Bayesian Analysis of Latent Underdispersion Using Discrete Order Statistics

Jimmy Lederman[1] and Aaron Schein[1]

[1]Department of Statistics, University of Chicago

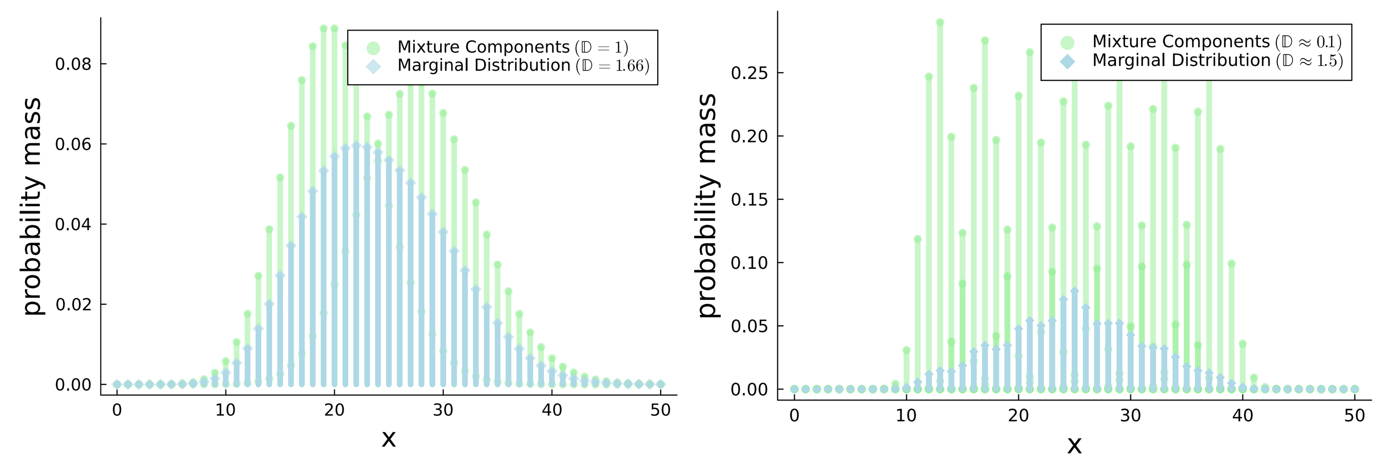
Conditional Underdispersion

Problem: no flexible tools exist to model underdispersed count data in probabilistic modeling frameworks.

Definition: A discrete random variable X is underdispersed with respect to the Poisson distribution if

$$\mathbb{D}[X] = \mathbb{V}[X]/\mathbb{E}[X] < 1$$

Count data which is *marginally* overdispersed may be consistent with a model that is underdispersed *conditionally*, once covariates or latent variables are observed.



