Structural equation modeling of immune response

Lorenz Wernisch

joint work with Jenny Houghton and David Dunn

MRC Biostatistics Unit, Cambridge

Parasitology, Cambridge University

Causal networks in biology

- Complex control mechanisms
- Networks are ideal for representation of complexity
- Network entities and wiring are real, not only conceptual or figment of imagination
- Obtaining suitable data is difficult, but rapid advances are made:
  - Single cell experiments
  - Flow cytometry
  - Lab robots
  - Microfluidic devices

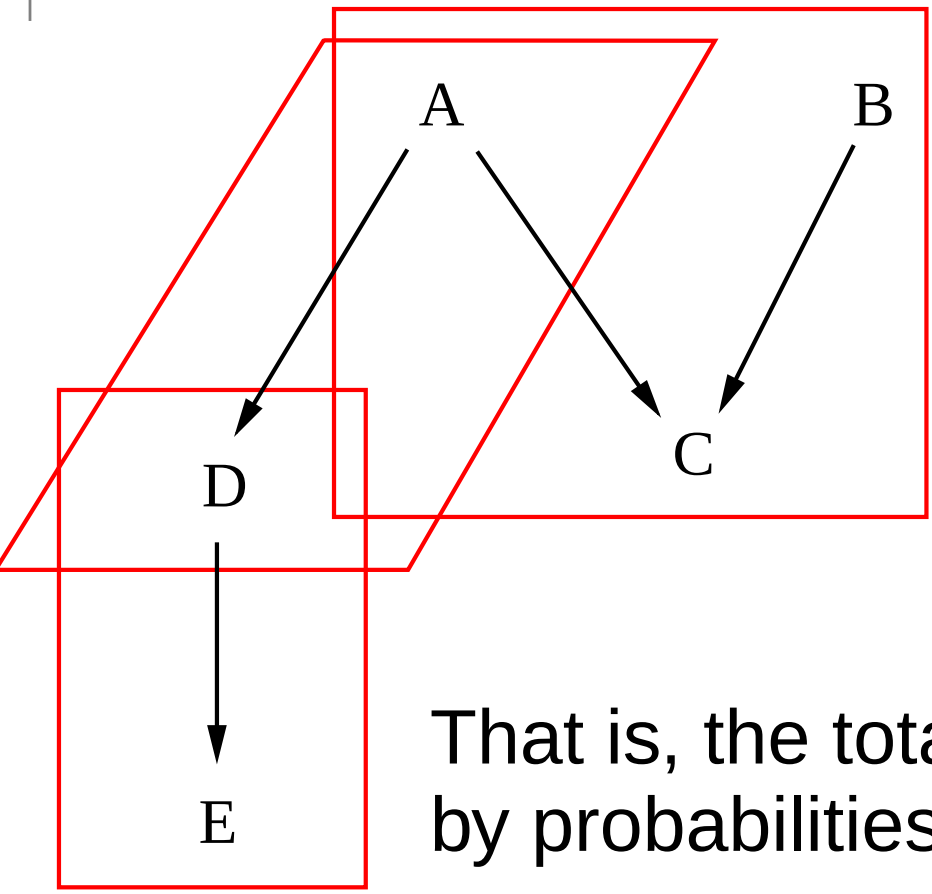
Choices for network models

- Data type type: discrete/continuous/mixed, linear/nonlinear, . . . ?
- **Steady state** assumption or **dynamic time-series**?
- **Network structure** known, partially known or needs to be inferred?
- Statistics on the **parameters** required?
- Major complication 1: suspect **hidden variables**
- Major omplication 2: **cyclic** causal dependencies are circular: A $\rightarrow$ B, B $\rightarrow$ C, C $\rightarrow$ A

Inference strategies for network structure

- **Scoring**
  - **Probabilistic score** measuring how well data are represented by graph
  - **Search strategy** for high scoring networks
- **Independence constraints**
  - **Independence oracle** provides information on statistical indepence of variables
  - **Graph algorithm** reconstructs network from independence relations

Standard Bayesian network



A probability distribution factors with or is represented by a directed **acyclic** graph (DAG) if

$$\mathrm{P}(X_1, \ldots, X_n) = \prod_i \mathrm{P}(X_i | \text{parents}(X_i))$$

That is, the total probability is given by probabilities on "families"

$$\mathrm{P}(A, \ldots, E) = \mathrm{P}(A)\mathrm{P}(B)\mathrm{P}(C|A, B)\mathrm{P}(D|A)\mathrm{P}(E|D)$$

