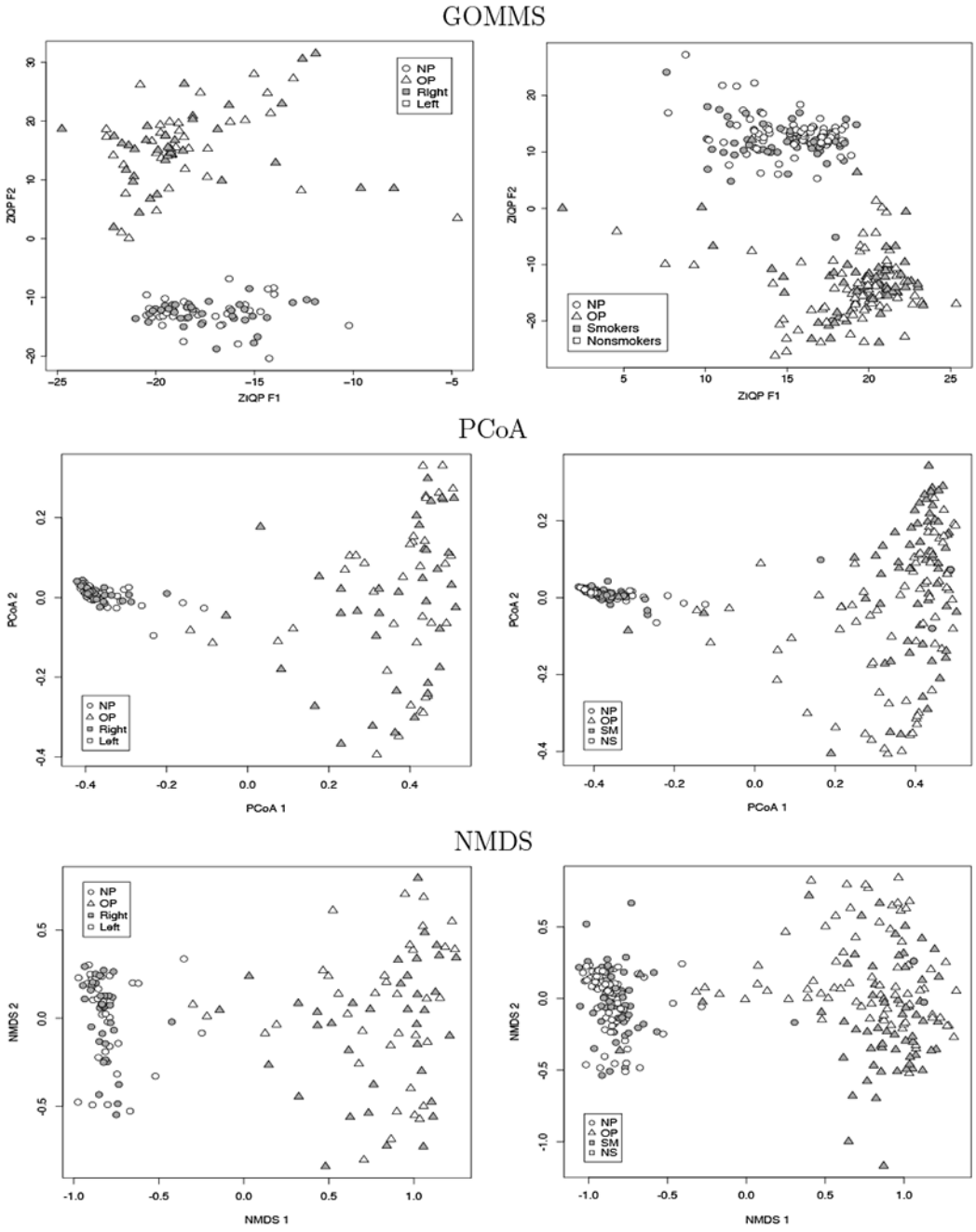

Figure 4.

Comparisons of bacterial community compositions from three methods, where circles and triangles represent samples from nasopharynx and oropharynx, respectively. Left panel: comparison between nasopharynx and oropharynx among the non-smokers. Dark gray color is for samples from the right side of nasopharynx and oropharynx and white color is for samples from the left side. Right panel: comparison between smokers and non-makers. Dark gray color represents samples of smokers and white color represents samples of nonsmokers.