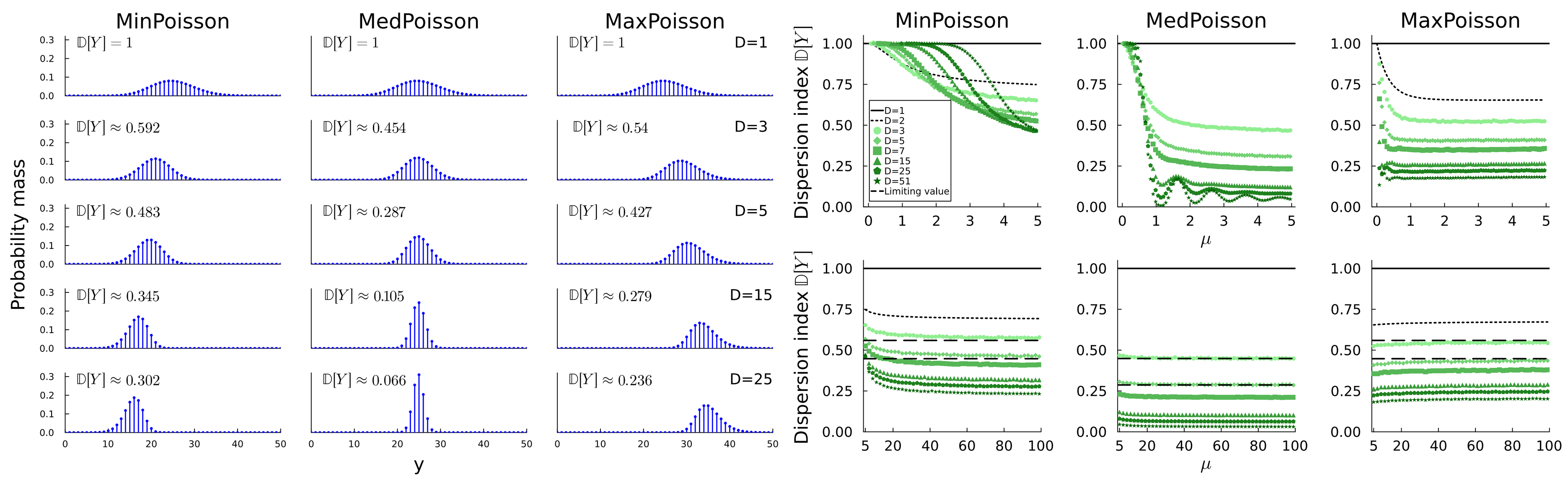
Key Benefit: Modeling conditional underdispersion allows for more precise probabilistic predictions than are possible with a Poisson likelihood.

Poisson Order Statistics are Underdispersed

We introduce discrete order statistics as a modular building block for probabilistic modeling. In particular, we build models with likelihoods of the form:

$$Y = Z_{(j,D)} \text{ for } j \in \{1,\ldots,D\}, \text{ where } Z_d \overset{\text{iid}}{\sim} \text{Poisson}(\mu)$$

This family of likelihoods gives rise to underdispersed count distributions.



D, the number of latent Poissons, controls the index of dispersion $\mathbb{D}[X]$.

As D grows larger, $\mathbb{D}[X]$ decreases as long as μ is sufficiently large.

Data Augmentation Connects Inference to the Poisson

We derive efficient MCMC-based inference for models with Poisson order statistic likelihoods via a data augmentation scheme with updates of the following form

$$Z_1 \ldots, Z_D \sim P(Z_1, \ldots, Z_D \mid \mu, Y) \quad (1)$$

$$\mu \sim P(\mu \mid Z_\bullet, Y), \text{ where } Z_\bullet = \sum_{k=1}^{D} Z_k \quad (2)$$

With the appropriate prior on μ, sampling from (2) is standard. Sampling from (1) is cumbersome due to the positive probability of exact ties.

Algorithm 4.1 Sampling from $P(Z_1, \ldots, Z_D \mid \theta, Z^{(j,D)} = Y) \propto \mathbb{1}(Y = Z^{(j,D)}) \prod_{d=1}^{D} f_\theta(Z_d)$

