Sparse Negative Binomial Signal Recovery for Genomic Variant Prediction in Diploid Species

Jocelyn Ornelas Munoz[⋆], Erica M. Rutter[⋆], Mario Banuelos[†], Suzanne S. Sindi[⋆], Roummel F. Marcia[⋆]

[⋆]Department of Applied Mathematics, UC Merced
[†] Department of Mathematics California State University, Fresno

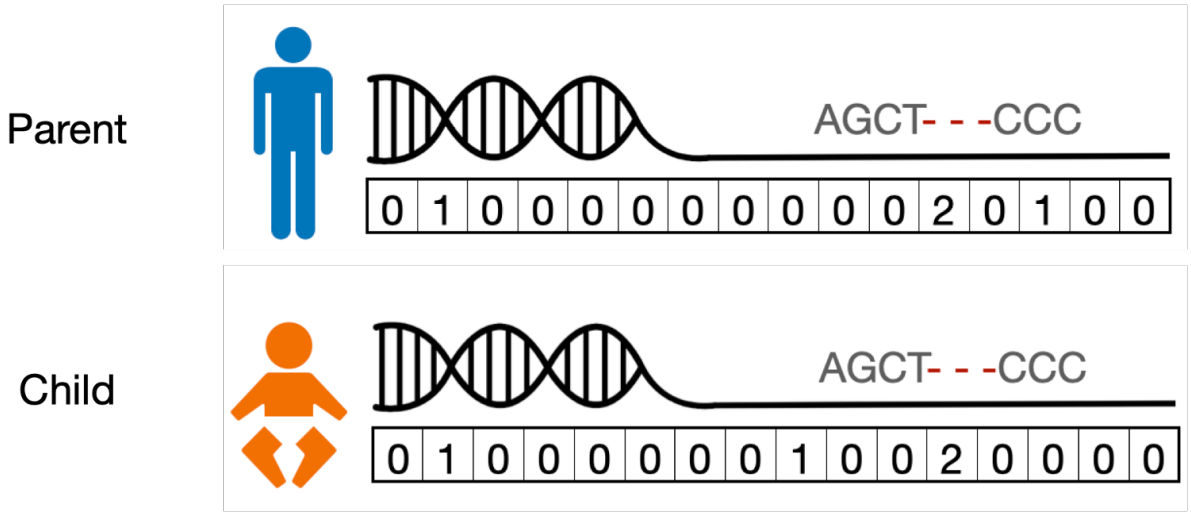
Wednesday, November 16, 2022

Overview

Goal: Reconstruct signals from noisy measurements as accurately as possible.

Problem Formulation

Suppose we have sequencing data for two individuals.



We denote the true signals:

- Parent : $\vec{f}_P \in \{0,1,2\}^n$
- Child : $\vec{f}_C = \vec{f}_H + \vec{f}_N \in \{0,1,2\}^n$

where $\vec{f}_H \in \{0,1,2\}^n$ and $\vec{f}_N \in \{0,1,2\}^n$ correspond to the vectors of inherited (H) and novel (N) SVs in the child, respectively.

Problem Formulation

For each individual signal we consider two indicator signals such that $\vec{f}_i = 2\vec{z}_i + \vec{y}_i$ for $i \in \{P, H, N\}$.

Homozygous indicator:

