Multiple pathogens infect multiple hosts: Inference for incidence, infection, & impact
Nature supports huge diversity of competitors

How do dozens to thousands of competitors coexist?
Hypothesis

• Competing species can coexist if each is attacked when it becomes abundant
• Requires a different pathogen to regulate each host
• If Janzen-Connell effects maintain diversity through pathogens, then
  – Pathogens effects are host-specific (N pathogens for N hosts)
  – Strongest effect when host is abundant
Inference for EID

‘ecological change and disease emergence are often mediated through complex processes that are not amenable to traditional causal inference’

Plowright et al. (2008) Frontiers Ecol
How to combine the evidence?

Plowright et al. (2008) *Frontiers Ecol*
Outline

• A basic model
• An application
• The dimensionality problem
• RJMCMC
• Evaluation
• Finding the important interactions
Outline

- A basic model
- An application
- The dimensionality problem
- RJMCMC
- Evaluation
- Finding the important interactions
Janzen-Connell for one pathogen, one host

Clark and Hersh (2009) *Bayesian Analysis*
Parameters, observations

Clark and Hersh (2009) *Bayesian Analysis*
A classical analysis

- Observations of both pathogen and survival:

\[
p_{D,S} = p(D,S|P = 1) = \left[\theta \phi s_1\right]^{y_{SD}} \left[\theta \phi (1 - s_1)\right]^{y_{(1-S)D}} \\
\times \left[(1 - \theta) s_0 + \theta (1 - \phi) s_1\right]^{y_{S(1-D)}} \\
\times \left[(1 - \theta)(1 - s_0) + \theta (1 - \phi)(1 - s_1)\right]^{y_{(1-S)(1-D)}}
\]

- Observations of survival only:

\[
p_S = p(S|P = 1) = \left[(1 - \theta) s_0 + \theta s_1\right]^{y_S} \left[1 - (1 - \theta) s_0 - \theta s_1\right]^{y_{1-S}}
\]

\[
y_S \quad \text{- no. seedlings in two } S \text{ classes}
\]

\[
y_{SD} \quad \text{- no. seedlings in four } (S, D) \text{ classes}
\]
A classical analysis

• Observations of both pathogen and survival:

\[
p_{D,S} = p(D, S| P = 1) = \left[ \theta \phi s \right]^{y_{DS}} \left[ \theta \phi (1 - s) \right]^{y_{D(1-S)}} \\
\times \left[ (1 - \theta)s_0 + \theta(1 - \phi)s \right]^{y_{(1-D)S}} \\
\times \left[ (1 - \theta)(1 - s_0) + \theta(1 - \phi)(1 - s) \right]^{y_{(1-D)(1-S)}}
\]

• Observations of survival:

\[
p_S = p(S| P = 1) = \left[ (1 - \theta)s_0 + \theta s \right]^{y_S} \left[ 1 - (1 - \theta)s_0 - \theta s \right]^{y_{1-S}}
\]

• Likelihood:

\[
\text{multinom} \left( y_{D,S} | n_{D,S}, p_{D,S} \right) \binom{y_S | n_S, p_S} \propto \prod_{D,S} p_{D,S}^{y_{D,S}} \times p_S^{y_S}
\]
A classical analysis

• Weak inference:
  – $\lambda \theta$ not independently identifiable
  – Too much uncertainty
  – 80% of seedlings attacked by > 1 pathogen
  – Effect of covariates?
  – Covariates vary among individuals
All hosts & pathogens must be modeled together

- Co-infection effects non-additive?
- All hosts provide information on incidence of all pathogens
- Environmental covariates affect pathogens, hosts, interactions
Outline

- A basic model
- An application
- The dimensionality problem
- RJMCMC
- Evaluation
- Finding the important interactions
Application

- Fungal pathogens on seedling hosts
- Experimental plots across moisture, light gradients
  - *Culture and DNA sequence* pathogens on live and dead hosts
- **Infer** pathogen incidence, infection, survival impact
- Complication: *co-infection*
Field methods

- Weekly mortality, biannual growth
- Dead and live hosts sampled for pathogens
- Covariates measured
- Survival submodel:

\[ \text{Bernoulli}(S_{hij} | s_{hijL}) \]
\[ \text{logit}(s_{hijL}) = x^{(s)}_{hijL} c_{hL} \]
Lab methods

- Disease symptoms
- Seedlings surface-sterilized
- Fragments plated on two media: AWA (fungi) and PARP (oomycetes)
- ID by morphology, DNA sequencing (ITS)
Detection varies