Input: observation $Y \in \mathbb{N}_0$, order $D \in \mathbb{N}$, rank $j \in [D]$, parent distribution f_θ
Initialize: $n^{(<Y)} = n^{(=Y)} = n^{(>Y)} = 0$
for $d = 1 \ldots D-1$ **do**
  $p_d^{(<Y)} \leftarrow P(Z_d < Y \mid \theta, n^{(<Y)}, n^{(=Y)}, n^{(>Y)}, Z^{(j,D)} = Y)$
  $p_d^{(=Y)} \leftarrow P(Z_d = Y \mid \theta, n^{(<Y)}, n^{(=Y)}, n^{(>Y)}, Z^{(j,D)} = Y)$
  $p_d^{(>Y)} \leftarrow P(Z_d > Y \mid \theta, n^{(<Y)}, n^{(=Y)}, n^{(>Y)}, Z^{(j,D)} = Y)$
  $c_d \sim \text{Cat}\left(p_d^{(<Y)}, p_d^{(=Y)}, p_d^{(>Y)}\right)$ where $c_d \in \{<Y, =Y, >Y\}$
  $n^{(c_d)} \leftarrow n^{(c_d)} + 1$
  if $n^{(=Y)} \geq 1$ and $n^{(<Y)} = n^{(<Y)}_{\max} = j-1$ **then**
    $Z_{d+1}, \ldots, Z_D \overset{\text{iid}}{\sim} \underset{[Y,\infty)}{\text{trunc}} f_\theta$
    break
  end if
  if $n^{(=Y)} \geq 1$ and $n^{(>Y)} = n^{(>Y)}_{\max} = D-j$ **then**
    $Z_{d+1}, \ldots, Z_D \sim \underset{[0,Y]}{\text{trunc}} f_\theta$
    break
  end if
  if $n^{(=Y)} = n^{(=Y)}_{\text{suff}} = \max(j - n^{(<Y)}, D - n^{(>Y)} - j + 1)$ **then**
    $Z_{d+1}, \ldots, Z_D \overset{\text{iid}}{\sim} f_\theta$
    break
  end if
end for
$n^{(=Y)} = \max(n^{(=Y)}, 1)$
$Z_1, \ldots, Z_{n^{(<Y)}} \overset{\text{iid}}{\sim} \underset{[0,Y-1]}{\text{trunc}} f_\theta$
$Z_{n^{(<Y)}+1}, \ldots, Z_{n^{(<Y)}+n^{(=Y)}} = Y$
$Z_{n^{(<Y)}+n^{(=Y)}+1}, \ldots, Z_d \overset{\text{iid}}{\sim} \underset{[Y+1,\infty)}{\text{trunc}} f_\theta$
Output: $\{Z_1, \ldots, Z_D\}$

Iteratively compute whether a latent Z_k is less than, greater than or equal to the observed value Y.

If Y is observed at least once and the remaining Z_k can be only less than or greater than Y, then we stop early.

Conditional on pre-computed information, sample independently from truncated distributions.

Using this algorithm, we can exactly sample from (1) for any discrete order statistic Y, including, for example, negative binomial order statistics.

Choosing which Order Statistic

While all Poisson order statistics can be used to model underdispersion, there are differences:

Maximum: computationally efficient, especially under **sparsity**. ($Y = 0 \implies Z_1, \ldots, Z_D = 0$)

Median: approximately mean-parameterized by latent μ, but more expensive computationally.

Minimum: computationally efficient, but not under sparsity.

