Equal extraction efficiency

Let i be the taxon index and j be the sample index. Let $\mathbf{X}$ be the $n \times p$ design matrix of the metadata. Let A_{ij} be the $\log_2$ absolute abundance of taxon i in sample j, so the absolute abundance is given by $2^{A_{ij}}$. Also, suppose the absolute abundance is related to the metadata by: $E(A_{ij}) = \sum_p \mathbf{X}_{jp}\beta_{ip} + \epsilon_{ij}$ where β_{ip} is the slope relating metadatum p to taxon i's $\log_2$ absolute abundance. We can group these slopes together for a taxon as a $p \times 1$ column vector: $\boldsymbol{\beta}_i = (\beta_{i1}, \beta_{i2}, ..., \beta_{ip})^T$. Let the $n \times 1$ vector $\mathbf{A}_i = (A_{i1}, A_{i2}, ..., A_{in})^T$ denote the vector of $\log_2$ absolute abundances for taxon i. Let D_j be the $\log_2$ total abundance in sample j, so $2^{D_j} = \sum_i 2^{A_{ij}}$. Let Y_{ij} be the $\log_2$ relative abundance of taxon i in sample j, so $2^{Y_{ij}} = 2^{A_{ij}}/2^{D_j} \iff Y_{ij} = A_{ij} - D_j$. As with the absolute abundances, let $\mathbf{Y}_i$ be the vector of $\log_2$ relative abundances for taxon i, and let $\mathbf{D}$ be the vector of $\log_2$ total abundances. Thus, $\mathbf{Y}_i = \mathbf{A}_i - \mathbf{D}$ for all i. If we observed a vector $\mathbf{A}_i$ and wanted to estimate $\boldsymbol{\beta}_i$, we would use the OLS solution: $\hat{\boldsymbol{\beta}}_i = (\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\mathbf{A}_i$. By the abundance decomposition above, we also have:

$$\hat{\boldsymbol{\beta}}_i = (\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\mathbf{A}_i = (\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\mathbf{Y}_i + (\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\mathbf{D} = \hat{\boldsymbol{\beta}}_i^{\text{rel}} + \hat{\boldsymbol{\beta}}^{\text{tot}}$$

where $\hat{\boldsymbol{\beta}}_i^{\text{rel}}$ is the result of regressing the $\log_2$ relative abundances on the design matrix, and $\hat{\boldsymbol{\beta}}^{\text{tot}}$ is the result of regressing the $\log_2$ total abundances on the design matrix. Note that since $\hat{\boldsymbol{\beta}}^{\text{tot}}$ is the same for all taxa, if the absolute abundance coefficient for one taxa i is d larger than the absolute abundance coefficient for another taxa i' (i.e., $\hat{\beta}_{ip} - \hat{\beta}_{i'p} = d$), the relative abundance coefficient for taxa i will be d larger than the relative abundance coefficient for taxa i' (i.e., $\hat{\beta}_{ip}^{\text{rel}} - \hat{\beta}_{i'p}^{\text{rel}} = d$). Thus, if all we have is the relative abundance data, we can't determine the absolute slopes $\hat{\beta}_{ip}$ themselves, but we can determine the relative coefficients $\hat{\beta}_{ip}^{\text{rel}}$, and the ordering of and spacing between these will be identical to the ordering of and spacing between the absolute coefficient for each metadatum. Also, since the OLS solution is unbiased for $\boldsymbol{\beta}_i$, we have:

$$\boldsymbol{\beta}_i = E(\hat{\boldsymbol{\beta}}_i) = E(\hat{\boldsymbol{\beta}}_i^{\text{rel}}) + E(\hat{\boldsymbol{\beta}}^{\text{tot}})$$