- **Cultures**: detection differs
- **DNA sequencing**: confident ID, available for a subset of seedlings $D_{hijk}^{(s)}$
- Detection in culture uncertain

$$p(D_{hijk}^{(c)} = 0 | I_{hijk} = 1) > 0$$
Incidence depends on soil moisture: \( \text{Bernoulli}(P_{jk} | \lambda_{jk}) \)

\[
\logit(\lambda_{jk}) = x_{jk}^{(\lambda)} a_k = a_k + a_{km} m_j
\]

Infection of host \( h \) by \( k \): \( \text{Bernoulli}(I_{hijk} | \theta_{hk}) \)

Survival and detection:

\[
p(S_{hiij}, D_{hijL}^{(c)} | I_{hijL}) = S_{hiij}^{S_{hij}} (1 - S_{hiij})^{1 - S_{hij}} \prod_{k \in L} \left( \phi_{k}^{(c)} \right)^{D_{hij}^{(c)}} (1 - \phi_{k}^{(c)})^{(\lambda_{hijk}^{(c)} - D_{hij}^{(c)})} I_{hijk}
\]

\[\Pr(D_{ij} | I_{ij} = 1)\]

\[\Pr(P_{j}) \quad \Pr(I_{ij} | P_{j} = 1) \]

\[\Pr(S_{ij} | I_{ij} = 0) \quad \Pr(S_{ij} | I_{ij} = 1)\]
Inference challenge: high pathogen diversity

Operational taxonomic units (OTUs) of fungi and oomycetes
Outline

• A basic model
• An application
• The dimensionality problem
• RJMCMC
• Evaluation
• Finding the important interactions
Dimensionality of the interactions

\[ \logit(s_{hijL}) = c_{h0} + \sum_{L=1}^{15} c_{hL} I_{hijL} + c_{hm} m_j + c_{hl} l_j \]

\[ = c_{h0} + c_{hL} + c_{hm} m_j + c_{hl} l_j \]

\( L - K \)-tuple of binary indicators in \( \{0, 1\}^K \)

For \( K \) pathogens on \( H \) hosts there are \( H \times 2^K \) combinations

10 hosts & pathogens require \( 10 \times 2^{10} = 10,240 \) models
Large model spaces increasingly common

- Complex systems (e.g., genomics, species and gene interactions)
- The dimensionality problem
  - $K$ parasites on $H$ hosts
- The multiplicity problem
  - Corrections for multiple comparisons (e.g., Bonferroni adjustment)
Traditional model selection criteria

- What they do: compare fit of two models to the same data set
- What they do not do (well):
  - Model evaluation
  - Provide probability statements
- Why not to use them?
  - Fit to one data set is usually not a good criterion for building a model
  - Index only relative
Why is it hard to select models?

• Scalability of AIC, BIC, DIC, …
  – 10 models require 45 comparisons, 100 models require 4950 comparisons
  – 10 hosts and pathogens require $10 \times 2^{10} = 10,240$ models = 52,423,680 comparisons
  – If we could make all of these comparisons, we’d still need an arbitrary cutoff for retention

• MCMC:
  – standard M-H simulates posterior within a model, does not change dimension
Required: move between models of different dimension

• Each infection combination represents a different model
How to move between dimensions?
Once there, where are we?

(where’s the posterior?)
Outline

- A basic model
- An application
- The dimensionality problem
- RJMCMC
- Evaluation
- Finding the important interactions
rjMcMC to evaluate high-dimensional model space

- $\theta_{hm}$ – vector of parameters for effects of each pathogen combination on survival of host $h$
- $M_{hm}$ – model indicator, dimension of $\theta_{hm}$
- Evaluate models of different dimension
- reversible jump Markov chain Monte Carlo
Posterior simulation

• **Metropolis**: a random walk through the posterior $p(\theta)$
  
  – Propose a parameter vector from symmetric $j(\theta)$

  $\theta' \sim j(\theta)$

  – Accept with probability

  $$a = \frac{p(\theta')}{p(\theta)}$$
Posterior simulation

- **Metropolis-Hastings:** also a random walk through the posterior $p(\theta)$
  - Propose a parameter vector from asymmetric $j(\cdot)$
    \[ \theta' \sim j(\theta) \]
  - Accept with probability
    \[ a = \frac{p(\theta')}{p(\theta)} \times \frac{j(\theta|\theta')}{j(\theta'|\theta)} \]
Posterior simulation