Tailored Models for Applications

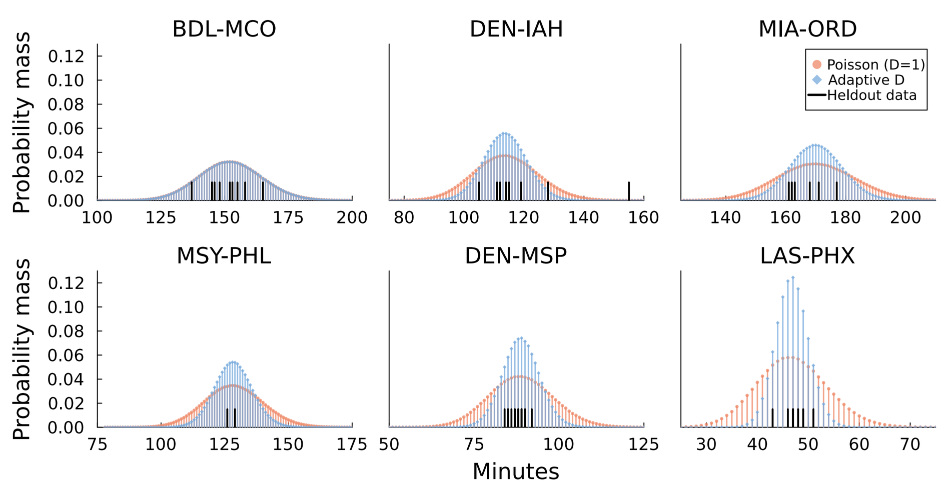
Predicting Flight Times

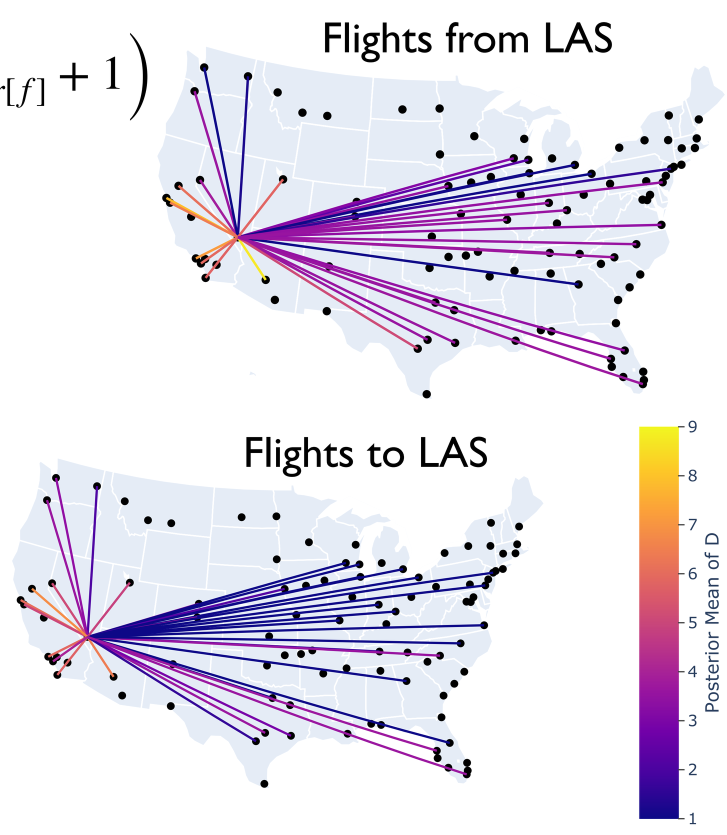
We model the number of minutes Y_f of flight f with origin $o[f]$, destination $d[f]$, and route $r[f] = (o[f], d[f])$ as

$$Y_f \sim \text{MedPoisson}\left(a_{o[f]} + b_{d[f]} + dist_{r[f]}\mu_{r[f]}, 2D_{r[f]} + 1\right)$$

$$D_r \sim \text{Binomial}\left(\frac{D_{max}-1}{2}, p\right)$$

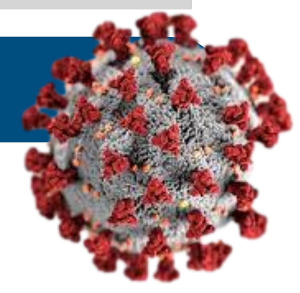
Some Example Posterior Predictive Distributions





A flexible dispersion parameter D gives varied uncertainty and learned latent structure.