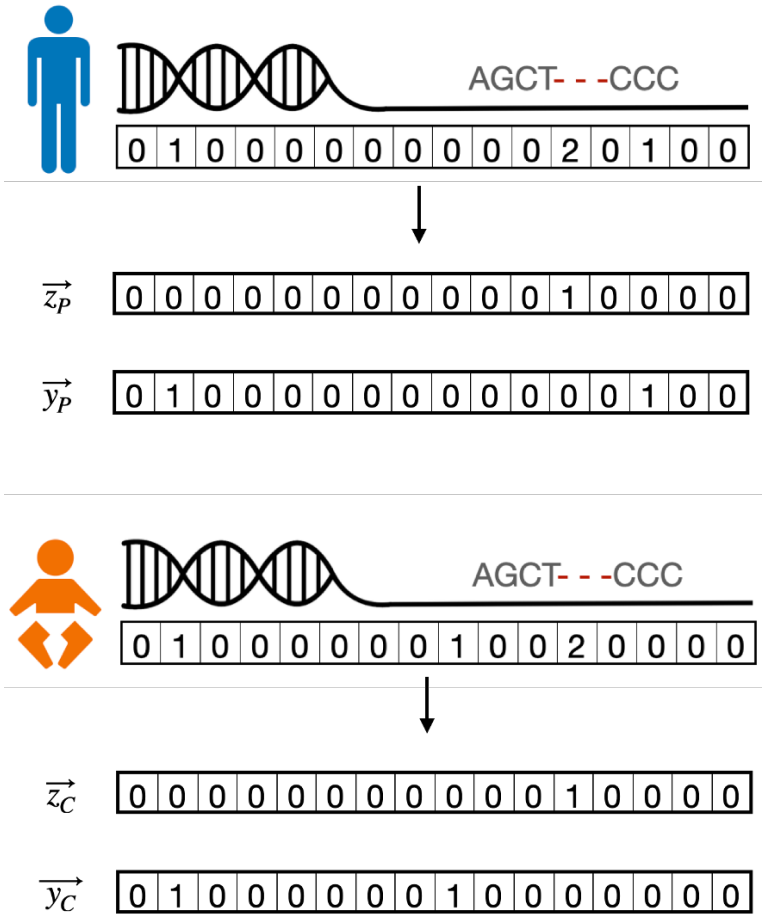
$\vec{z}_j = 1$ if 2 copies of an SV are present at location j

$\vec{z}_j = 0$ otherwise

Heterozygous indicator:

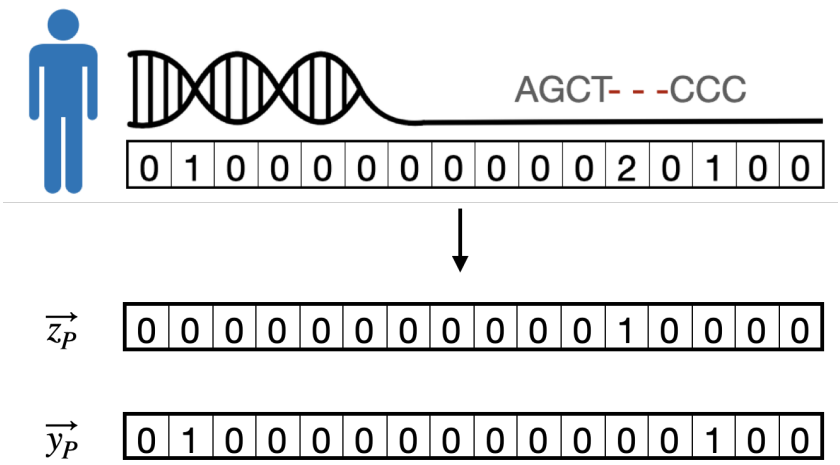
$\vec{y}_j = 1$ if 1 copy of an SV is present at location j

