Beyond Molecular Biology

Applying Gene Regulation Network Inference Methods in Ecology

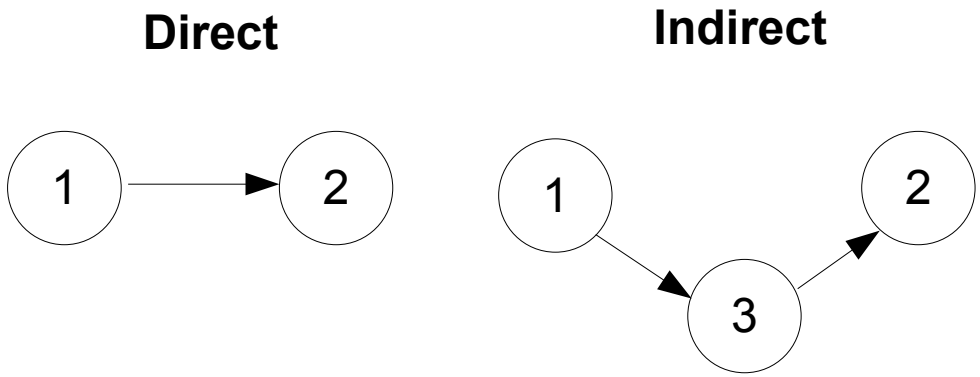
Frank Dondelinger
Ali Faisal
Dirk Husmeier
Colin Beale

Gene Regulation Networks

The Problem

- Large quantities of gene expression data (e.g. microarray data)
- How to infer the gene network?

Simple methods cannot distinguish between direct and indirect interactions



Need Machine Learning and Computational Statistics

Species Interaction Networks

Analogous Problem

- Large quantities of population data (e.g. from surveys)
- How to infer the species network?

Systems Biology	vs.	Ecology
Genes	⟷	Species
Expression Levels	⟷	Population Densities
Gene Regulation	⟷	Species Interactions
Different Conditions	⟷	Different Environments

The Data

In the real world:

- Population numbers gathered by ecologists
- Noisy, expensive to collect
- Usually estimates, or presence/absence values only

Simulation data:

- Same format, but based on a model
- Allows better evaluation of network reconstruction methods

Network Reconstruction Methods

Two sparse regression methods:

- Sparse Bayesian Regression (SBR)
- Least Absolute Shrinkage and Selection Operator (LASSO)

Bayesian network method:

Structure MCMC with Edge Reversal Move

Sparse Bayesian Regression

(Tipping and Faul 2003, Rogers and Girolami 2005)

Bayesian Linear Regression Model:

- Independent Gaussian priors for each weight

$$P(\mathbf{w}|\boldsymbol{\alpha}) = \prod_i N(w_i|0, \alpha_i^{-1})$$

(P(**w**) sparse after integrating out **α**)

- Optimise L2 log-likelihood of hyperparameters **α** indicating the strength of the priors.
- Obtains sparse solution: Most weights close to zero.

LASSO

(Tibshirani 1996, van Someren et al. 2006)

Linear regression model (L1 Regularisation)

$$\mathbf{w} = argmin\{\sum_i (y_i - \sum_j w_j x_{ij})^2\} \text{ with the constraint that } \sum_j |w_j| \leq t$$

Equivalent to adaptive ridge regression (Grandvalet 1998)

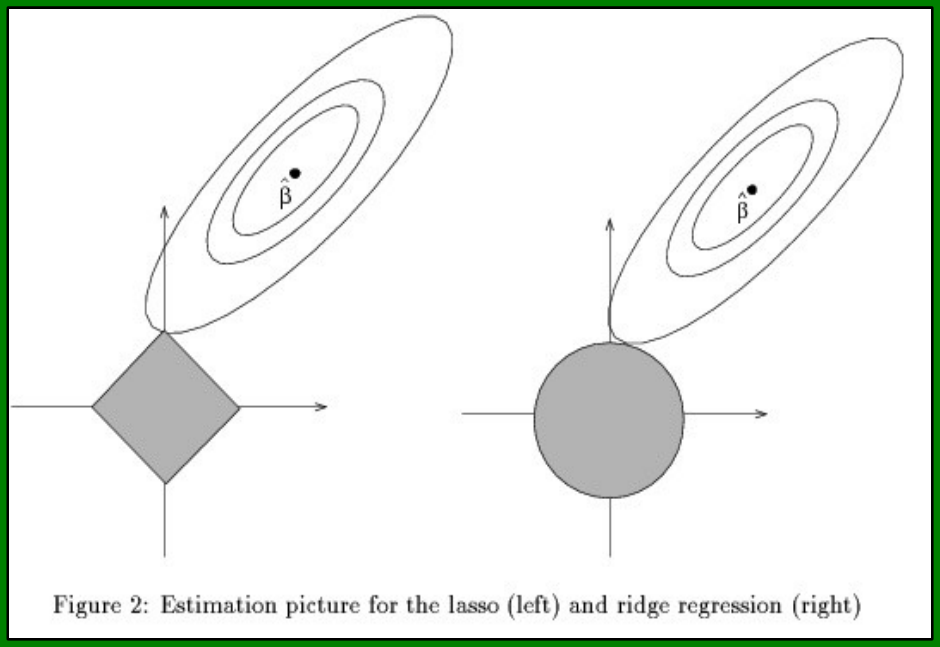


Figure 2: Estimation picture for the lasso (left) and ridge regression (right)

(Figure from Tibshirani 1996)

Advantages:

- Reduce weights as much as possible (shrinkage)
- Set some weights to zero (selection)

SBR vs LASSO

	SBR	LASSO
Setting Hyperparameters	ML Type II, Laplace Approximation	10-Fold Crossvalidation
Weight Prior	Using uniform hyperprior: $P(w_i) \propto \frac{1}{\lvert w_i \rvert}$ (improper)	Laplace Prior: $P(w_i) \propto e^{-\lvert w_i \rvert}$
Regularisation	$-\log P(w_i) \propto \log \lvert w_i \rvert$ $\nabla -\log P(w_i) \propto \frac{1}{\lvert w_i \rvert}$	$-\log P(w_i) \propto \lvert w_i \rvert$ $\nabla -\log P(w_i) \propto const$

Bayesian Networks

(Heckerman and Geiger 1994, Friedman et al. 2000)

Probabilistic graphical model where the joint probability decomposes as:

$$P(X_1 \ldots X_M) = \prod_i^M P(X_m | \Pi_m)$$

Want to learn the structure:

- Closed form for marginal likelihood under Gaussian assumption (BGe)
- Find posterior edge probabilities using MCMC

MCMC Structure Learning

(Madigan and York 1995)

Generate a Markov Chain of networks:

- At each step, add, delete or reverse an edge.