Bayesian complexity control

- Scoring network S using data $D = (D_1, \ldots, D_n)$
- Bayesian approach to network likelihood $p(D \mid S)$: **integrate out** unknown parameters:

$$p(D \mid S) = \int p(D \mid \theta, S) p(\theta)\, d\theta$$

- Inbuilt **cross validation**:

$$p(D \mid S) = p(D_1 \mid S)\, p(D_2 \mid D_1, S) \cdots p(D_n \mid D_1, \ldots D_{n-1}, S)$$

Predict data set D_i using $D_1, \ldots, D_{i-1}$ for training

Bayesian scoring

Given a candidate Bayesian network B on n nodes and data D, the probability of B is evaluated locally

$$p(B|D) = \frac{p(B)}{p(D)} \prod_{i=1}^{n} \frac{p(D(\text{parents}_i, \text{node}_i)|B)}{p(D(\text{parents}_i)|B)}$$

$p(D(X)|B)$ multivariate t**-distribution from integrating over** regression parameters and covariances in Normal-Wishart distribution

$p(B)$ is a **prior on networks**, independent knowledge about networks can be incorporated here

Details in papers by D Geiger and D Heckerman

Search strategies for scoring methods

- **Local moves**: remove, add arc, revert direction (test for acyclicity!)
- Due to factoring property only local score evalution needed
- **Hill climbing**: select a random node, calculate best move around this node
- For greedy methods **multiple start** from random networks important!
- **MCMC-MH sampling** from posterior of network probabilities using local moves

Constraint based: PC algorithm

- Start with complete graph
- Determine conditional independencies
- Remove edges between conditionally independent variables
- Determine v-structures (unmarried parents), they provide directions on edges
- Propagate direction of edges

TetradIV implementation (with many more algorithms)

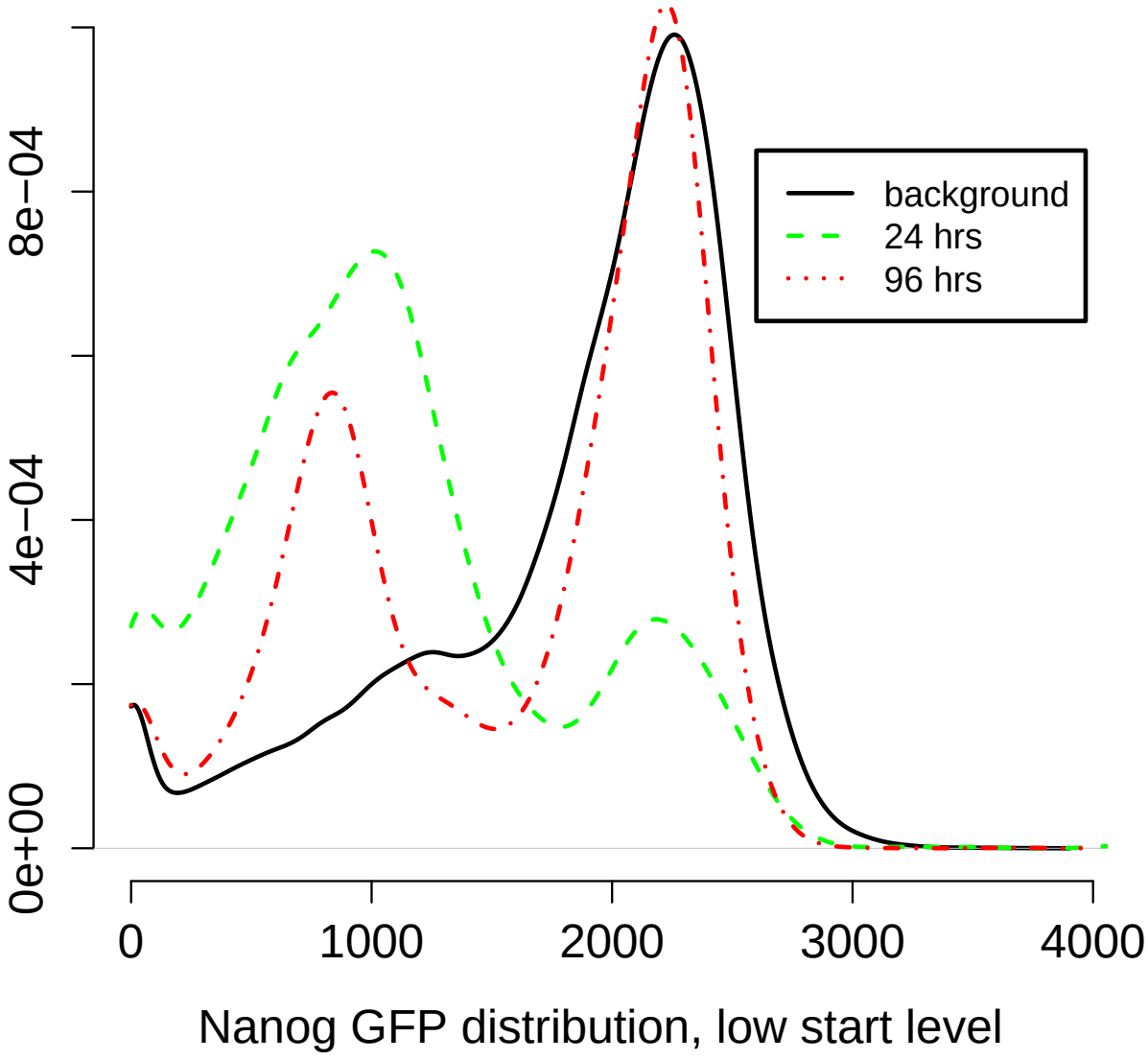
Application: differentiation of mouse ES cells