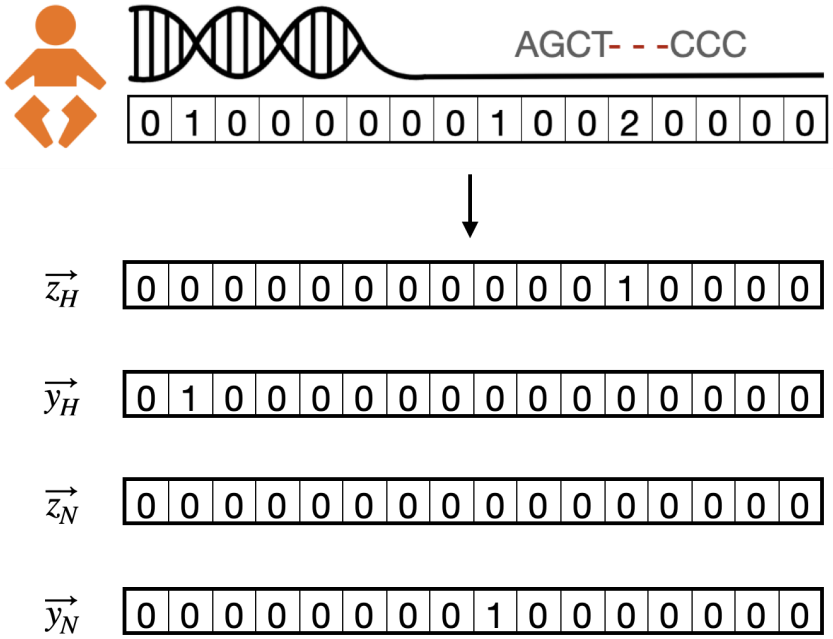
$\vec{y}_j = 0$ otherwise



Problem Formulation

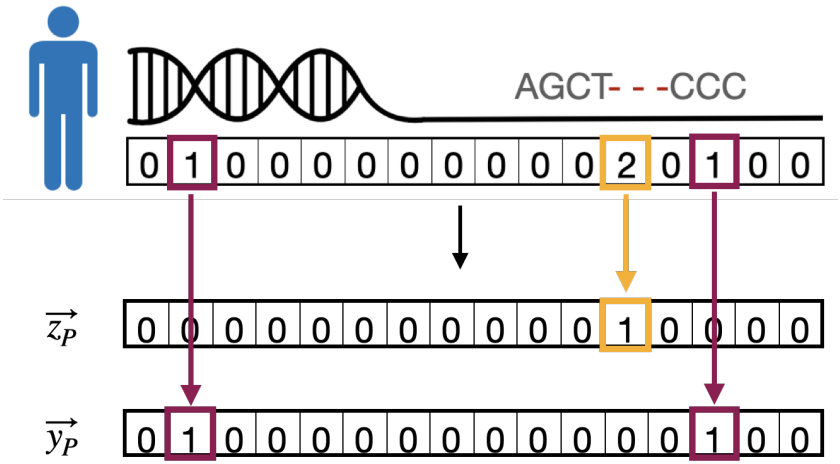
Next, we split the child's diploid indicators into inherited, H, and novel, N, indicators

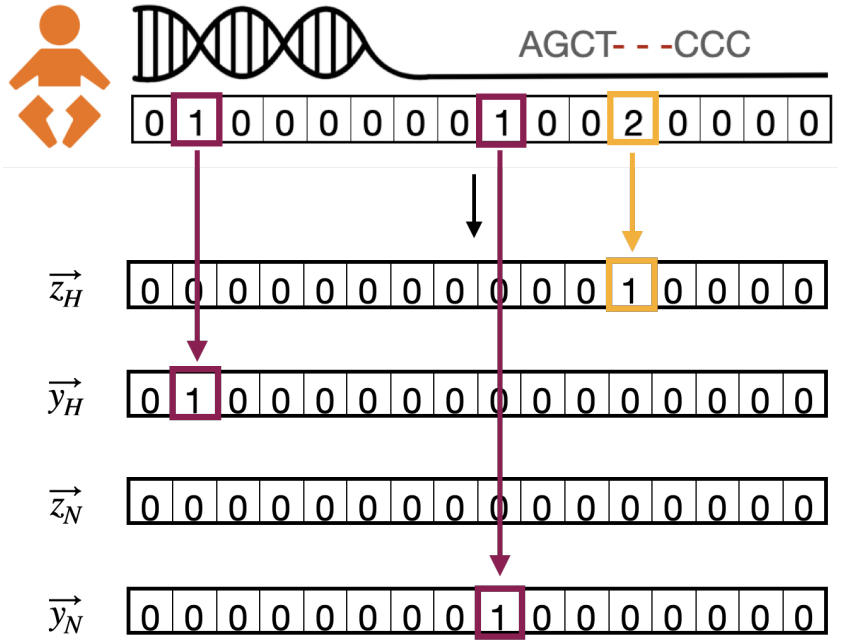




Problem Formulation

Next, we split the child's diploid indicators into inherited, H, and novel, N, indicators





Problem Formulation

We leverage relatedness in our model by incorporating the following assumptions:

Problem Formulation

We leverage relatedness in our model by incorporating the following assumptions:

- An individual can only have 0, 1, or 2 copies of an SV per candidate location

Problem Formulation

We leverage relatedness in our model by incorporating the following assumptions:

- An individual can only have 0, 1, or 2 copies of an SV per candidate location
- If the child has an inherited SV, the SV cannot also be novel

Problem Formulation

We leverage relatedness in our model by incorporating the following assumptions:

- An individual can only have 0, 1, or 2 copies of an SV per candidate location
- If the child has an inherited SV, the SV cannot also be novel
- If the child has an inherited homogenous SV, the parent has at least a heterogeneous SV

Problem Formulation

We leverage relatedness in our model by incorporating the following assumptions:

- An individual can only have 0, 1, or 2 copies of an SV per candidate location
- If the child has an inherited SV, the SV cannot also be novel
- If the child has an inherited homogenous SV, the parent has at least a heterogeneous SV
- If the parent has a homogeneous SV, the child must have at least a heterogeneous SV

Problem Formulation

We leverage relatedness in our model by incorporating the following assumptions:

- An individual can only have 0, 1, or 2 copies of an SV per candidate location
- If the child has an inherited SV, the SV cannot also be novel
- If the child has an inherited homogenous SV, the parent has at least a heterogeneous SV
- If the parent has a homogeneous SV, the child must have at least a heterogeneous SV
- If the child has a novel SV, the parent does not have that SV

Problem Formulation

We leverage relatedness in our model by incorporating the following assumptions:

- An individual can only have 0, 1, or 2 copies of an SV per candidate location
- If the child has an inherited SV, the SV cannot also be novel
- If the child has an inherited homogenous SV, the parent has at least a heterogeneous SV
- If the parent has a homogeneous SV, the child must have at least a heterogeneous SV
- If the child has a novel SV, the parent does not have that SV

Mathematically, these are written as constraints:

$$\mathcal{S} = \left\{ \vec{f} = \begin{bmatrix} \vec{z}_P \\ \vec{z}_H \\ \vec{z}_N \\ \vec{y}_P \\ \vec{y}_H \\ \vec{y}_N \end{bmatrix} \in \mathbb{R}^{6n} : \quad \begin{array}{l} \mathbf{0} \leq \vec{z}_i + \vec{y}_i \leq \mathbf{1} \\ \mathbf{0} \leq \vec{z}_H + \vec{y}_H + \vec{z}_N + \vec{y}_N \leq \mathbf{1} \\ \mathbf{0} \leq \vec{z}_H \leq \vec{z}_P + \vec{y}_P \leq \mathbf{1} \\ \mathbf{0} \leq \vec{z}_P \leq \vec{z}_H + \vec{y}_H \leq \mathbf{1} \\ \mathbf{0} \leq \vec{z}_N + \vec{y}_N \leq 1 - (\vec{z}_P + \vec{y}_P) \leq \mathbf{1} \end{array} \right\}$$

Observational Model

Under the Negative Binomial assumption, we can model the noise in our observation as

$$\begin{bmatrix} \vec{s}_P \\ \vec{s}_C \end{bmatrix} \sim \text{NegBin}\left(\begin{bmatrix} \vec{z}_P\,(2\lambda_P - \varepsilon) \;+\; \vec{y}_P\,(\lambda_P - \varepsilon) \\ \vec{z}_H\,(2\lambda_C - \varepsilon) \;+\; \vec{y}_H\,(\lambda_C - \varepsilon) \;+\; \vec{z}_N\,(2\lambda_C - \varepsilon) \;+\; \vec{y}_N\,(\lambda_C - \varepsilon) \end{bmatrix} \right)$$

where

- $\vec{s}_P,\ \vec{s}_C \in \mathbb{R}^n$: length-$n$ vector of counts for the parent and the child, respectively
- $\lambda_P,\ \lambda_C$: sequencing coverage of the parent and the child, respectively
- $\varepsilon > 0$: measurement errors incurred in the sequencing and mapping process

Observational Model

More generally, if we let

$$\vec{s} = \begin{bmatrix} \vec{s}_P \\ \vec{s}_C \end{bmatrix}, \quad \vec{z} = \begin{bmatrix} \vec{z}_P \\ \vec{z}_H \\ \vec{z}_N \end{bmatrix}, \quad \vec{y} = \begin{bmatrix} \vec{y}_P \\ \vec{y}_H \\ \vec{y}_N \end{bmatrix}, \quad \vec{f} = \begin{bmatrix} \vec{z} \\ \vec{y} \end{bmatrix}$$

we can express our observation model as

$$\vec{s} \sim \text{NegBin}(A\vec{f} + \varepsilon \mathbf{1})$$

where

- $\vec{s} \in \mathbb{Z}_+^{2n}$ is a length-$2n$ vector of observed coverage counts
- $\vec{f} \in \{0, 1\}^{6n}$ is the signal of interest
- $\mathbf{1} \in \mathbb{R}^{2n}$ is the vector of ones
- $A \in \mathbb{R}^{2n \times 6n}$ is the sequence coverage matrix