Sample from the chain to obtain post. edge probabilities

Problem: Edge reversals cause many rejections

Solution: Use better edge reversal method that samples new parents for nodes connected by reversed edge (Grzegorczyk and Husmeier 2008)

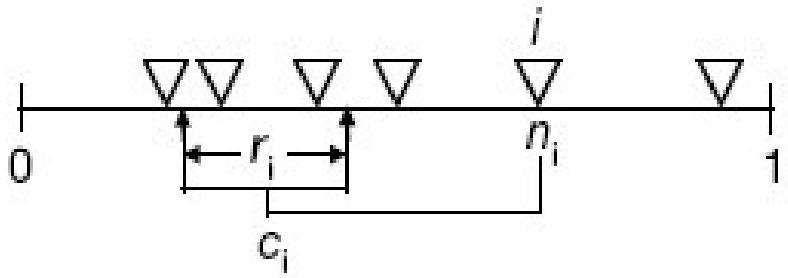
Simulation Model

Simulates population development of different species over 2D area

Two parts: Interaction model (food web) and population model

Interaction Model: Niche model

(Williams and Martinez 2000)



n_i – niche position of species i
c_i – centre of prey niche for species i
r_i – range of prey niche for species i

Shown to give a good fit to actual food webs

Simulation Model

Population Model:

(Engen and Lande 2003)

$$\frac{dX_i}{dt} = r_i + \frac{\sigma_d}{\sqrt{N_i}}\frac{dA_i(t)}{dt} + \sigma_e\frac{dB_i(t)}{dt} - \gamma X_i - \Omega(\mathbf{X}) + \sigma_E\frac{dE(t)}{dt}$$

X_i – log pop. density of species i
r_i – growth rate
σ_d – demographic std. dev.
N_i – pop. density of species i
$A_i(t)$ – demographic effect

σ_e – environmental std. dev. (species specific)
$B_i(t)$ – environmental effet (species specific)
γ – density dependence
$\Omega(\mathbf{X})$ – species interactions
σ_E – environmental effect std. dev. (global)
$E(t)$ – environmental effect (global)

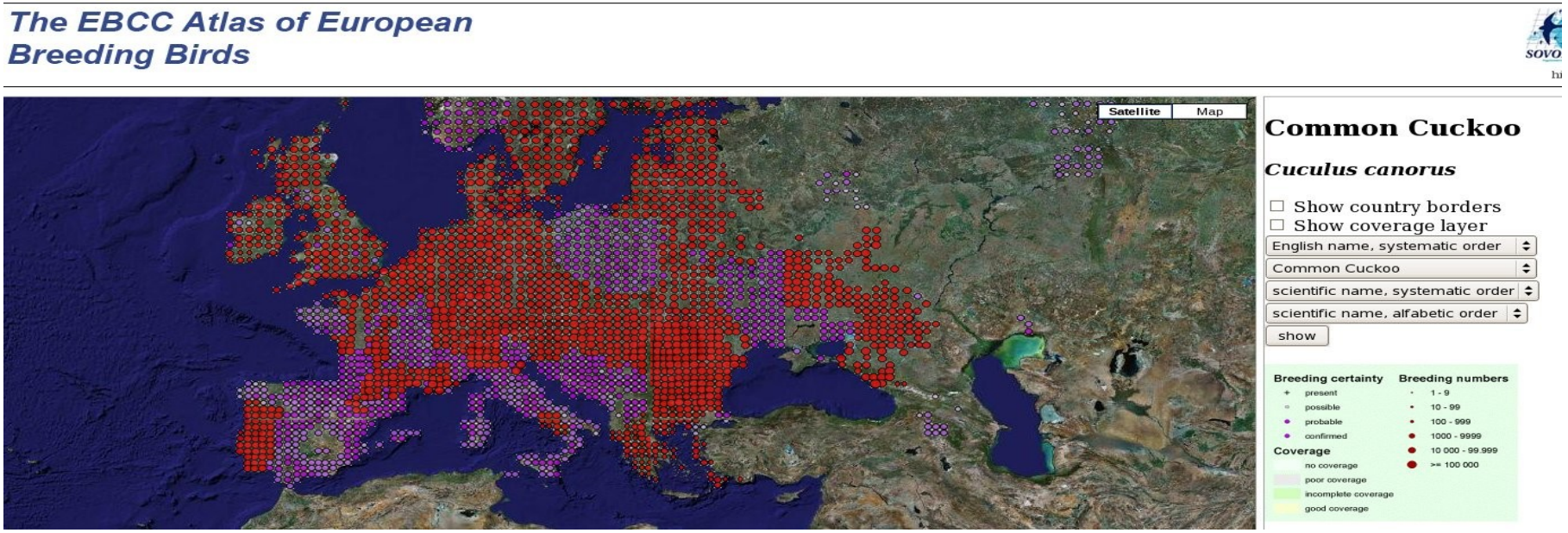
Modified to allow for species interactions

Includes exponential 2D species dispersal

Real Data

European Bird Atlas Data:

- Absence/Presence data for bird species in Europe
- Data at ~4000 grid points
- Each grid point corresponds to 50x50km square



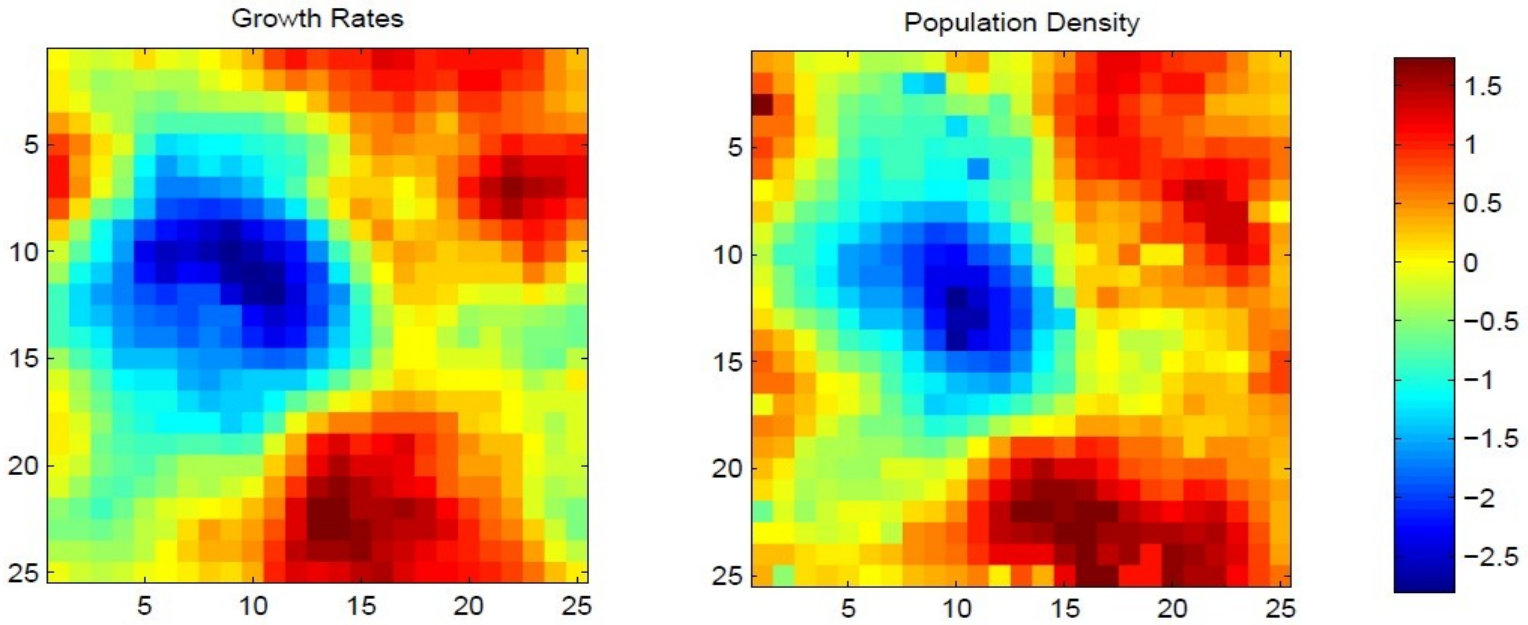
Purple – Probable presence **Red** – Confirmed presence