- **Pluripotent:** can differentiate into cell types
- **NOS network** (Nanog, Oct4, Sox2) main TFs which promote and maintain pluripotency (some other genes modifiers, eg Wnt)
- **Low or very high Oct4**: loss of pluripotency
- **Low Nanog**: high rate of differentiation
- Little known about topology or dynamics of NOS network

Flow cytometry

- Fluorescent protein (FP) as reporter
- Integration of FP into genome
- Estimate induction of protein from FP intensity
- 1000s of single cell measurements in one sample
- **Easy**: intensity distribution in a sample
- **Complicated**: tracking of single cells
- **Sorting**: pick out low or high intensity subpopulation and reculture

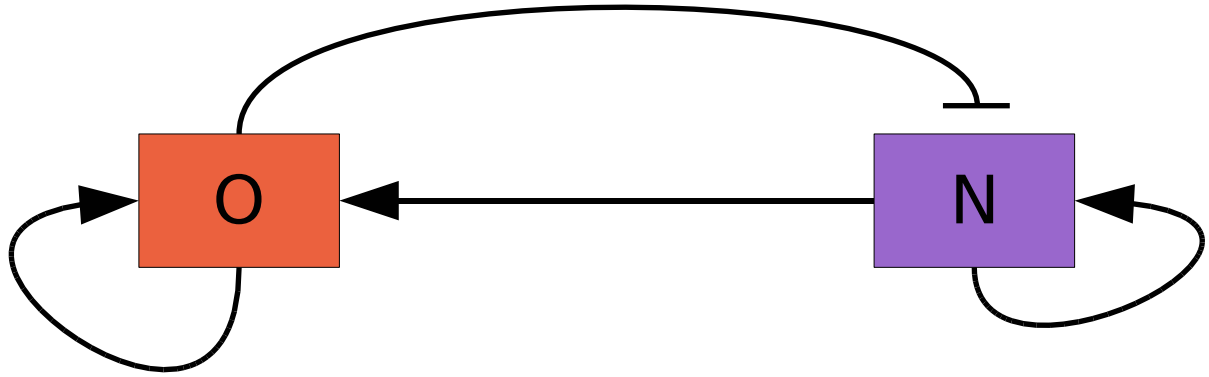
FACS sorting for low-Nanog ESCs



Low-Nanog cells with GFP reporter for Nanog binding cultured for 24 hrs and 96 hrs