Observational Model

The probability of observing a particular vector of counts $\vec{s}$ given the true signal $\vec{f}$ is given by

$$p(\vec{s}\,|A\vec{f}) = \prod_{l=1}^{2n} \left(\frac{1}{1 + (A\vec{f})_l + \varepsilon} \right) \left(\frac{((A\vec{f})_l + \varepsilon)}{1 + (A\vec{f})_l + \varepsilon} \right)^{s_l}$$

We use gradient-based maximum likelihood approach to recover the indicator variables $\vec{z}_i$, $\vec{y}_i$, $i \in \{ P, H, N \}$.

Observational Model

The probability of observing a particular vector of counts $\vec{s}$ given the true signal $\vec{f}$ is given by

$$p(\vec{s}\,|A\vec{f}) = \prod_{l=1}^{2n} \left(\frac{1}{1 + (A\vec{f})_l + \varepsilon} \right) \left(\frac{((A\vec{f})_l + \varepsilon)}{1 + (A\vec{f})_l + \varepsilon} \right)^{s_l}$$

We want to minimize the Negative Binomial negative log-likelihood function

$$F(\vec{f}) \equiv \sum_{l=1}^{2n} (1 + s_l) \log\left(1 + e_l^T A\vec{f} + \varepsilon\right) - s_l \log\left(e_l^T A\vec{f} + \varepsilon\right)$$

Optimization Framework

Our objective function takes the following form:

$$\begin{aligned} \underset{\vec{f} \in \mathbb{R}^{6n}}{\text{minimize}} \quad & F(\vec{f}) + \tau \text{pen}(\vec{f}) \\ \text{subject to} \quad & \vec{f} \in \mathcal{S} \end{aligned}$$

where

- $F(\vec{f})$ is the Negative Binomial negative log-likelihood function
- $\text{pen}(\vec{f}) = (\|\vec{z}_P\|_1 + \|\vec{z}_H\|_1 + \|\vec{y}_P\|_1 + \|\vec{y}_H\|_1) + \gamma(\|\vec{z}_N\|_1 + \|\vec{y}_N\|_1)$ is the ℓ_1 penalty term
- $\tau > 0$ and $\gamma \gg 1$ are regularization parameters
- $\mathcal{S}$ is the set of vectors satisfying the biological constraints

Optimization Framework

We use a second-order Taylor series expansion to approximate $F(\vec{f})$ at iterate $\vec{f}^k$ and solve the separable quadratic subproblem,

$$\vec{f}^{k+1} = \begin{array}{ll} \underset{\vec{f} \in \mathbb{R}^{6n}}{\arg\min} & \mathcal{Q}(\vec{f}) = \frac{1}{2}\|\vec{f} - \vec{r}^k\|_2^2 + \frac{\tau}{\alpha_k}\text{pen}(\vec{f}) \\ \text{subject to} & \vec{f} \in \mathcal{S} \end{array} \tag{1}$$

where $\vec{r}^k = [\vec{r}^k_{z_P}, \vec{r}^k_{z_H}, \vec{r}^k_{z_N}, \vec{r}^k_{y_P}, \vec{r}^k_{y_H}, \vec{r}^k_{y_N}]^T = \vec{f}^k - \frac{1}{\alpha_k}\nabla F(\vec{f}^k)$ and α_k is INSERT HERE.

