JUSTIN D. SILVERMAN

NOW:

MEDICAL SCIENTIST TRAINING PROGRAM

COMPUTATIONAL BIOLOGY AND BIOINFORMATICS

DUKE UNIVERSITY

SOON:

COLLEGE OF INFORMATION SCIENCE AND TECHNOLOGY

DEPARTMENT OF MEDICINE

PENN STATE

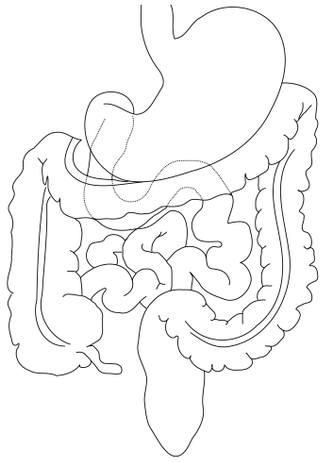
StatsAtHome.com

inschool4life

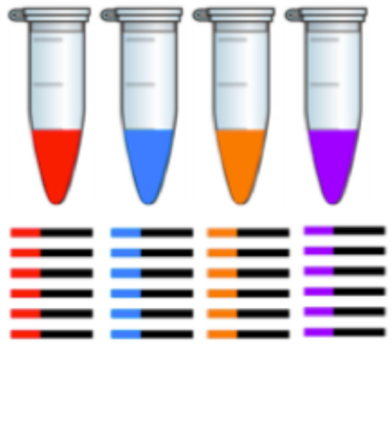
SCALABLE BAYESIAN MULTINOMIAL LOGISTIC-NORMAL MODELS FOR THE ANALYSIS OF SEQUENCE COUNT DATA

DATA COLLECTION AND SAMPLE PROCESSING

Sample Collection and Storage



DNA Extraction PCR Amplification



Sequencing

	Species 1	Species 2	Species 3	Sp
Sample 1	23	53	2	
Sample 2	69	64	70	
Sample 3	33	100	68	
Sample 4	5	63	57	
Sample 5	76	80	46	
Sample 6	58	7	37	
Sample 7	10	87	32	
Sample 8	21	89	73	

Adapted from Hamady. et al., *Nature Methods*, 2008

COMPOSITION: A CONTROVERSIAL TOPIC

THE DATA IS "COMPOSITIONAL"

It's all relative: analyzing microbiome data as compositions

Gregory B. Gloor PhD [a], Jia Rong Wu BSc [a], Vera Pawlowsky-Glahn PhD [b], Juan José Egozcue PhD [c]

Microbiome Datasets Are Compositional: And This Is Not Optional

Gregory B. Gloor[1*], Jean M. Macklaim[1], Vera Pawlowsky-Glahn[2] and Juan J. Egozcue[3]

NO ITS NOT

Susan Holmes @SherlockpHolmes · 5 Apr 2018

Replying to @timtriche @samclifford and 2 others

Absolutely not, microbiome data are not **compositional** and those methods don't apply,although it does apply to geostat data and other situations when one has a whole of exactly the same size. In microbiome data you have to control for different amounts of bacteria.

2 1 1

CHALLENGES OF COMPOSITION

	Species 1	Species 2	Species 3	Sp
Sample 1	23	53	2	
Sample 2	69	64	70	
Sample 3	33	100	68	
Sample 4	5	63	57	
Sample 5	76	80	46	
Sample 6	58	7	37	
Sample 7	10	87	32	
Sample 8	21	89	73	

CHALLENGES OF COMPOSITION

	Species 1	Species 2	Species 3	Sp
Sample 1	23	53	2	
Sample 2	69	64	70	
Sample 3	33	100	68	
Sample 4	5	63	57	
Sample 5	76	80	46	
Sample 6	58	7	37	
Sample 7	10	87	32	
Sample 8	21	89	73	

Row Sums are known to be arbitrary

CHALLENGES OF COMPOSITION

	Species 1	Species 2	Species 3	Sp
Sample 1	23	53	2	
Sample 2	69	64	70	
Sample 3	33	100	68	
Sample 4	5	63	57	
Sample 5	76	80	46	
Sample 6	58	7	37	
Sample 7	10	87	32	
Sample 8	21	89	73	

Row Sums are known to be arbitrary

Common practice is to "normalize"
(convert to **percentages** by dividing by row totals)

CHALLENGES OF COMPOSITION

	Species 1	Species 2	Species 3	Sp
Sample 1	23	53	2	
Sample 2	69	64	70	
Sample 3	33	100	68	
Sample 4	5	63	57	
Sample 5	76	80	46	
Sample 6	58	7	37	
Sample 7	10	87	32	
Sample 8	21	89	73	

Row Sums are known to be arbitrary

Common practice is to "normalize"
(convert to **percentages** by dividing by row totals)

Percentages = Relative Abundances = Compositions

CHALLENGES OF COMPOSITION

	Species 1	Species 2	Species 3	Sp
Sample 1	23	53	2	
Sample 2	69	64	70	
Sample 3	33	100	68	
Sample 4	5	63	57	
Sample 5	76	80	46	
Sample 6	58	7	37	
Sample 7	10	87	32	
Sample 8	21	89	73	

Row Sums are known to be arbitrary

Common practice is to "normalize"
(convert to **percentages** by dividing by row totals)

Percentages = Relative Abundances = Compositions

B+L+R=k

And all Positive

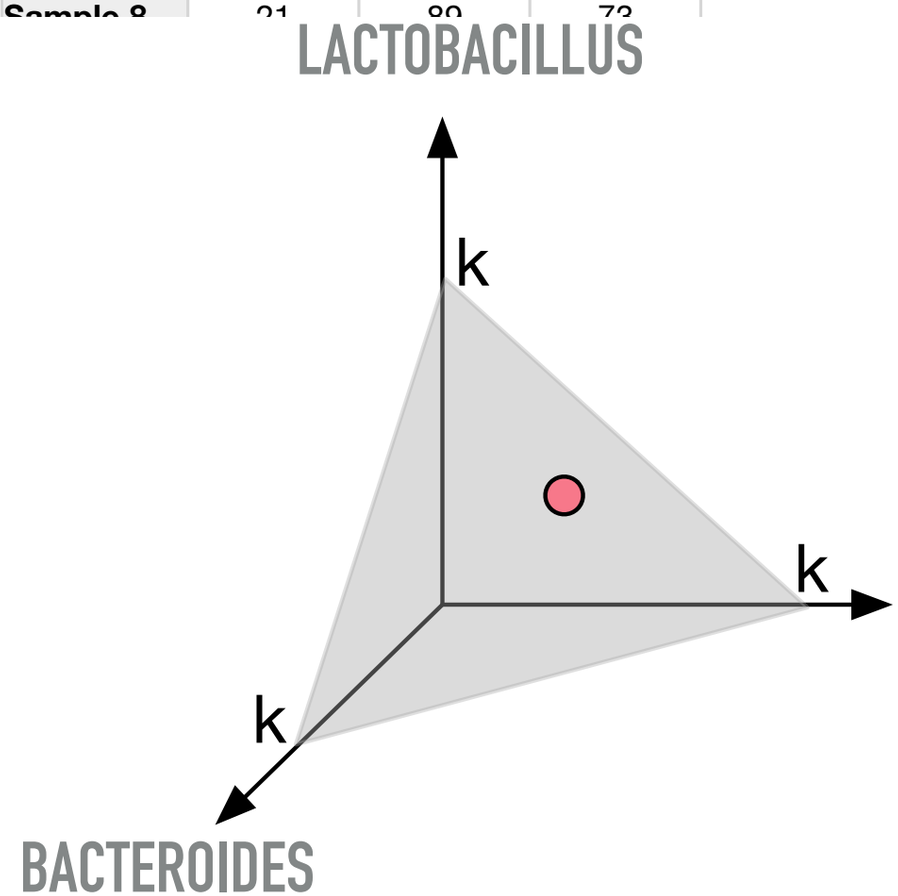
CHALLENGES OF COMPOSITION

	Species 1	Species 2	Species 3	Sp
Sample 1	23	53	2	
Sample 2	69	64	70	
Sample 3	33	100	68	
Sample 4	5	63	57	
Sample 5	76	80	46	
Sample 6	58	7	37	
Sample 7	10	87	32	
Sample 8	21	89	72	

Row Sums are known to be arbitrary

Common practice is to "normalize"
(convert to **percentages** by dividing by row totals)

Percentages = Relative Abundances = Compositions

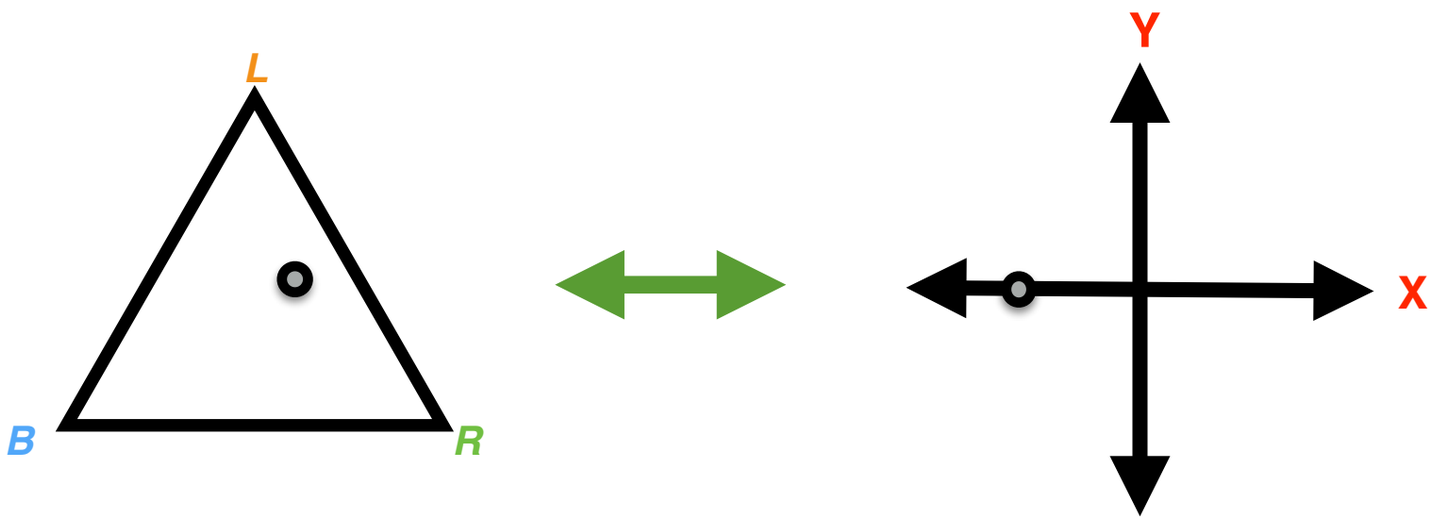


B+L+R=k

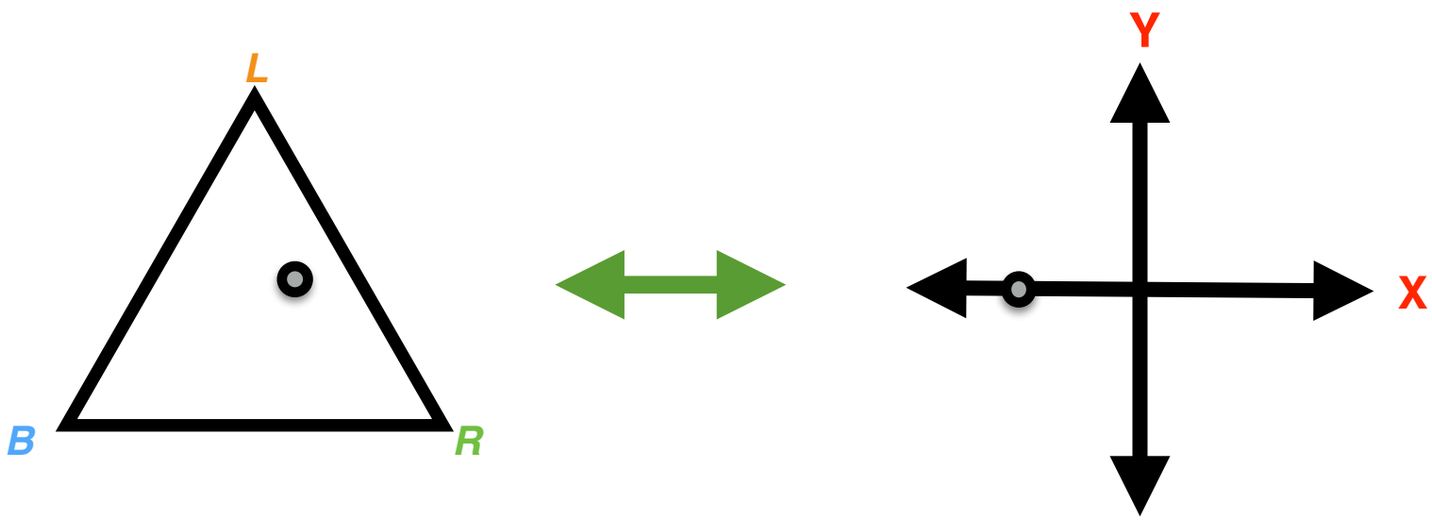
And all Positive

Example of problem: If B goes up, L+R must go down

HOW DO YOU DEAL WITH COMPOSITION?