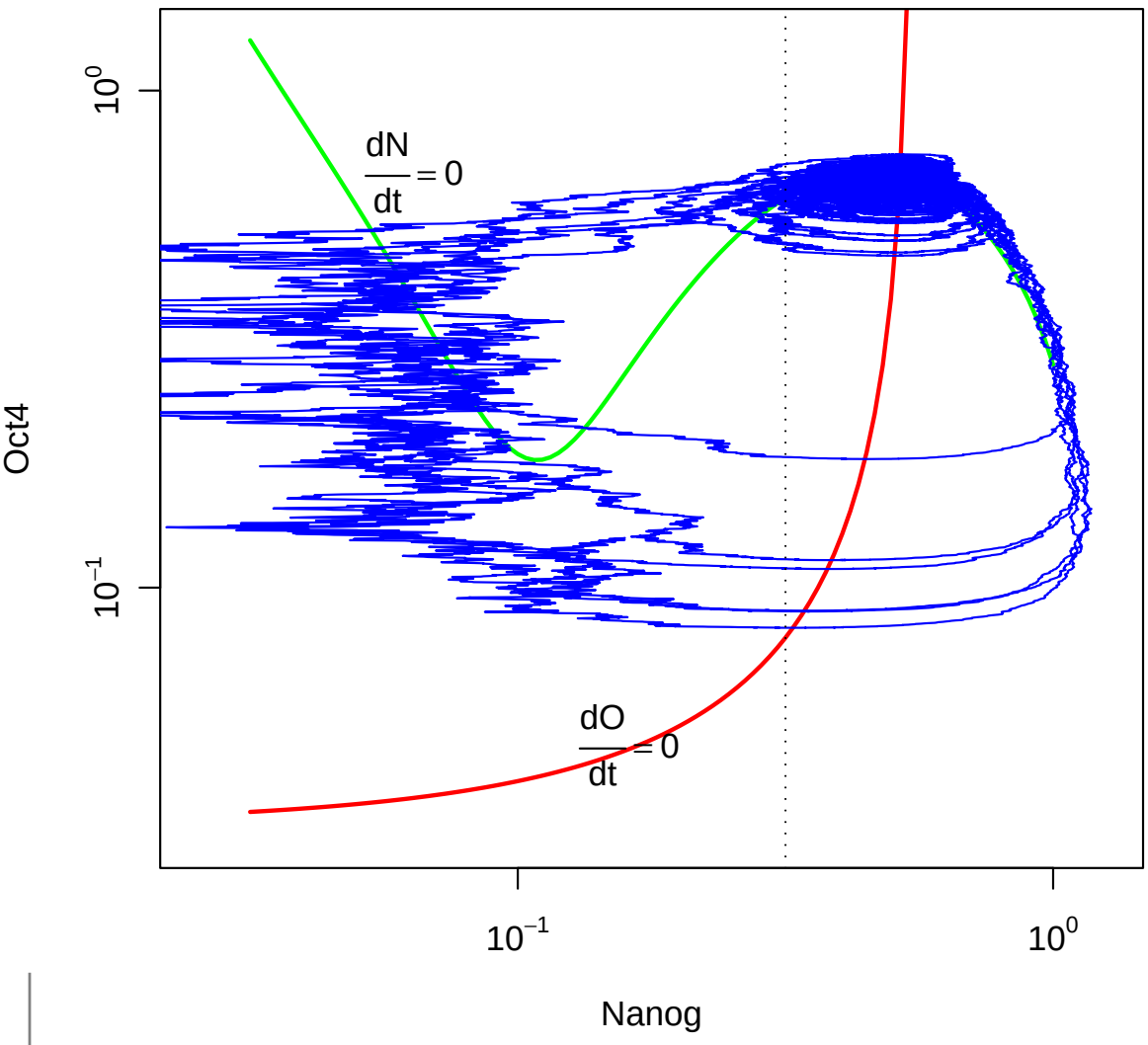
Experiments by Tibor Kalmar, Penny Hayward, and Alfonso Martinez-Arias (Cambridge University)

Stochastic DE model for Nanog-Oct4



$$dN = (\alpha_N + \frac{\beta_n N^2}{k_N^2 + N^2} - \frac{\delta O^{1.5}}{k_O^{1.5} + O^{1.5}} N - \gamma_N N)dt$$
$$+ D_1\, dB_1(t)$$
$$dO = (\alpha_O + \beta_O ON - \gamma_O O)dt + D_2\, dB_2(t)$$