Optimization Framework

Our objective function $\mathcal{Q}(\vec{f})$ is separable and decouples into the function $\mathcal{Q}(\vec{f}) = \sum_{j=1}^{n} \mathcal{Q}_j(\vec{z}_P, \vec{z}_H, \vec{z}_N, \vec{y}_P, \vec{y}_H, \vec{y}_N)$, where

$$\begin{aligned}\mathcal{Q}_j(\vec{z}_P, \vec{z}_H, \vec{z}_N, \vec{y}_P, \vec{y}_H, \vec{y}_N) = &\frac{1}{2}\Big\{((\vec{z}_P - \vec{r}^{\,k}_{\vec{z}_P})_j)^2 + ((\vec{z}_H - \vec{r}^{\,k}_{\vec{z}_H})_j)^2 + ((\vec{z}_N - \vec{r}^{\,k}_{\vec{z}_N})_j)^2 \\ &+ ((\vec{y}_P - \vec{r}^{\,k}_{\vec{y}_P})_j)^2 + ((\vec{y}_H - \vec{r}^{\,k}_{\vec{y}_H})_j)^2 + ((\vec{y}_N - \vec{r}^{\,k}_{\vec{y}_N})_j)^2\Big\} \\ &+ \frac{\tau}{\alpha_k}\Big\{|(\vec{z}_P)_j| + |(\vec{z}_H)_j| + \gamma|(\vec{z}_N)_j| \\ &+ |(\vec{y}_P)_j| + |(\vec{y}_H)_j| + \gamma|(\vec{y}_N)_j|\Big\}\end{aligned}$$

Optimization Framework

Since the bounds that define the region $\mathcal{S}$ are component-wise, then Equation (1) separates into subproblems of the form:

$$\begin{aligned}\vec{f}^{k+1} = \min_{\substack{z_P, z_H, z_N, \\ y_P, y_H, y_N \in \mathbb{R}}} & \frac{\tau}{\alpha_k} \left\{ |(\vec{z}_P)_j| + |(\vec{z}_H)_j| + \gamma |(\vec{z}_N)_j| + |(\vec{y}_P)_j| + |(\vec{y}_H)_j| + \gamma |(\vec{y}_N)_j| \right\} \\ & + \frac{1}{2} \Big\{ ((\vec{z}_P - \vec{r}^{\,k}_{\vec{z}_P})_j)^2 + ((\vec{z}_H - \vec{r}^{\,k}_{\vec{z}_H})_j)^2 + ((\vec{z}_N - \vec{r}^{\,k}_{\vec{z}_N})_j)^2 \\ & + ((\vec{y}_P - \vec{r}^{\,k}_{\vec{y}_P})_j)^2 + ((\vec{y}_H - \vec{r}^{\,k}_{\vec{y}_H})_j)^2 + ((\vec{y}_N - \vec{r}^{\,k}_{\vec{y}_N})_j)^2 \Big\} \\ & \text{subject to } \vec{f} \in S \end{aligned} \tag{2}$$

where f_i and r_{z_i}, r_{y_i} are scalar components of $\vec{f}_i$ and $\vec{r}_{z_i}, \vec{r}_{y_i}$, respectively, at the same location.

Optimization Approach

We solve our problem using an alternating block-coordinate descent approach, following the methods. We fix all but one individual and solve Equation (1) over both indicator variables for that individual.

Step 0: Compute the unconstrained minimizer of Equation (1) given by

$$\vec{f} = [\vec{r}_{z_P} - \frac{\tau}{\alpha_k}\mathbf{1}_n,\ \vec{r}_{z_H} - \frac{\tau}{\alpha_k}\mathbf{1}_n,\ \vec{r}_{z_N} - \frac{\tau}{\alpha_k}\gamma\mathbf{1}_n,$$
$$\vec{r}_{y_P} - \frac{\tau}{\alpha_k}\mathbf{1}_n,\ \vec{r}_{y_H} - \frac{\tau}{\alpha_k}\mathbf{1}_n,\ \vec{r}_{y_N} - \frac{\tau}{\alpha_k}\gamma\mathbf{1}_n]^T$$

where $\mathbf{1}_n \in \mathbb{R}^n$

Initialize indicator variables

$$\hat{z}_P^{(0)} = r_{z_P}^k - \frac{\tau}{\alpha_k} \qquad \hat{y}_P^{(0)} = r_{y_P}^k - \frac{\tau}{\alpha_k},$$
$$\hat{z}_H^{(0)} = \{0, r_{z_H}^k - \frac{\tau}{\alpha_k}, 1\} \qquad \hat{y}_H^{(0)} = \{0, r_{y_H}^k - \frac{\tau}{\alpha_k}, 1\},$$
$$\hat{z}_N^{(0)} = \{0, r_{z_N}^k - \frac{\tau}{\alpha_k}\gamma, 1\} \qquad \hat{y}_N^{(0)} = \{0, r_{y_N}^k - \frac{\tau}{\alpha_k}\gamma, 1\}$$

Optimization Approach

(cont.)