This implies that if we assume at least half of the features do not change with respect to a particular metadatum (i.e., $\beta_{ip} = 0$ for at least half the taxa i), the median absolute abundance coefficient will be 0 (i.e., $\text{med}(\beta_{1p}, \beta_{2p}, ...) = 0$), so $\text{med}(E(\hat{\beta}_{1p}^{\text{rel}}) + E(\hat{\beta}_p^{\text{tot}}), E(\hat{\beta}_{2p}^{\text{rel}}) + E(\hat{\beta}_p^{\text{tot}}), ...) = 0$. Since the term $E(\hat{\beta}_p^{\text{tot}})$ is the same in all of these, this implies $\text{med}(E(\hat{\beta}_{1p}^{\text{rel}}), E(\hat{\beta}_{2p}^{\text{rel}}), ...) = -E(\hat{\beta}_p^{\text{tot}})$. Thus, a test of $\beta_{ip} = 0$ is equivalent to a test of

$$E(\hat{\beta}_{ip}^{\text{rel}}) + E(\hat{\beta}_p^{\text{tot}}) = 0 \iff E(\hat{\beta}_{ip}^{\text{rel}}) = -E(\hat{\beta}_p^{\text{tot}}) = \text{med}(E(\hat{\beta}_{1p}^{\text{rel}}), E(\hat{\beta}_{2p}^{\text{rel}}), ...)$$

That is, testing whether one taxon's relative abundance slope for a metadatum is different from the median relative abundance slope for that metadatum is the same as testing whether that taxon's absolute abundance slope is different from 0.

Unequal extraction efficiency

Now, assume extraction efficiencies are not the same. Then, with A_{ij} as the $\log_2$ absolute abundance, Y_{ij} as the $\log_2$ observed relative abundance, S_i as the $\log_2$ sampling efficiency, and D_j as the $\log_2$ total efficiency-scaled abundance in sample j, we have $2^{D_j} = \sum_i 2^{A_{ij}S_i}$ and $2^{A_{ij}}2^{S_i}/2^{D_j} = 2^{Y_{ij}} \iff A_{ij} + S_i - D_j = Y_{ij}$. Let $\mathbf{S}_i$ be S_i repeated p times to match the dimensions of the taxon-specific slopes. With an equivalent decomposition to before, we have:

$$\hat{\boldsymbol{\beta}}_i = (\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\mathbf{A}_i = (\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\mathbf{Y}_i + (\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\mathbf{D} - (\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\mathbf{S}_i = \hat{\boldsymbol{\beta}}_i^{\text{rel}} + \hat{\boldsymbol{\beta}}^{\text{tot}} - \boldsymbol{\beta}_i^{\text{eff}}$$

and

$$\boldsymbol{\beta}_i = E(\hat{\boldsymbol{\beta}}_i^{\text{rel}}) + E(\hat{\boldsymbol{\beta}}^{\text{tot}}) - E(\hat{\boldsymbol{\beta}}_i^{\text{eff}})$$

Slightly different from before, if the absolute abundance slopes of two taxa are differenet by d (i.e., $\beta_{ip} - \beta_{i'p} = d$),

$$E(\hat{\beta}_{ip}^{\text{rel}}) - E(\hat{\beta}_{ip}^{\text{eff}}) - E(\hat{\beta}_{i'p}^{\text{rel}}) + E(\hat{\beta}_{i'p}^{\text{eff}}) = d$$

Unlike before, we do not immediately obtain $E(\hat{\beta}_{ip}^{\text{rel}}) = E(\hat{\beta}_{i'p}^{\text{rel}})$, so it is not immediately the case that a difference of d in absolute slopes corresponds to a difference of d in relative slopes. However, if we assume the sampling efficiencies are uncorrelated with the design matrix, $E(\hat{\beta}_{ip}^{\text{eff}}) = E(\hat{\beta}_{i'p}^{\text{eff}}) = 0$, and we do obtain this result. This seems plausible in practice: the sampling efficiency should just depend on the MGS protocol, and if the protocol is the same across the samples, there should be no correlation with the metadata.