Trajectory of Nanog-Oct4

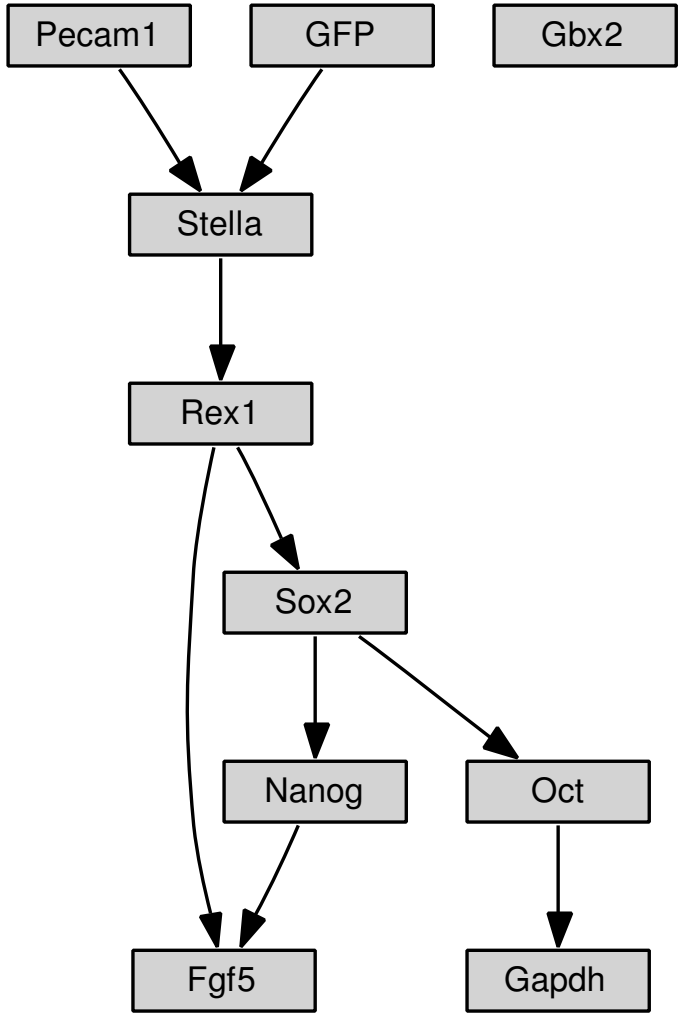


There is a stable equilibrium at the intersection of nullclines

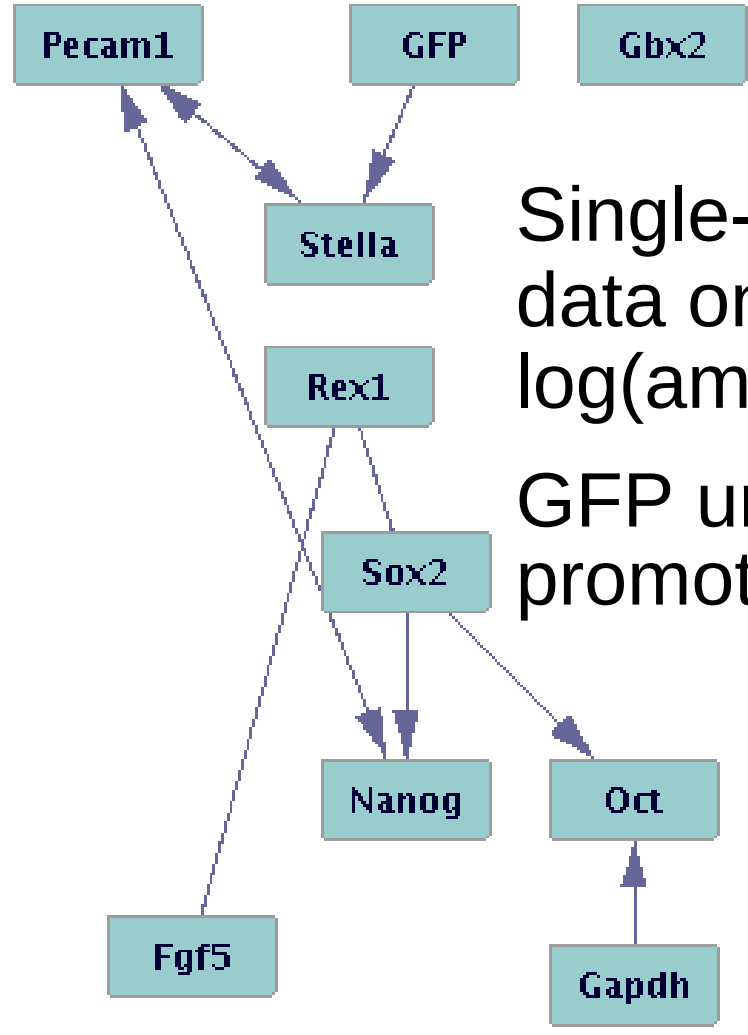
Noise pushes system away from equilibrium

A certain percentage of cell population is kept at low Nanog, susceptible to differentiation signals

Other players in Nanog system?