HOW DO YOU DEAL WITH COMPOSITION?



ALR $$(x, y) = \left(\log \frac{L}{R}, \log \frac{B}{R}\right)$$

CLR $$(x, y, z) = \left(\log \frac{L}{(LBR)^{1/3}}, \log \frac{B}{(LBR)^{1/3}}, \log \frac{R}{(LBR)^{1/3}}\right)$$

A PROBLEM WITH THE COMPOSITIONAL PERSPECTIVE

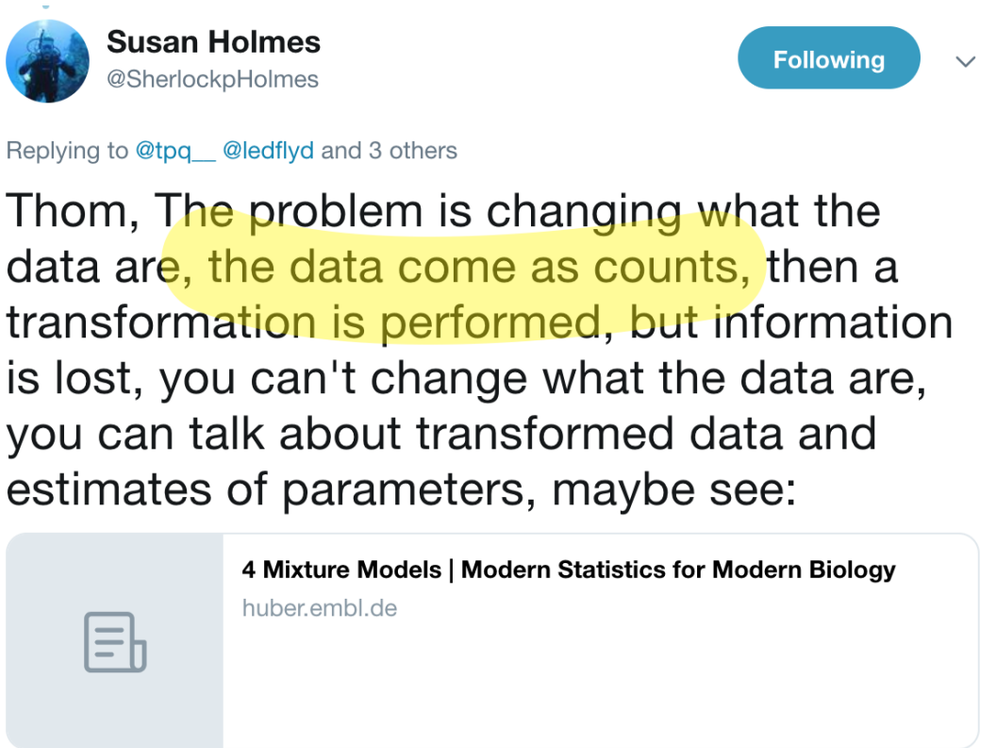
$$\log \frac{0}{x} = -\infty$$

A PROBLEM WITH THE COMPOSITIONAL PERSPECTIVE

$$\log \frac{0}{x} = -\infty$$

$$\log \frac{x}{0} \ldots \text{Oh Shit...}$$

OTHER SIDE OF THE AISLE



Susan Holmes
@SherlockpHolmes

Following

Replying to @tpq__ @ledflyd and 3 others

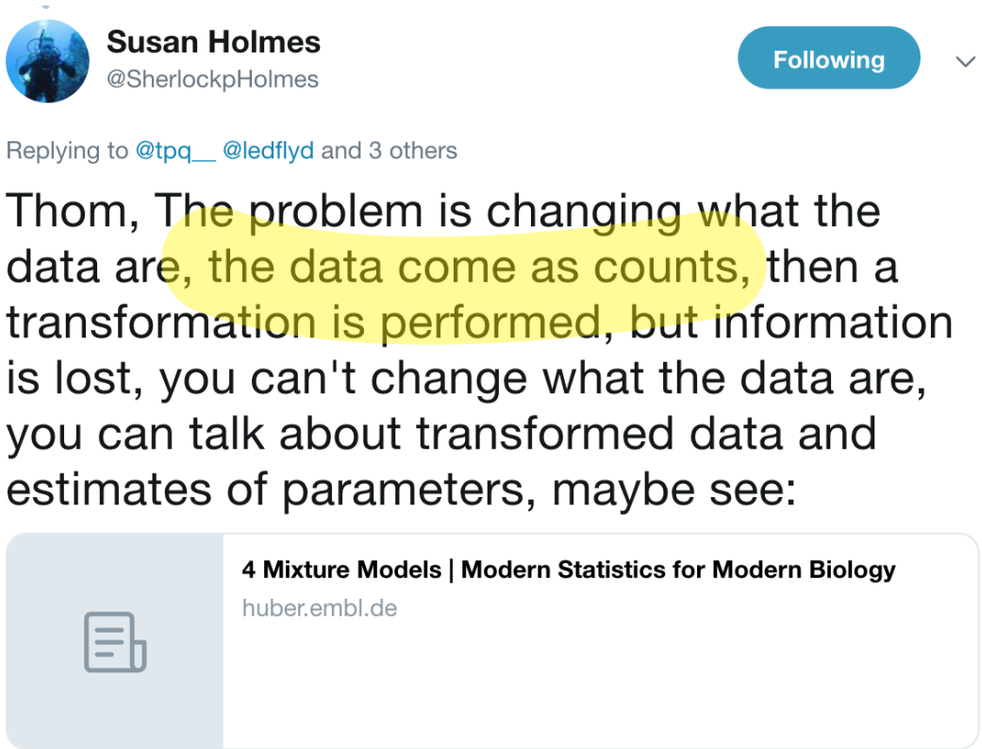
Thom, The problem is changing what the data are, the data come as counts, then a transformation is performed, but information is lost, you can't change what the data are, you can talk about transformed data and estimates of parameters, maybe see:

4 Mixture Models | Modern Statistics for Modern Biology
huber.embl.de

The data is count data

> A zero count can be because a taxa (e.g., species) had low, but non-zero, abundance.

OTHER SIDE OF THE AISLE



Susan Holmes
@SherlockpHolmes

Following

Replying to @tpq__ @ledflyd and 3 others

Thom, The problem is changing what the data are, the data come as counts, then a transformation is performed, but information is lost, you can't change what the data are, you can talk about transformed data and estimates of parameters, maybe see:

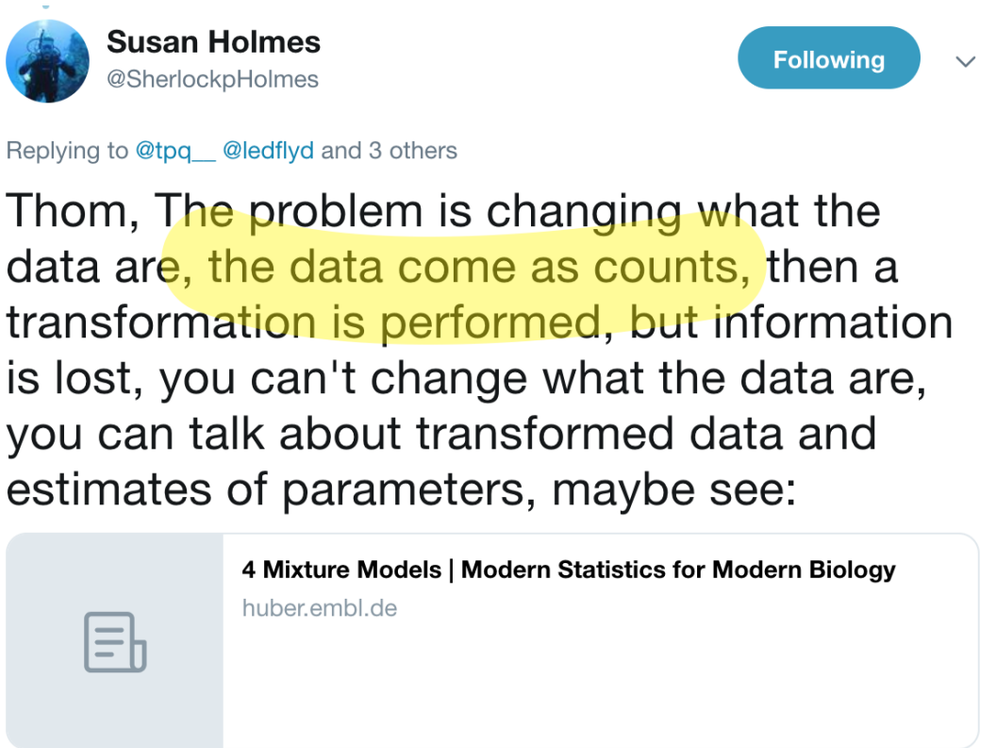
4 Mixture Models | Modern Statistics for Modern Biology
huber.embl.de

The data is count data

A zero count can be because a taxa (e.g., species) had low, but non-zero, abundance.

Model Random Counting
(*e.g.*, negative binomial or Poisson)

OTHER SIDE OF THE AISLE



Susan Holmes
@SherlockpHolmes

Following

Replying to @tpq__ @ledflyd and 3 others

Thom, The problem is changing what the data are, the data come as counts, then a transformation is performed, but information is lost, you can't change what the data are, you can talk about transformed data and estimates of parameters, maybe see:

4 Mixture Models | Modern Statistics for Modern Biology
huber.embl.de

The data is count data

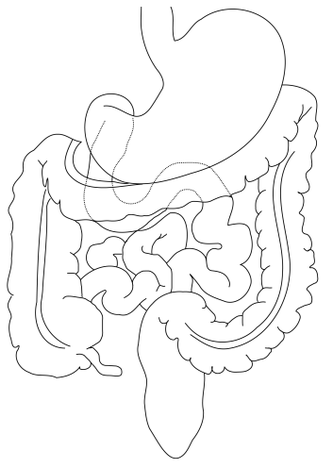
A zero count can be because a taxa (e.g., species) had low, but non-zero, abundance.

Model Random Counting
(*e.g.*, negative binomial or Poisson)

Yet often models each taxa as independent.