- **Reversible Jump MCMC**: a random walk through the posterior $p(\theta_\mu, M_\mu)$
  - Propose model & dimension variable
  \[ M' \sim J(M) \quad u \sim q(\theta_m, M_m) \]
  - Set $(\theta_m', u') = G(M, M') \forall \left\{ G(M, M') = G^{-1}(M', M) \right\}$
  - Accept with probability

\[
a = \frac{p(\theta')}{p(\theta)} \times \frac{j(\theta' | \theta)}{j(\theta' | \theta)} \times \frac{J(M | M')}{J(M' | M)} \times \frac{q(u')}{q(u)} \times \left| \frac{\partial G(\theta, M)}{\partial (\theta, M)} \right|
\]
Algorithm summary

- Propose a model
- Select a dimension-matching variable
- Evaluate parameter values from an invertible injection
- Acceptance criterion
Challenges

‘the application of reversible jump methodology has predominately remained within the domain of the MCMC expert, owing both to difficulties in constructing appropriate algorithms and to a common perception that it is particularly difficult to implement.’

Challenges

• MCMC won’t mix: there are no ‘local moves’
  – Metropolis can be optimized with arbitrarily small jumps
  – With RJMCMC we are changing dimensions
• Interpretation of parameters changes with model
• Multiplicity adjustment?
This algorithm

- Auxiliary variables and centering methods
- Model structure means that parameters have the same interpretation
- Multiplicity adjustment
Summary of inference goals

- Each host with each of $2^K$ pathogen combinations
  - Which are important?
  - Cannot test them all and compare, say, pairwise
- Can explore model and parameter spaces simultaneously, using RJMCMC
- Important relationships can be derived:
  - $\text{Pr}(M)$
  - $\text{Pr}(\theta|M)$
  - Total $\text{Pr}(S|P)$
  - Total $\text{Pr}(S|E)$
Outline

• A basic model
• An application
• The dimensionality problem
• RJMCMC
• Evaluation
• Finding the important interactions
Simulation for model evaluation

• Pathogen effect on survival
• Sample sizes like our experiment \((H = 6, K = 4, n = 700)\)
• Correct models identified, false positives when few detections in data set

Clark and Hersh, *Bayesian Analysis* (2009)
RJMCMC chains from simulation

Chains (left) and posterior densities (right) for models having the 10 highest posterior probabilities.

Chains are discontinuous as parameters are dropped and reinstated. Horizontal dashed lines are true values.

At right, prior densities are green, posterior densities black.

No. times the infection combination was detected is given at right.

Models 81 and 82 are false positives.

Clark and Hersh, *Bayesian Analysis* (2009)
Simulation studies recover parameter values

- 95% CIs include true values

Clark and Hersh, *Bayesian Analysis* (2009)
Simulation studies work with more spp

- 7 hosts, 7 pathogens = 896 models (20 correct)
- False positives when < 10 detections
- False negatives when effect is small

Clark and Hersh, *Bayesian Analysis* (2009)
Outline

• A basic model
• An application
• The dimensionality problem
• RJMCMC
• Evaluation
• Finding the important interactions
Range of infection rates: host-pathogen combinations

Clark and Hersh, *Bayesian Analysis* (2009)
Differences in survival effect

- Posterior model probabilities for host-specific combinations of infection
- Different hosts susceptible to different combinations of attack

Clark and Hersh, *Bayesian Analysis* (2009)
Predictive distributions

Survival given incidence marginalizes infection risk

\[
p(S_{hL} \mid P_L = 1) = \sum_{I_{hL} = 0,1} p(S_h \mid I_{hL}) p(I_{hL} \mid P_L = 1)
\]
Predictive distributions

Survival given environment marginalizes incidence:

$$p(S_h|m,l) = \sum_{P_k=0,1} \sum_{I_k=0,1} p(S_h|I_k,m,l)p(I_k|P_k)p(P_k|m)$$
Environment at $j$ affects incidence of pathogen $k$

Pathogens do best where hosts do best

Clark and Hersh, *Bayesian Analysis* (2009)
Predictive density for annual survival rate at different scales

Effects depend on scale of information about risk factor

Interactions change the risk factor

Dry bad for host and pathogen

Clark and Hersh, Bayesian Analysis (2009)
Conclusions

• The complexity challenge
  – Reduce huge no. of potential interactions to those that matter

• Janzen Connell
  – The importance of interactions
    • Without them, no specificity
    • With them, specificity