Spatial Autocorrelation

In real data and simulation:

- Discovered spurious interactions between species sharing the same habitat

In simulation, caused by growth rates:



Modeling Spatial Autocorrelation

In **regression**:

- Add autocorrelation variable α for current target species

If considering *n* neighbours:

$$\alpha = \sum_i^n w_i x_i$$

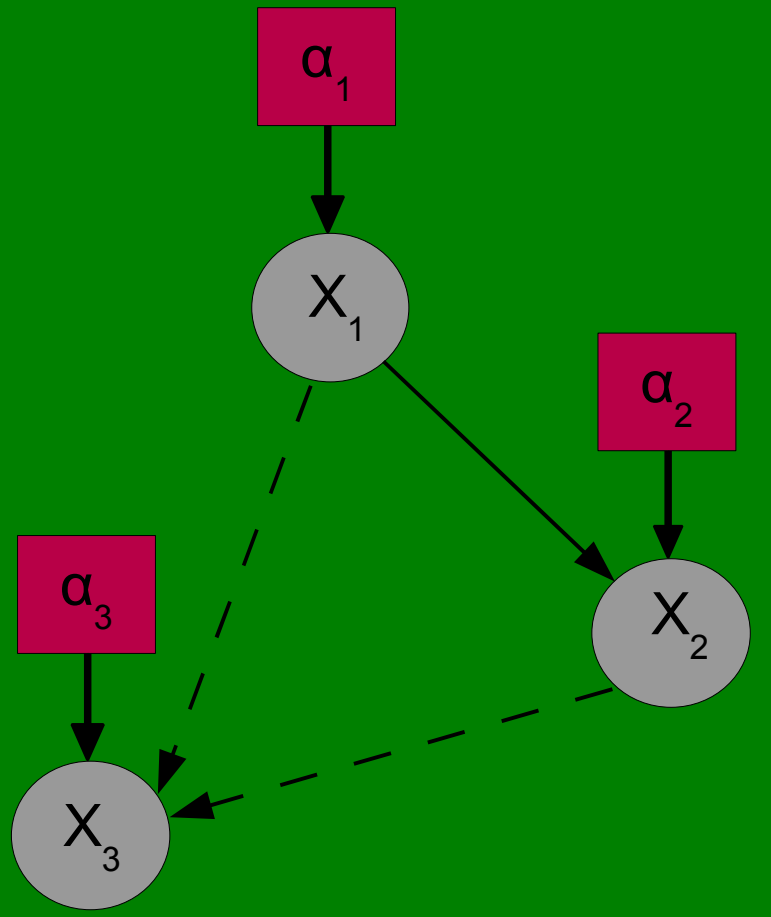
During regression, autocorrelation effects are caught by the weights for α and leave other weights to catch species interaction effects.

Modeling Spatial Autocorrelation

In **Bayesian networks**:

An equivalent approach would double the number of nodes in structure inference: **Not desirable**

Alternative: Add hard-wired autocorrelation nodes

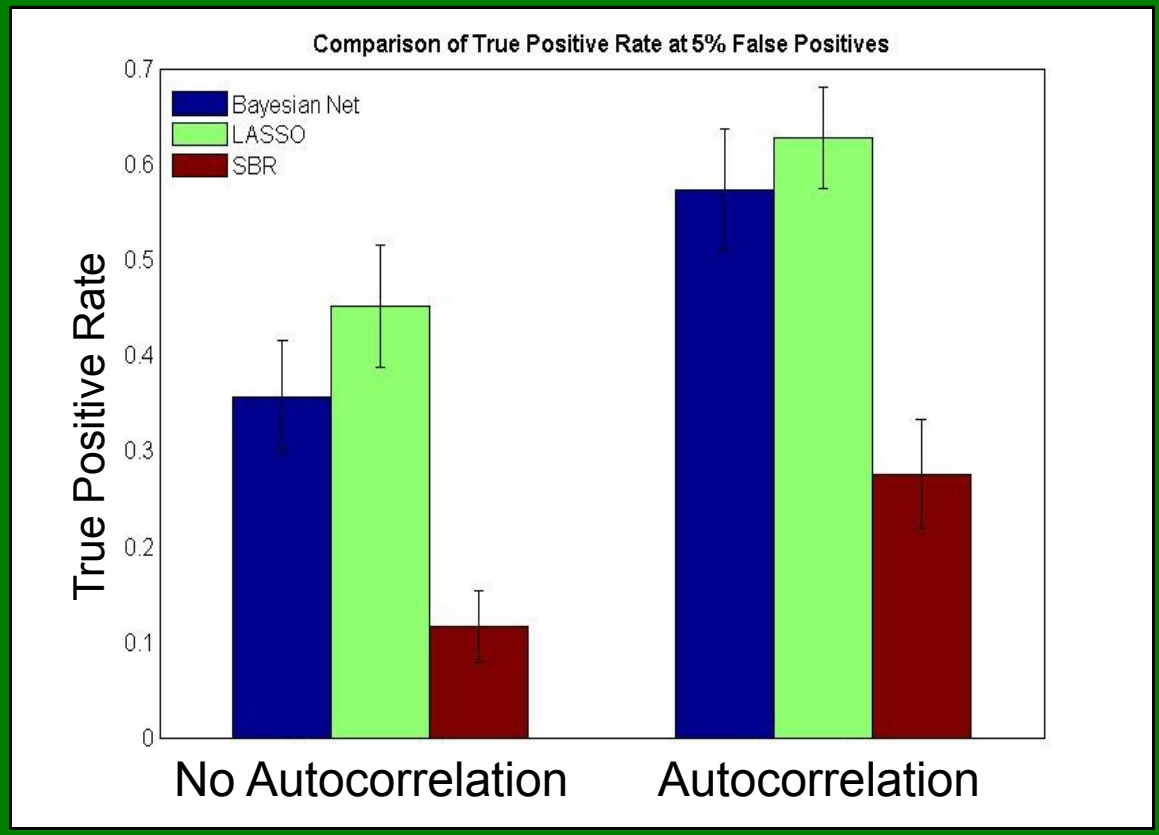


Simulation Results

Two measures:

- True positive rate at false positive rate 5% (TPFP5)
- ROC curve plotting true positives vs false positives