Bayesian score



PC (Tetrad IV)

Single-cell PCR data on 82 cells, log(amount RNA)

GFP under Stella promoter control

Bayesian networks for biological systems

Plus:

- Mature theory, easy to apply
- Efficient fast implementation of inference algorithms possible
- Easy to communicate to biologists
- **Causal interpretation** for biological control mechanisms **not too farfetched**

Minus:

- No feedback loops
- No simple (Bayesian) scoring method when latent variables are present

Schistosomiasis

- Caused by parasite *Schistosoma mansonii*
- Effects: fever, inner bleeding, liver damage
- 200 Mill at any one time, 1 Million death/year, widespread in tropical subtropical, Africa, Brazil
- Complex life cycle: larvae hatch from eggs in water, grow in snails, released in water again, picked up through skin by humans, through bloodstream in lungs, liver, vessels around gut, lay eggs, excreted
- Hard to recognise by immune system
- Treatment: praziquantel attacks worm tegument which makes it easily recognisable by immune system

Immune response

- Macrophage digest (bits of) invader, present it to T-helper cells
- T-helper cells sensitive to presented antigen, multiply and release cytokines: interleukines (IL)
- Interleukines (especially IL-4, IL-5, IL-6) stimulate B-lymphocyte production
- B-lymphocytes produces immunoglobulines, IgG, IgM, IgE (Y shaped)
- Igs stick to parasite, enable formation of membrane attack complex, or attract macrophages with cytolytic activities
- Some T-helper cells go in resting state for quick future response to infection (acquired immunity)

Data

- 460 infected individuals in study in Uganda (lab of David Lunn, Cambridge)
- Recorded: age, sex, infection intensity as measured by egg count in faeces, level of IgE
- Treated with drug that attacks worm tegument, triggering strong immune response
- Cytokines, interleukines, IGs measured 24h after treatment (stage A)
- Cytokines, interleukines, IGs measured 9 weeks after treatment (stage B)
- Infection intensity after 2 years by egg count
- **Large number of missing data** (20% of entries)

Structural Equation Model (SEM)

Variables x and regression coefficients A

$$x = Ax + \epsilon, \quad \epsilon \sim N(0, \Psi)$$

Might have **circular dependencies** among variables (nonrecursive). Some variables might be **latent** or unobserved.

Assume $(I - A)$ invertible, then

$$x = (I - A)^{-1}\epsilon$$

with covariance

$$\mathrm{Var}(x) = (I - A)^{-1}\Psi((I - A)^{-1})^T$$

Estimation

With J projection matrix and Jx observed variables, covariance induced on data is

$$C = \mathrm{Var}(x) = J(I - A)^{-1}\Psi((I - A)^{-1})^T J^T$$

ML minimisation of Gaussian $-2\log$ likelihood

$$\log|C| + \sum_i x_i^T C^{-1} x_i + \mathrm{const} = \log|C| + \mathrm{trace}(SC^{-1}) + \mathrm{con}$$

where S is the covariance matrix for centered data

Model fit, comparison

Model chi-square $\chi^2_M = (N-1)(-2\,\text{loglik})$

Estimation for **misspecification** of model

$$\hat{\delta} = \chi^2_M - \text{df}$$

Popular measures of model fit:

$$\text{RMSEA} = \sqrt{\hat{\delta}/(\text{df}(N-1))}$$

$$\text{CFI} = 1 - \hat{\delta_M}/\hat{\delta_B}$$

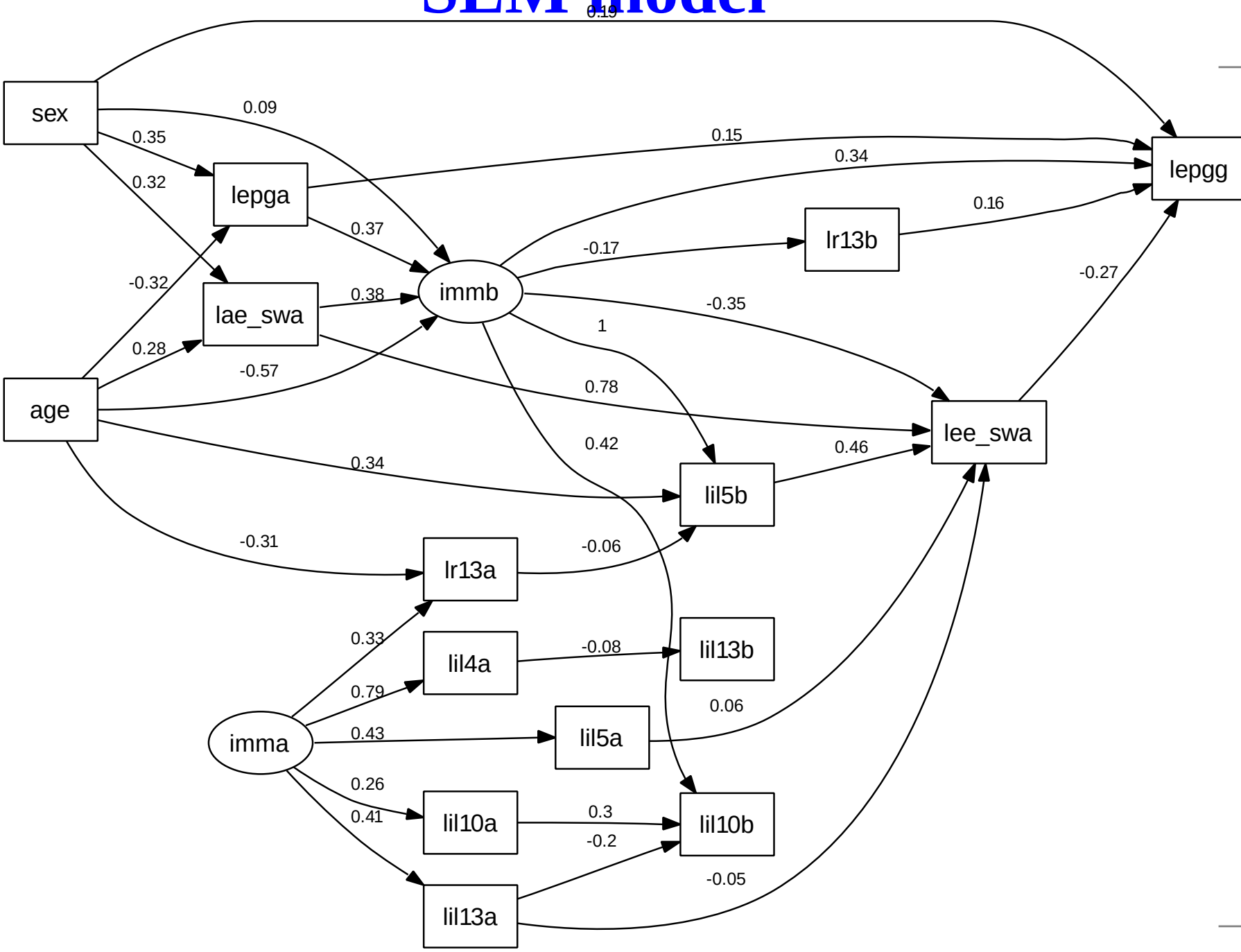
$\hat{\delta_B}$ for the base model (no arcs).

Also AIC, BIC etc

Estimation for Schistosoma data

- Time order on variables: pretreatment, 24h post treatment, 9 weeks post treatment, 2 years post treatment
- Prior knowledge on some expected relationships
- Latent variables of major interest, general immune response 24h (IMMA) and 9 weeks (IMMB)
- Missing data imputed by multiple imputation (MICE): random generation of missing values from predictive distribution based on regression on other variables
- All estimates on collection of random data tables, need to be summarised properly

SEM model


sex
age
lepga
lae_swa
immb
imma
lepgg
lr13b
lee_swa
lil5b
lr13a
lil4a
lil5a
lil10a
lil10b
lil13a
lil13b
0.19
0.09
0.35
0.32
0.15
0.34
0.16
0.37
-0.17
-0.27
-0.32
0.38
-0.35
1
0.28
-0.57
0.78
0.42
0.46
0.34
-0.31
-0.06
0.33
-0.08
0.79
0.06
0.43
0.26
0.41
0.3
-0.2
-0.05

Traditional vs Bayesian approach

- Signifiance via p-values with special heuristics for multiple imputation, probabilities preferred
- Proper credible intervals preferred to confidence intervals
- Model comparison via heuristic fit criteria, uncorrected ML unsuitable for (automated) model search
- Missing values are a pain, need to be imputed before SEM construction

Bayesian approach

SEM model

$$X = AX + \epsilon' 1, \quad \epsilon \sim N(0, \mathrm{diag}(\psi_1, \ldots, \psi_k))$$

with priors

$$(A_{(k)})_{P_{(k)}} \sim N(0, \epsilon_k \lambda I), \quad X^{(i)}_{H^{(i)}} \sim N(0, \Phi_{H^{(i)}})$$
$$\psi_k \sim \mathsf{Gamma}(\alpha, \beta), \quad \Phi \sim \mathsf{Wishart}(R, \rho)$$

$H^{(i)}$ is a binary indicator vector which value/variable is missing in data column i in X. Similarly, indicator vector $P^{(k)}$ for predictors of kth variable, F indicator for fixed values in A.