Author Manuscript Author Manuscript Author Manuscript Author Manuscript

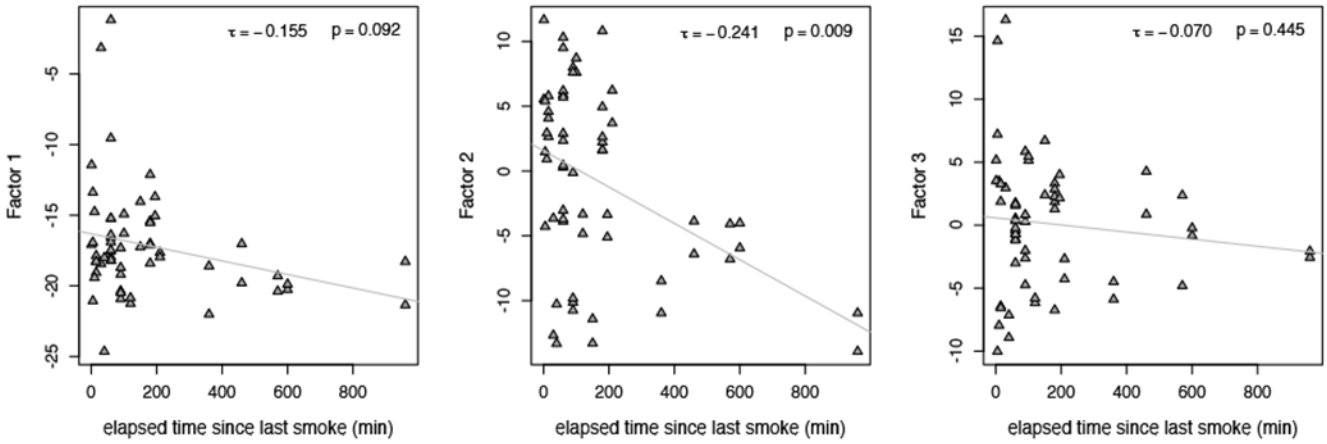

Figure 5.

Elapsed time since last smoke vs. the estimated factor loadings for oropharyngeal samples of the smokers, τ is Kendall's rank correlation coefficient and p is its corresponding p-value.

Author Manuscript

Author Manuscript

Author Manuscript

Author Manuscript

Table 1

Size/power comparison based on Hotelling's two-sample T^2 test for mean difference between two populations using the estimated coordinates at $\alpha = 0.01$. For each entry, the first number is the empirical test size, and the second number is the power.

	No difference in dispersions			Difference in dispersions		
	GOMMS	**NMDS**	**PCoA**	**GOMMS**	**NMDS**	**PCoA**
Sample size, n						
20	0.01/1.00	0.01/1.00	0.00/1.00	0.01/0.33	0.00/0.18	0.56/0.99
50	0.01/1.00	0.00/1.00	0.00/1.00	0.01/0.79	0.00/0.42	1.00/1.00
100	0.02/1.00	0.01/1.00	0.00/1.00	0.02/0.98	0.00/0.79	1.00/1.00
# of taxa, p						
50	0.00/0.98	0.00/0.99	0.01/1.00	0.04/0.60	0.00/0.39	0.76/1.00
100	0.01/1.00	0.00/1.00	0.00/1.00	0.01/0.79	0.00/0.42	1.00/1.00
200	0.00/1.00	0.00/1.00	0.01/1.00	0.03/0.94	0.00/0.40	1.00/1.00
% Differential abundant taxa						
10%	0.00/1.00	0.00/1.00	0.00/1.00	0.01/0.53	0.00/0.21	1.00/1.00
20%	0.01/1.00	0.00/1.00	0.00/1.00	0.01/0.79	0.00/0.42	1.00/1.00
30%	0.01/1.00	0.00/1.00	0.00/1.00	0.04/0.89	0.00/0.77	0.91/1.00
Sparsity, % of zeros						
30%	0.00/1.00	0.00/1.00	0.01/1.00	0.03/0.91	0.00/0.27	1.00/1.00
50%	0.01/1.00	0.00/1.00	0.00/1.00	0.02/0.81	0.00/0.46	1.00/1.00
70%	0.02/1.00	0.00/1.00	0.01/1.00	0.04/0.81	0.00/0.68	0.99/1.00