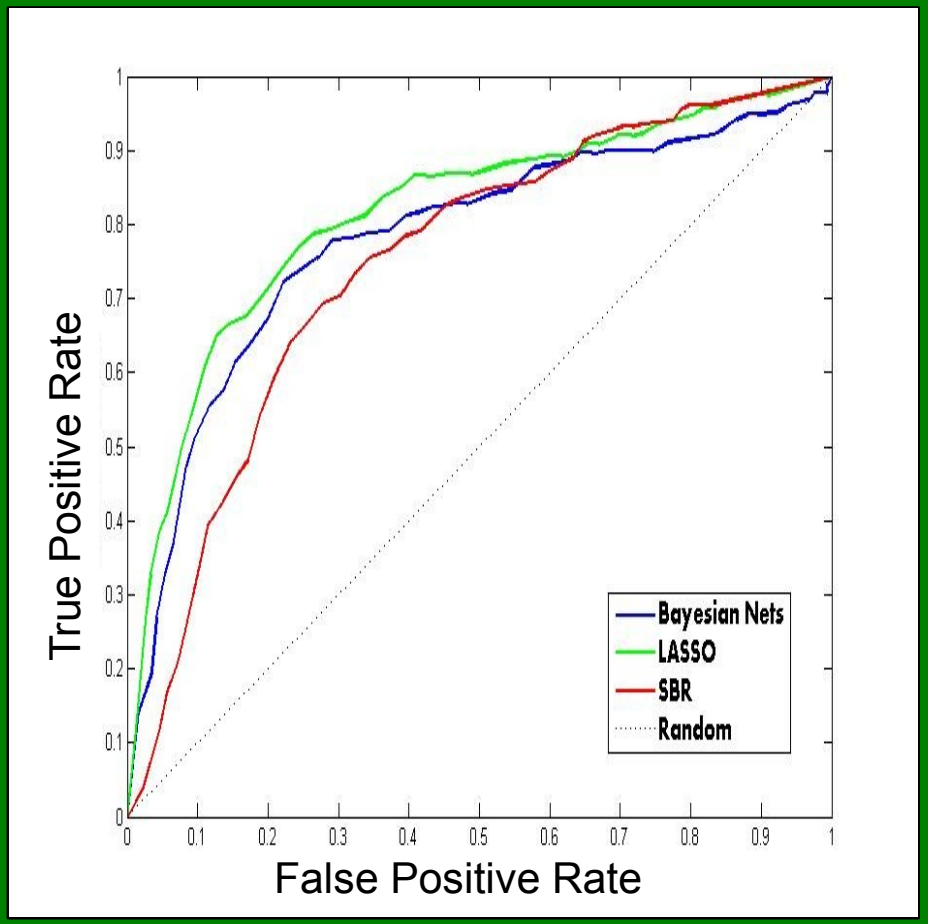
No Autocorrelation vs Autocorrelation TPFP5