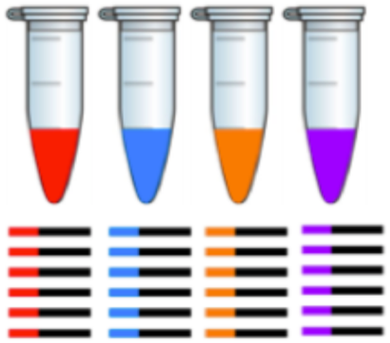
VIEWING AS RANDOM SAMPLING

Sample Collection and Storage



RANDOM SAMPLING

DNA Extraction PCR Amplification



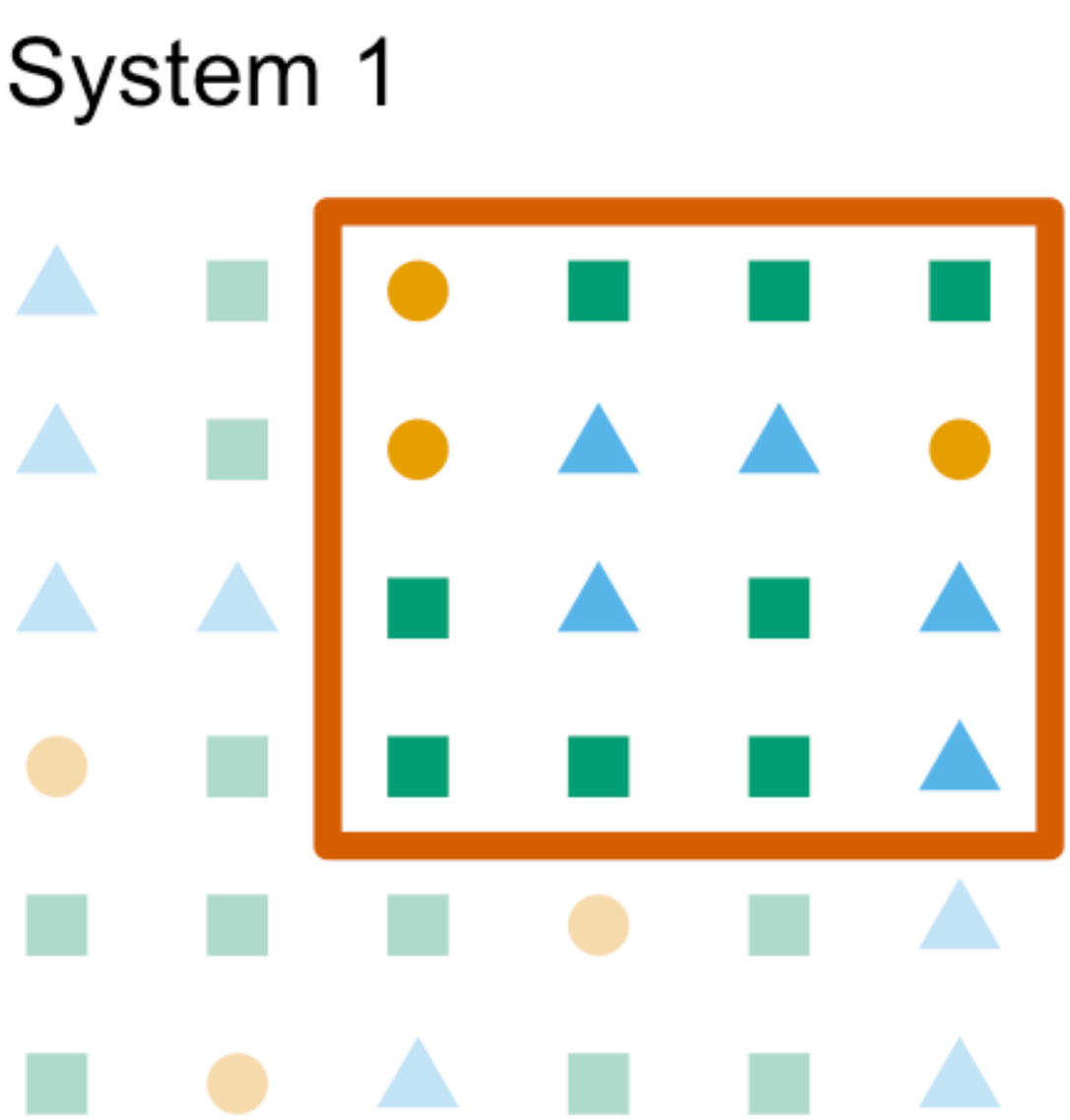
RANDOM SAMPLING

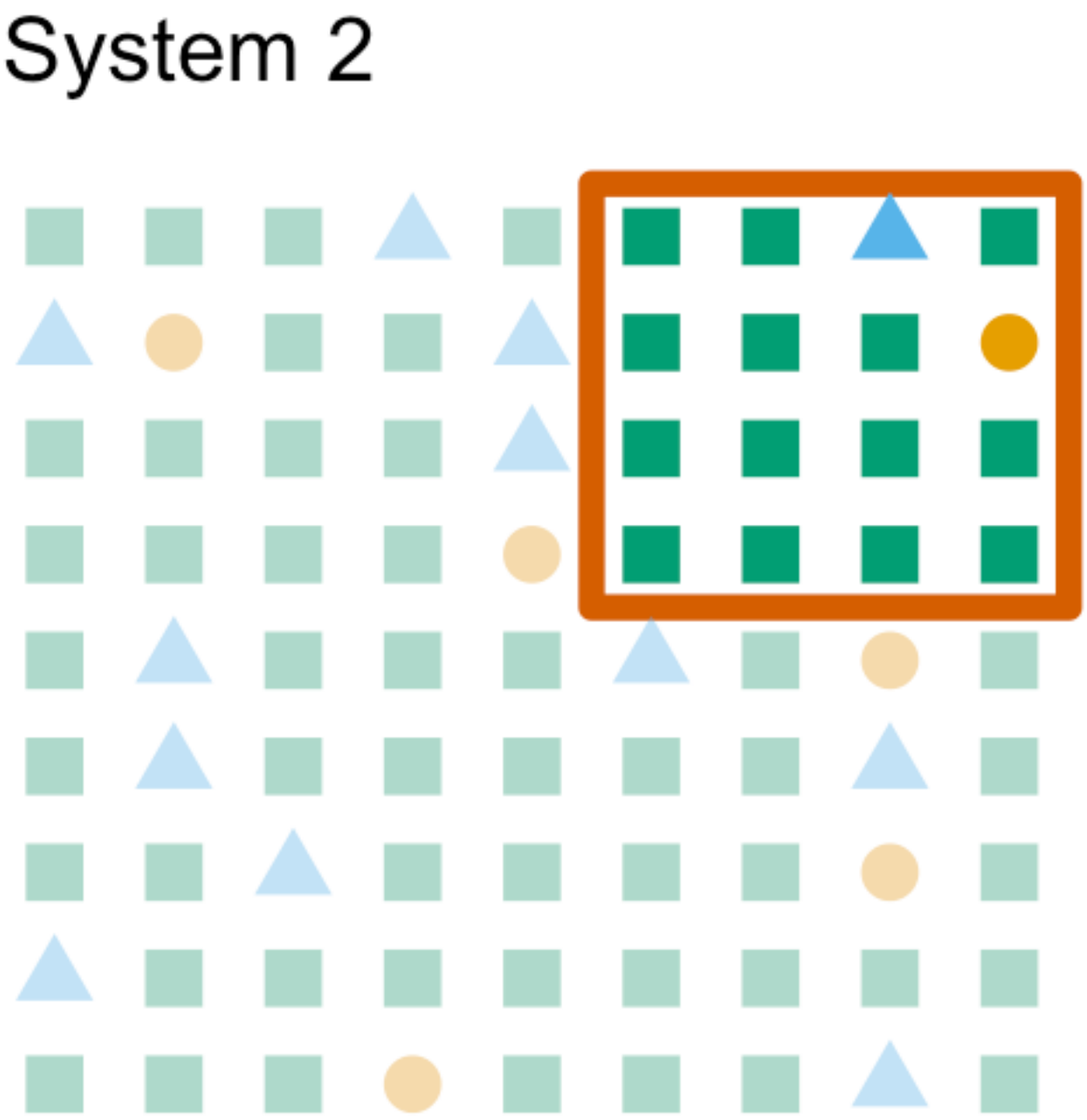
Sequencing

RANDOM SAMPLING

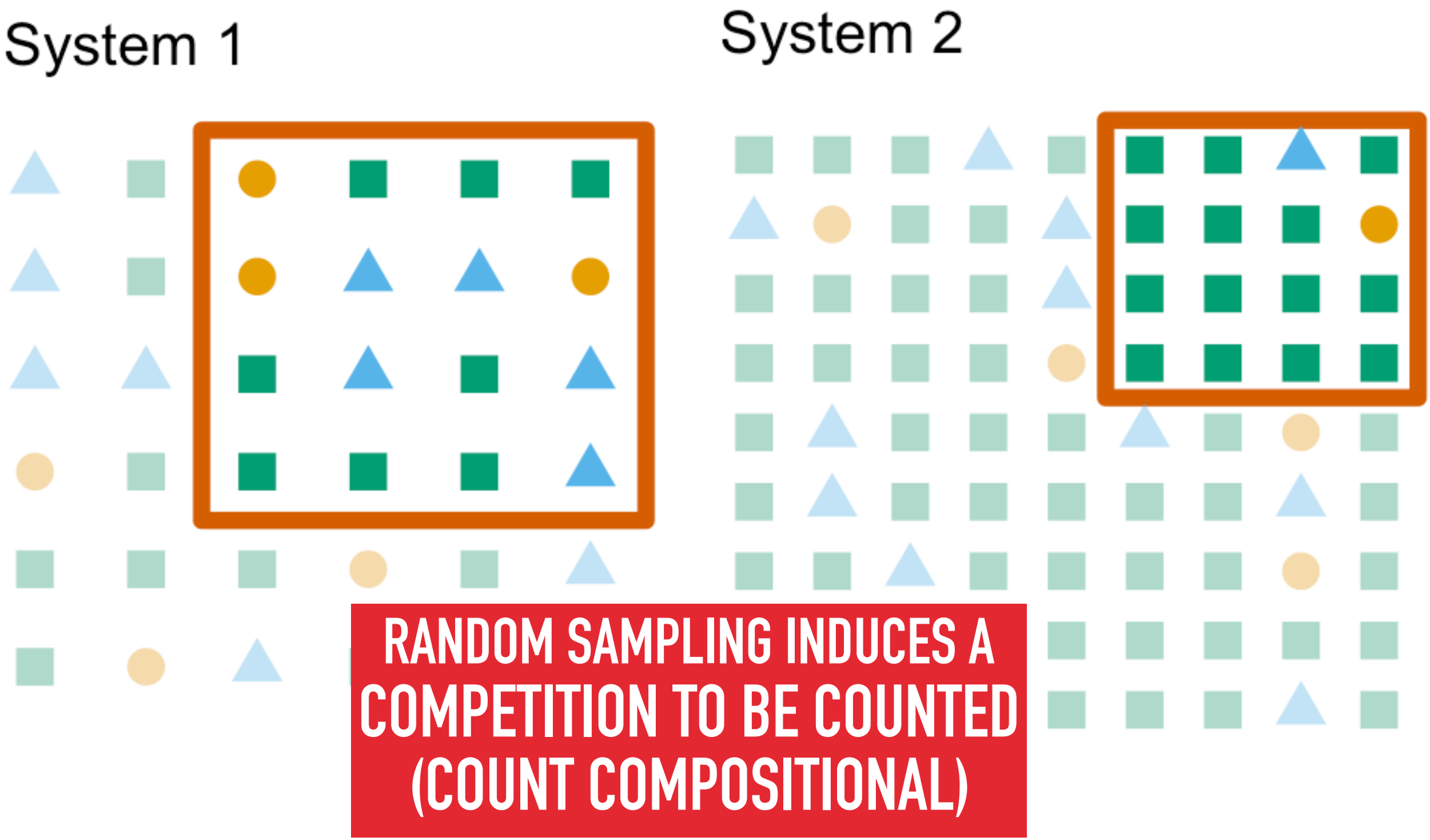
Adapted from Hamady. et al., *Nature Methods*, 2008