Step 1: Project $\hat{z}_P^{(k-1)}$ and $\hat{y}_P^{(k-1)}$ onto the feasible set S with fixed inherited and novel variables to obtain $\hat{z}_P^{(k)}$ and $\hat{y}_P^{(k)}$.

Optimization Approach

(cont.)

Step 1: Project $\hat{z}_P^{(k-1)}$ and $\hat{y}_P^{(k-1)}$ onto the feasible set S with fixed inherited and novel variables to obtain $\hat{z}_P^{(k)}$ and $\hat{y}_P^{(k)}$.

Step 2: Project $\hat{z}_H^{(k-1)}, \hat{y}_H^{(k-1)}$ onto the feasible set S with fixed parent and child's novel indicator variables to obtain $\hat{z}_H^{(k)}$ and $\hat{y}_H^{(k)}$.

Optimization Approach

(cont.)

Step 1: Project $\hat{z}_P^{(k-1)}$ and $\hat{y}_P^{(k-1)}$ onto the feasible set S with fixed inherited and novel variables to obtain $\hat{z}_P^{(k)}$ and $\hat{y}_P^{(k)}$.

Step 2: Project $\hat{z}_H^{(k-1)}, \hat{y}_H^{(k-1)}$ onto the feasible set S with fixed parent and child's novel indicator variables to obtain $\hat{z}_H^{(k)}$ and $\hat{y}_H^{(k)}$.

Step 3: Project $\hat{z}_N^{(k-1)}, \hat{y}_N^{(k-1)}$ onto the feasible set S with fixed parent and child's inherited indicator variables to obtain $\hat{z}_N^{(k)}$ and $\hat{y}_N^{(k)}$.

Optimization Approach

(cont.)

Step 1: Project $\hat{z}_P^{(k-1)}$ and $\hat{y}_P^{(k-1)}$ onto the feasible set S with fixed inherited and novel variables to obtain $\hat{z}_P^{(k)}$ and $\hat{y}_P^{(k)}$.

Step 2: Project $\hat{z}_H^{(k-1)}, \hat{y}_H^{(k-1)}$ onto the feasible set S with fixed parent and child's novel indicator variables to obtain $\hat{z}_H^{(k)}$ and $\hat{y}_H^{(k)}$.

Step 3: Project $\hat{z}_N^{(k-1)}, \hat{y}_N^{(k-1)}$ onto the feasible set S with fixed parent and child's inherited indicator variables to obtain $\hat{z}_N^{(k)}$ and $\hat{y}_N^{(k)}$.

Step 4: Repeat Steps $1, 2$, and 3 until the relative difference between consecutive iterates converges to $\|\vec{f}^{k+1} - \vec{f}^{k}\| / \|\vec{f}^{k}\| \leq 10^{-8}$.

Optimization Approach

Insert figure of feasible regions (or do it by steps)

Numerical Experiments

We implemented our method by modifying the existing SPIRAL approach to include the negative binomial statistical method. We compared the Poisson-based predictions (SPIRAL) with the Negative Binomial-based predictions (NEBULA).

Gracias!

Observational Model

[noframenumbering] Let $I_n \in \mathbb{R}^{n \times n}$ be the $n \times n$ identity matrix. Then we can write the sequence coverage matrix $A = [A_1\ A_2] \in \mathbb{R}^{2n \times 6n}$ with:

$$A_1 = \left[\begin{array}{c|c|c} (2\lambda_P - \varepsilon)I_n & 0 & 0 \\ \hline 0 & (2\lambda_C - \varepsilon)I_n & (2\lambda_C - \varepsilon)I_n \end{array}\right]$$

and

$$A_2 = \left[\begin{array}{c|c|c} (\lambda_P - \varepsilon)I_n & 0 & 0 \\ \hline 0 & (\lambda_C - \varepsilon)I_n & (\lambda_C - \varepsilon)I_n \end{array}\right]$$