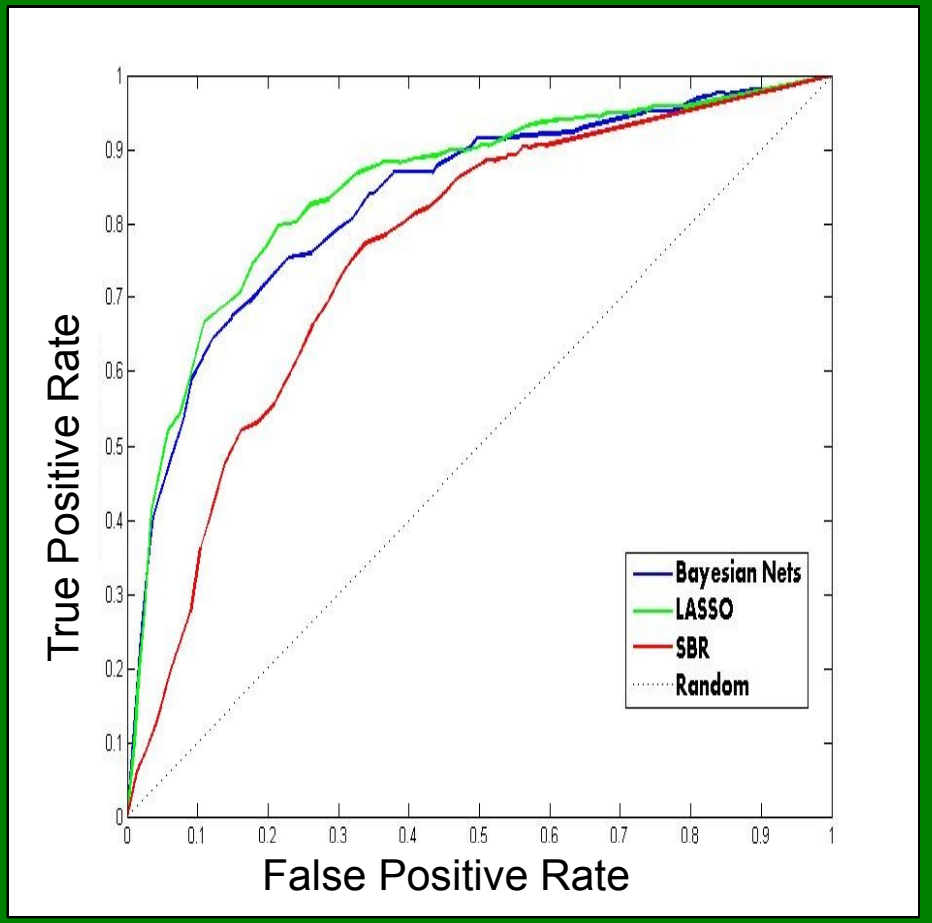
$p < 0.05$ for all 3 methods

Simulation Results

ROC Curves without Autocorrelation

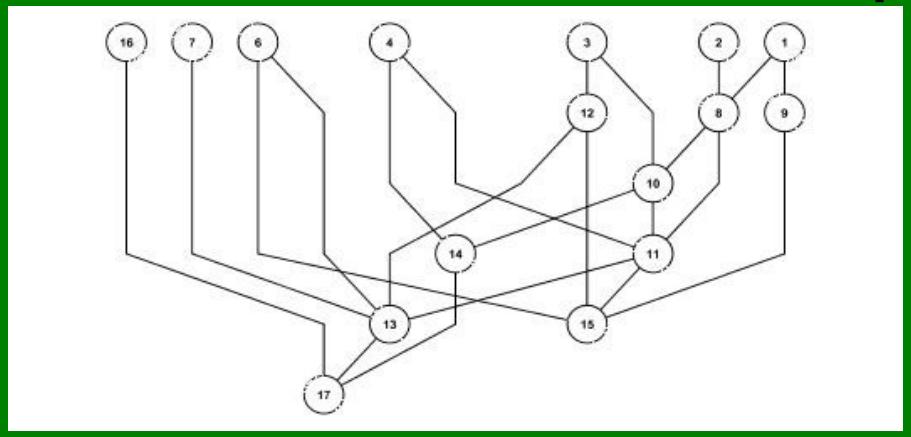


ROC Curves with Autocorrelation

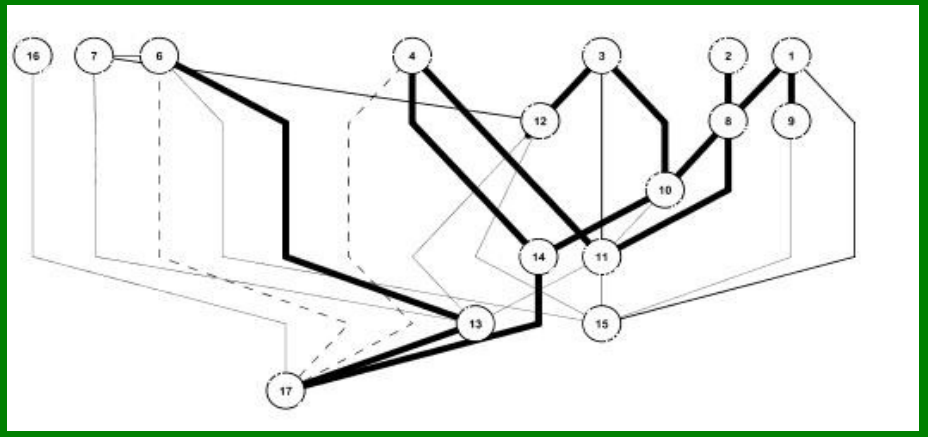


Simulation Results

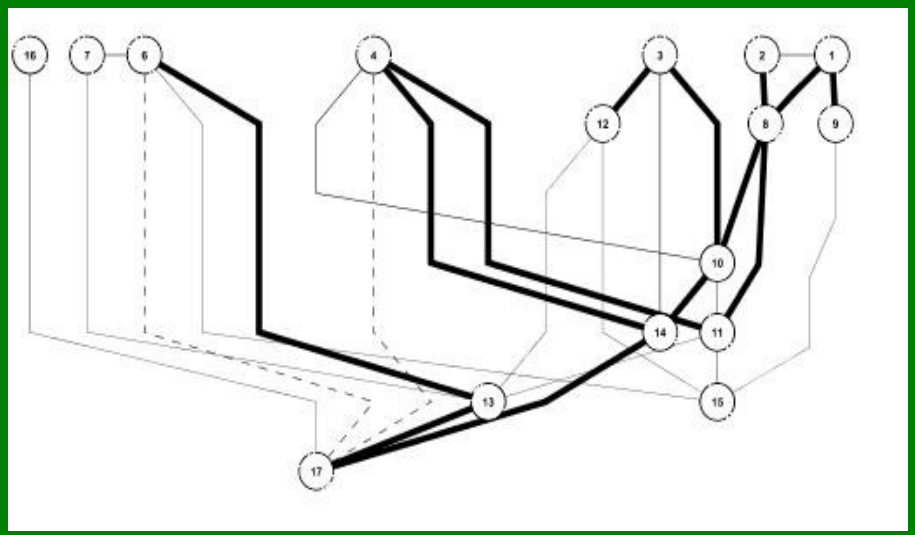
Example Network



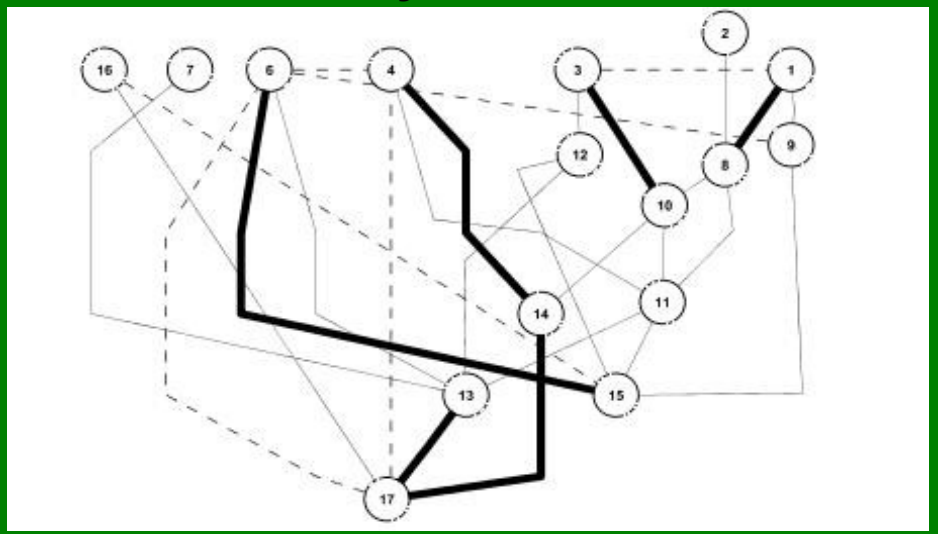
Real Network



Bayesian Net



LASSO

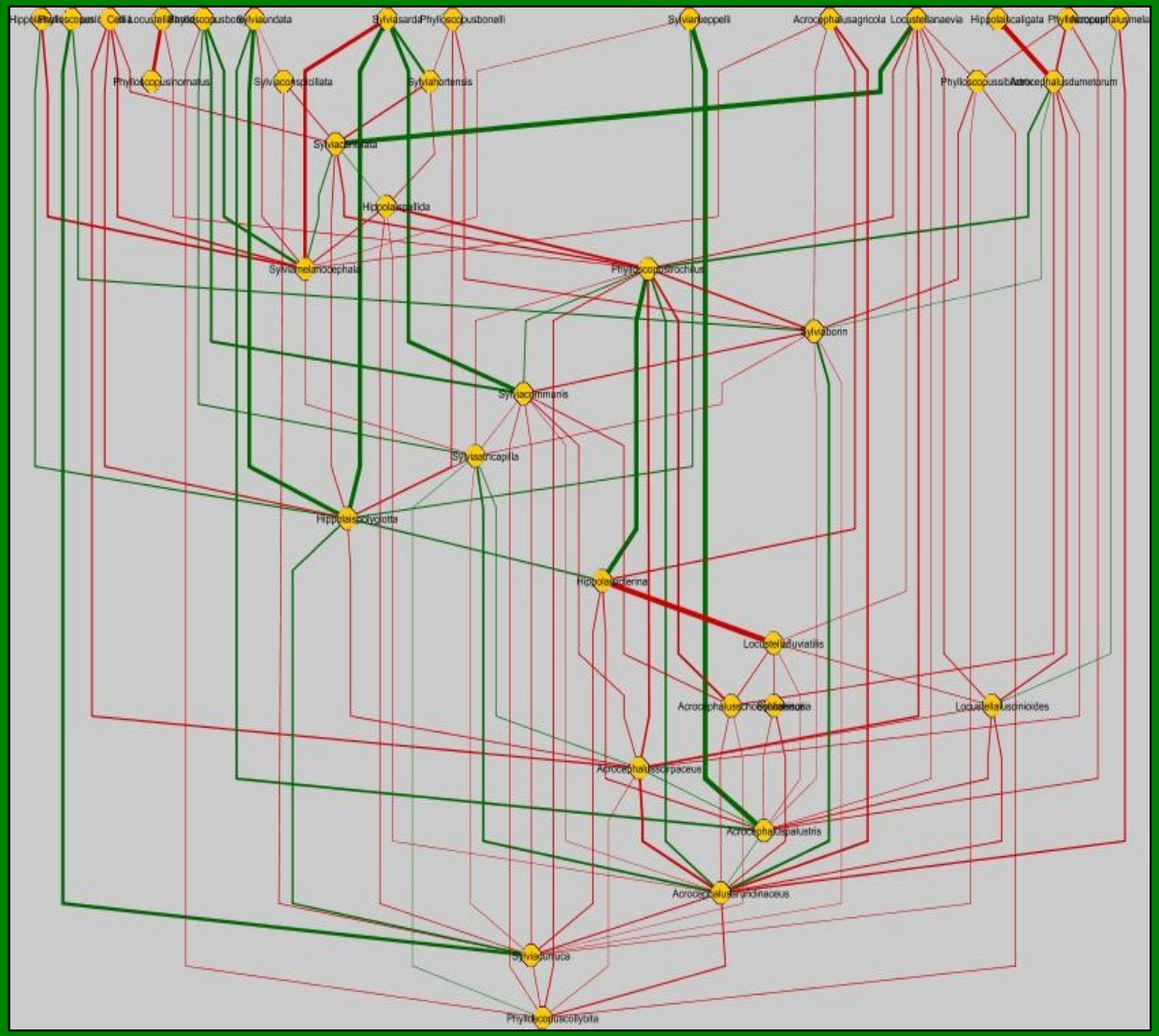


SBR

Real Data Results

Consensus Network (LASSO and Bayesian nets)

- Models autocorrelation
- Incorporates temperature and water presence
- Evidence for many interactions in literature



Data for figure obtained by Ali Faisal

Conclusions

- Machine Learning approaches viable for network inference in ecology