PROBLEM WITH MULTIVARIATE RANDOM SUBSAMPLING





PROBLEM WITH MULTIVARIATE RANDOM SUBSAMPLING

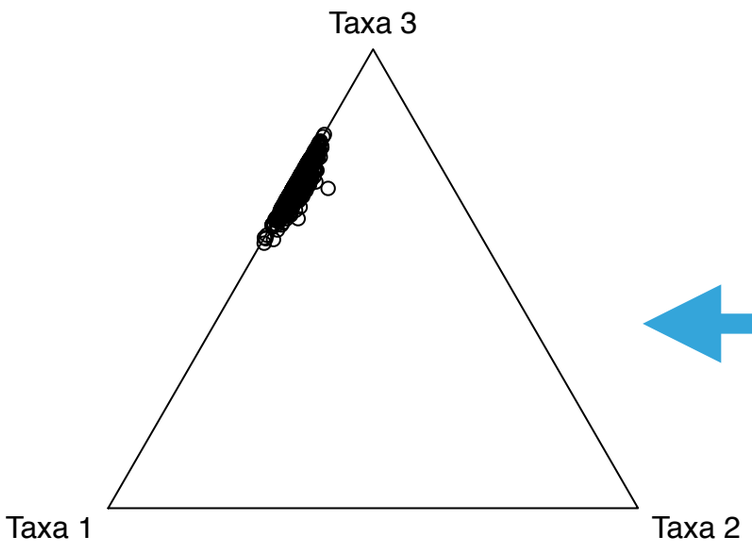


EXTRACTING MORE INFORMATION FROM COUNTS

Samples	1	2
Taxa 1	40	0
Taxa 2	0	0
Taxa 3	100	1

BAYESIAN MULTINOMIAL MODELS REFLECT INTUITION WE WANT

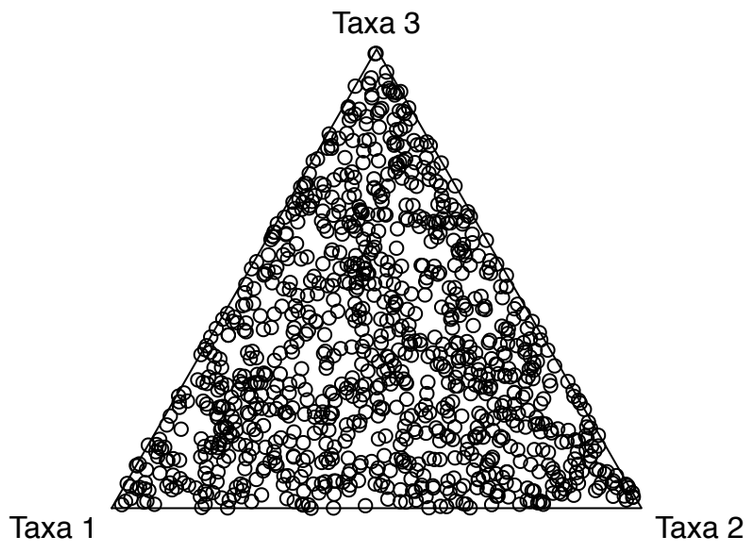
Posterior

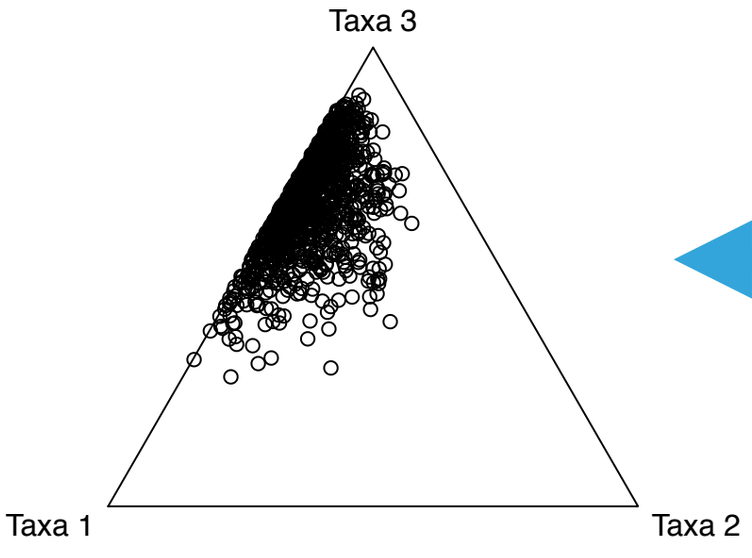


Data

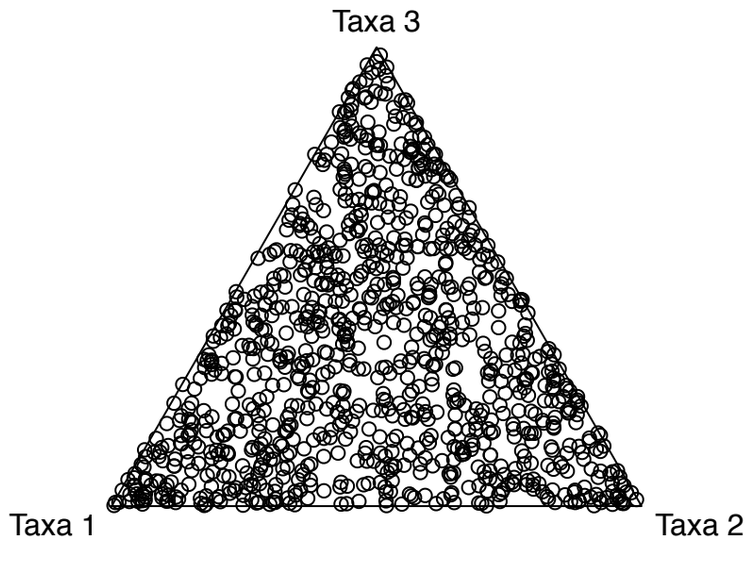
Samples	1
Taxa 1	40
Taxa 2	0
Taxa 3	100

Prior





Samples	1
Taxa 1	4
Taxa 2	0
Taxa 3	10



MULTINOMIAL-LOGISTIC NORMAL

$$Y \sim \text{Multinomial}(\pi)$$
$$\pi \sim \text{Logistic Normal}(\rho, \Xi)$$

$$Y \sim \text{Multinomial}(\pi)$$
$$\pi = \text{ILR}^{-1}(\eta)$$
$$\eta \sim \text{Multivariate Normal}(\mu, \Sigma)$$

ILR = "Isometric Log-Ratio" Transform

- Handles Zeros and Competition-to-be-counted
- Allows positive and negative covariation between taxa
- Models Multiplicative Errors

MODELING TIME-EVOLUTION

Observed Counts

↑

Addition of Technical Noise

↑

True State with Biological Noise

Priors

$$Y_t \sim \text{Multinomial}(\pi_t)$$

$$\pi_t = \text{ILR}^{-1}(\eta_t)$$

$$\eta_t = F_t'\theta_t + v_t \qquad v_t \sim N(0, V_t)$$

$$\theta_t = G_t\theta_{t-1} + \omega_t \qquad \omega_t \sim N(0, W_t)$$

$$\theta_0 \sim N(m_0, C_0)$$

$$V_1, \ldots, V_T, W_1, \ldots, W_T \sim p(\xi)$$

Silverman et al., Microbiome 2018

INFERENCE

THE COMPUTATIONAL BOTTLENECK

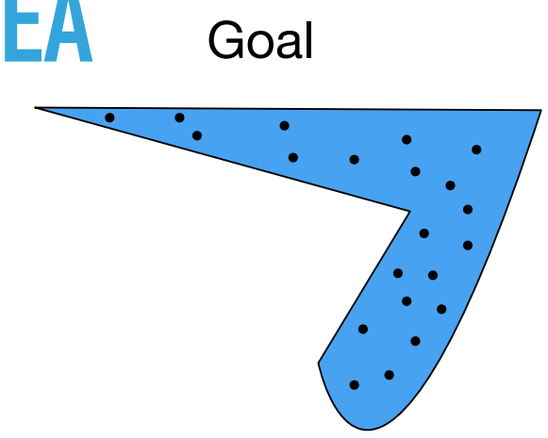
10 Taxa with **650 Samples**

As measured by Time to Effective Sample size of 2000

- Metropolis-within-Gibbs → >2 months
- Now on order of **milliseconds to seconds.**

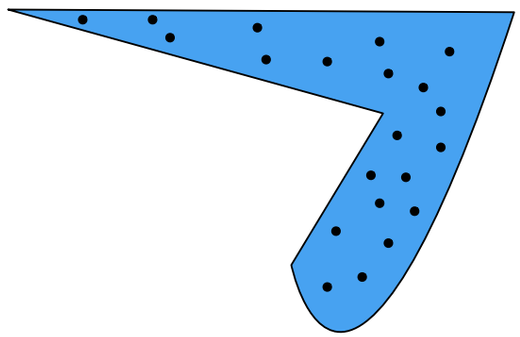
Can even scale to **5K x 20K** , ~ 1.4 days run-time