COVID-19 Forecasting

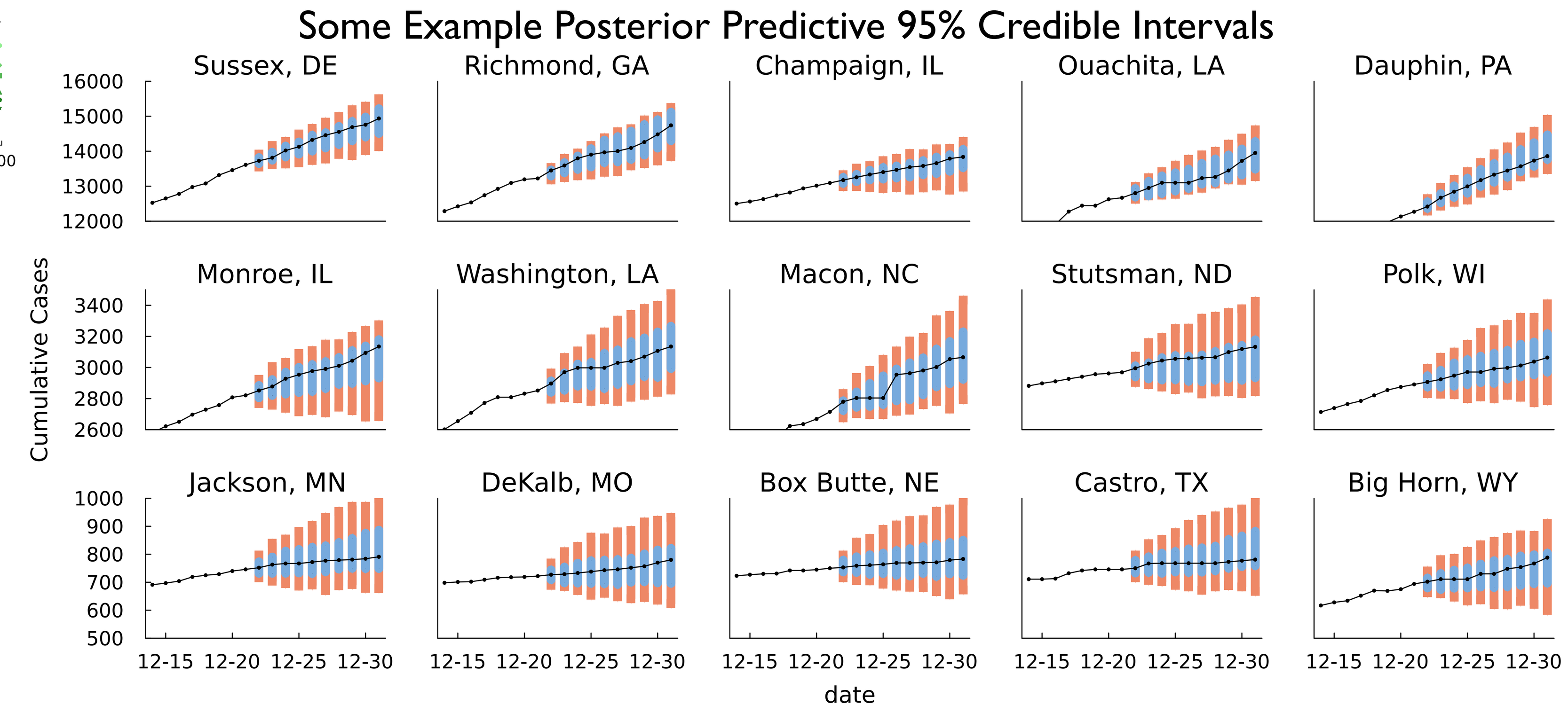


We model the cumulative COVID-19 cases Y_{ct} in each county c at time t as

$$Y_{ct} \sim \text{MedPoisson}\left(Y_{ct-1} + \log(pop_c)\left(\varepsilon + \alpha \sum_{k=1}^{K} \theta_{ck}\phi_{kt}\right), 2D_{ct} + 1\right)$$

$$D_{ct} \sim \text{Binomial}\left(\frac{D_{max}-1}{2}, \sigma\left(\sum_{q=1}^{Q} \beta_{cq}\tau_{qt}\right)\right)$$

Some Example Posterior Predictive 95% Credible Intervals



The underdispersed likelihood avoids artificially wide predictive intervals.