- Problem of Spatial Autocorrelation
- LASSO surprisingly effective
- Bayesian nets offer possibilities for incorporating prior knowledge
- Latent variable model holds some promise.

Acknowledgements

Thanks to:

- My supervisor, Dr. Dirk Husmeier of BioSS.
- Ali Faisal, who applied the methods to the bird species data set.
- Dr. Colin Beale of the Macaulay Institute.
- Dr. Marco Grzergrczyk, formerly of BioSS, now at TU Dortmund, for providing Matlab implementations of the Bayesian Net software.
- Dr. Jon Yearley at the University of Lausanne, for supplying the implementation of the simulation model.

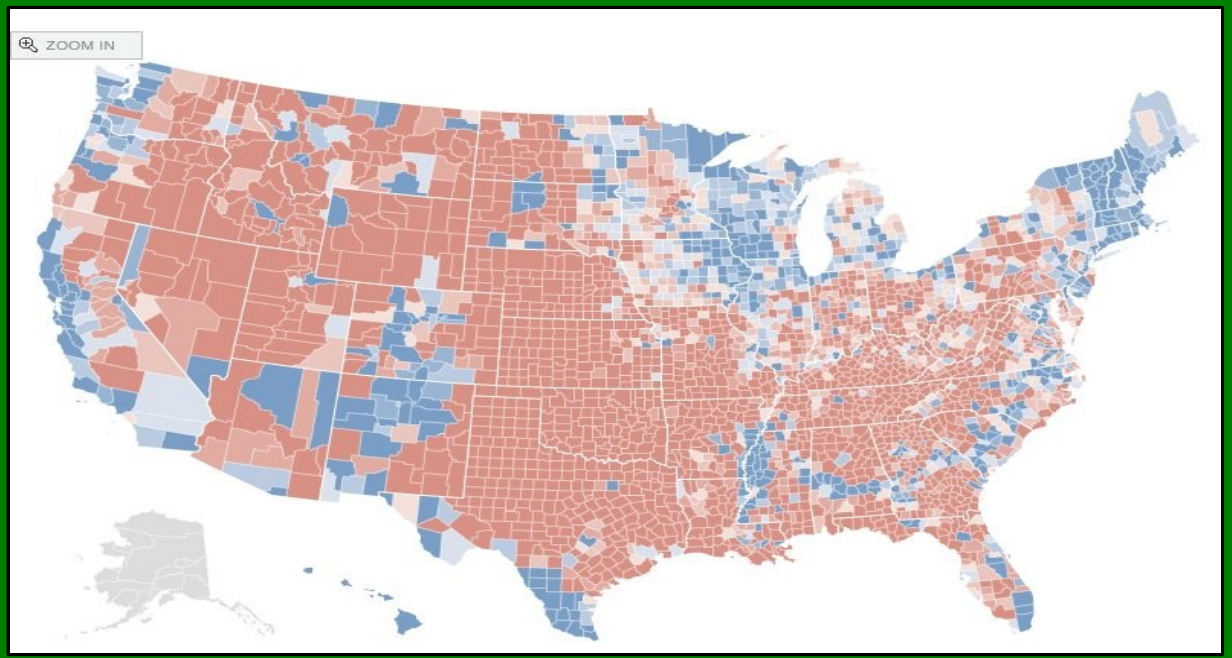
Any Questions?

The Autocorrelation Problem

Discovered spurious interactions between species sharing the same habitat:

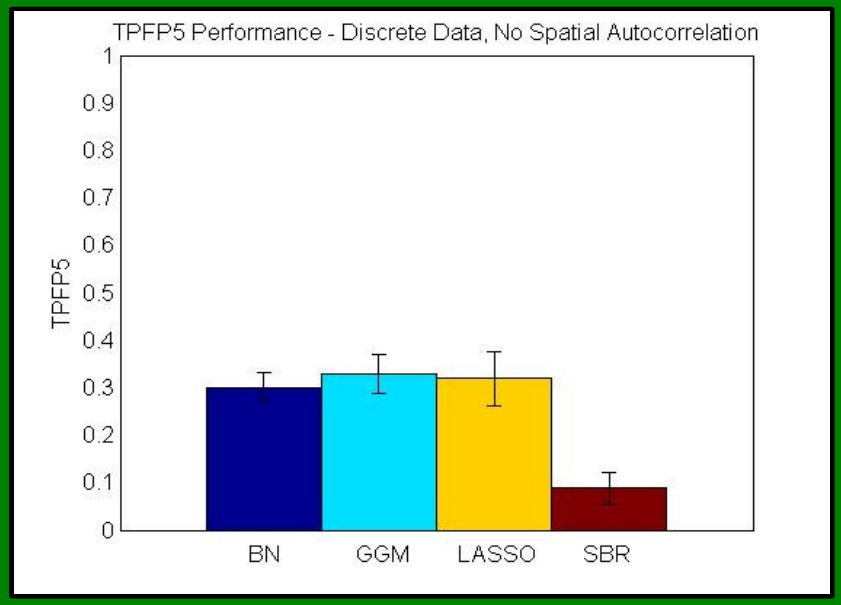
Spatial Autocorrelation

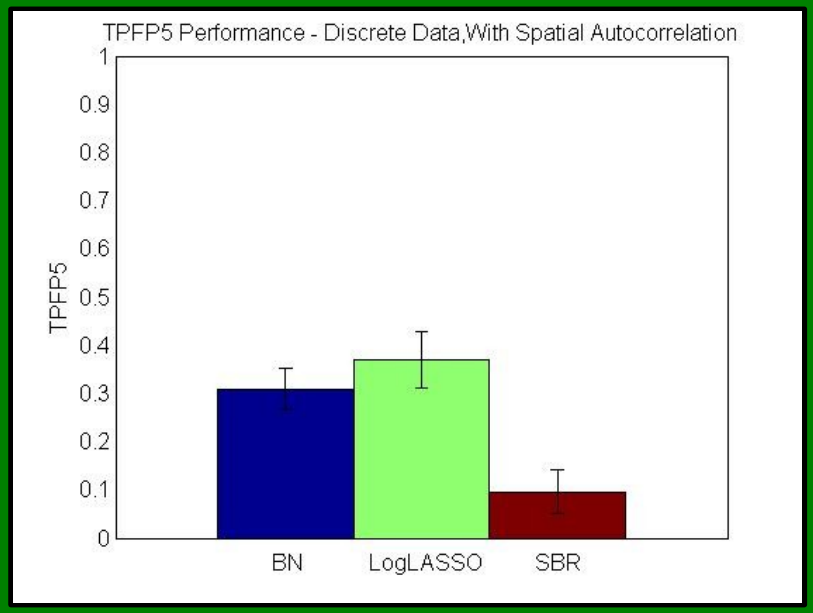
Example:



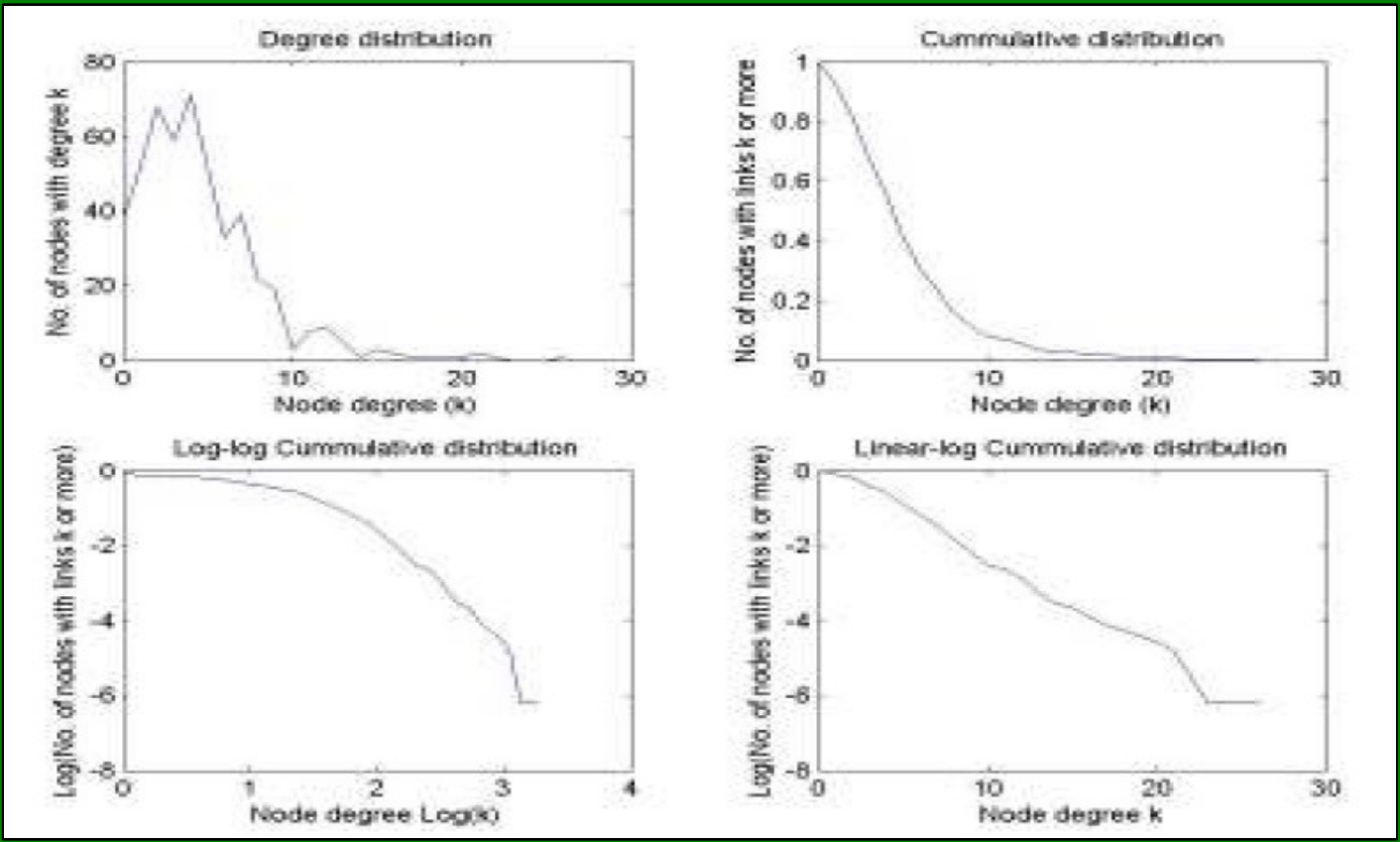
Discrete Data

- Discretise population densities using binomial observation process.
- Some information loss
- Autocorrelation effect disappears





Real Data Results



Degree distribution:

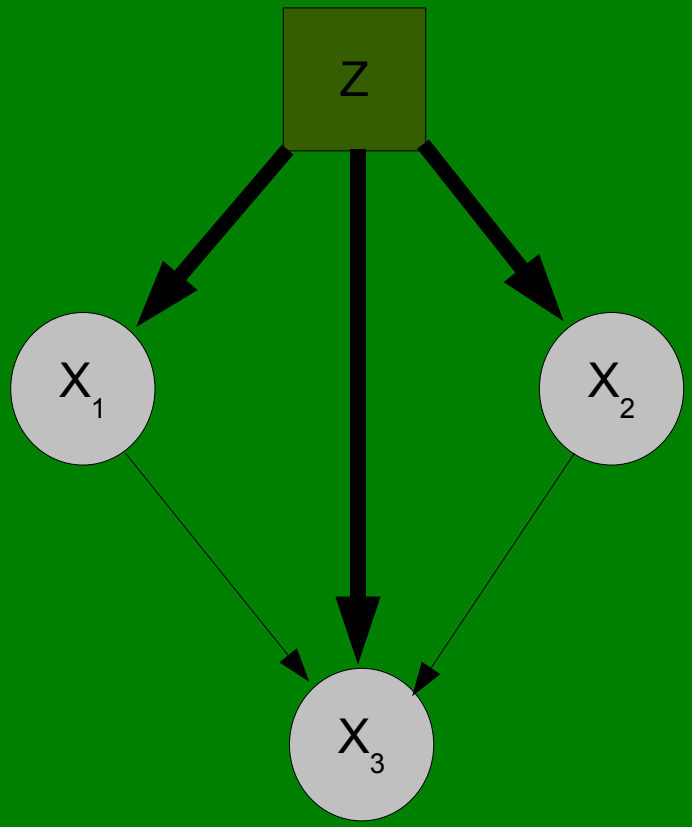
- Found to be approximately exponential
- Some disagreement about what the default distribution in ecological networks should be