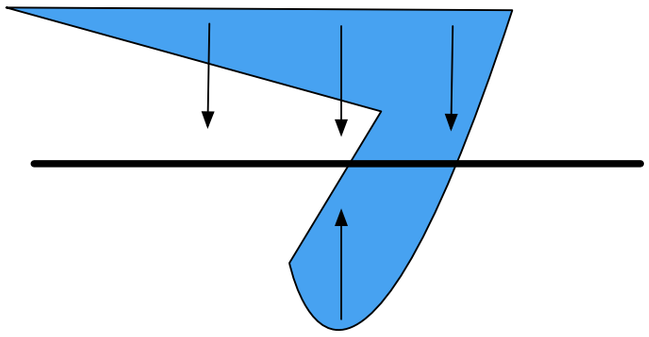
KEY IDEA



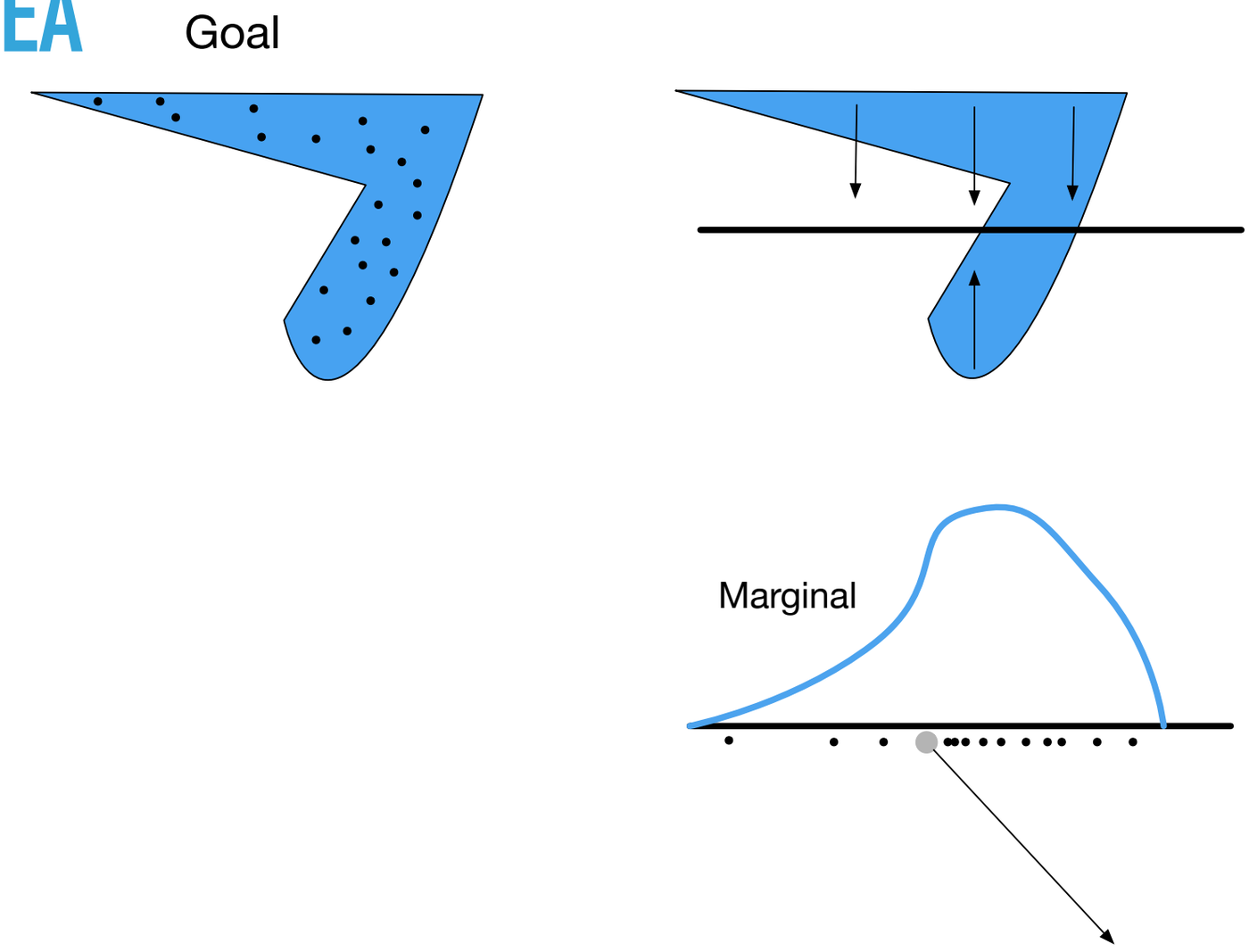
KEY IDEA

Goal

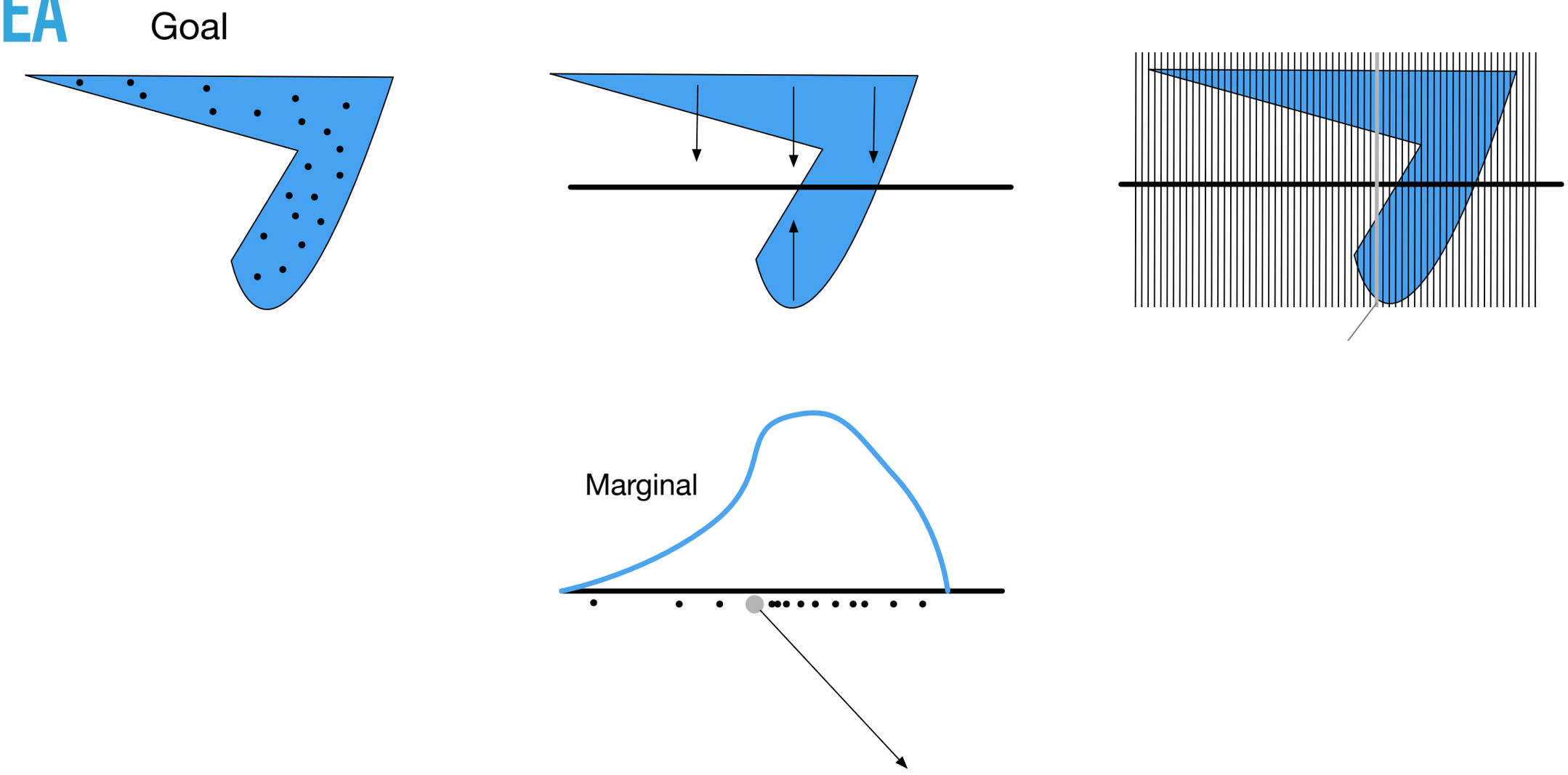




KEY IDEA

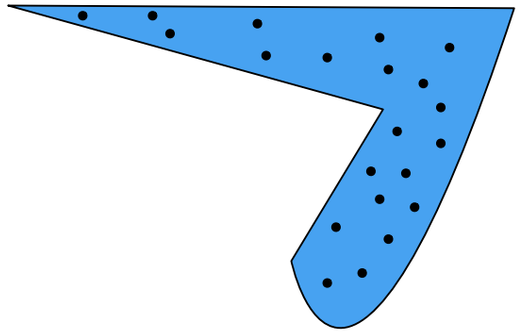


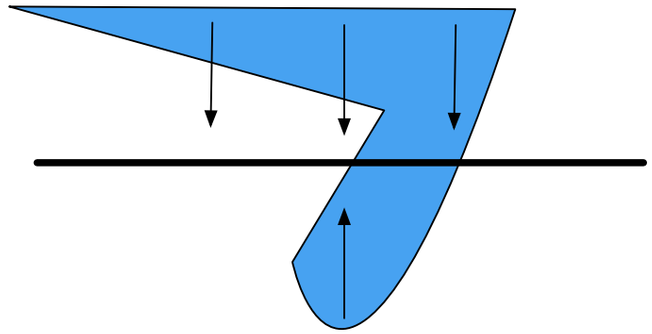
KEY IDEA

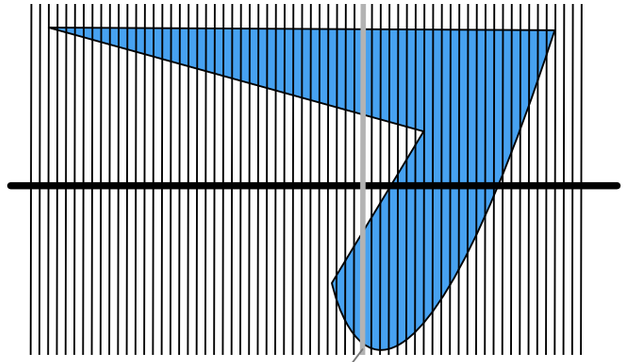


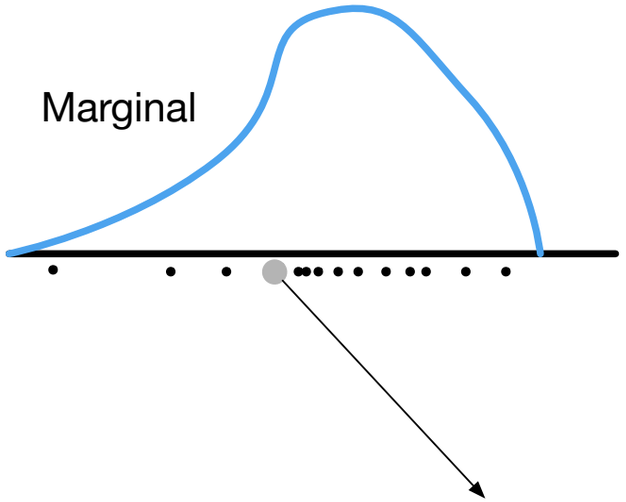
KEY IDEA

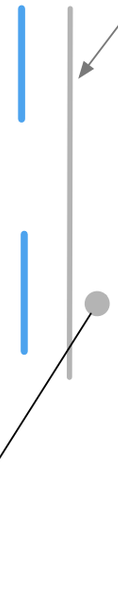
Goal



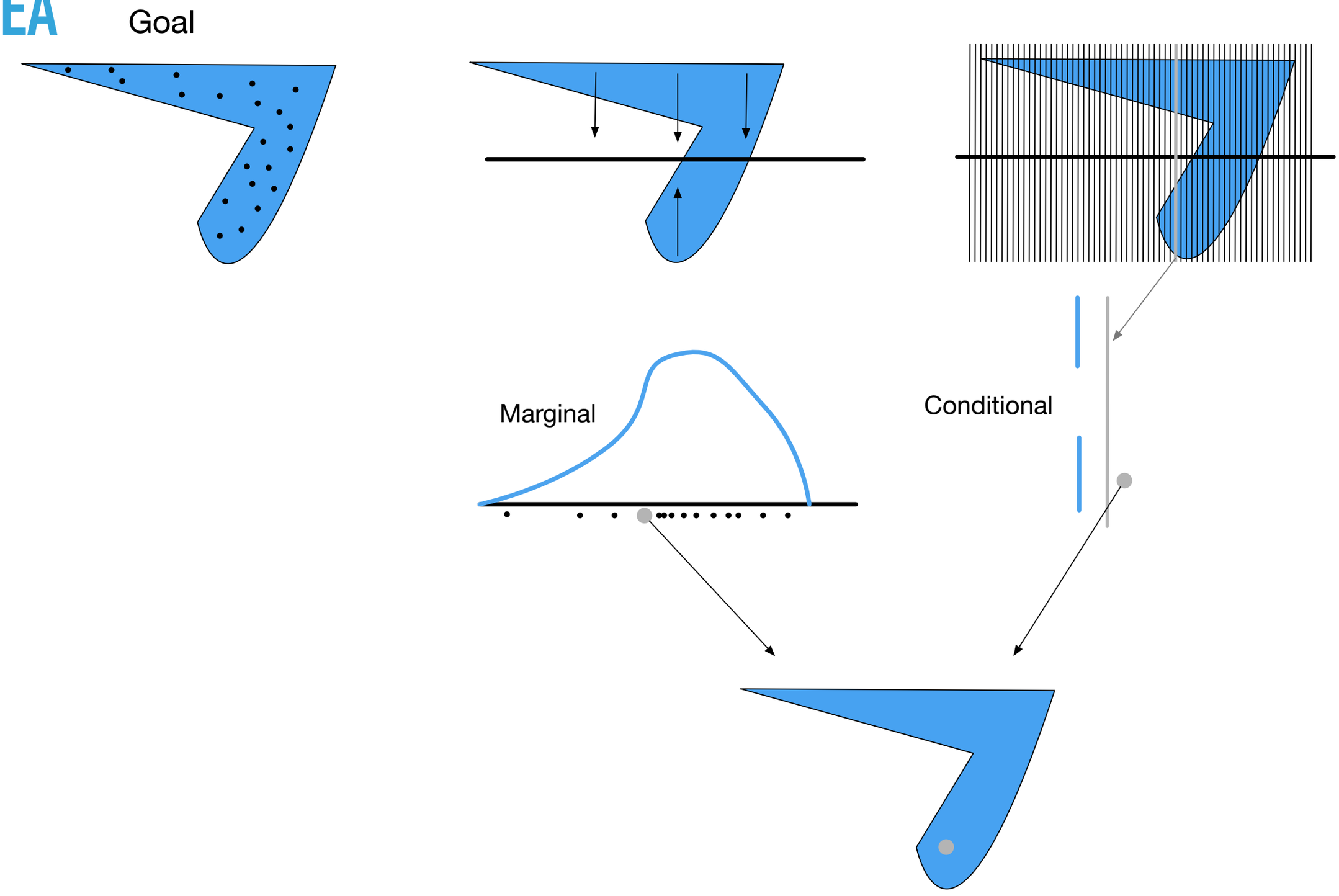






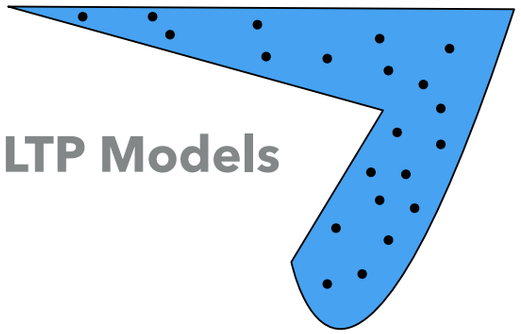


KEY IDEA

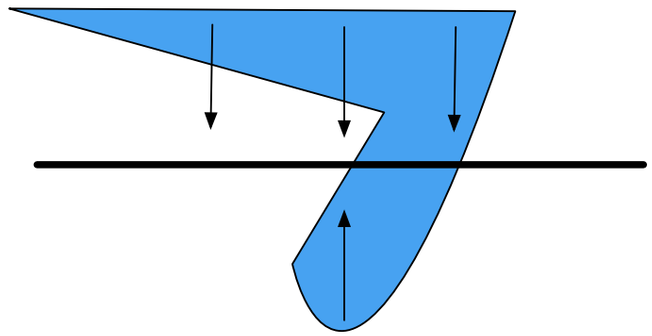


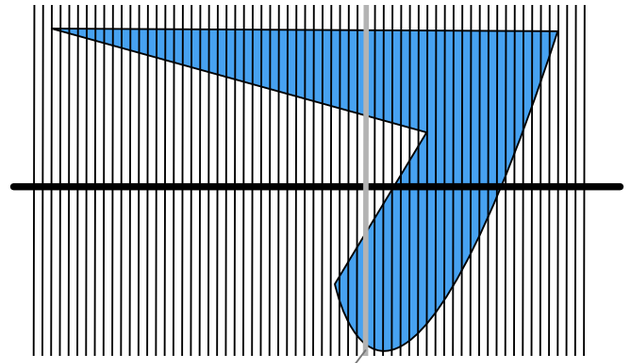
KEY IDEA

Goal

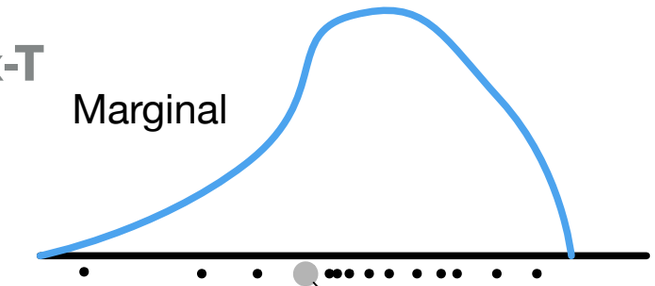