Conjugate analysis for coefficients A

$$X^P_{(k)} = X_{P_{(k)},\bullet}, \quad V_{A,k} = (\lambda^{-1} I + X^P_{(k)} (X^P_{(k)})')^{-1}$$

$$m_{A,k} = V_{A,k} X^P_{(k)} (X_{(k)} - (A_{(k)})_{F_{(k)}} (X_{(k)})_{F_{(k)}})'$$

$$(A_{(k)})_{P_{(k)}} \sim N(m_{A,k}, \psi_k V_{A,k})$$

Posterior for ψ_k^{-1}

$$\beta_k = \beta + \frac{1}{2}(X_{(k)}(X_{(k)})' - m'_{A,k} V_{A,k} m_{A,k})$$

$$\alpha_k = \alpha + n/2$$

$$\psi_k^{-1} \sim \text{Gamma}(\alpha_k, \beta_k)$$

Conjugate analysis for missing values in X

For the posterior for X note that

$$(I - A)_{\bullet,H^{(i)}} X^{(i)}_{H^{(i)}} = -(I - A)_{\bullet,1-H^{(i)}} X^{(i)}_{1-H^{(i)}} + \epsilon' 1$$

it is easy to see that therefore

$$V_{X,i} = (\Phi^{-1}_{H^{(i)}} + ((1 - A)_{\bullet,H^{(i)}})' \Psi^{-1}_{H^{(i)}} (1 - A)_{\bullet,H^{(i)}})^{-1}$$

$$m_{X,i} = -V_{X,i} ((1 - A)_{\bullet,H^{(i)}})' \Psi^{-1}_{H^{(i)}} (1 - A)_{\bullet,1-H^{(i)}} X^{(i)}_{1-H^{(i)}}$$

$$X^{(i)}_{H} \sim N(m_{X,i}, V_{X,i})$$

For Φ we use all the X data

$$\Phi^{-1} \sim \text{Wishart}(XX' + R, n + \rho)$$

Posterior for arcs

Adding new parents to a node k is expanding the basis of the regression.

Standard conjugate regression analysis gives posterior odds ratio

$$r_k = \frac{|V_{A,k,1}|\,\beta_{k,1}^{-\alpha_{k,1}}\,\Gamma(\alpha_{k,1})}{|V_{A,k,0}|\,\beta_{k,0}^{-\alpha_{k,0}}\,\Gamma(\alpha_{k,0})}\,\frac{1}{\pi^{(p_1-p_0)/2}}$$

where $p_1 - p_0$ is the number of added predictors (arcs)

Gibbs sampler

- Initialise A, X, structure (eg from traditional SEM analysis), then repeat:
- Draw coefficients A from posterior conditioned on structure (P matrix) and complete data X
- Draw missing values or hidden variables from posterior conditioned on structure and A
- Draw individual arcs from posterior conditioned on rest

Results on SEM structure very similar to previous analysis, similar means and sds.

$P(\text{IgE(B)} \rightarrow \text{final egg count}) = 0.87$
(p-value around 0.045)

Feedback loops

$$x_1 = ax_2 + \epsilon_1$$
$$x_2 = bx_1 + \epsilon_2$$

Real world interpretation: after random ϵ is set, a rapid equilibrium reached

Schisto study: ϵ random but persistent influences for each individual, on a slower time scale than equilibrium of components of immune system

Problem with simple regression: predictor (eg x_2) is correlated to error ϵ_1 via x_1

Simple regression

$$B = (I - A)^{-1} = \begin{pmatrix} 1 & -a \\ -b & 1 \end{pmatrix}^{-1} = \frac{1}{1-ab} \begin{pmatrix} 1 & a \\ b & 1 \end{pmatrix}$$

$$\mathrm{var}(X) = B \begin{pmatrix} \sigma_1^2 & 0 \\ 0 & \sigma_2^2 \end{pmatrix} B^T$$

$$= \frac{1}{(1-ab)^2} \begin{pmatrix} \sigma_1^2 + a^2\sigma_2^2 & b\sigma_1^2 + a\sigma_2^2 \\ b\sigma_1^2 + a\sigma_2^2 & b^2\sigma_1^2 + \sigma_2^2 \end{pmatrix}$$

$$\hat{\beta}_1 = \frac{\mathsf{cov}(x_1, x_2)}{\mathsf{var}(x_2)} = \frac{b\sigma_1^2 + a\sigma_2^2}{b^2\sigma_1^2 + \sigma_2^2} = \frac{b+a}{b^2+1}$$

Simple regression doesn't work with feedback loops!

Conclusions

- Networks immensly helpful for understanding complex biological control processes
- In practice not too much difference between frequentist and Bayesian approaches
- Main headache in (static) network modelling: **circular dependencies**