Latent Variables

Idea: Extend the Bayesian Network to include unobserved nodes

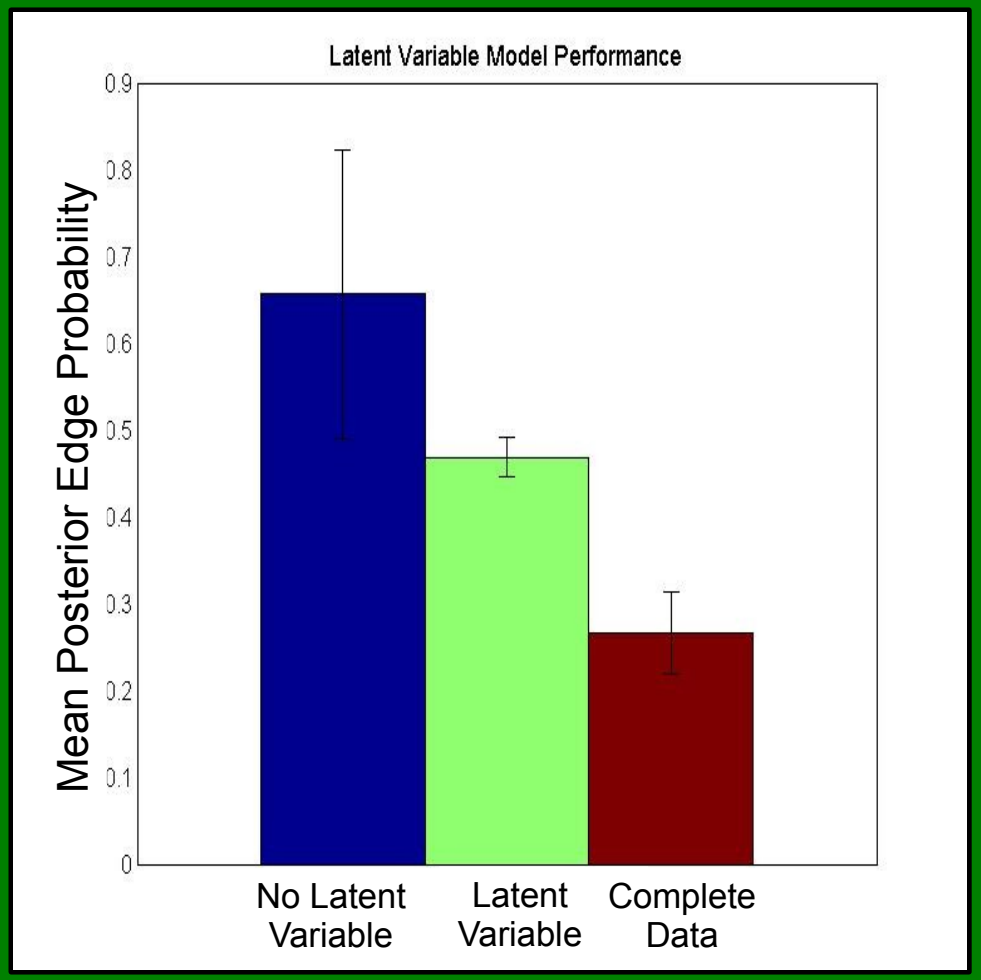
- Capture environmental effects
- Start with one latent variable and full connectivity
- Corresponds to mixture model

Allocation Sampler (Grzegorczyk et al. 2009)



Latent Variables

Simulation



Real Data

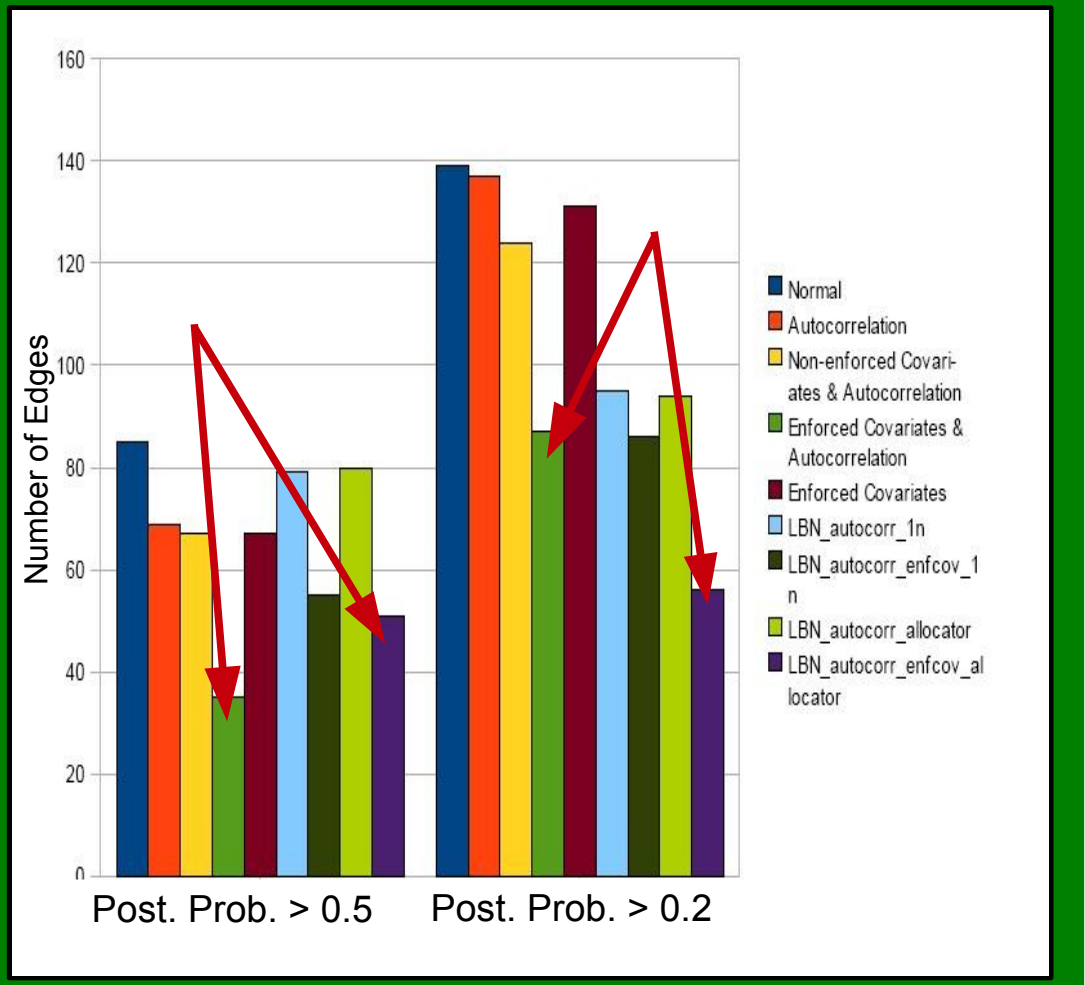


Figure created by Ali Faisal