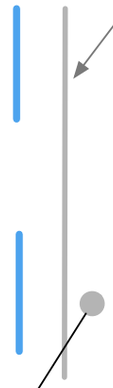
Marginally LTP Models

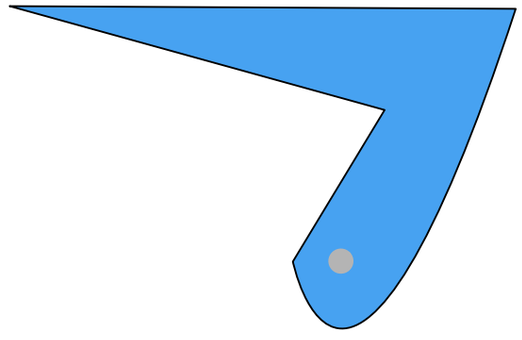




Latent Matrix-T Process (LTP)







(1) I have found that a huge class of models have identical marginal forms

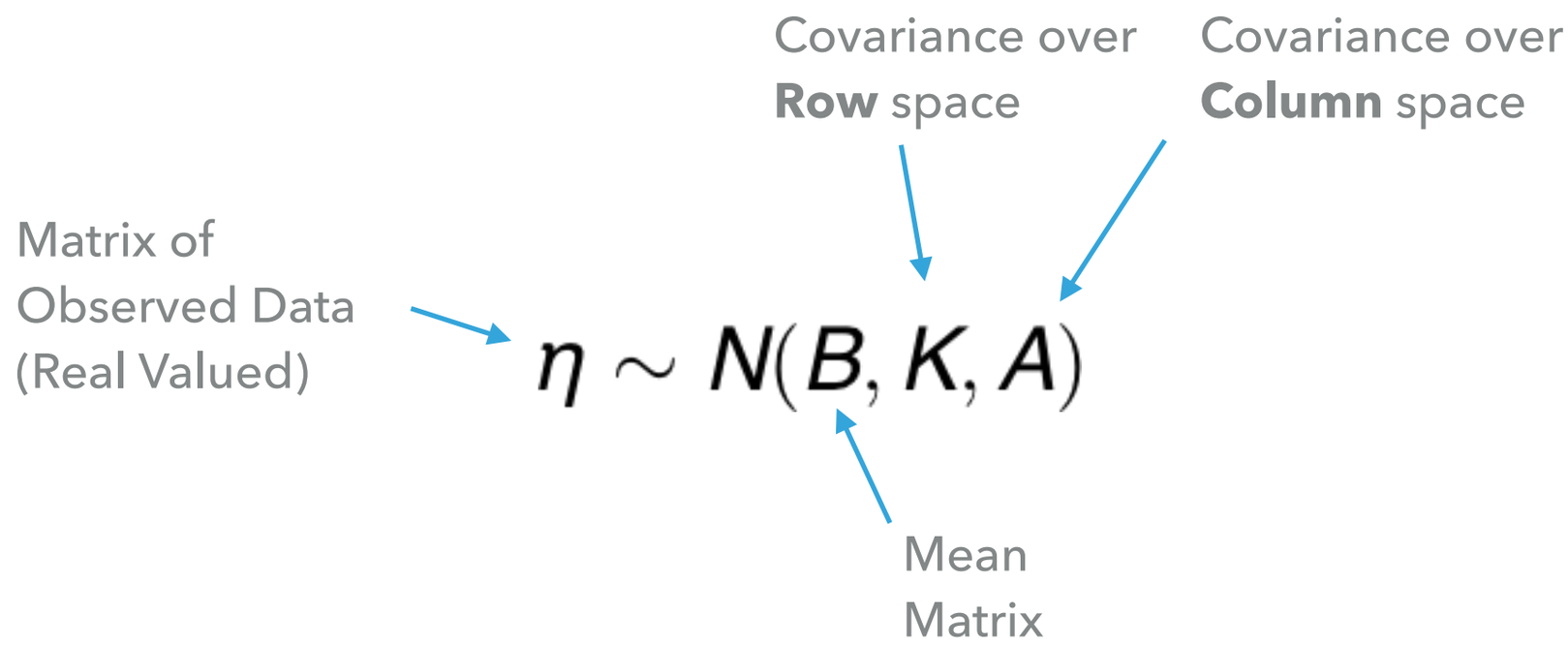
(2) I have found a highly accurate approximation for this marginal form

(3) These models often have conditionals that are easy to sample from.

MARGINALLY LATENT MATRIX-T PROCESSES MODELS

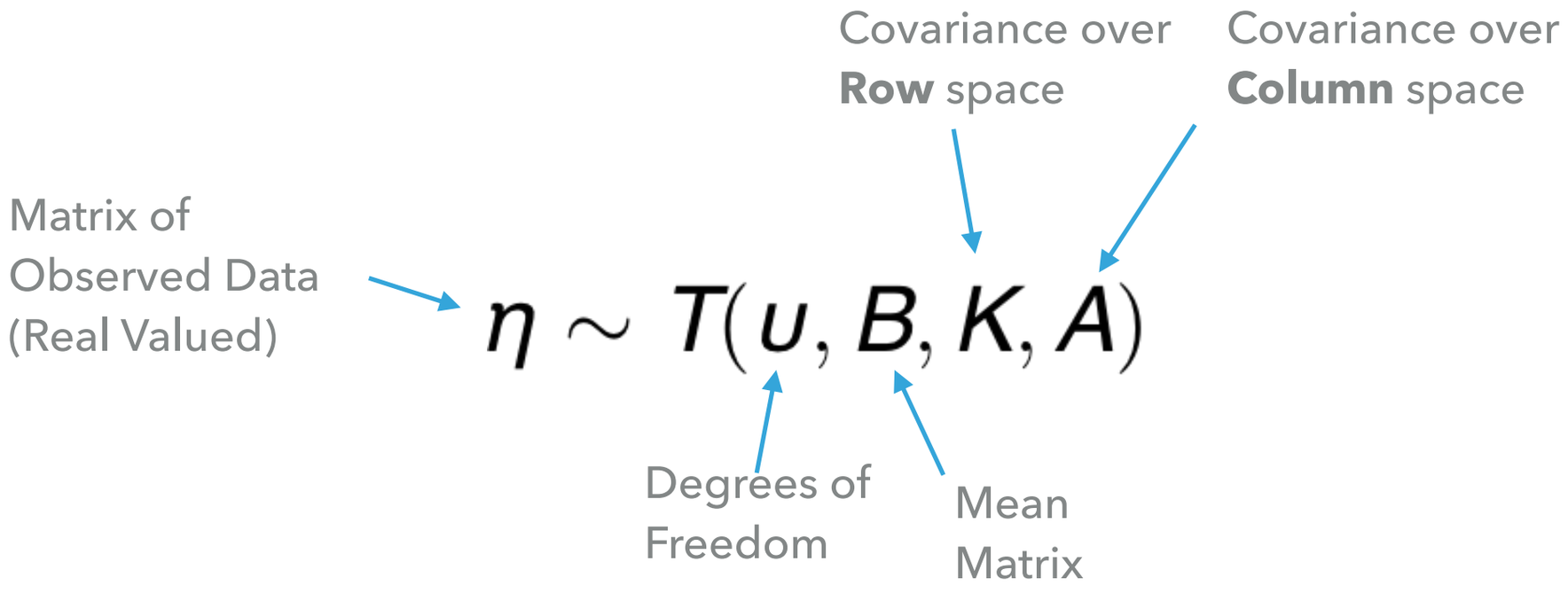
MARGINALLY LTP MODELS

MATRIX NORMAL PROCESS



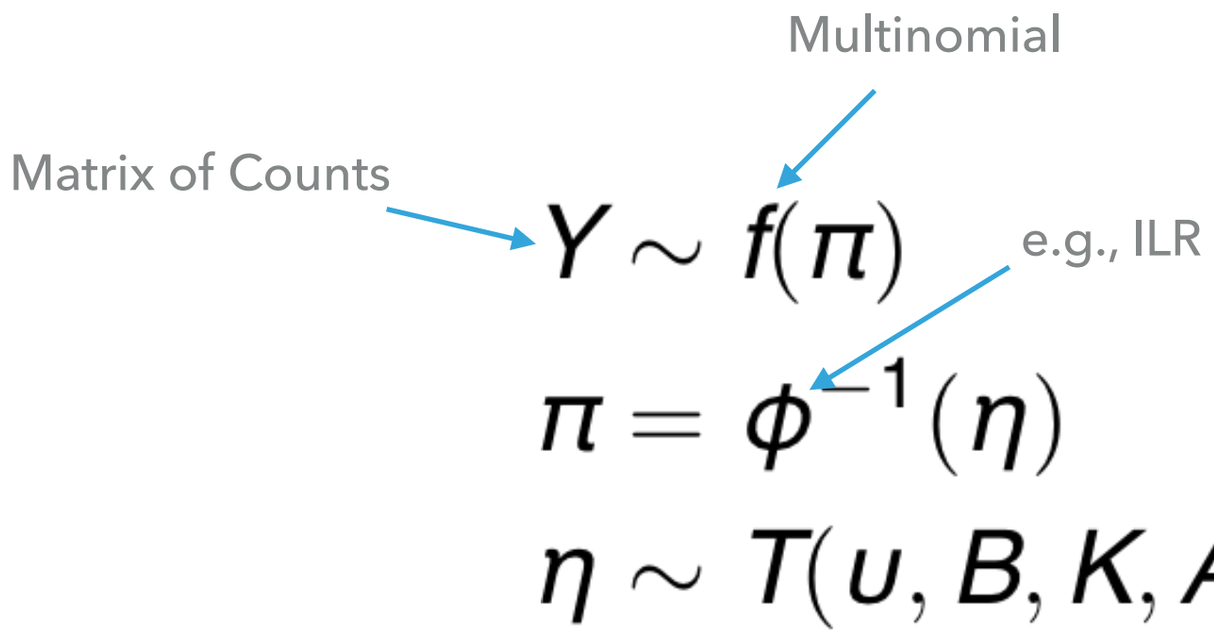
Silverman JD, Roche K, et al. 2019. *arXiv*