Modeling Abundance of Finnish Birds

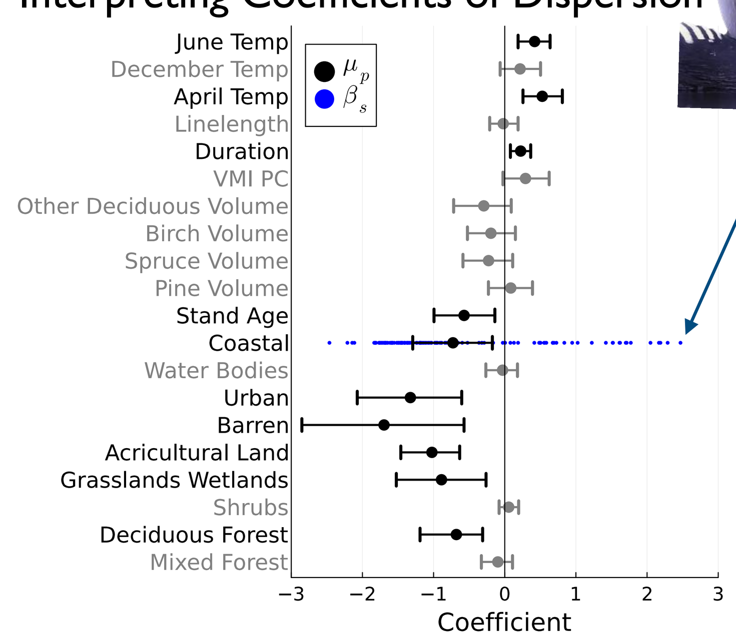
We model the number of birds Y_{ns} at sampling site n of species s, where each sampling site has covariate vector X_n, as

$$Y_{ns} \sim \text{MaxPoisson}\left(\sum_{k=1}^{K} \theta_{nk}\phi_{ks}, D_{ns} + 1\right)$$

$$D_{ns} \sim \text{Binomial}\left(D_{max} - 1, \sigma(X_n^T \beta_s)\right)$$

$$\beta_{sp} \sim N(\mu_p, \sigma_p^2) \quad \mu_p \sim N(0,1)$$

Interpreting Coefficients of Dispersion



Our model yields more precise predictions for waterbirds in coastal areas.

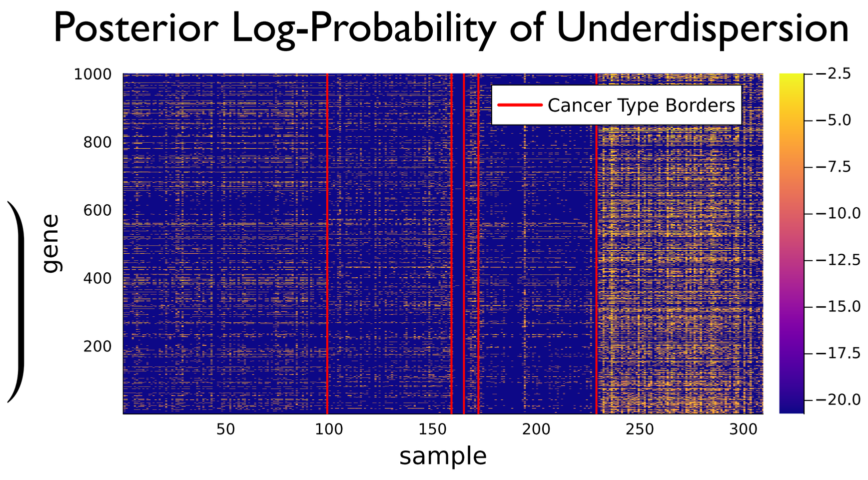
Investigating Dispersion in RNA-Seq Data

We model the RNA-sequencing count Y_{gs} for gene g at sample s as

$$Y_{gs} \sim \text{MedNB}\left(\sum_{k=1}^{K} \theta_{gk}\phi_{ks}, p_g, 2D_{gs} + 1\right)$$

$$D_{gs} \sim \text{Binomial}\left(\frac{D_{max}-1}{2}, \sigma\left(c_g + \sum_{q=1}^{Q} \beta_{gq}\tau_{sq}\right)\right)$$

Posterior Log-Probability of Underdispersion



We find little evidence of underdispersion in RNA sequencing data, but there are different patterns across cancer types.

Email questions to jlederman@uchicago.edu