MATRIX T-PROCESS



Silverman JD, Roche K, et al. 2019. *arXiv*

LATENT MATRIX-T PROCESS (LTP)



$$Y \sim f(\pi)$$

$$\pi = \phi^{-1}(\eta)$$

$$\eta \sim T(u, B, K, A)$$

Silverman JD, Roche K, et al. 2019. *arXiv*

AN EXAMPLE OF A MARGINALLY LTP MODEL

Multinomial Logistic Normal Process

$$Y_t \sim \text{Multinomial}(\pi_t)$$ Count Noise

$$\pi_t = \text{ILR}^{-1}(\eta_t)$$

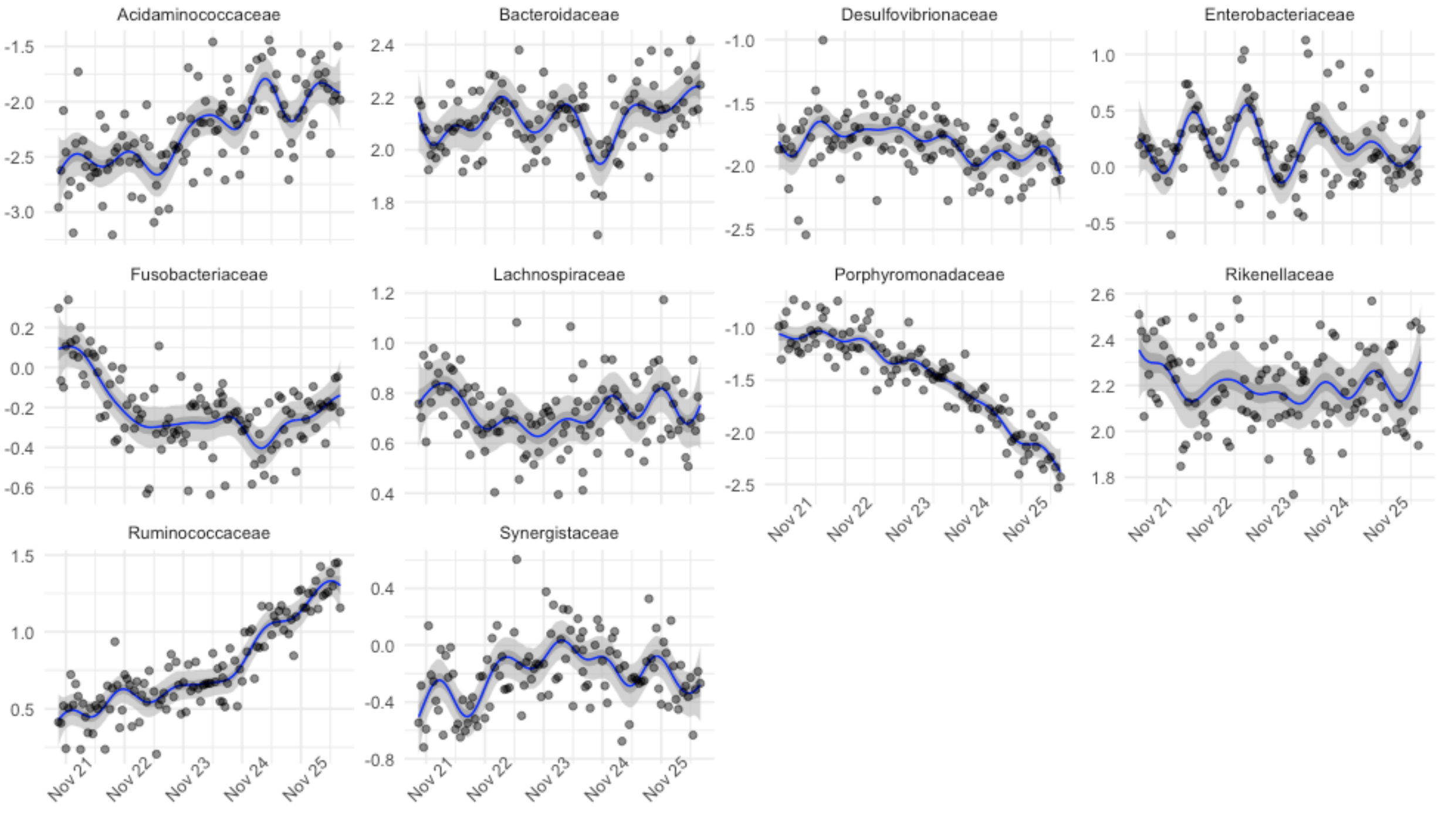
$$\eta_t \sim N(M_t, \Sigma)$$ Additional Noise

$$M \sim N(0, \Sigma, \Gamma)$$

$$\Gamma_{t,s} = \text{RBF(t,s)}$$ Smoothed State

$$\Sigma \sim IW(\Xi, \upsilon)$$ Unknown Covariance Between Log-Ratios

FOR TIME-SERIES ANALYSIS



A FEW MORE EXAMPLES

Generalized Multivariate
Dynamic Linear Models

$$
\begin{aligned}
Y &\sim f(\pi) \\
\pi &= \phi^{-1}(\eta) \\
\eta_t^T &= F_t^T \Theta_t + v_t^T, \quad v_t \sim N(0, \gamma_t \Sigma) \\
\Theta_t &= G_t \Theta_{t-1} + \Omega_t, \quad \Omega_t \sim N(0, W_t, \Sigma) \\
\Theta_0 &\sim N(M_0, C_0, \Sigma) \\
\Sigma &\sim IW(\Xi, u)
\end{aligned}
$$

Generalized Multivariate
Conjugate Linear Models

$$
\begin{aligned}
Y &\sim f(\pi) \\
\pi &= \phi^{-1}(\eta) \\
\eta_{\cdot j} &\sim N(\Lambda X_{\cdot j}, \Sigma) \\
\Lambda &\sim N(\Theta, \Sigma, \Gamma) \\
\Sigma &\sim IW(\Xi, u)
\end{aligned}
$$

And Many More ...

Silverman JD, Roche K, et al. 2019. *arXiv*

MULTINOMIAL LOGISTIC NORMAL MODELS WITH MARGINAL LAPLACE APPROXIMATION

C++, Eigen (+MKL)
R Interface using Rcpp

Extensively Unit Tested against
Independent Implementations

Silverman JD, Roche K, et al. 2019. *arXiv*

MULTINOMIAL LOGISTIC NORMAL MODELS WITH MARGINAL LAPLACE APPROXIMATION

Justin

Tukey

Gauss

C++, Eigen (+MKL)
R Interface using Rcpp

Extensively Unit Tested against Independent Implementations

Silverman JD, Roche K, et al. 2019. *arXiv*

MULTINOMIAL LOGISTIC NORMAL MODELS – BUT FAST

Benchmarking - Kim Roche

Silverman JD, Roche K, et al. 2019. *arXiv*

MULTINOMIAL LOGISTIC NORMAL MODELS – BUT FAST

Benchmarking - Kim Roche

Efficient

~ 5 orders of magnitude faster than HMC

~ 1-2 orders of magnitude faster than Variational Bayes (VB)

Silverman JD, Roche K, et al. 2019. *arXiv*

MULTINOMIAL LOGISTIC NORMAL MODELS – BUT FAST

Benchmarking - Kim Roche

Efficient

~ 5 orders of magnitude faster than HMC

~ 1-2 orders of magnitude faster than Variational Bayes (VB)

Accurate

- Point Estimation Accuracy (estimating posterior mean) is nearly perfect over all tested conditions (in contrast VB breaks down when many taxa)
- Uncertainty quantification (estimating posterior variance) only found to break down when > 93% zeros in dataset. (in contrast VB breaks down often)

Silverman JD, Roche K, et al. 2019. *arXiv*

STRAY

Public on GitHub

Many many different
multinomial logistic-normal
models scalable and accurately.

arXiv.org > stat > arXiv:1903.11695

Statistics > Methodology

Bayesian Multinomial Logistic Normal Models through Marginally Latent Matrix-T Processes

Justin D. Silverman, Kimberly Roche, Zachary C. Holmes, Lawrence A. David, Sayan Mukherjee

(Submitted on 27 Mar 2019 (v1), last revised 1 Apr 2019 (this version, v3))

Bayesian multinomial logistic-normal (MLN) models are popular for the analysis of sequence count data (e.g., microbiome or gene expression data) due to th
complex covariance structure. However, existing implementations of MLN models are limited to handling small data sets due to the non-conjugacy of the mu
introduce MLN models which can be written as marginally latent matrix-t process (LTP) models. Marginally LTP models describe a flexible class of generalize
series models. We develop inference schemes for Marginally LTP models and, through application to MLN models, demonstrate that our inference schemes a
magnitude faster than MCMC.

ACKNOWLEDGEMENTS

StatsAtHome.com

inschool4life

Wife and Collaborator (MERCK Biostatistics)
Rachel Silverman

University of Montana
Alex Washburne

NYU
Jamie Morton

UCLA
Liat Shenhav
Eran Halperin

Duke University
Lawrence David
Sayan Mukherjee
Kim Roche
Rachael Bloom
Heather Durand
Sharon Jiang
Brianna Petrone
Zach Holmes
Jeff Letourneau
Max Villa
Kevin Zhu
Eric Dallow

U. de Girona
Vera Pawlowsky-Glahn

U. de Catalunya Polytechnic
Juan Jose Egozcue

University of Western Ontario
Greg Gloor

University of Notre Dame
Johannes R Björk
Elizabeth Archie

BUT WHAT ABOUT THE CONDITIONALS?

Generalized Multivariate Conjugate Linear Models

$$Y \sim f(\pi)$$
$$\pi = \phi^{-1}(\eta)$$
$$\eta_{\cdot j} \sim N(\Lambda X_{\cdot j}, \Sigma)$$
$$\Lambda \sim N(\Theta, \Sigma, \Gamma)$$
$$\Sigma \sim IW(\Xi, \upsilon)$$

This is just the Solution to Bayesian Multivariate Linear Regression

$$\begin{aligned}
\upsilon_N &= \upsilon + N \\
\Gamma_N &= (XX^T + \Gamma^{-1})^{-1} \\
\Lambda_N &= (\eta X^T + \Theta\Gamma^{-1})\Gamma_N \\
\Xi_N &= \Xi + (\eta - \Lambda_N X)(\eta - \Lambda_N X)^T + (\Lambda_N - \Theta)\Gamma^{-1}(\Lambda_N - \Theta)^T \\
p(\Sigma|\eta, X) &= IW(\Xi_N, \upsilon_N) \\
p(\Lambda|\Sigma, \eta, X) &= N(\Lambda_N, \Sigma, \Gamma_N).
\end{aligned}$$

BUT WHAT ABOUT THE CONDITIONALS?

Generalized Multivariate
Dynamic Linear Models

$$
\begin{aligned}
Y &\sim f(\pi) \\
\pi &= \phi^{-1}(\eta) \\
\eta_t^T &= F_t^T \Theta_t + v_t^T, \quad v_t \sim N(0, \gamma_t \Sigma) \\
\Theta_t &= G_t \Theta_{t-1} + \Omega_t, \quad \Omega_t \sim N(0, W_t, \Sigma) \\
\Theta_0 &\sim N(M_0, C_0, \Sigma) \\
\Sigma &\sim IW(\Xi, u)
\end{aligned}
$$

B.2.1 Filtering Recursions for MDLM Model

(1) Posteriors at $t-1$:

$$
\begin{aligned}
p(\Sigma|H_{t-1}^T) &\sim IW(\Xi_{t-1}, \upsilon_{t-1}) \\
p(\Theta_{t-1}|\Sigma, H_{t-1}^T) &\sim N(M_{t-1}, C_{t-1}, \Sigma)
\end{aligned}
$$

(2) Priors at t:

$$
\begin{aligned}
a_t &= G_t m_{t-1} \\
R_t &= G_t C_{t-1} G_t^T + W_t \\
p(\Sigma|H_{t-1}^T) &\sim IW(\Xi_{t-1}, \upsilon_{t-1}) \\
p(\Theta_{t-1}|\Sigma, H_{t-1}^T) &\sim N(a_t, R_t, \Sigma)
\end{aligned}
$$

(3) One-step ahead forecast at t:

$$
\begin{aligned}
f_t^T &= F_t^T a_t \\
q_t &= \gamma_t + F_t^T R_t F_t \\
p(\Sigma|H_{t-1}^T) &\sim IW(\Xi_{t-1}, \upsilon_{t-1}) \\
p(\Theta_{t-1}|\Sigma, H_{t-1}^T) &\sim N(f_t, q_t \Sigma)
\end{aligned}
$$

(4) Posterior at t:

$$
\begin{aligned}
e_t &= \eta_t^T - f_t^T \\
S_t &= \frac{R_t F_t}{q_t} \\
m_t &= a_t + S_t e_t^T \\
C_t &= R_t - q_t S_t S_t^T \\
\upsilon_t &= \upsilon_{t-1} + 1 \\
\Xi_t &= \frac{1}{\upsilon_t}\left[\upsilon_{t-1}\Xi_{t-1} + \frac{e_t e_t^T}{q_t}\right] \\
p(\Sigma|H_{t-1}^T) &\sim IW(\Xi_t, \upsilon_t) \\
p(\Theta_{t-1}|\Sigma, H_{t-1}^T) &\sim N(m_t, C_t, \Sigma)
\end{aligned}
$$

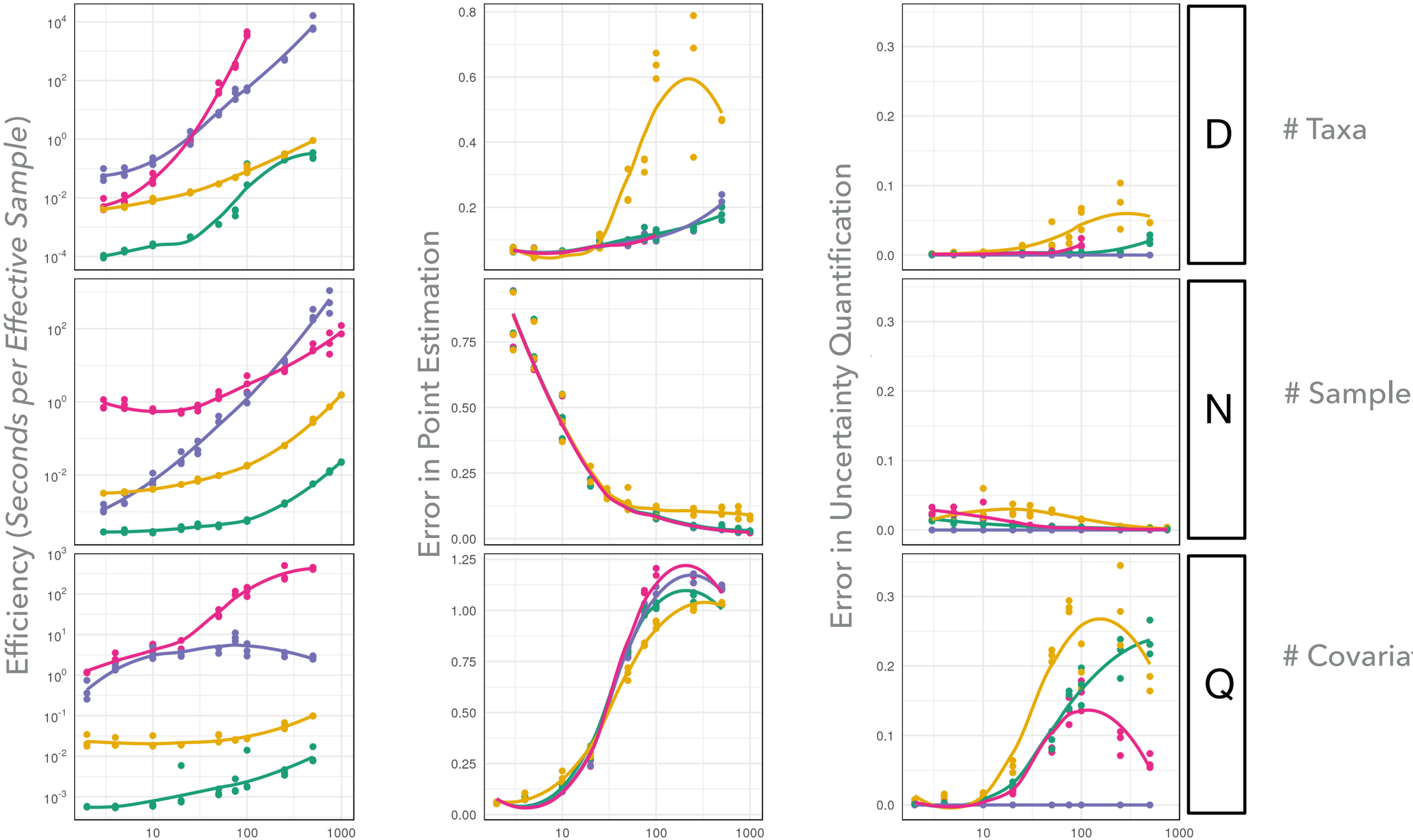
B.2.2 Simulation Smoothing Recursion

The recursions provided here follow directly from Prado and West [39, p. 268]
(1) Sample $\Sigma \sim IW(\Xi_T, \upsilon_T)$ and then $\Theta_T \sim N(M_t, C_t, \Sigma)$.
(2) For each time t from $T-1$ to 0, sample $p(\Theta_t|\Theta_{t+1}, H_T^T) \sim N(M_t^*, C_t^*, \Sigma)$ where

$$
\begin{aligned}
Z_t &= C_t G_{t+1}^T R_{t+1}^{-1} \\
M_t^* &= M_t + Z_t(\theta_{t+1} - a_{t+1}) \\
C_t^* &= C_t - Z_t R_{t+1} Z_t^T.
\end{aligned}
$$

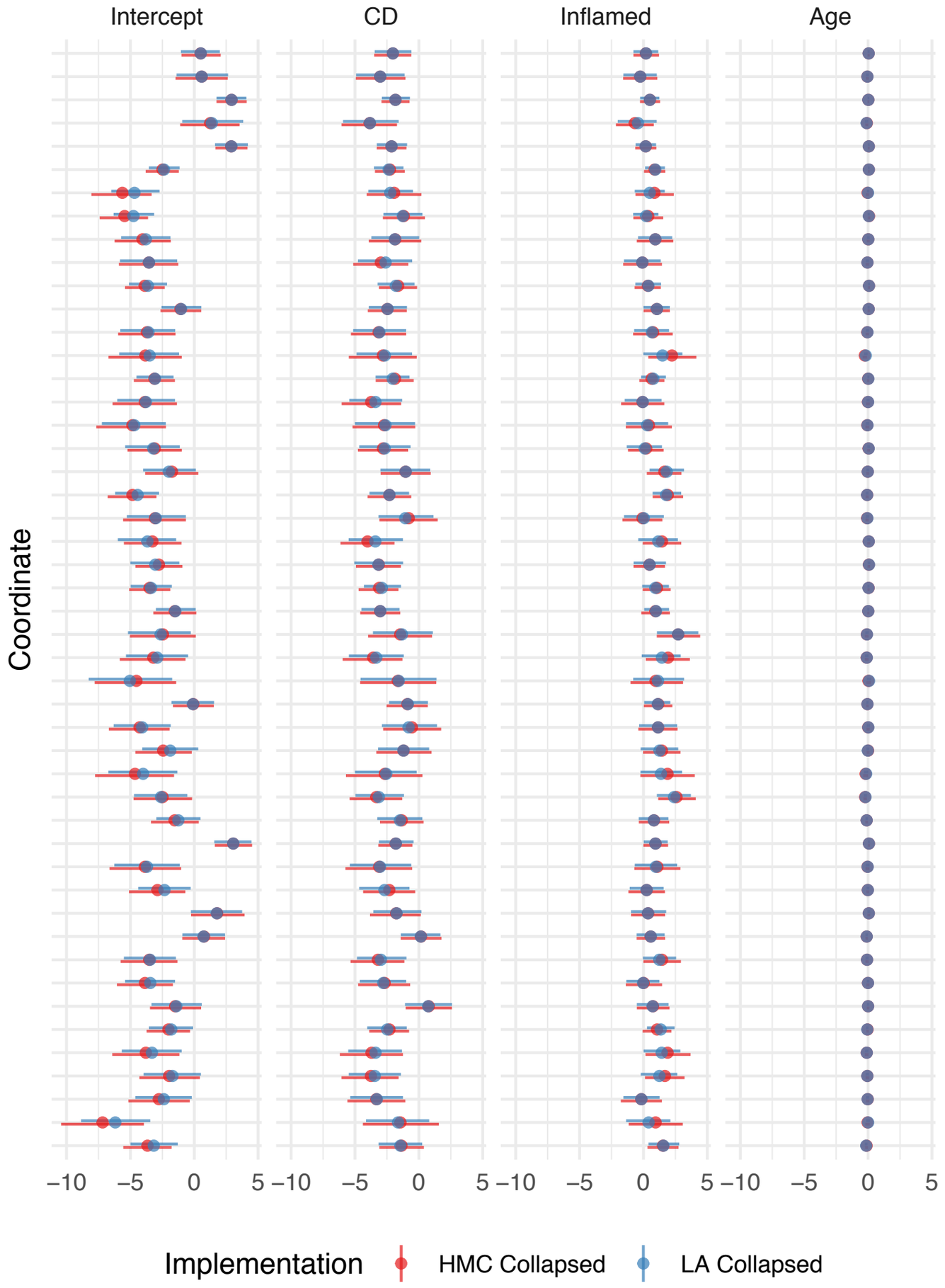
HMC Uncollapsed HMC Collapsed VB Collapsed LA Collapsed (stray)

BENCHMARKING RESULTS



STRAY

REAL DATA



NON-LINEAR TIME-SERIES MODEL FOR MICROBIOME

$$Y \sim f(\pi)$$

$$\pi = \phi^{-1}(\eta)$$

$$\eta \sim T(u, B, K, A)$$

$$f = \prod_{t=1}^{T} \text{Multinomial}(\pi_t)$$

$$\phi = \text{ILR}$$

$$B = 0_{D-1}$$

$$K_{i,j} = \kappa^2 \exp(-\gamma^2 [d_{\text{phylo}}(i, j)]^2)$$

$$A_{t,s} = \alpha^2 \exp(-\rho^2 (t - s)^2)$$

Silverman JD, Roche K, et al. 2019. *arXiv*

NON-LINEAR TIME-SERIES MODEL FOR MICROBIOME